LBNL-59356

Evolutionary Genomics of Life in (and from) the Sea

Jeffrey L. Boore[a,b,c,*], *Paramvir Dehal*[b], and *Susan I. Fuerstenberg*[c]

[a] Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

[b] DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598, USA

[c] Genome Project Solutions, Hercules, CA 94547, USA

[*] To whom correspondence should be addressed:

Addresss: DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

Phone: 925-296-5691

Fax: 925-296-5620

E-mail: jlboore@berkeley.edu

Abstract

High throughput genome sequencing centers that were originally built for the Human Genome Project (Lander et al., 2001; Venter et al., 2001) have now become an engine for comparative genomics. The six largest centers alone are now producing over 150 billion nucleotides per year, more than 50 times the amount of DNA in the human genome, and nearly all of this is directed at projects that promise great insights into the pattern and processes of evolution.

Unfortunately, this data is being produced at a pace far exceeding the capacity of the scientific community to provide insightful analysis, and few scientists with training and experience in evolutionary biology have played prominent roles to date. One of the consequences is that poor quality analyses are typical; for example, orthology among genes is generally determined by simple measures of sequence similarity, when this has been discredited by molecular evolutionary biologists decades ago.

Here we discuss the how genomes are chosen for sequencing and how the scientific community can have input. We describe the PhIGs database and web tools (Dehal and Boore 2005a; http://PhIGs.org), which provide phylogenetic analysis of all gene families for all completely sequenced genomes and the associated “Synteny Viewer”, which allows comparisons of the relative positions of orthologous genes. This is the best tool available for inferring gene function across multiple genomes.

We also describe how we have used the PhIGs methods with the whole genome sequences of a tunicate, fish, mouse, and human to conclusively demonstrate that two rounds of whole genome duplication occurred at the base of vertebrates (Dehal and Boore 2005b). This evidence is found in the large scale structure of the positions of paralogous genes that arose from duplications inferred by evolutionary analysis to have occurred at the base of vertebrates.

The Genomes

In only the last decade, the scientific world has gone from eager anticipation of having a few complete genome sequences of eukaryotes, those of *Drosophila*, *C. elegans*, and human, to now being awash in whole genome sequences. Soon there will be available at least draft, whole genome shotgun sequences of several scores of eukaryotes and many hundreds of prokaryotes (see http://www.genomesonline.org/). This will surely enable great leaps in understanding the biological world, both because of the reagents being made available for follow-on functional genomics studies in some organisms and because of the insights possible from comparing the sequences themselves.

Being effective at this requires a very broad range of biological and computational expertise. As never before, the advance at the technological leading edge of science needs the collective input of organismal biologists, paleontologists, taxonomists, molecular biologists, cytologists, ecologists, computer scientists, and a host of other specialists to find the biology in the sequences. Genome biology, as a field, needs to draw in a broad part of the scientific community for their input in our communal enterprise.

How are Genomes Chosen?

The high-throughput genome centers are soliciting help not only in the analysis after the sequencing, but also in determining how the fantastic sequencing “muscle” developed for the Human Genome Project can now best be applied to biological questions. Two of the major funding sources of genome sequencing in the United States, the National Institutes of Health (NIH) and the Department of Energy (DOE), each have programs for community input. Table 1 lists the URLs for these programs and for other related sites.

The NHGRI (one of the institutes of NIH) currently supports four major sequencing operations: (1) Washington University in St. Louis; (2) Baylor College of Medicine in Houston; (3) Agencourt Bioscience Corporation in Beverly, Massachusetts; and (4) the collaboration between the Whitehead and Broad Institutes in Boston. Their emphasis is on the “outstanding needs of biomedical research that can be addressed through comparative genomic analysis”, although this is defined quite broadly. They have established two “working groups” to guide their choices of genomic targets, one focused on “annotating the human genome” and the other on “comparative genome evolution”. These groups advance their own ideas as well as accept input from the scientific community. Their recommendations are passed to a separate coordinating committee for review, which then submits their recommended priorities to the National Advisory Council for Human Genome Research, which makes the final determination of sequencing targets. In addition, proposals that do not fit well into one of the two areas defined by the working groups may be submitted through their “white paper” process with current deadlines of January 10 and July 10. See instructions at <http://www.genome.gov/11509736>. This program will not consider the genomes of plants, algae, or any prokaryote, and does not accept proposals for EST sequencing, full-length cDNA sequencing, or the development of other genomic resources, although these may be funded by other programs of the NIH or other agencies.

The other major sequencing center in the United States is the DOE Joint Genome Institute (JGI), part of the University of California and operated under contract with the U.S. Department of Energy. Nearly all of the JGI’s sequencing is in response to input from the scientific community, either through the programs run by the DOE Office of Biological and Environmental Research (http://www.sc.doe.gov/production/ober/LSD/DNASeq.html) or by JGI’s “Community Sequencing Program” (CSP; http://www.jgi.doe.gov/CSP/index.html). The CSP accepts proposals annually and evaluates them through a peer-review system.

Approximately 20 billion nucleotides per year of DNA sequencing are allocated under this program.

In addition, there are several major U.S. grant programs that fund high-throughput DNA sequencing, including the National Science Foundation (NSF) program in Plant Genomics, the Microbial Genome Program funded by a joint venture between NSF and the U.S. Department of Agriculture (USDA), and various smaller programs run by DOE, NIH, and USDA.

What's wrong with BLAST?

The pace of producing genome sequences currently overwhelms our ability to interpret them in detail. One of the most urgent challenges is to transfer inferences of gene function across genomes by assigning orthologous relationships. Unfortunately, this most often is done simply by finding pairs of genes, one in each genome, that are reciprocally the most similar by BLAST (Altschul et al. 1990) score. This might produce acceptable results if all DNA sequences evolved at exactly the same rate, but this is not the case. Analogously, methods of phylogenetic reconstruction that rely solely on sequence similarity, like UPGMA ("unweighted pairwise group method of analysis"), would produce acceptable results if the molecular clock were perfect, but have been long-abandoned by the molecular evolution community (Prager and Wilson, 1978; Lin and Nei, 1991) because this assumption does not hold.

The types of errors commonly generated are illustrated in the hypothetical tree of gene sequences shown in Figure 1. In this case, we imagine a two-member gene family in the common ancestor of mouse and human. These genes diverge over time, such that the mouse "A" gene and the human "B" gene have the more rapid rates of sequence change, as indicated by the longer branches on the tree. Assigning orthology by reciprocal best BLAST score then would erroneously recognize the pair mouse-B and human-A, while making no assignment for mouse-A and human-B (since the best match to mouse-A is human-A, but this is not reciprocal, and similarly for human-B).

This can be corrected by performing a phylogenetic analysis on these gene families that would correctly group the pairs of "A" genes and the pairs of "B" genes, and so recognize the true pattern of orthology and paralogy. This has been recently done and presented through the product termed "PhIGs" (Phylogenetically Inferred Groups) (Dehal and Boore, 2006; http://PhIGs.org). As shown in the schema presented in Figure 2, PhIGs starts by performing a BLAST search of all genes in each pair of genomes, then does a global alignment of those with identified similarity and calculating a similarity score for each gene pair. Instead of implicitly assuming a molecular clock by inferring orthology at this point, though, PhIGs builds gene families that respect the known evolutionary relationships among the organisms and performs a phylogenetic analysis using a maximum likelihood method (Schmidt et al., 2002). The basis of the gene family clustering is shown at the bottom of Figure 2, where the relationships among the organisms are shown. PhIGs builds a graph with each protein sequence forming a node and the pairwise distances forming the edges. The shortest distance is found between any gene from either organism in Clade A and either organism in Clade B. This seeds a single-linkage clustering such that other genes are drawn in if they are more similar within each clade than the seed, and if they are more similar within the ingroup than to the outgroup genes. In practice, when more genomes are included, PhIGs starts at the base of the tree to build clusters, then works iteratively toward the tips of the tree of organisms drawing in genes not included in more basal groups. This recruits genes that are newly arising in evolution or that have diverged beyond recognition between long-separated lineages. Once the gene family clusters are built, then PhIGs uses Tree-Puzzle (Schmidt et al., 2002), a maximum likelihood phylogenetic reconstruction program, to infer the evolutionary relationships among all of the included genes. These are compared in an automated way to the tree of the evolutionary relationships among the organisms to determine when each gene duplication and loss occurred and to assign orthologous and paralogous relationships. Hidden Markov models are also built using Hmmer (http://hmmer.wustl.edu) so that users can assign individual genes from non-included organisms to clusters. The PhIGs analysis currently includes 409,653 genes from the complete draft sequences of 23 eukaryotes, organized into 42,645 gene families. Users can query by keyword searches on annotations, including terms from gene ontology (GO) or InterPro, or perform sequence similarity searches by BLAST and HMM to input sequences.

As an ancillary tool, PhIGs includes the "Synteny Viewer" (Figure 3). The user can specify an interval of genes in one genome, then specify one or more other genomes for the comparison, and the relative arrangements are shown for all orthologs.

An example of the use of PhIGs: Two rounds of whole genome duplication at the base of vertebrates

PhIGs has proven useful for addressing real issues in biology. The hypothesis that two rounds of whole genome duplication occurred at the base of the vertebrates (the so-called "2R hypothesis") has long been controversial (Ohno, 1970; Meyer and Schartl, 1999; McLysaght, Hokamp, and Wolfe, 2002; Friedman and Hughes, 2003). Although early anecdotal evidence in favor of the 2R hypothesis came from observations of a 1:4 ratio of some invertebrate-to-vertebrate genes (Popovici et al., 2001), this is now known to be true of only a very small proportion of genes (Meyer and Schartl, 1999; Friedman and Hughes, 2003; Dehal and Boore, 2005). Using PhIGs with the complete gene repertoires of four animals – a tunicate, fish, mouse, and human – allowed the determination of the subset of genes that have paralogs stemming from duplications at the base of the vertebrates (Dehal and Boore, 2005). When looking at the relative positions of these paralogs (but not those generated by later-occurring duplications), a distinctly four-fold pattern emerged.

The principle is shown in the hypothetical example in Figure 4. The genome is represented by the series of colored blocks, each being a gene (Figure 4A). The genome duplication shown in Figure 4B generates a complete set of duplicated genes; many of these supernumerary genes are then eliminated to generate the arrangements shown in Figure 4C. An additional genome duplication (Figure 4D), followed by further gene losses (Figure 4E) generates a pattern where few gene families have four members, but the large-scale pattern of the resulting paralogs is four-fold. This is the pattern of the paralogs in the human genome that result from duplications timed to be at the base of the vertebrates (Dehal and Boore, 2005).

Complete genome sequences are being produced faster than ever before. Input from the broad scientific community is urgently needed, both for the most sensible targeting of organisms as well as for interpreting their biological meaning. A significant part of the future for understanding the dynamics of ocean biosystems will be found in the data from high-throughput comparative genomics.

Acknowledgements

Thanks to the organizers of the DOBIS symposium in November, 2005. This work was supported by funding from the National Science Foundation (MCB-0242131, EF-0328516, and DEB-0445047) and was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231. This work was performed under the auspices of the US Department of Energy's

Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990) Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.

Dehal, P., and J. L. Boore (2005) Two rounds of genome duplication in the ancestral vertebrate genome. *PLoS Biology* **3(10):** e314.

Dehal, P., and J. L. Boore (2006) A phylogenomic gene cluster resource: The Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*. In press.

Friedman, R., and A. L. Hughes (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* **20:** 154-161.

Lander E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lin, J., and M. Nei (1991) Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. *Mol. Biol. Evol.* **8:** 356-365.

McLysaght, A., K. Hokamp and K. H. Wolfe (2002) Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31:** 200-204.

Meyer, A., and M. Schartl (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11:** 699-704.

Ohno, S. (1970) Evolution by Gene Duplication. Berlin: Springer-Verlag. 160 pp.

Popovici, C., M. Leveugle, D. Birnbaum and F. Coulier (2001) Homeobox gene clusters and the human paralogy map. *FEBS Lett*. **491:**237-242.

Prager, E. M., and A. C. Wilson (1978) Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. *J. Mol. Evol.* **11:** 129-142.

Schmidt, H. A., K. Strimmer, M. Vingron and A. von Haeseler (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502-504.

Venter J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, et al. (2001) The sequence of the human genome. *Science* **291:** 1304-1351.

Table 1. Genomics information on the web

Genomes Online	http://www.genomesonline.org/
NIH Genome Sequencing Input	http://www.genome.gov/10002189
NIH Genomes Status	http://www.genome.gov/10002154
DOE Sequencing Programs	http://www.sc.doe.gov/production/ober/LSD/DNASeq.html
NSF Plant Genome Research Program	http://www.nsf.gov/pubs/2005/nsf05603/nsf05603.htm
NSF/USDA Microbial Genome Program	http://www.nsf.gov/pubs/2003/nsf03526/nsf03526.htm
DOE Joint Genome Institute	http://www.jgi.doe.gov/
Washington University Genome Ctr	http://genome.wustl.edu/
Sanger Institute	http://www.sanger.ac.uk/
Broad Institute	http://www.broad.mit.edu/
Whitehead Institute	http://www.wi.mit.edu/
Baylor Genome Center	http://www.hgsc.bcm.tmc.edu/
NIH Intramural Sequencing center	http://www.nisc.nih.gov/
PhIGs, Phylogenetically Inferred Groups	http://phigs.org/

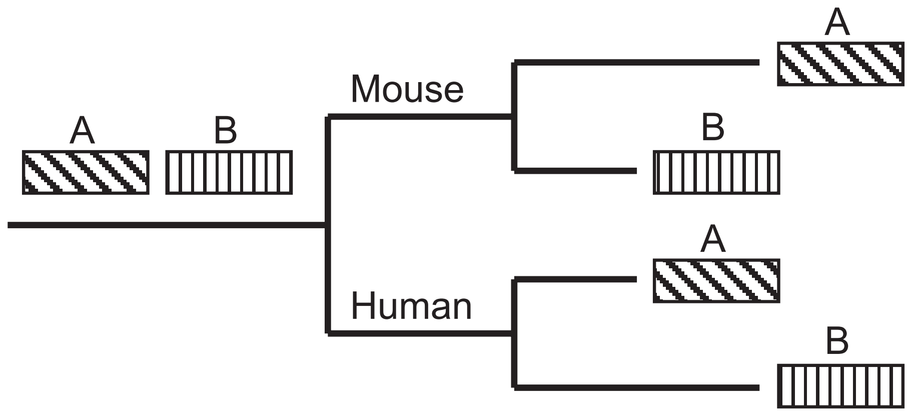

Figure 1. Phylogeny of a hypothetical gene family in human and mouse. See text.

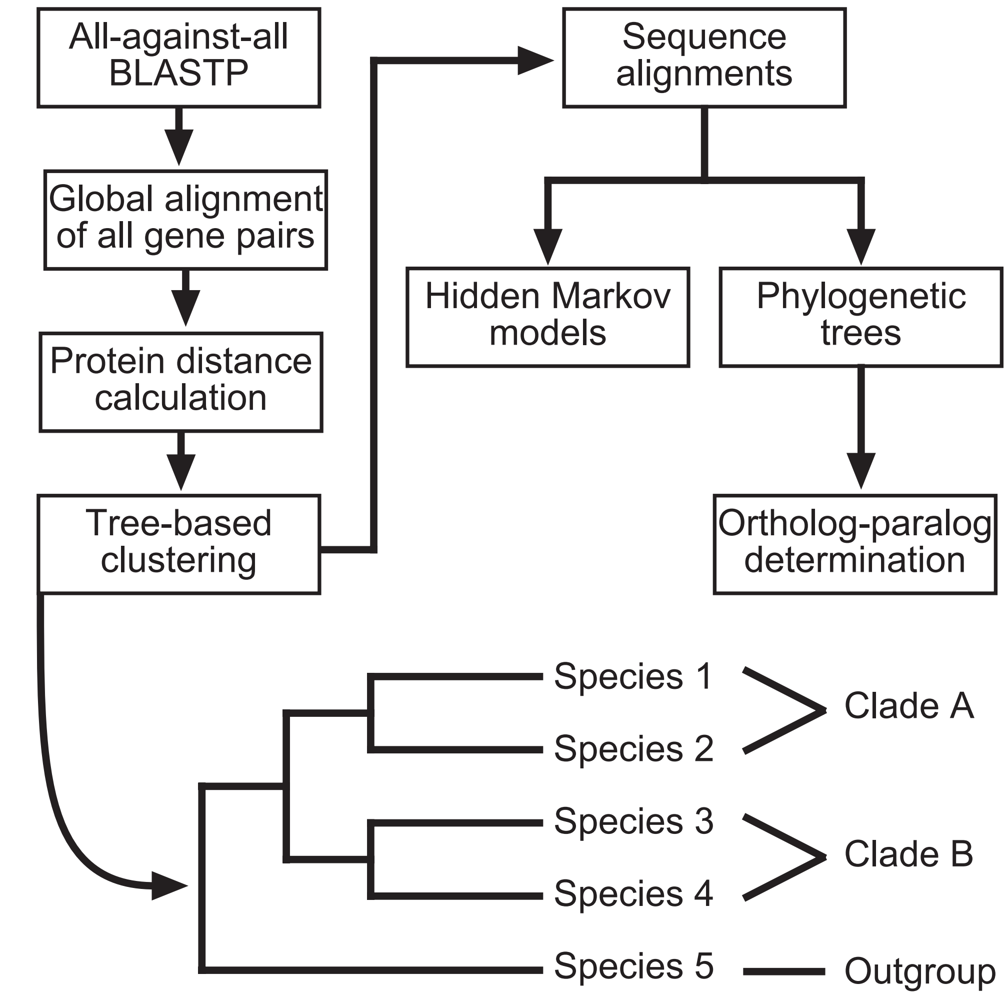

Figure 2. Schema for the PhIGs database and query tools. See text.

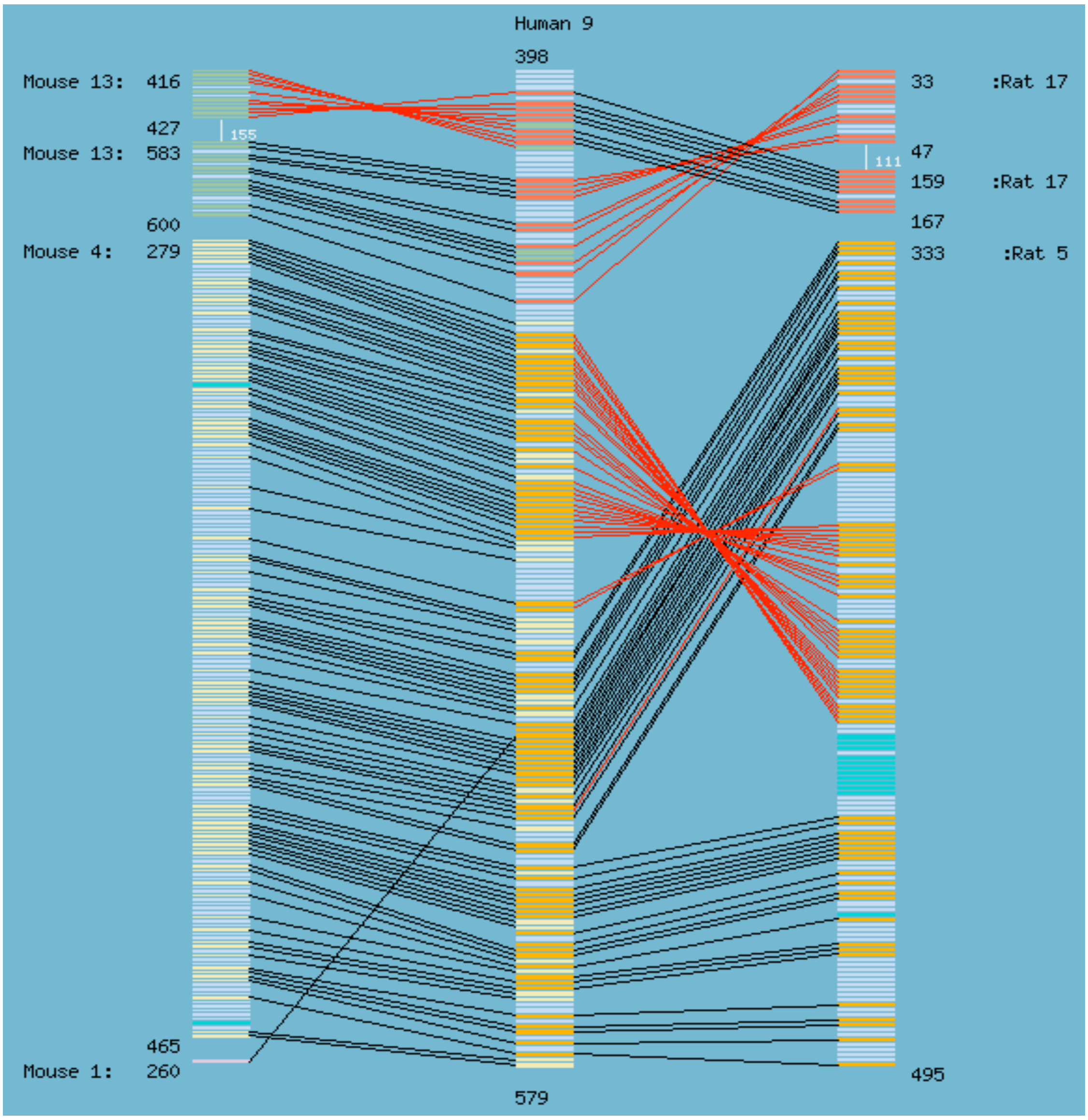

Figure 3. The "Synteny Viewer" allows simultaneous visualization of the relative arrangements of orthologous genes in multiple genomes.

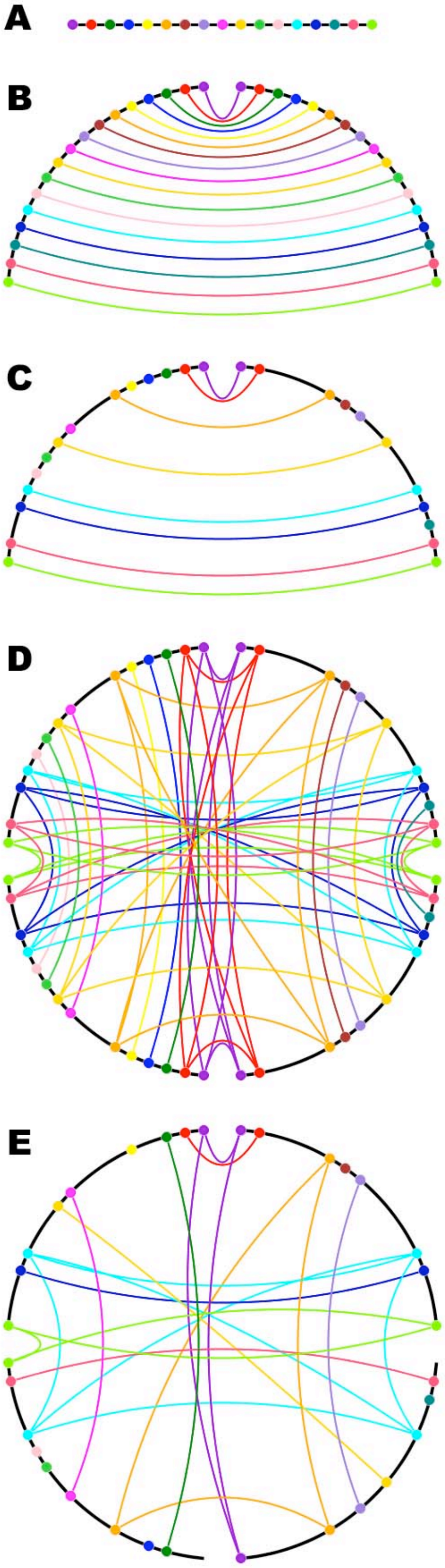

Figure 4. Model for generating four-fold paralogons after two rounds of genome duplication.