Title
The Genome of Amphioxus and the Ancestral Proto-chordate

Authors
Putnam, Nicholas
Kawashima, Takeshi
Terry, Astrid
et al.

Publication Date
2006-05-03

The Genome of Amphioxus and the Ancestral Proto-chordate

Nicholas Putnam [1], Takeshi Kawashima [2], Astrid Terry [1], Sky You [3], Yutaka Satou [2], Erika Lindquist [1], Igor Grigoriev[1], Jeremy Gibson-Brown[4], Marianne Bronner-Frasier[3], Peter Holland[5], Asao Fujiyama [6], Nori Satoh [2], Linda Holland [7], Daniel Rokhsar [1,8]

[1] DOE Joint Genome Institute, Walnut Creek CA, [2] Kyoto University, Japan, [3] California Institute of Technology, Pasadena CA, [4] Washington University, St. Louis MO, [5] University of Oxford, Oxford UK, [6] National Institute of Informatics, Japan, [7] UC San Diego, La Jolla CA, [8] University of California, Berkeley CA

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

U.S. D.O.E. JOINT GENOME INSTITUTE

Abstract

Lancelets ("amphioxus") are the modern survivors of an ancient chordate lineage, the cephalochordates, with a fossil record dating back to the Cambrian. We describe the structure and gene content of the highly polymorphic genome of the Florida lancelet, in the context of the known genomes from the other chordate lineages, including the tunicate *Ciona*, human, and other vertebrates. This whole-genome comparison illuminates the murky relationship between the three principal chordate subphyla (tunicates, cephalochordates, and vertebrates), and provides a new perspective on the nature of the chordate common ancestor and key events in the emergence of vertebrates. The amphioxus sequence allows reconstruction of not only the gene complement of the ancestral chordate, but also a partial reconstruction of its genomic organization and a description of subsequent duplications and reorganizations in the vertebrate lineage.

Generation of the *B. floridae* draft genome sequence and gene set.

The genome of B. floridae was sequenced by whole genome shotgun sequencing to a depth of coverage of 10X from shotgun libraries with mean insert sizes of 3kb, 8kb and 40kb in a total of 7.6 million shotgun reads. In addition, 77 thousand BAC clone end sequences have been determined. All libraries were created from genomic DNA of a single *B. floridae* individual, collected in Florida.

The shotgun reads were assembled with JAZZ, the JGI WGS pipeline. The result of initial assemblies showed that the two haplotypes in the genome are highly polymorphic: 5-10% of all base pair positions differ between haplotypes. As a consequence, resulting in the two haplotypes being assembled separately. Each locus in the haploid genome is represented doubly. The depth of coverage of each assembled haploid genome is approximately 5X, and the typical size of assembled scaffolds was 1.25 Mb. This allowed detection of potential global mis-joins by comparing the two haplotypes. We identified N sites in the assembly as potential misassembled based on a large scale lack of colinearity between the assembled haplotypes. The scaffolds were broken at these positions and re-assembled with the BAC and fosmid end sequence linking information. The resulting initial draft genome sequence is composed of 3,032 scaffolds with a mean (length-weighted) size of 1.58 Mbp, and mean (length-weighted) contig size of 16 Kbp.

Both homology-based and *ab initio* gene prediction was carried out with the JGI annotation pipeline. The resulting gene set contains 50K genes, with most genes represented double, once for each haplotype.

Chordate phylogeny: Cephalochordates branched early

In order to resolve the phylogenetic relationship of the cephalochordates to other groups, we constructed a concatenated multiple alignment of single-copy, slowly evolving genes from sequenced genomes. Using the draft genome of the sea anemone Nematostella vectensis, and representatives of the main deuterostome lineages for which whole genome datasets are available: Echinoderms (Sp), Cephalochordates (bf), Tunicaes (Ciona intestinalis), and vertebrates (the frog xenopus tropicalis and human). We chose as our sequences for this analysis a set of single copy genes present in all six included taxa, and which showed a slow rate of amino acid substitution. The rate of substitution was estimated by making clustalw multiple alignments of each genes, identifying the high-quality portion of the multiple alignment using Gblocks, and calculating the maximum-liklihood tree topology and branch lengths under a model of protein sequence evolution (JTT, Gamma+I, alpha=1). 282 genes with no branch longer that 0.3 expected changes per site.

This set of genes and multiple alignments were then used to reconstruct the relationships among the six taxa using, maximum parsimony, maximum likelihood and bayesian MCMC parameter estimation. All three methods support the single topology show in figure 1, with MCMC and parsimony bootstrap support values show above and below the branches, respectively. The branch lengths shown are from the MCMC estimates. Contrary to the classical consensus, cephalochordates are the living desendents of a group of chordates which diverged before the split of vertebrates from tunicates.

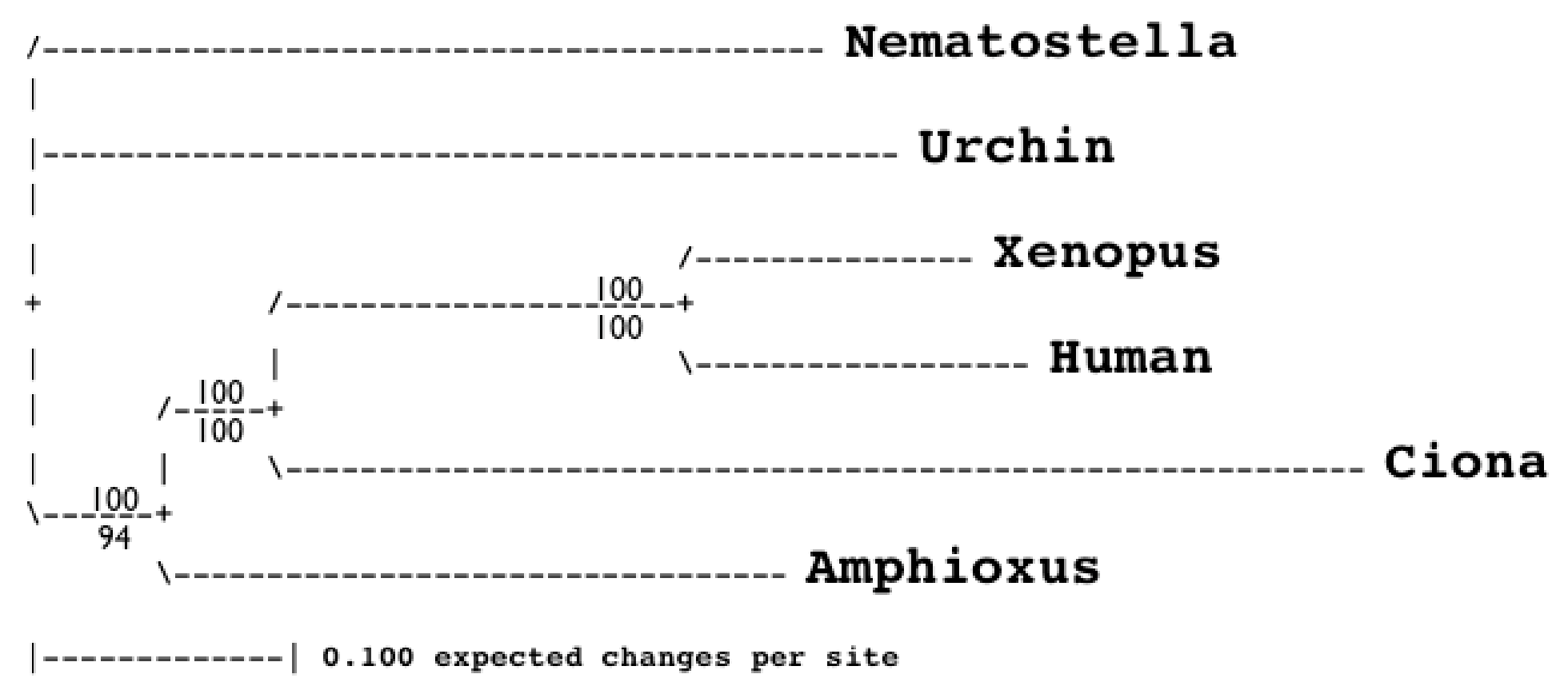


Fig 1.

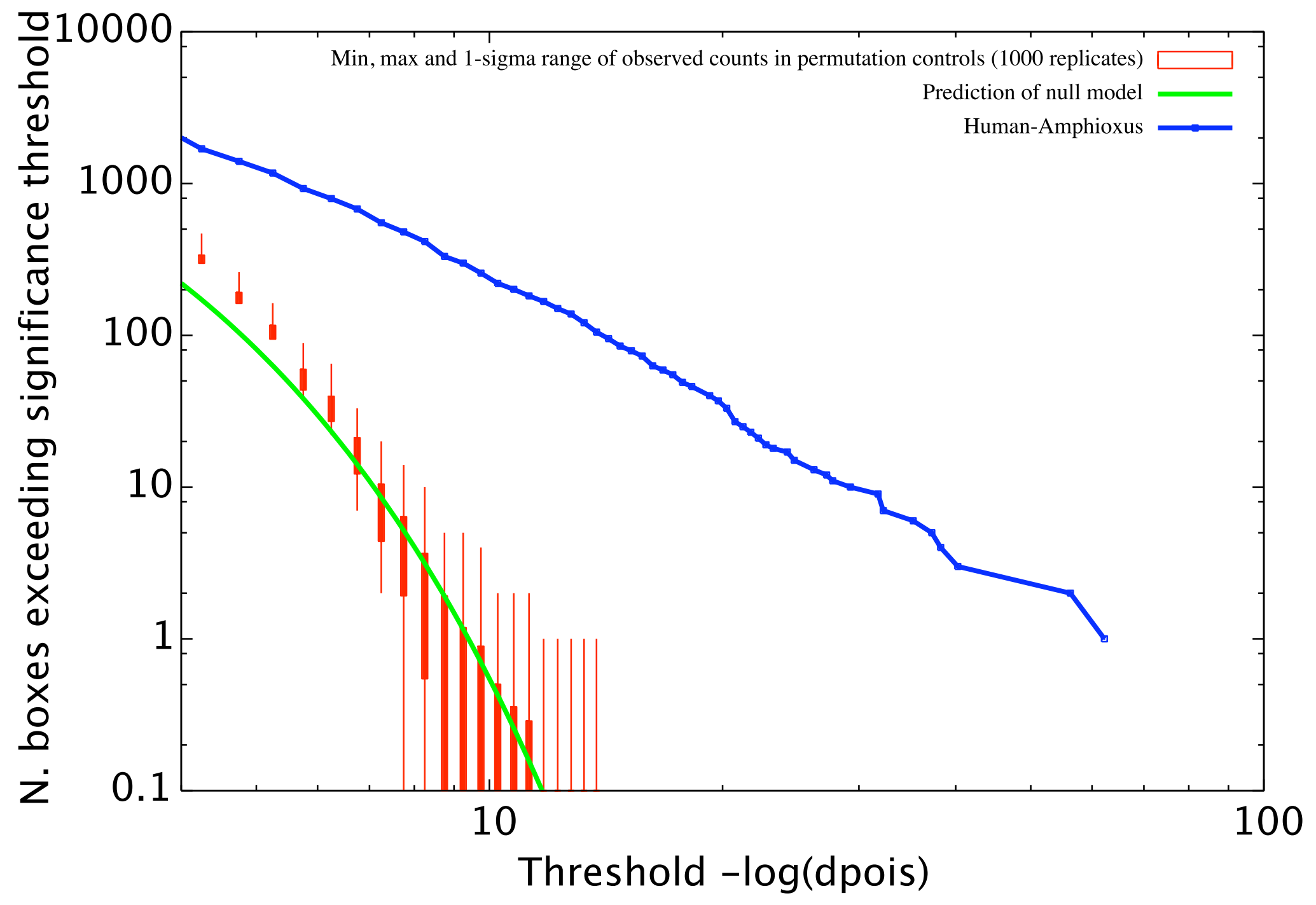
Gene linkages conserved with human reveals genome organization of the ancestral chordate.

Can we detect, by comparing the scaffolds of the draft amphioxus assembly with vertebrate genomes, any trace of the genome organization of the pre-cambrian common ancestor of the chordates? The surprising result is yes: there is a strong signal of conserved gene linkage that is well-enough preserved between humans and amphioxus that we can derive significant information about how the genes were grouped into chromosomes in the common ancestor.

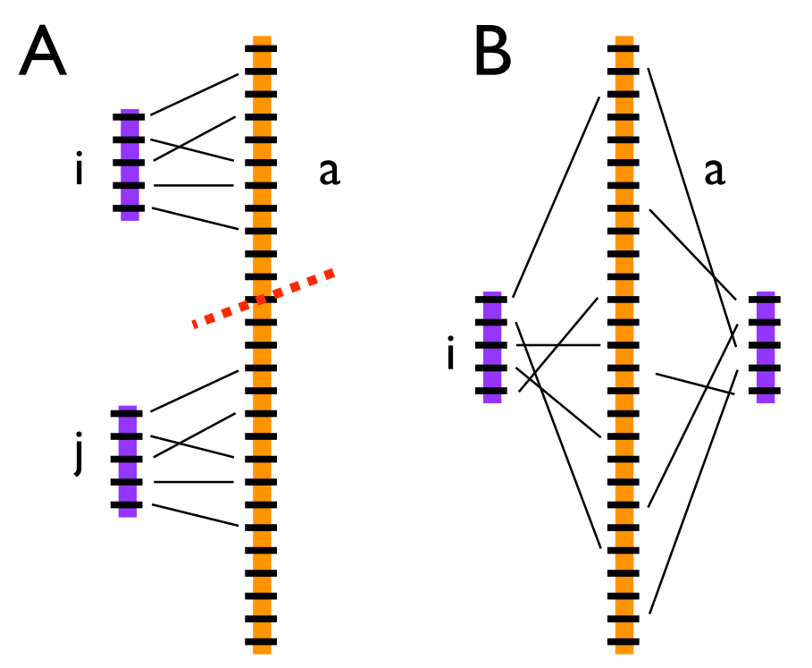
We compared the observed positional distribution of putative orthologs between the amphioxus and human genomes to the prediction of the null hypothesis that gene position has been completely randomized over evolution. The predicted peptide sequences of B. floridae were aligned to the NCBI Human genome 35 proteins using BLASTP.

All sequences hitting more than 10 genes in the opposite genome were eliminated from consideration. To normalize for the variable size of scaffolds, the number of observed independent hits between each scaffold and bins of 75 participating genes along the human genome was compared to the prediction of simple Poisson statistics. (To eliminate the effect of potentially independently-expanded tandem families, hits are considered non-independent if they involve the same gene.) The graph below shows, as a function of threshold p-value, the number of significant hits in blue, the number of hits predicted by the model, and the observed number of hits in randomization trials.

.



At a very stringent p-value threshold of 10^{-6}, (-log(p) > 13.8) there are 98 hits involving 86 amphioxus scaffolds, with a total length of 178 Mbp. In 1000 permutation trials, in which a total of more than 10^{8} scaffold-to-chromosome-bin hits were considered, there were no observations of a hit exceeding this threshold.

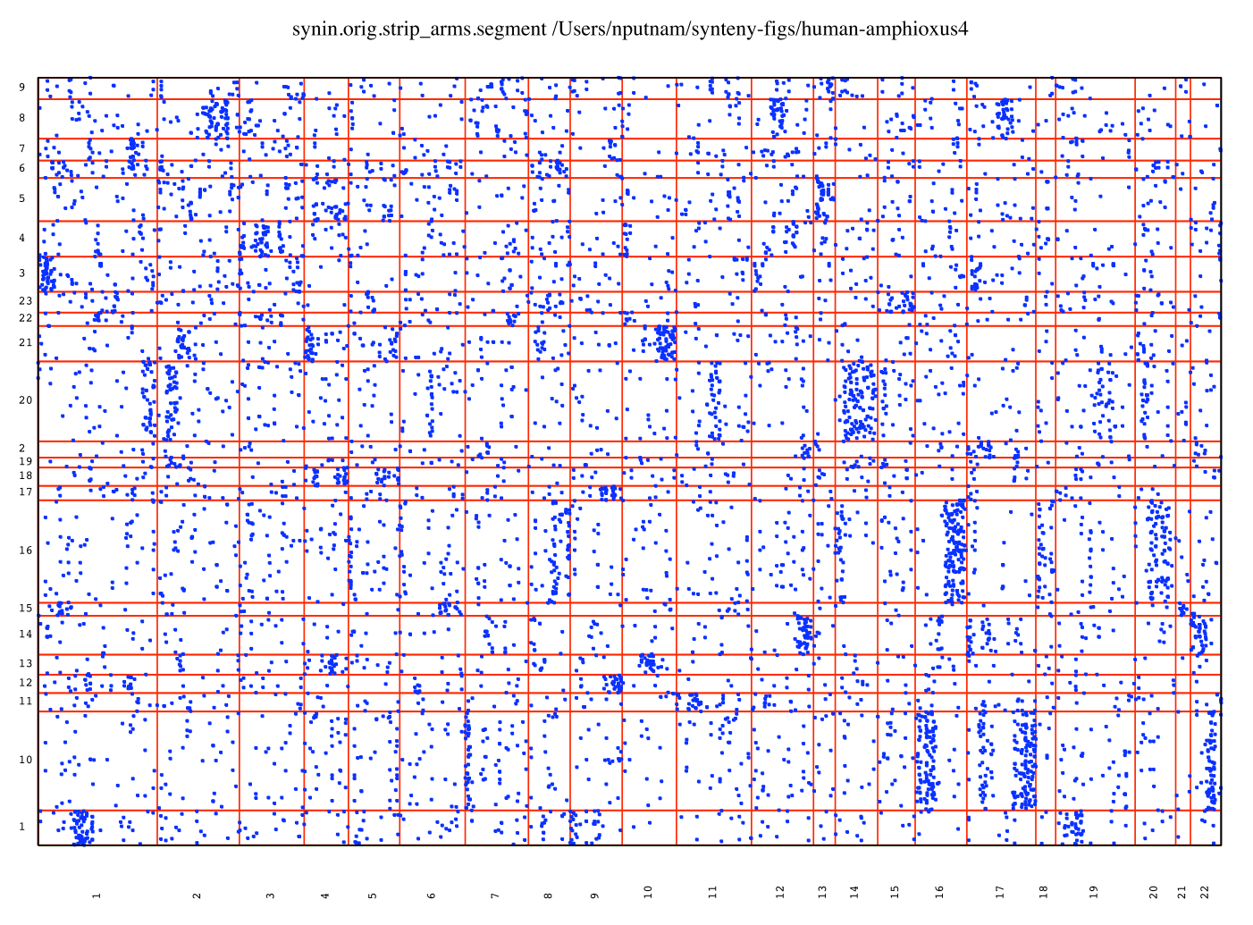


Local rearrangements reveal ancestral linkage relationships.

When genomic framgents are compared between species with highly-conserved gene order, (A) genomic segments *i* and *j* of sp 1 which have significant hits to the same genomic segment *a* species 2 may not be linked ancestrally: a recent fusion (red dashed line) may separate their hits on *a*. When comparing anciently diverged genomes, (B) segments *i* and *j* which have homologous genes distributed uniformly across the same segment *a* can be inferred to have shared linkage on the same chromosome for time long on the scale of .

Clustering identifies anciently-linked groups of scaffolds.

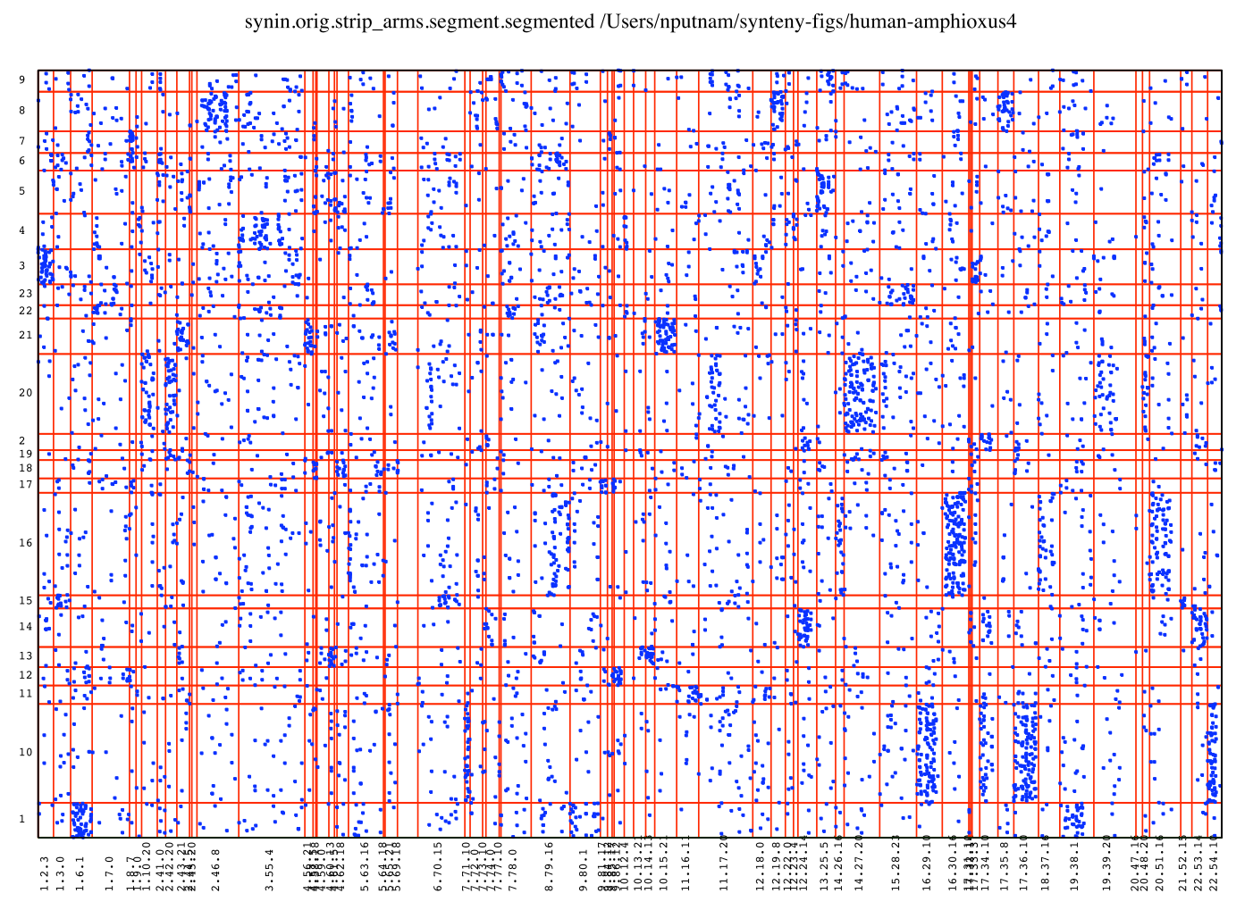
Scaffolds were clustered hierarchically based on the similarity (results shown here use Pearson correlation coefficient) of their significant hits to binned human chromosomes, with the pairwise centroid-linkage method. (See previous section for significance criteria.) The result is not highly sensitive to the distance metric or clustering algorithms used.



-Blue dots indicate BLASTP alignments with a score >=50 between human and amphioxus peptides.
-Horizontal position indicates the rank order position of the human gene in the human genome.
-Vertical position indicates rank order position of the amphioxus gene in the amphioxus scaffolds, which have been ordered according to result of the hierarchical clustering.
-Vertical red lines indicate boundaries between human chromosomes.
-Horizontal red lines indicate boundaries between clusters of scaffolds at a threshold correlation of 0.5.

Recent chromosomal breakpoints identified with an HMM:

Recent chromosomal fusion events on the lineage leading to humans results in the discontinuities in the pattern of dots in the figure above. The hidden states of the model are combinations of the scaffold clusters defined previously. An empirically-optimized set of transition probabilities and noise frequency allows automatic identification of these discontinuities in one genome, given a clustering of the sequences in the other. The breaks identified in human chromosomes based on the above clustering of amphioxus scaffolds are represented in the following figure by added vertical red lines:



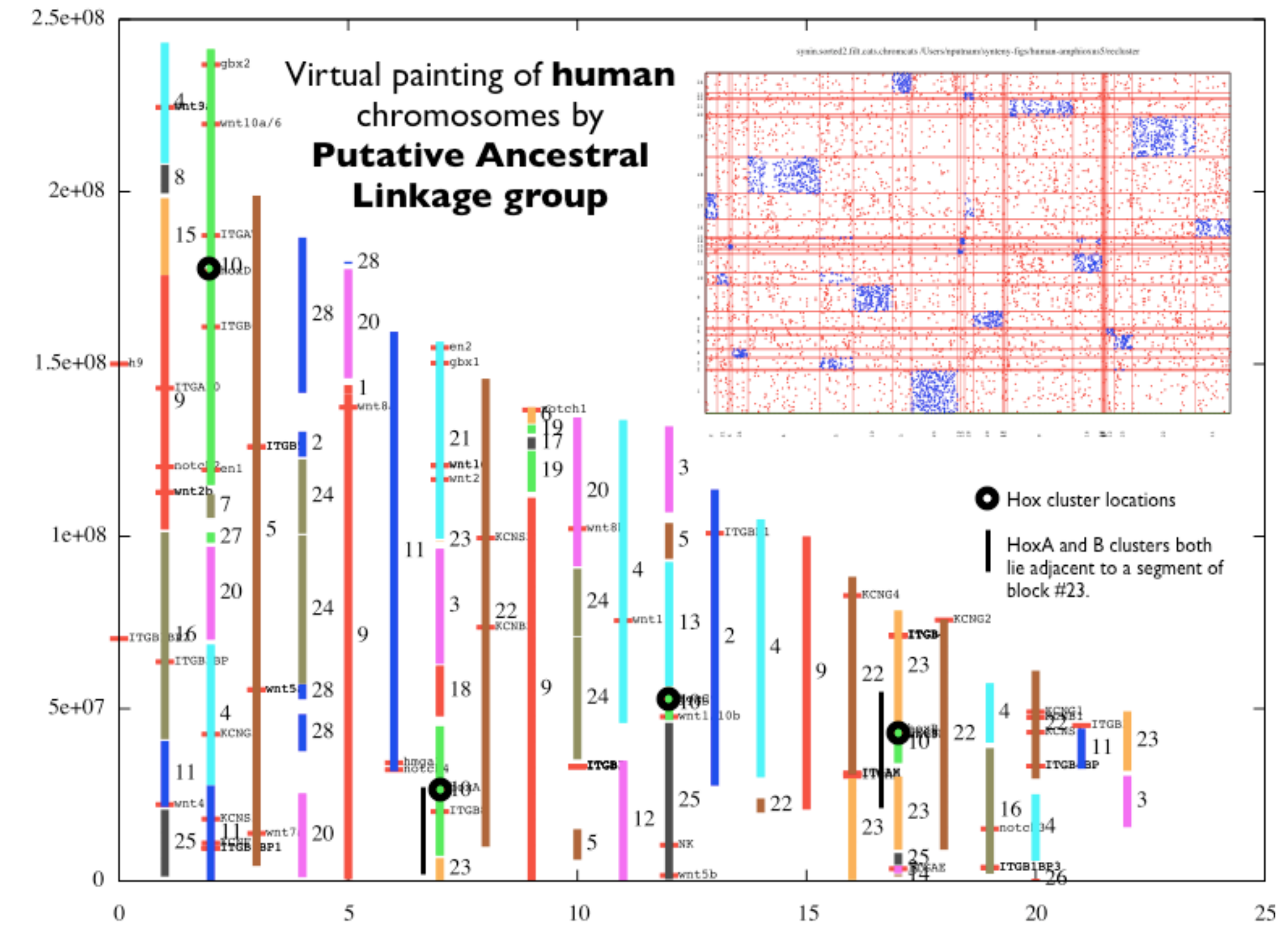
Re-clustering scaffolds increases sensitivity.

Improved discrimination of significant conserved synteny relationships is achieved by re-clustering the scaffolds based on the similarity of their significant hits to the chromosome segments identified by the HMM because fewer chromosome breakpoint boundaries fall within a bin and decrease the signal-to-noise.

Clustering human chromosome segments identifies paralogous regions.

Improved discrimination of significant conserved synteny relationships is achieved by re-clustering the scaffolds based on the similarity of their significant hits to the chromosome segments identified by the HMM, since fewer recent chromosome breakpoint boundaries fall within a bin, thereby decreasing signal-to-noise. This clustering groups together segments derived from the same putative ancestral linkage group (PAL).

Human autosome segments assigned to Putative Ancestral Linkage (PAL) groups:



-The human autosomes, segmented as described above, and labeled by PAL. Segments are color coded, but the colors are not unique.
-The inset shows the locations of protein-protein alignment hits between clustered amphioxus scaffolds (vertical axis) and clustered human chromosome segments (horizontal axis). Hits falling in bins with a significant excess of hits relative the null model are shown in blue, others in red. Horizontal and vertical red lines indicate cluster boundaries.
-The locations of the four hox gene clusters are indicated with bold circles. The positions of selected other previously-studied groups of paralogous genes are also indicated.

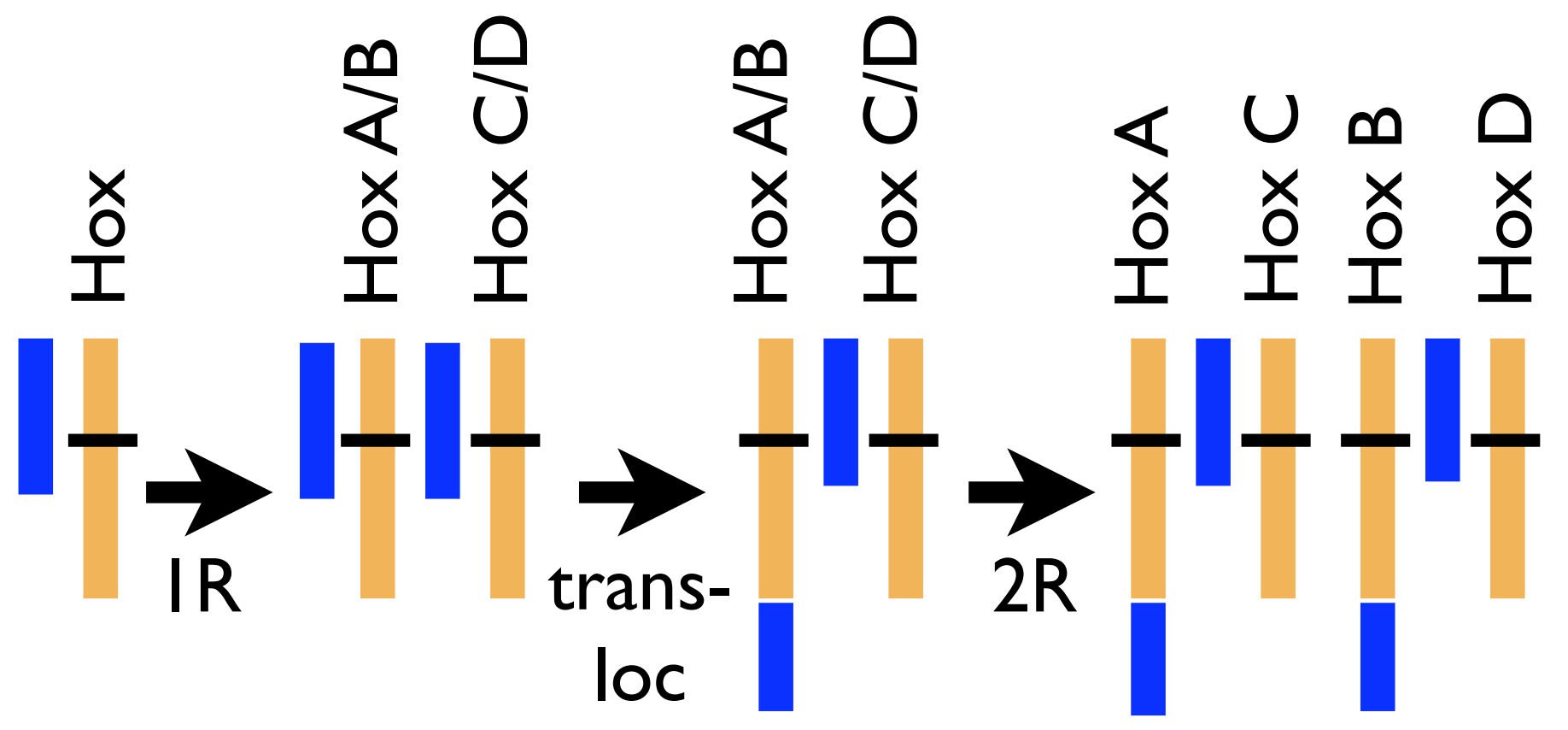
Identification of chromosomal translocation events on the vertebrate stem, confirmed in the Zebrafish genome.

It was noted that the hoxA and hoxB clusters, on human chromosomes 7 and 17 respectively, lie within PAL 10 and adjacent to PAL 23. This similarity could be due to independent fusion events, or could have been created by the same duplication (segmental, or of the whole genome) which created these two clusters from an ancestral hoxA/B cluster. The latter hypothesis has at least two significant predictions:

1. The association of the hoxA and hoxB clusters with PAL 23 should be present in all vertebrates, except where subsequent genome re-arrangements have separated them again.
2. The PAL 23 segments on chr7 and chr17 should have a number of retained paralogous gene pairs typical for paralogous segments, as opposed to segments derived from the same ancestral segment by fragmentation rather than duplication.

Both of these predictions are met by the PAL10/PAL23 association: 1) When the amphioxus scaffold clusters (i.e. these PALs) described here are compared to the zebrafish genome (Zv7), there are significant hits of PAL23 to zebrafish chromosomes 3, 12 and 19, which contain the hoxBa, hoxBb and hoxAa clusters. 2) for all pairs of segments assigned to the same PAL (see above), the rate of paralogous gene pair retention was estimated based on the number of instances of one gene from each segment hitting the same amphioxus gene assigned to the same PAL, divided by the number of genes in the smaller of the two segments. Across all such pairs, the mean rate of retention was 19%, and between the chr 7 and chr 17 PAL 23 segments is was 35% (7 out of 20 genes had a paralogous pair).

The ancient translocation scenario is illustrated below.



Contact: Nik Putnam <NHPutnam@lbl.gov>

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48; Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098; and Los Alamos National Laboratory under contract No. W-7405-ENG-36.