Automated identification of conserved synteny after whole genome duplication

Julian M. Catchen, John S. Conery and John H. Postlethwait

Genome Res. published online May 22, 2009
Access the most recent version at doi:10.1101/gr.090480.108

Automated identification of conserved synteny after whole genome duplication

Julian M. Catchen[1,2], John S. Conery[1], John H. Postlethwait[2,*]

May 6, 2009

[1] Department of Computer and Information Science, University of Oregon, USA

[2] Institute of Neuroscience, University of Oregon, USA

[*] Corresponding author.

Phone: 541-346-4538; Fax: 541-346-4548; Email: jpostle@uoneuro.uoregon.edu

Running Title: Automated identification of conserved synteny

Keywords: genome duplication, conserved synteny, ohnologs, gene family evolution

Abstract

An important objective for inferring the evolutionary history of gene families is the determination of orthologies and paralogies. Lineage specific paralog loss following whole genome duplication events can cause anciently related homologs to appear in some assays as orthologs. Conserved synteny – the tendency of neighboring genes to retain their relative positions and orders on chromosomes over evolutionary time – can help resolve such errors. Several previous studies examined genome-wide syntenic conservation to infer the contents of ancestral chromosomes and provided insights into the architecture of ancestral genomes, but did not provide methods or tools applicable to the study of the evolution of individual gene families. We developed an automated system to identify conserved syntenic regions in a primary genome using as outgroup a genome that diverged from the investigated lineage before a whole genome duplication event. The product of this automated analysis, the Synteny Database, allows a user to examine fully or partially assembled genomes. The Synteny Database is optimized for the investigation of individual gene families in multiple lineages and can detect chromosomal inversions and translocations as well as ohnologs (paralogs derived by whole-genome duplication) gone missing. To demonstrate the utility of the system, we present a case study of gene family evolution, investigating the ARNTL gene family in the genomes of *Ciona intestinalis*, amphioxus, zebrafish, and human.

1 Introduction

An important objective for inferring the evolutionary history of gene families and chromosome segments is the determination of orthology and paralogy. A stepwise approach generally uses BLAST (basic local alignment search tool) (Altschul et al., 1997) to define coarse relationships among genes followed by phylogenetic reconstruction to suggest more detailed hypotheses of descent. Events such as gene duplications or whole genome duplications (WGD), with associated differential gene loss, introduce noise into these analyses. Anomalies, such as lineage specific paralog loss, can cause anciently related homologs to appear to be orthologs, thereby confusing sequence similarity with functional homology (Postlethwait, 2006). Such errors can confound attempts to create non-human animal disease models and can obscure recent, species-specific evolutionary change among sister lineages.

Orthologs are two genes, one in each of two species, that descended from a single gene in the last common ancestor of those two species. Paralogs are a set of genes derived by duplication within a lineage, and together, a group of paralogs can be co-orthologous to their unduplicated ortholog in a related species. Ohnologs are a special subset of paralogs that result from a whole genome duplication event (Wolfe, 2000). The differential loss of genes that follows a duplication event can create *ohnologs gone missing* when different ohnologs are lost reciprocally in different lineages.

Understanding and distinguishing ohnologs gone missing from orthologs is a pervasive problem in vertebrate genomics due to multiple genome duplication events. Two rounds of whole-genome duplication events called R1 and R2 likely occurred at the base of the vertebrate lineage after the divergence of non-vertebrate chordates and prior to the appearance of jawed vertebrates (Garcia-Fernàndez and Holland, 1994; Spring, 1997; Dehal and Boore, 2005). A third duplication called R3 likely occurred in the teleost lineage after the divergence of ray-finned and lobe-finned fishes (Amores et al., 1998; Taylor et al., 2003; Jaillon et al., 2004) but before the radiation of the teleosts. Additional genome duplications punctuated the evolution of other lineages, like salmonids, catastomids, goldfish,

Xenopus laevis, and even a rodent (Allendorf and Thorgaard, 1984; Mungpakdee et al., 2008a,b; Uyeno and Smith, 1972; Risinger and Larhammar, 1993; Larhammar and Risinger, 1994; David et al., 2003; Schmid and Steinlein, 1991; Gallardo et al., 1999). Given the pervasive nature of genome duplication in chordates, and the importance of teleost fish and *Xenopus laevis* as model organisms, it is important to develop automated methods to identify true orthologs among groups of paralogs and to distinguish them from more ancient, non-orthologous homologs.

Figure 1 illustrates the problem of distinguishing orthologs following duplication and lineage-specific loss of a gene g and some of its neighboring genes after WGD (R1), speciation (S) and a second WGD event (R2) in one of the descendant lineages. In an idealized case, chromosomes would experience few changes in gene order or gene content, as illustrated by genes of the same color in Figure 1. The most common fate of genes created by a WGD event, however, is pseudogenization and nonfunctionalization (Li, 1980; Watterson, 1983). Surviving duplicates can develop new functions (Ohno, 1970) or partition or lose their existing functions (Force et al., 1999; Lynch and Force, 2000; Winkler et al., 2003; Postlethwait et al., 2004; Jovelin et al., 2007; Jarinova et al., 2008; Chain et al., 2008; Conant and Wolfe, 2008). From the time of the duplication event to the present, duplicated genes can alter their expression patterns (Force et al., 1999) or their exon structure (Altschmied et al., 2002), or their activities (Zhang et al., 2002; Zhang, 2003), and such changes can alter protein-protein interactions or subsequent developmental or physiological functions.

In the case of differential gene loss and gene rearrangements in lineages S1 and S2, most reciprocal best hit BLAST algorithms (Wall et al., 2003) would associate gene *g2* with *g1a* and *g1b* and most phylogenetic methods, due to a lack of data, would find that the most likely hypothesis of descent was that genes *g2*, *g1a*, and *g1b* shared their most recent common ancestor; in other words, these methods would incorrectly infer that *g1a* and *g1b* were orthologs of *g2*. The erroneous assignment of orthology presents a problem because it implies that the last common ancestor at time S had a single gene with a set of functions

that evolved to *g1* (and its subsequent duplicates, *g1a* and *g1b*) in S2 and *g2* in S1, but in fact, no such gene actually existed.

To address this problem, and to better infer orthologies and paralogies, we can take advantage of conserved synteny – the tendency of neighboring genes to retain their relative positions and orders on chromosomes over evolutionary time. In a WGD event, duplicated chromosomes (homeologs) initially have gene orders identical to each other and to their immediate ancestor. Between the time of duplication and speciation events, however, genes can be lost from one homeolog or the other (unless preserved by structures such as embedded regulatory elements (Kikuta et al., 2007)), and inversions and other chromosome rearrangements can occur independently on the two duplicated homeologs. These events occurring in the chromosomal vicinity of the gene in question give an identity to all of the genes in the neighborhood. In the example given in Figure 1, we could test the hypothesis that genes *g1a* and *g1b* are co-orthologous to gene *g2* by first examining the neighbors of *g1a* and *g1b* – ensuring that a sufficient number of gene neighbors are also paralogous – and then by checking those neighboring paralogs to ensure that they are orthologous to the neighbors of *g2*. The conserved syntenic region defined by such genes would confirm (or in this case, reject) the co-orthology of genes *g1a* and *g1b* to *g2*. This approach complements the use of BLAST and phylogenetic reconstruction and provides additional evidence to infer the evolutionary history of gene families independent of sequence identities.

Several previous studies examined syntenic conservation at a genomic level to determine the nature of the ancestral chromosomes for that organism's lineage. Evidence for two rounds of genome duplication in stem vertebrates came from a whole-genome analysis of human, mouse, and fugu pufferfish using the urochordate *Ciona intestinalis* as an outgroup (Dehal and Boore, 2005). Analysis of the *Tetraodon nigroviridis* (green spotted pufferfish) genome and the construction of a dense meiotic map for medaka, supported earlier conclusions (Amores et al., 1998; Postlethwait et al., 1998; Woods et al., 2000; Postlethwait et al., 2002; Taylor et al., 2003; Van de Peer et al., 2003) that a third genome duplication had occurred in the teleost fish. Analysis of *Tetraodon* and medaka provided evidence for a

twelve chromosome ancestral vertebrate genome by calculating conserved syntenic regions between the fish and human genomes (Jaillon et al., 2004; Naruse et al., 2004). Subsequent work reconstructed the ancestral vertebrate genome using data from human, chicken and medaka genomes (Nakatani et al., 2007) and, in opposition to earlier work (Jaillon et al., 2004; Naruse et al., 2004; Woods et al., 2005), concluded that the osteichthyan ancestor had approximately 40 chromosomes. These studies provided insights into the architecture of the ancestral genome, but were not convenient for the study of the evolution of individual gene families because the methods employed did not form individual syntenic clusters (Jaillon et al., 2004; Dehal and Boore, 2005; Nakatani et al., 2007); instead, they used hand-curated data (Jaillon et al., 2004; Nakatani et al., 2007); or they downplayed portions of the genome that did not fit into the analysis (Dehal and Boore, 2005).

We developed an automated system to identify conserved syntenic regions in a primary genome using an outgroup genome that diverged from the investigated lineage before a whole genome duplication. Our Synteny Database allows for the analysis of fully or partially assembled genomes (Bridgham et al., 2008), and is optimized for the investigation of individual gene families in multiple lineages. The Synteny Database specializes in comparing genomes that have undergone one or more whole-genome duplications; it is able to detect chromosome inversions and translocations as well as ohnologs gone missing in the gene families investigated. To demonstrate the utility and use of the system we present a case study of the evolution of the ARNTL gene family in the amphioxus, *Ciona intestinalis*, zebrafish, and human genomes.

2 Results

The prediction of syntenic clusters allows us to enumerate regions of the genome that have been conserved since the last whole genome duplication (relative to the unduplicated outgroup). These syntenic clusters, in turn, depend on the identification of paralogous genes within a genome along with their corresponding orthologous genes in the outgroup genome.

We built an analysis pipeline to satisfy each of these two dependencies: first, identifying paralogous and orthologous genes and second, discovering clusters of conserved synteny.

Our modified Reciprocal Best Hit (mRBH) Analysis Pipeline identifies paralogous gene groups in a primary genome (rather than a single 'best' hit) and then *anchors* those gene groups to an ortholog in an outgroup genome using a BLAST-based approach. The pipeline naturally creates paralogous groups relative to the last whole genome duplication that occurred in the primary genome but not in the outgroup genome. For example, if the primary genome has experienced a duplication since it diverged from the outgroup genome, then the pipeline will produce gene groups of size two. If, on the other hand, a duplication occurred before two species diverged, then the pipeline reverts to a simple ortholog pipeline with a one-to-one correspondence between genes in the primary and outgroup genomes. In practice, recent tandem gene duplication, gene loss, and sequence divergence heavily influence the number of genes per group.

Given a set of paralogous gene groups in the primary genome that are co-orthologous to a single gene in the outgroup, we wish to look for regions of conserved synteny among paralogous chromosome segments within the primary genome and between orthologous chromosome segments in the primary and outgroup genomes. Our second analysis pipeline, which populates the Synteny Database, employs a sliding window analysis to identify chromosome regions in the primary and outgroup genomes that have been conserved since the last whole genome duplication event while allowing for small-scale changes in gene order, gene orientation, and gene loss. The result is a set of paralogous cluster pairs within the primary genome and a set of orthologous cluster pairs between the primary and outgroup genomes.

The Synteny Database employs a web-based interface to provide paralogous and orthologous gene groups and syntenic clusters to the researcher in a format searchable by gene name or genomic location. The user can also access several web-based visualization tools, including linear and circular plots of paralogs and orthologs to render gene groups and syntenic clusters. The following case study extensively utilizes these web-based tools and

illustrates how researchers can use the Synteny Database to infer gene family histories.

2.1 Case Study: the ARNTL gene family

The Synteny Database provides a useful data set for the examination of the evolutionary history of the ARNTL gene family. The aryl hydrocarbon receptor nuclear translocator-like gene (*ARNTL* or *BMAL1*) is a helix-loop-helix protein that forms a heterodimer with CLOCK to regulate the circadian clock, a system that provides daily periodicity for biochemical, physiological, and behavioral activities (Ikeda and Nomura, 1997; Gekakis et al., 1998; Pando and Sassone-Corsi, 2002). We will test the ability of the mRBH Analysis Pipeline to identify orthologs and paralogs of the ARNTL gene family in the basally diverging chordate amphioxus, the urochordate *Ciona intestinalis* (a sea squirt), the ray fin fish *Danio rerio* (zebrafish), and the lobe fin fish *Homo sapiens*. Then, using the Synteny Database, we will search for conserved chromosome segments surrounding the orthologous or paralogous ARNTL genes. If the amphioxus, *Ciona*, zebrafish, and human ARNTL gene families descended from a single, ancestral gene in the last common ancestor, then we would expect the genomic neighborhood of the ARNTL genes to reflect the existence of R1 and R2 in the vertebrate lineages and R3 in teleost fish. We will use this syntenic conservation to verify each orthologous and paralogous relationship in the ARNTL gene tree and in the process confirm or reject our orthology and paralogy assignments. The full case study is available in the Supplemental Material, here we will discuss two parts to highlight several features the Synteny Database detects: the paralogy assignment in the human genome and one orthology assignment between the human and zebrafish genomes.

2.1.1 ARNTL Paralogs in the Human Genome

We examine the origins of ARNTL paralogs in three steps: output from the mRBH Analysis Pipeline, a comparison of those results to phylogenetic analysis, and inferences obtained from the Synteny Database. According to the results of the mRBH Analysis Pipeline, *ARNTL*, located on human chromosome 11 (Hsa11), has a single paralog in the human

genome, *ARNTL2*, on chromosome 12 (Hsa12) (Hogenesch et al., 2000). Because the genome assembly of *Ciona intestinalis* (Satou et al., 2003) does not contain an ARNTL ortholog, the mRBH pipeline incorrectly anchored the human ARNTL orthologs to the nearest related extant gene in the *Ciona* genome (*Q4H3W4_CIOIN*), which is in reality the ortholog of the human *ARNT* and *ARNT2* genes – ancient paralogs of the ARNTL genes. These conclusions were confirmed by building a phylogenetic tree, which shows that amphioxus, which diverged more basally than *Ciona* in chordate history (Philippe et al., 2005; Blair and Hedges, 2005), has an ortholog of human *ARNT* and *ARNT2* as well as an ortholog of *ARNTL* and *ARNTL2* (Fig. 2A). This analysis emphasizes the problem illustrated by Figure 1: reciprocal BLAST procedures can assign false orthologies in the case of lost gene duplicates. Because the current genome assembly of *Ciona* lacks an ARNTL ortholog, we will use the amphioxus genome as an outgroup to search for syntenic conservation among the human ARNTL paralogs.

2.1.2 Paralogy of Human ARNTL Chromosome Segments

The Synteny Database generates several visualizations, including dotplots, circle plots, and gene traces that the user can download in raster (PNG) and vector (PDF) formats. To our knowledge, this is the only site that provides public access to such visualization tools. A particularly useful display is a dotplot, which plots genes (grey dots) according to their order and relative distance along a user-selected index chromosome displayed along the horizontal axis of the plot in megabases. The paralogs (red dots) of each gene on the index chromosome are plotted vertically above or below on the appropriate chromosomes, ordered with respect to the location of the gene on the index chromosome rather than their order on their native chromosome. Users can specify genes to be circled on the plot and a grey disc shows the index chromosome's centromere, when known. The dotplot readily identifies regions of the index chromosome that are duplicated by a large-scale event, such as a WGD. A paralogy dotplot for Hsa11 (Fig. 2B) showed this duplication pattern within a large region encompassing *ARNTL*. More than 60 megabases (Mb) of Hsa11 contained

genes with paralogs on Hsa12 (green dots), spanning the region that includes *ARNTL2* and providing evidence that this region of Hsa11/Hsa12 was produced in a large-scale duplication event. Hsa19 also showed many paralogs from this region.

While dotplots enhance visualization of data across the entire genome, a gene trace provides a more detailed view of a conserved region. The Synteny Database identified a conserved region of nine pairs of Hsa11/Hsa12 paralogs near *ARNTL* using a sliding window size of 50 (Fig. 2C). To evaluate the relationship of window size and shared gene pairs, we performed a permutation analysis (see Methods). In brief, with longer windows, the likelihood of finding a pair of orthologs that are syntenic in two species will increase solely by chance. According to the permutation analysis, the nine pairs of genes found using the 50-gene window demonstrates conservation from the last common ancestor of the *ARNTL* chromosome segments. Each grey square in a gene trace represents a gene with order, but not distance or size, maintained along the chromosome. Colored genes are members of this particular paralogous cluster while grey genes are not. Lines connect members of the cluster representing paralogs and are colored according to how the sliding window analysis detected them. The colored lines connecting paralogs make chromosome rearrangements readily apparent.

2.1.3 ARNTL Paralogs in Teleost Fish

The hypothesis that teleost fish experienced a third genome duplication after splitting from the lineage that led to humans (Amores et al., 1998; Postlethwait et al., 1998; Taylor et al., 2003; Jaillon et al., 2004; Naruse et al., 2004), predicts that there should be two orthologs (co-orthologs) of each human ARNTL gene in the zebrafish and other teleosts, except for post-duplication gene loss. Additionally, we would expect to find conserved paralogous regions around each pair of zebrafish co-orthologs as well as conserved orthologous regions around each zebrafish/human ortholog pair. To test these predictions, we first queried the mRBH Analysis Pipeline results to identify zebrafish orthologs of human *ARNTL* and *ARNTL2* and then used the Synteny Database to search for conserved synteny in regions

surrounding those orthologs. The ortholog circle plot of Figure 3A summarizes the human and zebrafish syntenic clusters identified by the pipeline. The circle plot, which is a third visualization available from the Synteny Database, displays chromosomes drawn around the circumference of a circle while arcs join orthologous gene pairs positioned relative to their location on the chromosome. Orthologous gene arcs are colored according to their syntenic cluster membership. Users can specify chromosomes, or portions of chromosomes, from the primary genome, or between the primary and outgroup genomes to include in customized circle plots.

The results of the mRBH Analysis Pipeline identified three paralogous zebrafish genes: *arntl1a*, *arntl1b*, and *arntl2*. The output suggested the unexpected result that all three are co-orthologous to human *ARNTL* and none of them were orthologous to *ARNTL2*. Three zebrafish ARNTL genes have been reported in the literature: *arntl1a* and *arntl1b* were said to be orthologous to human *ARNTL* while *arntl2* was thought to be orthologous to *ARNTL2* (Cermakian et al., 2000; Ishikawa et al., 2002; Wang, 2009). The fact that the pipeline yielded results different from the published results raised two questions; first, given two copies of the ARNTL genes (*ARNTL* and *ARNTL2*) in the ancestral vertebrate lineage, the R3 duplication event should have produced four copies of the ARNTL paralogs in teleosts, not three. We infer that the fourth zebrafish gene has been lost or modified so greatly that the pipeline could not find it by sequence similarity search. A second question is: why did the pipeline anchor zebrafish *arntl2* to a human ortholog different from the published conclusion? The pipeline properly assigned the three zebrafish *arntl* genes to a single paralogous group – with *arntl1a* and *arntl1b* being highly related to one another, followed by *arntl2*. When the automated system attempted to anchor the three zebrafish genes to their human orthologs, however, it made an erroneous assignment. In this case, the rate of change of human *ARNTL2* relative to its zebrafish ortholog was sufficiently fast that an RBH-based method does not possess enough power to detect the proper ortholog successfully. A phylogenetic analysis (Fig. 2A) confirmed the published results and led us to tentatively reject the assignment from the orthology pipeline.

We next sought to use conserved synteny to provide an independent line of evidence not based on sequence similarities. The first step was to confirm the orthology assignment of the zebrafish *arntl1* genes, and the Synteny Database provided strong syntenic support showing that *arntl1a* and *arntl1b* are co-orthologs of human *ARNTL* (see the Supplement for detail). The next step was to confirm the orthology of zebrafish *arntl2*, which is described below.

2.1.4 Orthology of Zebrafish *arntl2* Chromosome Segments

Searching for syntenic conservation to support the *ARNTL2* orthology assignment, we examined the pipeline results with an orthology dotplot of Hsa12. The dotplot revealed strong conservation along more than 80% of the length of Dre4 (Supplementary Fig. 3A), as well as weak conservation with Dre18 and Dre25. The search for a conserved syntenic cluster between the human *ARNTL2* and zebrafish *arntl2* genes led to an illuminating situation. The orthology dotplot identified both Dre18, which harbors *arntl2*, and Dre4, without an *arntl*-related gene, as the likely R3 paralogons of Hsa12 (Supplementary Fig. 3A). The Synteny Database found a conserved region on Hsa12 surrounding *ARNTL2* and orthologous to Dre4 (Fig. 3B), and also found a second region on Hsa12 that is 12Mb distant from *ARNTL2* that shows strong syntenic conservation with Dre18 (Fig. 3C). The Dre4/Hsa12 conserved region contains 38 pairs of orthologous genes while the Dre18/Hsa12 cluster contains 18 orthologous gene pairs providing strong support. So, the gene traces connect the region on Hsa12 with *ARNTL2* to a region on Dre4 without an *arntl*-related gene (Fig: 3A, orange lines), and they connect a second region on Hsa12, without *ARNTL2*, to a region on Dre18 that *does* contain *arntl2* (Fig. 3A, green lines). This result poses the question: if Dre4 and Dre18 are paralogons from the R3 duplication event, why do they show syntenic conservation with different regions of Hsa12? One hypothesis to explain these results is that there was an inversion on the ancestral chromosome in the lineage leading to humans after the lobe fin and ray fin fish lineages diverged. This inversion event would have separated the two regions we see on modern Hsa12. If we return to the paralogous cluster that linked Hsa11 with Hsa12 (Fig. 2C), we find that several paralogs within that region of Hsa11 con-

nect it to the Hsa12/Dre18 region, including *TPH1*/*TPH2*, and *CSRP3*/*CSRP2* on Hsa11 and Hsa12 respectively. Given two regions on Hsa12, one that is orthologous to Dre4 and the other orthologous to Dre18, with both of those regions on Hsa12 paralogous to Hsa11, the architecture suggests that an inversion on ancestral Hsa12 must have occurred that moved *ARNTL2* relative to other genes after the lineage leading to humans split from the lineage leading to zebrafish (see Supplementary Fig. 4 for additional evidence supporting an inversion). Furthermore, the strongly conserved region on Dre4 suggests that the fourth zebrafish ARNTL gene (which would have been called *arntl2b*) is an ohnolog gone missing (Postlethwait, 2006). The original position of *arntl2b* was likely either directly upstream of zebrafish gene *si:dkey-207j16.2* or *si:ch211-234f20.7* on Dre4 (Fig. 3B) depending on the layout of the ancestral chromosome prior to the transposition event.

Having established good syntenic support showing co-orthologous regions between zebrafish chromosomes 4 and 18 and Hsa12, the last task is to test for paralogy of Dre4 and Dre18 and we show this analysis in the Supplemental Material.

In summary, analysis using the Synteny Database suggests the following model (Fig. 3D). A single ancestral *ARNTL* gene, whose descendant still exists in amphioxus (but does not appear in the genome assembly of *Ciona intestinalis*), was duplicated in R1. Because only two copies of that gene remain in the human genome (*ARNTL* and *ARNTL2*), we infer that the second copy of the ancient *ARNTL* gene was lost prior to R2. The remaining pair of genes was duplicated again in R3 after the lineage leading to humans split from the lineage leading to teleost fish. Three of these four predicted genes remain in zebrafish today, *arntl1a*, *arntl1b*, and *arntl2*, and a fourth copy was lost, although it was probably located on Dre4 as inferred from orthologies of neighboring genes. These results are consistent with the recent work by (Wang, 2009).

2.1.5 Lessons the ARNTL study reveals about the functioning of the Synteny Database

Exercising the Synteny Database with the ARNTL gene family in this case study allowed several observations. First, the mRBH Analysis Pipeline worked well to identify the *ARNTL* paralogous gene groups in both the human and zebrafish genomes. The limits of the power of the RBH methodology, however, were illustrated by its inability to properly assign the zebrafish *arntl2* gene to its human ortholog. Second, the Synteny Database had the strength to rectify the reduced ability of the RBH methodology by identifying conserved synteny not only where reciprocal best hit analysis was strong and all of the expected R2 and R3 duplicate genes were present, but also when RBH evidence was weak and some genes had been lost. In the former case the Database showed clear syntenic conservation for *ARNTL* and its co-orthologs, *arntl1a* and *arntl1b*, and in the later case, the Database was able to buttress the weak evidence from the mRBH pipeline for orthology between the zebrafish *arntl2* gene and its human ortholog. Third, the Synteny Database was able to identify the likely location of lost ohnologs, for example the lost *arntl2b* gene in zebrafish. Fourth, the Synteny Database identified chromosome rearrangements including inversions, translocations, and transpositions, such as the inversion the Database identified on Hsa12.

3 Discussion

In this study, we introduced the Synteny Database: an automated system to identify conserved syntenic regions among sequenced genomes. A unique attribute of this system is that it was designed from the outset to cope with gene duplications, especially whole genome duplication events. Studies that specifically search for syntenic conservation in support of orthology or paralogy of a particular gene or gene family are often done by hand, and usually use a basic RBH algorithm to infer homology within a region of interest. Because the search for neighboring orthologs or paralogs is laborious and error-prone, the labor involved often limits the number of genes an investigator can reasonably study. The Synteny Database,

with its single-linkage clustering algorithm, can identify paralogy for larger groups of genes providing more targets for conserved areas. In addition, because all orthologous and paralogous relationships are pre-computed, the Synteny Database can rapidly present the results of a comprehensive search for conserved synteny. The power of this approach is evident in the ARNTL case study in which the automated system was able to identify, first, a region on Dre4 where a member of the ARNTL gene family had been lost during evolution and, second, a transposition on Hsa12 that had moved the syntenically conserved region for *ARNTL2* twelve megabases upstream on the human chromosome relative to the zebrafish paralogons.

The Synteny Database provides syntenic clusters produced using several different sliding window sizes from 50 to 200 genes. The sliding window method allows the investigator to search for conservation in broad areas using a large window size and, when areas of interest are found, to use a smaller window size to focus on strongly conserved syntenic regions. While the permutation analysis (Fig. 4) showed that all window sizes provided statistically significant results when compared to a randomized distribution, a sliding window size of 50 genes yielded the best results relative to the randomized background.

One weakness in the mRBH Analysis Pipeline, and in RBH-based algorithms in general, is fallibility when handling substantial evolutionary rate variation among a set of genes. This problem appears when only the domain that defines the gene family remains sufficiently intact to be identified by a BLAST local alignment. The rapidly evolving gene can be assigned to a paralog with the most conserved version of the family domain, rather than the gene with which it shares its pre-duplication ancestry. In such cases, the analysis of conserved syntenies automated by the Synteny Database can usually provide data that illuminates gene histories.

In this study, we focused on amphioxus, *C. intestinalis*, human, and zebrafish genomes to examine the ARNTL gene families, but the Synteny Database is also populated with other sequenced genomes, including stickleback, medaka, fugu, and mouse. The Synteny Database can analyze any genome that has been at least partially assembled into scaffolds or

a subset of chromosomes and is optimized for the investigation of individual gene families in multiple lineages. Note that the accuracy of the output depends on the accuracy of available genome assemblies. Presently, the human and mouse assemblies are of high quality, and the zebrafish assembly will soon reach this quality. Furthermore, tandem duplicated regions are often not well assembled even in the human genome, which can lead to the failure to assemble genes embedded within tandem duplications and apparent gene loss (She et al., 2004). In addition, copy number variation within a species can result in apparent gene duplication or gene loss if the genome sequenced is from a single individual polymorphic for such variants (Kidd et al., 2008; Sharp et al., 2006).

The Synteny Database presents results in an online, searchable database. In addition to the tools used to draw the gene trace images shown in the case study, the Synteny Database provides several uniquely-available tools to study the genome-wide distribution of genes, including dotplots and circle plots users can customize in a variety of ways. We recently rebuilt the databases for the mRBH Analysis Pipeline and the Synteny Database using data from Ensembl version 52, including the latest releases of the human, mouse, and zebrafish genomes, as well as version 2 of the amphioxus genome. The Synteny Database is available for public use at http://teleost.cs.uoregon.edu/synteny_db/.

4 Methods

To enumerate regions of the genome that have conserved gene content since the last whole genome duplication (relative to an unduplicated outgroup) we built two analysis tools, the mRBH Analysis Pipeline, which relies on BLAST (Altschul et al., 1997) to associate homologous genes through a modified reciprocal best hit (RBH) algorithm (Wall et al., 2003), coupled with the Synteny Database, which uses a sliding window analysis to create clusters of paralogous and orthologous genes.

4.1 mRBH Analysis Pipeline

The mRBH Analysis Pipeline begins by taking the protein sequence of every gene in the primary genome and performing a BLASTp search against all other proteins in the primary genome. In the case of multiple splice variants, the pipeline performs a search for each transcript. Following the within-primary-genome search, the pipeline conducts a BLAST search using each protein from the primary genome as query against the outgroup genome and, any sequences found are then used as queries to search back into the primary genome (a retro-BLAST).

The pipeline uses the collected BLAST results to build paralogy groups. Although reciprocal best hit relationships are often used to identify orthologous genes between species (Wall et al., 2003), the mRBH method requires modification to identify paralogous genes. Given the paralogs **A**, **B**, and **C**, only two of them can be reciprocal *best* hits. Allowing for transitivity, however, can accomodate multiple duplciation events: if genes **A** and **B** are traditional reciprocal best hits, then if gene **C**'s best hit is either **A** or **B** and **A** or **B**'s next best hit is **C**, then genes **A**, **B** and **C** should all be considered reciprocal best hits. The pipeline employs a single-linkage clustering algorithm (Van de Peer, 2004), implemented by traversing a directed graph, to achieve this goal. See the Supplemental Material for more detail.

The mRBH Analysis Pipeline uses WU-BLAST (Gish, 2003) with the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) and records only BLAST hits with an E-Value below $1x10^{-5}$. We also employed a gap opening penalty of 11 and a gap extension penalty of 1. We experimented with different substitution matrices and BLAST parameters, but, given the wide variation in rates of divergence between genes or gene families across the genome, a general approach provided the best results.

4.1.1 Noise Reduction

One of the major issues governing the size of the paralog groups the pipeline builds is how many BLAST hits to make available to the single-linkage clustering algorithm. If

a gene has one or more conserved domains or even if a gene contains weakly conserved motifs, then BLAST will pick up those regions in its search for statistically significant local alignments. Because each BLAST hit is a potential edge in the directed graph, the system must limit those edges to hits that are likely to provide information to infer real paralogy and orthology, not simply a small, well-conserved protein domain. Several heuristic approaches can eliminate noise from BLAST results (Li et al., 2005; Hahn et al., 2007); the mRBH Analysis Pipeline requires that any local alignment (or more accurately, any set of non-overlapping high-scoring pairs) produced by BLAST between two genes covers at least 50% of the length of the longer of the two genes. Prior to executing the single-linkage clustering algorithm, the pipeline checks every BLAST hit and marks those that do not meet these criteria.

4.1.2 Outgroup Anchoring

Prior to executing outgroup anchoring, the analysis pipeline constructs paralogous groups from the primary genome. The system then checks each member of each group to determine its top BLAST hit in the outgroup genome. If a group member does not have a BLAST hit in the outgroup, the pipeline drops that group member from further consideration. If members of a paralogous group have best BLAST hits to different genes in the outgroup, then the pipeline splits the group, with each subset of the original group being *anchored* to the appropriate (orthologous) outgroup gene. BLAST hits for outgroup genes are then checked to ensure that the outgroup gene hits the original gene in the primary genome (although it does not have to be the top hit). If an outgroup gene does not retro-BLAST back to a gene in the original paralogy group, then the gene from the primary genome is eliminated from the group. Finally, the system performs the outgroup anchoring analysis on all genes in the primary genome that had not been assigned to a paralogous group, i.e. singletons, to attempt to identify orthologs for all genes. The end result is a series of paralogous gene groups from the primary genome each anchored to a single gene in the outgroup.

4.2 The Synteny Database

The second analysis pipeline populates the Synteny Database by taking a set of paralogy groups along with its corresponding outgroup genes and searching for conserved syntenic areas within the primary genome and between the primary and outgroup genomes. The algorithm employs a sliding window analysis where window size is measured in numbers of contiguous genes. The pipeline places the window on the first gene on the first chromosome and moves this window until it finds a pair of genes, one on each of two chromosomes, that are members of the same paralogy group. It then places the second window at the starting location of the gene on the second chromosome and marks the start of a syntenic cluster. The software then continues to search for paralogous genes located within the space bounded by the two windows. If another gene pair is found, the windows are advanced to the starting positions of the new pair of paralogous genes and the search continues. If the search reaches the tail of either window without finding another pair of paralogous genes then the pipeline marks the cluster closed and records it. The position of the first window is then reset to the first gene that was not part of the last syntenic cluster and the search is restarted. The analysis pipeline repeats this process until all paralogous genes on the first and second chromosome have been examined. To identify inverted regions of conserved synteny, the pipeline runs the two windows in opposing directions and again records found clusters. Finally, the analysis pipeline merges clusters from the two analyses that occupy areas on the chromosome within a sliding window's length of one another. The software continues this analysis on every pair of chromosomes in the primary genome – comparing the first and third chromosomes, the first and fourth chromosomes, and so on, coming up with a genome-wide representation of paralogons. To identify conserved syntenies between species, the system performs the entire analysis again, this time comparing each chromosome of the primary genome to every chromosome of the outgroup genome. For this study we experimented with four window sizes, 25, 50, 100, and 200 genes in length.

4.3 Permutation Analysis

It is important to question whether paralogons defined by the Synteny Database are the result of a large-scale duplication event or could have originated by chance alone. To examine this question, we performed a permutation analysis to test the statistical significance of observed genomic data. For each primary genome, we took all of the paralogous genes defined by our mRBH Analysis Pipeline and randomized their locations throughout the genome. We then re-executed our clustering algorithm and recorded the results – repeating this process 100 times. For each sliding window length, we plotted with error bars the average number of clusters of a particular size that were detected after randomizing genomic data (cluster size was measured as the number of gene pairs contained within the cluster). We also plotted the actual number of clusters of a particular size found in our original data.

Figure 4 plots the results of a permutation analysis of the human genome with amphioxus as outgroup. As the length of the gene window increased, the pipeline generated larger clusters from the randomized data. With a window size of 25 genes, the largest cluster created from the randomized data contained only three gene pairs. With a window size of 200 genes, however, the simulation generated clusters from randomized data that were as large as any actual cluster produced in the original analysis. A t-test showed, however, that the mean cluster size of our actual data was statistically significantly larger than the mean cluster size of the permuted data for all four sliding window sizes (p-values of $1.7\text{x}10^{-126}$, $1.0\text{x}10^{-239}$, $2.8\text{x}10^{-207}$, and $8.6\text{x}10^{-41}$ for window sizes of 25, 50, 100, and 200 genes, respectively). We conclude that analyses should usually use the 50 or 100-gene windows for most reliable results.

4.4 Data Sources

For this study, Ensembl (Birney et al., 2004; Kasprzyk et al., 2004) provided data for the *Homo sapiens* genome, using NCBI v36 obtained from Ensembl version 41; the *Danio rerio* genome, using Zv7 from the Sanger Institute obtained from Ensembl 46; the *Gasterosteus aculeatus* genome, using BROAD version S1 obtained from Ensembl 41; the *Mus musculus*

genome, using NCBI version m36 obtained from Ensembl 41; the *Ciona intestinalis* genome, using JGI version 2 obtained from Ensembl 43. We also obtained version 1 of the *Branchiostoma floridae* genome, which was produced by and obtained from the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/).

5 Acknowledgements

J. Catchen was supported in part by an IGERT grant from NSF in Evolution, Development, and Genomics (DGE 9972830). The work was supported by grant 5R01RR020833 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

6 Figure Legends

Figure 1. Differential gene loss following whole genome duplication creates *ohnologs gone missing*. This image shows the evolutionary history of a gene *g* and neighbors undergoing a whole genome duplication event (R1), a speciation event (S) and a second WGD event (R2) that occurrs in only one of the descendant lineages, S2. If no changes to the order or composition of genes on the chromosomes occurrs over time, most algorithms would find that *g1a* and *g1b* are co-orthologous to *g1* and that *g2a* and *g2b* are co-orthologous to *g2*. In a more realistic evolutionary history, gene losses and chromosomal rearrangements follow the genome duplication event, including a loss of *g1* from the S1 lineage and *g2a* and *g2b* from the S2 lineage. Due to these losses, orthology assignment algorithms will usually infer that *g1a* and *g1b* are co-orthologous to *g2*, incorrectly assigning co-orthology where there is none. Extinct genes shown in grey.

Figure 2. Analysis of the ARNTL gene family. (A) ARNTL phylogenetic tree based on maximum likelihood showing that *Danio rerio* (Dre) *arntl1a* is paralogous to *arntl1b* and that both of these genes are co-orthologous to human (Hsa) *ARNTL*. The tree suggests that Dre *arntl2* is orthologous to Hsa *ARNTL2*. Abbreviations: chicken (Gga), amphioxus (Bfl), *Ciona intestinalis* (Cin). The tree was generated with Phyml (Guindon and Gascuel, 2003) using a maximum likelihood algorithm with a GTR model and gamma-distributed rate variation. Bootstrap values are reported on the internal nodes. (B) Human chromosome 11 (Hsa11) paralogy dotplot. Each gene on Hsa11 is represented as a gray dot with its corresponding paralogs plotted as red crosses directly above or below the Hsa11 gene but shown on the paralog's respective chromosome. *ARNTL* (Hsa11) and *ARNTL2* (Hsa12) are circled. A large region of conserved synteny inhabits the short arm of Hsa11 (the centromere is a gray circle) and Hsa12 (paralogs indicated by green crosses). Other extensive paralogons are on Hsa1 and Hsa19. (C) The *ARNTL* and *ARNTL2* paralogous syntenic cluster in humans is characterized by an inversion of six pairs of genes with *ARNTL* and *ARNTL2* serving as the pivot (50-gene sliding window).

Figure 3. Conserved syntenies for *ARNTL* genes. (A) A circle plot summarizing human and zebrafish *ARNTL* family clusters. Arcs along the circumference of the circle represent chromosomes, while arcs within the circle connect pairs of orthologs. (B) The *ARNTL2* orthologous syntenic cluster showing strong syntenic conservation between Hsa12 and Dre4. Several genes that are part of the original Hsa11/Hsa12 paralogous cluster (Fig. 2C) are labeled. A transposition moved two parts of the Dre4/Hsa12 cluster relative to one another (orange and blue lines). The fourth *ARNTL* gene in zebrafish (putative *arntl2b*) would have existed directly upstream of either *si:dkey-207j16.2* or *si:ch211-234f20.7* on Dre4 before its loss. (C) The *arntl2* orthologous syntenic cluster showing syntenic conservation between portions of Hsa12 and Dre18. The zebrafish *arntl2* gene did not appear in this cluster because the pipeline misidentified it (see text); its position in the cluster is marked with an arrow. Human orthologs in the Dre18/Hsa12 cluster fall approximately 25Mb from *ARNTL2* on Hsa12 (Fig. 2D) due to an inversion occurring after the zebrafish and human lineages diverged. (D) A gene tree showing the inferred evolutionary history of the *ARNTL* gene family in the amphioxus (Bfl), zebrafish (Dre), and human (Hsa) lineages. S represents a speciation event while $R1$, $R2$, and $R3$ represent three whole genome duplications in the lineages leading to human and zebrafish. Genes in pale, strikethrough text have been lost.

Figure 4. A permutation analysis of all syntenic clusters that the Synteny Database found in the human genome using amphioxus as outgroup. We permuted the location of paralogous group members throughout the genome and re-clustered the randomized data, repeating the randomization and cluster analysis 100 times for each window size. The mean number of clusters found for a particular cluster size are plotted with error bars. The number of clusters the Synteny Database found in actual human genome data is plotted in red crosses.

References

Allendorf, F. W. and Thorgaard, G. H., 1984. Tetraploidy and the evolution of salmonid fishes. In Turner, B., editor, *The Evolutionary genetics of fishes*, pages 1–53. Plenum Publishing, New York.

Altschmied, J., Delfgaauw, J., Wildea, B., Duschla, J., Bouneaub, L., Volffa, J.-N., and Schartl, M., 2002. Subfunctionalization of duplicate *mitf* genes associated with differential degeneration of alternative exons in fish. *Genetics*, **161**:259–267.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., 1997. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**:3389–3402.

Amores, A., Force, A., Yan, Y.-L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y.-L., *et al.*, 1998. Zebrafish *hox* clusters and vertebrate genome evolution. *Science*, **282**(5394):1711 – 1714.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., *et al.*, 2004. An overview of Ensembl. *Genome Research*, **14**(5):925–928.

Blair, J. E. and Hedges, S. B., 2005. Molecular phylogeny and divergence times of deuterostome animals. *Molecular Biology and Evolution*, **22**(11):2275–2284.

Bridgham, J. T., Brown, J. E., Rodríguez-Marí, A., Catchen, J. M., and Thornton, J. W., 2008. Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genetics*, **4**(9).

Cermakian, N., Whitmore, D., Foulkes, N. S., and Sassone-Corsi, P., 2000. Asynchronous oscillations of two zebrafish clock partners reveal differential clock control and function. *Proceedings of the National Academy of Sciences of the USA*, **97**(8):4339–4344.

Chain, F., Ilieva, D., and Evans, B., 2008. Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology*, **8**(1):43.

Conant, G. C. and Wolfe, K. H., 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet*, **9**(12):938–950.

David, L., Blum, S., Feldman, M. W., Lavi, U., and Hillel, J., 2003. Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. *Molecular Biology and Evolution*, **20**(9):1425–1434.

Dehal, P. and Boore, J. L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, **3**(10):1700–1708.

Force, A., Lynch, M., Pickett, F. B., Amores, A., lin Yan, Y., , and Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**:1531–1545.

Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A., and Köhler, N., 1999. Discovery of tetraploidy in a mammal. *Nature*, **401**:341.

Garcia-Fernàndez, J. and Holland, P. W. H., 1994. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature*, **370**:563–566.

Gekakis, N., Staknis, D., Nguyen, H. B., Davis, F. C., Wilsbacher, L. D., King, D. P., Takahashi, J. S., and Weitzcircadian, C. J., 1998. Role of the CLOCK protein in the mammalian circadian mechanism. *Science*, **280**(5369):1564–1569.

Gish, W., 2003. WU BLAST.

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5):696–704.

Hahn, M. W., Han, M. V., and Han, S.-G., 2007. Gene family evolution across 12 drosophila genomes. *PLoS Genetics*, **3**(11):e197.

Henikoff, S. and Henikoff, J., 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA*, **89**:10915–10919.

Hogenesch, J. B., Gu, Y., Moran, S. M., Shimomura, K., Radcliffe, L. A., Takahashi, J. S., and Bradfield, C. A., 2000. The basic helix-loop-helix-pas protein MOP9 is a brain-specific heterodimeric partner of circadian and hypoxia factors. *Journal of Neuroscience*, **20**.

Ikeda, M. and Nomura, M., 1997. cDNA cloning and tissue-specific expression of a novel basic helix–loop–helix/pas protein (BMAL1) and identification of alternatively spliced variants with alternative translation initiation site usage. *Biochemical and Biophysical Research Communications*, **233**(1):258–264.

Ishikawa, T., Hirayama, J., Kobayashi, Y., and Todo, T., 2002. Zebrafish CRY represses transcription mediated by CLOCK-BMAL heterodimer without inhibiting its binding to dna. *Genes to Cells*, **7**:1073–1086.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., *et al.*, 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**(7011):946–957.

Jarinova, O., Hatch, G., Poitras, L., Prudhomme, C., Grzyb, M., Aubin, J., Bérubé-Simard, F.-A., Jeannotte, L., and Ekker, M., 2008. Functional resolution of duplicated *hoxb5* genes in teleosts. *Development*, **135**:3543–3553.

Jovelin, R., He, X., Amores, A., lin Yan, Y., Shi, R., Qin, B., Roe, B., Cresko, W. A., and Postlethwait, J. H., 2007. Duplication and divergence of *fgf8* functions in teleost development and evolution. *Journal of Experimental Zoology*, **308B**(6):730 – 743.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E., *et al.*, 2004. Ensmart: A generic system for fast and flexible access to biological data. *Genome Research*, **6**(1):31.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., *et al.*, 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**:56–64.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., *et al.*, 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, **17**:545–555.

Larhammar, D. and Risinger, C., 1994. Molecular genetic aspects of tetraploidy in the common carp *Cyprinus carpio*. *Molecular Phylogenetics and Evolution*, **3**(1):59–68.

Li, W.-H., 1980. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics*, **95**(1):237–258.

Li, W.-H., Gu, Z., Cavalcanti, A. R., and Nekrutenko, A., 2005. Detection of gene duplications and block duplications in eukaryotic genomes. *Journal of Structural and Functional Genomics*, **3**:27–34.

Lynch, M. and Force, A., 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**:459–473.

Mungpakdee, S., Seo, H.-C., Angotzi, A. R., Dong, X., Akalin, A., and Chourrout, D., 2008a. Differential evolution of the 13 Atlantic salmon Hox clusters. *Molecular Biology and Evolution*, **25**(7):1333–1343.

Mungpakdee, S., Seo, H.-C., and Chourrout, D., 2008b. Spatio-temporal expression patterns of anterior Hox genes in Atlantic salmon (*Salmo salar*). *Gene Expression Patterns*, **8**:508–514.

Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S., 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Research*, **17**:1254–1265.

Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J., and Mitani, H., 2004. A medaka gene map: The trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Research*, **14**(5):820–828.

Ohno, S., 1970. *Evolution by Gene Duplication*. Springer-Verlag.

Pando, M. P. and Sassone-Corsi, P., 2002. Unraveling the mechanisms of the vertebrate circadian clock: zebrafish may light the way. *BioEssays*, **24**:419–426.

Philippe, H., Lartillot, N., and Brinkmann, H., 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Molecular Biology and Evolution*, **22**(5):1246–1253.

Postlethwait, J., Amores, A., Cresko, W., Singer, A., and Yan, Y.-L., 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends in Genetics*, **20**(10):481–490.

Postlethwait, J. H., 2006. The zebrafish genome in context: ohnologs gone missing. *Journal of Experimental Zoology (Mol Dev Evol)*, **308B**(5):563–577.

Postlethwait, J. H., Amores, A., lin Yan, Y., and Austin, C., 2002. Duplication of a portion of human chromosome 20q containing topoisomerase (Top1) and Snail genes provide evidence on genome expansion and the radiation of teleost fish. In Shimizu, N., Aoiki, T., Hirono, I., and Takashima, F., editors, *Aquatic Genomics*, pages 20–34. Springer.

Postlethwait, J. H., Yan, Y.-L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force1, A., Gong, Z., *et al.*, 1998. Vertebrate genome evolution and the zebrafish gene map. *Nature Genetics*, **18**:345 – 349.

Risinger, C. and Larhammar, D., 1993. Multiple loci for synapse protein SNAP-25 in the tetraploid goldfish. *Proceedings of the National Academy of Sciences of the USA*, **90**:10598–10602.

Satou, Y., Imai, K. S., Levine, M., Kohara, Y., Rokhsar, D., and Satoh, N., 2003. A genomewide survey of developmentally relevant genes in *Ciona intestinalis* I. Genes for bHLH transcription factors. *Development Genes and Evolution*, **213**:213–221.

Schmid, M. and Steinlein, C., 1991. Chromosome banding in Amphibia. XVI. High-resolution replication banding patterns in *Xenopus laevis*. *Chromosoma*, **101**(2):123–132.

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., Hennekam, R. C., *et al.*, 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, **28**:1038 – 1042.

She, X., Jiang, Z., Clark, R. A., Liu, G., Cheng, Z., Tuzun, E., Church, D. M., Sutton, G., Halpern, A. L., and Eichler, E. E., *et al.*, 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**:927–930.

Spring, J., 1997. Vertebrate evolution by interspecific hybridisation – are we polyploid? *Federation of European Biochemical Societies*, **400**(1):2–8.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A., and de Peer, Y. V., 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Research*, **13**:382–390.

Uyeno, T. and Smith, G. R., 1972. Tetraploid origin of the karyotype of catostomid fishes. *Science*, **175**(4022):644–646.

Van de Peer, Y., 2004. Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics*, **5**:752–763.

Van de Peer, Y., Taylor, J. S., and Meyer, A., 2003. Are all fishes ancient polyploids? *Journal of Structural and Functional Genomics*, **3**:65–73.

Wall, D. P., Fraser, H. B., and Hirsh, A. E., 2003. Detecting putative orthologs. *Bioinformatics*, **19**(13):1710–1711.

Wang, H., 2009. Comparative genomic analysis of teleost fish *bmal* genes. *Genetica*, **136**(1):149–161.

Watterson, G. A., 1983. On the time for gene silencing at duplicate loci. *Genetics*, **105**(3):745–766.

Winkler, C., Schäfer, M., Duschl, J., Schartl, M., and Volff, J.-N., 2003. Functional divergence of two zebrafish midkine growth factors following fish-specific gene duplication. *Genome Research*, **13**:1067–1081.

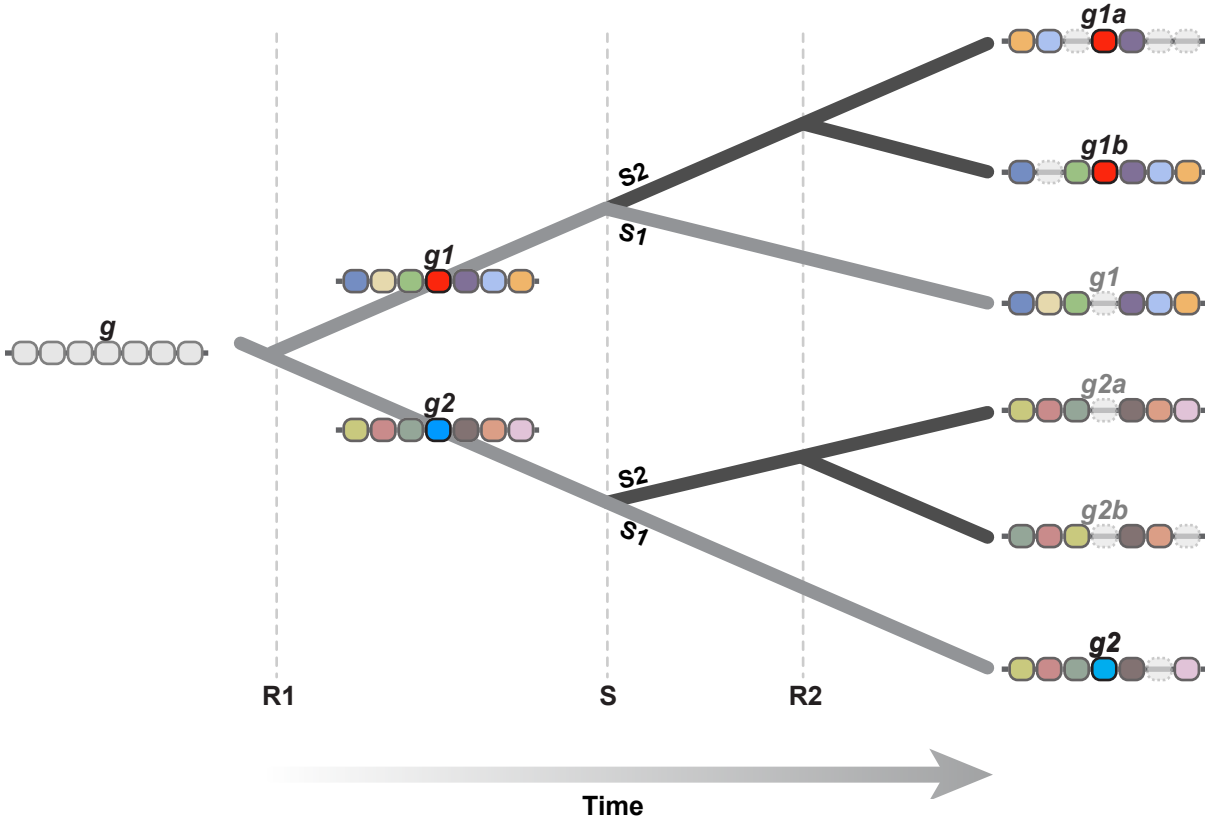
Wolfe, K., 2000. Robustness – it's not where you think it is. *Nature Genetics*, **25**:3–4.

Woods, I. G., Kelly, P. D., Chu, F., Ngo-Hazelett, P., Yan, Y.-L., Huang, H., Postlethwait, J. H., and Talbot, W. S., 2000. A comparative map of the zebrafish genome. *Genome Research*, **10**:1903–1914.

Woods, I. G., Wilson, C., Friedlander, B., Chang, P., Reyes, D. K., Nix, R., Kelly, P. D., Chu, F., Postlethwait, J. H., and Talbot, W. S., *et al.*, 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Research*, **15**:1307–1314.

Zhang, J., 2003. Parallel functional changes in the digestive rnases of ruminants and colobines by divergent amino acid substitutions. *Molecular Biology and Evolution*, **20**(8):1310–1317.

Zhang, J., Zhang, Y., and Rosenberg, H. F., 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics*, **30**:411–415.



Catchen - GENOME/2008/090480, Figure 2