

Abstract

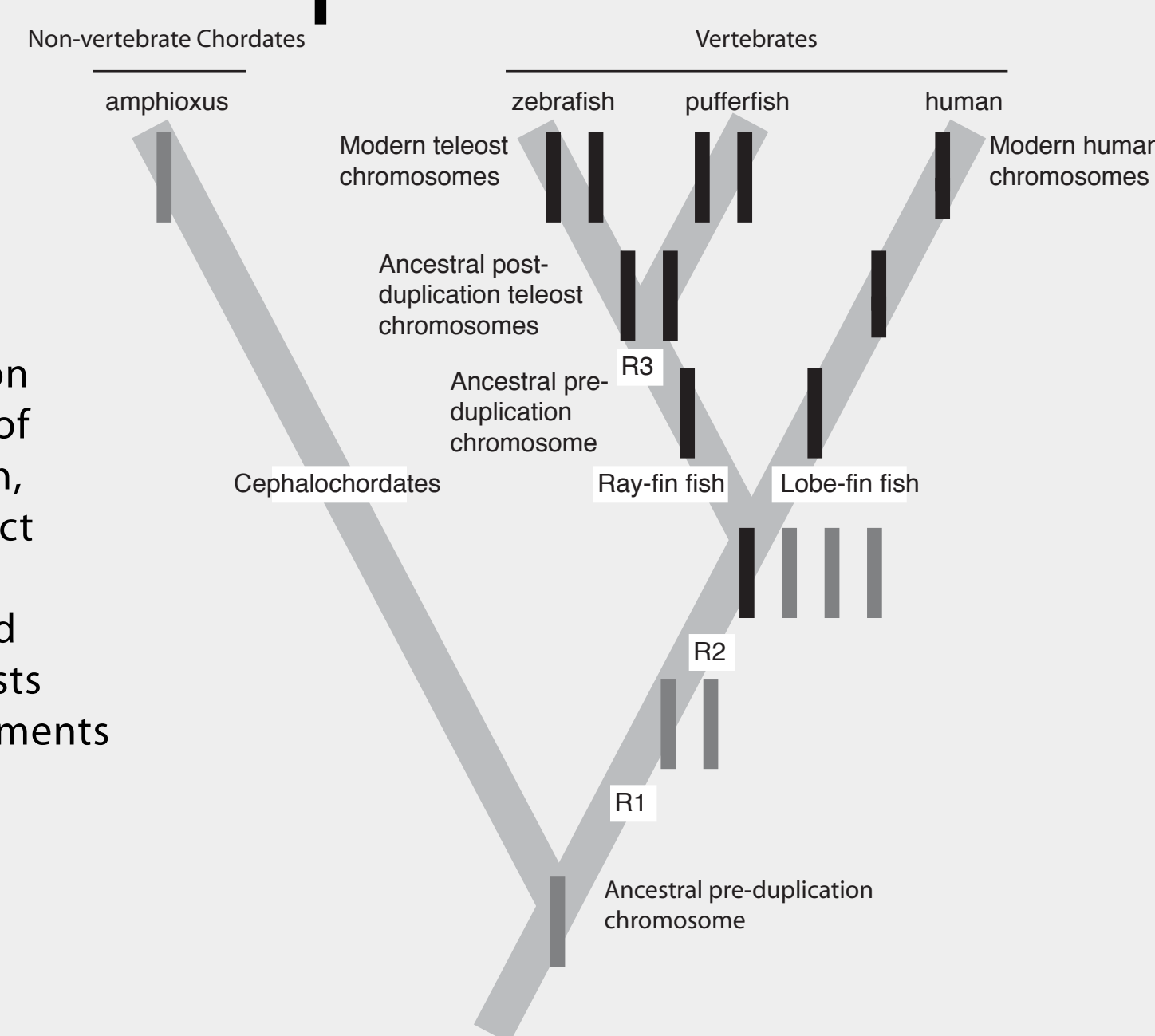
To explain the evolutionary mechanisms by which populations of organisms change over time, it is necessary to first understand the pathways by which genomes have changed over time. Understanding genome evolution requires comparing modern genomes to ancestral genomes, which thus necessitates the reconstruction of those ancestral genomes. Here we describe automated approaches to infer the nature of ancestral genomes from modern sequenced genomes. Because several rounds of whole genome duplication have punctuated the evolution of animals with backbones, and current methods for ortholog calling do not adequately account for such events, we developed ways to infer the nature of ancestral chromosomes after genome duplication. We apply this method here to reconstruct the ancestors of a specific chromosome in the zebrafish *Danio rerio*.

Background

Genome Duplication

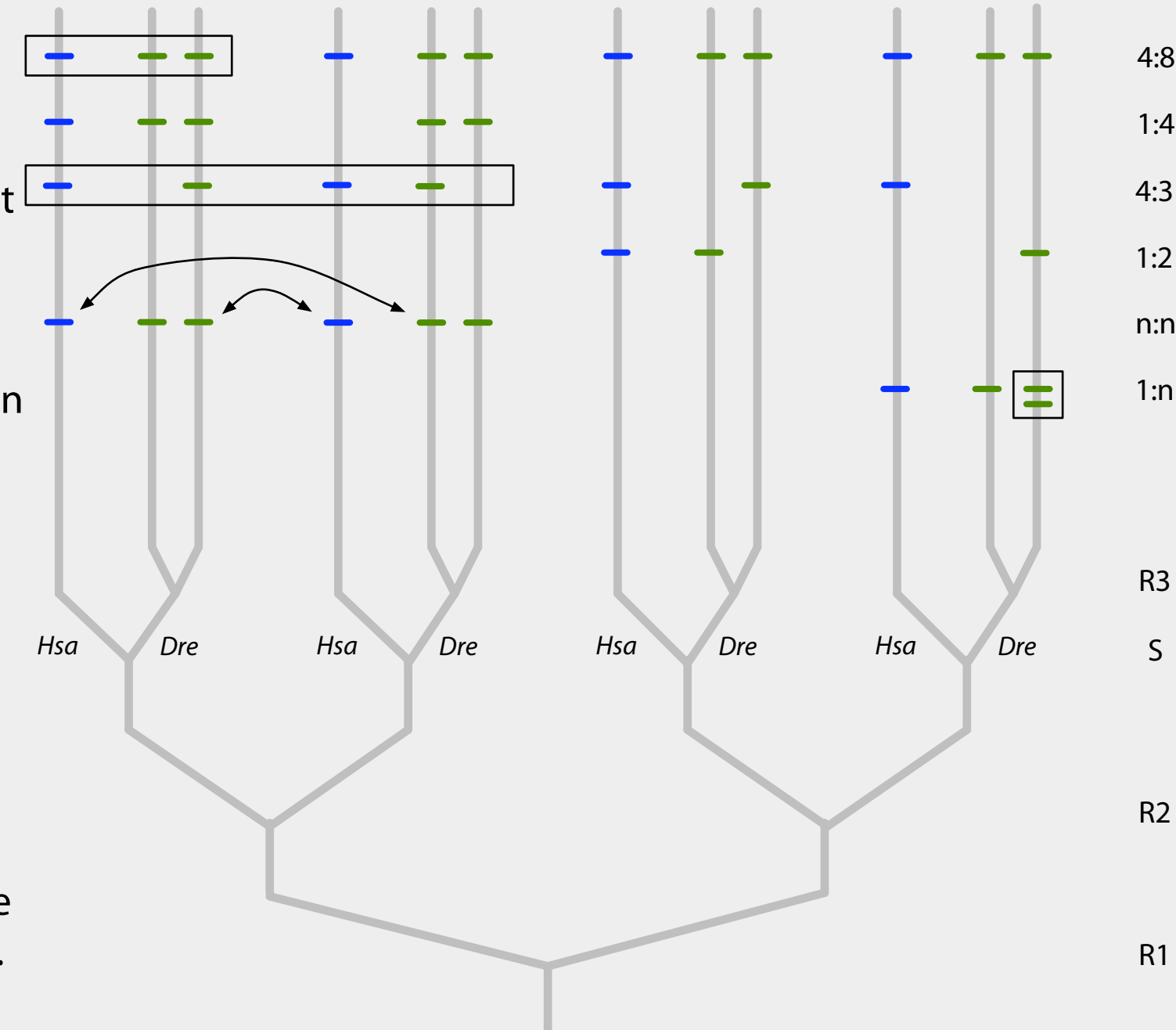
Two consecutive genome duplication events, called R1 and R2, occurred in early vertebrate evolution about 500 million years ago or more.

An additional round of genome duplication (R3) occurred at the base of the radiation of teleost fish (the crown group of ray-finned fish, like zebrafish, tilapia, and pufferfish, distinct from basally diverging ray-finned fish, like sturgeon and gar). This R3 event generated duplicate chromosome segments in teleosts corresponding to single chromosome segments in humans and other mammals.



Reconstructing ancestral chromosomes requires us to examine modern genomes from several organisms and infer their ancestral states. At its core, this is an exercise in large-scale sequence matching, but we must be able to differentiate between several different types of evolutionary events.

The types of evolutionary relationships our software analysis must detect are shown to the right. The figure is a gene tree that shows evolutionary events that led to the current genomes for human and zebrafish: two rounds of genome duplication (R1 and R2), speciation (S), and a third round of genome duplication (R3) in the teleost lineage.

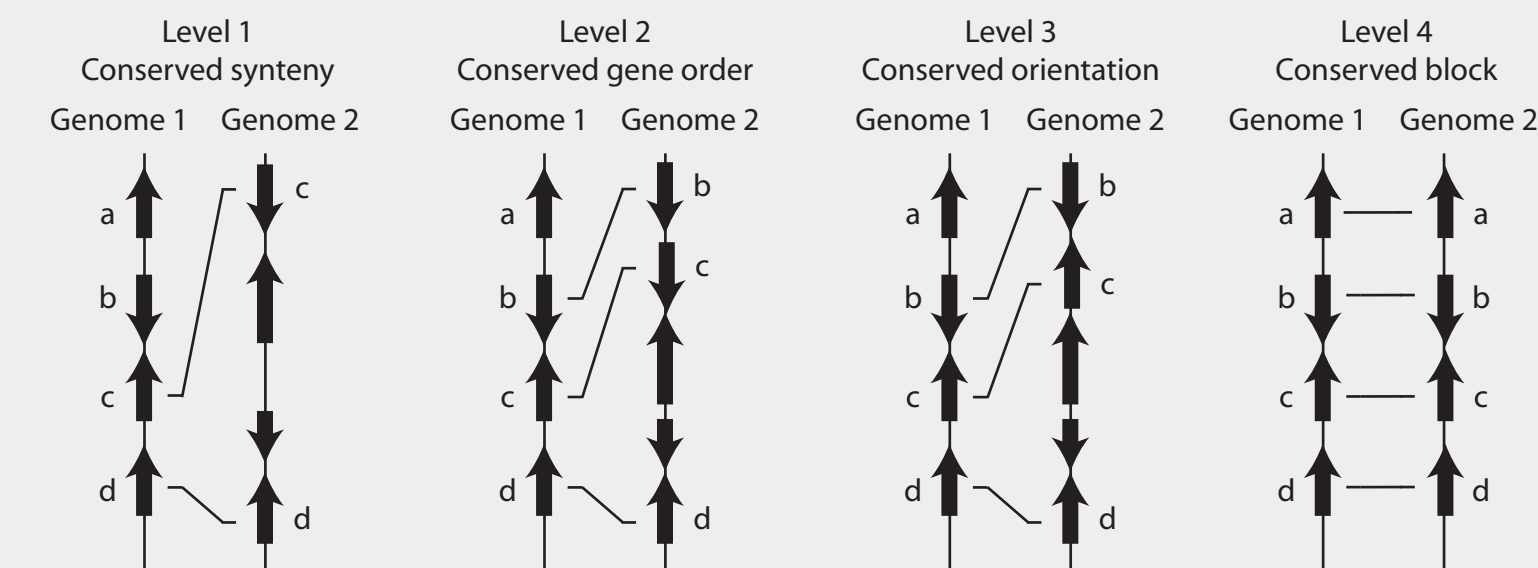


The 4:8 row is ideal, and might be seen for genes that encode essential functions, e.g. for HOX clusters: all 4 Hsa descendants and 7 of the 8 Dre descendants of the original ancestral clusters have survived.

In most cases, the number of identifiable genes that have survived is much lower.

Syntenic Conservation

Conserved syntenies suggest that all members of the group of genes were syntenic in the last common ancestor of the two genomes.



Level 1 is the conservation of syntenies between a pair of genomes. In a conserved synteny, two or more genes that are syntenic in one genome have orthologs that are also on a single chromosome in another genome.

Level 2 is a group of three or more genes that shows conserved syntenies could in addition display conserved gene order: the order of genes in one genome is the same as the order of their orthologs in the other genome.

Level 3 is the conservation of transcription orientation within a group of genes that possesses conserved gene order. Loss of conserved orientation reflects an inversion involving a single gene in the lineage of one genome relative to the other.

Level 4 involves a conserved genomic block in which all genes in the block in one species have orthologs that are in the same order and have no intervening or missing genes in the other species.

Definitions

Hsa: *Homo sapiens* (human)
Dre: *Danio rerio* (zebrafish)
Tni: *Tetraodon nigroviridis* (pufferfish)
Tni7 or DreLG3: refers to a specific chromosome for that species. 'LG' refers to a Linkage Group, which is a way to refer to a chromosome that is not fully assembled.

Ortholog: genetic entities in different genomes descended from a single entity in the last common ancestor of those genomes.
Paralog: genes within the same genome descended from a single gene that was duplicated.
Paralogon: a pair of chromosomes within an organism that were produced by a full chromosome duplication.
Syntenic: conservation of gene order on chromosomes.

Inferring Ancestral Chromosomes

Julian Catchen^{1,2}, John Conery¹, and John Postlethwait²

¹ Department of Computer and Information Science, University of Oregon, Eugene OR
² Institute of Neuroscience, University of Oregon, Eugene OR

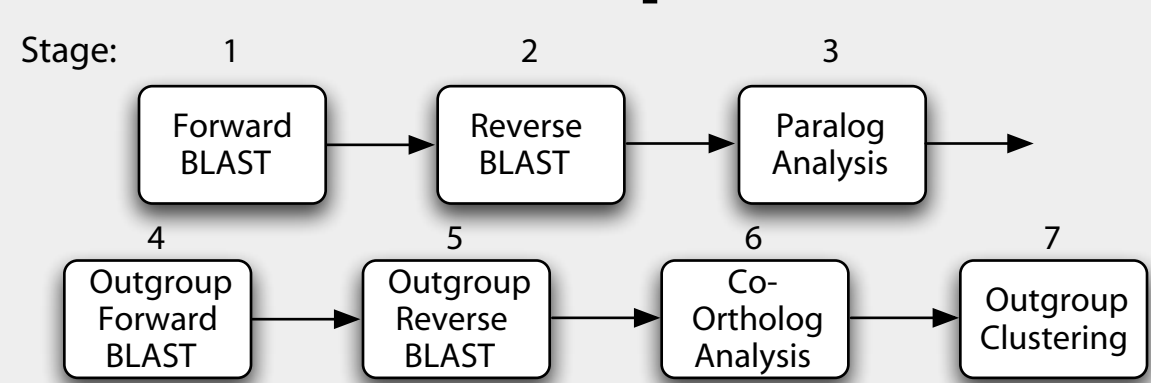
Analysis Pipelines

To infer the content of ancestral chromosome sequences, we must conduct two major analyses, the identification of paralogs within a species and the identification of orthologs between species.

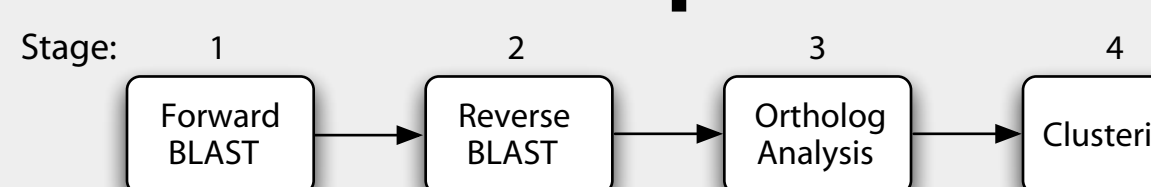
The **PIP Software Pipeline** conducts each analysis, managing data and executing the bioinformatic applications that process it. PIP (pipeline interface program) is a generic framework that allows us to create many different 'pipelines' by combining arbitrary analysis stages in different orders. The initial data set (gene sequences and annotations) are stored in a relational database. PIP then runs each application in turn, passing results from one analysis stage as inputs to the next analysis stage, archiving the results of each stage in the database.

Each of the two major analysis steps, identifying paralogs and identifying orthologs, is embodied in its own pipeline as shown to the right.

Orth Pipeline

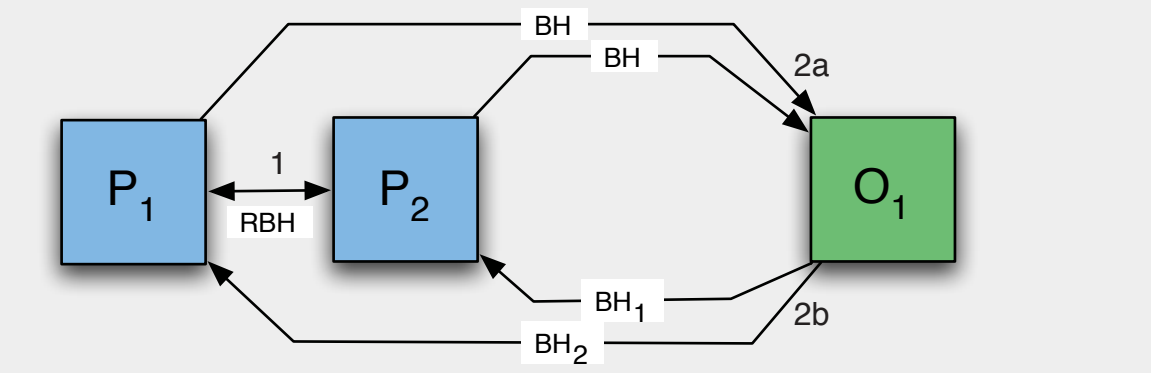


Para Pipeline

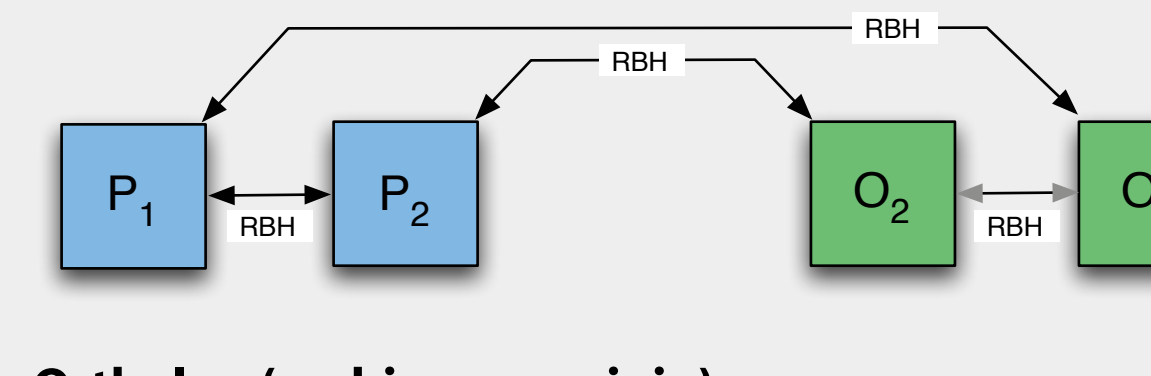


RBH Defined

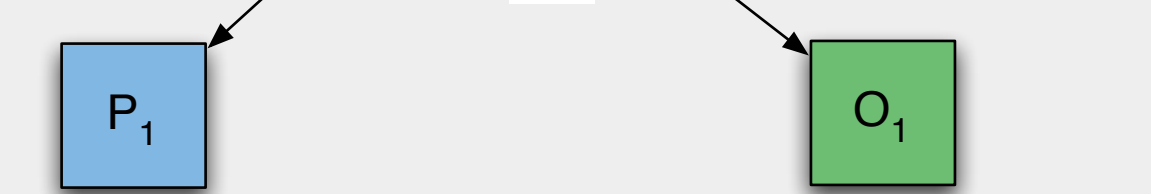
Co-orthologs (mostly R3 Duplicates)



Ancient paralogs (R1 or R2 Duplicates)



Ortholog (ambiguous origin)



If P₁ and P₂ in the primary organism match a single ancient outgroup gene (O₁) and gene O₁ has two best hits genes P₁ and P₂, then P₁ and P₂ are co-orthologs of outgroup gene O₁, and were produced either in R3 or in a recent tandem duplication event.

If the two sister genes (P₁ and P₂) in the primary organism have as RBH different ancient outgroup genes (O₁ and O₂), then the pipeline has identified a pair of ancient paralogs produced in the first or second round of genome duplication.

If only one member of the pair has an RBH with the outgroup, then the origins of this gene pair are ambiguous.

Results

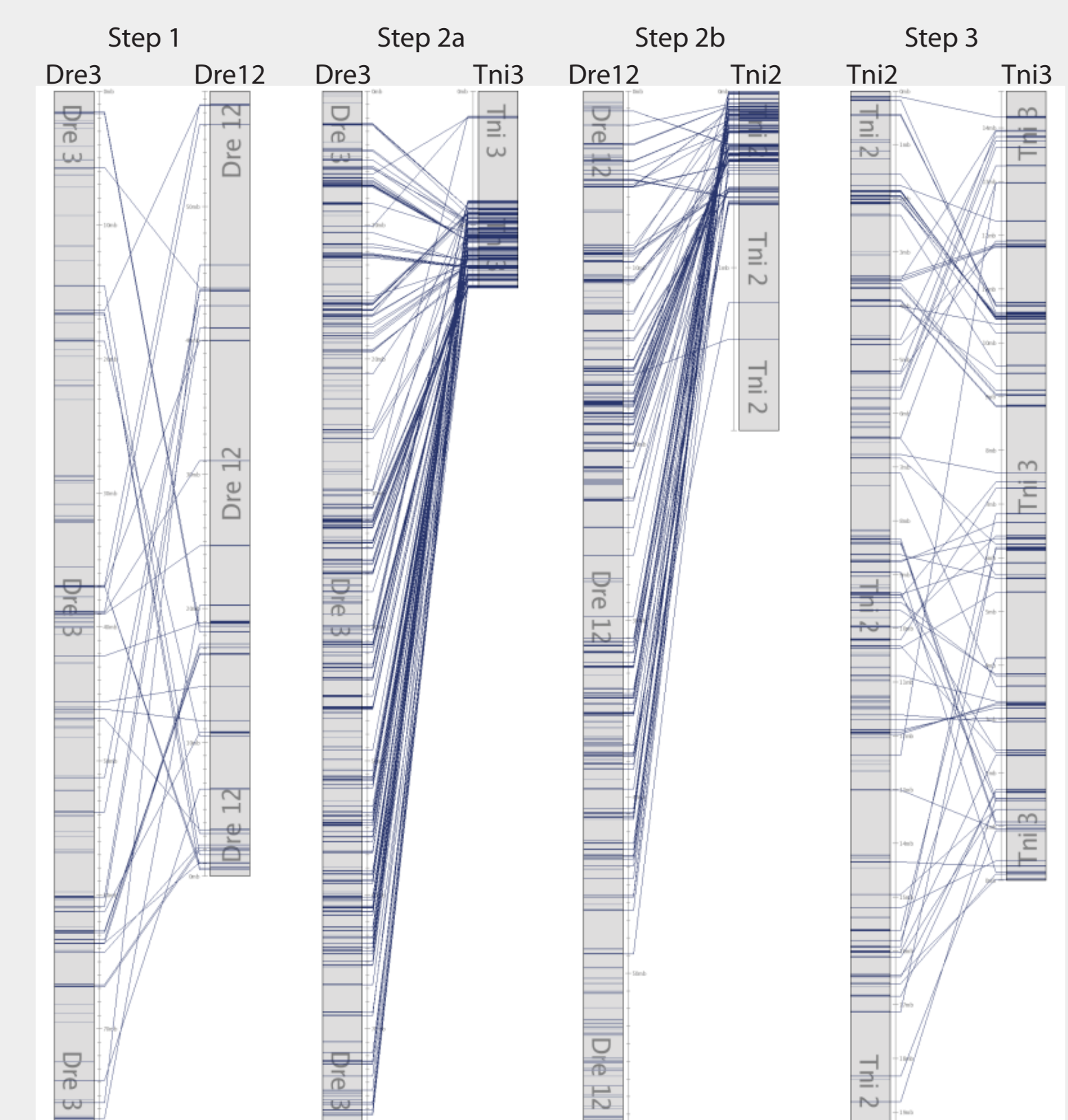
Our initial focus was to reconstruct the ancestral chromosome and gene orders for the *Danio rerio* Linkage Group 3 (DreLG3), one of the 25 zebrafish chromosomes.

Step 1. Identify paralogous genes within the *Danio rerio* genome to infer chromosome segments that constitute the most likely paralogon. PIP identified Dre12.

Step 2. Take genes from DreLG3 and DreLG12 and search for orthologous genes in the pufferfish. PIP identified Tni3 as orthologous to Dre3, and Tni2 as most closely related to Dre12.

Step 3. Identify paralogous genes within the pufferfish genome. PIP identified Tni3 as the Tni2's paralogon.

Our data demonstrates the principle of **transitive homology**.



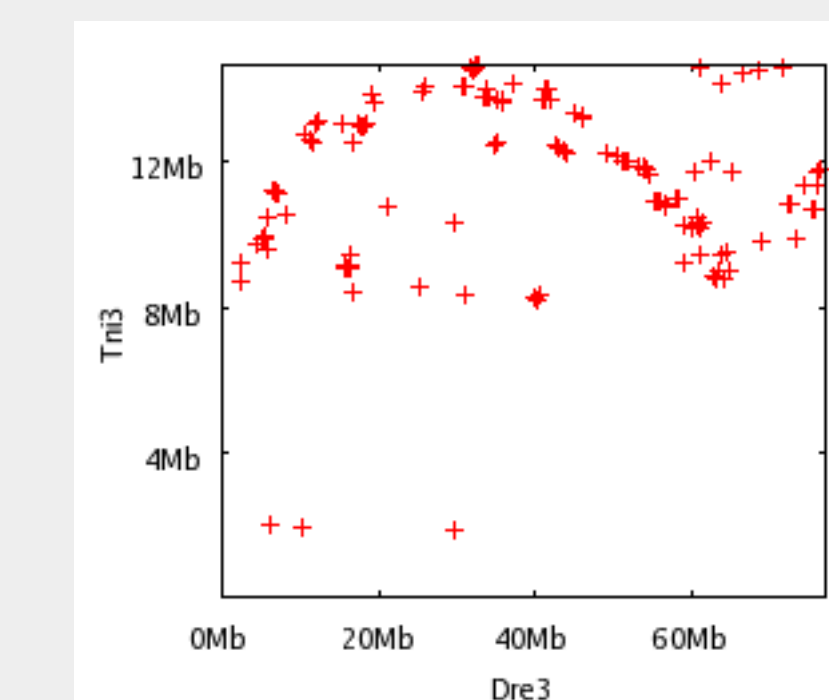
To cluster the predicted orthologs, we construct a gene homology matrix (GHM), as pictured to the right.

Each point in the GHM plot represents an orthologous gene pair.

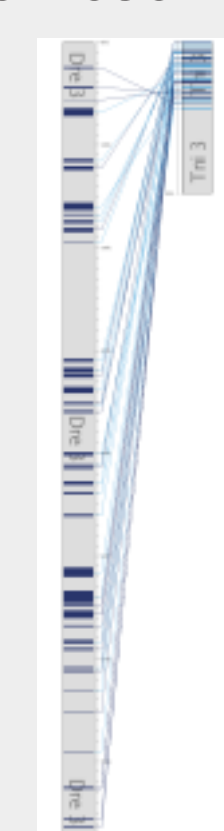
Clusters appear as diagonal lines on the GHM plots; an uninterrupted slope indicates conserved gene order, and the reversal of slope from positive to negative indicates an inversion.

The comparative chromosome map shows the detected clusters. Inverted clusters are drawn in light blue.

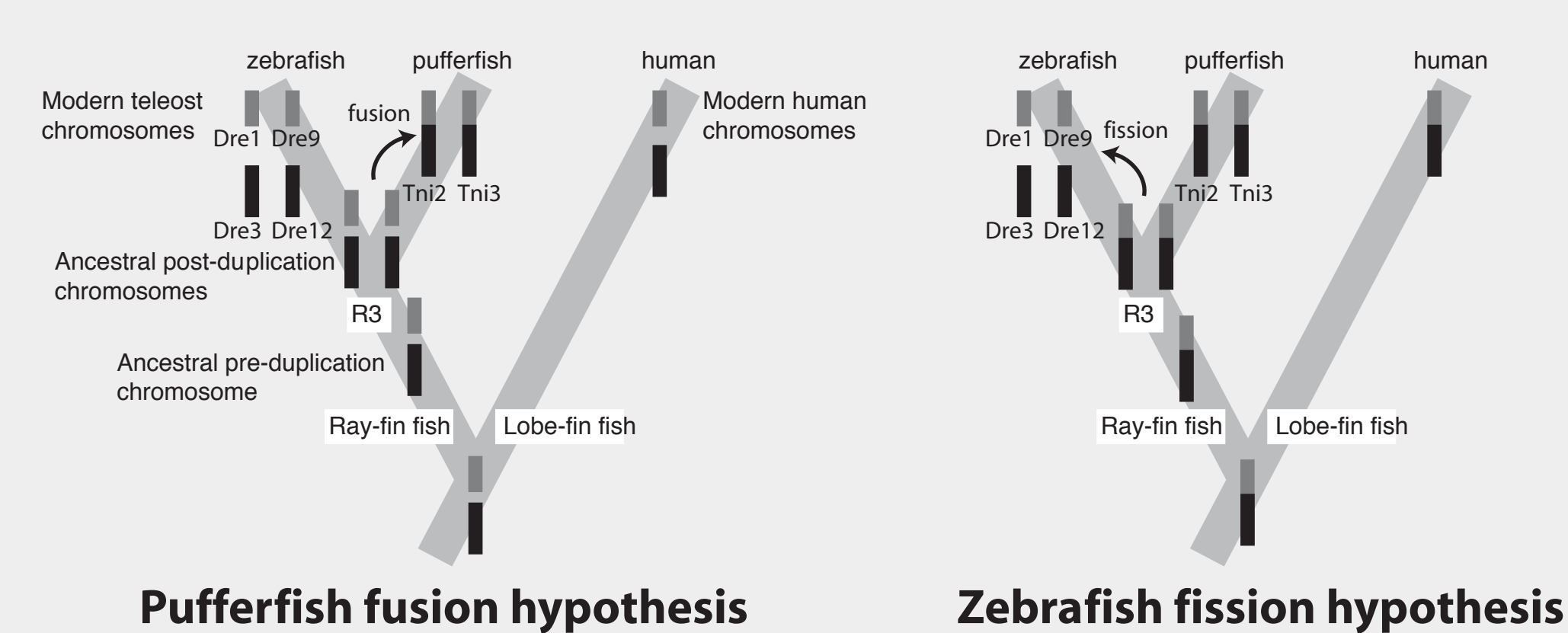
Clustering Analysis with Gene Homology Matrix



Comparative Chromosome Map



Discussion



The pipeline shows that the non-DreLG3 portion of Tni3 (corresponding to DreLG1), and the non-DreLG12 portion of Tni2 (orthologous to DreLG9) are both orthologous to the long arm of human chromosome two. This type of relationship would be expected according to the zebrafish fission hypothesis but not according to the pufferfish fusion hypothesis. Therefore, we conclude that the ancestral preduplication chromosome that was the ancestor to DreLG3 consisted of a chromosome that was substantially similar to the sum of the genetic content of pufferfish chromosomes Tni2 and Tni3.

