Amphioxus genome supplementary materials

Supplementary Note 1. Additional amphioxus background information

The importance of amphioxus in chordate evolution is celebrated in "The Amphioxus Song" written by Philip Pope. A RealAudio formatted performance of the song by folk singer (and marine biologist) Sam Hinton is available at
http://www.sandiegohistory.org/audio/hinton/amphioxus.ram.
Lyrics (and further historical information) can be found at
http://www.molecularevolution.org/resources/amphioxus/

"The Amphioxus Song" or, "It's a Long Way From Amphioxus"

Lyrics by Philip H. Pope, 1921. (sung to the tune of "Its a Long Way to Tipperary")

A fish-like thing appeared among the annelids one day.
It hadn't any parapods nor setae to display.
It hadn't any eyes nor jaws, nor ventral nervous cord,
But it had a lot of gill slits and it had a notochord.

Chorus:
It's a long way from Amphioxus. It's a long way to us.
It's a long way from Amphioxus to the meanest human cuss.
Well, it's goodbye to fins and gill slits, and it's welcome lungs and hair!
It's a long, long way from Amphioxus, but we all came from there.

It wasn't much to look at and it scarce knew how to swim,
And *Nereis* was very sure it hadn't come from him.
The mollusks wouldn't own it and the arthropods got sore,
So the poor thing had to burrow in the sand along the shore.

He burrowed in the sand before a crab could nip his tail,
And he said "Gill slits and myotomes are all to no avail.
I've grown some metapleural folds and sport an oral hood,
But all these fine new characters don't do me any good.

(chorus)

It sulked awhile down in the sand without a bit of pep,
Then he stiffened up his notochord and said, "I'll beat 'em yet!
Let 'em laugh and show their ignorance. I don't mind their jeers. *
Just wait until they see me in a hundred million years. *

My notochord shall turn into a chain of vertebrae
And as fins my metapleural folds will agitate the sea.
My tiny dorsal nervous cord will be a mighty brain
And the vertebrates shall dominate the animal domain.

(chorus)

* note -- the two lines marked by asterisks are not the original words, which are:
I've got more possibilities within my slender frame
Than all these proud invertebrates that treat me with such shame.

Supplementary Note 2. Genome sequencing and assembly

2.1 BAC library.

A Bacterial Artificial Chromosome (BAC) library (CHORI-302) was constructed by Chung-Li Shu and Kazutoyo Osoegawa in Pieter de Jong's laboratory, BACPAC Resources, Children's Hospital Oakland Research Institute. This library provides an estimated 17-fold coverage of the genome. The average size of the inserts in the library is 142 kb. The library is available through Children's Hospital Oakland Research Institute (http://bacpac.chori.org/amphiox302.htm). BAC clones were end sequenced at RIKEN-GSC and JGI-Stanford Human Genome Center and these sequences are deposited in the NCBI Trace Archive.

2.2 Genome assembly and validation.

Supplementary Table S4 summarizes the whole genome shotgun libraries and data set. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession ABEP00000000. The versions described in this paper include the first version, ABEP01000000, and Version 2, ABEP02000000.

Assembly Versions 1 and 1.5: Haplotypes assembled apart

The shotgun reads were initially assembled using JAZZ [1]. We estimated that the haploid genome size is approximately 500 Mbp, and that the two haploid genomes of the donor animal are represented independently in the assembly as separate haplotypes, each approximately 85% complete, with at least one locus assembled for more than 95% of genomic loci (see below).

Comparison of the haplotypes by aligning the nucleotide sequence of the reconstructed genome fragments highlighted a number of likely mis-assemblies (global mis-joins) in the initial scaffolds. A manual search for potential mis-joins based on an assumption of long-range co-linearity of the two haplotypes identified a total of 331, or about one every 2.5 Mb of assembled sequence. When an assumption of haplotypic co-linearity implied a mis-assembly in one of two scaffolds, both scaffolds were broken. Short internal inversions were allowed if flanked by co-linear regions. The resulting fragments were ordered and oriented with JAZZ using BAC- and fosmid-end read pairs to link them. Roughly one third (129) of the identified breaks came back together in the new assembly, and 202 remain split. While this approach may lead us to break correctly assembled regions where the assumption of long-range co-linearity is violated by large-scale structural polymorphisms, the predominance of short inversion polymorphisms (of order 1kb) over intermediate length scale differences (of order 10kb - 100 kb) indicates that structural variation at the scaffold length scale may be rare.

The resulting assembly (version 1) contains a total of 3,032 scaffolds (total length 923,340,504 bp) composed of 81,073 contigs (total length of 831,199,399 bp). Half of the total contig sequence is contained in the largest 9,177 contigs (contig N50), which are each longer than 25,666 bp (contig L50). Half of the total scaffold sequence is in the largest 174 scaffolds, which are each longer than

1.6 Mb. This assembly as available for download and interactive examination at http://www.jgi.doe.gov/Brafl1/Brafl1.home.html and deposited at GenBank as accession ABEP01000000. There were 332,545 unplaced reads with a total trimmed length of 194 Mb, or 4.4% of the 7.6 million-read starting data set. Supplementary File sf1.txt contains a list of the read identifiers for these unplaced reads; their sequences are available from the NCBI Trace Archive, along with the complete WGS data set.

Following automatic and manual annotation of gene models (Supplementary Note 3), examples of mis-assemblies in version 1 of a specific type were identified: allelic pairs of genomic sequence tens to hundreds of kilobases in length are sometimes assembled in tandem in the same scaffold, rather than on different scaffolds. Mis-assemblies of this type ("stutters") were identified in the largest 1000 scaffolds, spanning 97% of the assembly, by visual inspection of the sequence self-alignment dot plot and the 8kb mate-pair assembly layout. 574 stutters spanning ~6% of assembly version 1 were identified and removed from these scaffolds to create assembly version 1.5, with all break points but one chosen to avoid overlap with gene models to preserve annotations.

Assembly version 1 allelic structure

In preliminary shotgun assemblies, we noted that previously cloned and sequenced *Branchiostoma floridae* genomic sequences typically aligned to two locations in the assembly. This could in principle be explained by (a) separate assembly of two divergent haplotypes from the diploid donor, (b) a recent whole genome or large-scale segmental duplication in amphioxus, or (c) an aberrantly allotetraploid mutant individual donor genome. A recent genome duplication affecting the entire species (hypothesis (b)) can be ruled out because numerous FISH studies with different individuals have identified only single physical positions on the chromosomes for markers such as the HOX and paraHOX clusters. In addition to being *a priori* very unlikely, the amphioxus-specific allotetraploidy hypothesis (c) does not agree with the observed distribution of the number of single nucleotide differences in windows along the two haplotypes (see Supplementary Note 4 below). We therefore conclude that the assembly typically represents the two alleles at a given genomic locus with independent sequences.

To estimate how commonly two haplotypes are collapsed together due to locally low heterozygosity (or any other reason), we examined the depth of coverage of reads placed by the assembler in (1) the 200 largest scaffolds and (2) all the smaller scaffolds in the version 1 assembly. The distributions (shown in Supplementary Figure S22) are peaked at a depth of ~5, and do not show a peak near double this depth as would be expected for regions that represent the collapse of two haplotypes onto a single sequence. This distribution is, however, quite broad, and it is possible that there is some contribution from local regions assembled at twice the nominal depth of coverage for each haplotype. 4% of the length of these scaffolds is spanned by segments composed of two or more consecutive 5kb bins with mean read coverage exceeding 8X suggesting that this fraction of the assembly may represent regions where both alleles are collapsed in to a single consensus.

Unassembled reads

Since the assembler is by design biased against assembling potentially repetitive sequences appearing at unusually high depth in the shotgun reads, we tested the hypothesis that genomic

regions of low heterozygosity have been left unassembled. We compared the distribution of the number of "overlaps" (sequence alignments to other reads in the shotgun data set) for all reads to that for the unassembled reads (Supplementary Figure S24). Overlaps are semi-global alignments of at least 85% identity of at least 250 bp in length. The distribution for all reads is peaked around 10, as expected for a random shotgun at 5X coverage (each read has ~5 alignments to the left, and ~5 to the right). Rather than having unusually many overlaps, the distribution for the unplaced reads shows that the vast majority of these reads have unusually few overlaps. This is a result either of sequence contamination, cloning and/or sequencing bias, or of a high prevalence in these reads of 16-mers that are found at extremely high copy number in the data set, representing highly repetitive sequences rather than two haplotypes.

Completeness

Of 11,628 expressed sequence tag (EST) cluster assemblies (see Supplementary Note 3) with an apparently complete open reading frame, 95.3% had at least one alignment of at least 80% identity, spanning more than 50% of its length to the assembly, indicating that the assembly is approximately 95% complete in capturing at least one allele of known genes.

Genome size estimate

Since 69% (15,123) of the estimated total of 21,955 genes are represented by an allelic pair of gene models (See Supplementary Note 3), assuming that the fraction missing from the assembly of each haplotype is approximately independent, we estimate that each haplotype is (0.69)^(1/2), or ~83% complete. This would lead us to expect that 0.17 x 0.17 or ~3% of loci would be missing by chance from both haplotypes, which is close to, and consistent with, the observed 4.7% of EST clusters that do not align to the assembly. Based on the total length of the assembled contigs, and the estimate that each haplotype is 83% complete, the genome size of the assembled portion of the genome is estimated at 831 / (2 x 0.83) = 500 Mb.

In light of these analyses, the assembly appears to capture two independent haplotypes with substantial completeness, in a manner that is not obviously biased by local fluctuations in heterozygosity.

Assembly Version 2: Allelic redundancy removed

As an additional resource we created a non-redundant representation (Version 2) of the genome sequence which is a mosaic of the two haplotypes found in assembly Version 1. The 1000 longest scaffolds of assembly version 1.5 were aligned to one another using MegaBLAST[2], and manually curated into 398 connected sets of allelic scaffolds, examples of which are shown in Supplementary Figure S63. In this process, 132 potential mis-joins were identified in Version 1.5 scaffolds and broken. Each of the 398 sets of allelic scaffolds was merged into a non-redundant representative sequence which is a mosaic of the two haplotypes, created by concatenating segments of the scaffolds in the set. The mosaic was constructed to switch between haplotypes only between gene models, and to minimize the number of transitions between haplotypes. Among the possible tilings with the minimum number of transitions, we selected that which minimizes the total length of sequence gaps in the merged sequence. This method is similar in spirit to that applied to the *Ciona savignyi* genome by Small *et al* [3]. Assembly version 2 spans 522 Mb, with scaffold N/L50 = 62 / 2.6 Mb and contig N/L50 = 4916 / 28kb. The net assembly length is slightly longer than the estimated haploid genome size, which could be accounted for by contributions from internal assembly gaps, residual allelic redundancy and haplotype-unique sequences. A mapping of the gene models created for assembly version 1 (Supplementary Note 3) is available from the amphioxus genome portal in gff format.

Assembly Validation

To assess the accuracy of the shotgun assembly, we compared assemblies version 1 and 2 to the 33 clones from CHORI-302 BAC library that were sequenced as part of the genome project (Supplementary Note 2.1). These clones have a total length of 5586473 base pairs, and include two allelic pairs. (Supplementary Table S17)

To asses the sequence fidelity at the base-pair level, we divided the 33 clones into 5568 non-overlapping fragments of 1000 base pairs (discarding 33 incomplete bins from the ends of the clones), and aligned each to assembly version 1 using BLAT[4]. The total alignment length (matches + mismatches) of the best hits of each 1kb query against the assembly spans 99.0% of the query sequences, with a mismatch frequency of 0.25 %. 26,041 bases (0.47%) of the query sequences were spanned by alignment gaps, and represent insertions in the reference sequence relative to the assembly.

Since assembly 1 represents the two haploid genome copies separately, and each haplome is largely but not entirely complete, some reference sequences have their best alignment to the other haplotype, and consequently these rates of discrepancy include contributions from sequencing and assembly errors as well as from polymorphisms. To estimate the sequencing error rate alone, we considered only those reference sequence bins representing genomic loci assembled from both haplotypes in assembly version 1 (bins with exactly two BLAT alignments to assembly version 1 which span 75% of the length of the bin, and aligning with at least 85% identity). Within the best BLAT alignments of these reference segments, 87% have one or fewer mismatches per aligned 1000 bp (PhredQ >= 30), and the mean mismatch rate is 0.15%.

To assess the accuracy of the assembly at longer length scales, we aligned the full length sequence of each of the same 33 reference sequence clones to assembly version 2. The best alignment of each to the assembly is represented as a dot plot in supplemental figures S30-S62. The alignment of each clone to the assembly spans the entire length of the clone or nearly so. This process detected no apparent global mis-assemblies, indicating fewer than 1 error of this type per 5 Mbp, or fewer than 100 such errors in the whole genome should be expected. The comparisons reveal 52 local sequence stutters in the assembly relative to the finished clones ranging in size from 1 to 10 kb, indicating a frequency of such errors of approximately 1 per 100 Kbp.

2.3 *De novo* identification of repetitive sequences in the assembly

A set of 562,796 vector-and-quality-trimmed fosmid-end read sequences (roughly ~0.6X coverage) were aligned to the assembly with BLAT[4]. Alignments were filtered to preserve only those which span 90% of the length of the read with at least 85% identity. We collected 11,483 stretches of the genome that were (a) covered throughout by at least 6 reads and (b) contained a region where the local depth of coverage is 16 or greater. These were made non-redundant by assembly with PHRAP (Phil Green, Unpublished) to yield 75 repetitive elements contained within the assembly, ranging from 104 bp to 4,006 bp in length (mean 545 bp). These sequences are contained in Supplementary File sf2.fasta.

2.4 Transposable elements in the amphioxus genome

General description of amphioxus TEs

Transposable elements (TEs) constitute >28% of the amphioxus genome (Supplementary Table S5) and belong to >500 families. In terms of their bulk contribution to the genome size, DNA transposons are twice more abundant than retrotransposons. In the oldest families of amphioxus TEs, copies of fossilized transposons are 19-22% divergent from their consensus sequences (*e.g.*, Harbinger-N5_BF and Harbinger-N15_BF). CR1 non-LTR retrotransposons form the most diverse superfamily of amphioxus TEs, composed of more than 100 families (elements from different families are less than 75% identical to each other). Most non-autonomous non-LTR retrotransposons, also known as SINEs, identified in amphioxus have been retrotransposed by reverse transcriptases and endonucleases encoded by CR1 elements, including SINE2 and SINE3 with their internal pol III promoters derived from tRNA and 5S rRNA, respectively (these SINEs and CR1s share common 3'-terminal portions).

Methods

Transposable elements were identified using WU-BLAST (http://blast.wustl.edu) and its implementation in CENSOR (http://girinst.org/censor/). First, we detected all fragments of the amphioxus genome coding for proteins similar to transposases, reverse transcriptases, and DNA polymerases representing all known superfamilies of TEs. The detected DNA sequences have been clustered based on their pairwise identities using BLASTclust (standalone NCBI BLAST). Each cluster has beet treated as a potential family of TEs described by its consensus sequence. The consensus sequences were built automatically based on multiple alignments of the cluster sequences expanded in both directions and manually modified. Using WU-BLAST/CENSOR we identified fragments of the amphioxus genome similar to the consensus sequences that were considered as copies of TEs. Second, given the identified consensus sequences, we detected automatically insertions longer than 50-bp present in the identified copies of the protein-coding TEs. The insertions have been treated as potential TEs, clustered based on their pairwise DNA identities and replaced by their consensus sequences built for each cluster. After manual refinements of the consensus sequences, the identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously. Identified TEs are deposited in Repbase[5].

Supplementary Note 3. Annotation of protein coding genes

3.1 Expressed sequence tag sequencing.

cDNA libraries for amphioxus were prepared from gastrula and neurula stage embryos, and from larvae as described by [6], and a single library was created for *Petromyzon marinus* (sea lamprey). Approximately 32,000 expressed sequence tags (ESTs) were attempted from each library. (Supplementary Table S6).

3.2 Gene prediction, functional annotation and quality control.

The JGI Annotation Pipeline was used for annotation of the amphioxus v1.0 assembly described here. The pipeline includes the following steps: (1) repeat masking, (2) mapping ESTs, full length cDNAs, and putative full length known genes, (3) gene structure prediction using several methods, (4) protein functional annotation using several methods, and (5) combining gene predictions into a non redundant representative set of gene models, which are subject to genome-scale analysis. The genomic sequence, predicted genes and annotations of amphioxus, together with available evidence, are available at the JGI Genome Portal (www.jgi.doe.gov/Amphioxus) and from GenBank under accession number ABEP01000000.

Transposons were masked using RepeatMasker [7] tools and a custom library of manually curated repeats (see above Supplementary Note 2.4). 480,070 ESTs were clustered into 77,402 consensus sequences and both individual ESTs and consensus sequences were mapped onto genome assembly using BLAT[4].

Gene predictors used for annotation of amphioxus v1.0 included *ab initio* FGENESH [8], homology-based FGENESH+ [8], homology-based GENEWISE [9], and EST-based ESTEXT (Grigoriev, unpublished, available upon request).

A set of 4,272 genes derived from EST clusters with a putative full length open reading frame (ORF) was directly mapped to the genomic sequence to build gene models. FGENESH was trained on this set to achieve sensitivity and specificity of 78% and 78%, respectively on exon prediction. To generate homology-based gene models, proteins from the NCBI NR database were aligned against genomic sequence using BlastX[10]. High quality seed proteins were then used to build models using FGENESH+ and GENEWISE. GENEWISE gene models were then filtered to remove models with frameshifts and internal stop-codons and extended to include start and stop codons where possible. FGENESH, FGENESH+ and GENEWISE gene models were then processed using ESTEXT to correct them according to splicing patterns observed in available ESTs and to extend 3' and 5' UTR of the genes when possible. Annotation summary can be found in Supplementary Table S7A

All gene models were annotated by homology to other proteins from NCBI NR, SwissProt and KEGG databases. Using InterproScan [11] we predicted proteins domains. Using these sources of information, annotation of each protein was mapped to the terms of Gene Ontology [12], KOG clusters of orthologs [13], and KEGG pathways [14].

The large set of all predicted models was reduced to a non-redundant set of 50,818 representative models (Filtered Models), where every (haploid) locus is described by a single best gene model according to the criteria of homology and EST support. For this set of representative gene models we assigned 27,600 peptides (54%) to Gene Ontology (GO) terms [12] , 29,030 peptides (57%)

proteins to KOG clusters [13], and 21,060 peptides (41%) to KEGG pathways [14]. Supplementary Table S7B summarizes the functional characterization of the predicted genes.

3.3 Estimation of haploid gene number and identification of allelic pairs of gene models.

Co-linear blocks of highly-similar predicted genes in near-identical order in different locations in assembly version 1 where tabulated using the SEGMENTATOR algorithm, as previously described for identifying conserved gene order between species[15]. The threshold for similarity was a Smith-Waterman alignment p-score < 1e-20, and at most two intervening genes were allowed in constructing colinear segments.

31,813 gene models of the full set of 50,818 predictions participated in a conserved colinear block of a least two "rungs" (i.e. group of putative orthologs with representation on both haplotypes) after discounting tandemly repeated genes with similar sequence. To assess the number of genes represented in each colinear block, for each rung we counted the minimum of the number of genes participating from each haplotype as the number of genes. For example, if a predicted gene on one scaffold of the assembly is part of block of conserved gene order on another scaffold, where it hits two adjacent models, these two models are assumed to be spurious fragmentations of the corresponding allele, and only a single gene is counted. Similarly, if a tandemly repeated gene family shows three copies in one allele, and a greater number of gene models on the other allele, the greater number of models on the other allele is assumed, for counting purposes, to be due to gene modeling artifact.

Counting genes in this way, we compiled a list of gene models which represent 15,123 genes predicted from both alleles by the 31,813 models. This provides an estimate of the over-prediction due to gene fragmentation in the annotation of 5.2%, since 31,813(2 x 15,123) = 1.052. Of the remaining models, which are not confirmed by the occurrence of a similar model in consistent context in two haplotypes, we removed as likely artifacts and/or un-masked repetitive elements 12,173 models which had no identifiable homolog in another sequenced chordate genome. The majority of the gene models excluded by this criterion fall in (sequence-based) clusters of amphioxus gene models with between 30 and 1,244 similar sequences. This is consistent with their being dominated by repetitive elements. (For a description of the method of ortholog identification, see Supplementary Note 7.) The remaining 6,832 gene models represent genes captured from one allele only. Thus the estimated total gene number is 15,123+6,832= 21,955.

Supplementary file sf3.txt correlates gene models with their identified allelic pairs. In some cases, a single model on one haplotype is associated with multiple gene models on the other haplotype, due either to the one gene model being a fusion of genes correctly called as separate genes on the other side, or a single gene being fragmented into multiple genes on the other side.

Supplementary file sf4.txt lists the genes models which have apparent orthologs in other bilaterian genomes, but do not have an identified allelic pair in the annotation.

Supplementary Note 4. Analysis of allelic polymorphisms

Sequence divergence parameters (including Ka, Ks) of allelic gene pairs were estimated by the method of Yang and Nielsen[16,17] (Supplementary Figures 1 and 7). Identified allelic gene pairs were also used as anchors for aligning allelic genomic scaffolds with LAGAN[18].

To estimate the polymorphism rate at neutrally varying sites, we examined a high-confidence set of 3,381 allelic gene pairs (see Supplementary Note 3) which (a) were identified as allelic pairs with very high confidence, being within a strictly co-linear block of at least 10 identified allelic loci and (b) had at least 100 four-fold synonymous positions in the alignment between the two alleles. We found that 77,532 out of 801,442 sites (9.7%) differed between alleles, and that the median synonymous-site divergence between alleles was 9.0%.

To estimate the polymorphism rate of the two haplotypes across the whole genome, we compiled a sample of 100,000 randomly selected stretches of the genome assembly, each 2kb in length and free of any sequence gaps. These sequences were aligned to the entire assembly using BLAT [4] with default settings for nucleotide alignment. 23,041 of these had a best non-self (not overlapping the identical sampled sequence) hit to the assembly that (a) did not span any sequence gaps, and (b) spanned at least 1 kb of the sample query sequence. BLAT reported that alignment of this 46.08 Mb of sequence spanned 43.62 Mbp (Category "A" in Supplementary Figure S2), with 36.47 million matching aligned bases, and 1.41 million mismatches (Category "B"), indicating a genome-wide mean mismatch rate between haplotypes of 3.72%. The aligned sample sequences contained 5.74 Mbp in internal alignment gaps (Categories "C" and "D"). Of this, 2.98 Mb (6.83%) represents sequence not aligning because it has been inserted relative to a shorter sequence on the opposite haplotype, and the mean length of these insertions was 20 bp (Category "C"). The remaining 5.74-2.98=2.76 Mb represents additional unaligned sequence (not spanned by a BLAT pairwise alignment, but internal to the overall alignment of the allelic regions, Category "D"), which could be caused by short, local sequence inversion, not modeled by BLAT. The mean spacing of these indels and small inversions was one every 139 base pairs. In summary, between mismatches (3.72% of homologous nucleotide positions), insertions (6.83% of all nucleotide positions) and inversions or distinct sequence (6.33% of all nucleotide positions), we find that on average only 83.6% of base pairs in one haplotype have a homologous position with the same nucleotide in the other haplotype. If indels and micro-inversions are treated as point differences equivalent to mismatches, the combined rate of variation was 4.0%.

To characterize the variation in polymorphism rate over longer length scales, allelic gene pairs identified as described in Supplementary Note 3 were used as anchors for aligning allelic genomic scaffolds with LAGAN[18]. 1000 paired intergenic sequences, with mean length 100.7 Kb and median length 58 kb, were analyzed for the distribution of single base and indel differences. Supplementary Figures S4, S5, S6, S25, and S26 summarize the distributions of various measures of the sequence divergence observed between these aligned intergenic allelic sequences: number of mismatches in 50bp windows; autocorrelation of haplotype divergence at short short genomic distance; distribution of the lengths of ungapped alignments between haplotypes; size distribution of short indel polymorphisms; size distribution of longer polymorphic indels.

Supplementary Note 5. Deuterostome relationships

Supplementary Table S8 lists the gene set release version numbers used for each genome.

Deuterostome relationships.
Sets of orthologous genes were collected by grouping together mutual-best BLAST[10] hits between Nematostella vectensis (sea anemone) and each other gene set. Individual multiple alignments were created with CLUSTALW[19], manually reviewed, trimmed with Gblocks[20], and concatenated. 364 ortholog sets had representation from all genomes (alignment 1), and 1,090 had up to one missing (alignment 2). The alignments were analyzed by Bayesian and maximum likelihood methods using mrbayes[21,22] and PHYML[23]. Supplementary Note 5 contains more details of the data sources, data compilation and analysis.

Topology constraints

In all analyses of the concatenated protein multiple sequence alignment, we constrained the topology of the trees considered to include only those which respect the following constraints, which represent undisputed phylogenetic relationships: that the two tunicates (*Ciona* and *Oikopleura*) be placed together; that the two non-deuterostome out-groups (*Nematostella* and *Drosophila*) be placed together; that the five vertebrates be placed together, with topology (*Petromyzon* , ((*Takifugu*, *Gasterosteus*), (*Homo*, *Gallus*))).

Maximum likelihood analysis

We ran ProtTest v1.2.6[24] on alignment 1 to select the model of evolution for maximum likelihood analysis, which was carried out with PHYML v2.4.4[23]. The model used was: WAG[25] substitution matrix, 29% invariant sites and gamma-distributed rates with shape parameter 0.76, estimated using four rate categories. For both the identification of the maximum likelihood topology, and for the calculation of bootstrap support values, the likelihoods of all tree topologies satisfying the topological constraints was calculated, with branch length optimization turned on. Bootstrap support values indicate the percentage of 500 bootstrap replicates, constructed with SEQBOOT[17], which resolve the labeled nodes as shown in the maximum likelihood tree.

Bayesian analysis

Bayesian analysis was carried out with MrBayes v3.1.2[21,22] using the WAG substitution matrix, with the fraction of invariant sites and the shape parameter of the rate distribution allowed to vary. Twenty independent runs with random starting trees converged on the same topology and apparently stationary distributions of likelihood and rate parameters within 5000 Monte Carlo iterations. 100% of sampled trees over one million monte carlo iterations after this burn-in period had the topology shown in Figure 1 of the main text. The means of the posterior distributions for the shape parameter and invariant fraction were 0.744 and 0.277 respectively.

Oikopleura dioica gene set creation

707,762 whole genome shotgun reads from the *Oikopleura dioica* genome project (Seo, 2001 #294) were downloaded from the NCBI Trace Archive on 19 May 2006 and assembled with JAZZ[1] into 88 Mb of scaffold sequence, with N50 scaffold size of 38 Kbp. Human protein sequences from

ENSEMBL and *Ciona intestinalis* sequences from JGI were aligned to the assembled *Oikopleura* sequence by tblastn, and 8,911 gene models were created with Genomescan[26].

Supplementary Note 6. Intron evolution

Based on mutual-best Smith-Waterman alignment of protein sequences with human genes, a collection of 2,576 orthologous gene sets were constructed with representation from human, stickleback, fugu, chicken, amphioxus, *Nematostella*, sea urchin, *Ciona,* and the gastropod *Lottia gigantea* (see Supplementary Table S8). The starting gene sets were the same as those described in Supplementary Note 5, with the substitution of *Lottia gigantea* as a representative protostome outgroup in place of *Drosophila*, which has been shown to be highly derived with respect to intron conservation[27].

Performing ClustalW alignments and identifying high reliability intron splice sites (below) yields 5,337 splice sites for which a 9-way comparison can be made. If the *Oikopleura* gene models are also included, 1,508 usable splice sites are available for analysis in the 10 species alignment.

"Highly reliable" splice sites are defined as being within a gap free region flanked by fully conserved amino acids and at least 2 amino acids from the edge of such a region. Also, at least 5 of the 10 flanking amino acids on either side of the site should be full conserved or similar as measured by the BLOSUM62 matrix. Finally, no splice site should exist in any species within 4 amino acids from a highly reliable splice site. This last criterion removes seemingly different splice sites that may be due to problems in gene modeling due to ambiguous or cryptic splice sites.

PAUP 4b[28] was used to reconstruct the relationships among the 9 species using both simple parsimony and weighted parsimony in which intron gains were weighted as more costly (by a constant weight factor) than intron losses, for weight factors varying from 2 to 9. All these methods indicated strong support for the same consensus tree topology obtained from the sequence-based phylogeny, with parsimony bootstrap values for monophyletic vertebrates, olfactores and chordates greater than 95%.

We also analyzed the data sets with MrBayes[21,22], using a general binary character model which allows different gain and loss rates (although not different gain and loss rates on different parts of the tree, and therefore does not capture a striking feature of the data, namely the higher rate of intron turnover in the tunicates[27]). Within this model, there is also strong support for vertebrates, olfactores, and chordates, with these groups appearing as monophyletic in more than 99% of topologies sampled in the Bayesian run after apparent equilibrium is reached.

Parsimony analysis (weight=5) on the 1508, 10-species data set (with the highly divergent species *Oikopleura* included) gives high bootstrap support for monophyletic tunicates and olfactores, and 53% bootstrap support for monophylectic chordates, indicating that the placement of *Ciona* sister to the vertebrates in the analysis of the larger character matrix is unlikely to be a spurious result to due to long branch attraction.

In all intron analyses, *Nematostella* and *Lottia* were constrained to be the outgroups to the other (deuterostome) species, which is important since models of intron evolution in which gain and loss are treated differently were used. If no such constraint is imposed, the effect in both the parsimony

and Bayesian analyses is a consensus tree that still shows monophyletic vertebrates, olfactores, and chordates (all with > 99% bootstrap support in unweighted parsimony analysis), but *Lottia* and sea urchin are brought together with high support (96% in unweighted parsimony analysis).

Supplementary Note 7. Chordate gene families

Two methods were used to approximate the gene complement of the chordate ancestor by clustering genes in modern chordates according to their putative orthology relative to the chordate ancestor. Both methods gave approximately the same reconstruction, indicating that the results are robust to details of methods.

7.1 Chordate gene families reconstructed by c-score clustering

As in [27] we define the "C-value", for BLAST hit between peptides x and y (from genomes X and Y, respectively) as follows:

$$c(x,y) == b(x,y) / \max\{b(x,z) \text{ for } z \text{ in } Y, \text{ or } b(w,y) \text{ for } w \text{ in } X\}$$

where b(x,y) is the BLAST score between x and y.

Note that pairs (x,y) with c(x,y)=1 are mutual best hits (including degenerate cases), and c close to unity generalizes the concept of a mutual best hit. Gene families can be constructed by single link clustering. Supplementary Table S9 shows human-amphioxus clustering by this method; in this calculation, tandem gene families have been collapsed and are represented by the single gene in the tandem cluster with the highest c-score. The number of human-amphioxus clusters was constant at ~8,000 for C >~ 0.7.

7.2 Chordate gene families reconstructed by hierarchical "metazome" clustering

Genes from human, chicken, fugu, stickleback, *Ciona*, amphioxus, sea urchin, *Drosophila* and *Nematostella* (gene sets are listed in Supplementary Table S8) were clustered hierarchically to reconstruct ancestral gene sets for ancestral genomes as previously described in the context of reconstructing the ancestral eumetazoan gene set[27], with the following modification: when forming clusters to represent the genes of a given ancestor, instead of merging clusters based on mutual best hits (MBH) by BLAST of individual genes, we computed mutual best hits between the gene clusters already computed at the next highest nodes of the tree. For example, when considering the merger of tetrapod clusters with bony fish clusters to create ancestral jawed vertebrate clusters, we grouped BLAST hits by cluster, and identified mutual best hit relationships between clusters rather than considering separately the mutual best hits between human-fugu, chicken-stickleback, etc. This prevents "spiraling" due to the non-transitivity of the MBH relationship, in which, for example, a human gene Hs1 has MBH in fugu to Tr1, which has MBH in chicken Gg1, which has MBH in human to a different human gene Hs2, and so on.

By this method we defined 9,975 ancient chordate gene families, and identified gains of novel genes, and losses of ancestral genes on the chordate stem, and on the lineages leading to the major chordate sub-groups. Each family nominally represents the modern descendants of a single gene in the chordate ancestor. These families account for 13,401 human genes, 10,094 chicken genes, 14,286 stickleback genes, 15,766 fugu genes, 7,216 Ciona genes, and 13,610 amphioxus genes (after removing redundant alleles.)

Of the 9,975 gene families, 8,437 have at least one amphioxus member and at least one from Ciona and/or a vertebrate. These are the ancestral chordate gene families discussed in the main text. These correspond roughly to the human-amphioxus clusters with C-value ~ 0.7 in the method described in Supplementary Note 7.1. Of the 9,975 families, 859 appear to have been "lost" in amphioxus (or are missing, poorly predicted, highly divergent, or otherwise incorrectly omitted from the correct gene family) and 679 "lost" in the vertebrate/urochordate ("Olfactores") clade. There are 972 apparent vertebrate gains and 110 apparent olfactores gains.

The 679 apparent olfactores losses were estimated as follows: There are 826 apparent olfactores losses relative to the chordate clusters. 147 of these have amphioxus gene(s) with a c-score > .75 hit to human, suggesting that while not grouped together in the "metazome" clusters they may have orthologs that were missed. This leaves 679 apparent olfactores losses. Again, this set likely contains poor gene predictions and/or highly divergent sequences as well as bona fide losses.

2,427 of the 9,975 families lack Ciona genes. These may represent losses, but also might represent gene families that are present, but not placed in the correct cluster due to the length of the Ciona branch. We searched for possible Ciona members of these gene families as follows: 600 families have a human gene with an e-value < 10^-3 BLAST hit to some Ciona gene, and 1,077 have human or amphioxus with a BLAST hit to a Ciona gene, leaving some 1,350 with no detectable relationship by BLAST to a Ciona gene.

Apparent chordate stem gene losses were identified by identifying clusters that include a sea urchin gene and a *Drosophila* or *Nematostella* gene or genes, but lack genes from any of the representative chordate genomes (126 clusters). Only 8 of these clusters are high confidence chordate losses, in the sense outlined above for olfactores losses, and have no BLAST hit with high c-score between a non-chordate cluster member and a chordate gene. These are listed in Supplementary Table S10.

55 gene families appear to have duplicated on the chordate stem (i.e., after the last common deuterostome ancestor, but before the amphioxus-olfactores split) to create 76 new genes in the chordate ancestor.

7.3 Functional categories enriched in amphioxus-specific ancient paralogs

The ontology terms with a significant enrichment among the chordate families which have expande d anciently on the lineage
leading to amphioxus are, in order of enrichment factor: Gprotein coupled receptor; Receptor; Ion transport; Transport;
Oxidoreductase; Other transporter; Cell surface receptor mediated signal transduction; Transporte r; Extracellular matrix; G-
protein mediated signaling; Anion transport; Other metabolism; Oxygenase; Lipid, fatty acid and st eroid metabolism.

The selective forces responsible for these functional enrichments in cephalochordates are unknown .

Supplementary Note 8. Chordate linkage groups

Chromosome segmentation. The human, chicken, and stickleback chromosomes were segmented iteratively by comparison to one another and to the scaffolds of the fugu genome assembly. In each iteration, scaffolds or chromosome segments of genome B are clustered hierarchically, with the Pearson correlation coefficient of the ortholog concentration log-likelihood score on the segments of genome A as the distance metric, and the average pairwise linking method as previously described [27]. Ortholog concentration log-likelihood score is defined to be -log($m\ p$) or zero, whichever is less, where m is the number of pairwise segment comparisons between the genomes in question (*i.e.* a multiple test correction) and p is the probability of the observed number of shared orthologs relative to the null model that the two segments draw their genes independently from the set of genes in the common ancestor. Breakpoints along the chromosomes and chromosome segments of genome A are then identified as discontinuities in the pattern of how their orthologs fall in genome B, using a hidden Markov model algorithm, described in ([27]). Iteration was repeated until comparison to amphioxus, stickleback, and fugu revealed no further breaks and identified a total of 111 breakpoints along the human genome, which bound 135 segments. These segments are referred to by chromosome name, followed by a period, followed by the number of the segment along the chromosome, e.g. "9.1" indicates the segment at the tip of the p arm of human chromosome 9. When a pair of segments on one human chromosome have syntenic orthologs in chicken, fish and amphioxus genomes, we assume them to have been created by intra-chromosomal inversions in the human lineage, and we treated them as a single merged segment, and denote it with a slash as in "9.1/3". There are a total of 105 merged segments, of which 99 had more than 10 ancestral chordate genes represented on them.

Construction of Chordate Linkage Groups. For whole-genome synteny analysis, orthology between genomes was based on c-score clustering as previously described[27], with c-score threshold of 0.75 when comparing human and amphioxus, and 0.95 when comparing human to other vertebrates. To define initial CLGs (Supplementary Figure 64), human chromosome segments and amphioxus scaffolds were clustered using the same method as for segmentation, and clusters defined by cutting the clustering trees at a correlation threshold of 0.25. Statistical significance of ortholog concentration between regions of one genome and another was computed with Fishers Exact test using the null hypothesis that inferred ancestral genes (ortholog clusters) with representative descendants on the two genome segments are chosen independently. Variation in ortholog cluster size makes these p-values an approximation; to limit the magnitude of this effect, we excluded any ortholog clusters with 25 or more genes. We applied a Bonferroni correction for the total number of pairwise tests.

Additional individually significant conserved synteny after segmentation. As described in the main text, 111 amphioxus scaffolds out of 239 that contain twenty or more orthologs of vertebrate genes has a significant concentration of human orthologs on one or more chromosomes, relative to random expectation ($p<0.05$, multiple test corrected). After dividing the human chromosomes into segments representing ancient blocks of conserved linkage (Methods), this number increases to 145 scaffolds with significant conserved macrosynteny relative to human segments.

Coverage of human genome by segments. The segments of the human genome assigned to a CLG by the clustering method span 3.0 Gbp, or 98% of the euchromatic human genome, and 96% of ENSEMBL human gene predictions. 15,021 of these genes have an assigned amphioxus ortholog by the C-value method with $C>0.75$ (see above), and 12,682 are included in this analysis, the rest falling into large gene families and excluded from this analysis. Of these, 53% have (at least one of)

their amphioxus ortholog(s) in a conserved linkage position, (i.e., on a scaffold grouped into the same CLG).

Conserved micro-synteny. Embedded within the 17 ancestral linkage groups are relatively short stretches of conserved gene order (micro-synteny) between amphioxus and human. If we allow at most 10 (or 20) genes to intervene between consecutive pairs of orthologs on the respective genomes, there are 417 (or 1044) human genes that participate in collinear clusters of three or more chordate gene families, compared to 33 (or 154) in a control genome in which gene order in human is scrambled. Thus scrambling of gene order has likely occurred on both the amphioxus and human lineages, and in fact the scrambling of gene order with general preservation of linkage is what allows a single amphioxus scaffold (only a small fraction of a chromosome) to identify conserved linkage with human segments.

Oxford grid methodology. For every scaffold-segment pair, we tabulated the number of predicted ancestral chordate genes with descendants found in both the amphioxus scaffold and human segment. This number of shared orthologous genes was compared to a null model in which the scaffolds and segments have gene content independently drawn from the ancestral set. The "Oxford grid" containted in Supplementary File SF5 in which numbers indicate the number of inferred ancestral chordate genes with orthologous descendants on both the human chromosomal segment (row) and amphioxus scaffold (column). Boxes colored yellow reflect p-value < 0.05, where Bonferroni multiple test correction has been applied as if every pairing in the grid were an independent test (N=45,639). Empty boxes indicate segment-scaffold pairs without orthologs. This grid illustrates not only that there are many scaffold-segment pairs with a highly significant excess of shared ancestral genes, but that the amphioxus scaffolds and human chromosome segments can be grouped into classes, such that scaffold-segment pairs drawn from the same class are likely to have a significant excess of shared ancestral genes.

Reconstruction of teleost proto-chromosomes

We clustered fugu and stickleback genome segments by the same procedure used to reconstruct ancestral chordate chromosomes, with orthology defined as c-score >= 0.75. Supplementary Figure 13 shows that this procedure produces 12 clearly-defined clusters, which we interpret as the 12 chromosomes of the proto-bony fish karyotype, before the teleost-specific genome duplication. Supplementary Table 15 lists the stickleback chromosome segment boundaries.

Reconstruction of bony vertebrate proto-chromosomes

We searched for evidence of fusions between the 69 post-2R vertebrate proto-chromosomes (2rVPCs; Supplementary Table 1) as follows: For each 2rVPC X, we identified inferred ancestral chordate genes (see Supplementary Note 7) which (a) have an amphioxus descendant on a scaffolds grouped into the parent CLG of X, (b) have a human descendant on one of the human chromosome segments assigned to X in Supplementary Table 1, (c) do not have any additional human descendants on paralogous segments. We found 3,486 such diagnostic genes, a mean of 51 diagnostic ancestral chordate genes per 2rVPC. For each pair of 2rVPC X and Y, we then identified those human, chicken and reconstructed ancestral teleost chromosomes which had a significant enrichment (by Fisher's exact test) for genes descended from the diagnostic gene sets of X and Y. Pairs of 2rVPCs which co-occur on one or more chromosomes in each of the human, chicken and reconstructed ancestral teleost genomes are likely already to have been merged in the last common

ancestor of the bony vertebrates. For 6 2rVPCs (11c, 12c, 14a, 15a, 17c, and 1b), no significant concentration of diagnostic genes was found on any chicken chromosome. For pairs involving these 2rVPCs, co-occurrence on a human chromosome and a fish proto-chromosome was interpreted to imply a fusion before the bony vertebrate common ancestor. 39 of the 69 2rVPCs participate in such an implied merger of segments after 2R and prior to the bony vertebrate LCA, including ancient pairings of this kind, including 13 pair-wise and 6 three-way groupings (Supplementary Table 11). Several 2rVPCs participate in more than one ancient pairing, which could indicate fissions of these segmentts between 2R and the bony vertebrate LCA. 12 2rVPCs could not be tested due to a lack of a significant hit to any teleost or human chromosome, and could therefore not be evaluated by this criteria. We estimate therefore that the ancestral bony vertebrate had between 37 (if none of the 12 untested segments was it's own chromosome in the bony vertebrate LCA) and 49 chromosomes (if they all were).

Identification of fusions on the amniote stem

Diagnostic collections of ancestral chordate genes were created for each of the bony vertebrate LCA chromosomes by merging the diagnostic clusters of their constituent 2rVPCs. In these cases where one 2rVPC appears to have undergone fission, and participates in two independent bony vertebrate LCA chromosomes, the diagnostic set was subdivided accordingly. For example, 2rVPC 5d occurs with 5a on human chromosome X and chicken chromosome 1, and also with 4d on human 1 and chicken 23. The diagnostic genes with ortholgos on human X or chicken 1 were combined with the diagnostic genes of 5a form one new set of diagnostic genes, while those with orthologs on human 1 or chicken 23 were grouped with those of 4d. Pairs of these diagnostic sets were then compared to the human and chicken chromosomes, to identify the 4 apparent mergers, and one apparent fission, of bony vertebrate LCA chromosomes on the amniote lineage. (Supplementary Table 12).

Fluorescent in situ hybridization (FISH). Chromosome preparation was performed as previously described[29] and modified as follows.

To obtain metaphase plates with non-overlapping chromosomes, after an initial treatment with 0.02% colchicine (Sigma) in sea water (25 minutes), embryos were transferred to a 1.5-ml microfuge tube and treated with hypotonic mixtures (seawater and 0.075M KCl in ratios 2:1 and 1:1 for 5 min each). Two color FISH was carried out as previously described[30] with a modified hybridization solution. The hybridization mix consisted of 50% formamide, 2xSSC, 10% dextran sulfate, 0.1µg/µl sheared B. floridae adult DNA, 0.15% SDS. The DIG and biotin-labeled probes were added to the hybridization mix at a final concentration of 2-10ng/µl. (See Supplementary Table S2 and Supplementary Figure S8)

Supplementary Note 9. Timing of whole genome duplications

9.1 Phylogenetic analysis

We applied a common protocol to ORFs derived from lamprey ESTs and peptides from *Ciona* and *Fugu* genes:

1. For each of 1,093 ancient chordate clusters with paralogs arising from the 2R event or events, take one amphioxus gene (ignoring the other allele, and/or amphioxus specific gene duplicates) and one pair of human genes on distinct chromosomal segments that are apparently paralogous based on earlier analysis. (If the family has three or more paralogs, other genes are not considered further.)
2. Add a lamprey, *Ciona*, or *Fugu* gene "X" to one of these clusters if it's best BLAST hit among human genes is in that cluster. If more than one lamprey/ciona/fugu gene satisfies this condition for the same cluster, add only the one with the best score. This procedure forms sets of quartets (Bf,Hs1,Hs2,X).
3. For each quartet, create a multiple alignment with Clustalw[19], trim the alignment with stringent options in Gblocks[20].
4. Evaluate the likelihoods of the three unrooted tree topologies for this tree with PHYLIP[17]. If one topology has maximum likelihood in >50% of 500 bootstrap replicates with PHYLIP, it is considered resolved.

If the two rounds of whole genome duplication occurred before the divergence of a given lineage, then we would expect that a gene X from a family that duplicated on the vertebrate stem would more often be closely associated with a specific jawed-vertebrate gene in a family of ohnologs, and so would be related to their closest human (Hs1, Hs2) and amphioxus (Bf) homologs according to the nested relationship ((Hs1, X), Hs2), Bf). If, on the other hand, large-scale duplication occurred both before and after the divergence of a given lineage, then we would expect at least 33% of individual genes to be related to pairs of human genes according to (((Hs1, Hs2), X), Bf) (Supplementary Figure S11).

Results are summarized in Supplementary Table 1 of the main text.

9.2 Elephant shark conserved synteny and 2R.

Analysis of the draft sequence of cartilaginous fish *Callorhynchus millii* (the elephant shark) has provided evidence that both rounds of genome duplication affecting bony vertebrates preceded the divergence of bony and cartilaginous fish[31]. We used the whole genome data of that study to compare gene linkage in elephant shark with our reconstructed chordate linkage groups and 2R paralogy groups. By BlastP, we compared 1,731,890 shotgun reads to the human proteome, and identified 2,096 fosmid clones whose end sequences have distinct unique homolog (*i.e.*, no other hits with > 65% of the best score) in the human gene set. These genes are ~35-40 kb apart in the shark genome.

If ancestral synteny relationships in the shark genome have not been lost through genomic rearrangements (as they nearly have all been in *Ciona*), and if 2R preceded the cartilaginous/bony split, then we expect that the shark gene pairs would be found predominantly in the same ancestral chordate linkage groups, and also within the same paralogy group (*i.e.* the same row and column of Table 1 of the main text). If, on the other hand, one or both rounds of genome duplication were specific to the bony vertebrates and had not been shared with cartilaginous fish, then duplication and intervening loss after the split with shark would lead to an excess of clones between different paralogous segments derived from the same chordate linkage group. 1,123 clones (54%) hit different genes belonging to the same chordate linkage group. Of these 980 (87%) hit two genes in the same paralogy group within a given CLG. It is most parsimonious to conclude that both rounds of genome duplication preceded the divergence of bony vertebrates and sharks. (The alternate hypothesis of a bony-vertebrate-specific genome duplication would require highly correlated reciprocal gene loss across the entire genome.)

Supplementary Note 10. Conserved non-coding sequences

10.1 Computational detection of conserved non-coding sequences.

Whole genome alignments of the Amphioxus genome version 1 scaffold sequences and Human genome reference sequence (NCBI v35) were performed using the VISTA pipeline infrastructure [32], based on efficient combination of global and local alignment methods. First, we obtained a map of large blocks of conserved synteny between the two species by applying Shuffle-LAGAN glocal chaining algorithm[18] to local alignments produced by translated BLAT[4]. Then, in each syntenic block we apply Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions. After that, VISTA with a similarity cut-off of 60% identity over 50bp was used to identify candidate conserved non-coding sequences.

The human genome coordinates of all alignments were used in the UCSC table browser to identify and remove: (i) elements overlapping known genes, human mRNAs, spliced ESTs, Ensembl genes, human repeats, simple repeats, retroposed genes, and high copy number regions in the human genome (defined by more than one overlapping self-chain hit); (ii) elements not overlapping the "most conserved" track in the UCSC genome browser; (iii) elements not conserved between human and at least one of the fish, chicken, or frog genomes. The amphioxus genome coordinates of all remaining alignments were used to exclude elements with BLASTX alignments to proteins encoded by the amphioxus genome (Supplementary Table S11).

Supplementary Note 11. NK quadruple conserved synteny.

In addition to the genome-wide synteny analysis, a detailed analysis was carried out on four version 1 scaffolds that make up the NK homeobox cluster in amphioxus[33]. To ensure accuracy, phylogenetic reconstruction, rather than BLAST, was used to assign orthology and detect gene duplication. Phylogenetic analysis was carried out for 122 amphioxus Filtered Models from version 1 scaffolds 56, 124, 185, and 294 which showed strong evidence for being real genes (based only upon EST matches and/or BLAST score and not upon prior data from previous BAC analysis) and which had BLAST matches to known genes in a range of invertebrates and vertebrates including humans. Trees with support values less than 50% at key nodes were discarded. Amphioxus-specific duplicates, or identical amphioxus genes (which may have been due to assembly errors) were counted only once per family. The locations of human orthologs resulting from duplications at the base of the vertebrate lineage were determined for the remaining 82 gene families. The 82 amphioxus genes correspond to 111 human genes, indicating that amphioxus genes typically have a single ortholog in humans. The observed chromosomal distribution of these human genes was significantly different from the expected distribution calculated from the total number of “known genes” on each human chromosome (Chi-squared test $p<0.001$). A histogram (Supplementary Figure 7) showing the number of orthologs on each human chromosome, scaled by the total number of genes on that chromosome, revealed peaks on human chromosomes 4, 5, 8 and 10, in agreement with the genome-wide analysis of CLG #7.

Figure S1.

Scatter plot of non-synonymous (dN) and synonymous (dS) substitution rates between alleles.

Supplementary Figure I

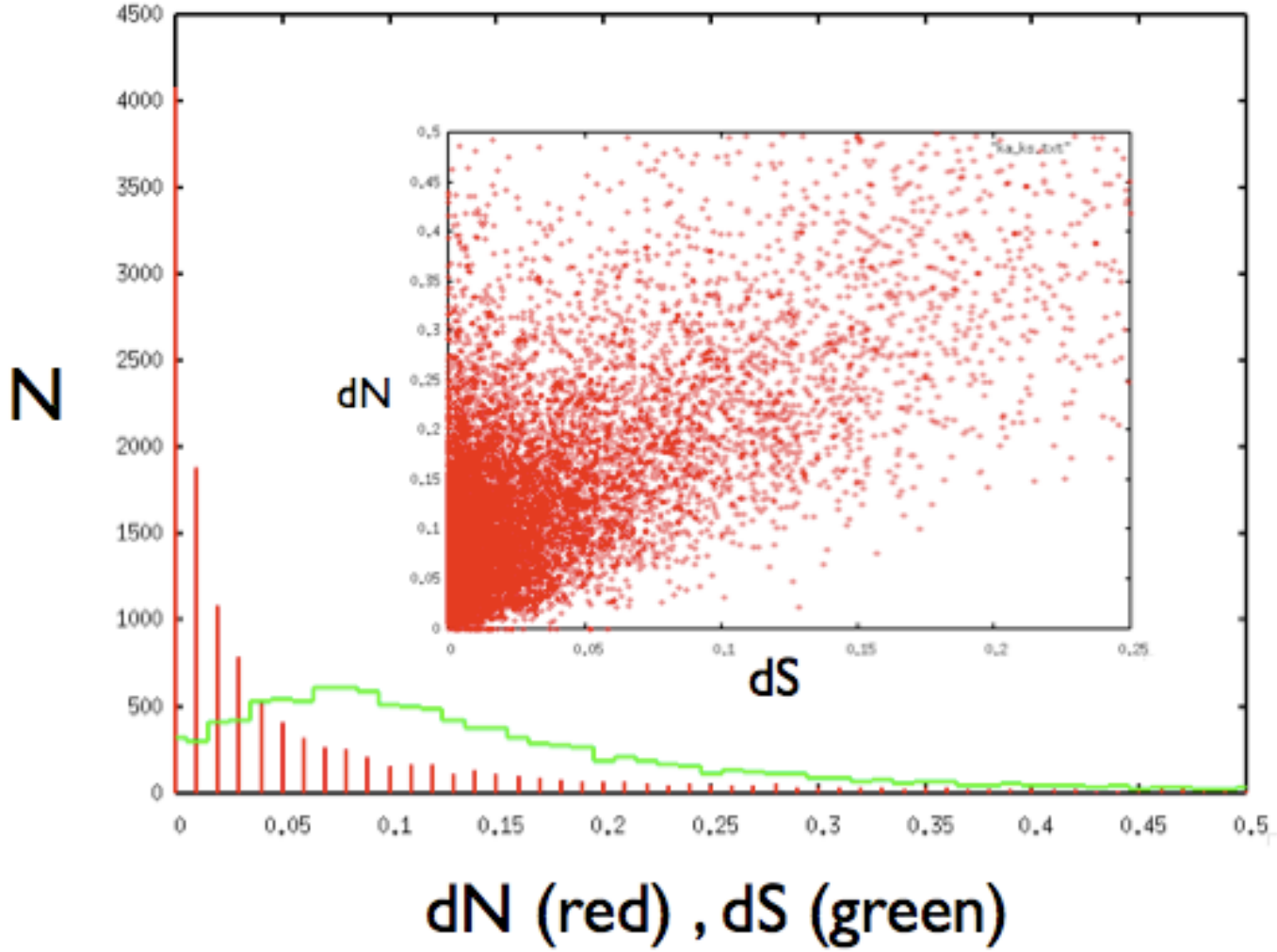


Figure S2.

Classes of variation observed in alignments of a sample of 23,014 allelic pairs of genomic loci, in which randomly-selected 2,000 bp query segments of the genome have their best non-self BLAT alignment to the assembly in regions free of assembly gaps. The total span (A) summed over the sample was 43.62 Mbp. The total length of gap-free aligned regions (B) was 37.88 Mbp. The total length of (net) inserted sequence into the query (C) was 2.98 Mbp. The remaining 2.76 Mbp was contained in internal aligned regions (D).

Supplementary Figure 2

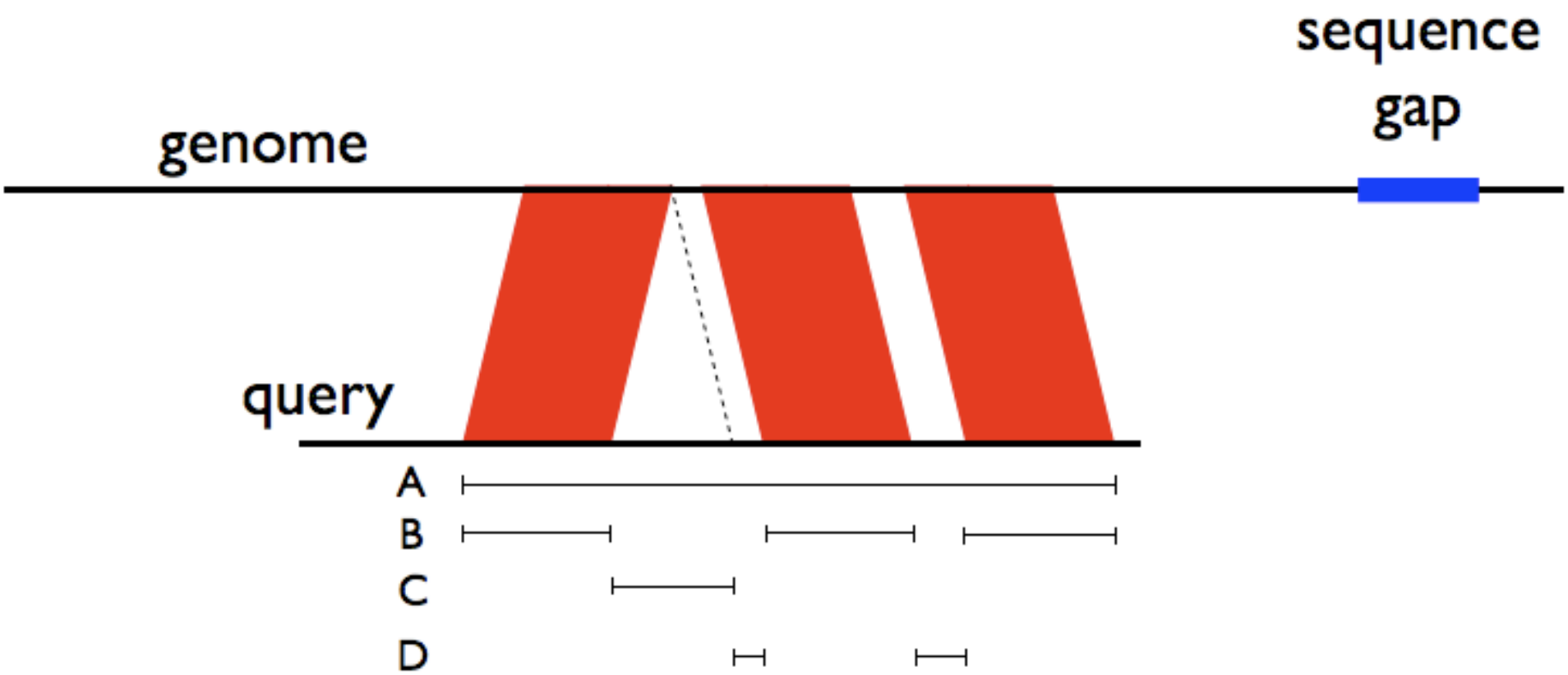


Figure S3.

(B) Human paralogs on the same chromosomal segment (pink) are highly enriched for recent gene duplications, and have few ancient duplications, as measured by four-fold synonymous transversions. Conversely, paralogs lying on different segments (blue) are enriched for ancient gene duplications (4DTv ~0.42). Paralogous gene pairs joining different segments from the same reconstructed chordate linkage group (green) are overwhelmingly from ancient duplications.

Supplementary Figure 3

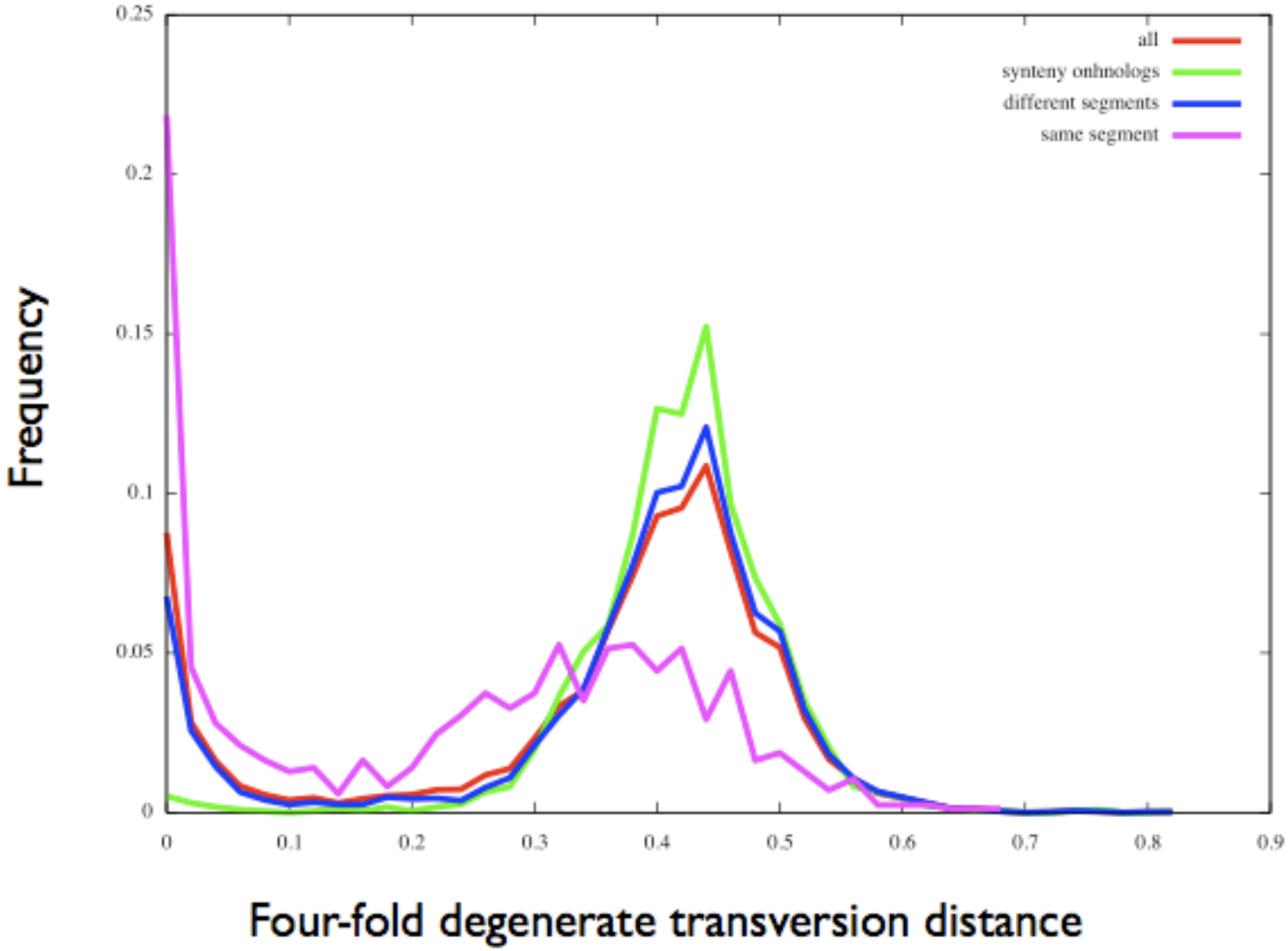


Figure S4. Distribution of mismatches in 50 bp windows in aligned, gap-free, allelic intergenic sequences. Note logarithmic scale on y-axis. Below ~80% identity, or 10 mismatches per 50 bp window (98% of the 380,000 sampled windows) the distribution matches a geometric distribution with mean of 1.6 mismatches per window (*i.e.*, 96.2% identity). The peak above ~20 mismatches corresponds to non-allelic variation.

Supplementary Figure 4

Distribution of mismatches in 50 bp windows in aligned allelic intergenic sequences.

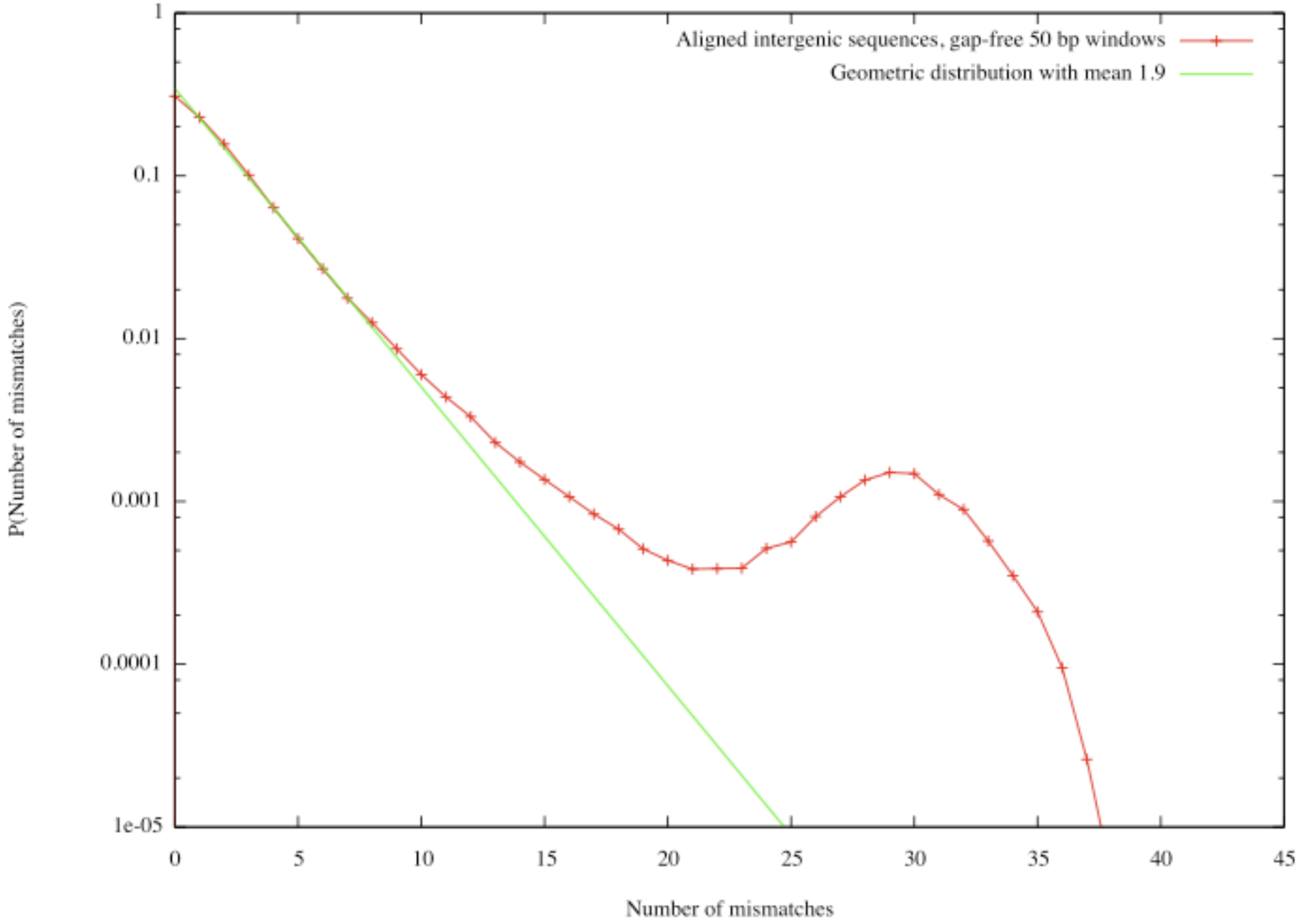


Figure S5. Local heterozygosity is correlated a short length scales. Correlations in local heterozygosity measured by the probability that there are no allelic mismatches in a 10 bp window at position x+d given that there are no mismatches at position x.

Supplementary Figure 5

Local heterozygosity is correlated at short length scales.

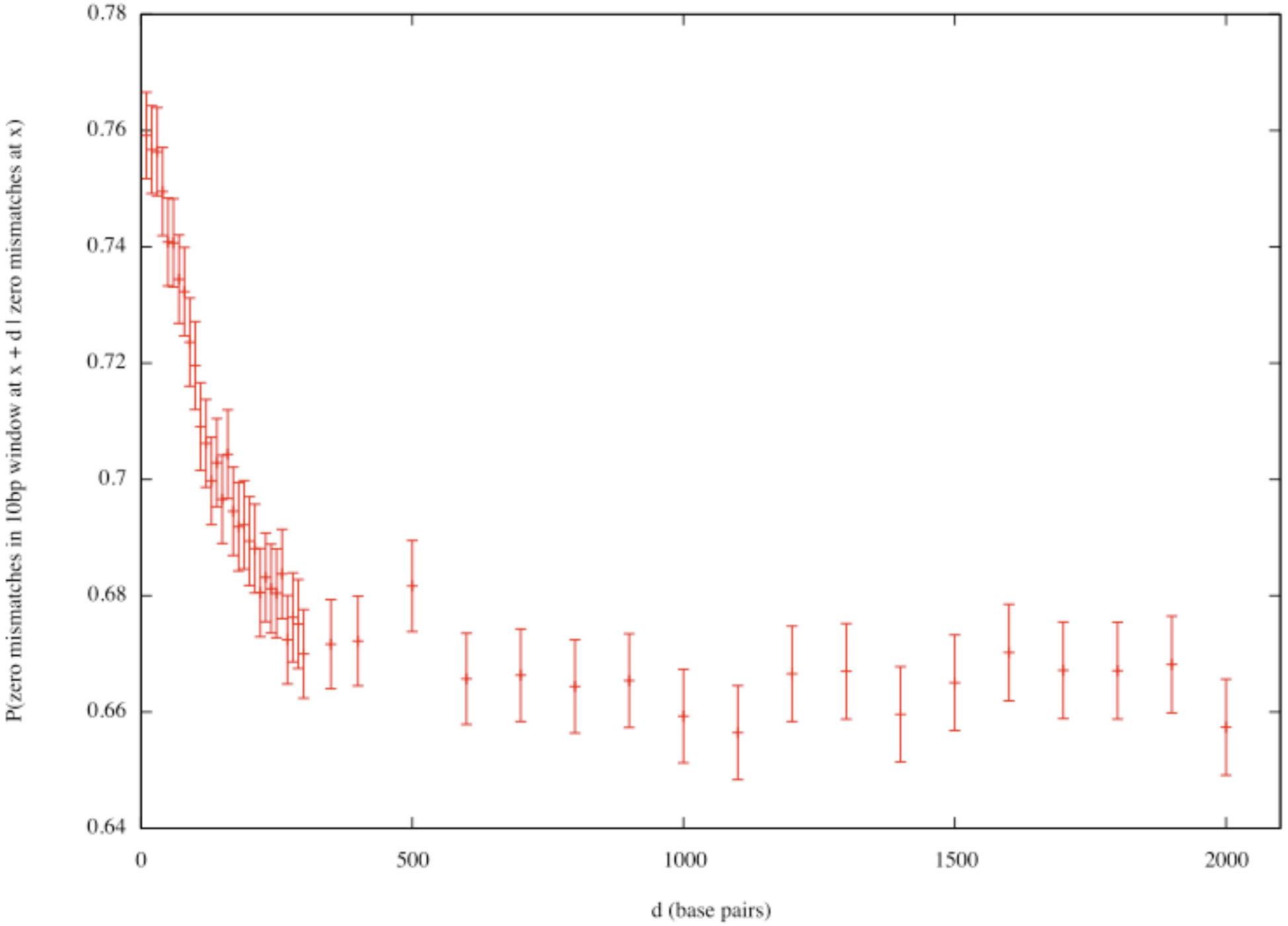


Figure S6. Length distribution of gap-free aligned regions in allelic, intergenic sequence.

Supplementary Figure 6

Distribution of lengths of ungapped alignments between haplotypes.

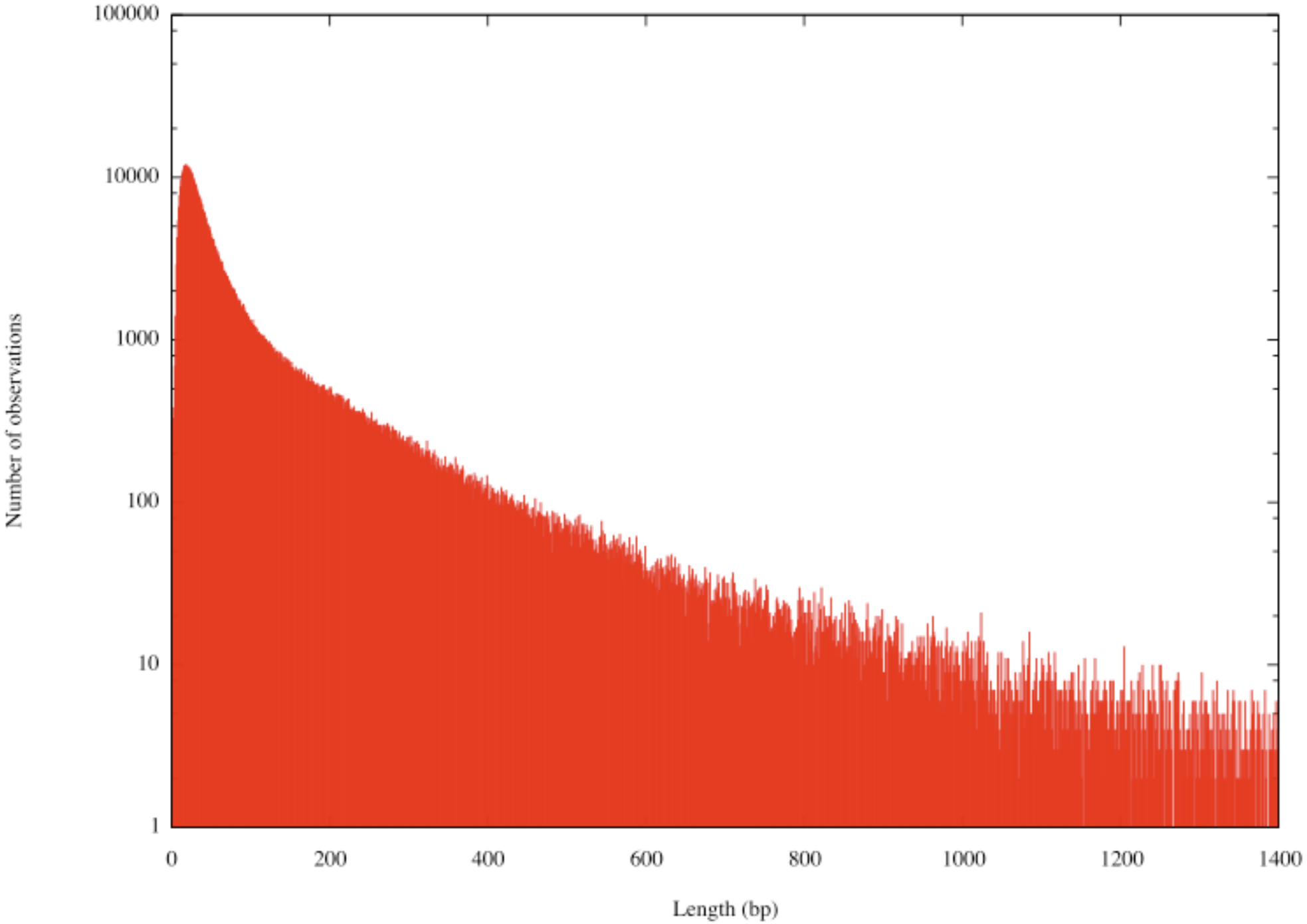


Figure S7. Ka/Ks distribution for amphioxus alleles, human-chimp and human-mouse orthologs. The distribution of Ka/Ks for amphioxus allelic pairs, along with that of human-chimpanzee and human-mouse ortholog pairs. Ka/Ks was estimated by the method of Yang and Nielsen[16] via the PHYLIP package[17] based on full-length pairwise alignments of protein sequence in each case, but the genes used were restricted to those with at least 100 amino acid positions after a 4-species multiple sequence alignment was trimmed with Gblocks.

Supplementary Figure 7

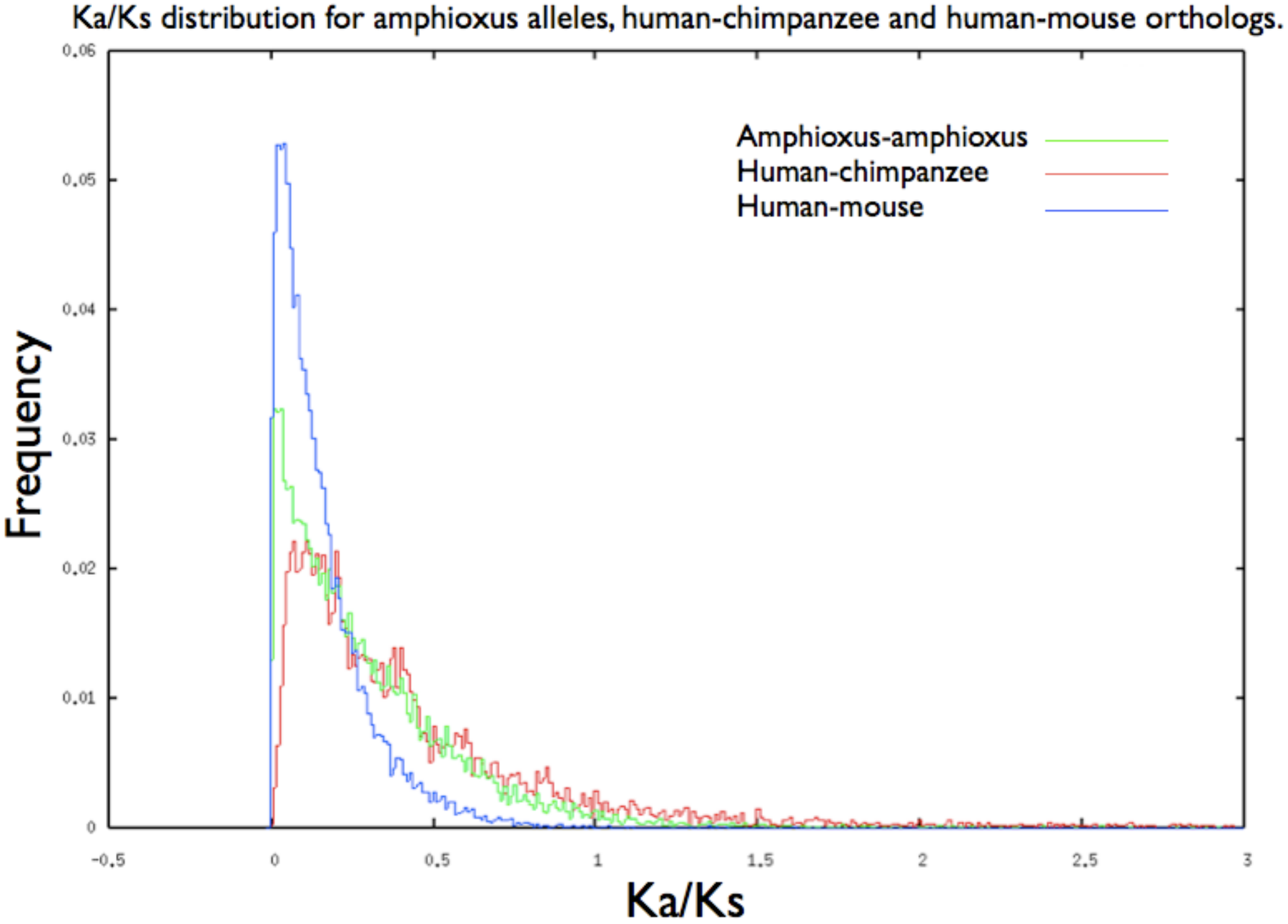


Figure S8.

Phylogenetic tree inferred from presence and absence of introns in conserved positions, with Bayesian (black) and maximum parsimony bootstrap support (red) values shown for chordates and olfactores. Nematostella and Lottia were constrained to to out-group positions (See Supplemental Note 6).

Supplementary Figure 8

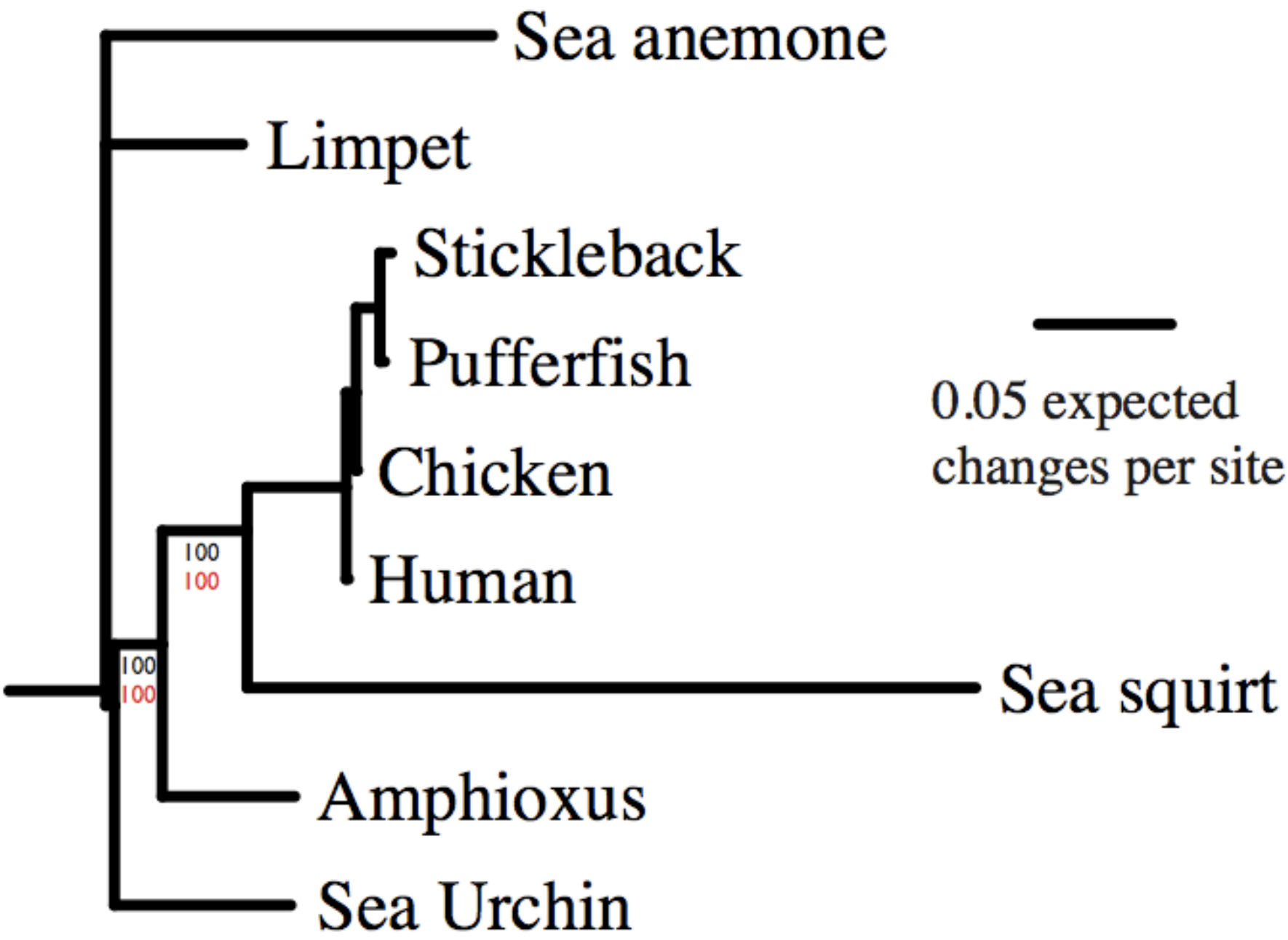


Figure S9.

For the 17 CLGs, the distribution of the number of independent chromosomal segments in the jawed-vertebrate common ancestor as inferred by multi-species synteny conservation alone (blue), distribution of ohnologs alone (green), and the combined analysis (red). The distribution of the number of independent chromosomal segments descended from each CLG at the time of the fish-tetrapod split as inferred by parsimony using multi-species synteny and gene duplicates separately, and in a combined analysis (Methods). Because the multi-species synteny analysis relies heavily on comparison to ray-finned fish chromosomes, which appear to have undergone substantial scrambling of gene order within chromosomes, the estimated number of inferred ancestral jawed vertebrate segments per CLG is expected to be a lower bound. In contrast, we expect the number of independent segments inferred by gene duplicates alone to be inflated due to small chromosome segments with too few ancient duplicates to be constrained with statistical significance. When combined, however, these complementary signals reveal a striking dominance of quadruple conserved synteny over other multiplicities.

Supplementary Figure 9

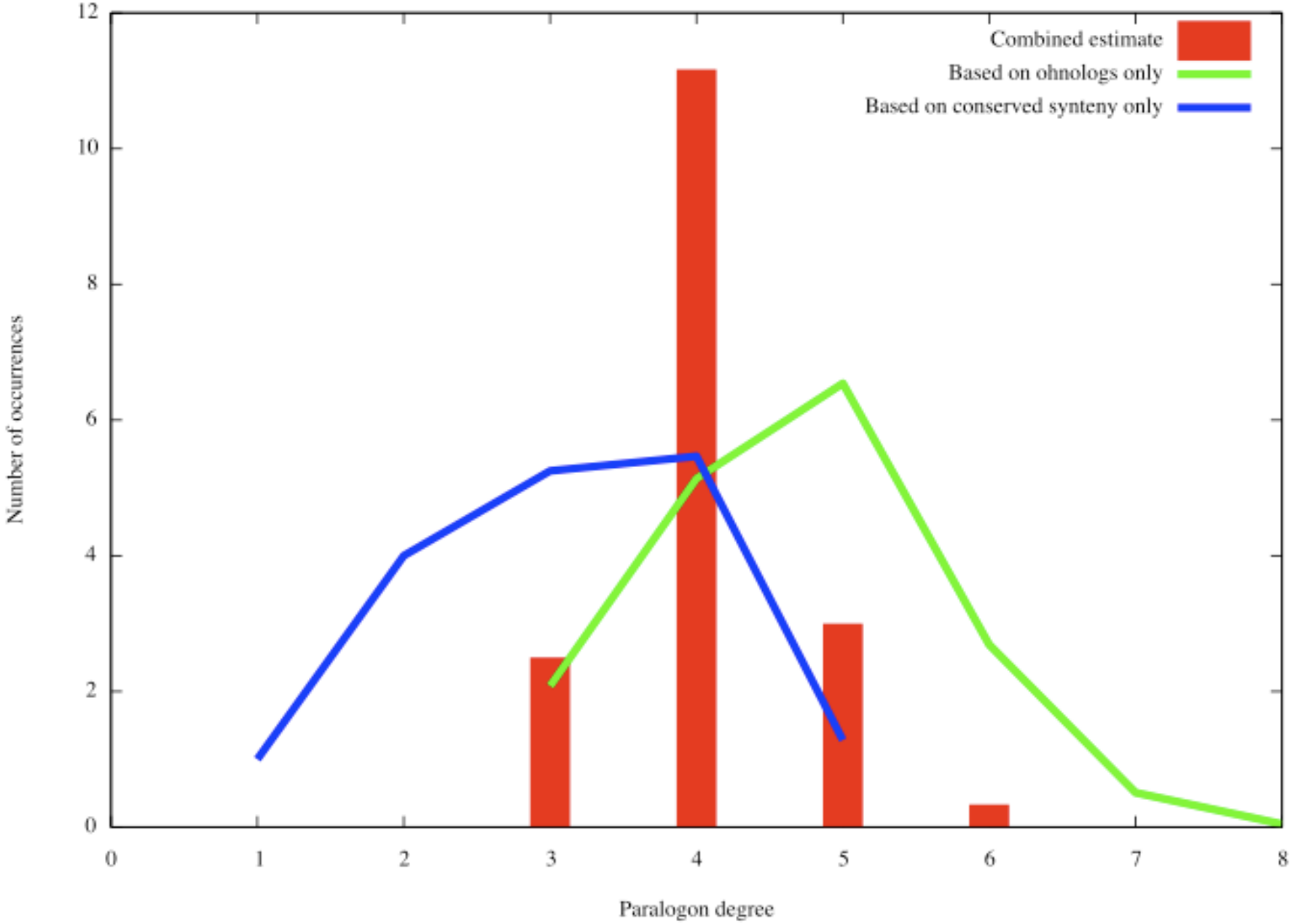


Figure S10. Amphioxus - Chicken Dot Plot

Dots represent the rank order position of orthologous gene pairs in chicken chromosomal segments (vertical position) and the 461 largest amphioxus scaffolds (horizontal position). Amphioxus scaffolds are shown in the order shown in Figure 3b of the main text, and chicken chromosomal segments have been ordered by hierarchical clustering based on comparison to the 17 clusters of amphioxus scaffolds, using the clustering method described in Methods.

Supplementary Figure 10

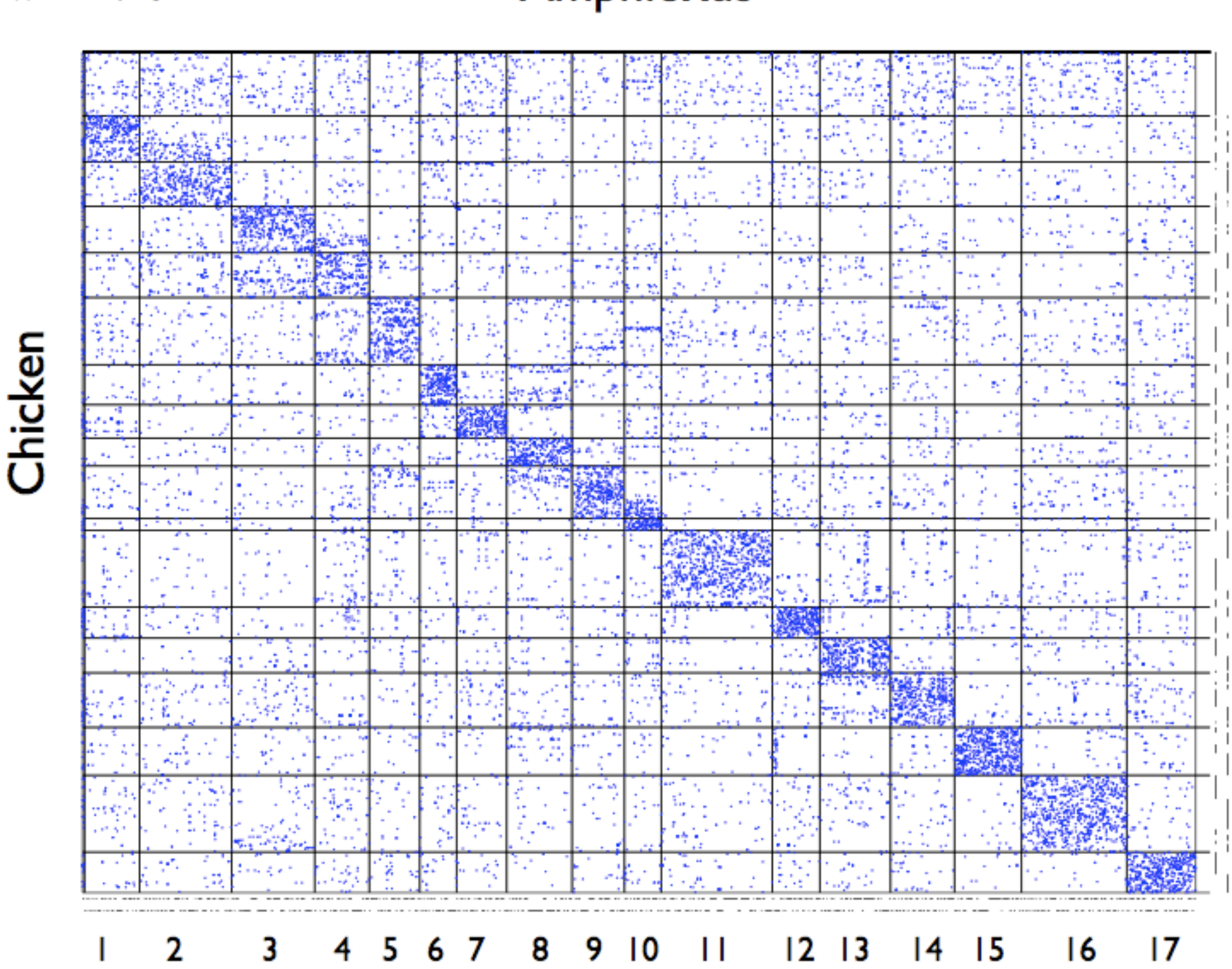


Figure S11. Amphiouxs - Pufferfish Dot Plot

Dots represent the rank order position of orthologous gene pairs in pufferfish scaffolds (vertical position) and the 461 largest amphioxus scaffolds (horizontal position). Amphioxus scaffolds are shown in the order shown in Figure 3b of the main text, and pufferfish scaffolds have been ordered by hierarchical clustering based on comparison to the 17 clusters of amphioxus scaffolds, using the clustering method described in Methods.

Supplementary Figure 11

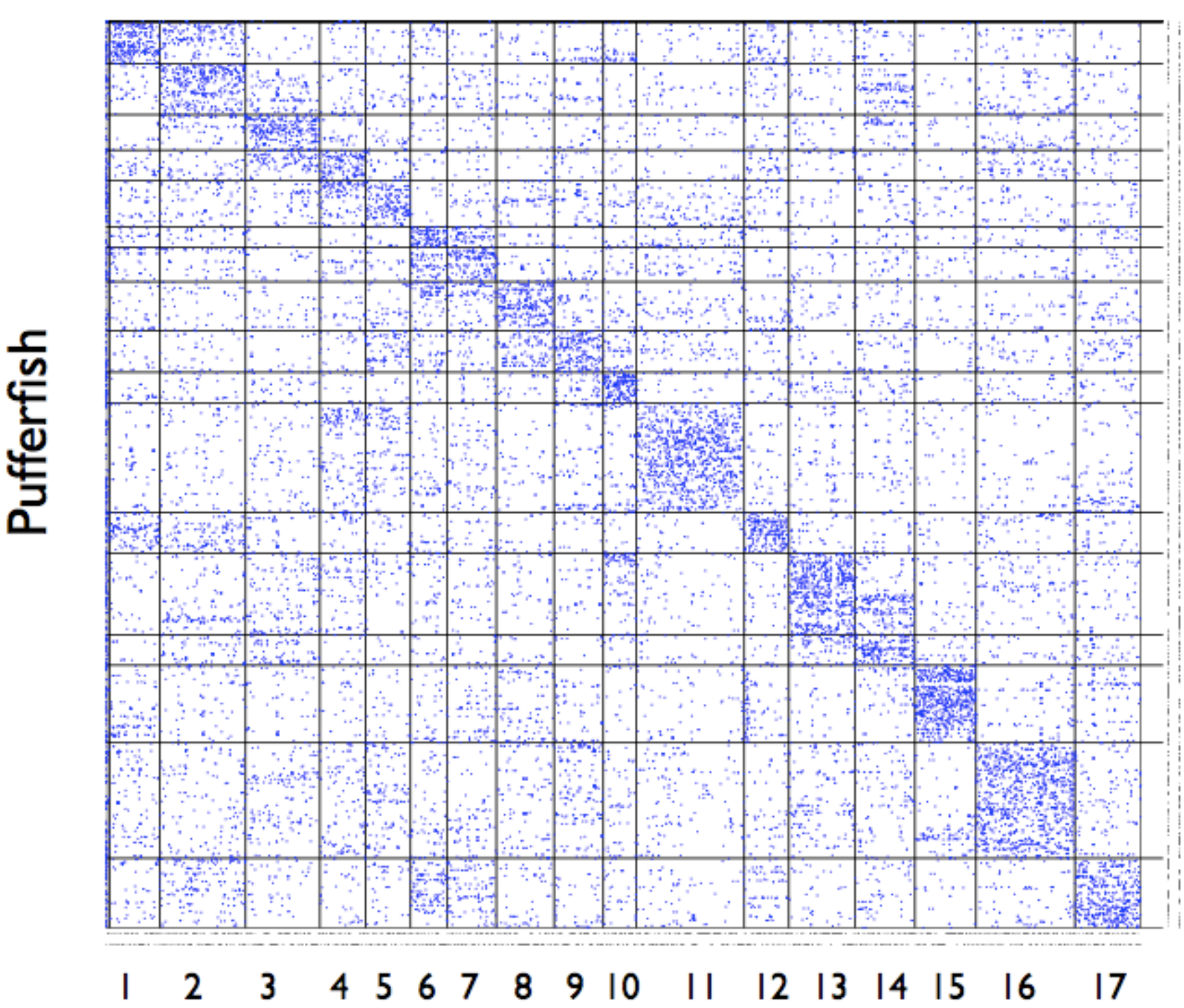


Figure S12. Amphioxus - Stickleback Dot Plot

Dots represent the rank order position (index) of orthologous gene pairs in stickleback chromosomal segments and scaffolds (vertical position) and the 461 largest amphioxus scaffolds (horizontal position). Amphioxus scaffolds are shown in the order shown in Figure 3b of the main text, and stickleback chromosomal segments and scaffolds have been ordered by hierarchical clustering based on comparison to the 17 clusters of amphioxus scaffolds, using the clustering method described in Methods.

Supplementary Figure 12

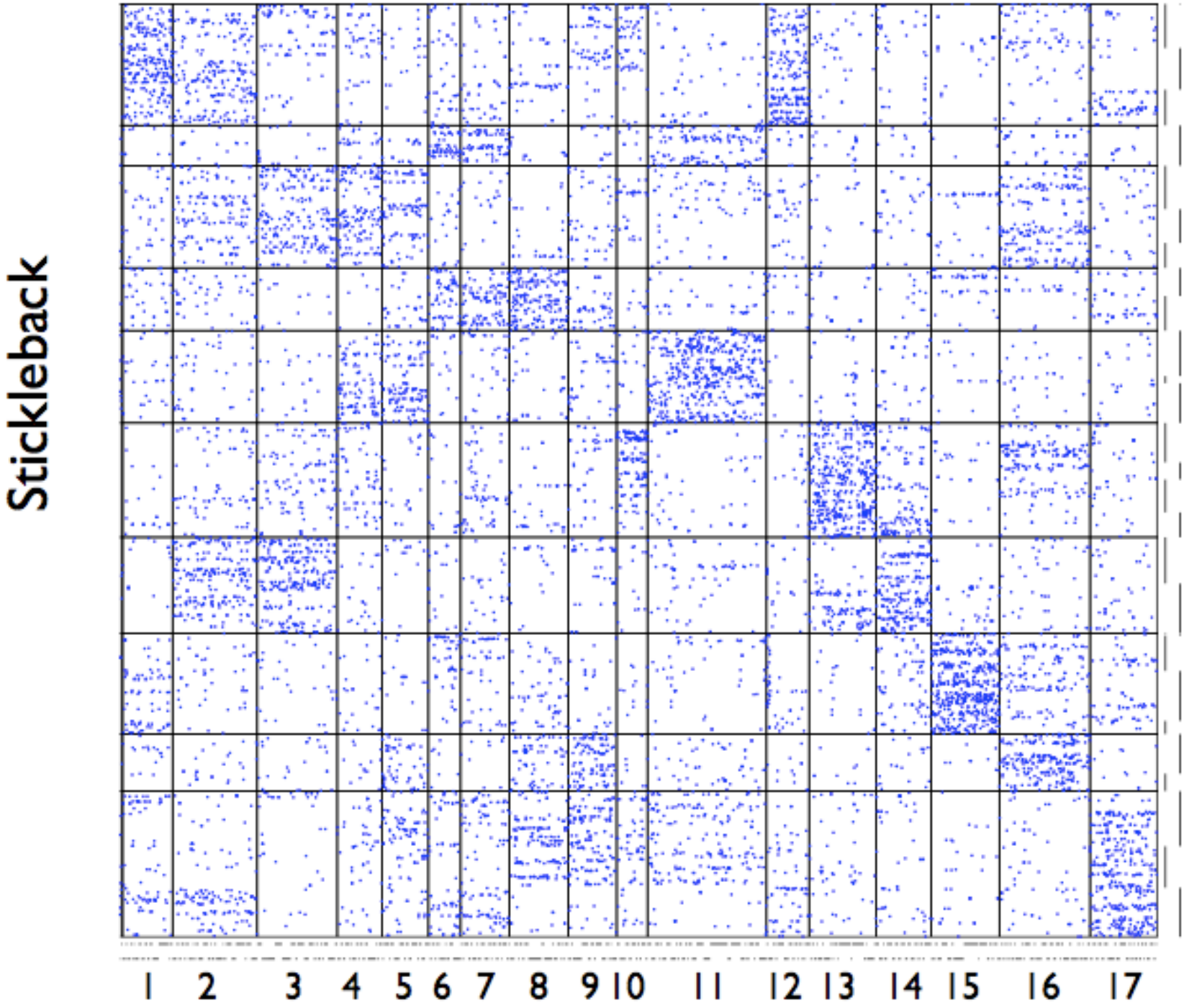


Figure S13. Puffefish - Stickleback Dot Plot

Dots represent the rank order position (index) of orthologous gene pairs in stickleback chromosomal segments and scaffolds (vertical position) and pufferfish scaffolds (horizontal position). Pufferfish and stickleback chromosomal segments and scaffolds have been ordered by hierarchical clustering based on comparison to each other, using the clustering method described in Methods.

Supplementary Figure 13

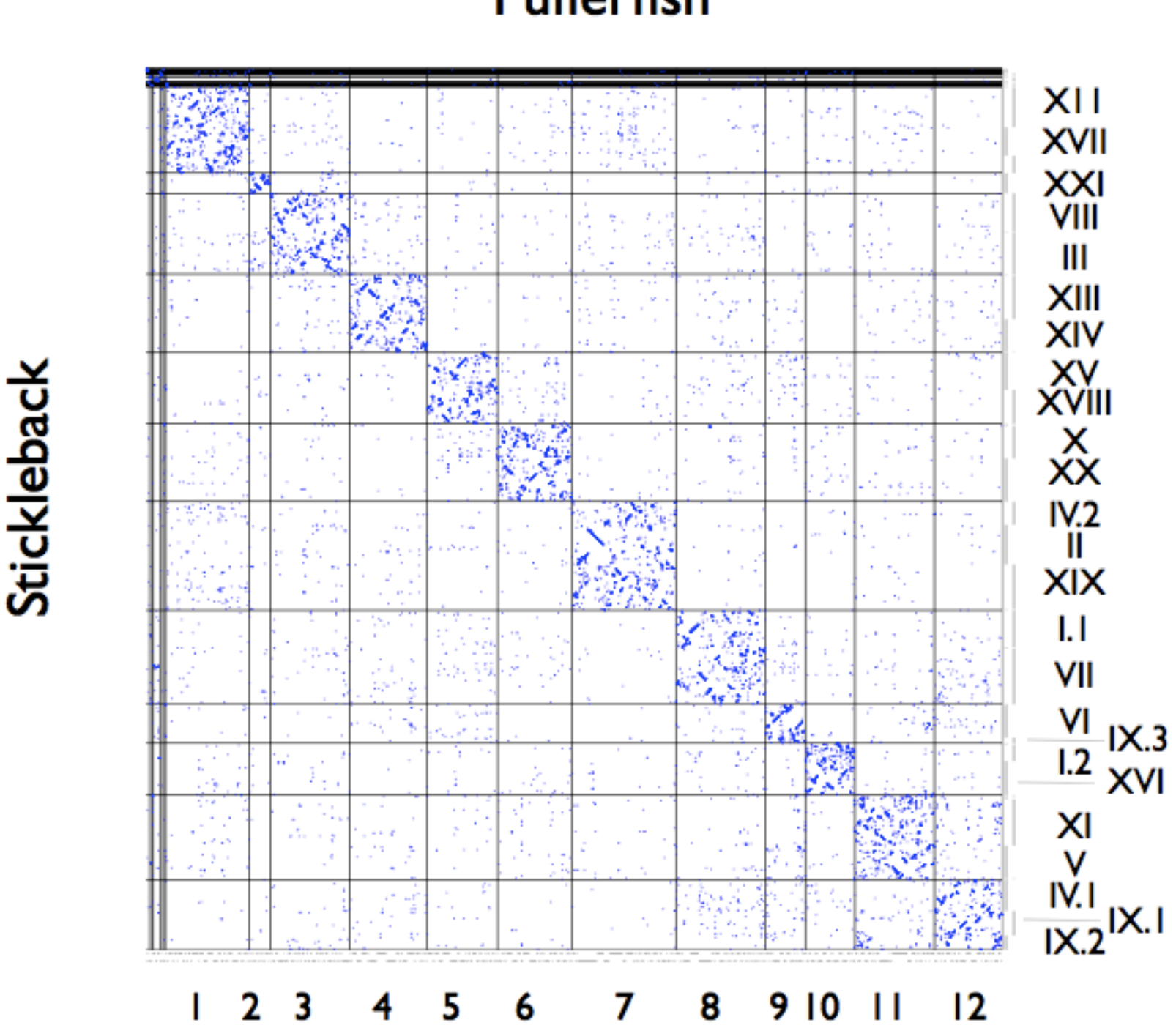


Figure S14. Amphioxus - *Ciona* Dot Plot
Dots represent the rank order position of orthologous gene pairs in Ciona intestinalis chromosomal segments and scaffolds (vertical position) and the 461 largest amphioxus scaffolds (horizontal position). Amphioxus scaffolds are shown in the order shown in Figure 3b of the main text, and Ciona chromosomal segments and scaffolds have been ordered by hierarchical clustering based on comparison to the 17 clusters of amphioxus scaffolds, using the clustering method described in Methods.

Supplementary Figure 14

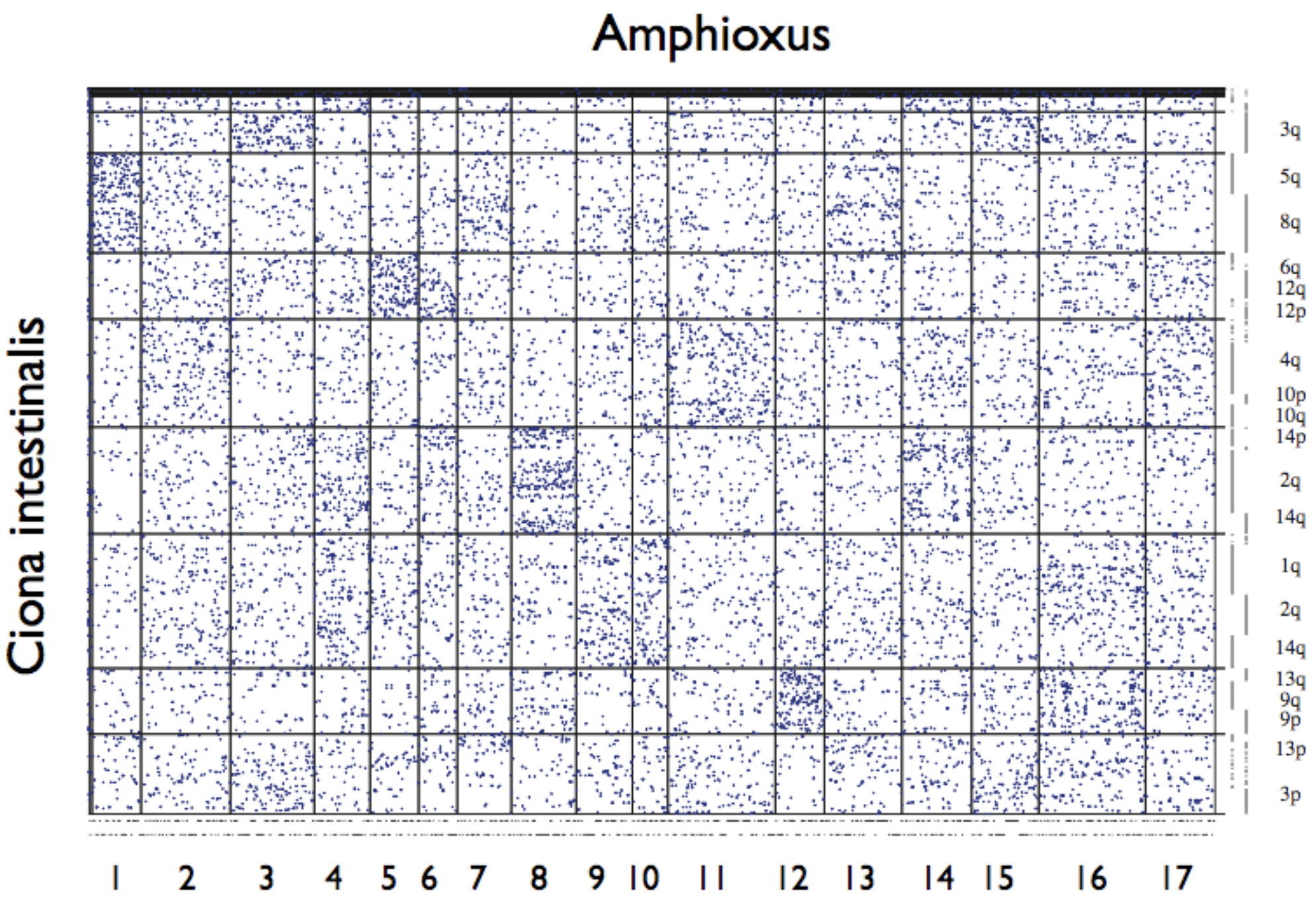


Figure S15. Histogram of human chromosomal orthologs of genes linked to NKhomeobox clusters in amphioxus.

Supplementary Figure 15

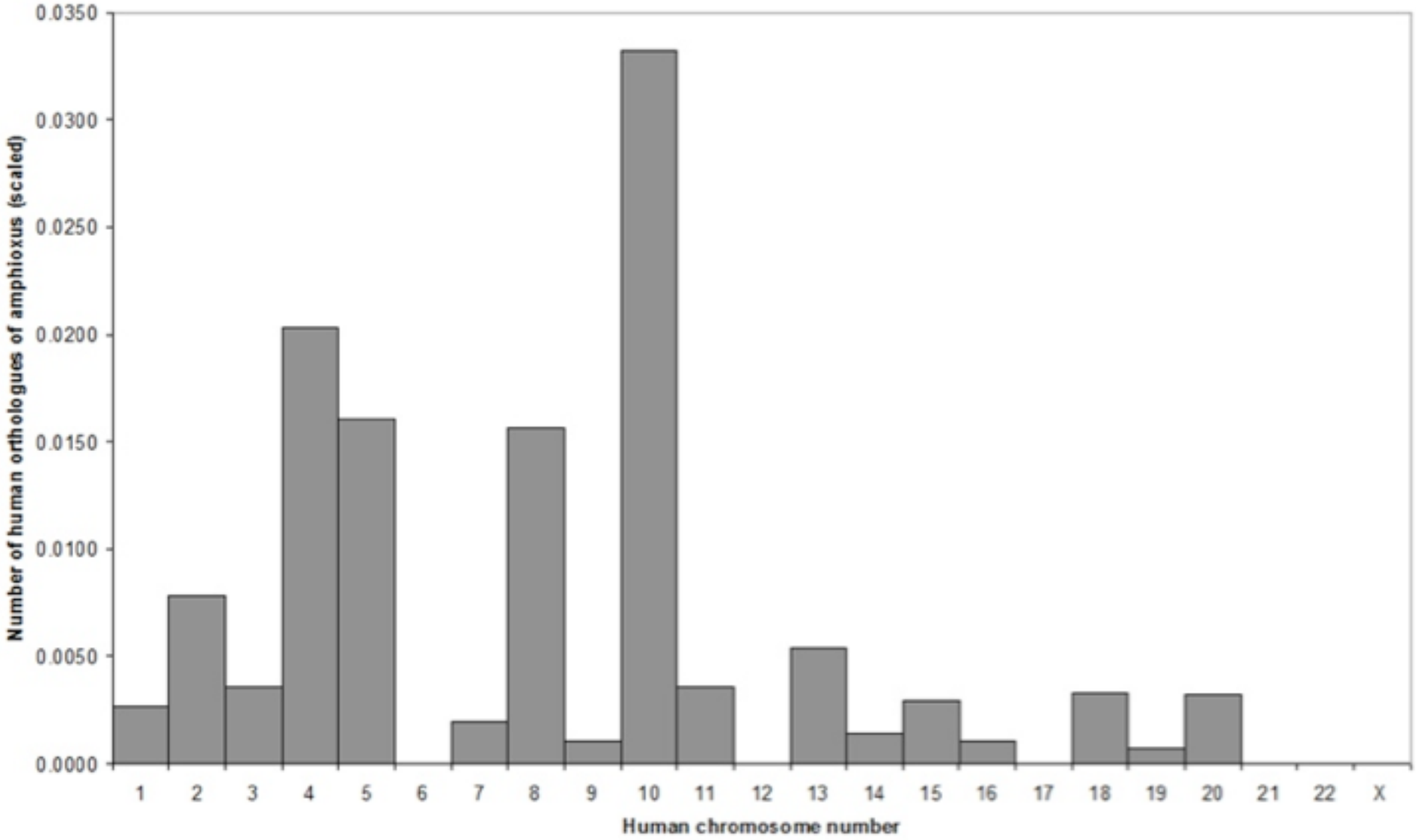


Figure S16. Fluorescent in situ hybridization (FISH) maps scaffolds of one CLG to one amphioxus chromosome. (a) FISH with 007_B24 (red) and 005_C15 (green). Arrowhead indicates strong DAPI-staining (centromere bands). (b) The locations of 29 FISH-mapped BACs (left side) are shown relative to DAPI-counterstained chromosomes with centromere bands (at an upper region). Clones that mapped proximal to the centromere are represented by a pair of black circles. For example, corresponding cytogenetics Clones that mapped distal to the centromere are represented by a pair of red circles, and those that mapped to the middle portion of the arms are shown by a pair of gray circles. Arrows on the right side of the scaffold number indicate the orientation of the mapped sequences.

Supplementary Figure 16

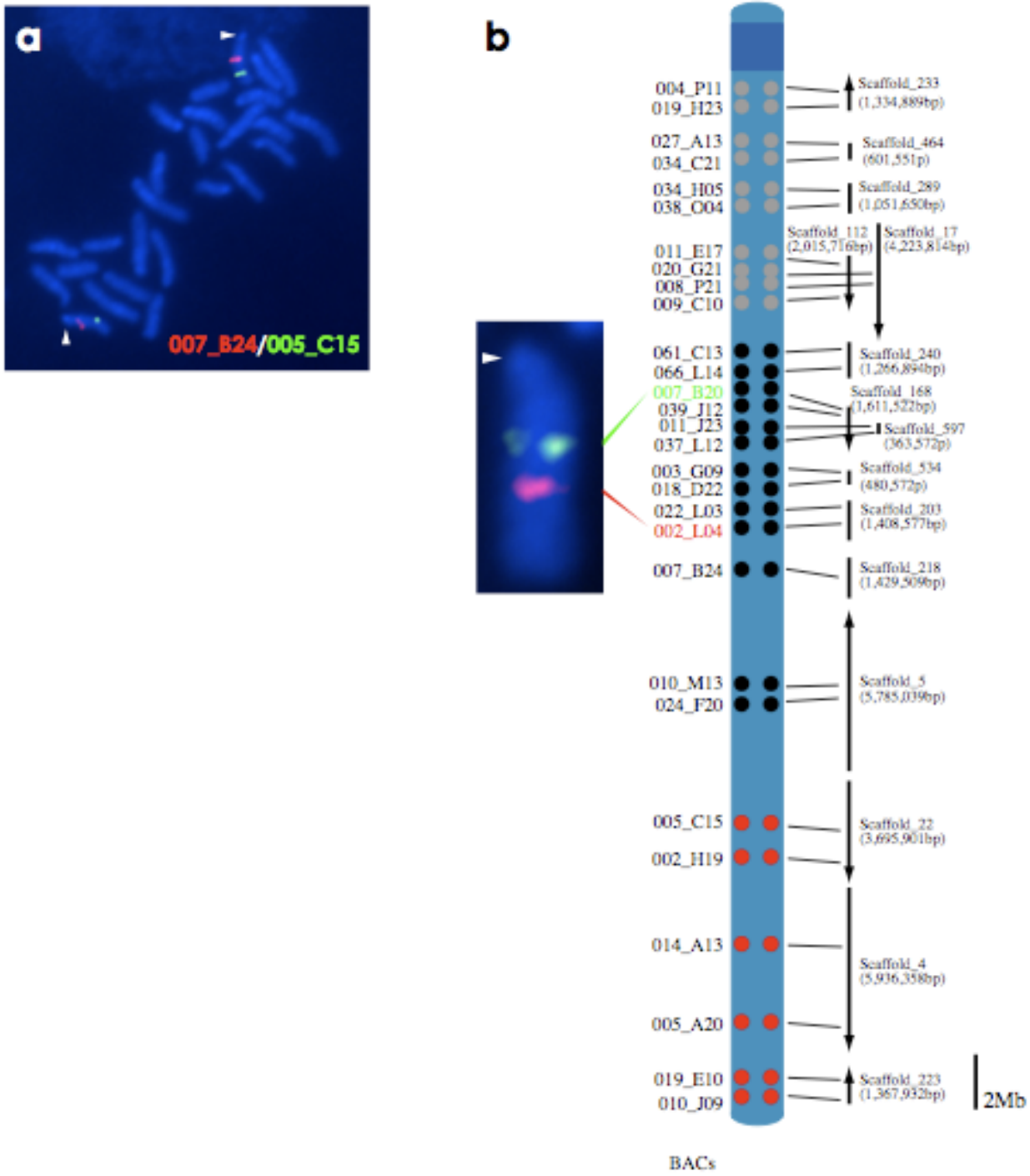


Figure S17. Color code representing patterns of multi-species synteny conservation, used in Supplementary Figures S18-S20.

Each color is followed by a list of the conservation profiles that it is used to represent. Each conservation profile consists of a string of four characters. The four positions in each string refer to conservation in the stickleback, pufferfish, amphioxus and *Nematostella vectensis* genomes respectively. A "1" at a particular position indicates that two segments of the human genome co-occur in the same chromosome, or cluster of scaffolds in the corresponding genome with high statistical significance. A "0" indicates that the two segments each have a significant hit in the corresponding genome, but that the hits do not coincide. A "?" indicates that one or more of the segments lacks any significant conserved synteny in the other genome.

Supplementary Figure 17

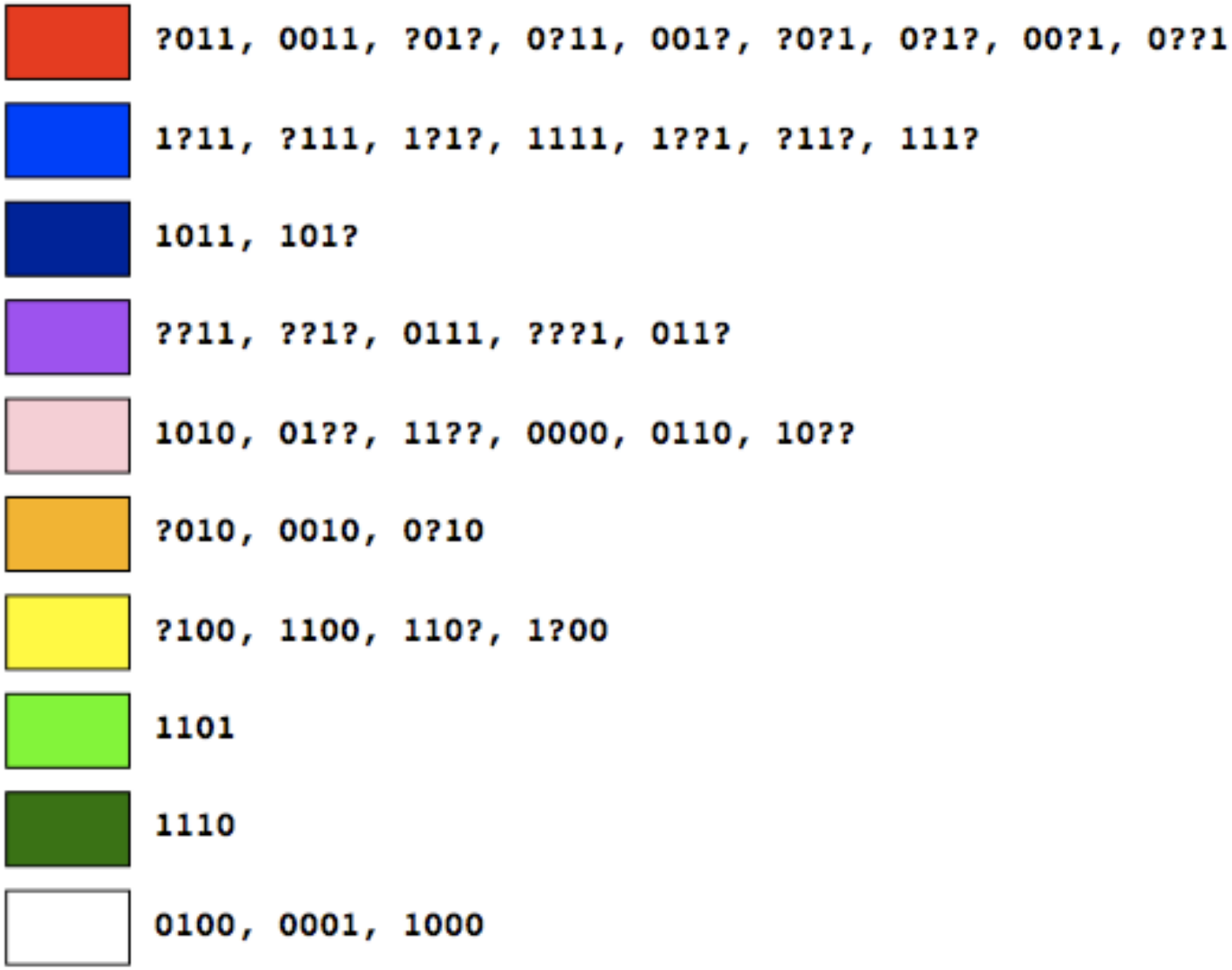
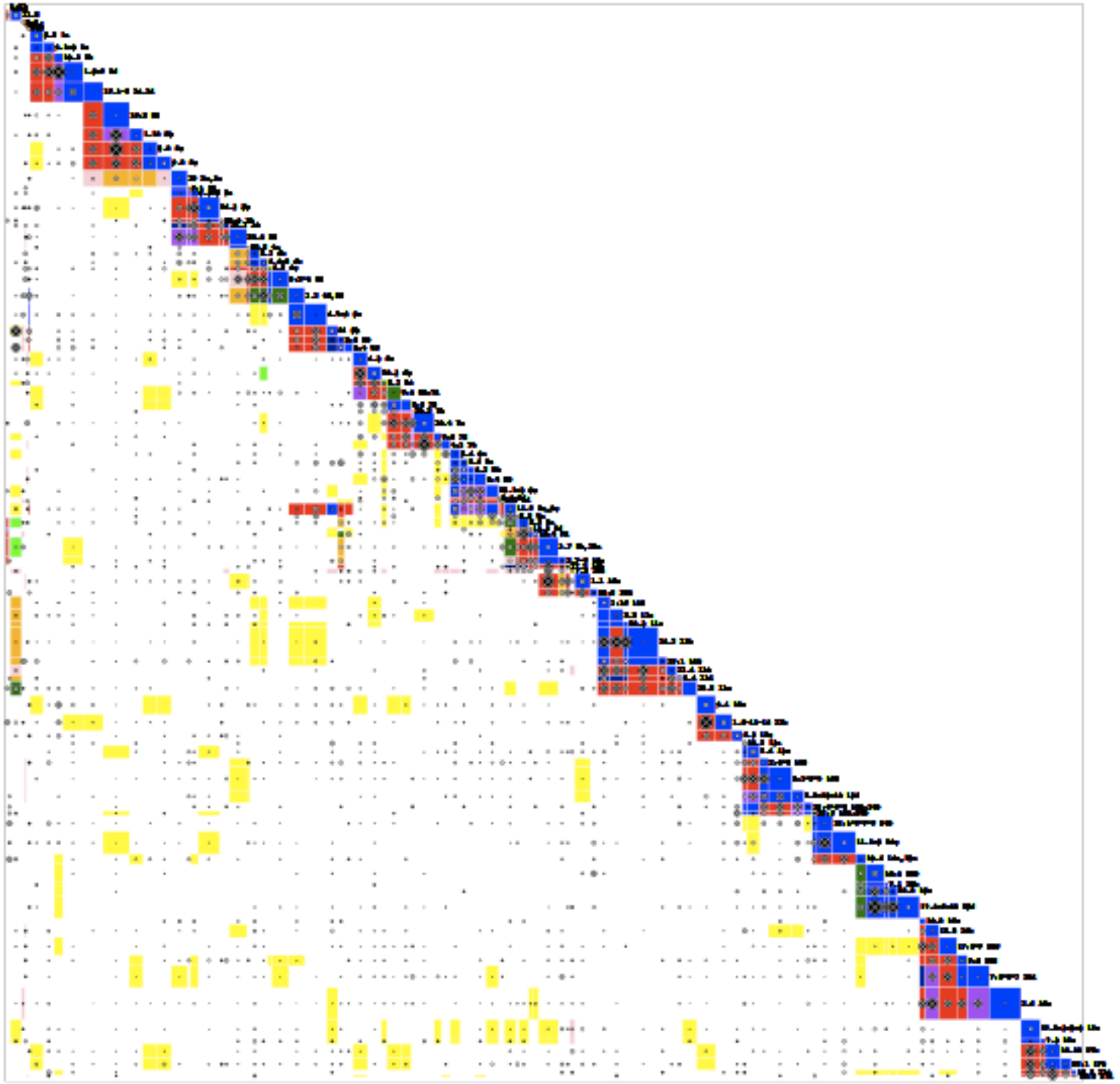


Figure S18. All-to-all comparison of human genome segments. Rows and columns represent the segments of the human genome, derived as described in the text. The color at the intersections of rows and columns indicate whether the segments are not related by conserved synteny (white), show conserved synteny to the same regions of fish, amphioxus and *Nematostella* genomes (blue), show conserved synteny to two distinct regions in fish genomes, but the same regions of amphioxus and *Nematostella* genomes (red). Orange indicates that only amphioxus places the segments together, indicating either a chromosomal fusion on the amphioxus lineage, or a vertebrate stem divergence, but without statistically-significant confirmation in *Nematostella*. Circles indicate the number of ohnolog pairs relating two segments, with area proportional to observed number divided by the expected number under a random model, given the number of genes on each segment. The inscribed circle provides a visual indication of the uncertainty in the estimate, by showing the size that the circle would take for N - sqrt(N) ohnologs. Circles marked with an "X" have a significant ($p<0.05$) excess of ohnologs relative to the expectation of the random model.

Supplementary Figure 18



Figures S19. Relationships among the segments assigned CLGs 1-9. Detailed views of the matrix shown in Supplementary Figure S18, showing each of the first 9 chordate linkage groups and their component human chromosomal segments.

Supplementary Figure

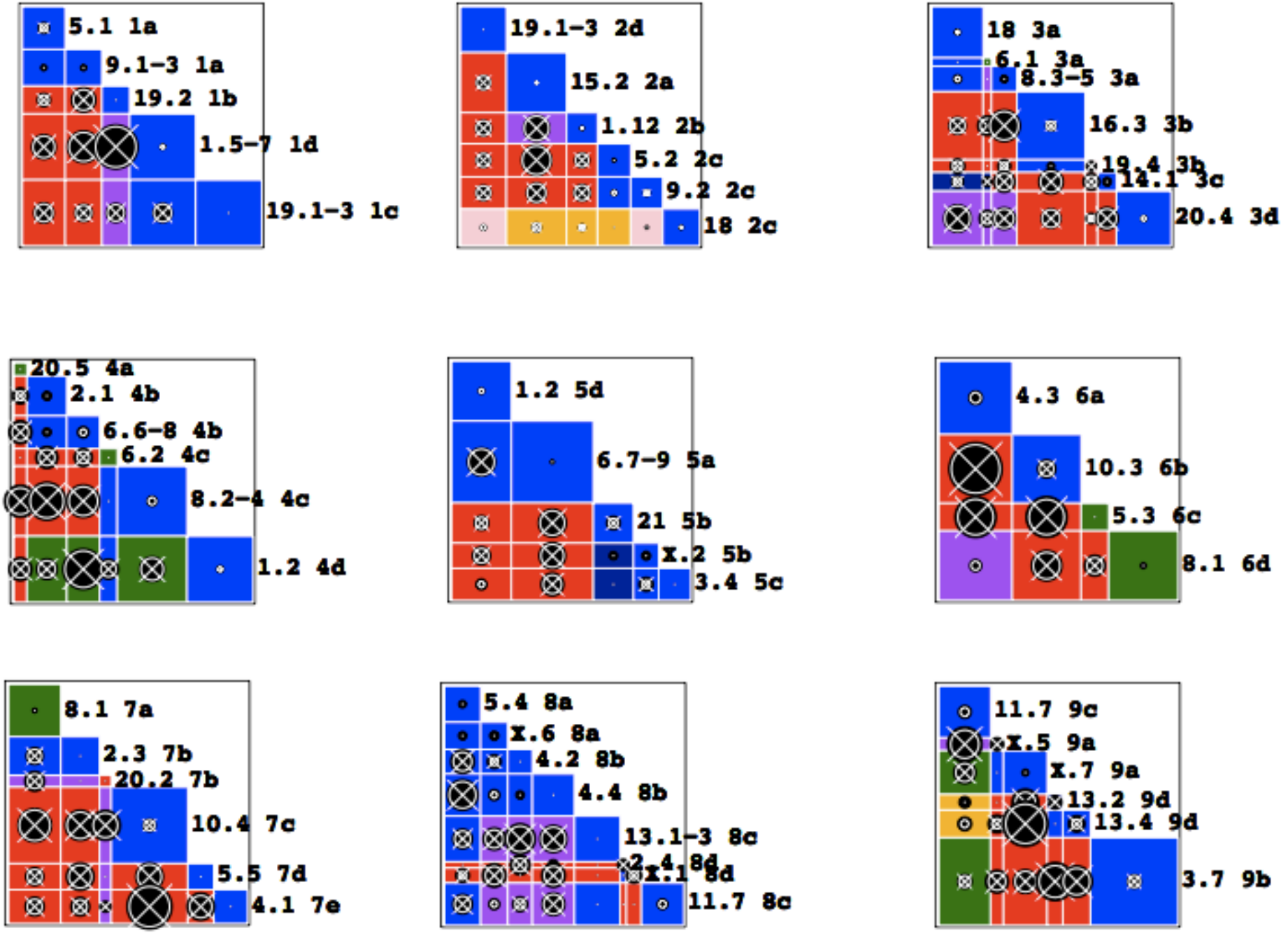


Figure S20. Relationships among the segments assigned CLGs 10-17. Detailed views of the matrix shown in Supplementary Figure S18, showing chordate linkage groups 10-17 and their component human chromosomal segments.

Supplementary Figure 20

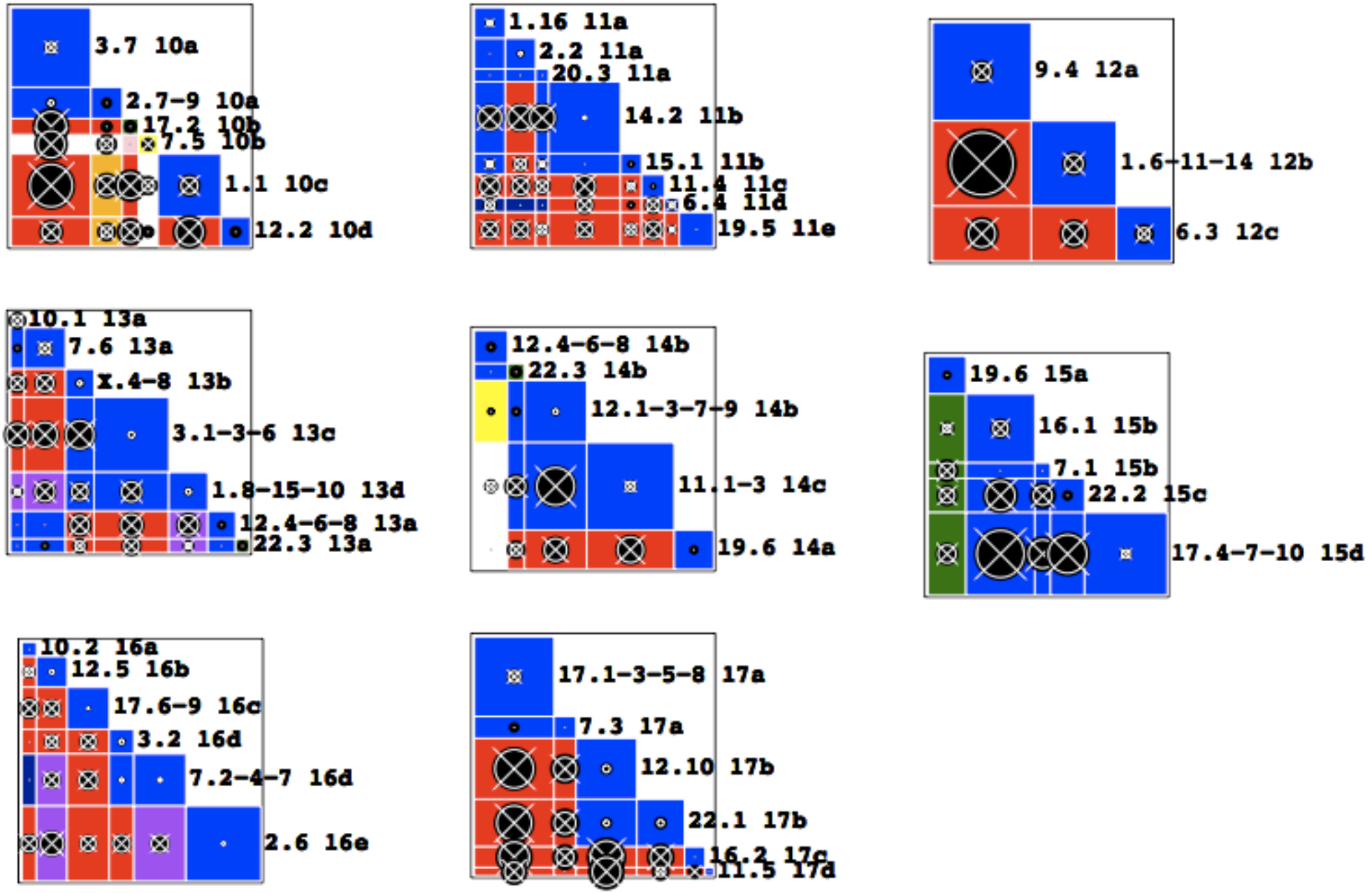


Figure S21. Lineage X divergence relative to 2R predicts gene tree topology statistics. For the three possible positions of the divergence of a linage "X" relative to two rounds of gene duplication, the figure illustrates the resulting gene tree topology (top row of trees), and lists the predicted ratio of three-gene trees, including oneortholog in X, and two randomly selected genes from the set of paralogs created by the duplications (upper row of ratios and percentages). The bottom row of trees and the lower row of rations shows the impact on the predicted tree statistics of the loss of one of theparalogs created by the first round of gene duplication, as is expected for many individual genes following whole genome duplication.

Supplementary Figure 21

Lineage X divergence relative to 2R predicts gene tree topology statistics.

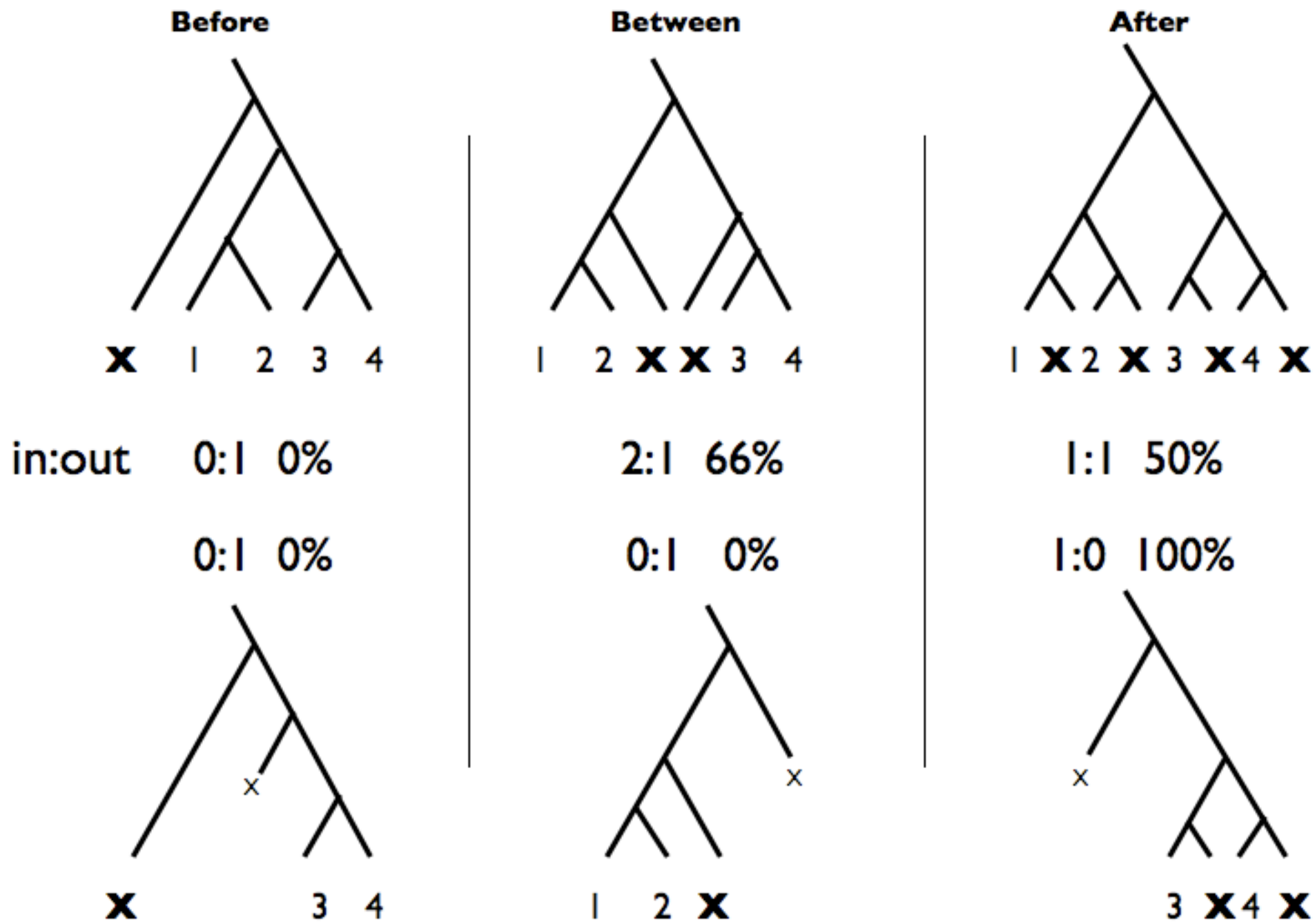


Figure S22. Depth of coverage in the WGS assembly. The distribution of the number of placed reads covering each position in the WGS assembly, for the 200 largest scaffolds in red (approximately half of the total length of the assembly), and all the other scaffolds (green). The peak at zero depth represent the gaps in the scaffolds.

Supplementary Figure 22

Depth of coverage in the WGS assembly.

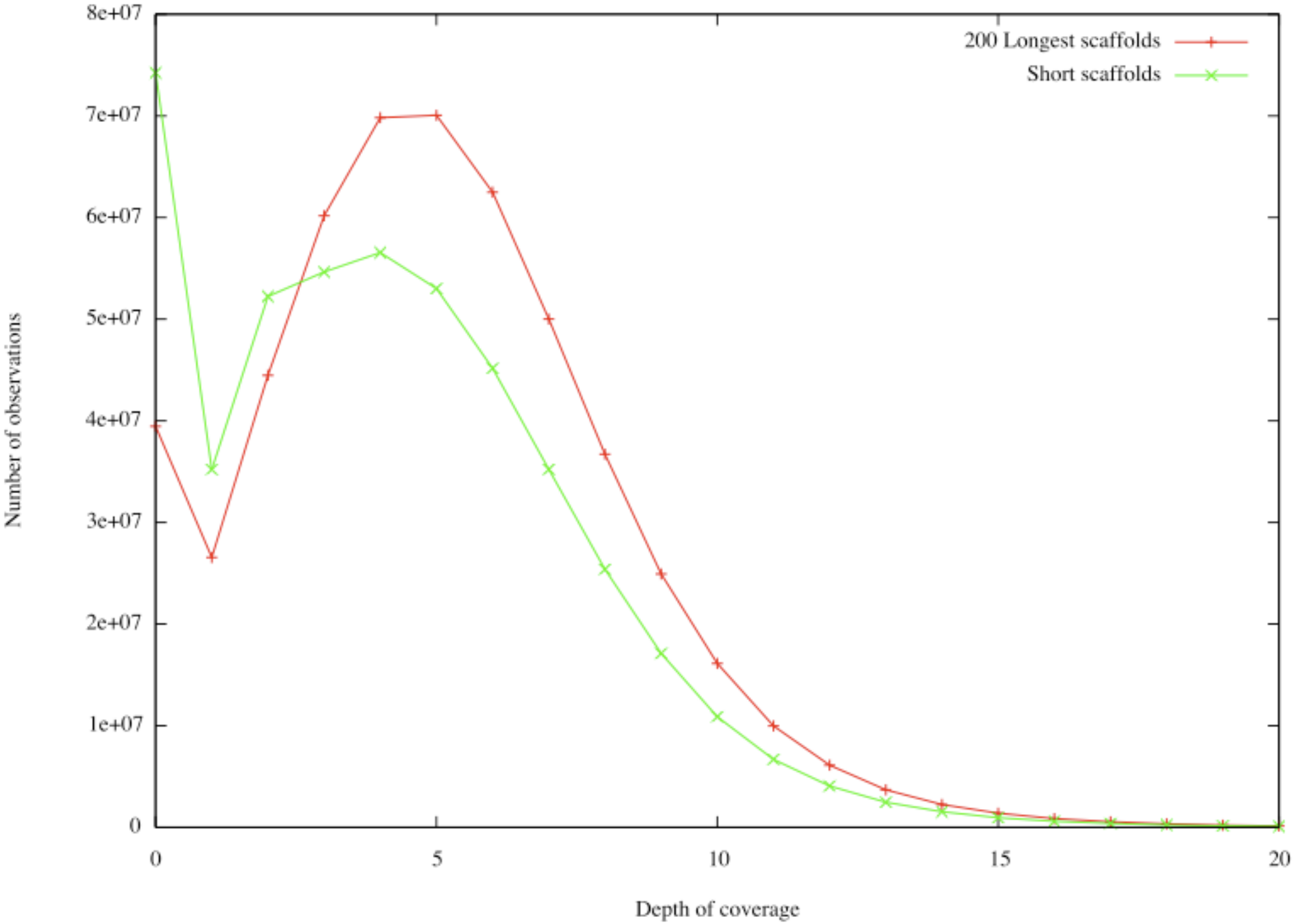


Figure S23. Depth of coverage in 10 kb windows along the WGS assembly. The same as Supplementary Figure S22 except in non-overlapping 10 kb windows along the assembly.

Supplementary Figure 23

Depth of coverage in 10kb windows along the WGS assembly.

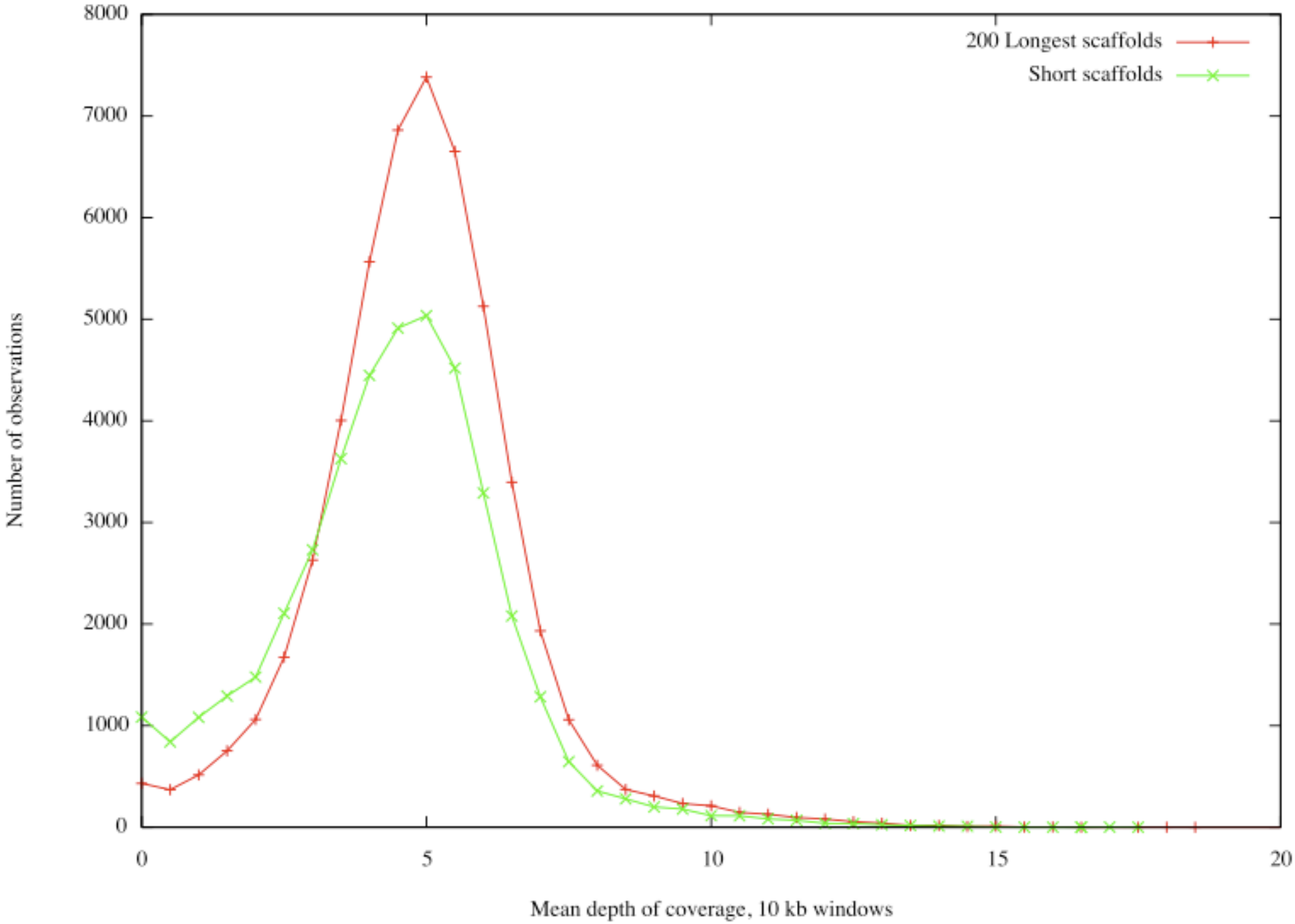


Figure S24. Number of overlapping reads for all (green) and WGS-unplaced (red) reads. All against all read-read alignments were performed as part of the JAZZ assembly process as described in [1]. A peak at 2d overlaps per read is expected for a random distribution of shotgun reads and depth d. For all reads (green), the peak occurs at ~11-12 overlaps/read, consistent with ~5.7X depth per haplotype. Reads that are unplaced in the assembly (red) are enriched for low overlap per read.

Supplementary Figure 24

Number of overlapping reads for all, and WGS-unplaced reads.

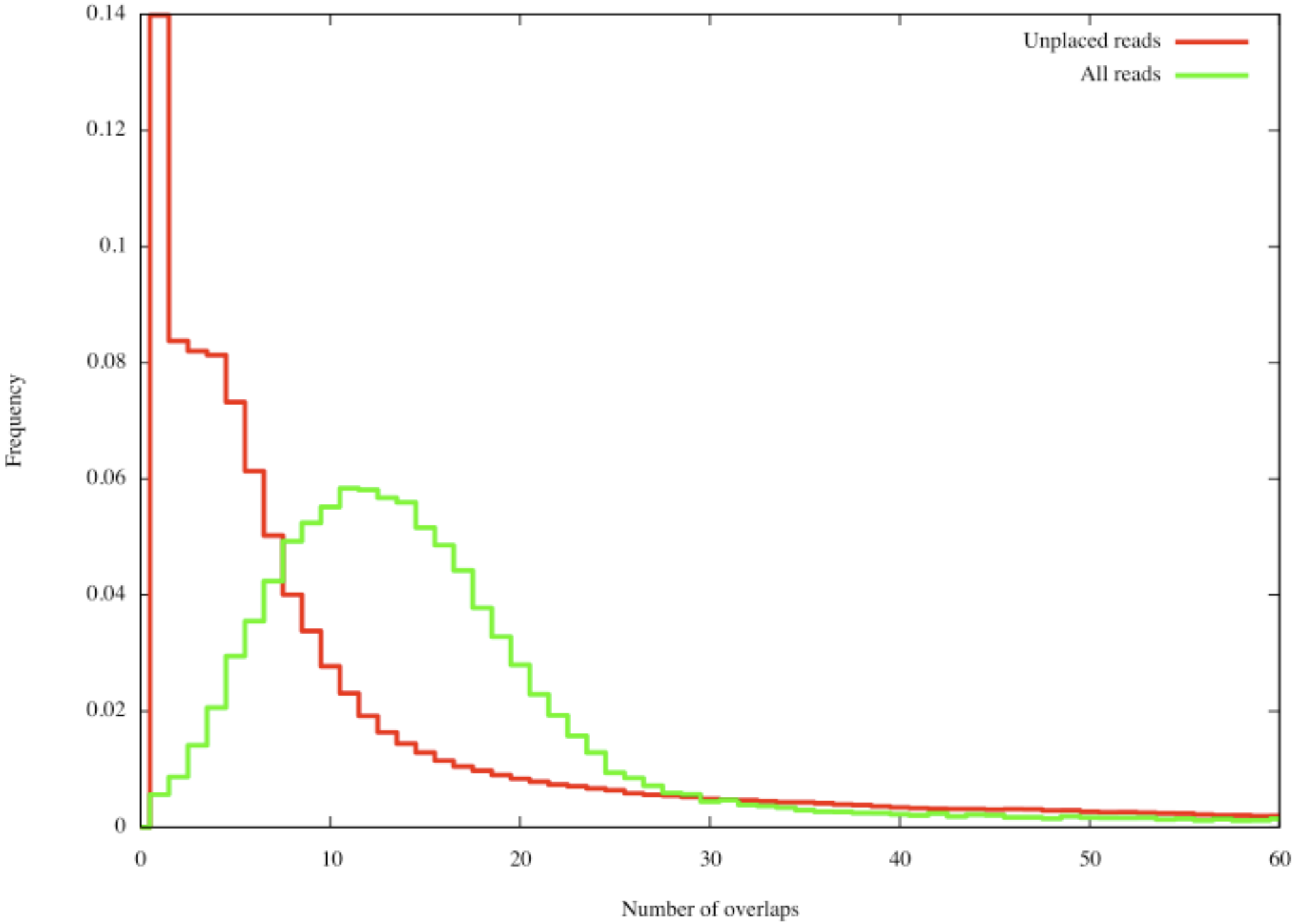


Figure S25. Distribution of sizes of short indel polymorphisms. Small indel polymorphisms are the most common (note logarithmic scale on y-axis).

Supplementary Figure 25

Distribution of sizes of short indel polymorphisms.

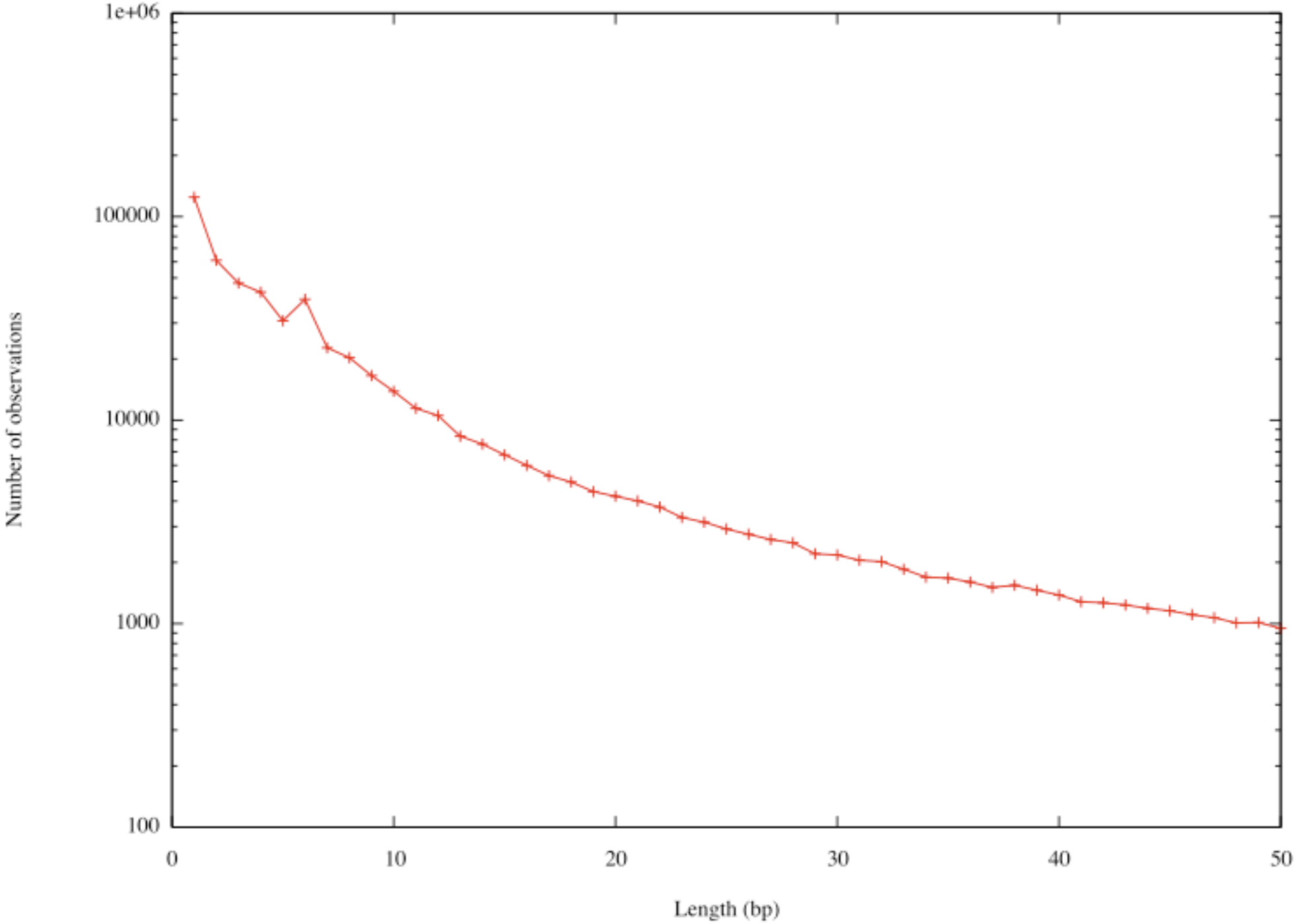


Figure S26. Distribution of polymorphic indel sizes. Indel polymorphisms show declining frequency with length of indel over kilobase scales. Note logarithmic scale on y-axis. Discrete peaks at 210, 348, 441, and 580 bp reflect polymorphisms in the presence/absence of specific short interspersed repetitive elements.

Supplementary Figure 26

Distribution of polymorphic indel sizes.

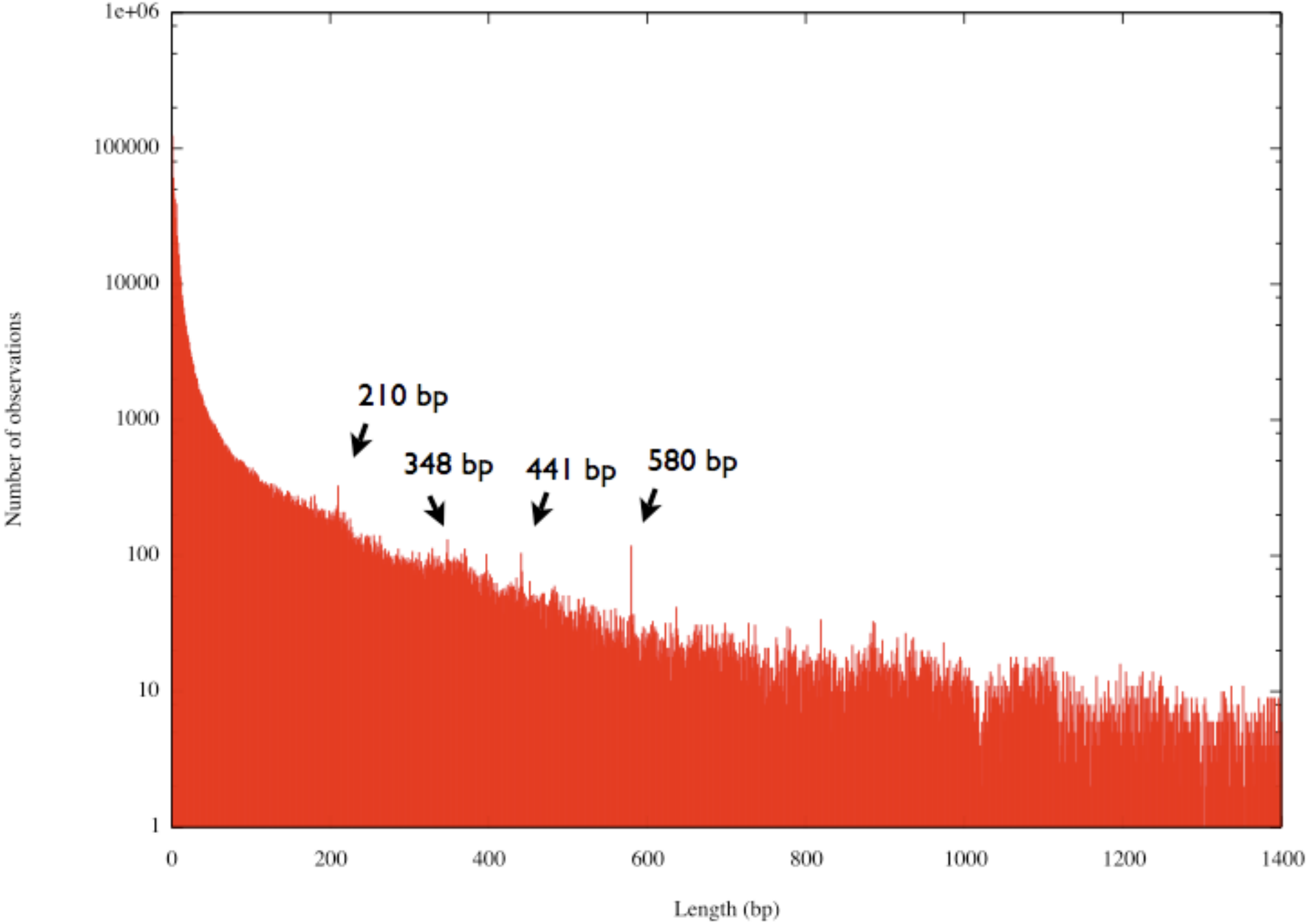


Figure S27. Example of quadruple conserved synteny. Example of quadruple conserved synteny of the human genome (portions of chromosome 4, 5, 8, and 10, compared with amphioxus version 1 scaffolds. Note extensive loss of ancient gene duplicates; For genes shown in the figure, 30% of ancestral chordate genes survive as multiple “ohnologs” in the modern human genome.

Supplementary Figure 27

Example of quadruple conserved synteny.

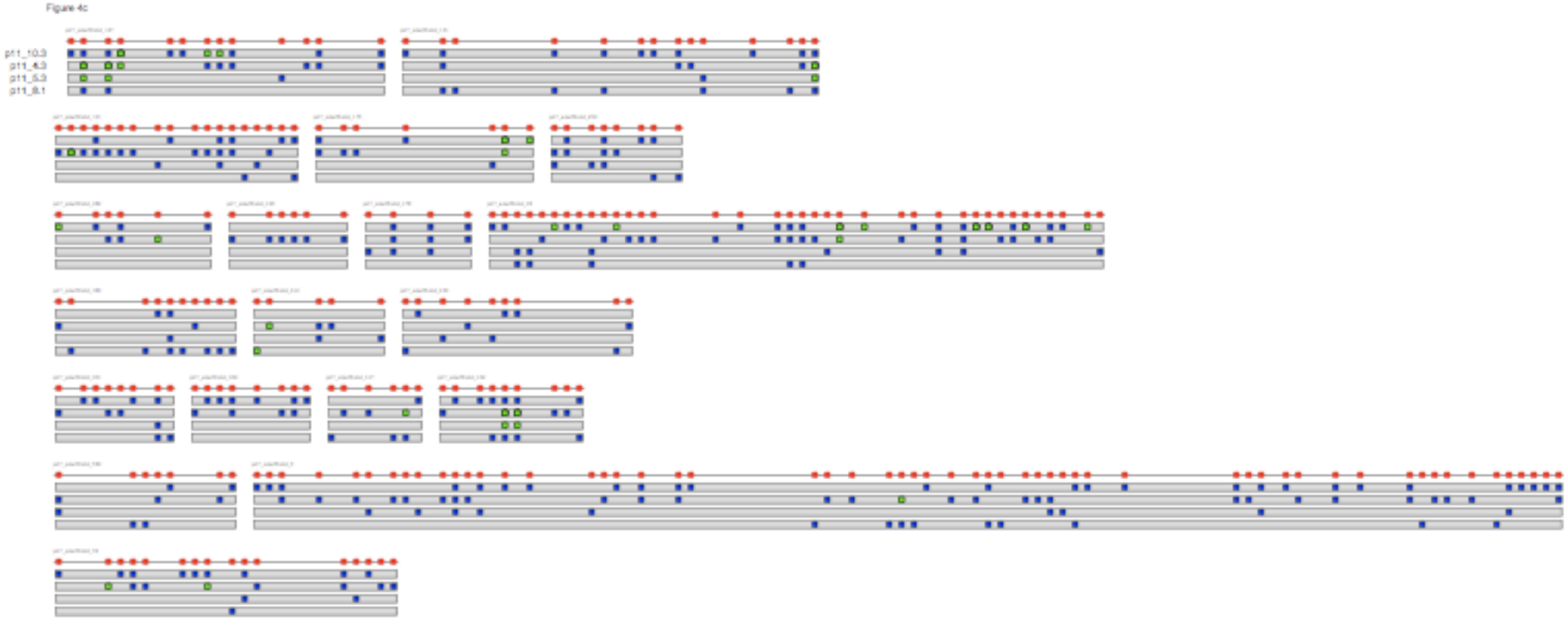


Figure S28. Mapping human genome segment divergences to the vertebrate stem. fish, indicating duplication on the Middle panel shows case of two distinct pairs of teleost fish-tetrapod chromosomal segments that arose by duplication from a single ancestral segment on the jawed vertebrate stem. In this case, each human segment (white and black) have a corresponding distinct segment in teleost fish, but both show conserved synteny to the same amphioxus and sea anemone scaffolds. The panel to the left shows two tetrapod segments that show conserved synteny to the same segments of teleosttetrapod stem. The panel to the right shows the case of a chromosomal fusion on the amphioxus lineage, which mimics the syntenic relationships between amphioxus, tetrapods, and teleost fish that are found in the middle panel. Unlike the middle panel, however, the sea anemone outgroup shows that these segments fused in amphioxus.

Supplementary Figure 28

Mapping human genome segment divergences to the vertebrate stem.

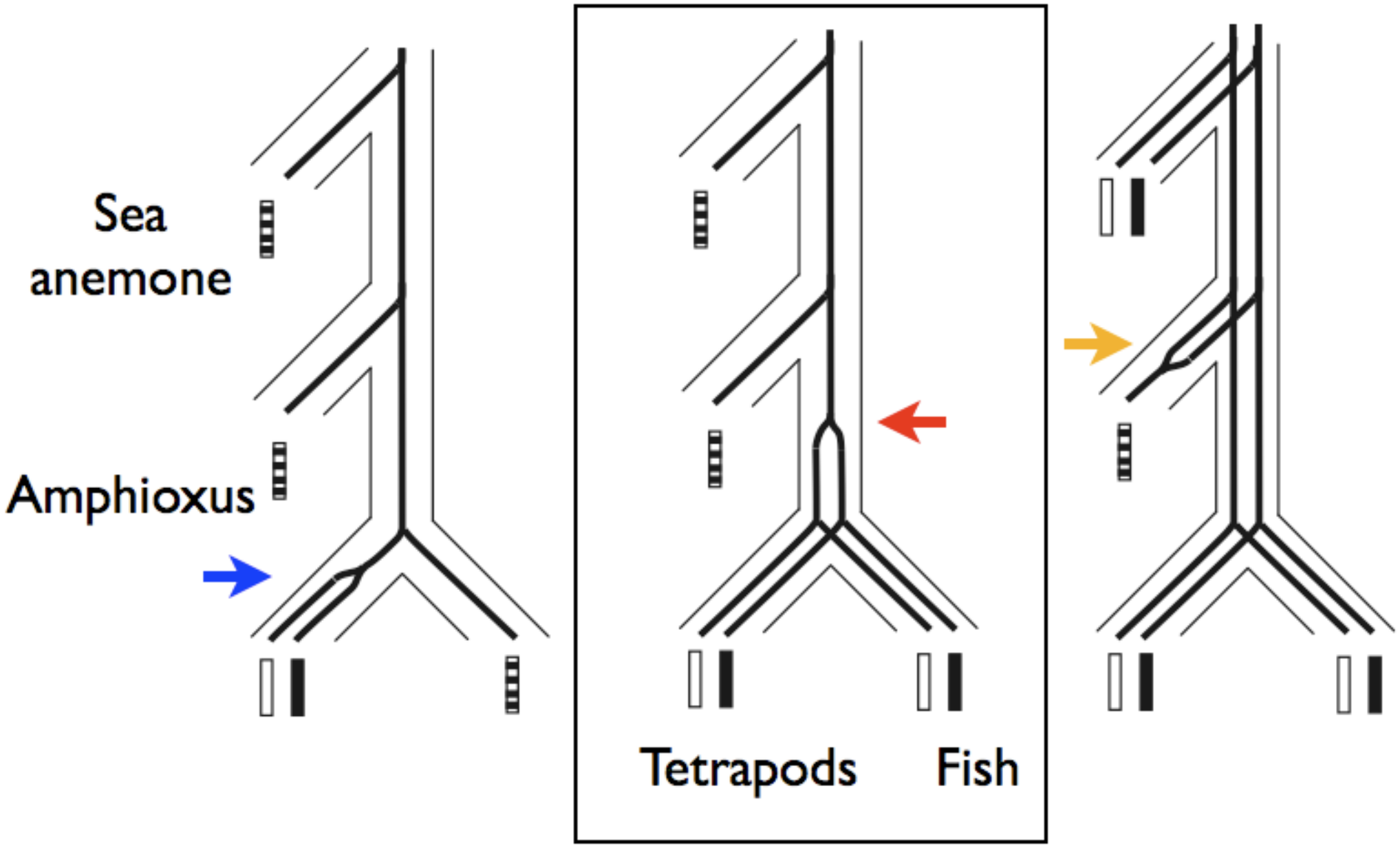
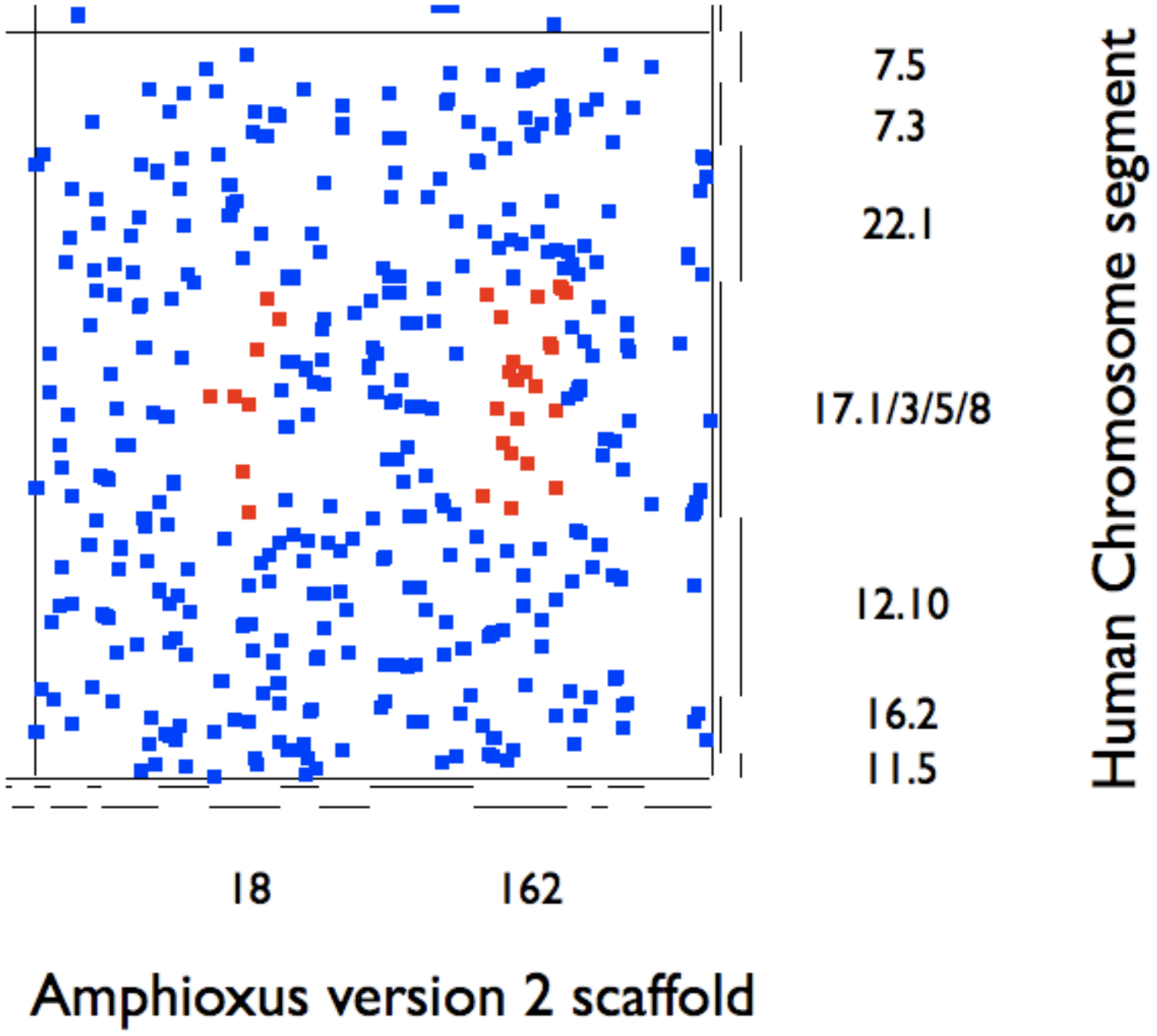


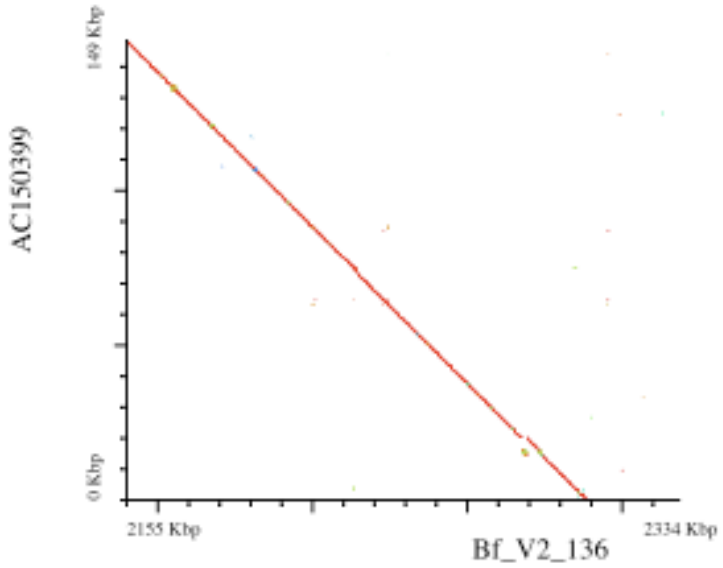
Figure S29. Detail of Supplemental Figure 64

Supplementary Figure 29

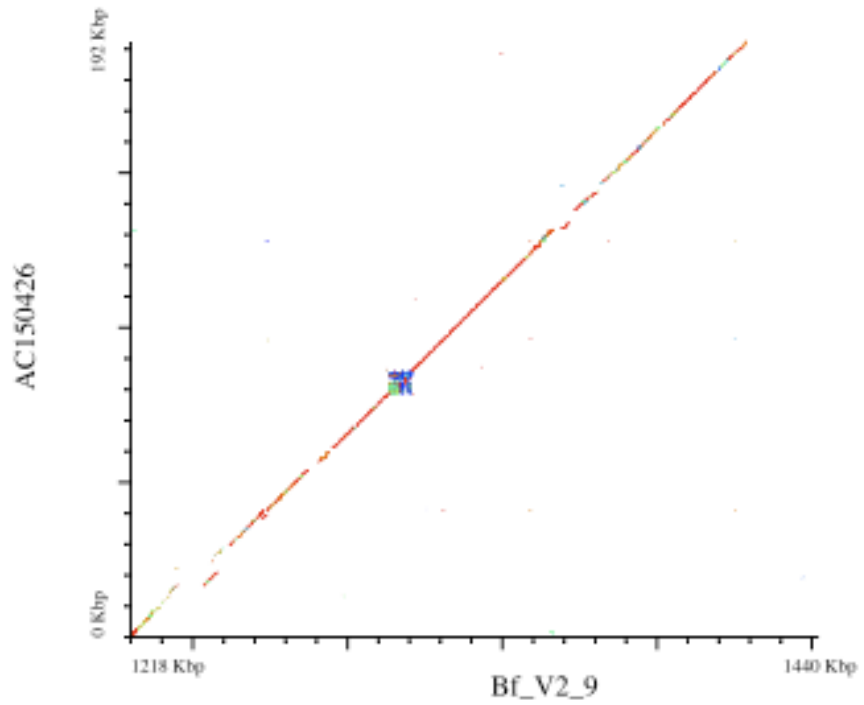


Figures S30-S62. Alignments of finished BAC clones to assembly version 2.

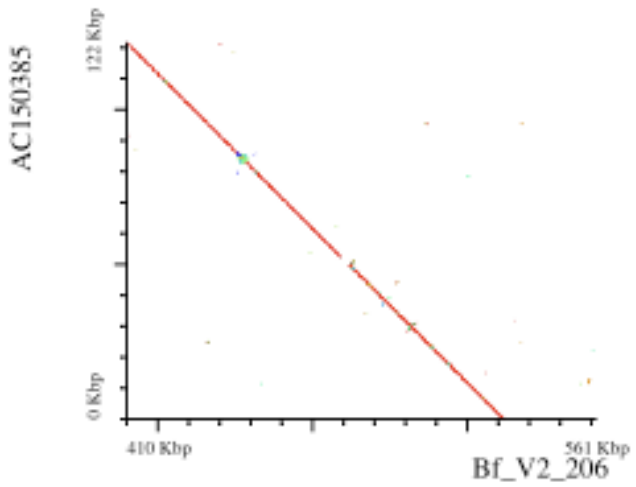
Supplementary Figure 30



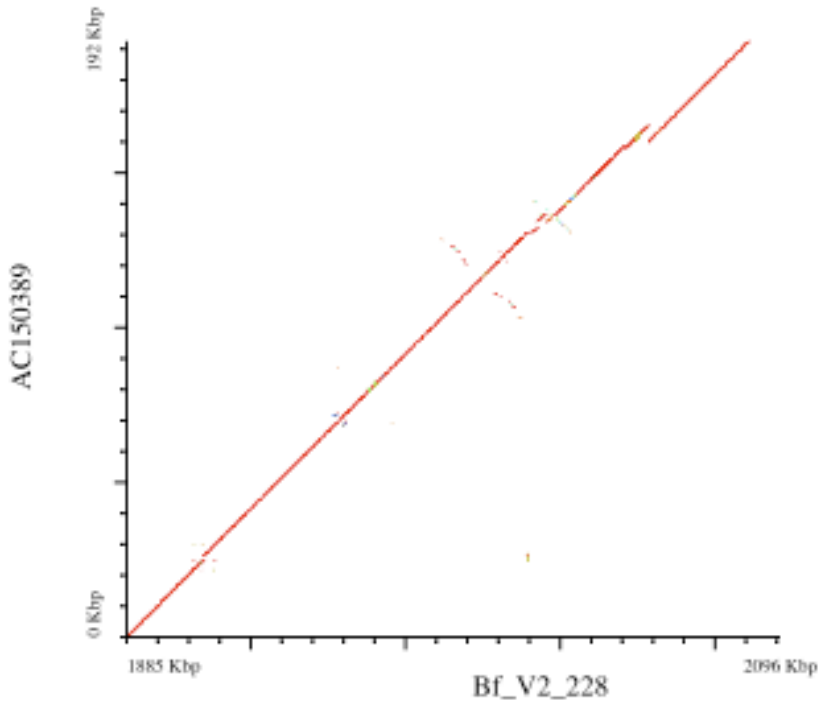
Supplementary Figure 31



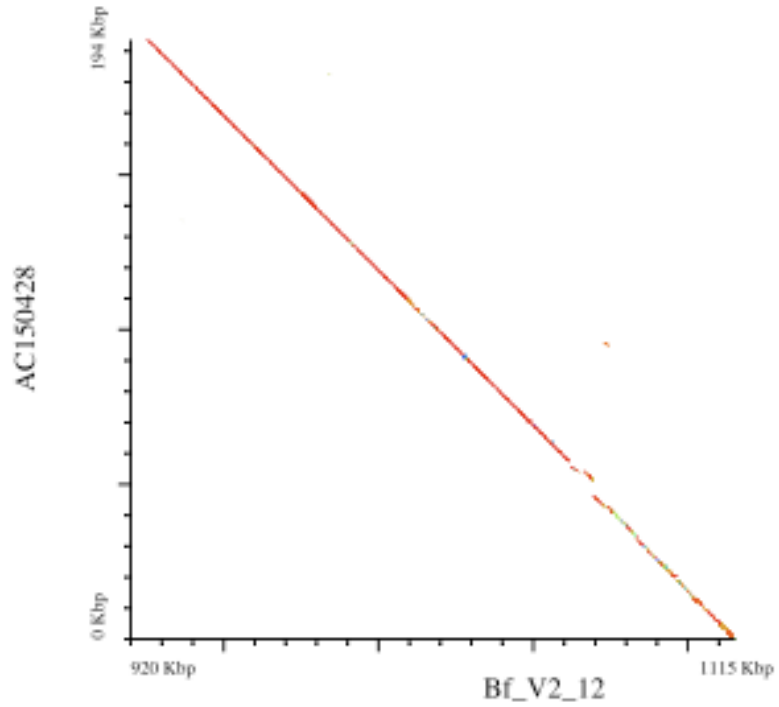
Supplementary Figure 32



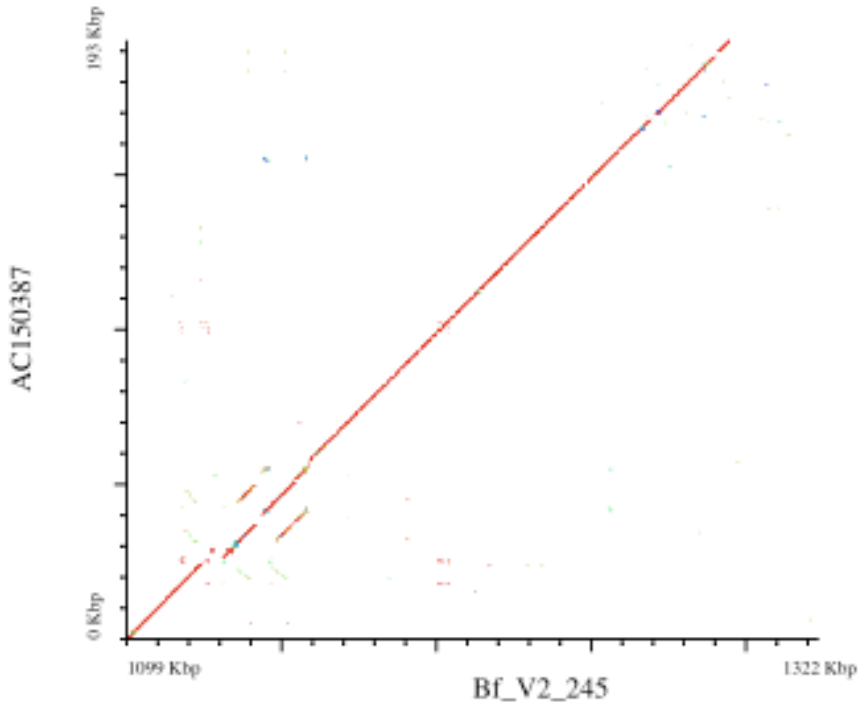
Supplementary Figure 33



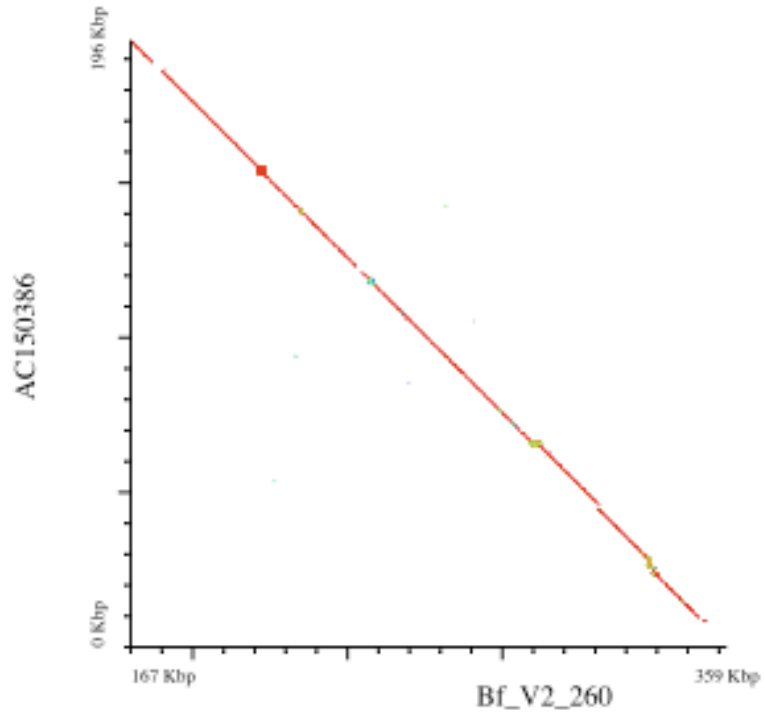
Supplementary Figure 34



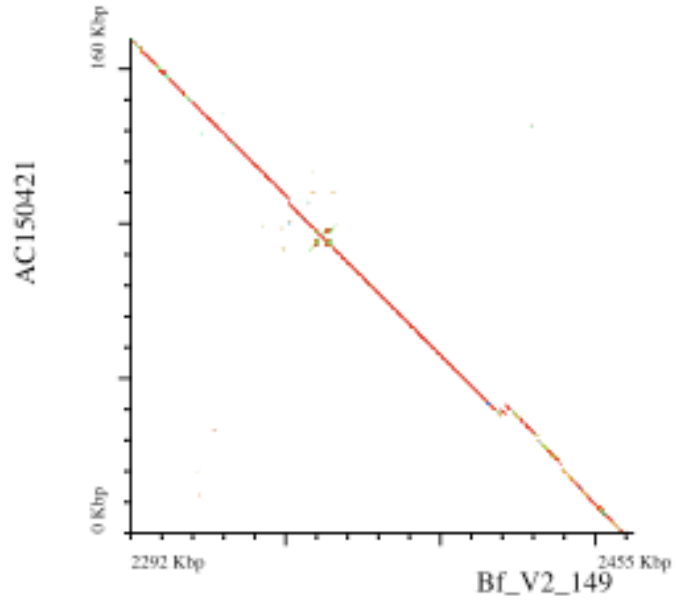
Supplementary Figure 35



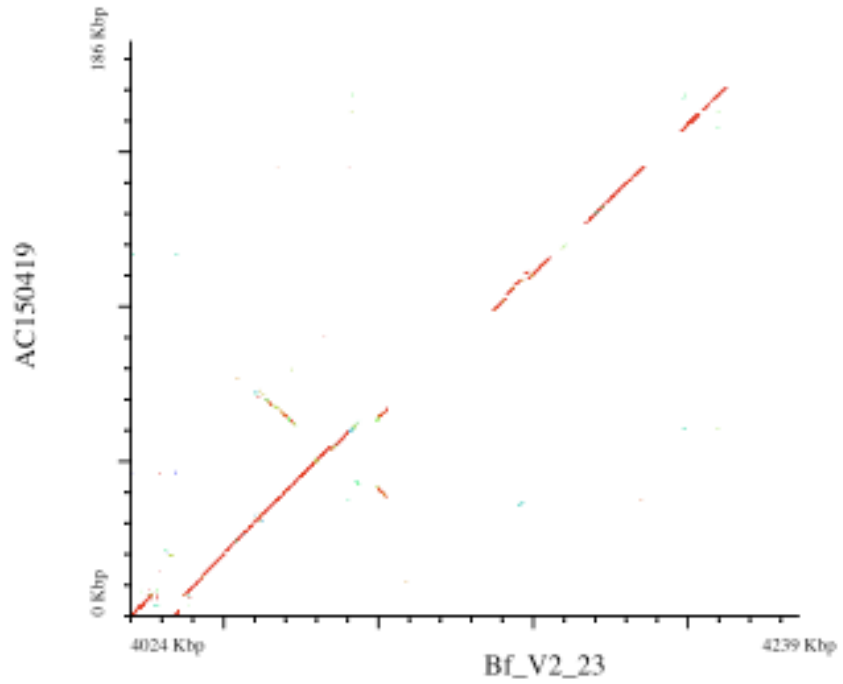
Supplementary Figure 36



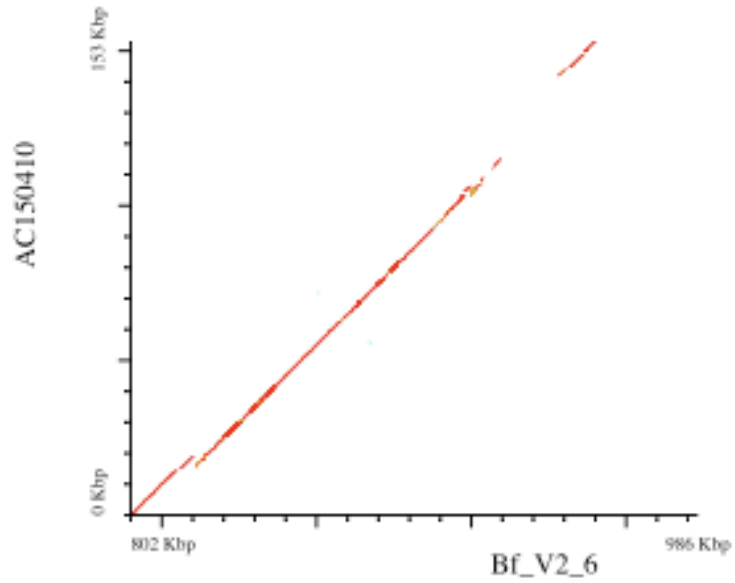
Supplementary Figure 37



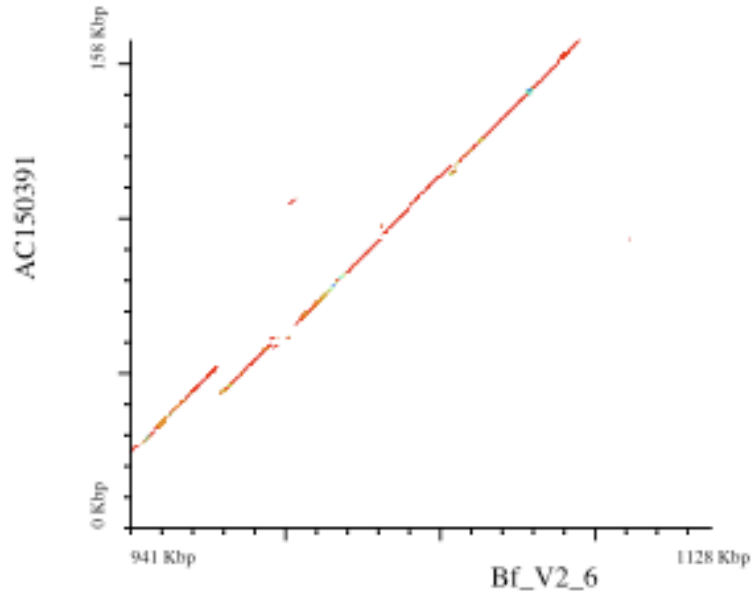
Supplementary Figure 38



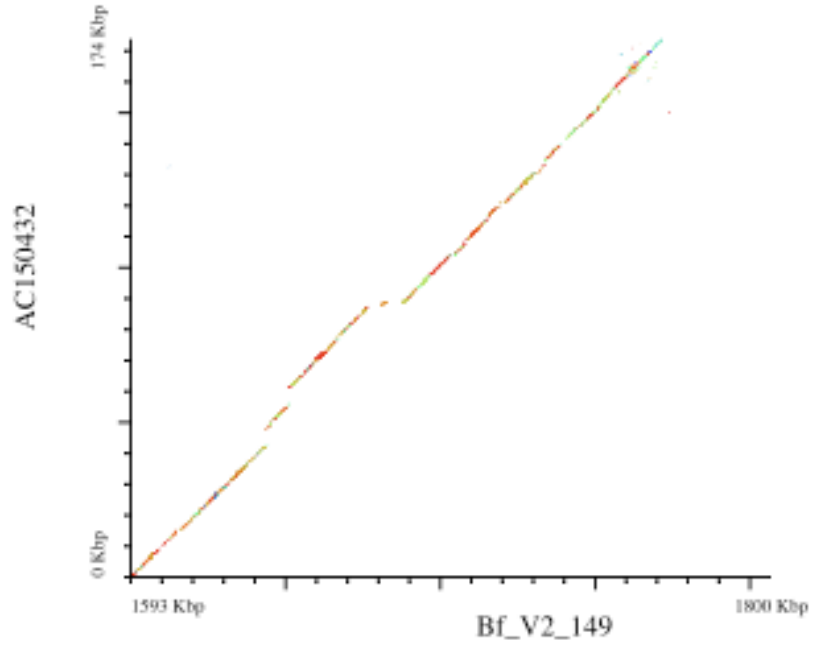
Supplementary Figure 39



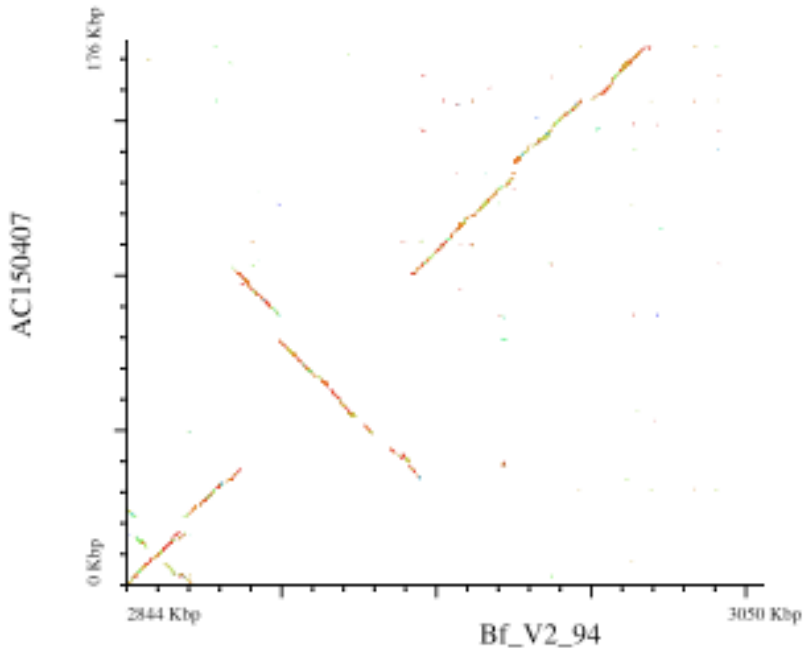
Supplementary Figure 40



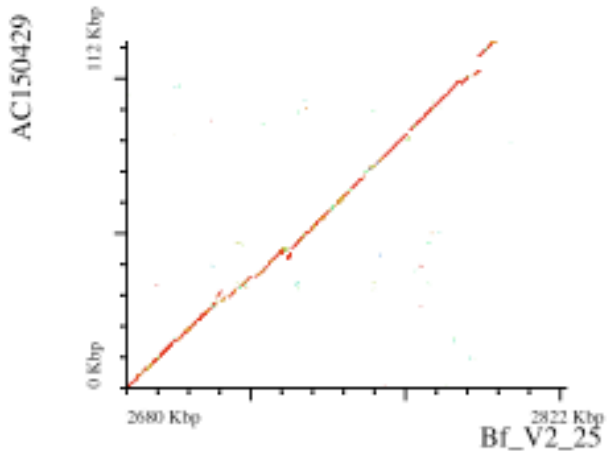
Supplementary Figure 4I



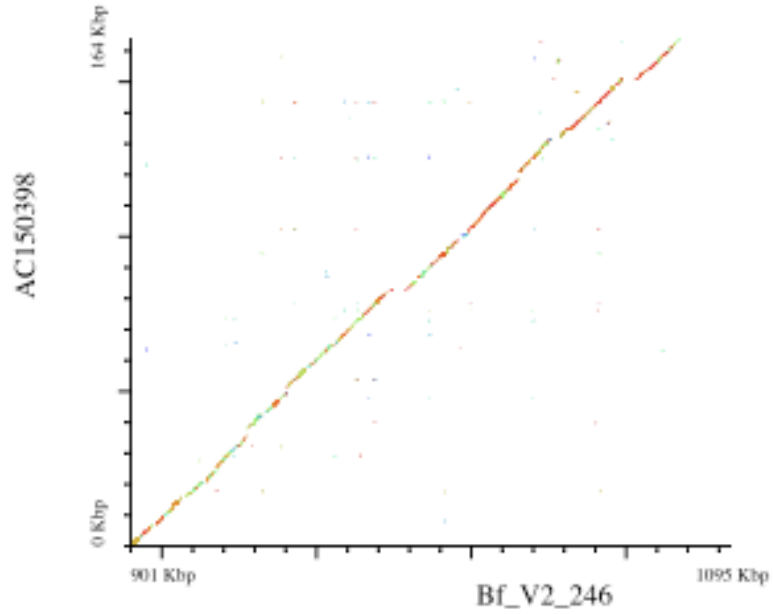
Supplementary Figure 42



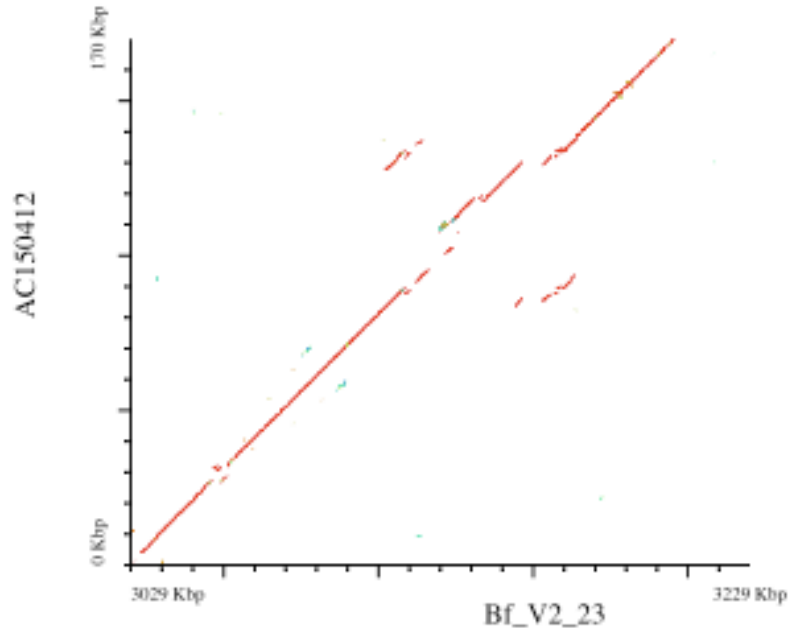
Supplementary Figure 43



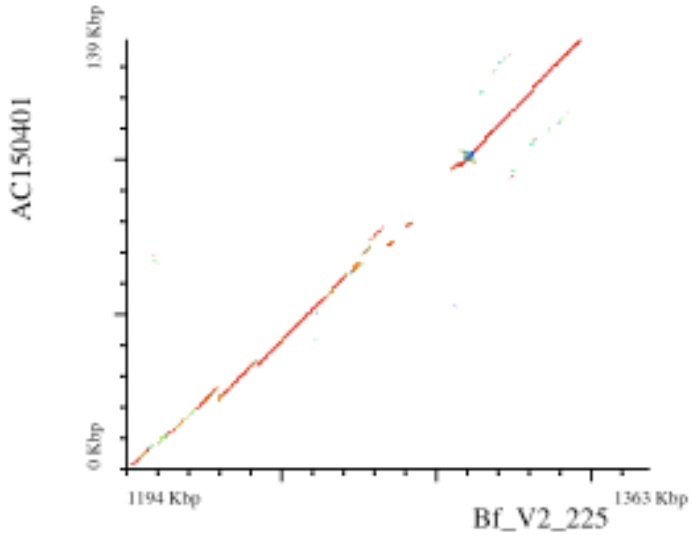
Supplementary Figure 44



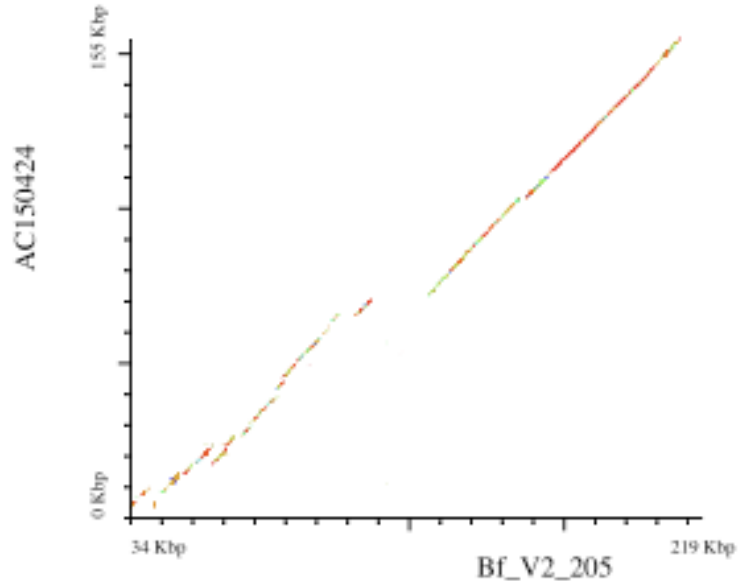
Supplementary Figure 45



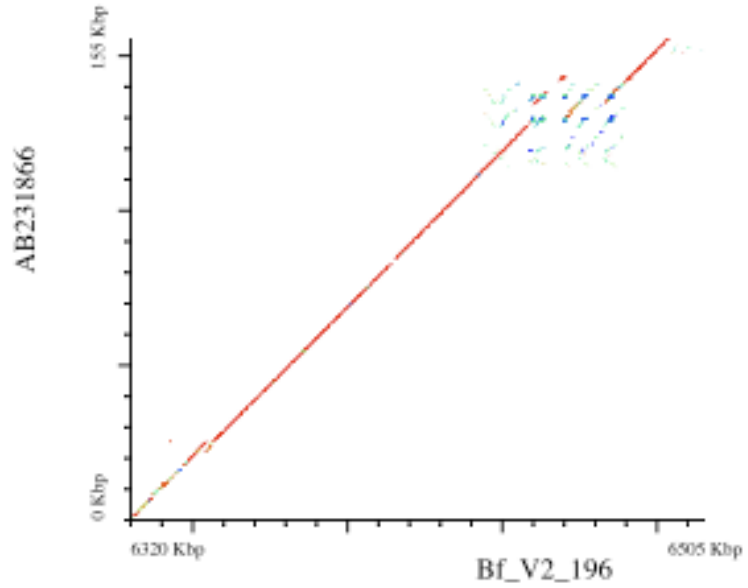
Supplementary Figure 46



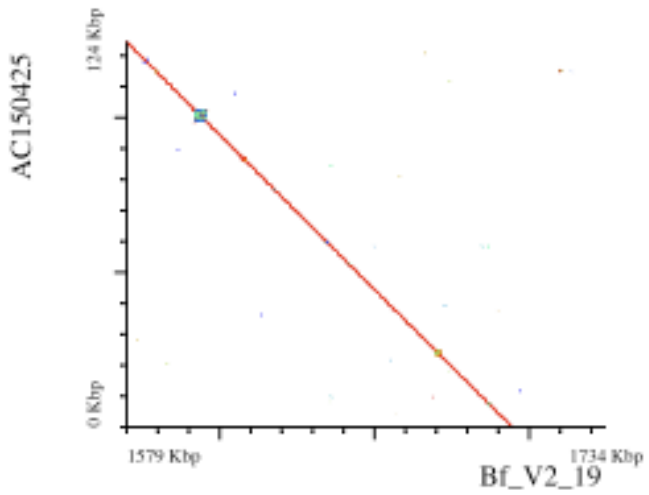
Supplementary Figure 47



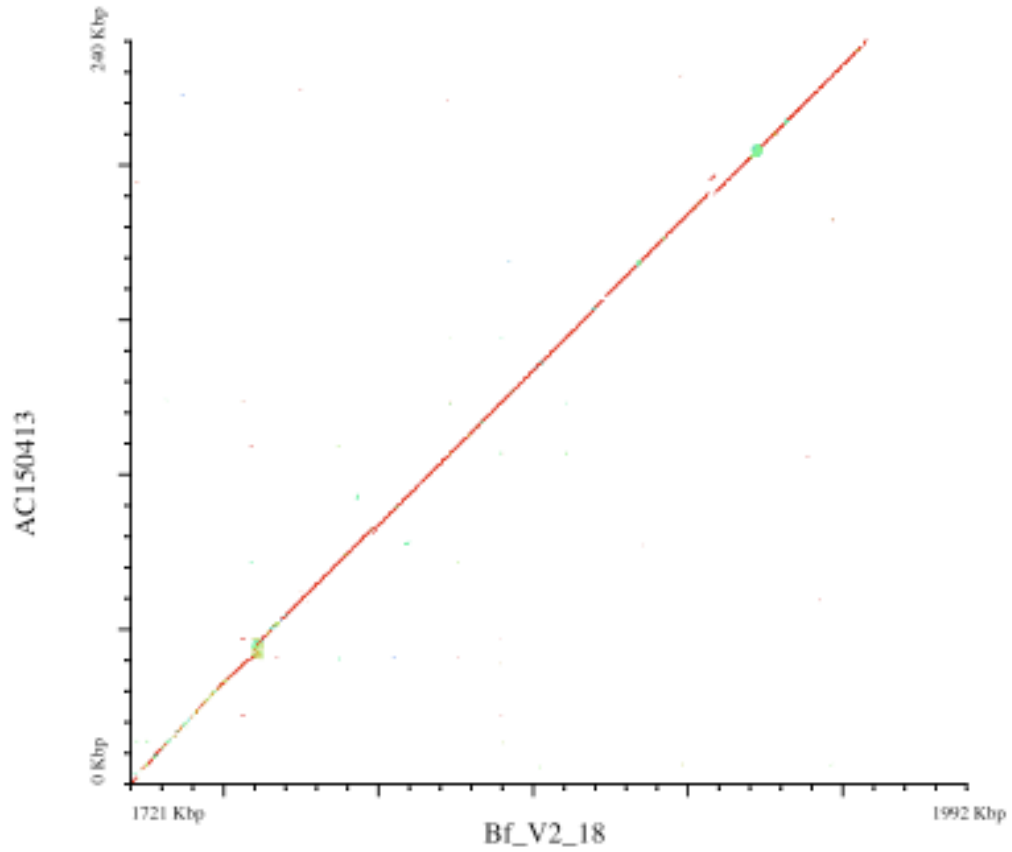
Supplementary Figure 48



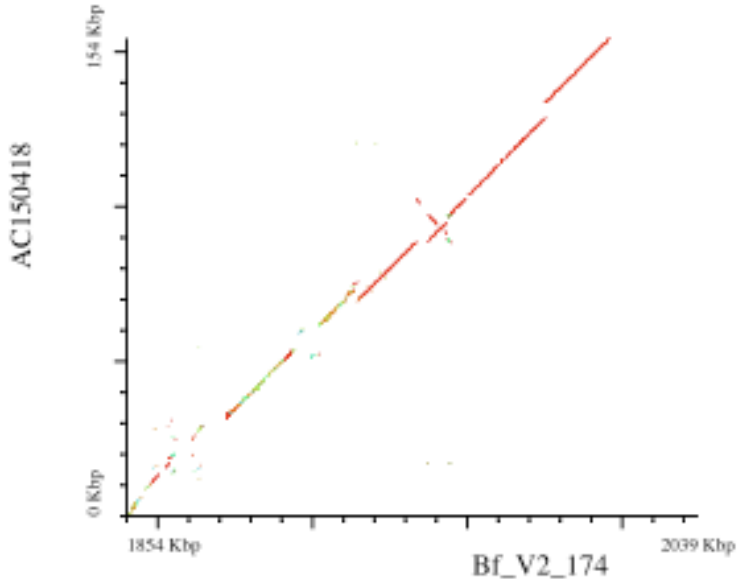
Supplementary Figure 49



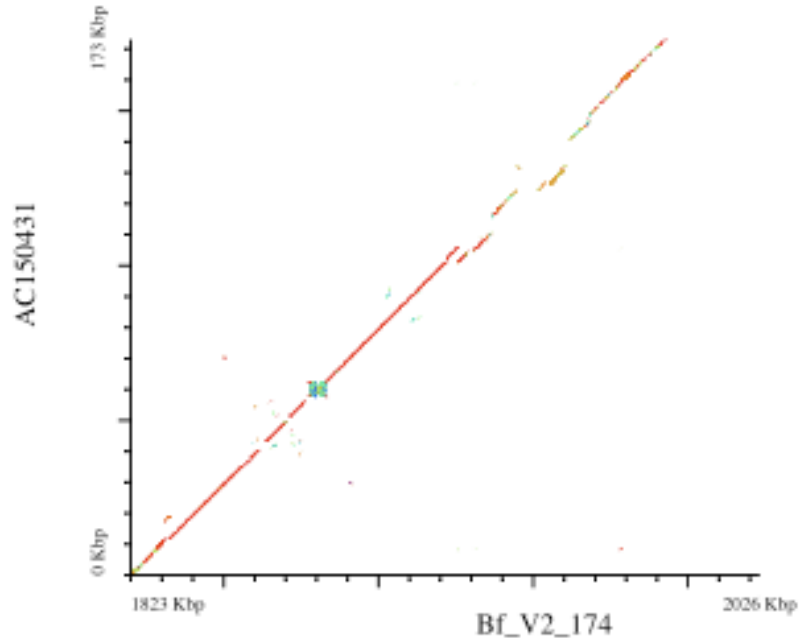
Supplementary Figure 50



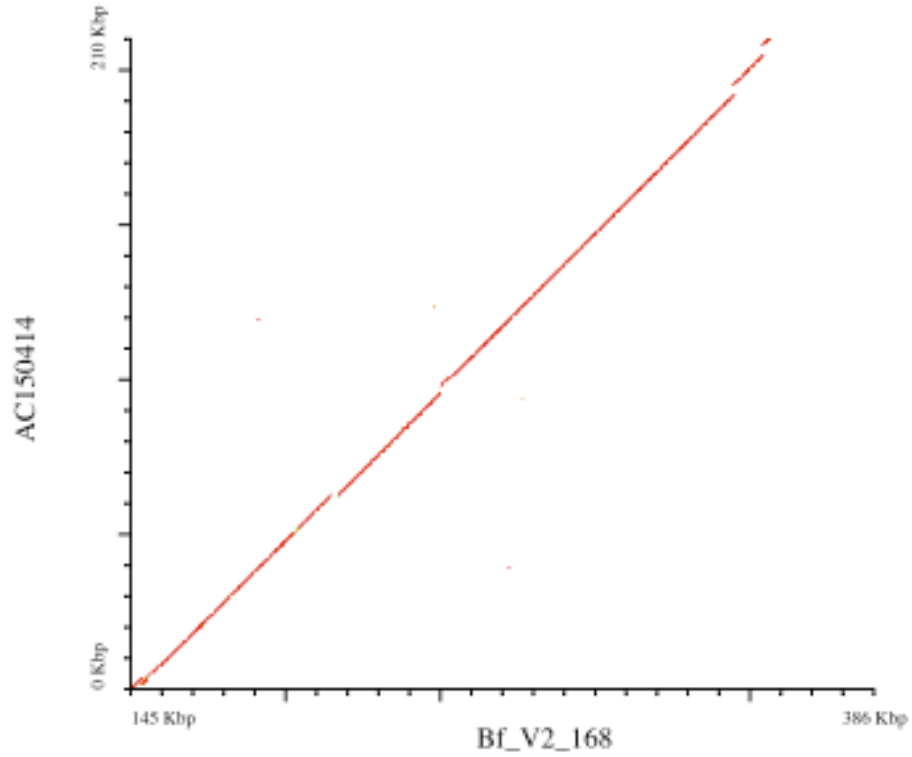
Supplementary Figure 51



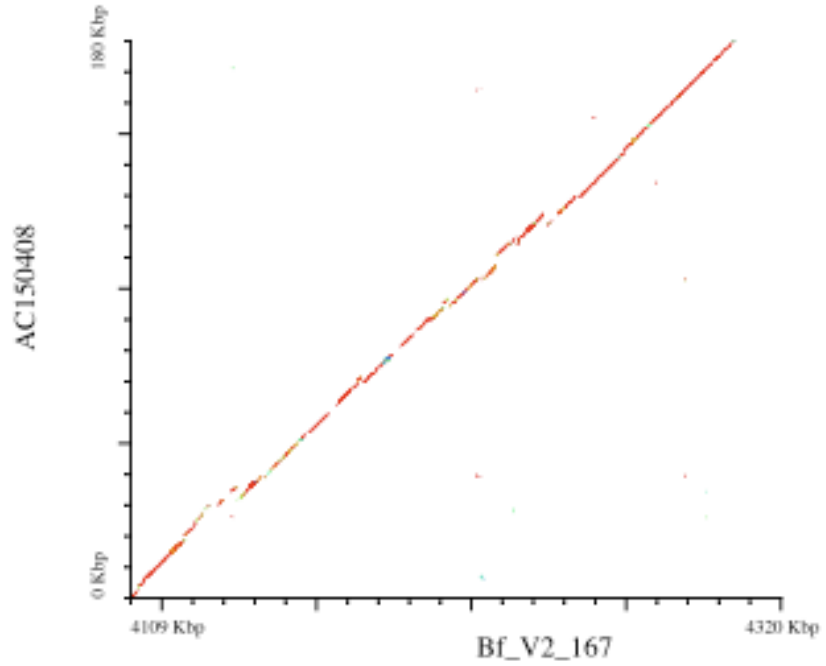
Supplementary Figure 52



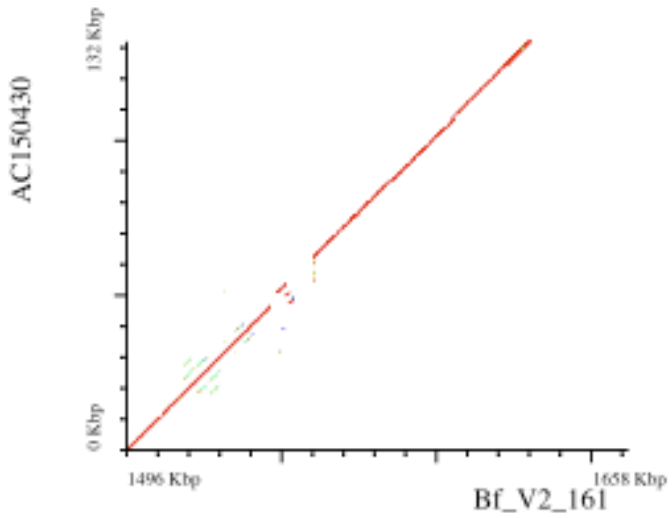
Supplementary Figure 53



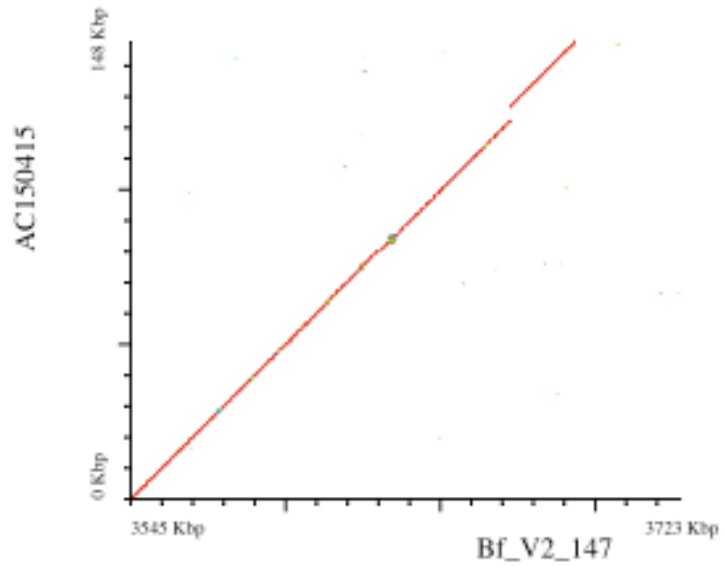
Supplementary Figure 54



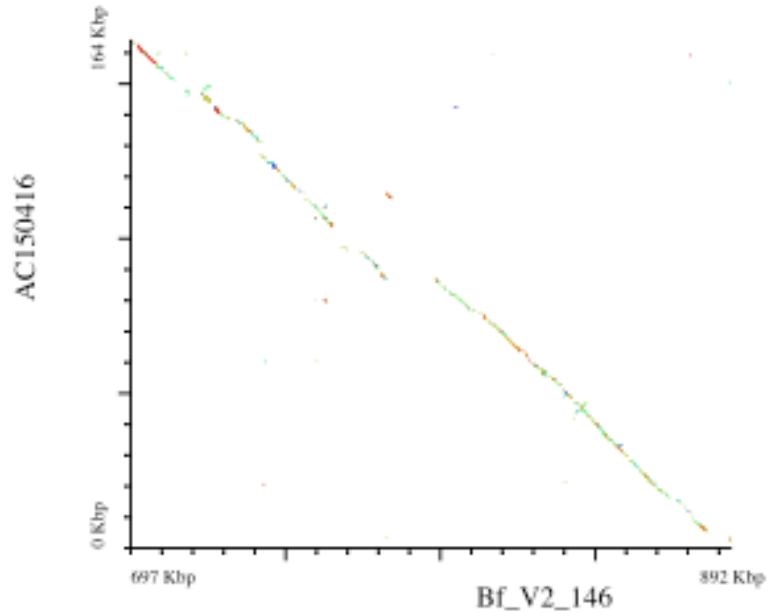
Supplementary Figure 55



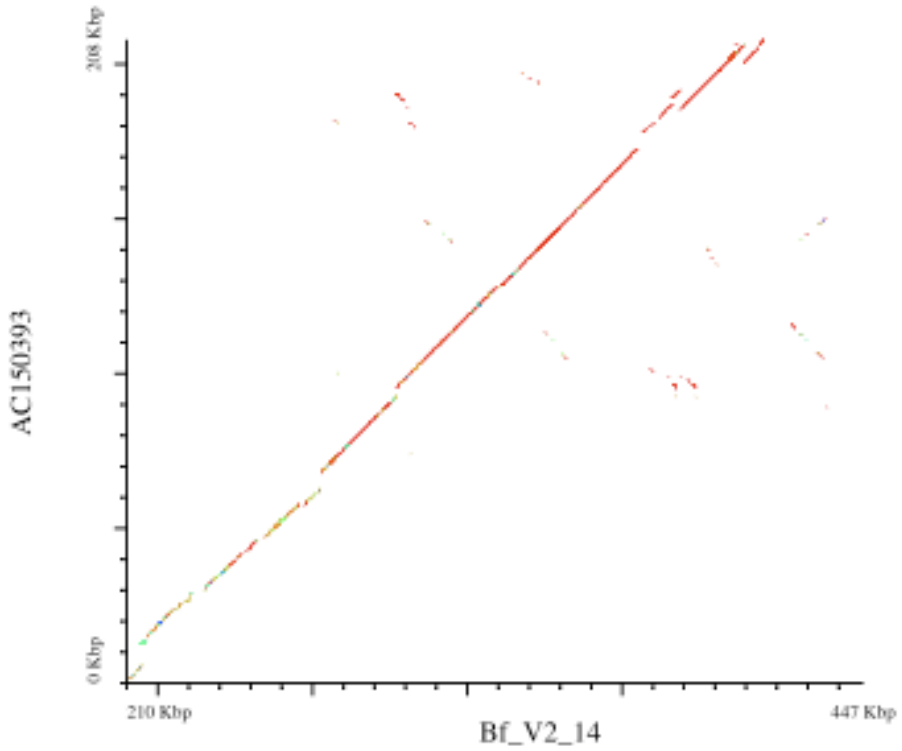
Supplementary Figure 56



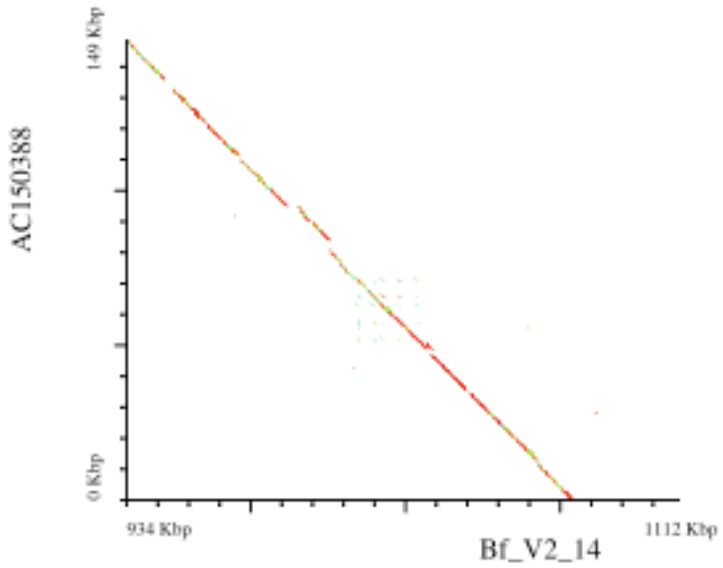
Supplementary Figure 57



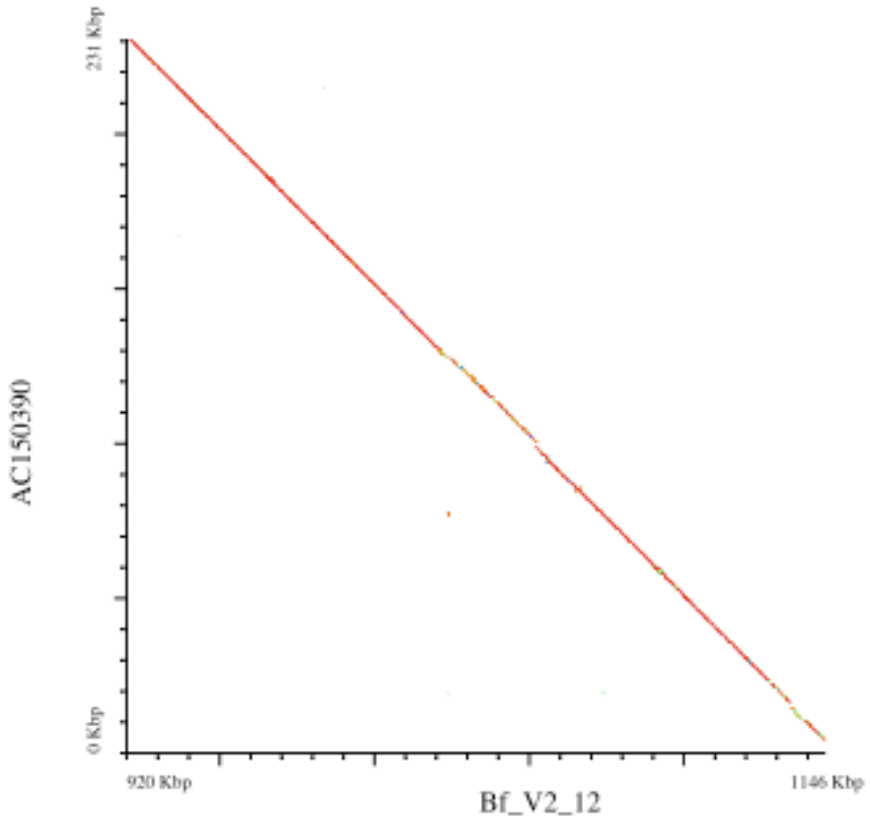
Supplementary Figure 58



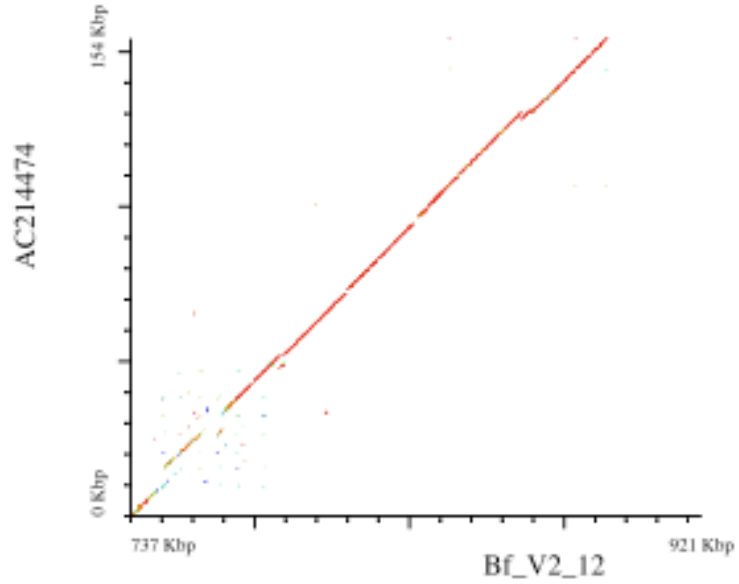
Supplementary Figure 59



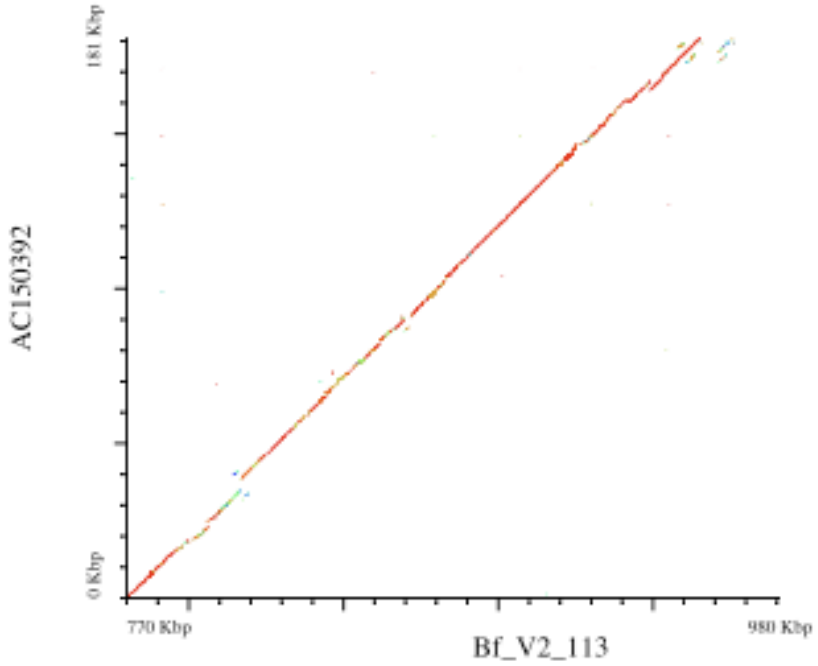
Supplementary Figure 60



Supplementary Figure 6l



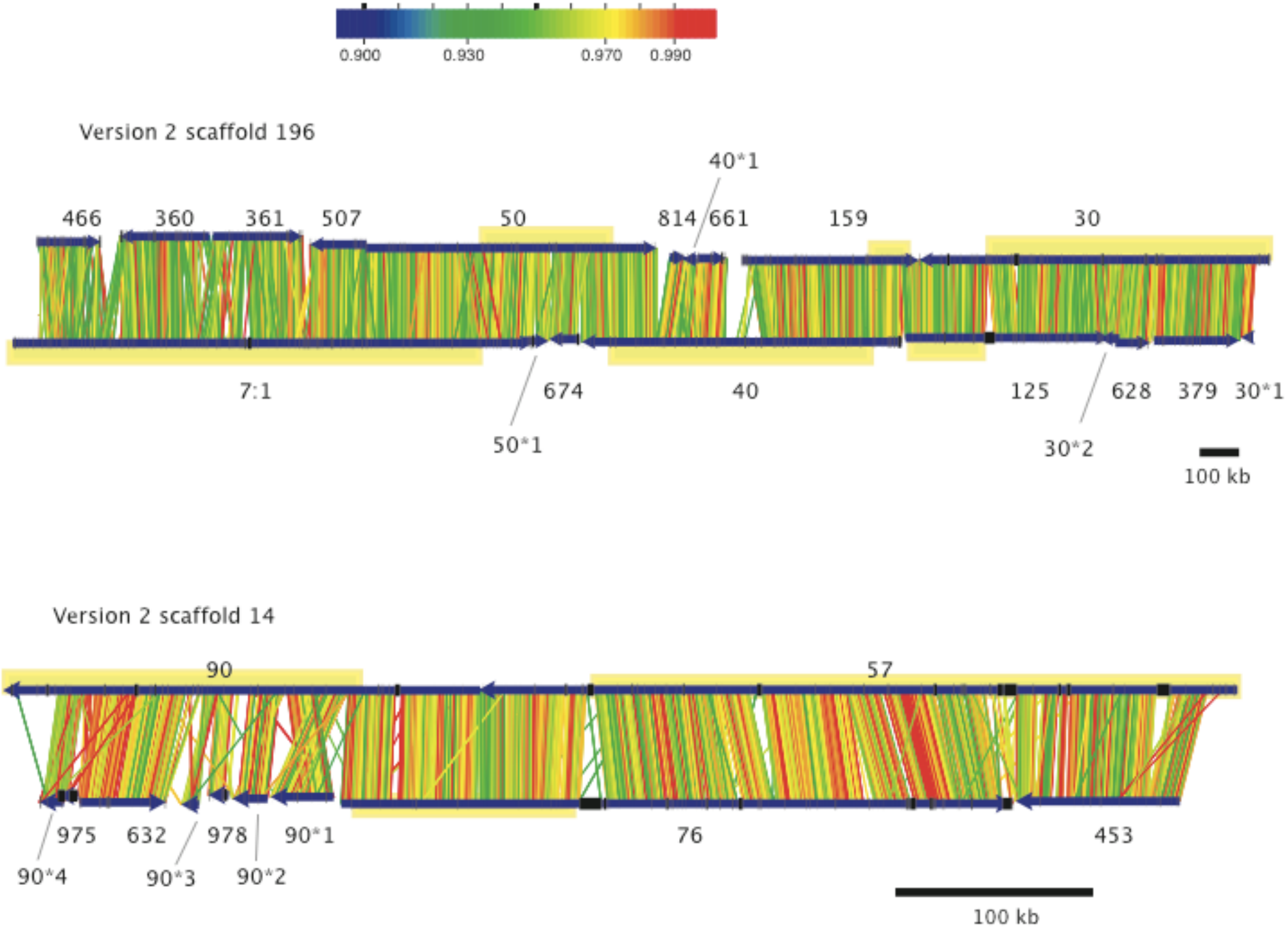
Supplementary Figure 62



Supplementary Figure 63: Construction of assembly version 2

Blue arrows represent scaffolds of assembly version 1, labeled by scaffold number, and showing hashes representing sequence gaps. Allelic "stutter" mis-assemblies (see Supplementary Note) have been excised, and are labeled with the source version 1 scaffold number followed by an asterisk, followed by the index number of the deletion (e.g. 30*2 denotes the second excised sequence from scaffold 30). A scaffold number followed by a colon indicates that a version 1 scaffold has been broken, and the number following the colon indicates the index number of the fragment shown. Colored lines represent individual MegaBLAST alignment segments, color coded by percent identity. The segments of the version 1 scaffolds which are tiled together to produce each version 2 scaffold are highlighted in yellow.

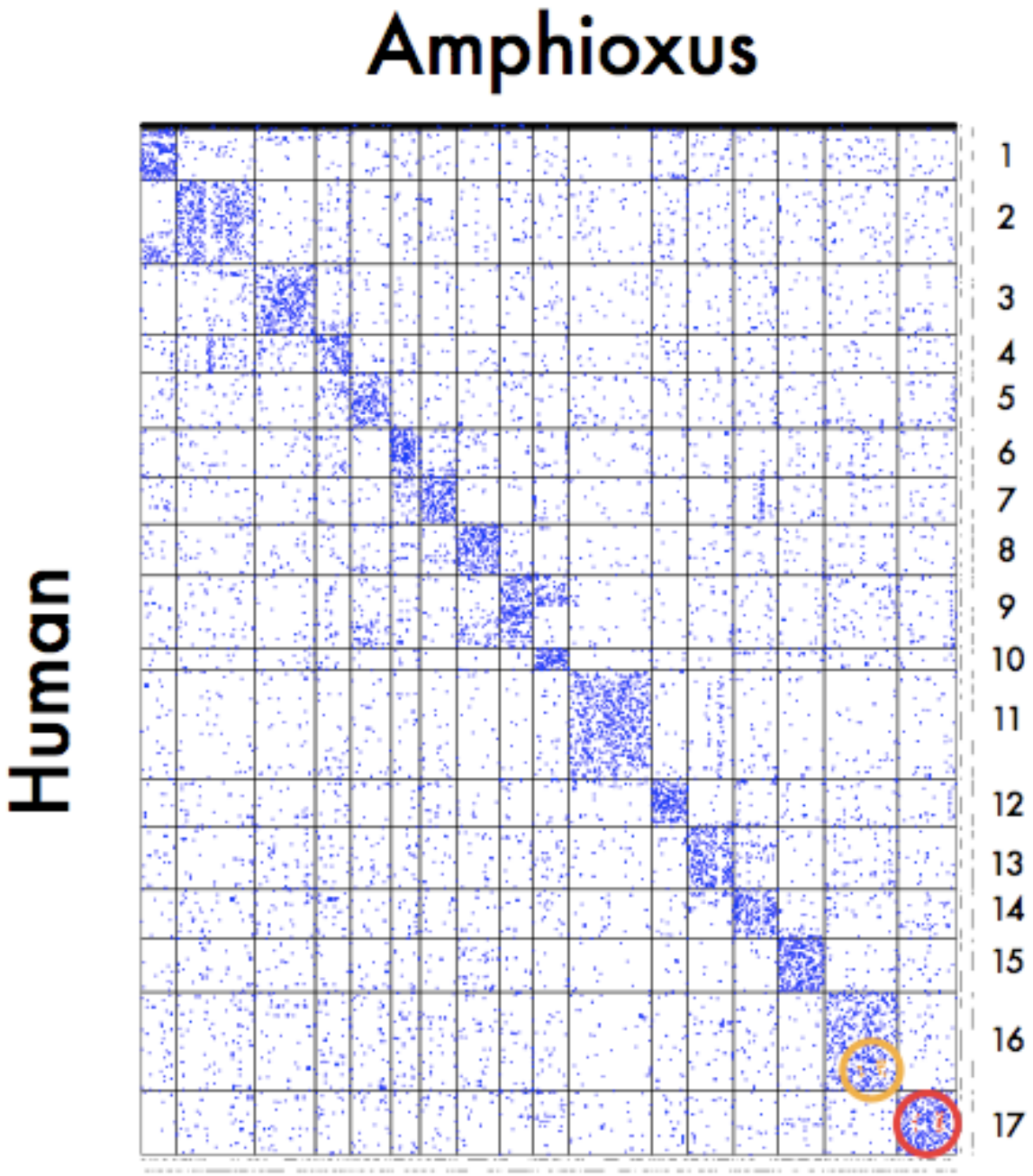
Supplementary Figure 63



Supplementary Figure 64:

Dots represent the relative position of orthologous gene pairs in human chromosomal segments (vertical position) and the 196 largest amphioxus version 2 scaffolds (horizontal position), in increasing index (i.e., gene number) along the chromosome/scaffold. Amphioxus scaffolds and human chromosomal segments have been ordered to show 17 clusters of human chromosomal segments and amphioxus scaffolds determined by hierarchical clustering. Alternating bars along the right and lower edges indicate the extent of individual segments and scaffolds respectively. The orthology relationships represented by red and yellow lines in (A) are shown as read and yellow dots, and their locations are hilighted with colored circles. More detailed views can be found in Supplementary Figure 29, which shows a detailed view of the 17 th CLG, and in Supplementary File 5, which contains an Oxford grid tabulating the number of orthologs for each scaffold-segment pair.

Supplementary Figure 64



References

1. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-10 (2002).
2. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-14 (2000).
3. Small, K. S., Brudno, M., Hill, M. M. & Sidow, A. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biol* **8**, R41 (2007).
4. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
5. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-7 (2005).
6. Yu, J. K. et al. Axial patterning in cephalochordates and the evolution of the organizer. *Nature* **445**, 613-7 (2007).
7. Smit, A. & Green, P. RepeatMasker at http://repeatmasker.genome.washington.edu. (2002).
8. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-22 (2000).
9. Birney, E. & Durbin, R. Using GeneWise in the Drosophila annotation experiment. *Genome Res* **10**, 547-8 (2000).
10. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
11. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-8 (2001).
12. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
13. Koonin, E. V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5**, R7 (2004).
14. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-80 (2004).
15. Tuskan, G. A. et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596-604 (2006).
16. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43 (2000).
17. Felsenstein, J. (Distributed by the author., 2004).
18. Brudno, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721-31 (2003).
19. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
20. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).
21. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5 (2001).
22. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-4 (2003).
23. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
24. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-5 (2005).
25. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-9 (2001).
26. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res* **11**, 803-16 (2001).
27. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94 (2007).
28. Swofford, D. L. (Sinauer Associates, Sinderland, Massachusetts, 2003).
29. Castro, L. F. & Holland, P. W. Fluorescent in situ hybridisation to amphioxus chromosomes. *Zoolog Sci* **19**, 1349-53 (2002).
30. Shoguchi, E. et al. Fluorescent in situ hybridization to ascidian chromosomes. *Zoolog Sci* **21**, 153-7 (2004).

31. Venkatesh, B. et al. Survey Sequencing and Comparative Analysis of the Elephant Shark (Callorhinchus milii) Genome. *PLoS Biol* **5**, e101 (2007).
32. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273-9 (2004).
33. Luke, G. N. et al. Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc Natl Acad Sci U S A* **100**, 5292-5 (2003).

Table 1. The most parsimonious partitioning of human chromosomal segments into paralogy groups derived from two rounds of whole genome duplication.

	A	B	C	D	E
1	5.1 9.1/3	19.2	19.1/3	1.5/7	
2	15.2	1.12	5.2 18 9.2	19.1/3	
3	8.3/5 6.1 18	16.3 19.4	14.1	20.4	
4	20.5	2.1 6.6/8	6.2 8.2/4	1.2	
5	6.7/9	X.2 21* (or C, or as E)	3.4	1.2	
6	4.3	10.3	5.3	8.1* (or merged with 4.3)	
7	8.1	2.3 20.2* (or in D, or as F)	10.4	5.5	4.1
8	X.6 5.4	4.4 4.2	11.7 13.1/3	2.4 X.1	
9	X.7 X.5	3.7	11.7	13.2 13.4	
10	3.7 2.7/9	7.5 17.2	1.1	12.2	
11	2.2 1.16 20.3	14.2 15.1	11.4	6.4	19.5
12	9.4	1.6/11/14	6.3		
13	10.1 7.6 22.3 12.4/6/8	X.4/8	3.1/3/6	1.8/15/10	
14	19.6	12.1/3/7/9 22.3 12.4/6/8	11.1/3		
15	19.6	16.1 7.1	22.2	17.4/7/10	
16	10.2	12.5	17.6/9	3.2 7.2/4/7	2.6
17	17.1/3/5/8 7.3	22.1 12.10	16.2	11.5	

Supplementary **TableS2. FISH data-sets that showed signals on the same chromosomes**

Scaffold	Rhodamine (Red)	Scaffold	FITC (Green)	order*
4	CH302-005_A20	4	CH302-014_A13	C_F_R
4	CH302-005_A20	5	CH302-024_F20	C_F_R
4	CH302-005_A20	17	CH302-020_G21	C_F_R
4	CH302-014_A13	203	CH302-022_L3	C_F_R
4	CH302-005_A20	223	CH302-019_E10	C_R_F
5	CH302-010_M13	5	CH302-024_F20	C_R_F
5	CH302-010_M13	22	CH302-005_C15	C_R_F
17	CH302-008_P21	5	CH302-024_F20	C_R_F
17	CH302-008_P21	17	CH302-020_G21	not clear
17	CH302-020_G21	17	CH302-008_P21	not clear
17	CH302-008_P21	112	CH302-011_E17	C_F_R
17	CH302-008_P21	240	CH302-066_L14	C_R_F
22	CH302-002_H19	4	CH302-014_A13	C_R_F
22	CH302-002_H19	5	CH302-024_F20	C_F_R
22	CH302-002_H19	17	CH302-020_G21	C_F_R
22	CH302-002_H19	22	CH302-005_C15	C_F_R
22	CH302-002_H19	240	CH302-066_L14	C_F_R
22	CH302-005_C15	534	CH302-018_D22	C_F_R

112	CH302-009_C10		4	CH302-014_A13	not clear
112	CH302-009_C10		5	CH302-024_F20	C_R_F
112	CH302-011_E17		112	CH302-009_C10	C_R_F
112	CH302-009_C10		233	CH302-019_H23	C_F_R
168	CH302-007_B20		22	CH302-005_C15	C_R_F
168	CH302-011_J23		168	CH302-007_B20	C_R_F
168	CH302-007_B20		168	CH302-011_J23	not clear
168	CH302-011_J23		240	CH302-066_L14	C_F_R
203	CH302-002_L4		5	CH302-010_M13	C_R_F
203	CH302-002_L4		112	CH302-009_C10	C_F_R
203	CH302-002_L4		168	CH302-007_B20	C_F_R**
203	CH302-002_L4		240	CH302-066_L14	C_F_R
203	CH302-002_L4		534	CH302-018_D22	C_F_R
218	CH302-007_B24		5	CH302-024_F20	C_R_F
218	CH302-007_B24		17	CH302-008_P21	C_F_R
218	CH302-007_B24		22	CH302-005_C15	C_R_F**
218	CH302-007_B24		203	CH302-022_L3	C_F_R
218	CH302-007_B24		240	CH302-066_L14	C_F_R
223	CH302-010_J9		223	CH302-019_E10	C_F_R
233	CH302-004_P11		233	CH302-019_H23	C_R_F
233	CH302-004_P11		289	CH302-038_O4	C_R_F

233	CH302-004_P11	289	CH302-038_O4	C_R_F
240	CH302-061_C13	5	CH302-024_F20	C_R_F
240	CH302-061_C13	112	CH302-011_E17	C_F_R
240	CH302-061_C13	240	CH302-066_L14	not clear
240	CH302-061_C13	534	CH302-018_D22	C_R_F
289	CH302-034_H5	4	CH302-005_A20	C_R_F
289	CH302-034_H5	17	CH302-008_P21	C_R_F
464	CH302-027_A13	17	CH302-020_G21	C_R_F
464	CH302-027_A13	233	CH302-019_H23	C_F_R
464	CH302-034_C21	289	CH302-038_O4	C_R_F
464	CH302-034_C21	534	CH302-018_D22	C_R_F
534	CH302-003_G9	5	CH302-024_F20	C_R_F
534	CH302-018_D22	168	CH302-011_J23	C_F_R
534	CH302-003_G9	223	CH302-019_E10	C_R_F
597	CH302-037_L12	112	CH302-009_C10	C_F_R
597	CH302-039_J12	168	CH302-007_B20	C_F_R

*For example, C_R_F indicates the order of centromere, rhodamine, and FITC.

** These FISH images are shown in Supplementary Figure S8.

Supplementary Table S3 Developmental gene linkages

Linkage	Scaffold (Version 2)	Gene family	Gene name	Protein IDs
Otx - Gsc	25	Gsc	*Gsc*	290661, 290663
		Otx	*Otx*	290652, 290662
Mnxa - Ro	205	Mnx	*Mnxa*	290473, 290474
		Ro	*Ro*	290436, 88105
NKx2-1 – NKx2-2	245	NK2-1	*NKx2-1*	116884, 291128
		NK2-2	*NKx2-2*	291064, 291065
NKx6 – NKx7	36	NK6	*NKx6*	290430, 291130
		NK7	*NKx7*	91037, 290480
En – Nedxa – Nedxb - Dll	9	En	*En*	290466, 291073
		Nedx	*Nedxa*	290469, 291199
			Nedxb	290470, 291206
		Dlx	*Dll*	290298, 290299
Hox cluster[a]	12	Hox	*Hox1* to *Hox15*	See Supplementary Note 2.2
ParaHox cluster[b]	150	Gsx	*Gsx*	69519
		Xlox	*Xlox*	69517
		Cdx	*Cdx*	69516
NK cluster remains[c]	36	Lbx	*Lbx*	290586, 291158
		Tlx	*Tlx*	290614
	39	NKx1	*NKx1a*	290589
			NKx1b	290608
		Ventx	*Vent1*	116892, 290585
			Vent2	289443, 291146
	26	NK4	*NKx4*	124258
		NK3	*NKx3*	290584
Wnt genes	149	Wnt10	*Wnt10*	118306

		Wnt3	*Wnt3*	118309
		Wnt6	*Wnt6*	57222
		Wnt1	*Wnt1*	113720
		Wnt9	*Wnt9*	67024

Ancient Developmental Gene Linkages: Otx and goosecoid are immediate neighbors (5 kb apart; scaffold 25), as are Mnx and ro (17 kb apart; scaffold 205), Nkx2-1 and Nkx2-2 (50 kb; scaffold 245), and Nkx7 with Nkx6 (scaffold 36, separated from the previously reported Lbx/Tlx gene pair by just two intervening genes). In addition, there is a cluster of four homeobox genes comprising En and Dll, with tandem duplicates of CG13424 (Nedxa and Nedxb) between them, in total spanning 130 kb (scaffold 9).

Supplementary Table S4 Whole genome shotgun sequencing summary

Insert type	JGI Library Identifiers	Mean insert size (bp)	Attempted Reads	Vector and quality trimmed reads	Sequence depth	Passing clones	Clone depth
3 kb plasmid	ASWX, AFPZ, ATUP	3,109	4,077,002	3,672,127	4.9X	1,696,599	10.5
6 kb plasmid	ATGN, ATWW, ASFW, AFSA, ATGI, ATWX	6,507	4,194,546	3,379,715	5.0X	1,641,896	21.4
37 kb (fosmid)	AWYB AWXX APWS APNK AFSB	35,303	640,513	562,798	0.7X	259,341	18.3
BAC-end		142,000		76,799	0.12 X	24,576	7.0
Total				**7,691,439**	**10.8X**		

Supplementary Table S5. Transposable elements in the amphioxus genome

Classes of TEs	Percent of the genome %
Total DNA transposons	**15**
"cut and paste":	
Mariner (*Tc1, Pogo* groups)	0.4
hAT	1.2
Kolobok	0.1
PiggyBac	1.0
Harbinger	3.0
P	0.1
MuDR	0.3
En/Spm	0.3
Chapaev	<0.01
Unclassified	6.5
"self-synthesizing" *Polintons*	0.1
"rolling circle" *Helitrons*	1.5
Total retrotransposons	**9**
LTR retrotransposons:	
Gypsy	0.3
BEL	0.1
Copia	0.1
ERV	0.1
DIRS	0.1
Non-LTR retrotransposons:	
CR1 (CR1 and L2 groups)	2.5
RTE (RTE and RTEX)	0.4
L1	0.3

Jockey	1.3
I (I, LOA)	0.1
NeSL	<0.01
SINE2	2.0
SINE3	0.1
Penelope	1.9
Unclassified TEs	**4**
Total TEs	**28**

Supplementary Table S6 Expressed sequence tag (EST) sequencing summary

Library identifier	sample	Total ESTs sequenced	short	insertless	contaminant	ESTs used in clustering
CAXC	*B. floridae* neurula	32,256	13%	1%	22%	66%
CAXF	*B. floridae* gastrula	32,256	17%	5%	23%	47%
CAXG	*B. floridae* larvae	33,024	7%	1%	26%	62%
CAAA	*P. marinus* mixed Embryonic stages, 2-12 days of development	38,400	9%	2%	6%	78%

"short" have vector sequence detected at the 3' end of a sequencing read, indicating that the cloned sequence is less than ~650 bp.
"insertless" have less than 150 bp of high quality, non-vector sequence.
"contaminant" are those ESTs which align to *E. coli*, or to mitochondrial or ribosomal sequence of the target organism.

Supplementary Table S7A Gene structure annotation summary

Data	Total
ESTs	480,070
EST clusters	77,402
putative ORFs from EST cluster consensus	4,272
known genes	96
Gene Catalog	50,818

Supplementary Table S7B Functional annotation summary

Functional Assignment	Gene models	Percentage	Distinct categories
Gene Ontology assignments (GO)	27,600	54%	3852
KEGG assignments	21,060	41%	866
KOG assignments	29,030	57%	4396

Supplementary Table S8 Gene sets used for comparative analyses

Species	Common name	Source
Takifugu rubripes	Japanese pufferfish	JGI v4.0
Gasterosteus aculeatus	Three spined stickleback	Broad S1 from ENSEMBL 41
Gallus gallus	chicken	Assembly WASHUC, Mar 2004 from Ensembl build 41 (gene build december 2005)
Homo sapiens	human	NCBI 36 from ENSEMBL 41
Ciona intestinalis	sea squirt	JGI finalized models Dec 2005
Oikopleura dioica	larvacean	NCBI Trace Archive; See Supplementary Note 5
Branchiostoma floridae	amphioxus	this study
Strongylocentrotus purpuratus	purple sea urchin	NCBI gene build 2 version 1 on Baylor's assembly Spur_v2.1
Drosophila melanogaster	fruit fly	BDGP 4 from ENSEMBL 41
Petromyzon marinus	sea lamprey	ESTs from this study
Saccoglossus kowalevskii	acorn worm	ESTs
Lottia gigantea	Limpet	JGI v1.0
Nematostella vectensis	Sea anemone	JGI v1.0

Supplementary Table S9: C-score clusters, based on hits with evalue <= 0.001

minimum C-value	# clusters	# clusters with <9 members	# human genes in clusters	# amphioxus genes in clusters (including allelic variants.)
1.0 (mutual best hits only)	8,111	8,107	8,220	9,901
0.9	8,173	8,040	11,602	14,612
0.8	8,401	8,081	14,079	19,625
0.7	8,009	7,609	15,813	24,812
0.6	7,268	6,875	16,855	29,829
0.0 (all BLAST hits)	1,825	1,685	18,581	42,992

Supplementary Table S10 Table of apparent chordate gene losses

Drosophila gene ID	Sea urchin protein ID	Drosophila gene name	Drosophila gene function (from flybase.org)
CG15654	XP_787203.2		open rectifier potassium channel activity.
CG5393	XP_001197867.1	apontic	RNA-binding transcription factor
CG11293	XP_001201086.1		-
CG15284	XP_001199322.1	partner of burs	neuropeptide hormone
CG16777	XP_001199210.1		-
CG7685	XP_001199885.1		alpha-glucosidase activity
CG31133	XP_785096.1		serine-tRNA ligase activity; ATP binding.

Supplementary Table S 11

1d	12b	3	1	8
2d	1c	3	19	28
2c	1a	4	5 9	Z Z_random
4b	11a	5	2	3
5d	4d	6	1	23
4c	3a	6	6 8	2
5d	5a	5 6	X	1
14b	13a	7 1	12 22	1
9b	10a	8 3	3	9
7c	6b	9 11	10	6
17c	15b	11	16	
1b	14a	11	19	
9d	7b	12	2	4
9c	11c	8	11	
8c	11c	8	11	
9c	8c	8	X 11	1
8b	6a	12	4	4
8b	7e	12	4	4
7e	6a	12	4	4
8a	6c	12	5	13
8a	7d	12	5	13
7d	6c	12	5	13
4b	11d	5	6	3
5a	11d	5	6	3
5a	4b	5	6	3
4c	12c	6	6	

5a	12c	6	6
3a	12c	6	6
3b	15a	6	19
15a	11e	6	19

Supplementary Table S 12

9d-7b	9b-10a	X	9
15c	14b-13a	22	1
9c-8c	5a-4b-11d	X	1
9d-7b	5a-4b-11d	X	1
9d-7b	9c-8c	X 13	1 4

Supplementary Table S13 Conserved non-coding elements in human and amphioxus

Human chromosome	from	to	Neighbouring Gene in Human	Percent Identity	Length	Amphioxus scaffold (Version1)	From	to	Com scor
chr13	94156903	94157110	SOX21	68.3	208	scaffold_50	1169971	1170173	1420
chr10	76835035	76835166	ZNF503	79.6	137	scaffold_9	3320557	3320693	1090
chr8	37652068	37652163	ZNF703	84.5	97	scaffold_9	3320560	3320656	8196
chr4	4976727	4976823	MSX1	71.4	98	scaffold_56	2362024	2362121	6997
chr14	60191711	60191798	SIX1	68.2	88	scaffold_52	2878536	2878622	6001
chr20	29655769	29655845	ID1	73.8	80	scaffold_166	1326698	1326773	5904
chr13	99413906	99413998	ZIC5	62.4	93	scaffold_40	1809962	1810050	5803
chr14	60188067	60188154	SIX1	64.8	88	scaffold_52	2880653	2880738	5702
chr10	124885142	124885219	BUB3	65.5	87	scaffold_406	567486	567572	5698
chr12	113586552	113586623	TBX3	70.8	72	scaffold_89	309180	309250	5097
chr9	107334635	107334713	KLF4	62.5	80	scaffold_114	550565	550642	5000
chr15	50877303	50877377	ONECUT1	59.5	79	scaffold_66	258311	258385	4700
chr1	90906604	90906670	BARHL2	64.2	67	scaffold_432	475978	476042	4301
chr10	114927547	114927612	TCF7L2	61.8	68	scaffold_9	3118949	3119004	4202
chr1	60180747	60180808	SIX1	60.9	69	scaffold_52	2889041	2889107	4202
chr	85757666	85757728	NKX61	63.6	66	scaffold_294	852642	852707	4197.
chr1	53253623	53253688	ONECUT2	60.3	68	scaffold_66	273303	273367	4100
chr1	113587672	113587734	TBX3	63.5	63	scaffold_89	310103	310164	4000
chr2	60710389	60710443	BCL11A	71.4	56	scaffold_136	1537170	1537225	3998
chr2	176839843	176839897	HOXD4	63.9	61	scaffold_402	180723	180783	3897
chr16	50234206	50234266	SALL1	58.5	65	scaffold_343	581922	581984	3802
chr3	144224977	144225040	PAQR9 / CHST2	59.4	64	scaffold_379	748738	748791	3801
chr2	54003009	54003062	PSME4	64.3	56	scaffold_473	147449	147504	3600
chr9	100571572	100571622	TMEFF1 / PRG-3	63	54	scaffold_218	562121	562170	3402
chr19	34403346	34403400	UQCRFS1	60.7	56	scaffold_469	71095	71149	3399
chr3	172261919	172261969	TNIK	62.3	53	scaffold_40	873391	873438	3301

chr2	1716994	1717040	Mouse PXDNL alignment	66	50	scaffold_15	2177531	2177575	3300
chr2	58584124	58584168	FANCL	68.1	47	scaffold_82	392305	392351	3200
chr13	59760566	59760615	TDRD3	61.5	52	scaffold_96	883301	883350	3198
chr1	10631300	10631349	SRG / PEX14	61.5	52	scaffold_367	514337	514387	3198
chr3	144224841	144224888	PAQR9 / CHST2	61.5	52	scaffold_30	396381	396432	3198
chr1	10631056	10631101	SRG / PEX14	62.7	51	scaffold_96	1876956	1877006	3197.
chr20	22496053	22496102	FOXA2	60.8	51	scaffold_42	883668	883714	3100
chr3	144218340	144218386	PAQR9 / CHST2	60.8	51	scaffold_30	390184	390234	3100
chr3	144226491	144226538	PAQR9 / CHST2	60.8	51	scaffold_379	747153	747199	3100
chr17	44007332	44007382	HOXB	60.8	51	scaffold_402	208920	208964	3100
chr2	57884100	57884150	EFEMP1 / VRK2	60.8	51	scaffold_218	577131	577174	3100
chr4	85755832	85755882	NKX61	60.8	51	scaffold_294	856715	856763	3100
chr4	160623342	160623388		62	50	scaffold_500	124798	124842	3100
chr10	124543866	124543915		60	50	scaffold_386	604793	604842	3000
chr6	150021599	150021647		60	50	scaffold_330	760848	760887	3000
chr1	10626884	10626931	SRG / PEX14	60	50	scaffold_96	1875233	1875282	3000
chr1	90906542	90906591	BARHL2	60	50	scaffold_432	476060	476108	3000
chr18	53247909	53247954	ONECUT2	62.5	48	scaffold_66	235430	235473	3000
chr12	94725217	94725260	NTN4 / SNRPF	61.2	49	scaffold_442	285358	285406	2998
chr15	50961156	50961204	ONECUT1	61.2	49	scaffold_5	3901544	3901591	2998
chr2	55086346	55086394	AK131467	61.2	49	scaffold_42	722465	722510	2998
chr9	107308843	107308890	KLF4	61.2	49	scaffold_114	524039	524082	2998
chr1	90904095	90904141	BARHL2	61.2	49	scaffold_432	479703	479750	2998
chr19	47122782	47122825		65.9	44	scaffold_42	169984	170027	2899
chr2	44190630	44190677	LRPPRC / PPM1B	60.4	48	scaffold_60	1517863	1517907	2899
chr4	174717890	174717937	SCRG1 / HAND2	60.4	48	scaffold_20	3100299	3100340	2899
chr18	43010923	43010967	FUSSELL18	63	46	scaffold_101	1556357	1556402	2898
chr12	56757363	56757402		66.7	42	scaffold_321	109272	109313	2801
chr17	44022952	44022997	HOXA7	60.9	46	scaffold_402	190240	190278	2801
chr14	36216910	36216944	PAX9 / SLC25A21	74.3	35	scaffold_42	928661	928694	2600

*** Based on precomputed MULTIZ alignments in the UCSC genome browser**

Supplementary Table S14

Human genome segment boundaries.

Segment ID	Cytological Coordinates	Chromosome	Start	End
1.1	1p36.12-36.33	1	1	23526132
1.2	1p34.2-36.12	1	23526132	42588011
1.3	1p34.2	1	42588011	42689030
1.4	1p34.2	1	42689030	43180612
1.5	1p22.3-34.2	1	43180612	85630086
1.6	1p22.2-22.3	1	85630086	91246779
1.7	1p21.1-22.2	1	91246779	102063750
1.8	1p13.3-21.1	1	102063750	110948178
1.9	1p13.2-13.3	1	110948178	112039666
1.1	1p13.2	1	112039666	114548603
1.11	1p12-13.2	1	114548603	120491741
1.12	1p12-q23.1	1	120491741	155279431
1.13	1q23.1-23.3	1	155279431	159410831
1.14	1q23.3-32.1	1	159410831	199491263
1.15	1q32.1-32.3	1	199491263	210027080
1.16	1q32.3-44	1	210027080	247249719
2.1	2p23.2-25.3	2	1	28049984
2.2	2p14-23.2	2	28049984	69324397
2.3	2p14-q11.2	2	69324397	97999791
2.4	2q11.2	2	97999791	101351072
2.5	2q11.2-13	2	101351072	113047483

2.6	2q13-36.1	2	113047483	222822858
2.7	2q36.1-37.1	2	222822858	232953730
2.8	2q37.1	2	232953730	233181891
2.9	2q37.1-37.3	2	233181891	242951149
3.1	3p24.3-26.3	3	1	15435942
3.2	3p21.31-24.3	3	15435942	48087963
3.3	3p12.3-21.31	3	48087963	75560151
3.4	3p12.3-q21.1	3	75560151	123747788
3.5	3q21.1	3	123747788	124038889
3.6	3q21.1-22.1	3	124038889	135038112
3.7	3q22.1-29	3	135038112	199501827
4.1	4p15.2-16.3	4	1	25986602
4.2	4p15.2-q12	4	25986602	57101698
4.3	4q12-27	4	57101698	122974672
4.4	4q27-35.1	4	122974672	184807545
4.5	4q35.1-35.2	4	184807545	191273063
5.1	5p15.33-q13.2	5	1	68487890
5.2	5q13.2-31.1	5	68487890	131346486
5.3	5q31.1-31.3	5	131346486	139835480
5.4	5q31.3-35.1	5	139835480	167951722
5.5	5q35.1-35.3	5	167951722	180857866
6.1	6p24.3-25.3	6	1	7925759
6.2	6p22.2-24.3	6	7925759	24949289
6.3	6p21.2-22.2	6	24949289	37743189
6.4	6p21.1-21.2	6	37743189	44494690
6.5	6p12.3-21.1	6	44494690	50919969
6.6	6p12.3-q14.1	6	50919969	76010435
6.7	6q14.1-22.31	6	76010435	120493901
6.8	6q22.31-23.3	6	120493901	136923570

6.9	6q23.3-27	6	136923570	170899992
7.1	7p21.3-22.3	7	1	7245460
7.2	7p11.2-21.3	7	7245460	55291577
7.3	7p11.2-q11.23	7	55291577	77205318
7.4	7q11.23-21.3	7	77205318	97651696
7.5	7q21.3-22.1	7	97651696	102443158
7.6	7q22.1-36.1	7	102443158	148578580
7.7	7q36.1-36.3	7	148578580	158821424
8.1	8p23.3-q11.23	8	1	52934748
8.2	8q11.23-21.11	8	52934748	75203543
8.3	8q21.11-22.2	8	75203543	99158104
8.4	8q22.2-24.3	8	99158104	145137892
8.5	8q24.3	8	145137892	146274826
9.1	9p22.3-24.3	9	1	15455933
9.2	9p22.3-q22.31	9	15455933	93154261
9.3	9q22.31-32	9	93154261	114841269
9.4	9q32-34.3	9	114841269	140273252
10.1	10p13-15.3	10	1	15173895
10.2	10p13-q11.21	10	15173895	42977351
10.3	10q11.21-23.31	10	42977351	91148004
10.4	10q23.31-26.3	10	91148004	135374737
11.1	11p15.4-15.5	11	1	3646292
11.2	11p15.4	11	3646292	8513570
11.3	11p15.4-q12.3	11	8513570	61969973
11.4	11q12.3-13.1	11	61969973	66588478
11.5	11q13.1-13.2	11	66588478	67708722
11.6	11q13.2-13.4	11	67708722	70921654
11.7	11q13.4-23.1	11	70921654	111081603
11.8	11q23.1-25	11	111081603	134452384

12.1	12p13.31-13.33	12	1	5733870
12.2	12p12.3-13.31	12	5733870	16622256
12.3	12p11.21-12.3	12	16622256	32686135
12.4	12p11.21-q13.11	12	32686135	46754658
12.5	12q13.11-14.1	12	46754658	56444814
12.6	12q14.1-15	12	56444814	67572059
12.7	12q15-21.33	12	67572059	90031496
12.8	12q21.33-23.1	12	90031496	99176013
12.9	12q23.1-23.3	12	99176013	107108041
12.1	12q23.3-24.33	12	107108041	132349534
13.1	13p13-q14.11	13	1	41837067
13.2	13q14.11-14.3	13	41837067	51381672
13.3	13q14.3-31.3	13	51381672	92425235
13.4	13q31.3-34	13	92425235	114142980
14.1	14p13-q12	14	1	29709897
14.2	14q12-32.33	14	29709897	106368585
15.1	15p13-q15.3	15	1	41538402
15.2	15q15.3-26.3	15	41538402	100338915
16.1	16p11.2-13.3	16	1	28448217
16.2	16p11.2-q11.2	16	28448217	45205208
16.3	16q11.2-24.3	16	45205208	88827254
17.1	17p13.1-13.3	17	1	6848164
17.2	17p13.1	17	6848164	8263123
17.3	17p13.1	17	8263123	8349403
17.4	17p11.2-13.1	17	8349403	18829718
17.5	17p11.2-q12	17	18829718	33764827
17.6	17q12-21.33	17	33764827	45884125
17.7	17q21.33-22	17	45884125	53088870
17.8	17q22-23.2	17	53088870	57957458

17.9	17q23.2-24.1	17	57957458	59989546
17.1	17q24.1-25.3	17	59989546	78774742
19.1	19p13.2-13.3	19	1	9866833
19.2	19p13.12-13.2	19	9866833	15445888
19.3	19p13.11-13.12	19	15445888	19486034
19.4	19p13.11-q13.12	19	19486034	40572552
19.5	19q13.12-13.32	19	40572552	53188588
19.6	19q13.32-13.43	19	53188588	63811651
20.1	20p13	20	1	2550878
20.2	20p12.3-13	20	2550878	5926099
20.3	20p12.3-q11.21	20	5926099	29625029
20.4	20q11.21-13.33	20	29625029	60346742
20.5	20q13.33	20	60346742	62435964
22.1	22p13-q12.3	22	1	31349522
22.2	22q12.3-13.2	22	31349522	41827288
22.3	22q13.2-13.33	22	41827288	49691432
X.1	Xp22.2-22.33	X	1	15752846
X.2	Xp11.3-22.2	X	15752846	45951242
X.3	Xp11.3	X	45951242	46928149
X.4	Xp11.21-11.3	X	46928149	56042521
X.5	Xp11.21-q13.1	X	56042521	70406305
X.6	Xq13.1-22.3	X	70406305	106924338
X.7	Xq22.3-28	X	106924338	151401568
X.8	Xq28	X	151401568	154913754

Supplementary Table S15

Stickleback segment boundaries.

Segment	Chromosome	Start	End
groupIX.1	groupIX	1	12739387
groupIX.2	groupIX	12739387	18260863
groupIX.3	groupIX	18260863	20249479
groupI.1	groupI	1	19964834
groupI.2	groupI	19964834	28185914
groupIV.1	groupIV	1	17714141
groupIV.2	groupIV	17714141	32632948

Supplementary Table S16

Chicken segment boundaries.

Segment	Chromosome	Start	End
1.1	1	1	32179740
1.2	1	32179740	40853667
1.3	1	40853667	43982312
1.4	1	43982312	44818075
1.5	1	44818075	49701333
1.6	1	49701333	69028920
1.7	1	69028920	115004230
1.8	1	115004230	126498076
1.9	1	126498076	145801726
1.1	1	145801726	158501106
1.11	1	158501106	162452226
1.12	1	162452226	170671740
1.13	1	170671740	184388247
1.14	1	184388247	end
2.1	2	1	39491254
2.2	2	39491254	39527937
2.3	2	39527937	42629862
2.4	2	42629862	55946861
2.5	2	55946861	67057040
2.6	2	67057040	89645598
2.7	2	89645598	89645598
2.8	2	89645598	97401111

2.9	2	97401111	end
3.1	3	1	38232351
3.2	3	38232351	79429206
3.3	3	79429206	end
4.1	4	1	17087066
4.2	4	17087066	34590573
4.3	4	34590573	61870015
4.4	4	61870015	73600663
4.5	4	73600663	end
5.1	5	1	21539089
5.2	5	21539089	end
6.1	6	1	18748287
6.2	6	18748287	end
8.1	8	1	12093392
8.2	8	12093392	12093392
8.3	8	12093392	14122324
8.4	8	14122324	19913306
8.5	8	19913306	end
13.1	13	1	7773480
13.2	13	7773480	10458622
13.3	13	10458622	end
14.1	14	1	14266398
14.2	14	14266398	end
Z.1	Z	1	7368121
Z.2	Z	7368121	25068728
Z.3	Z	25068728	end

Supplementary Table S17

Reference sequence accession numbers

AC214474	AC150426
AC150388	AC150428
AC150391	AC150430
AC150392	AC150385
AC150432	AC150429
AC150387	AC150407
AC150408	AC150399
AC150414	AC150431
AC150416	AC150418
AC150421	AC150419
AC150425	AC150412
AC150398	AC150413
AC150401	AC150386
AC150410	AC150389
AC150424	AC150393
AC150415	AC150390
AB231866	

Supplementary Table S18

Enriched functional categories. The ten most significantly enriched PANTHER functional categories [87] in chordate gene families with retained 2R duplicates (Methods). Significance of the enrichment is measured relative to the null hypothesis that retention in multiple copies occurs at random, independent of gene function. For example, 9% of chordate gene families have an annotation that implicates them in "signal transduction", the most significantly enriched ontology term; of these, 38% were retained in multiple copies after 2R by the more stringent criteria, requiring paralogous pairs in conserved syntenic position (more than twice the overall retention rate of 18%). All are significant at the level of p < 1e-15.

Number chordate gene families	Number of chordate families with retained duplicates (synteny-confirmed based on 2R reconstruction)	% chordate families with retained duplicates (with synteny)	PANTHER Functional annotation category
799 (9%)	438 (306)	55 (38)	Signal transduction
602 (7%)	315 (236)	52 (39)	Developmental processes
299 (4%)	174 (135)	58 (45)	Cell surface receptor mediated signal transduction
474 (6%)	236 (176)	50 (37)	Transcription factor
296 (4%)	163 (107)	55 (36)	Intracellular signaling cascade
375 (4%)	189 (145)	50 (39)	mRNA transcription regulation
518 (6%)	239 (178)	46 (34)	mRNA transcription

241 (3%)	132 (93)	55 (39)	Cell communication
149 (2%)	94 (75)	63 (50)	Mesoderm development
165 (2%)	100 (68)	61 (41)	Neuronal activities