

AUTOMATED METHODS TO INFER ANCIENT HOMOMOLOGY AND SYNTENY

by

JULIAN M. CATCHEN

A DISSERTATION

Presented to the Department of Computer and Information Science  
and the Graduate School of the University of Oregon  
in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy

June 2009

“Automated Methods to Infer Ancient Homology and Synteny,” a thesis prepared by Julian M. Catchen in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Computer and Information Science. This thesis has been approved and accepted by:

---

Dr. John S. Conery, Chair of the Examining Committee

---

Date

Committee in charge:           Dr. John S. Conery, Chair  
                                          Dr. John H. Postlethwait  
                                          Dr. Virginia M. Lo  
                                          Dr. Arthur M. Farley  
                                          Dr. William A. Cresko

Accepted by:

---

Dean of the Graduate School

Copyright 2009 Julian M. Catchen

An Abstract of the Thesis of  
Julian M. Catchen for the degree of Doctor of Philosophy  
in the Department of Computer and Information Science  
to be taken June 2009  
Title: AUTOMATED METHODS TO INFER ANCIENT HOMOMOLOGY  
AND SYNTENY

Approved: \_\_\_\_\_  
Dr. John S. Conery, Chair

Establishing homologous (evolutionary) relationships among a set of genes allows us to hypothesize about their histories: how are they related, how have they changed over time, and are those changes the source of novel features? Likewise, aggregating related genes into larger, structurally conserved regions of the genome allows us to infer the evolutionary history of the genome itself: how have the chromosomes changed in number, gene content, and gene order over time?

Establishing homology between genes is important for the construction of human disease models in other organisms, such as the zebrafish, by identifying and manipulating the zebrafish copies of genes involved in the human disease. To make such inferences, researchers compare the genomes of extant species. However, the dynamic nature of genomes, in gene content and chromosomal architecture, presents a major technical challenge to correctly identify homologous genes. This thesis presents a system to infer ancient homology between genes that takes into account a

major but previously overlooked source of architectural change in genomes: whole-genome duplication. Additionally, the system integrates genomic conservation of synteny (gene order on chromosomes), providing a new source of evidence in homology assignment that complements existing methods. The work applied these algorithms to several genomes to infer the evolutionary history of genes, gene families, and chromosomes in several case studies and to study several unique architectural features of post-duplication genomes, such as Ohnologs gone missing.

## CURRICULUM VITAE

NAME OF AUTHOR: Julian M. Catchen

PLACE OF BIRTH: Bronxville, New York

DATE OF BIRTH: June 3, 1978

### GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon  
Pennsylvania State University

### DEGREES AWARDED:

Doctor of Philosophy in Computer and Information Science,  
2009, University of Oregon

Master of Science in Computer and Information Science,  
2006, University of Oregon

Bachelor of Science in Computer Science,  
2000, Pennsylvania State University

### AREAS OF SPECIAL INTEREST:

Evolution of genome architecture  
Whole genome duplication  
Conserved synteny  
Genetic networks  
Bioinformatics

## PROFESSIONAL EXPERIENCE:

Graduate Research Fellow, Evolution, Development, and Genomics IGERT Program, University of Oregon, Eugene, OR, 2006 - present

Graduate Research Fellow, Postlethwait Lab, University of Oregon, Eugene, OR, 2003 - 2006

Teaching Assistant, Department of Computer and Information Science, University of Oregon, 2002 - 2003

Software Engineer, Intel Corporation, Chandler, AZ, 2000 - 2002

Software Developer, IBM Corporation, Poughkeepsie, NY, 1999

## GRANTS, AWARDS AND HONORS:

Upsilon Pi Epsilon Honor Society for the Computing Sciences, October, 2006

## PUBLICATIONS:

C. Sullivan, J. Charette, J. Catchen, C. Lage, G. Giasson, J. Postlethwait, P. Millard, and C. Kim. Zebrafish toll-like receptor-4 gene history is predictive of divergent functions. Submitted, 2009.

J. Catchen, J. Conery, and J. Postlethwait. Automated identification of conserved synteny after whole genome duplication. *Genome Research*, In Press. 2009.

R. Jovelin, Y. Yan, X. He, J. Catchen, A. Amores, H. Yokoi, C. Cañestro, J. Postlethwait. Evolution of developmental regulation in the vertebrate FgfD subfamily. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, In Press. 2009.

C. Cañestro, J. Catchen, A. Rodríguez-Marí, H. Yokoi, and J. Postlethwait. Consequences of lineage-specific gene loss on functional evolution of surviving ohnologs in vertebrate genomes: ALDH1A and retinoic acid signaling. *PLoS Genetics*, 5(5):e1000496, 2009.

H. Yokoi, Y. Yan, M. Miller, R. BreMiller, J. Catchen, E. Johnson, and J. Postlethwait. Expression profiling of zebrafish *sox9* mutants reveals that *Sox9* is required for retinal differentiation. *Developmental Biology*, 329(1):1–15, 2009.

J. Catchen, J. Conery, and J. Postlethwait. Inferring ancestral gene order. *Methods in Molecular Biology*, 452:365–383, 2008.

J. Bridgham, J. Brown, A. Rodríguez-Marí, J. Catchen, and J. Thornton. Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genetics*, 4(9):e1000191, 2008.

J. Conery, J. Catchen, and M. Lynch. Rule-based workflow management for bioinformatics. *VLDB Journal*, 14(3):318–329, 2005.



## ACKNOWLEDGMENTS

I am grateful to Dr. John Conery for providing me with the opportunity to enter the field of computational biology; for investing his time and energy in my training; for his technical skills; for Ireland. I feel truly privileged to have had the opportunity to work with Dr. John Postlethwait, who took a risk and gave me a seat at his lab bench; who provided me with half a decade of quiet mentoring and support; who taught me how to do science.

I would like to thank Cristian Cañestro, Angel Amores, Tom Titus, and all the members of the Postlethwait Lab, for their instruction, collaboration, and friendship. The IGERT program in Evolution, Development, and Genomics provided me with much of my education in biology – I became a better student the day I began attending Friday afternoon journal club. I owe thanks to Star Holmberg for shepherding me through this arduous process and for providing support when I needed it most. I would like to acknowledge those institutions that supported me financially, including the National Institutes of Health and the National Science Foundation.

Finally, I would like to thank my friends and family: Mom, Dad, Aaron, for calling me when I didn't call you; Anna, for picking me up off the ground; Kevin, for a thousand coffees; and, the Graduate Teaching Fellows Federation, AFT Local 3544, for making me feel like a citizen.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION . . . . .	1
1.1 Gene Relationships . . . . .	2
1.2 Whole-Genome Duplication . . . . .	4
1.3 Assigning Orthology . . . . .	11
1.4 Ohnologs Gone Missing . . . . .	14
1.5 Conserved Synteny . . . . .	19
1.6 Contributions and Outline . . . . .	21
II. RELATED WORK . . . . .	25
2.1 Stand-alone Tools . . . . .	26
2.2 Whole-Genome Studies of Conserved Synteny . . . . .	34
2.3 Studies Related to Ohnologs Gone Missing . . . . .	37
III. THE RBH ANALYSIS PIPELINE . . . . .	44
3.1 Methods . . . . .	45
3.2 Results . . . . .	60
3.3 Case Study: Inferring Ancestral Gene Order . . . . .	76
3.4 Summary . . . . .	84
IV. THE SYNTENY DATABASE . . . . .	86
4.1 Methods . . . . .	88
4.2 Case Study: The ARNTL Gene Family . . . . .	104
4.3 Case Study: The MSX Gene Family . . . . .	125

Chapter	Page
V. IDENTIFYING OHNOLOGS GONE MISSING . . . . .	143
5.1 Methods . . . . .	147
5.2 Results . . . . .	156
5.3 Summary . . . . .	167
VI. CONCLUSION . . . . .	169
6.1 Future Work . . . . .	171
APPENDICES . . . . .	174
A. IMPORTANT BIOLOGICAL CONCEPTS . . . . .	174
B. SINGLE LINKAGE CLUSTERING ALGORITHM . . . . .	180
C. SLIDING WINDOW ALGORITHM . . . . .	182
D. MICRO-SYNTENY ALGORITHM . . . . .	184
BIBLIOGRAPHY . . . . .	185

## LIST OF FIGURES

Figure		Page
1.1	The evolutionary history of a hypothetical gene . . . . .	3
1.2	An illustration of subfunctionalization . . . . .	6
1.3	Whole-Genome Duplications in the chordate lineages . . . . .	7
1.4	<i>Hox</i> Clusters: the signature of chordate whole-genome duplications . . .	9
1.5	The Reciprocal Best Hit Algorithm . . . . .	12
1.6	Differential gene loss following whole genome duplication creates <i>ohnologs gone missing</i> . . . . .	16
1.7	Four categories of conservation . . . . .	19
3.1	Anchoring paralogous genes to the outgroup . . . . .	45
3.2	RBH Analysis Pipeline Scheme . . . . .	47
3.3	Output of the Local Minimum Alignment algorithm . . . . .	49
3.4	Examples of the BLAST Clustering algorithm . . . . .	51
3.5	The single linkage clustering algorithm of the RBH Analysis Pipeline . .	55
3.6	Summary of RBH Analysis Pipeline Results . . . . .	61
3.7	<i>Danio rerio</i> primary genome anchored to the <i>Homo sapiens</i> outgroup genome . . . . .	66
3.8	<i>Tetraodon nigroviridis</i> primary genome anchored to the <i>Homo sapiens</i> outgroup genome . . . . .	67
3.9	<i>Gasterosteus aculeatus</i> primary genome anchored to the <i>Oryzias latipes</i> outgroup genome . . . . .	68
3.10	<i>Homo sapiens</i> primary genome anchored to the <i>Mus musculus</i> outgroup genome . . . . .	69
3.11	Orthology dotplots reveal duplication signal . . . . .	71
3.12	BLAST search results for <i>msxb</i> . . . . .	72
3.13	The RBH Analysis Pipeline web interface . . . . .	75
3.14	Search for paralogous and orthologous chromosome segments . . . . .	78
3.15	Two hypotheses for the reconstruction of ancestral chromosomes . . . . .	81
3.16	Ancestral chromosome reconstruction . . . . .	83
4.1	The PIP-based pipeline that populates the Synteny Database . . . . .	88
4.2	Sliding Window Analysis . . . . .	89
4.3	Syntenic cluster detection . . . . .	92
4.4	The <i>HOXB4</i> paralogous syntenic cluster in human . . . . .	96
4.5	Synteny Database Web Interface . . . . .	97
4.6	A permutation analysis of all syntenic clusters . . . . .	101

Figure	Page
4.7 Analysis of the ARNTL gene family . . . . .	106
4.8 Evolutionary relationships between ARNTL genes . . . . .	110
4.9 Conserved syntenies in <i>ARNTL</i> evolution . . . . .	113
4.10 Dre7 paralogy dotplot . . . . .	114
4.11 Conserved syntenies for zebrafish <i>arntl</i> paralogs . . . . .	115
4.12 Conserved syntenies for <i>ARNTL</i> genes . . . . .	118
4.13 Support for an inversion on Hsa12 . . . . .	119
4.14 The Dre18 paralogy dotplot . . . . .	120
4.15 A syntenic cluster between Dre18 and Dre7 . . . . .	120
4.16 Conserved syntenies in stickleback . . . . .	122
4.17 Analysis of the MSX gene family . . . . .	127
4.18 Conserved syntenies for <i>MSX2</i> -related genes . . . . .	130
4.19 Dre14 orthology dotplot against the human genome . . . . .	132
4.20 Conserved syntenies for <i>Msx3</i> . . . . .	135
4.21 NSG gene family tree . . . . .	139
4.22 Evolutionary history of the MSX Gene Family . . . . .	140
5.1 Reciprocal Gene Loss . . . . .	144
5.2 Micro-synteny search algorithm . . . . .	148
5.3 Reconciliation . . . . .	150
5.4 Teleost OGM Schematic . . . . .	151
5.5 The Teleost OGM Pipeline . . . . .	152
5.6 Human OGM Schematic . . . . .	154
5.7 Human OGM Pipeline . . . . .	155
5.8 Reciprocal synteny of <i>MATN3</i> . . . . .	157
5.9 Hsa2 versus <i>Danio rerio</i> dotplot . . . . .	161
5.10 Reciprocal synteny of <i>ALDH1A2</i> . . . . .	163
5.11 Ohnologs gone missing as identified by the Teleost OGM Pipeline . . . . .	165
A.1 Two illustrations of a gene . . . . .	175
A.2 An illustration of the transcription and translation process . . . . .	177

## LIST OF TABLES

Table	Page
V.1 Cases of reciprocal gene loss between human genes, teleost species <b>A</b> , and teleost species <b>B</b> , as discovered by the Teleost OGM Pipeline. . . . .	160

# CHAPTER I

## INTRODUCTION

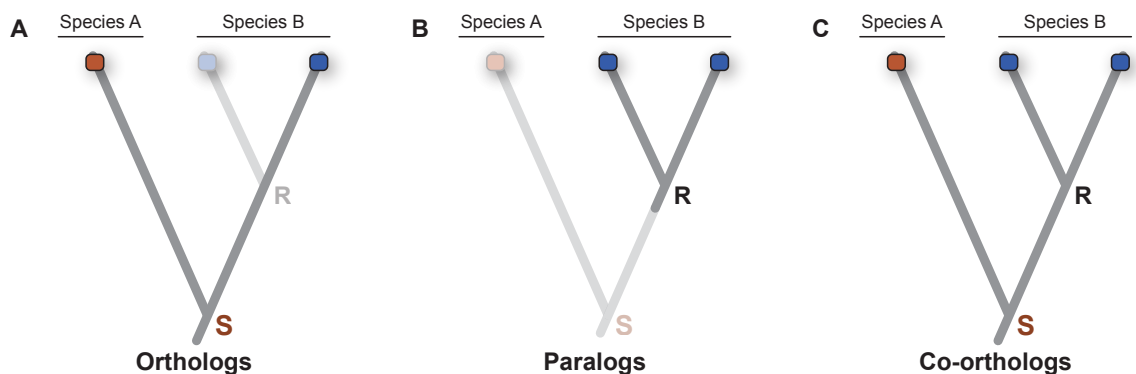
Inferring ancient homology among genes and identifying conserved syntenic regions within a genome provide us answers to two types of questions: theoretical and practical. In the former case, establishing homologous, or evolutionary, relationships among a set of genes allows us hypothesize about their histories: how are they related, how have they changed, and are those changes the source of novel features? Likewise, aggregating related genes into larger, conserved syntenic regions of the genome allows us to infer the evolutionary history of the genome itself: how have the chromosomes changed in number and makeup over time? In the latter, practical case, inferring homology between genes can be used to build human disease models in other organisms, such as the zebrafish, by identifying and manipulating genes involved in the disease. To make these inferences, researchers compare the genomes of extant species. However, the dynamic nature of genomes, in gene content and chromosomal architecture, presents a major technical challenge to correctly identify homologous genes; without confidence in gene homology the reliability of evolutionary inferences and

disease models is undermined. This work presents a system to infer ancient homology between genes that takes into account one major source of architectural change in genomes: whole-genome duplication. Additionally, the system integrates genomic conservation of synteny, providing a new source of evidence in homology assignment that complements existing methods. These algorithms are then applied to several genomes to infer the evolutionary history of genes, gene families, and chromosomes in several case studies. In the following sections, we will introduce some terminology (Section 1.1); discuss the nature of, and evidence for, whole-genome duplications (Section 1.2); describe conserved synteny (Section 1.5); and discuss some of the implications whole-genome duplication has on the evolution of gene families (Section 1.4). See Appendix A for a basic introduction to gene architecture and the processes of transcription and translation.

## 1.1 Gene Relationships

We begin by describing several common relationships among genes that are required to present the system described in this work. Having earlier defined homologous genes as those sharing an evolutionary relationship, we can be more specific and refer to **homologs** as genes that are related by a common ancestor in the past. A gene that is present in two species and was a single gene in their last common ancestor is known as an **ortholog** (Fig. 1.1A). For example, if we could access the genome of the last common ancestor of humans and mice, we would be able to take





**FIGURE 1.1:** The evolutionary history of a hypothetical gene is pictured. (A-C) A single, ancestral gene existed at the base of the tree. (A) The ancestral gene has undergone a speciation event (**S**) and a single copy of the gene exists in *Species A* (red square) and in *Species B* (blue square). The red and blue genes are orthologs. (B) We consider a case where the gene in *Species B* has been duplicated (**R**) resulting in two copies of the gene in *Species B* (blue squares). These two genes are paralogs. (C) The genes in *Species B* are co-orthologous to the gene in *Species A*.

a gene from that ancestor and find the modern descendant of it in both human and mouse. This human gene and mouse gene would be orthologous to one another. There is not usually a one-to-one correspondence between ancient, extinct genes and their contemporary representatives, however, as genes commonly duplicate over time (they are also shuffled, recombined, and destroyed by a fascinating set of permutation mechanisms). Most commonly, a single gene will be duplicated in place (a tandem duplication); sometimes a region of a chromosome is duplicated and rarely, an entire genome is duplicated. Genes that result from these duplication events are known as **paralogs** (Fig. 1.1B). Thus, orthologs are two genes that arise from a speciation event, and paralogs are two genes that arise from a gene duplication event within a

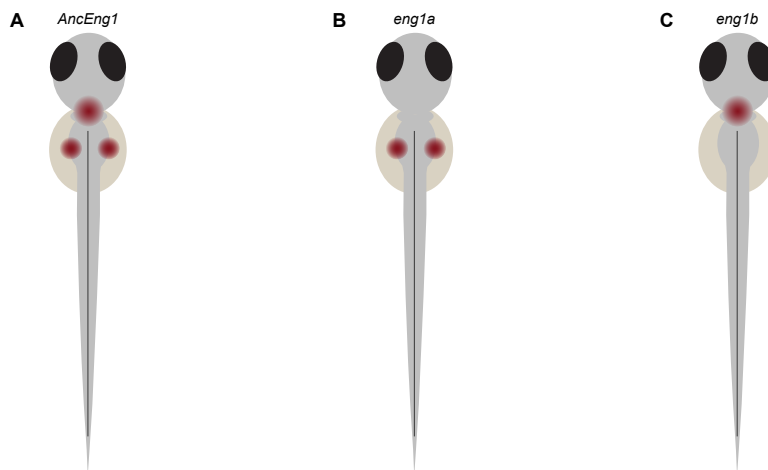
lineage. When a set of paralogs in one organism, and the ortholog of those paralogs in another organism are still related to a single gene in the last common ancestor they are known as **co-orthologs** (Fig. 1.1C). Co-orthologs, paralogs, and orthologs are all more specific cases of homologs. With some terminology in hand, we next define and present the evidence for whole-genome duplication events.

## 1.2 Whole-Genome Duplication

The complexity of the vertebrates (mammals, birds, reptiles, amphibians, and fish) is one of the great phenomena the theory of evolution seeks to explain. Whole-genome duplication (WGD) has been proposed as an initiating mechanism which can lead to complexity [76]. When a whole-genome duplication occurs (a polyploidization event), the number of chromosomes – including all of the genes and regulatory mechanisms for those genes – are doubled. With an entirely new set of genes selective pressure on them is relaxed. So, if one copy of a pair of duplicate genes experiences a mutation that negatively affects its fitness, the other copy still exists to maintain the essential function of the pair. As mentioned above, genes created by a duplication event are referred to as paralogs, however, genes resulting from a WGD are referred to as **ohnologs** [119]. The most common fate of ohnologs is pseudogenization and nonfunctionalization [64, 117, 67], however, some duplicates do obtain a selective advantage and preserve themselves. This selective advantage is described by two models: in the neofunctionalization model [76] one of the duplicate genes retains the ancestral

gene function while the second duplicate is free to develop an entirely new function. Alternatively, in the duplication-degeneration-complementation (DDC) model [33], the functions of the ancestral gene are partitioned between the two paralogs so that both copies are required to maintain the functionality of the original gene (also known as subfunctionalization). While the former process requires a rapid acquisition of new function to preserve the duplicated gene, the latter process allows both paralogs to persist and undergo incremental change.

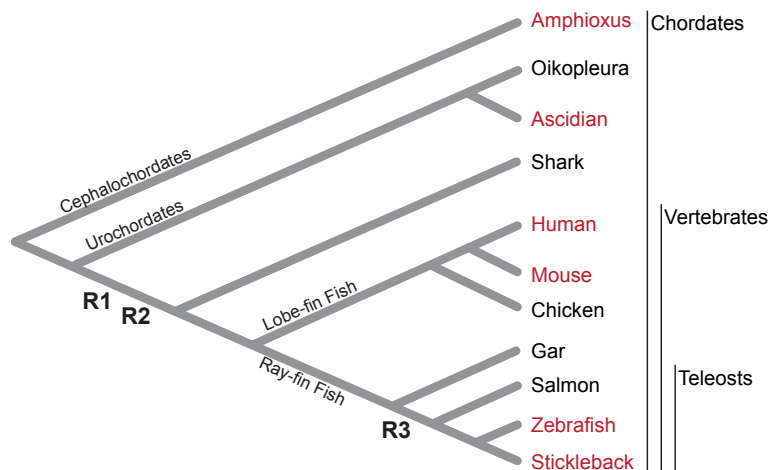
Figure 1.2 illustrates one way subfunctionalization manifests itself in practice. During the development of an organism, such as a zebrafish, genes are expressed at different times and only in certain cells – depending on what the purpose of the gene is during development. These expression patterns can be detected through laboratory experiments. Duplicating and subfunctionalizing a gene allows a finer-grained control over its expression patterns. One example of this is the *engrailed-1* genes of the zebrafish. Two *engrailed-1* genes exist in the zebrafish, *eng1a* and *eng1b*, resulting from a duplication event. If it were possible to look at the non-duplicated ancestor of *eng1a* and *eng1b*, which we call *AncEng1* in this example, (Fig. 1.2A) we would find the gene expressed at the base of the brain and in the fin buds. In the zebrafish, the expression of the gene has been split between the two copies of the gene – *eng1b* is expressed at the base of the brain and *eng1a* is expressed in the fin buds (Fig. 1.2B and C), thus both copies of the gene must be preserved by natural selection to maintain the ancestral expression pattern. (Although in this example we



**FIGURE 1.2:** An illustration of the subfunctionalization of the *engrailed-1* genes in the zebrafish. (B) and (C) show a picture of a developing zebrafish embryo as seen from above while (A) shows a hypothetical, pre-duplication ancestor of the zebrafish. (A) The ancestral, unduplicated *engrailed-1* gene (*AncEng1*) is expressed at the base of the brain and in the fin buds of the fish during development (expression is represented by red circles). (B) After subfunctionalization *eng1a* is expressed only in the fin buds while (C) *eng1b* is expressed only at the base of the brain. Both *eng1a* and *eng1b* are required to maintain the ancestral expression of the *AncEng1* gene.

show *engrailed-1* gene expression in a hypothetical zebrafish ancestor, these results were originally observed using mouse as an outgroup, allowing the ancestral zebrafish expression to be inferred [33].)

One feature common to duplicate genes resulting from a WGD is evolutionary rate asymmetry – one of the duplicates evolves at a faster rate than the other [17, 107]. Experiments in yeast indicated that rate increases occur soon after the WGD event in one of the duplicates and have been cited as evidence for widespread neofunctionalization [17]. Additionally, high-level surveys of gene function indicate that one of the two duplicates may have obtained a specialized function, or may be less well characterized in the literature [17, 107]. However, systematic functional experiments in yeast,

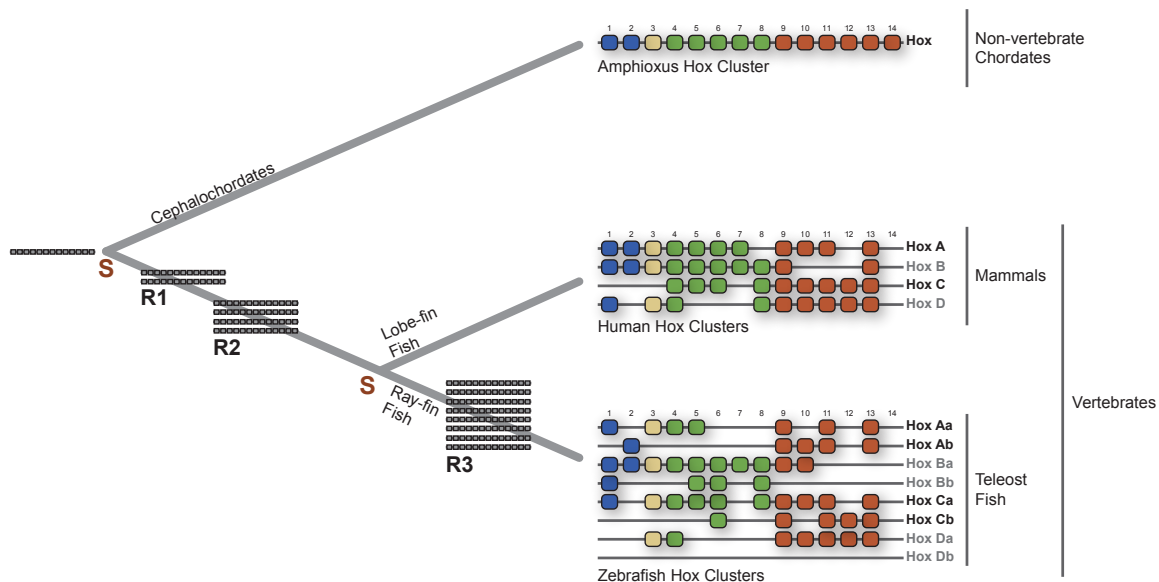


**FIGURE 1.3:** Whole-genome duplications in the chordate lineages. Two rounds of whole-genome duplication (**R1** and **R2**) likely occurred after the divergence of the cephalochordates and the urochordates. A third genome duplication likely occurred after the ray-fin and lobe-fin fish diverged, at the base of the teleost radiation (**R3**). Species names in red font represent a subset of the lineages examined in this work.

where the functions of the duplicated genes are compared directly to the orthologous gene in an unduplicated lineage consistently show evidence of subfunctionalization [110, 107], and in vertebrates, many individual cases of subfunctionalization have been characterized [33, 118, 82, 54, 52]. Subfunctionalization may not chiefly present an opportunity for genes to develop new functions, but instead may allow genes that have already accumulated multiple functions over long periods of time to separate those functions into distinct physical genes. Consistent with this idea, a recent large survey in yeast found that duplicate genes resulting from a WGD event diverge more often with respect to regulatory control, and less often in their biochemical functions [116]. Given the breadth of the evidence, it is likely that neofunctionalization and subfunctionalization are both active evolutionary processes.

Two rounds of whole-genome duplication are proposed to have occurred at the base of the vertebrate lineage – after development of neural crest cells and prior to the appearance of jawed vertebrates [36, 100, 25]. There is some controversy as to whether these events, which are referred to as **R1** and **R2**, happened in quick succession or were separated in time [86, 58]. A third duplication is thought to have occurred in the teleost fish (**R3**), after the ray-fin fish diverged from the lobe-fin fish [7] at the base of the teleost radiation (Fig. 1.3). (At over 20,000 species, the teleost radiation is responsible for the largest living group of vertebrates [82, 104].) Additional genome duplications have punctuated the evolution of other lineages, like fungi, salmonids, catostomids, goldfish, and the frog, *Xenopus laevis* [56, 1, 71, 72, 108, 90, 59, 23, 94].

Individual, or tandem gene duplication is a continuous process in evolving lineages. When a tandem duplication occurs, a new copy of the gene is deposited near the original, interrupting the original gene order. In contrast, when a whole-genome duplication occurs, all of the chromosomes are copied and immediately after the duplication event each pair of chromosomes is identical in gene content, order, and orientation. It follows that a genome that has undergone a full duplication should look significantly different in its architecture than one that has only undergone tandem duplications. Synteny, which refers to the co-localization of genes on the same chromosome, would be constantly interrupted by a series of tandem duplications, whereas it would be perfectly conserved immediately after a WGD. Although chromosome breaks and mutations continually change the underlying genome over time,



**FIGURE 1.4:** *Hox* Clusters: the signature of chordate whole-genome duplications. At the tips of the tree are the *Hox* clusters in a subset of the chordate lineages, including amphioxus, human, and zebrafish. The number of clusters each organism possesses reflects the number of genome duplications: the early-branching cephalochordates did not experience any genome duplications and have a single *Hox* cluster. The lineage leading to humans underwent two duplications (R1 and R2) and have four *Hox* clusters. The lineage leading to the teleost fish underwent three duplications (R1, R2, and R3) and have eight *Hox* clusters.

a duplication signal should be detectable. In fact, evidence of such a signal led to the proposal of the R1 and R2 whole-genome duplications in the ancestral human lineage.

The Hox clusters are a group of 39 genes grouped in four clusters in mammalian lineages with each cluster located on a different chromosome [35]. The Hox genes are responsible for patterning the basic body plan during early development. Interestingly, the expression of the Hox genes, both temporally and spatially, is correlated with their order along the chromosome. So, as development progresses along the

anterior to posterior axis of the body (head to tail), the Hox genes are expressed in order along the chromosome [35]. In fact, it is likely this requirement of time and space expression that has conserved the Hox genes in a large variety of lineages [43], from invertebrates like fruit flies, to fish, birds, and mammals. While there are at least four clusters in all vertebrate lineages, invertebrates, such as fruit flies, only have a single cluster; amphioxus, the most closely-related invertebrate to the vertebrates, has only a single cluster as well [36, 43, 6]. The four-to-one ratio of Hox clusters in mammals led to the proposal of two rounds of whole-genome duplication at the base of the vertebrate radiation and carried the implication that complexity in vertebrate body plans was rooted in the duplication of genes that controlled the early patterning of the body. The identification of seven Hox clusters in teleost fish [7, 104, 8] provided the initial evidence of a third round of duplication at the base of the teleost radiation (interestingly, the eighth Hox cluster in zebrafish has been reduced to a single microRNA [120]). Figure 1.4 illustrates these whole-genome duplications, how they would effect the number of Hox clusters and the modern membership of the Hox clusters in amphioxus, humans, and zebrafish.

While the Hox clusters were remarkable, they represented only 39 genes (in mammals), and could not make an unequivocal case for genome duplication [46]. The clusters could have been produced by a series of tandem duplications, with natural selection favoring the clustering of genes over time, or, there may have been small-scale duplications within the genome [99]. In time, additional studies in mammals

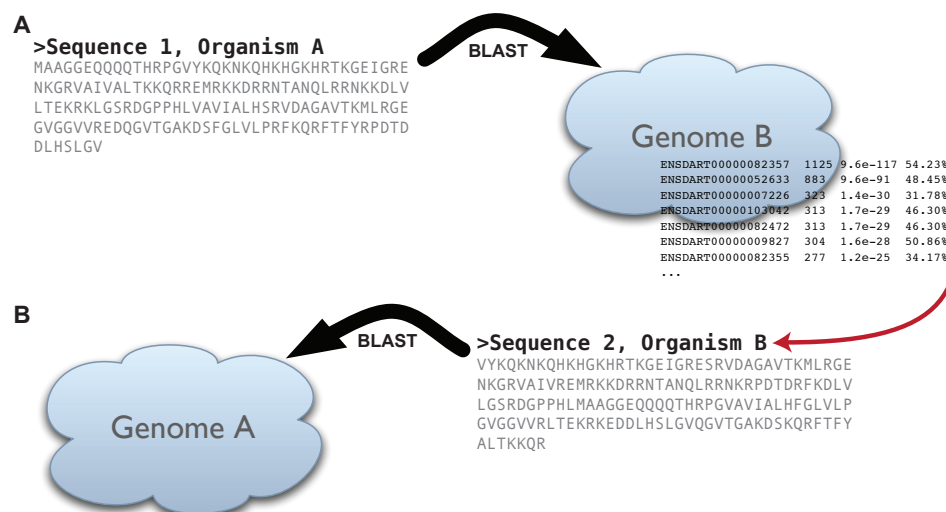


and fish provided more data in support of whole-genome duplication, including phylogenetic studies of larger numbers of gene families [79, 105, 104], and eventually whole-genome analysis [51, 74, 121, 25, 73, 86].

Whole-genome duplications are disruptive events that create branches in the evolutionary history of gene families. These events are pervasive on the tree of life and introduce noise into processes that are used to assign orthology. After introducing two methods that provided much of the evidence in support of 1R, 2R, and 3R we will examine some additional implications of whole-genome duplications.

### 1.3 Assigning Orthology

In discussing the Hox clusters in the previous section, we presented the various *Hox* genes from different species as orthologs. But, how do we actually know that one gene is related to another by ancestry? From a biological perspective, one way to characterize genes is by their expression: when during an organism's development is a gene expressed and where within the organism is the gene expressed? However, gene expression is quite susceptible to evolutionary change so we instead want to rely on a character that changes at a slower rate and hence, provides more inferential power. Amino acid sequences, which define a gene's product (its protein), are slow to change due to the degenerate nature of the genetic code [65] and are widely used. For closely related organisms, the nucleotide sequence of the gene itself is often used.



**FIGURE 1.5:** The Reciprocal Best Hit Algorithm. (A) Given the sequence of a gene in organism A (Sequence 1), we use it as a query to search the genome of organism B using BLAST. (B) We take the best hit generated by the search (Sequence 2) and now use it as a query to search the genome of organism A. If this second search returns our original query gene, we have a reciprocal best hit and may infer that these genes are orthologs.

One of the most commonly used methods to assign orthology between genes is to search a database of gene sequences (or protein sequences) for a gene whose sequence is the most similar to a query gene. Sequence similarity is determined by an alignment algorithm and a measure of statistical significance used to infer biological relatedness. The algorithm searches for the gene (a *hit*) that aligns best to the query gene; it then turns the hit into the query gene and repeats the search. If the second search turns up the original query gene, then the algorithm has found a reciprocal best hit (RBH) [114] and we infer that the pair of genes are orthologs (Fig. 1.5). In plain terms, given genes **A** and **B**, if **B** is **A**'s best hit, and if **A** is **B**'s best hit – where best hit means “has the most similar sequence” – then we consider them orthologs. The most

commonly used algorithm to perform this searching via alignment is BLAST (Basic Local Alignment Search Tool) [5].

Another important set of methods used to assign orthology is phylogenetic inference. We have already informally used phylogenetic trees to talk about gene relations and genome duplications (Figures 1.1 and 1.4). The leaves of a phylogenetic tree represent contemporaneous organisms, or characters of those organisms such as genes or proteins. From the leaves, a series of branches move backwards in time to the root of the tree – internal nodes in the tree represent ancestral organisms. Examining the tree from its root out to the leaves describes a precise ordering of speciation, from the ancient ancestral organism, to its modern-day descendants. A variant on a species tree is a gene tree, in which nodes represent a family of genes and the internal nodes represent ancestral versions of those genes. A species tree appeared in figure 1.4 while a gene tree appeared in figure 1.1. To create a phylogenetic tree we must choose a tree topology, determine the lengths of the branches of the tree, and decide what genes to place at each leaf node. Needless to say, this is a large and active area of research that is beyond the scope of this document. However, the most robust and consistent methods are based on statistical inference. Given a set of data (nucleotide or protein sequences) and a model of evolution, these methods calculate the likelihood of observing the data given the model. The evolutionary model has a set of parameters to represent factors such as the background frequency of individual nucleotides (what percentage of the genes are adenine nucleotides?), and how likely

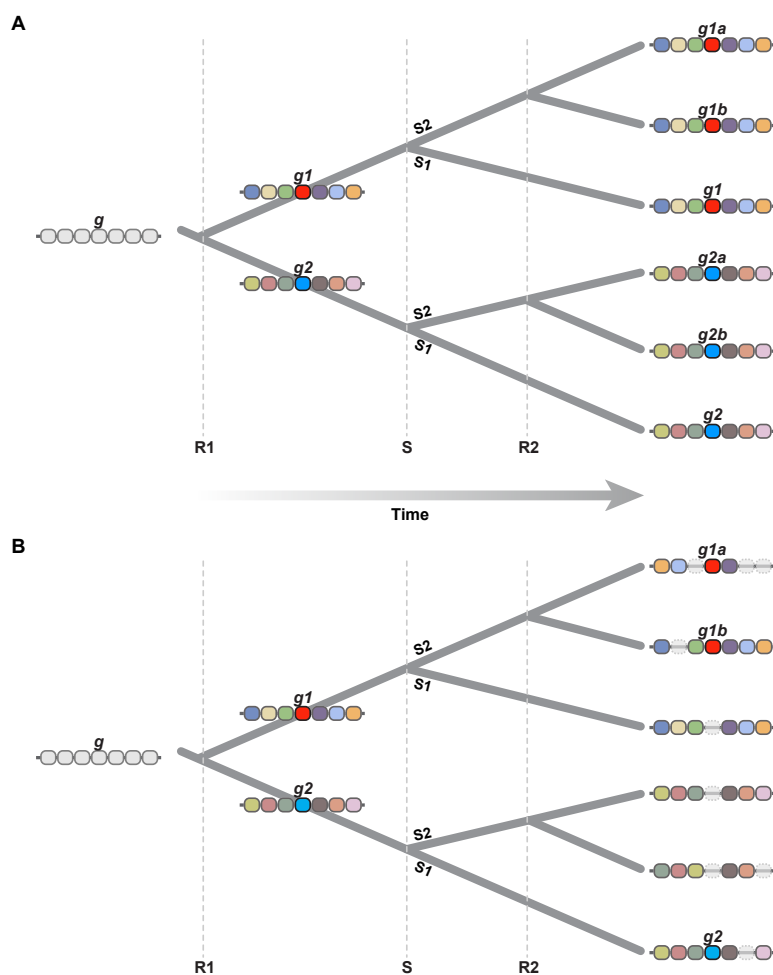
one nucleotide is to mutate into another, a tree topology (describing branching order), and a set of branch lengths, where each branch length is proportional to the number of mutations that have occurred along it. With a given set of parameter values for the evolutionary model, the algorithm can calculate the probability of the data occurring. The algorithm then tries to optimize this set of variables choosing a tree and a set of parameters that makes the data the most likely. The final tree is considered a hypothesis of descent for the species (or genes) on the tree. Two of the most commonly used algorithms are maximum likelihood (see [44, 34] for an introduction) and Bayesian inference (see [29, 122] for an introduction). Commonly used programs that implement the two algorithms to generate phylogenetic trees are Phyml [39] and MrBayes [45], respectively.

## 1.4 Ohnologs Gone Missing

As described above, one of the most common fates of genes that undergo a whole-genome duplications is pseudogenization or nonfunctionalization. When a gene is *lost*, it is no longer read and transcribed by the machinery of the cell; although the code of the gene may still be present in the DNA (a pseudogene), its instructions are no longer useful. This can happen in several ways, the most common occurs when the nucleotides marking the coding start site of the gene are mutated (like writing junk to the pointer marking the head of a linked-list). Another common way a gene is lost is when a mutation changes a structurally important amino acid making the

resulting protein ineffective (nonfunctionalization); although the gene is read and transcribed in this case, the produced protein is not functional in the organism. If a gene's function is important, negative selection will eventually purge malfunctioning copies of it from the population. However, if that gene has a duplicate that maintains the original function, there will be no selective pressure to prevent the accumulation of mutations eventually making the gene unrecognizable from background noise in the genome.

Over time, speciations occur in the post-WGD lineages parallel with the continuing loss of duplicate genes, with different duplicates lost in different lineages. This is again illustrated with the Hox genes: following the R3 duplication in the teleost fish, different species of fish lost different members of their seven Hox clusters [8]. Further, if we consider the R1/R2 duplication events and compare the Hox clusters in human and zebrafish, we again see different Hox genes retained in different lineages (the human and zebrafish Hox clusters are shown in Fig. 1.4). Recall that genes created in a WGD are known as ohnologs, and the differential loss of genes that follows a duplication event can create *ohnologs gone missing* when different ohnologs are lost in different lineages [84]. Figure 1.6 illustrates the problem ohnologs gone missing cause when trying to assign orthology between genes.



**FIGURE 1.6:** Differential gene loss following whole genome duplication creates *ohnologs gone missing*. (A) An idealized gene tree that focuses on gene  $g$  and its nearest neighbors on the chromosome. The tree shows several evolutionary events affecting  $g$  including a duplication event (**R1**), followed by a speciation event (**S**) that splits the lineage into *Species 1* and *Species 2*, and finally a second duplication in one of the lineages (**R2**). The lineages originating from ancient gene  $g$  lead to two sets of co-orthologs:  $g1$ , in *Species 1*, co-orthologous to  $g1a$  and  $g1b$  in *Species 2*, and  $g2$  co-orthologous to  $g2a$  and  $g2b$ . Neighboring genes of the same color are also co-orthologous. The illustration shows perfectly conserved synteny in the regions surrounding the descendants of  $g$ . (B) A more realistic gene tree that shows differential gene loss and rearrangements in the two organisms. Gene  $g1$  was lost from the *Species 1* lineage and genes  $g1a$  and  $g1b$  were lost from the *Species 2* lineage. Due to the loss of genes, many orthology assignment algorithms will incorrectly infer that  $g2$  is co-orthologous to  $g1a$  and  $g1b$  due to missing data. However, when considering the conserved synteny of neighboring genes it is clear that these genes are not true co-orthologs.

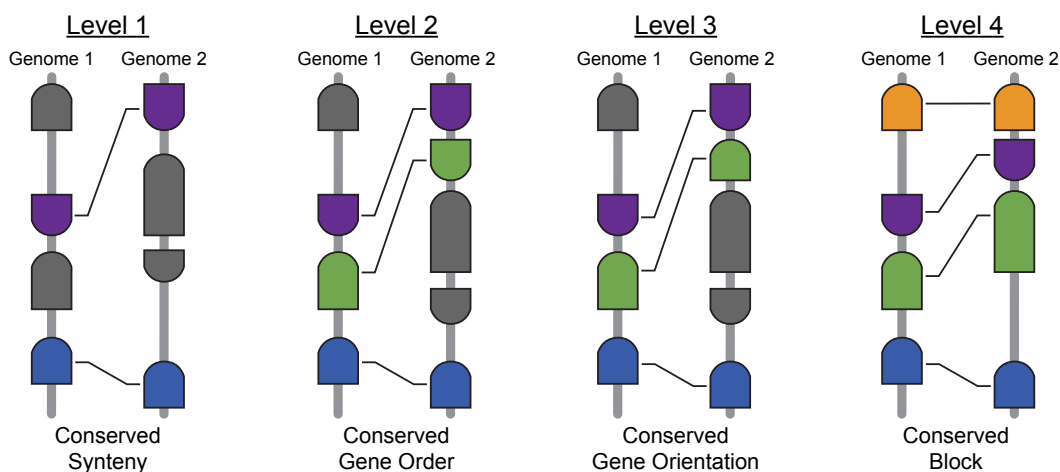
Figure 1.6A shows the evolutionary history of a gene  $g$  and its nearest chromosomal neighboring genes as it undergoes a WGD event (R1), a speciation event (S), and a second WGD event (R2) occurring in only one of the descending lineages. To identify the contemporary descendants of  $g$ , most RBH algorithms would find that genes  $g1a$  and  $g1b$  in lineage S2 were co-orthologous to  $g1$  in lineage S1. Likewise, genes  $g2a$  and  $g2b$  would be found to be co-orthologous with  $g2$ . Figure 1.6B depicts the same WGD and speciation events as A but includes differential gene loss and gene rearrangements on the chromosomes in lineages S1 and S2. Given Figure 1.6B, most RBH algorithms 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 assume that  $g1a$  and  $g1b$  were orthologs of  $g2$ .

Whole-genome duplication events provide opportunities for neofunctionalization and subfunctionalization (Section 1.2); between the time of a duplication event and the time two lineages (S1 and S2) diverge, a pair of duplicated genes ( $g1$  and  $g2$ ) can alter their expression patterns [33] or the complement of exons they possess [2], or their activities [124, 123] and such changes can alter protein-to-protein interactions or subsequent developmental or physiological functions. Therefore, subsequent reciprocal lineage-specific loss of one duplicate (say the  $g1$  copy in S1 and the  $g2$  copies in S2) can provide trees that suggest orthology where none exists. 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 S1 and *g2* in S2, but in fact, no such gene actually existed.

One interesting example of ohnologs gone missing has recently been documented in the model organism, *Arabidopsis thaliana*, a small flowering plant [11]. In *Arabidopsis*, the *HPA1* and *HPA2* paralogs are responsible for the production of histidine, an important amino acid necessary for growth and development. *HPA1* has been retained in one strain of this species (the Col strain), but has incurred a large deletion in a second strain (the Cvi strain). Likewise, *HPA2* has been retained in the Cvi strain and lost in the Col strain. If these two strains of *Arabidopsis* are bred, and the resulting offspring receives both disabled copies of *HPA1* and *HPA2* then the plant will not be viable [11]. If enough genetic incompatibilities accumulate, eventually the Col and Cvi strains of *Arabidopsis* will speciate. Finally, if we consider this process in the light of the teleost fish, with 3 whole-genome duplications and the most species of any vertebrate group, we can deduce that differential gene loss has affected a number of gene families and accounting for ohnologs gone missing is an important aspect in determining the evolutionary history of genes. In the following section we will examine how the signal of whole-genome duplications – conservation of synteny – can help us account for ohnologs gone missing.





**FIGURE 1.7:** Four increasingly stringent categories of conservation. Connected, colored genes are orthologs.

## 1.5 Conserved Synteny

Species that are evolutionarily related exhibit the property of conserved synteny: the tendency of neighboring genes to retain their relative position and ordering on the chromosomes over evolutionary time. Species exhibit this property in proportion to their evolutionary distance from one another. As we discussed in section 1.2, in a WGD event, duplicated chromosomes (homeologs) initially have their gene orders intact. 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 [57]), 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 more detail, we classify *conservation* into four increasingly stringent categories, from *conservation of synteny* to *block conservation* (Fig. 1.7). In the first category we have two or more genes from a single chromosome in one genome orthologous to two genes on the same chromosome in a second genome. The second category of conservation contains the same properties as the first, but regions also exhibit conservation of gene order. The third category adds conservation of transcription orientation (which strand of DNA the gene is read from) while the fourth category represents a conserved block – including conserved gene order, transcription orientation, and no intervening genes.

To address the problem of ohnologs gone missing, we can take advantage of conserved synteny to infer when genes are truly orthologous or paralogous. To be explicit, an RBH algorithm might falsely associate one set of co-orthologs due to ohnologs gone missing, but if we examine the neighboring genes of those co-orthologs, we will be able to find many more co-orthologous if the original co-orthologous relationship is true. In the example given in Figure 1.6B, 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 them are also paralogous and then by checking those neighboring paralogs to ensure they are orthologous to the neighbors of *g2*. The conserved syntenic region, which such genes would define, would confirm (or in this case, reject) the co-orthology of genes *g1a* and *g1b* to *g2*.

Even in the absence of missing genes, if a subset of a gene family is highly diverged, there may not be enough signal in the data for a phylogenetic algorithm to properly assign orthologs and paralogs to the correct branches of a gene tree [14]. In these cases (more of which will be presented in the following work), conserved synteny can be used to disambiguate the assignments.

## 1.6 Contributions and Outline

An important objective for inferring the evolutionary history of gene families and chromosome segments is the determination of orthology and paralogy relationships. A stepwise approach generally uses BLAST [5] 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 loss of genes, introduce noise into this process. Anomalies, such as lineage specific paralog loss, can cause anciently related homologs to appear to be orthologs, thereby confusing sequence similarity with functional homology [84]. Such errors can confound attempts to create non-human animal disease models and can make it more difficult to identify recent, species-specific evolutionary change among sister lineages.

Chapter II of this dissertation contains work related to three main areas: orthology assignment and synteny discovery algorithms, studies making use of conserved synteny at a genomic level, and studies related to the identification of lost genes. We examine

these studies with several goals, first, from the perspective of design choices: is it better to design stand-alone applications, or custom research systems? Second, what are the trade-offs in algorithm design, is it better to use heuristic algorithms that can incorporate additional biological knowledge, or to employ more formal, abstract methods? Third, when conducting a whole-genome analysis, should the data be curated in some way? Finally, we look at how the genomic distance of organisms under study affects the types of algorithms that can be employed.

Chapters III, IV, and V each present a major contribution of this work. Chapter III presents the Reciprocal Best Hit Analysis Pipeline, an automated system that can assign co-orthology to genes that have undergone a whole-genome duplication. The algorithm, which identifies duplicate genes in a primary genome relative to an outgroup genome, includes two novel components, the single-linkage clustering algorithm to group paralogs, and the gap statistic for noise-reduction. We present the results of the pipeline as applied to several vertebrate genomes, including several teleost fish, as well as humans, mouse, and the cephalochordate, amphioxus. The results of the algorithm are made available through a web interface, which we will describe as well as several visualization tools. Finally, we will apply the RBH analysis pipeline to a case study in order to determine the ancestral state of a teleost/human chromosome.

Chapter IV presents the Synteny Database, an automated system that uses the dataset produced by the RBH Analysis Pipeline to discover regions of conserved synteny within a genome. Given a primary genome that has undergone a whole-genome duplication, along with an outgroup genome that has not, the Synteny Database will find regions of conserved synteny within the primary genome, and between the primary and outgroup genomes, while allowing for small-scale changes in gene order, gene orientation, and gene loss in the conserved regions. The Synteny Database includes a searchable database of syntenic clusters and a series of programs to render those clusters and make them available via the World Wide Web, which we will describe. We then use the Synteny Database to study the evolutionary history of the ARNTL and MSX gene families in several genomes utilizing syntenic clusters to disambiguate orthology assignments in the MSX gene family that have persisted in the literature.

Last, in Chapter V, we present a pair of algorithms to investigate several genomes for ohnologs gone missing. Building on the syntenic clusters discovered by the Synteny Database, we use the Teleost OGM Pipeline to identify ohnologs that have been lost in one of several teleost genomes using the human genome as a reference. This analysis relies on two components, the micro-synteny algorithm and the reconciliation algorithm, to identify several unique architectural features in post-duplication genomes, such as reciprocal gene loss. Our second algorithm, the Human OGM

Pipeline, also utilizing the micro-synteny and reconciliation components, chains together syntenically conserved regions from multiple teleost genomes to predict the locations of ohnologs gone missing in the human genome. Both of these pipelines are built to analyze an arbitrary number of teleost genomes to produce independent lines of evidence from multiple genomes in support of an ohnolog gone missing and we present the results of examining the human genome as well as the zebrafish, stickleback, and medaka genomes. We will have some concluding remarks to make in Chapter VI.

## CHAPTER II

### RELATED WORK

In the following chapter we discuss studies related to the three main contributions of this work: the Reciprocal Best Hit Pipeline, the Synteny Database and our examination of ohnologs gone missing. While it would be convenient to group related work strictly according to the later chapters in this dissertation, many studies overlap in their goals and methods. For this reason, we group related work into three functional areas: studies that have produced general, stand-alone tools that have been released to the research community, whole-genome studies regarding the underlying architecture of a particular species, and studies meant to identify lost genes in different genomes. Grouping these studies into three areas allows us to examine design decisions in different contexts; with regard to stand-alone tools, we look at trade-offs in algorithm design including the complexity of existing algorithms, the parameters that govern them, and the use of statistical measures of significance. Whole-genome studies allow us to examine what types and how much data our system should handle, and studies that look at lost genes allow us to discuss how biological realities of the

genome restrict the data that we can examine. Finally, at the end of this chapter we discuss how these design decisions influenced our major contributions in this work.

## 2.1 Stand-alone Tools

We first examine several stand-alone tools that have been released to the research community. Since BLAST, or BLAST-like algorithms are ubiquitous in this research area, we will first take a very brief look at the algorithm that underlies it. Following that we will examine stand-alone methods to assign orthology and paralogy that take three different approaches: sequence similarity comparisons, clustering methods, and phylogenetic methods. Following that, we will look at two stand-alone methods to identify conserved synteny, the first utilizing a global algorithm and the second utilizing a local, greedy algorithm.

Methods that perform sequence similarity comparisons base their results on the Basic Local Alignment Search Tool (BLAST) [4, 3, 5, 38]. Written by Stephen Altschul and colleagues, BLAST was first released in 1990, later revised in 1997, and continues to enjoy wide use today. BLAST provides a fast, heuristic algorithm to identify potentially homologous subsequences; given a query sequence, it can search a database of sequences and find statistically significant matches by aligning the query to sequences in the database. In more detail, the BLAST algorithm has three phases, compiling a list of high-scoring words within the query sequence, searching the database for occurrences of these words, and extending the word pairs into larger



alignments. A dynamic programming algorithm is used to align the sequences in which the word pairs were located, based on scores from a substitution matrix (an empirical measure of how likely one amino acid is to be replaced by another), and the final alignment is checked against a distribution of alignment scores to determine its statistical significance, referred to as an E-value. Given a query sequence, BLAST is an effective tool able to search databases containing millions of sequences in order to identify hits – genes that are likely to be homologous to the query.

Remm, Storm, and Sonnhammer presented one of the earliest and still commonly used programs to assign paralogy and orthology between genes, INPARANOID [89, 10]. Their algorithm initially uses BLAST to identify candidate homologs between two gene datasets; given datasets  $A$  and  $B$ , sequence similarity scores are calculated between all genes in set  $A$  versus set  $A$ , all genes in  $A$  versus set  $B$ ,  $B$  versus  $A$ , and finally  $B$  versus  $B$ . Reciprocal best hits (see Section 1.3) are recorded when unambiguous and several variables are used as cut-offs to limit the genes considered in these pairwise comparisons, including a BLAST-score cutoff, and a minimum length for the alignments considered between homologs. Next, the reciprocal best hits are used as seeds to create an initial set of clusters, and INPARANOID then uses a series of heuristic rules to merge additional genes into the clusters, combine clusters and divide existing clusters. These rules use the BLAST score as a measure of distance between genes and assume that the evolutionary rate between paralogs and orthologs is equal. INPARANOID's heuristic, BLAST-based approach is fast and can examine

large datasets in a reasonable amount of time; its assumption of an equal evolutionary rate among paralogs and orthologs may be problematic (see Section 1.2) and the use of arbitrary, manual cut-off limits can cause some inconsistency in what results are considered for the clustering portion of the algorithm.

In contrast to INPARANOID, Li, Stoeckert, and Roos implemented a novel clustering method in OrthoMCL that dispenses with heuristic clustering rules [66]. Whereas INPARANOID is limited to working with two species at a time, OrthoMCL is meant to work with multiple species. OrthoMCL also uses BLAST to obtain initial pairwise homology scores for all of the genes considered and it uses reciprocal best hits to identify initial sets of paralogs and orthologs. From these initial predicted paralogs and orthologs, OrthoMCL normalizes the scores between genes from different genomes (relying on BLAST's measure of statistical similarity), and then models the homology of the genes considered as a graph, with each node representing a gene, and edges connecting nodes as BLAST hits weighted by the BLAST score. At this point, OrthoMCL diverges from INPARANOID by feeding this graph into a Markov clustering algorithm [32]. The Markov clustering method can be considered as similar to hierarchical or *k-means* clustering, however, in practice it is implemented quite differently – simulating random walks through the graph in order to identify natural clusters. The MCL algorithm represents the graph as a matrix, and simulates random walks through the the graph by iteratively performing matrix transformations. The intuition underlying the algorithm is that natural clusters in the graph,

such as evolutionarily-related genes, will be highly connected, while connections linking natural clusters will be much more sparse. OrthoMCL's matrix transformations exacerbate the natural structure of the graph until separate clusters become disconnected and these disconnected subgraphs define the final groupings of orthologs and paralogs.

OrthoMCL applies a novel clustering method to group families of genes, but no matter what system is used to cluster, arrange, or categorize orthologs and paralogs, the previous methods are ultimately limited by the amount of information available in a BLAST local alignment. Phylogenetic approaches remain the most reliable methods to determine proper orthology or paralogy, however, these methods remain hard to automate and apply to large quantities of data. Dufayard, et al. present an algorithm to assign orthology and paralogy by automating the process of reconciling species and gene trees (introduced in Section 1.3) [28]. Dufayard's algorithm starts with a set of broad gene families as determined by BLAST. They do not attempt to rigorously define the gene families, simply relying on transitive BLAST hits (if gene *A* hits gene *B*, and gene *B* hits gene *C*, then *A*, *B*, and *C* are considered a gene family) [80]. Phylogenetic trees are built for each family and a species tree must be provided describing the order of descent for the species being considered. Given these inputs, the algorithm attempts to reconcile the species tree with each gene tree: if a particular gene tree is missing representatives from certain species, the algorithm inserts nodes to represent those lost genes; if the branch lengths separating taxa on the species

tree are not proportional to the branch lengths separating individual genes on the gene tree then the algorithm infers that the genes can not be orthologs, but must be ancient paralogs, and inserts the appropriate nodes in the gene tree to represent this inference. While quite powerful, there are many cases that are not deterministically reconcilable when comparing gene and species trees, particularly since the source trees being compared are reliant on the underlying phylogenetic algorithm used to construct them. For these reasons, the system presented by Dufayard includes a graphical user interface to manually examine and curate the results of the algorithm where appropriate.

While the previous algorithms focused on assigning orthology and paralogy between genes, we now turn to algorithms that attempt to identify conserved synteny. The i-ADHoRe algorithm, first published by Simillion and colleagues [98] and recently updated [97], is one of the primary stand-alone synteny detection algorithms. i-ADHoRe uses a very broad BLAST-based approach to identify homologs in a number of genomes; given a number of genomic segments, such as chromosomal fragments, the program searches for homologous genes on the fragments that are colinear to one another. Colinearity can be visualized by placing two genomic fragments on the horizontal and vertical axis of a matrix. Cells in the matrix are marked positive if a pair of homologous genes (one on each chromosomal segment) line up. Large areas of colinearity would appear as diagonal lines through such a matrix and can be interpreted as conserved synteny. The i-ADHoRe algorithm searches each pair of genomic

segments for a pair of homologs that are a minimum distance apart and uses these genes to form an initial cluster; additional homologs are added to the cluster as long as they are less than the minimum distance from an existing member of the cluster. A linear regression is calculated to determine how well the genes in the cluster fit onto a diagonal line and the cluster may be discarded if the fit does not surpass a user-specified limit. The minimum distance is then exponentially increased and additional genes are added to the cluster if they do not negatively affect the colinearity of the existing cluster [111]. A statistical test next assess how likely the cluster is to form by chance and if the cluster is significant it is converted into an alignment profile. An initial profile is created from two genomic segments, however, once created, the profile can be used as a generalized form of the detected cluster to search for additional colinear regions. As additional regions are found they are merged into the profile (similar in some ways to progressively aligning multiple sequences) and the process continues until all genomic segments have been searched. The result are clusters of colinear genes from two or more regions of one or more genomes.

SynBlast, by Lehmann, et al. takes a hybrid, greedy approach to detecting synteny [60]. Algorithms to detect conserved synteny, such as i-ADHoRe, start with a fully annotated genome enabling them to examine a totality of the data. SynBlast, however, does not rely on this data, instead opting to perform its own translating BLAST (tBLASTn – a BLAST variant that uses a protein as a query sequence searching against a nucleotide database, with BLAST translating the protein into all

its possible nucleotide components) to detect genes within an unannotated genome. Generally, only a fraction of genes in a genome have been verified by functional laboratory experiments, the remainder are predicted by gene detection algorithms that search the genomic sequence for transcription start sites and exon/intron boundaries to create gene models. The model prediction algorithms are not perfect and sometimes multiple gene models can be predicted for a single gene, or exons can be missed, or other similar errors can occur. SynBlast starts with a user-supplied region of a genome, say a target gene and the neighboring genes within a megabase up and downstream of the target, and then does a translating BLAST to search the raw nucleotide sequence of the genome for hits. The algorithms described previously search only the set of gene models for hits; BLAST may identify several significant local alignments in a single gene, but algorithms such as i-ADHoRe simply consider the whole gene a BLAST hit (which might then be used to find reciprocal best hits). SynBlast instead takes the raw, local alignments from BLAST and attempts to order them itself into larger syntenic regions in order to avoid including any data from errant gene models. In this way, it greedily orders those raw BLAST results into a syntenic region. The results are then presented to the user to evaluate any conservation of synteny for the original target gene.

Stand-alone tools have several requirements that many specialized research systems do not, primarily the algorithms they are based on must be general enough to

accommodate a number of different types of data; in the areas of homology and synteny detection, this means genomic datasets of varying completeness and of varying evolutionary relatedness. Some algorithms can be quite successful when comparing relatively close relatives but may fail when applied to highly divergent species. Phylogenetics is widely accepted as the most reliable means to assign homology, but the models and optimization algorithms used by phylogenetic methods are very sensitive to the underlying data – the number of species included and the evolutionary distance between those species; this makes deploying phylogenetic algorithms in an automated way very difficult. There is a trade-off in designing a stand-alone algorithm between the complexity of the method and its performance against the data it processes. The algorithms based on sequence similarity presented here all rely on BLAST, and the amount of inferential power of any BLAST-based algorithm is ultimately limited by the evolutionary signal that can be inferred from the statistical significance of BLAST's local alignments. Given that OrthoMCL and INPARANOID both rely on BLAST alignments, does the performance of OrthoMCL's novel clustering method warrant its complexity over INPARANOID's simple set of heuristic clustering rules? Finally, many stand-alone algorithms want to provide their users with some type of assurance of their correctness, usually in the form of a statistical measure. i-ADHoRe can process data from multiple genomes in search of conserved synteny and will discard many found clusters based on measures of statistical significance. But, correctly implementing meaningful statistical measures is hard. SynBlast, on the other hand,

tries to analyze only the smallest subset of genomic data in a very detailed way without making any judgements about the significance of the synteny the algorithms identifies. For these reasons, many researchers instead choose to design integrated research systems to apply only to immediate problems and in the next section we will examine several such cases.

## 2.2 Whole-Genome Studies of Conserved Synteny

Several studies have examined syntenic conservation at a genomic level, often coinciding with the release of a new genome sequence, to determine the architecture of the ancestral chromosomes for that organism's lineage. In search of evidence for two rounds of genome duplication in vertebrates, Dehal and Boore performed a whole-genome analysis of four chordate genomes, including human, mouse, and fugu, with the urochordate, *Ciona intestinalis*, as outgroup [25]. The authors used a clustering method based on BLASTp scores (and verified with phylogenetic trees) to create gene families and then used a sliding window analysis to find conserved syntenic regions in the vertebrate genomes. These conserved regions were found to occur most often in groups of four, a pattern that Dehal and Boore attributed as evidence for the R1 and R2 whole-genome duplication events early in the chordate lineage.

With the release of the *Tetraodon nigroviridis* (green-spotted pufferfish) genome, Jaillon and colleagues provided support for the R3 duplication event in the teleost fish



and gave a hypothesis for a twelve chromosome ancestral vertebrate genome by calculating conserved syntenic regions between the pufferfish and human genomes [51]. To identify conserved syntenic regions Jaillon identified reciprocal best hits between several vertebrate species (using a hard cutoff on the raw BLAST score) and then manually curated the list by removing any groups of orthologs not present in all species. They then used a manual, rule-based approach to piece the conserved syntenic regions into the proposed ancient proto-chromosomes. This rule-based approach identified parts of the *Tetraodon* genome where two segments of the genome were shown to be orthologous to a single region in the human or mouse genomes – dubbed by the authors as doubly-conserved synteny (DCS). Following up on Jaillon and Dehal’s work, in [73], Nakatani et al. reconstructed the ancestral vertebrate genome using data from human, chicken and medaka genomes. Three reconstructions were completed including the amniote (birds, mammals, reptiles, dinosaurs), osteichthyan (bony vertebrates), and gnathostome (jawed vertebrates) ancestral genomes. Nakatani built groups of orthologous genes using a method similar to Dehal and Boore [25], and then built syntenic regions from those orthologs using the DCS method introduced by Jaillon. From there the actual reconstructions were performed in two steps. First, a statistical method was used to determine which syntenic regions within a genome were paralogous (testing whether the orthologs within the conserved regions occurred due to a duplication or simply due to chance). Second, syntenic regions were drawn

as nodes in an undirected graph and nodes were connected based on paralogous relationships. These connected portions of the graph were considered proto-chromosomes in the ancestral genome being reconstructed. Interestingly, Nakatani found that the osteichthyan ancestor had approximately 40 chromosomes contradicting the earlier study by Jaillon [51] (among others) who predicted 12 ancestral chromosomes.

Kikura, et al. examined syntenic conservation between zebrafish and human genomes in [57] proposing that the conservation of syntenic regions are driven by highly conserved non-coding elements (HCNEs) belonging to duplicated genes. These HCNEs, regulatory regions located far upstream of the target gene, were preserved by natural selection to maintain the function of the gene, along with any unrelated genes located within the area between an HCNE and its target gene. The authors determined conservation of synteny by aligning raw genomic sequence from the zebrafish and human genomes together and then piecing together the small, genomically conserved regions that could be identified into syntenic blocks.

These studies provided excellent insights into the architecture of the ancestral genome and in each case the authors built custom research systems to study the conservation of synteny. One of the major advantages to a genome-wide study is that the researcher only needs to be able to detect enough of a signal in their data to provide evidence for or against their hypothesis. Examining multiple genomes increases the total pool of available data and allows for algorithmic simplifications by doing such things as eliminating noisy data. These simplifications become problematic, however,

if one wants to build an automated system to provide similar information about conserved synteny, but apply it on the level of individual gene families. In this case, one cannot hand-curate the data [51, 73], or discard portions of the genome that did not fit into the analysis [25]. Additionally, you must make the data available in a form that allows it to be studied on the level of gene families, not simply make genome-wide measures of it [51, 25, 73]. In the next section, we will continue to discuss whole-genome and multi-genome studies as well as some studies that focus on individual gene families; this work goes beyond conserved synteny and attempts to identify ohnologs gone missing.

## 2.3 Studies Related to Ohnologs Gone Missing

We will group the literature that focuses on gene loss in general, and in some cases on identifying ohnologs gone missing into three categories: studies examining and cataloging pseudogenes in mammalian species, studies that identify specific cases of reciprocal gene loss in species that have experienced whole-genome duplications, such as in yeast and teleost fish, and studies that examine individual gene families and identify specific cases of ohnologs gone missing.

The identification of pseudogenes in human and other mammalian genomes is where much of the work in the study of lost genes has centered. These efforts generally focus on identifying recently duplicated genes that have been pseudogenized; as opposed to ohnologs lost from the R1, R2, or R3 WGD events, the remnants of

recently pseudogenized genes can still be detected in the raw genome sequence. In order to better understand algorithms that detect ohnologs gone missing, we will briefly describe two such studies. The general approach, as used by Suyama and colleagues in [103], is to use a BLAST-like tool to search the raw genome for sequences that are similar to existing genes. Gene fragments found in the search are interpreted as recently duplicated genes that experienced disabling mutations. Conservation of synteny was employed at a cursory level to distinguish functional genes from true pseudogenes (recent pseudogenes are frequently the product of retrotransposition, which places the duplicate far away from the original copy). Using this technique, the authors were able to identify almost 10,000 such pseudogenes in the human and mouse genomes. In a novel variation of this technique, Zhu et al. sought to identify the loss of well-established genes in the human genome – genes that had been present in the last common ancestor of the human and rodent lineages approximately 75 million years ago [125]. This work extends the earlier gene loss studies by searching for lost genes that had much more ancient origins (although the study only showed that the genes were in existence at a time still much more recent than the major vertebrate genome duplications). Their method took advantage of conservation of synteny on a gene-by-gene basis; given an existing gene in mouse, with an ortholog in dog (the outgroup), but without an ortholog in human, they authors attempted to identify the remnants of the gene by searching the raw human genome for remnants of components of the gene such as exons and 5' and 3' untranslated regions. When they could

identify the remnants of these components, they relied on the conservation of synteny of the exons and untranslated regions to determine if they had found the correct pseudogene. Using this method, they were able to identify 26 genes that had been lost in the human lineage but were still present in the mouse and dog genomes, indicating that these 26 genes had been present in the last common ancestor of human, mouse, and dog.

Scannell and colleagues compared the syntenic conservation in six species of yeast, three of which had undergone a WGD and three of which had not, in search of reciprocal gene loss using their very pretty tool, the Yeast Gene Order Browser [93]. They assigned orthology by a mix of reciprocal best hit BLAST analysis and through manual curation of the datasets. Syntenic conservation between any two of the six genomes was determined by aligning the homologs from all six species and then checking that for any homolog there was at least one more homolog on the same chromosome no more than 20 genes apart and with no more than six intervening homologs that pair to other yeast species [16]. The authors were able to identify 14 different classes of gene loss using this method, the most common of which occurred in 72% of cases with the same gene lost in all species that experienced a WGD; the remainder of the cases present a number of patterns of differential gene loss among paralogs in the duplicated yeast species.

Working in the teleost fish, Sémon and Wolfe compared syntenically conserved regions in the zebrafish and pufferfish using the human genome as an outgroup [95]

in search of differential gene loss. Given a particular human gene, in principle there should be two zebrafish orthologs and two pufferfish orthologs due to the R3 WGD. In the case that each teleost fish lost at least one of the orthologs, the authors wanted to determine if both fish lost the same copy (orthologs) or different copies (paralogs) – the latter case demonstrating reciprocal gene loss. For every human gene they examined 40 genes upstream and downstream and ranked which pufferfish and zebrafish chromosomes contained the most orthologs from this region. Taking the two pufferfish and two zebrafish chromosomes with the most orthologs to the human region, they compared the four fish chromosomes to determine their paralogy. Having determined which chromosomes to compare, if 30% of the human orthologs from the defined region were present within the fish syntenic regions, they considered the region to be syntenically conserved. Using this method the authors determined that approximately 7% of all loci in the zebrafish and pufferfish had experienced differential gene loss.

Beyond whole-genome studies, the characterization of individual gene families often includes a study of orthologs gone missing. The general approach of these studies is to identify all the members of a gene family in a number of different lineages and then to assign orthology and paralogy among the family members in the different lineages in order to infer the evolutionary history of the gene family. This is typically done in two parts; first, a phylogenetic tree is built to determine the branching order of the genes. Many times, due to the divergence of the sequences, a lack of species

to sample, or a missing outgroup, the trees will not be definitive. For this reason, conservation of synteny is engaged in order to provide supporting evidence and to infer where ohnologs gone missing would have previously been present in the genome. In one such example, Braasch, Volf, and Schartl examined the evolutionary history of the endothelin system which is involved in the regulation of neural crest cells during development [14]. There are three known endothelin genes in tetrapod lineages, such as human (*Edn1*, *Edn2*, *Edn3*), and the authors identified five to six copies of these genes in the different teleosts. Using manual methods to determine syntenic conservation, the authors demonstrated that one of the teleost endothelin genes, *Edn4*, had become an ohnolog gone missing in the tetrapod lineages. In a separate study, an ohnolog gone missing of the *Msx* gene family was also identified in the human lineage (*Mxs3*) despite being present in mouse and the teleost fish [83], and an ohnolog gone missing for the *ALDH1A* gene family was identified in medaka, while still present in other teleost fish and human lineages [18].

One of the interesting results of the studies of pseudogenes in species such as human and mouse is the sheer number of genes resulting from non-whole-genome duplications. However, algorithms that search for the remnants of genes in the genome can only be applied to recently duplicated genes due to the effects of unrestrained mutations in disabled genes. The only way to study recently duplicated genes is to compare closely related species. To study species more distantly related, such as Sémon and Wolfe's study of reciprocal gene loss in humans and teleost fish, you must

infer gene loss by utilizing the conserved synteny of existing genes. In studies of ohnologs gone missing in individual gene families, determining reciprocal best hits by manually searching with BLAST is a very error-prone process; without a uniformly applied method it is easy to misinterpret RBH relationships and it is difficult to uncover many of the more complicated conserved syntenies that exist in distantly related species.

When considering the body of work presented in this chapter, it highlights several trade-offs in algorithm design. We would like algorithms that can be applied to genomes at a variety of evolutionary distances and as such would like to avoid the arbitrary parameters present in many heuristic approaches. We would also like to incorporate knowledge of whole-genome duplications into our implementation, however, preventing a purely abstract approach. We would like to provide data from our analyses at a fine granularity, allowing inferences to be drawn not only about genes that have some type of genome duplication signature (orthology or conserved synteny), but also those that do not. We therefore wish to design algorithms that work well with entire genomes, not hand-curated subsets. Finally, many of the whole-genome analyses make high-level inferences about the evolution of the genome itself, but say little about individual gene families. Likewise, many studies of the evolution of individual gene families would benefit greatly from a set of automated, consistent algorithms that could help make orthology assignments. Much of the work on gene loss has focused on recent gene losses, with very little focus on trying to identify

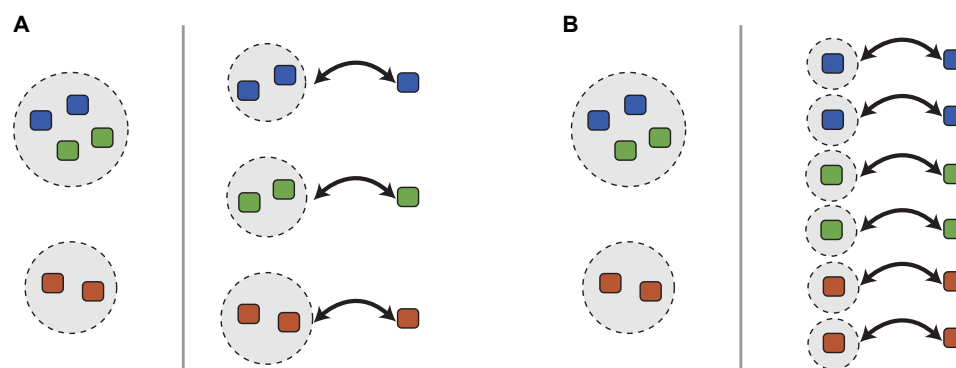


much more ancient gene losses. In the remainder of this work we will attempt to apply the insights of these earlier studies in our own algorithms to assign orthology and paralogy, determine conserved synteny and to discover ohnologs gone missing. We start, first, with our Reciprocal Best Hit Analysis Pipeline.

# CHAPTER III

## THE RBH ANALYSIS PIPELINE

A prerequisite to examine conserved synteny or search for orthologs gone missing is an accurate assignment of orthology between genes. As we discussed previously, a Reciprocal Best Hit (RBH) algorithm can assign orthology between genes and can be applied to large datasets in an automated way. It is an appropriate tool to study and compare the genomes of multiple species in order to make inferences about their ancestral architecture. In this chapter, we describe the RBH Analysis Pipeline, a high-throughput ortholog assignment algorithm that accounts for the effects of the R1, R2, and R3 whole-genome duplications in the vertebrates and features an effective paralog clustering method and a novel noise reduction algorithm. After describing the method, we present the application of the method to several teleost and tetrapod genomes and use the resulting data to infer the organization of an ancestral teleost chromosome by examining zebrafish and pufferfish co-orthologs of human genes.



**FIGURE 3.1:** Anchoring paralogous genes to the outgroup. In each illustration, the left image shows two paralogous groups formed from the primary genome. The right image shows those genes anchored to the outgroup in two cases: (A) the primary genome has experienced a WGD that the outgroup genome has not (creating paralogous groups of size 2), and (B) the primary and outgroup genomes have experienced the same number of WGD (creating a one-to-one correspondence between primary and outgroup genes).

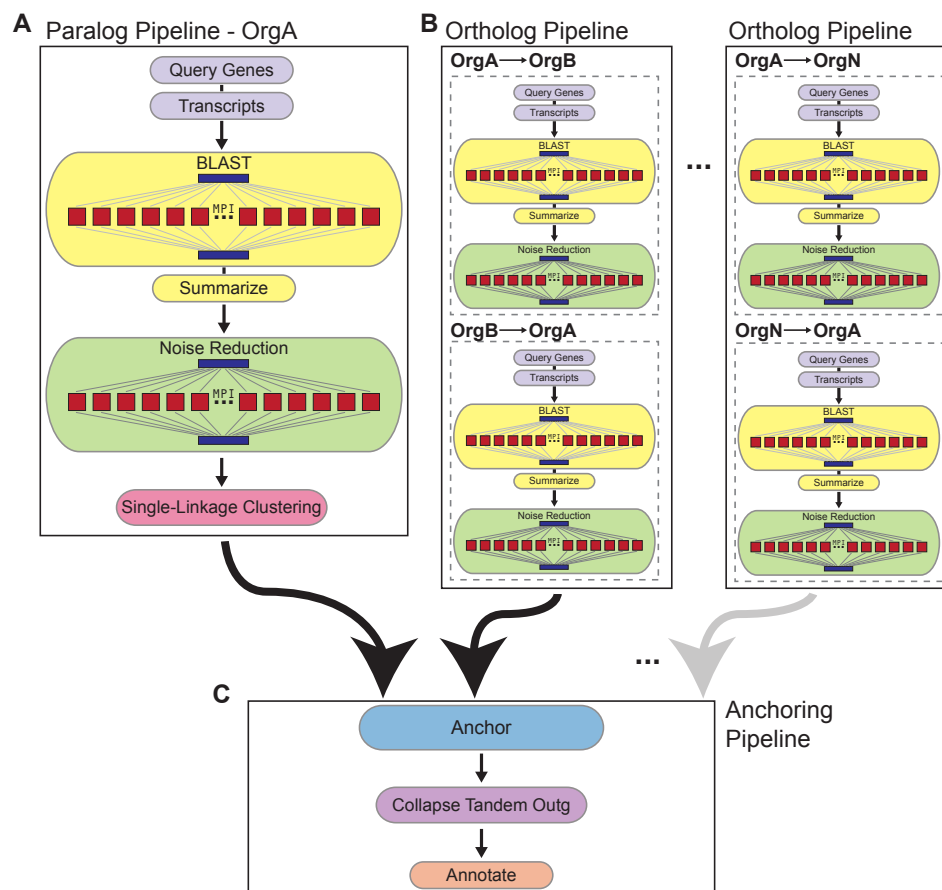
### 3.1 Methods

The RBH Analysis Pipeline identifies paralogous gene groups in a primary genome and then *anchors* those gene groups to an ortholog in an outgroup genome using a BLAST-based approach. The result of this anchoring is a mapping between genes in the primary genome and their orthologs in the outgroup genome. Paralogous groups are created by the pipeline relative to the last whole genome duplication present in the primary genome but absent in the outgroup genome using a single linkage clustering algorithm [109]. For example, if the primary genome has experienced a duplication since it diverged from the outgroup genome, as in the teleost fish (R3) compared to the unduplicated outgroup, humans, then the pipeline will produce gene groups of size two – each group corresponding to its single ortholog in the outgroup (Fig. 3.1A).

If, on the other hand, both genomes have experienced the same duplications in their history, such as human and mouse (R1 and R2), then the pipeline reverts to a simple ortholog pipeline with a one-to-one correspondence between genes in the primary and outgroup genomes (Fig. 3.1B). In practice, the number of genes per group is heavily influenced by recent tandem gene duplication, gene loss, and sequence divergence.

### 3.1.1 Pipeline Interface Program

Each of the systems described in this work are built on PIP (Pipeline Interface Program) [22], a generic framework that allows us to create many different pipelines by combining arbitrary analysis stages in different orders. Data are fed into each analysis stage from a relational database, the analysis stage then transforms the data in some way, and the results are stored back in the database. In this way stages are chained together, with the data flowing between them transformed at each step. If the analysis being performed is embarrassingly parallel, then PIP can execute the stage in parallel using the MPI libraries [70]. Dependencies are defined for each stage and PIP monitors the database tables in a way analogous to how Make [101] monitors object and source files. When Make notices that a particular object file has become older than the source it was built from, it recompiles it. Similarly, when PIP notices that data has been updated in a database table that the current analysis stage depends on, PIP re-executes the dependent stage of the pipeline.



**FIGURE 3.2:** RBH Analysis Pipeline Scheme. The RBH analysis pipeline is composed of three PIP-implemented pipelines; the Paralog Pipeline (A) is combined with an arbitrary number of Ortholog Pipelines (B) by the Anchor Pipeline (C).

The RBH Analysis Pipeline is composed of three PIP-based pipelines that use a series of modular stages to create paralogous groups in the primary genome; to compare those groups to an arbitrary number of outgroup genomes; and finally, to anchor genes from the primary genomes to each of the outgroup genomes (Fig. 3.2).

### 3.1.2 The Paralog and Ortholog Pipelines

The paralog pipeline (Fig. 3.2A) begins by loading all of the gene names for the primary genome, which we refer to as the *query genes*, and then loads the protein sequence for each of those genes. In the case a gene has multiple splice variants, a transcript of each variant is loaded. Next, the pipeline performs a BLAST search, using each transcript as a query, against all other proteins in the primary genome. The BLAST stage is parallelized to decrease execution time – as the zebrafish genome contains approximately 35,000 transcripts and the human genome contains about 56,000, the BLAST search stage is an intensive operation. Following the within-primary-genome search, the pipeline summarizes the search results, removing self-hits from the list of BLAST results (in a within-genome search, the best search result will always be the query gene’s own sequence) and combining the reported results where BLAST found significant local alignments in more than one area of the same gene (multiple high-scoring pairs, or HSPs). At this point, for each query gene in the primary genome, there exists a list of BLAST hits of possible paralogs. Depending on the architecture of the gene, it may have a lot of hits or it may have none at all; many genes share common domains or even common sequence motifs despite a lack of orthology. Because BLAST’s local alignment algorithm may identify these regions, it is necessary to differentiate hits that may indicate orthology or paralogy from those that the pipeline considers equivalent to background noise (such as a simple, common

HIT	SCORE	E-VALUE	LENGTH	PERCENT IDENT	HSPs	MIN ALN COVERAGE	STATUS
ENSDART00000082357	1125	9.6e-117	437	54.23%	1	93.76%	keep
ENSDART00000052633	883	9.6e-91	419	48.45%	1	89.69%	keep
ENSDART00000007226	323	1.4e-30	343	31.78%	1	74.82%	keep
ENSDART00000103042	313	1.7e-29	162	46.30%	1	35.49%	drop
ENSDART00000082472	313	1.7e-29	162	46.30%	1	35.49%	drop
ENSDART00000009827	304	1.6e-28	116	50.86%	1	26.62%	drop
ENSDART00000082355	277	1.2e-25	240	34.17%	1	49.40%	drop
ENSDART00000076161	266	1.9e-24	271	33.58%	1	51.32%	keep
ENSDART00000025449	266	1.9e-24	263	33.08%	1	52.76%	keep
ENSDART00000091286	254	3.7e-23	217	35.48%	1	44.36%	drop
ENSDART00000103132	254	3.7e-23	217	35.48%	1	44.36%	drop
ENSDART00000014696	254	3.7e-23	153	39.87%	1	30.22%	drop
ENSDART00000012470	253	6.5e-28	94	59.57%	2	46.04%	drop
ENSDART00000003506	251	7.7e-23	141	43.26%	1	29.74%	drop
ENSDART00000080466	249	5.3e-26	94	58.51%	1	20.14%	drop

**FIGURE 3.3:** Output of the Local Minimum Alignment algorithm for zebrafish *hoxb3a*.

domain or motif). This operation is performed in the noise reduction stage and we will describe the algorithm used next.

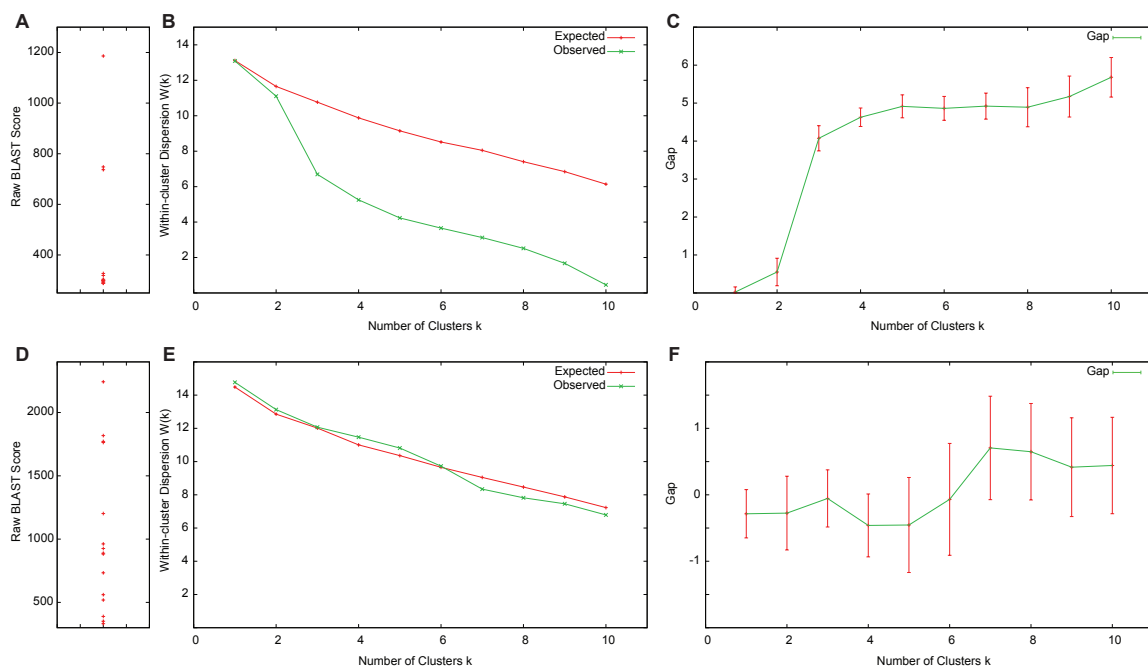
### Noise Reduction

Several heuristic approaches have been applied to eliminate noise from BLAST results. Two of the most common approaches involve measuring the size of the alignment between a query gene and the search hit. In both cases, the idea is to avoid short alignments that may only indicate a shared protein domain or sequence motif. The first heuristic, which we will call a global alignment cutoff, is based on aligning the full length of both genes; for any two genes that BLAST found a local alignment, a full, global alignment is performed and then checked to ensure that the alignment includes at least 80% of the length of the longer sequence with at least 30% sequence identity [63]. An alternative heuristic, used by INPARANOID [89] as well as in an

earlier version of this work [20], which we refer to as a local alignment cutoff, simply looks at the local BLAST alignment and checks that the alignment covers at least 50% of the length of the longer gene. A third alternative, uses an order of magnitude cutoff [40], considering BLAST hits noise if their score is an order of magnitude smaller than the best BLAST score.

There are two major problems with these approaches. First, the cutoffs are arbitrary and not based on any objective criteria. Alignment length and sequence identity will be higher or lower in proportion to the evolutionary distance between the genomes being compared. Not only will these criteria change with respect to the overall evolutionary distance of the genomes, but they will vary with respect to individual gene families – some gene families will be highly conserved and some less so, and therefore any single cutoff value is likely to be inaccurate. The second major problem is that these methods tend to create inconsistent results. Figure 3.3 shows the list of BLAST hits for the zebrafish query gene, *hoxb3a* and the affect of applying a local alignment cutoff of 50% to those results. The algorithm determines that the first three BLAST hits meet the stated criteria, the next four BLAST hits fail the criteria, the following two hits meet the criteria, and the remainder do not. As we will make clear in the following section, any RBH-based algorithm relies on a precise ordering of BLAST hits according to statistical significance. An RBH algorithm that includes the first three hits as well as the eighth and ninth hits (as would be the case with *hoxb3a*) would violate this requirement and is hence, inconsistent. Instead, an





**FIGURE 3.4:** Examples of the BLAST Clustering algorithm. (A-C) Zebrafish *sox9a*. The algorithm is able to determine that three clusters is optimal after calculating the gap statistic; data from the lowest scoring cluster is discarded. (D-F) Human *ALDH1A2*. The algorithm is unable to determine the optimal number of clusters due to the even range of the BLAST scores and therefore no data is discarded.

algorithm engaged in noise reduction should identify a single value; any results above or equal to this value should be considered significant, and any results below this value should be considered insignificant.

We created a novel noise reduction algorithm that employs a standard hierarchical clustering algorithm to separate insignificant BLAST hits from the BLAST search results for each query gene. In order to avoid the problem of arbitrary cutoffs and to handle the comparison of genomes at different evolutionary distances, we decide how to cluster the BLAST results by permuting the search results to create a null distribution and then apply the gap statistic [106] to choose the optimum number

of clusters to employ. Once the data is properly clustered, we can discard the least significant cluster of search results as background noise.

We will present the algorithm in more detail using the example of zebrafish *sox9a* (Fig. 3.4A-C). Given *sox9a*, we have a set of BLAST search results and a raw BLAST score associated with each one. When we plot those scores (Fig. 3.4A) we see that the search results are naturally clustered into three groups. The highest ranked gene found in the search is *sox9b*, the R3 paralog of *sox9a*. The next cluster is formed by two ancient paralogs of *sox9a*, followed by a third cluster composed of a number of hits made up of small, local alignments to more distantly related genes. Although the clusters are naturally visible in this example, we require a method that can determine the proper number of clusters to use to reliably exclude insignificant BLAST hits.

Tibshirani, Walther, and Hastie provide just such a method with the gap statistic [106]. First, given  $n$  data points, we cluster the data using a hierarchical clustering method 10 separate times; during the first iteration we place the data into a single cluster ( $k = 1$ ), during the second iteration we place the data into two clusters ( $k = 2$ ), and so on until  $k = 10$ . (We could let  $k$  range much higher than 10, however, in practice, we gain little additional precision by using more than 10 groups to cluster BLAST results.) At each iteration we measure the fit of the clusters to the data (the within-cluster dispersion) in the following way. If we have placed the data into  $k$  clusters, for each cluster  $r$  we measure the pairwise distance of all the points in that cluster:

$$D_r = \sum_{i,i' \in C_r} d_{ii'}$$

where  $d$  is the squared Euclidean distance. We then sum the average fit of each of our  $k$  clusters ( $n$  is the number of data points in a particular cluster  $r$ ) giving the within-cluster dispersion:

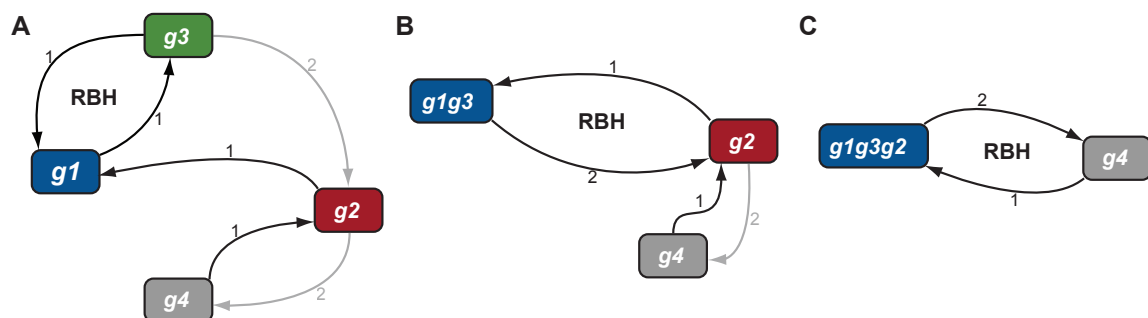
$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r.$$

Applying this method to *sox9a* we get the “observed” curve in Fig. 3.4B. This curve represents the *fit* or *tightness* of our clusters for different values of  $k$ . Now, we want to know the number of clusters to use to best fit our data. As we increase  $k$  at each iteration, we expect our measure of fit ( $W_k$ ) to improve (obviously the best fit would occur when each data point is in its own cluster). To determine the optimal number of clusters to use we will compare our observed  $W_k$  values to those calculated from a randomly distributed set of data points. So, given our BLAST scores for *sox9a*, we generate the same number of data points over the same range by randomly drawing them from a uniform distribution. We then cluster them 10 times and calculate  $W_k$  just as we did before repeating the simulation  $B = 10$  times (Fig. 3.4B, “expected” curve). The gap is defined as the difference between our observed and expected curves:

$$Gap(k) = (1/B) \sum_b \log(W_{kb}^*) - \log(W_k).$$

Finally, after calculating the standard deviation of  $W_k$  over our 10 simulations we choose the smallest number of clusters  $k$  such that the  $Gap(k)$  is larger than  $Gap(k+1)$  minus the error of  $Gap(k+1)$  (Fig. 3.4C). For *sox9a*, going from  $k = 1$  to  $k = 2$  clusters reduces our measure of within-cluster dispersion, and going to  $k = 3$  greatly reduces  $W_k$ ; this is reflected in the large jump in the gap measure (Fig. 3.4C). However, after  $k = 3$  the within-cluster dispersion keeps improving, but not at a rate that is faster than in the randomly generated null distribution and from this data, the algorithm determines that  $k = 3$  is the optimal number of clusters for this dataset. Looking at a second example for the human *ALDH1A2* gene (Fig. 3.4D-F), the BLAST scores are not nearly as distinctly distributed. In this example, the algorithm is not able to determine the optimal number of clusters to use as  $G(k) \not\geq G(k+1) - s(k+1)$ .

The noise reduction stage of the paralog pipeline applies this algorithm to the BLAST results of every query gene. The analysis stage utilizes R [87] to perform the hierarchical clustering portion of the algorithm and is parallelized for speed. For each query gene, if an optimal number of clusters can be found for the BLAST data, BLAST hits that fall in the cluster with the lowest set of scores are discarded as insignificant alignments. If the algorithm is unable to determine the optimal number of clusters to use, none of the data is discarded. This novel algorithm provides consistent clustering results and requires no arbitrary configuration variables allowing it to be applied to a wide variety of datasets at different evolutionary distances. These



**FIGURE 3.5:** The single linkage clustering algorithm of the RBH Analysis Pipeline. BLAST search results are represented as a directed graph, with each node representing a gene and each directed edge in the graph representing a BLAST hit between two genes (the label of the edge represents the rank order of the BLAST hit). A cycle of length 2 formed between two nodes represents a generalized reciprocal best hit (gRBH).

filtered BLAST search results provide us with enough data to build paralogous groups for the primary genome, a task achieved by the single linkage clustering algorithm implemented in the next pipeline stage.

### 3.1.3 The Single Linkage Clustering Stage

The pipeline has now conducted a BLAST search for every gene in the primary genome, summarized, and then filtered low scoring alignments from the search results. The final paralog pipeline stage uses the collected BLAST results to build paralogy groups. Although reciprocal best hit (RBH) relationships are often used to identify orthologous genes between species [114], accommodating multiple duplication events requires a more general definition of RBH. Strictly speaking, given the paralogous genes **A**, **B**, and **C**, only two of them can be reciprocal best hits. However, we can accommodate multiple duplication events by allowing for transitivity in our BLAST

hits – that is, 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 generalized reciprocal best hits (gRBH). More formally, the analysis pipeline employs a single linkage clustering algorithm to achieve this goal [109]. As shown in Figure 3.5, we can represent our BLAST search results as a directed graph, with each node representing a gene and each directed edge in the graph representing a BLAST hit between two genes (the label of the edge represents the strength of the BLAST hit – a rank of 1 is the best BLAST hit, a rank of 2 is the second-best, and so on). A cycle formed between two nodes represents a reciprocal best hit, however, we must consider edges by their rank. That is, we cannot form a cycle using an edge of rank 2 if we have not first examined the edge of rank 1 in the graph. Given this algorithm, we traverse the graph collapsing nodes each time we encounter a gRBH; repeating the procedure until no more nodes can be collapsed. Figure 3.5A displays a portion of such a graph showing genes  $g1$ ,  $g2$ ,  $g3$ , and  $g4$  and the edges between them. The cycle between genes  $g1$  and  $g3$  shows that they are generalized reciprocal best hits. The pipeline then collapses the  $g1$  and  $g3$  nodes (Fig. 3.5B) and establishes a new gRBH cycle in the graph – representing a best hit from the  $g1$  or  $g3$  gene to  $g2$ . Another iteration reveals a third gRBH between the  $g1g3g2$  and  $g4$  nodes (Fig. 3.5C). As the original graph (Fig. 3.5A) illustrates, genes  $g1$  and  $g4$  have no direct connection.

Pseudocode for the single linkage clustering algorithm is available in Appendix B. As the code shows, there are three major loops utilized in the implementation of the algorithm. Given a particular gene, the inner most loop examines all of that gene's BLAST hits looking for gRBH cycles. The second most inner loop iterates over all of the genes in the primary genome. Finally, the outer loop continues executing the two inner loops as long as a gRBH cycle is found in the previous execution. Given  $n$  as the number of genes, in the worst case scenario, this algorithm performs on the order of  $O(n^3)$ , although in practice, that limit is never reached (the number of BLAST hits per gene is limited by BLAST E-value, only a fraction of the genes in the primary genome are paralogs, and the number of genes in the genome is biologically limited to approximately 50,000).

At the conclusion of the single linkage clustering stage the paralog pipeline has built a set of paralogous groups from the genes in the primary genome. The remainder of the RBH Analysis Pipeline focuses first on collecting BLAST hits between genes in the primary and outgroup genomes (the ortholog pipeline) and then anchoring paralogous groups in the primary genome to their orthologs in the outgroup genome (the anchor pipeline).

### **3.1.4 The Ortholog Pipeline**

The modularity of PIP allows us to arbitrarily recombine pipeline stages, a feature that makes it easy to describe the second major pipeline, the ortholog pipeline

(Fig. 3.2B), which simply reuses the first five stages of the paralog pipeline. Genes from the primary genome are again loaded and a BLAST search is now performed for every query gene against the outgroup genome – referred to as the *forward* search; the results are summarized and low scoring alignments are filtered. Next, during the *reverse* search, all of the hits generated by the forward search (a subset of the outgroup genome) are loaded as query genes for a BLAST search back into the primary genome (a retro- or reverse-BLAST). The final results are again summarized and filtered. This stage can be repeated multiple times for different outgroup genomes. For example, the *Danio rerio* genome may first be run against the human genome, then against the stickleback, and so on. The result of the ortholog pipeline runs are combined with the paralog pipeline output in the anchoring pipeline.

### 3.1.5 The Anchoring Pipeline

Prior to executing the anchoring pipeline, the paralog pipeline has constructed a number of paralogous groups from the primary genome and the ortholog pipeline has amassed a catalog of BLAST hits to one or more outgroup genomes. This final component of the RBH Analysis Pipeline will anchor genes in the primary genome to their orthologs in the outgroup genome by examining BLAST hits between the two genomes. The first stage of the anchoring pipeline, the anchor stage (Fig. 3.2C), checks each member of each paralogous group to determine its top BLAST hit in the first 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 (Fig. 3.1). The BLAST hits for the outgroup genes are then checked to ensure that the outgroup gene retro-BLASTs back to 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. The size and membership of each paralogous group is relative to the last whole genome duplication that occurred in the primary genome and did not occur in the outgroup genome.

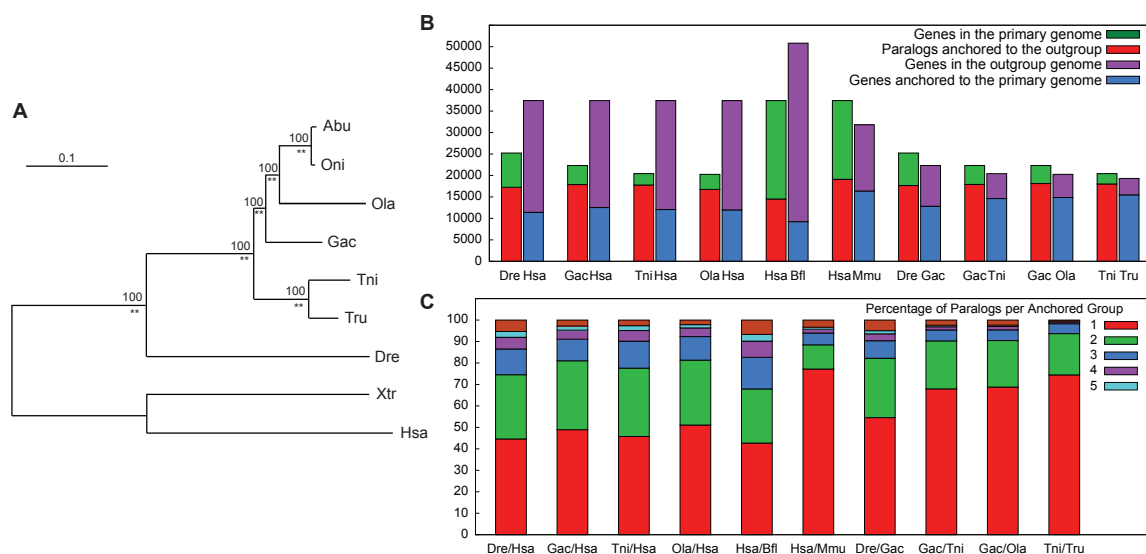
The second stage of the anchoring pipeline looks for outgroup genes that are recent tandem duplicates of each other. These genes are located on the same chromosome generally within a megabase of their duplicate. The system will search for outgroup genes that are very close to one another on the same chromosome and then check if the two paralogy groups in the primary genome originated from the same group (before being split during the anchoring stage). In these cases, the system will merge the two paralogous groups. Finally, the annotation stage of the anchoring pipeline

merges results produced by splice variants of the same gene and stores additional data related to the primary and outgroup genes for use by the web interface. In the next section, we present the results of applying the RBH Analysis Pipeline against several teleost, mammalian, and chordate genomes.

## 3.2 Results

We executed the RBH Analysis Pipeline using several teleost fish as the primary genome and using the human genome as the outgroup. The analysis included zebrafish, *Danio rerio* (Dre), stickleback *Gasterosteus aculeatus* (Gac), green-spotted pufferfish, *Tetraodon nigroviridis* (Tni), and medaka, *Oryzias latipes* (Ola). In addition we used the human genome as a primary genome against the cephalochordate amphioxus genome, as well as against the mouse, and we ran several of the teleosts as the primary genome against a second teleost as an outgroup. Much of this data was generated in order to find regions of conserved synteny (Chapter IV) and to make inferences regarding ohnologs gone missing (Chapter V). Immediately, however, we will examine a subset of the data to identify some general trends and then use the data in order to infer the ancestral gene order of a zebrafish and pufferfish chromosome.

The phylogenetic tree produced by Hoegg and colleagues [41] (Fig. 3.6A) shows the teleost fish as a distinct clade with human as the most basally diverging species in the tree. Within the teleost clade, stickleback is most related to medaka, while pufferfish and fugu (Tru) are the most closely related pair of teleosts; the zebrafish is the



**FIGURE 3.6:** Summary of RBH Analysis Pipeline Results. (A) A phylogenetic tree showing the evolutionary relationships between several teleost fish, including *Oryzias latipes* (Ola), *Gasterosteus aculeatus* (Gac), *Tetraodon nigroviridis* (Tni), *Takifugu rubripes* (Tru), *Danio rerio* (Dre), and human (Hsa). Based on the tree from [41]. (B) A summary of gene counts showing the total number of genes in each genome for which the RBH Analysis Pipeline established an orthologous relationship. (C) The percentage of genes in paralogy groups of a distinct size.

earliest branching of the teleost fish on the tree. Figure 3.6B, shows pairs of columns, the left column representing the primary genome, the right column representing the outgroup genome. Each column represents the total size of the genome (green/purple) along with the number of genes within the genome that were anchored (red/blue). The first four column pairs represent the results from analyzing teleost fish with a human outgroup, and since the teleosts experienced the R3 duplication while the human lineage did not, it makes sense that a higher percentage of teleost genes are anchored than human genes, indicating that multiple teleost genes are being anchored to a single human gene. In fact, the ratio ranges between 1.4 (Ola) and 1.51 (Dre)

teleost genes anchored to one human gene. Likewise, an analysis using human as the primary genome and amphioxus (Bfl) as the outgroup produced the highest ratio of 1.57. Given that the R1 and R2 WGD events are the most ancient [78], and therefore the hardest to detect, the human/amphioxus ratio is still higher than any of the ratios detected between teleost and human genomes (where the teleost genomes have experienced the more recent R3 WGD) – consistent with the fact that the human genome has experienced the R1 and R2 WGD events while amphioxus has not. The primary to outgroup gene ratio is smallest for the human/mouse (Mmu) comparison (1.16), and the teleost/teleost results also have a smaller ratio – consistent with comparing genomes that have the same number of duplication events in their history.

Figure 3.6C shows the percentage of paralogs in the primary genome that are in a group of a particular size. While the highest percentage of genes are found in groups of size one (a single primary gene anchored to a single outgroup gene), the teleost/human datasets exhibit the largest percentage of genes in a group of size two. Likewise, the human/amphioxus analysis shows the highest percentage of primary genes in groups of size three and four. These results are consistent with the relative distribution of whole-genome duplications in the primary versus outgroup genome, which the RBH Analysis Pipeline is built to detect.

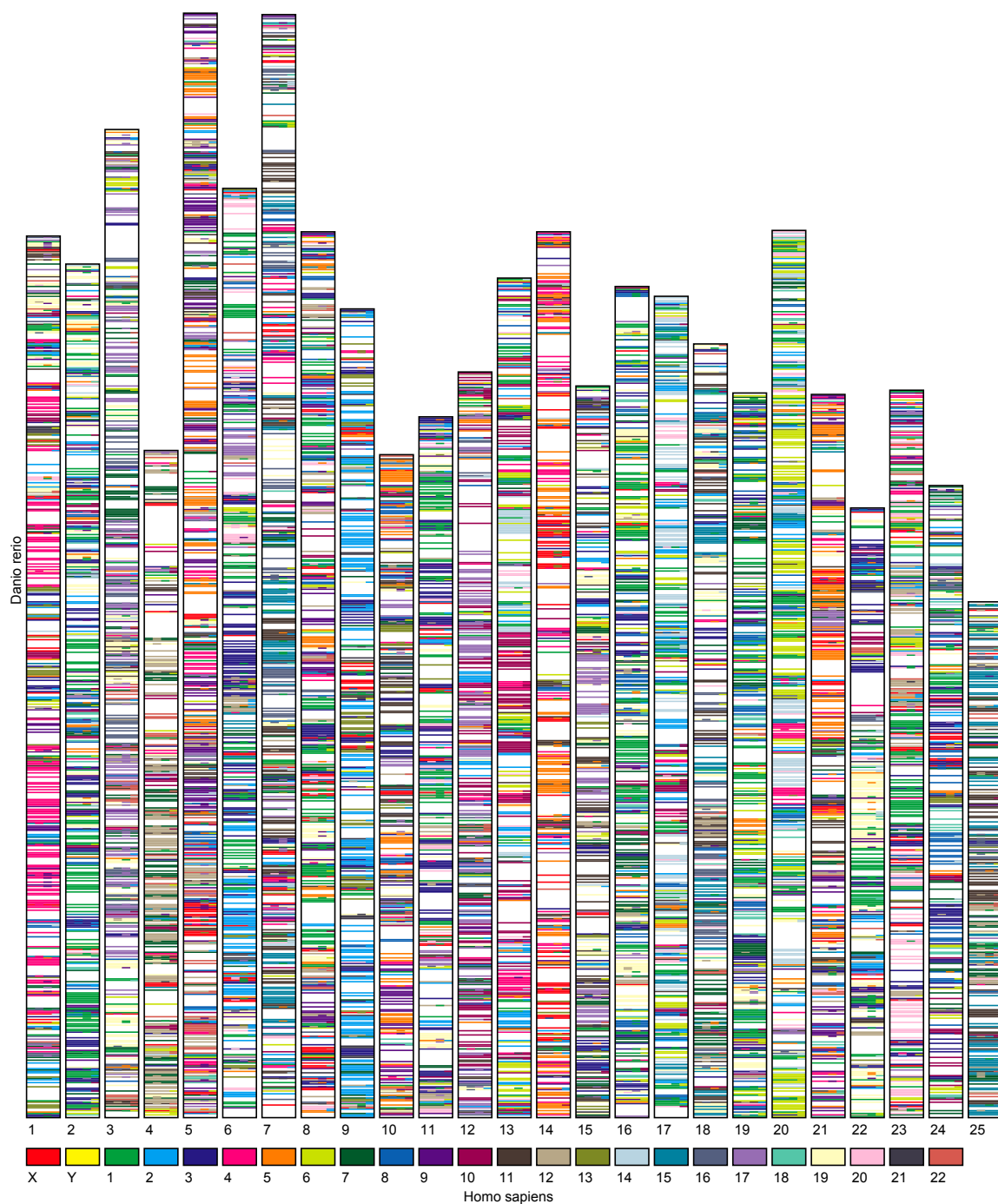
When we look at the results of the pipeline analyses for pairs of species that are expected to have a one to one orthology (Hsa/Mmu, Dre/Gac, Gac/Ola, Gac/Tni, Tni/Tru), we find that a higher percentage of primary genes are in groups of size

one. Moreover, we find that the percentage of single orthologs is proportional to the evolutionary distance between the genomes: human and mouse are the most closely related and have the most single orthologs, followed by Tru/Tni, Gac/Ola, Gac/Tni, and Dre/Gac. While a naive interpretation of the R1, R2, and R3 duplication events would lead us to expect all of teleost/human gene groups to have a primary to outgroup ratio of two to one, and similarly, would lead us to believe the human/amphioxus gene groups to have a ratio of four to one, in practice this is not the case. As we described in Section 1.2, gene loss is widespread in the time following a duplication event; the more diverged the species being compared the fewer genes retained in duplicate. Given that teleosts and human are diverged by several hundred million years (and human and amphioxus are even further diverged), we will not find a perfect two to one ratio (or four to one). However, if we were to examine a fish that was much more closely related to the teleosts but had not experienced the R3 duplication event, say the Semionotiformes (gars) [49, 68], we would expect to find a ratio very near two to one (the gar genome has not yet been fully sequenced).

Besides tallying up the number of primary and outgroup genes the RBH Analysis Pipeline found, we can also examine the spatial distribution of orthologs. In genomes that have experienced a duplication relative to the outgroup we expect to find orthologs distributed in a way that reflects the duplication of the chromosomes they resided on. Figure 3.7 shows the 25 zebrafish chromosomes along with the paralogs

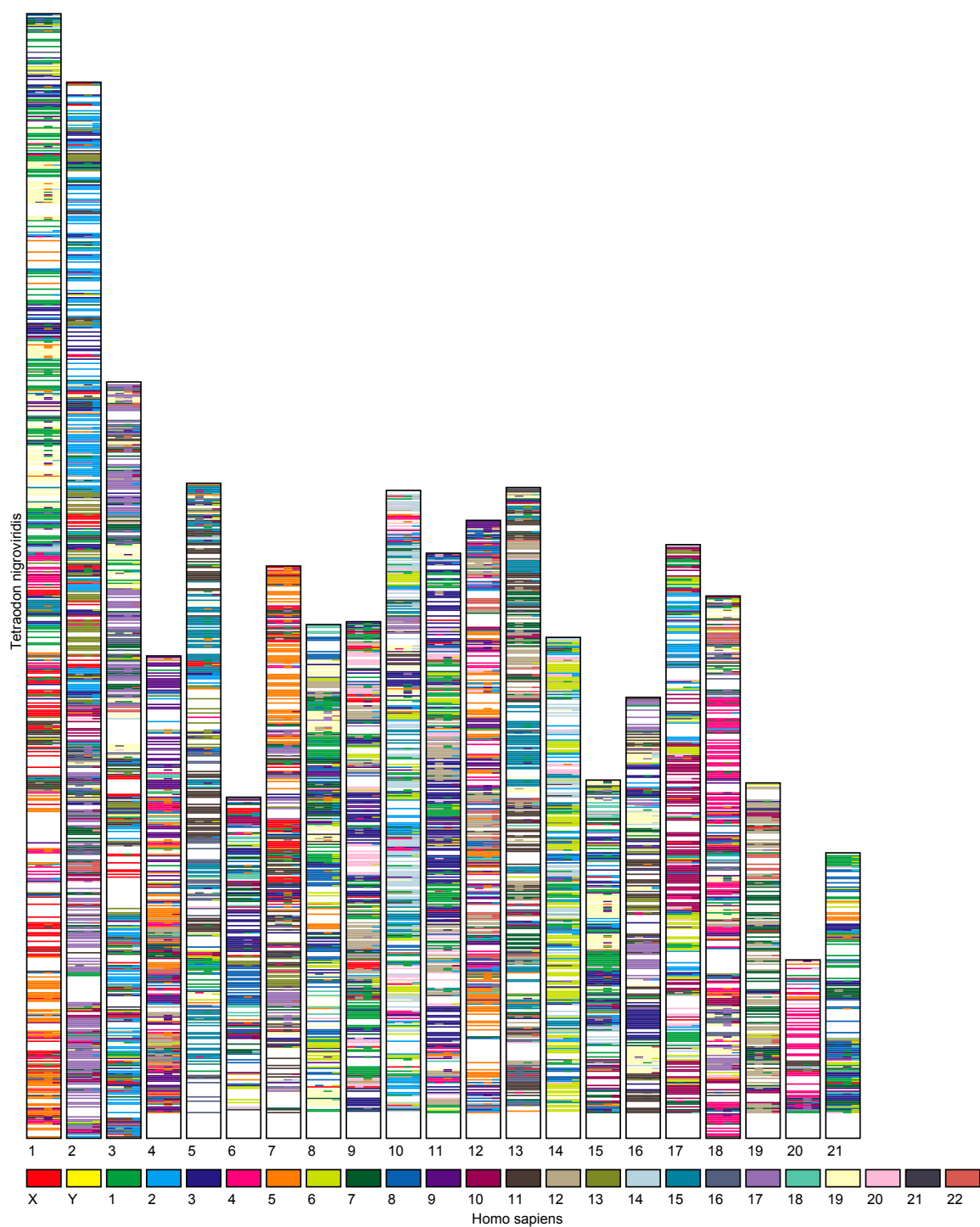
the pipeline detected. The paralogs are colored according to the location of their outgroup ortholog in the human genome, so, zebrafish orthologs of human chromosome 1 (Hsa1) would be colored green in the image. If there was a perfectly preserved ordering of zebrafish genes relative to their human orthologs, and the zebrafish and human genomes experienced the same number of WGD, then we would expect each zebrafish chromosome to be a single, solid color, corresponding to its human orthologous chromosome. On the other hand, if the coloration of the zebrafish chromosomes was totally random, then there would be no evidence of conserved synteny. Evidence for a WGD in the zebrafish would appear as multiple zebrafish chromosomes (or portions of those chromosomes) with the same coloring, indicating that both regions contain orthologs located on the same chromosome in the human genome. Looking at Figure 3.7, genes on zebrafish chromosome 1 (Dre1) show strong conservation (pink) to human chromosome 4 (Hsa4) and Dre14 shows weaker conservation to Hsa4. Zebrafish chromosomes 3 and 12 also show strong conservation (purple) to Hsa17 indicating that Hsa17 exists in duplicate in the zebrafish, on Dre3 and Dre12 and Hsa4 exists on Dre1 and Dre14. While the zebrafish genome appears to have experienced many architectural rearrangements relative to the human genome (hence the fragmented nature of the coloration), pairs of chromosomes can be identified. The pufferfish genome is much less fragmented than the zebrafish genome (Fig. 3.8) and shows a much stronger duplication signal. Human chromosome 2 is split between Tni2

and Tni3 (light blue) while Hsa5 is split between Tni1, Tni4, Tni7, and Tni12 (orange). When we look at genomes that have the same number of relative duplication events, such as stickleback and medaka (Fig. 3.9, or human and mouse (Fig. 3.10) we see a very clear one-to-one ratio between regions of the genome. Although there have still been rearrangements, the regions do not exist in duplicate. In the next section we will introduce an additional type of visualization that will confirm this fact.

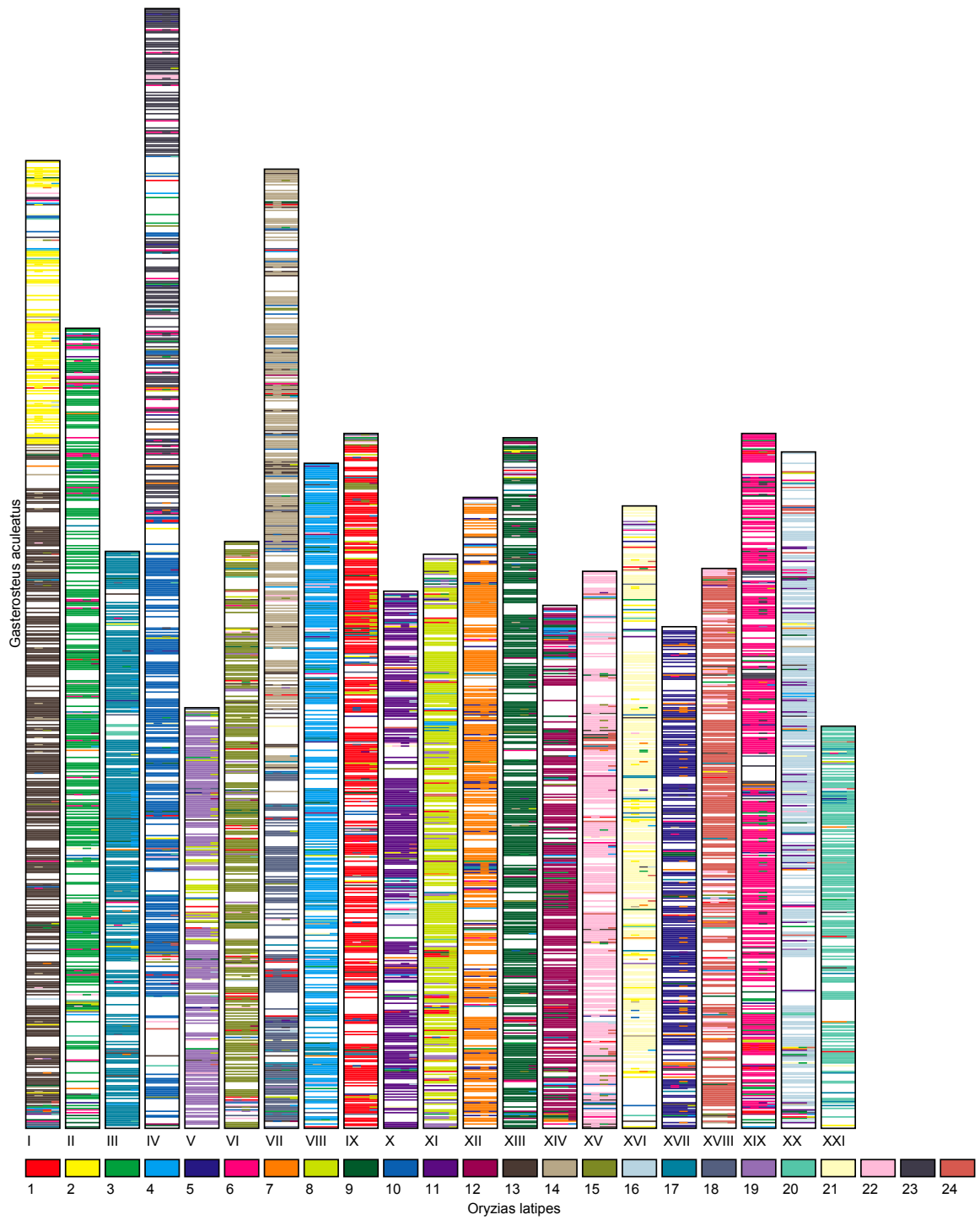


**FIGURE 3.7:** *Danio rerio* primary genome anchored to the *Homo sapiens* outgroup genome.

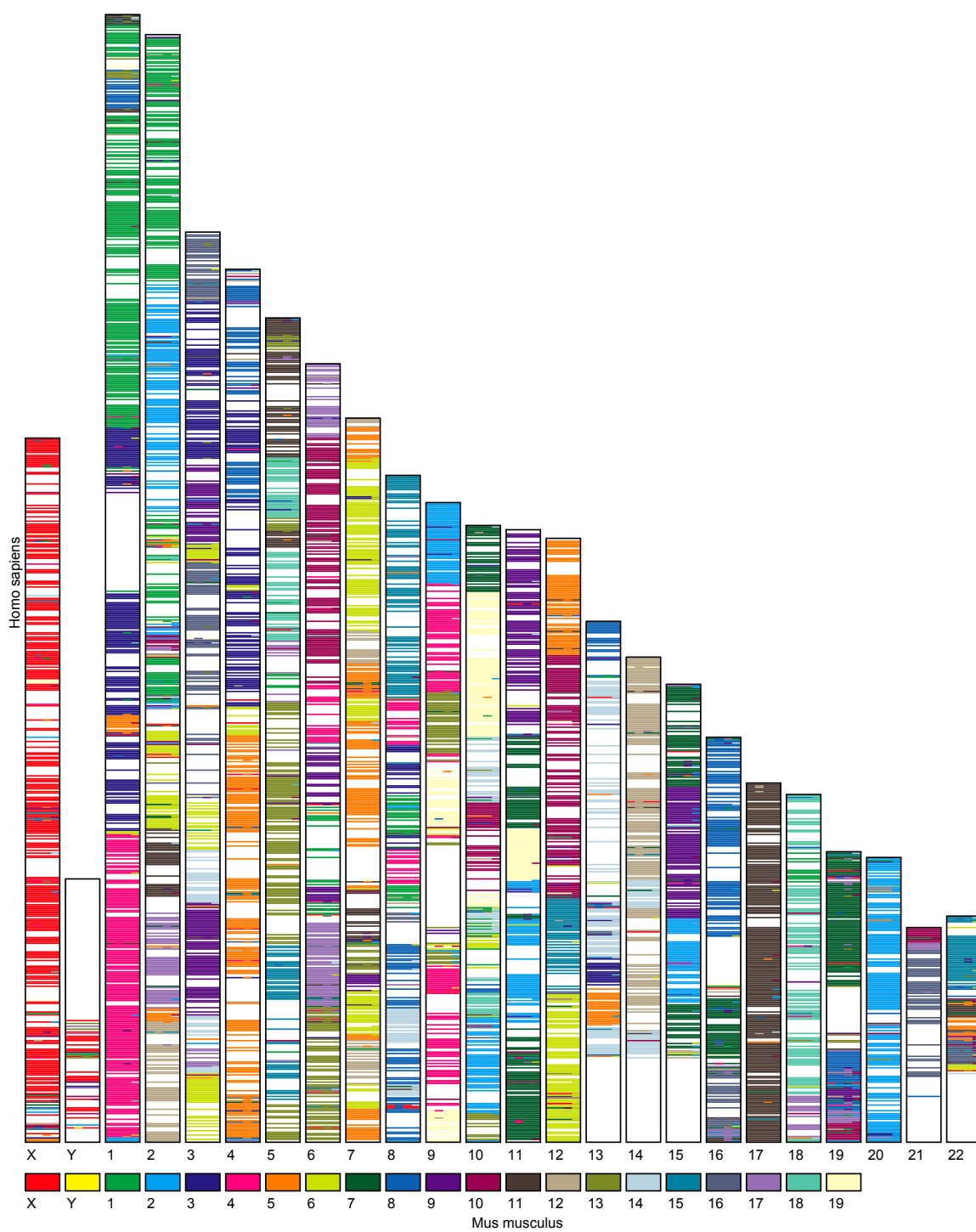




**FIGURE 3.8:** *Tetraodon nigroviridis* primary genome anchored to the *Homo sapiens* outgroup genome.



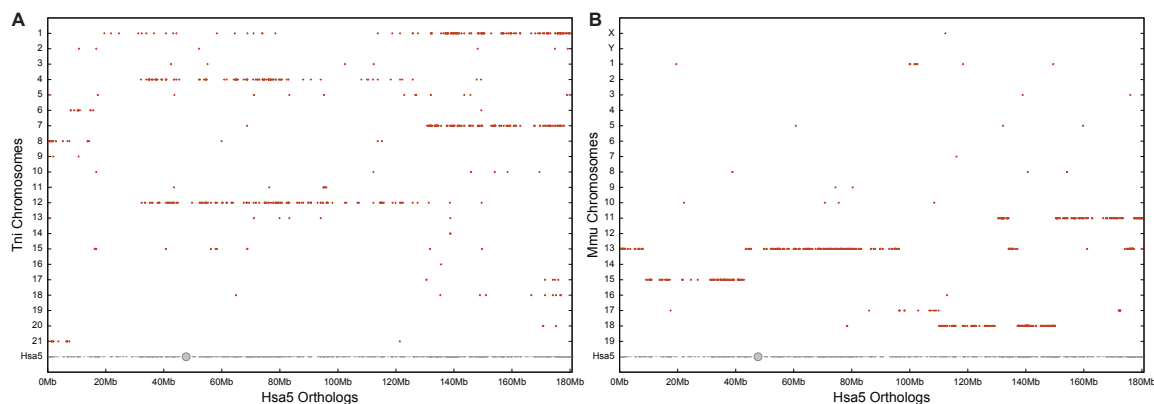
**FIGURE 3.9:** *Gasterosteus aculeatus* primary genome anchored to the *Oryzias latipes* outgroup genome.



**FIGURE 3.10:** *Homo sapiens* primary genome anchored to the *Mus musculus* out-group genome.

### 3.2.1 Dotplots

Plots showing the distribution of orthologs across a genome are broadly informative (Figs. 3.7-3.10), but lack the detail to make strong statements about how the physical layout of genes has changed across genomes. However, using a dotplot to visualize the paralogs and orthologs detected by the RBH Analysis Pipeline, changes in genome architecture can easily be detected, including patterns indicative of a whole-genome duplication. First introduced for the display of synteny by Dehal and Boore [25], for a particular chromosome a dotplot displays the distribution of orthologs or paralogs of that chromosome across the rest of the genome. If we return to our claim from the previous section that orthologs from human chromosome 5 (Hsa5) are distributed across pufferfish chromosomes 1, 4, 7, and 12, a dotplot image can make the evidence in favor of this claim visible. In the plot (Fig. 3.11A), Hsa5 is displayed along the X-axis and genes that reside on Hsa5 are drawn as grey dots. Pufferfish orthologs are displayed directly above their human copies on their natural pufferfish chromosome, however, the genes are ordered with respect to the genes on Hsa5. The advantage of this approach is that if a single region in human exists in a duplicated state in pufferfish, then orthologs will be displayed in parallel along their duplicated chromosome segments. This duplication signal is exactly what we see in Figure 3.11 where the upper 50 megabases of chromosome 5 is duplicated over pufferfish chromosomes 1 and 7, while a separate, 100 megabase region of Hsa5 is duplicated over pufferfish chromosomes 4 and 12 – a clear signal of the R3 WGD. Figure 3.11B,



**FIGURE 3.11:** Orthology dotplots reveal duplication signal. (A) A dotplot showing all detected *Tetraodon nigroviridis* (Tni) orthologs of genes on human chromosome 5 (Hsa5). Hsa5 is represented in duplicate in Tni, with portions on Tni chromosomes 1 and 7, 4 and 12, and 8 and 21. (B) A dotplot showing all detected mouse orthologs to genes on Hsa5. Human chromosome 5 is unduplicated in the mouse, represented by portions of mouse chromosomes 11, 13, 15, 17, and 18.

shows a comparison of Hsa5 instead with the mouse genome. Here both genomes have experienced the same number of WGD and although Hsa5 has been rearranged onto several different chromosomes in the mouse since the human/mouse speciation (or vice versa), the ancestral copy of the chromosome clearly only exists as a single copy in the two genomes. If synteny was not conserved in the human, mouse, and pufferfish genomes (regardless of WGD), we would expect to see a random pattern of red crosses in both Fig. 3.11A and Fig. 3.11B demonstrating that there was no relationship between orthologs and their location in the genome.

A Forward BLAST Hit #1:						B Forward BLAST Hit #2:					
QUERY	HIT	SCORE	E-VALUE	LENGTH	PERCENT IDENT	QUERY	HIT	SCORE	E-VALUE	LENGTH	PERCENT IDENT
<i>msxb</i> (Dre1)	<i>Msx2</i> (Mmu13)	558	1.1e-55	281	49.82%	<i>msxb</i> (Dre1)	<i>Msx3</i> (Mmu7)	511	4.2e-53	155	69.03%
Reverse BLAST:						Reverse BLAST:					
QUERY	HIT	SCORE	E-VALUE	LENGTH	PERCENT IDENT	QUERY	HIT	SCORE	E-VALUE	LENGTH	PERCENT IDENT
<i>Msx2</i> (Mmu13)	<i>msxd</i> (Dre21)	641	9.6e-65	237	60.34%	<i>Msx3</i> (Mmu7)	<i>msxc</i> (Dre13)	546	1.6e-54	213	56.81%
<i>Msx2</i> (Mmu13)	<i>msxa</i> (Dre14)	607	4.3e-61	231	56.28%	<i>Msx3</i> (Mmu7)	<i>msxb</i> (Dre1)	512	7e-51	185	61.08%
<i>Msx2</i> (Mmu13)	<i>msxc</i> (Dre13)	591	2.3e-59	236	57.20%	<i>Msx3</i> (Mmu7)	<i>msxe</i> (Dre14)	486	4.3e-48	166	64.46%
<i>Msx2</i> (Mmu13)	<i>msxb</i> (Dre1)	558	8e-56	281	49.82%	<i>Msx3</i> (Mmu7)	<i>msxd</i> (Dre21)	469	1.2e-48	136	72.06%

**FIGURE 3.12:** BLAST search results for zebrafish *msxb* against the mouse genome. (A) The top BLAST hit for *msxb* is mouse *Msx2*; reverse-BLASTing *Msx2* back against the zebrafish genome returns the zebrafish *msx* paralogs. (B) The second best BLAST hit for *msxb* is mouse *Msx3*; reverse-BLASTing *Msx3* back against the zebrafish genome returns the same zebrafish *msx* paralogs in a different order. Although a mouse *Msx3* BLAST search hits the same zebrafish genes as mouse *Msx2*, all of the *Msx3* hits have a lower score than the *Msx2* hits.

### 3.2.2 The effect of rate asymmetry on reciprocal best hit

#### BLAST

As we discussed in the introduction (Section 1.2), one feature common to duplicate genes resulting from a WGD is evolutionary rate asymmetry – one of the duplicates evolves at a faster rate than the other and experimental evidence suggests that rate increases occur soon after the WGD event in one of the duplicates. This phenomenon is one of the major limiting factors for an RBH-based orthology assignment algorithm. When a single copy of a gene is present in two genomes the RBH method will reliably determine that the genes are orthologous. However, when duplicate paralogs of the genes exist due to a WGD, rate asymmetry can cause incorrect assignments to be made once the genes are sufficiently diverged. An example of this effect can be seen with the MSX gene family in zebrafish and mouse. We will discuss the function and

evolutionary history of this gene family in detail in Chapter IV, but for our purposes here we will present the BLAST results for one of the zebrafish paralogs (Fig. 3.12).

There are five MSX paralogs in the zebrafish, and three paralogs in the mouse. Zebrafish genes *msxa* and *msxb* are co-orthologous to mouse *Msx2*, *msxc* and *msxd* are co-orthologous to mouse *Msx3*, and *msxe* is orthologous to *Msx1*. The RBH Analysis Pipeline, however, finds that *msxa*, *msxb*, *msxc*, and *msxd* are all co-orthologous to *Msx2*, which is incorrect. This misassignment is caused by rate asymmetry.

If we examine the BLAST search results for *msxb* against the mouse genome, we find that its top BLAST hit is mouse *Msx2*. The reverse-BLAST search, using *Msx2* as a query against the zebrafish genome, returned *msxd*, *msxa*, *msxc*, and *msxb* in that order (Fig. 3.12A). Now, the second best BLAST hit for *msxb* is its correct ortholog, *Msx3*, and performing a reverse-BLAST with *Msx3* as a query against zebrafish returned *msxc*, and *msxb* as the top two hits (Fig. 3.12B). However, the scores for these two hits were both lower than all four of the BLAST hits for *Msx2*. Therefore, the RBH analysis pipeline erroneously grouped *msxb* and *msxc* with mouse *Msx2*. The mouse *Msx3* gene has apparently diverged far enough from its zebrafish orthologs that there is now greater similarity between all four zebrafish paralogs with mouse *Msx2* than with *Msx3*. The pipeline does not have the power to make the proper assignment.

Every orthology inference method has its limitations, but the effects of rate asymmetry on RBH BLAST have not been described previously. Rate asymmetry becomes

problematic when comparing genomes that are highly diverged. The most constructive approach to this problem is to add RBH comparisons between additional species that are more closely related. Comparing the zebrafish MSX genes first against the more closely related gar fish, and then comparing gar to the mouse genome would be one plausible approach to solve this problem. Another approach, which we will discuss in detail in the next chapter is to use the conserved synteny of neighboring genes to aid in making the proper assignments.

### 3.2.3 Data Sources

For this chapter, Ensembl [12, 55] provided data for the *Homo sapiens* genome, using NCBI v36 obtained from Ensembl version 52; the *Danio rerio* genome, using Zv7 from the Sanger Institute obtained from Ensembl 52; the *Gasterosteus aculeatus* genome, using BROAD version S1 obtained from Ensembl 52; *Tetraodon nigroviridis* genome, using TETRAODON 8 obtained from Ensembl 52; *Oryzias latipes* genome, using version HdrR obtained from Ensembl 52; and the *Mus musculus* genome, using NCBI version m37 obtained from Ensembl 52. We also obtained version 2 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/>).



**RBH Analysis Pipeline** Synomix Tools

[circle plots](#) | [dotplots](#) | [syntney db](#)

▼ **Select Data Sources**

Organism:  Outgroup:  Variant:

► **Filter Results**

◀ 1 ▶ (1 groups) groups per page 10

Group	Primary Genes		Outgroup Gene
<a href="#">2559 detail</a>	<a href="#">ENSDARG00000043923</a> (sox9b) Dre: 3 position: 60,693k	<a href="#">ENSDARG00000003293</a> (sox9a) Dre: 12 position: 538k	<a href="#">ENSG00000125398</a> (SOX9) Hsa: 17 position: 67,629k

◀ 1 ▶ (1 groups) groups per page 10

**FIGURE 3.13:** The RBH Analysis Pipeline web interface.

### 3.2.4 User Interface

The results of the RBH analysis pipelines are made available through a web-based interface (Fig. 3.13). This interface provides an extensive filtering interface allowing a researcher to view results according to a particular gene, chromosome, or chromosomal region. In addition, for every orthology assignment, details are made available showing the BLAST search results, the noise reduction algorithm, and the output of the single linkage clustering algorithm. Subsets of results can be exported directly from the website to a spreadsheet program such as Microsoft Excel.

In addition, several visualization tools have been made available through the web as well. The researcher can generate dotplots for any primary or outgroup chromosome, can highlight particular genes in the plots, and can export the images in raster

or vector format. Two other visualization tools are also available allowing the export of gene homology matrices and circle plots as well. These later two visualizations will be described later in this chapter while all three types of visualizations are used extensively in the case studies of this work. Having presented the results of the RBH Analysis Pipeline as well as the Pipeline's limitations, we next use data generated by the Pipeline to infer the architecture of an ancestral teleost chromosome.

### 3.3 Case Study: Inferring Ancestral Gene Order

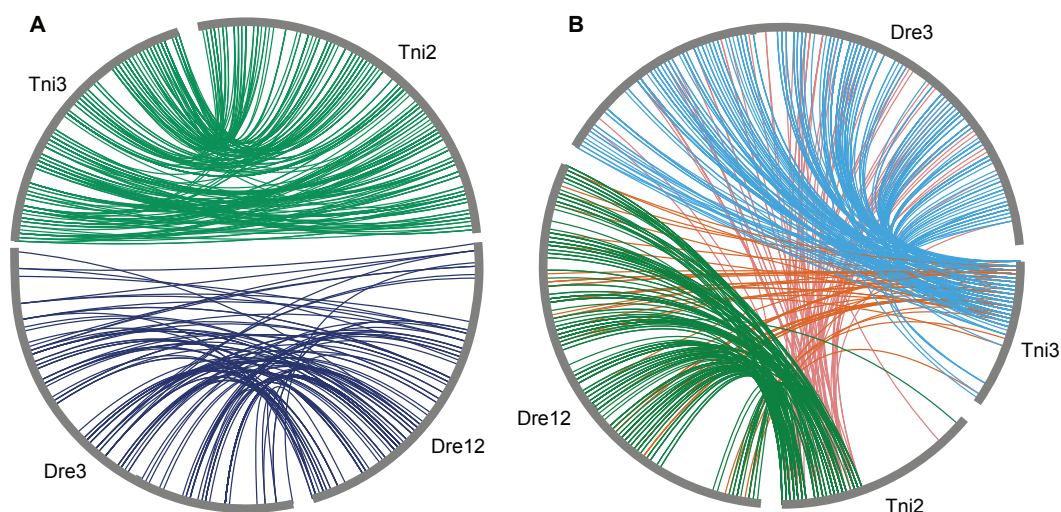
By assigning paralogy and orthology between genes and visualizing the distribution of those genes across genomes, we are able to use the RBH Analysis Pipeline to infer conserved gene orders within a primary genome and between a primary and out-group genome. Regions of conserved gene order in the genome may reflect either the affect of selection preserving the order, or simple failure by chance to fix chromosome rearrangements in a population over time. If fully conserved genomic blocks persist in different lineages over increasing time periods, then selection becomes an increasingly probable mechanism for the maintenance of conserved blocks. As we discussed in the Introduction (Section 1.2), one of the best-studied examples of conserved gene order in vertebrate genomes are the HOX clusters, which provide an example of gene order conserved due to functional constraints.

To investigate whether we could identify additional genomic regions containing conserved gene order we applied the RBH Analysis Pipeline to two teleost genomes

and the human genome. This work, originally published in [19], investigated the conservation of gene orders in the teleost genomes, inferring ancestral gene orders in the pre-duplication teleost genome, and inferred genome content in the last common ancestor of teleost fish and mammals by comparing the ancestral teleost genome to the human genome.

Inferring the gene content of the last common ancestor of teleosts and mammals requires three organisms: a primary organism (zebrafish in this case) and two outgroups. The recent outgroup is an organism that diverged from our primary organism after the R3 duplication event, and we will use the green-spotted pufferfish *Tetraodon nigroviridis*, whose genome sequence is nearly complete [51]. An organism that diverged from our primary organism prior to the most recent duplication can be used as an ancient outgroup, in this case we use the human genome because of its high quality of annotation. We executed the RBH Analysis Pipeline with zebrafish as the primary genome and anchored it to both pufferfish (recent) and human (ancient) outgroup genomes. After collecting the data, we proceed in the following way:

1. We compare the gene content of chromosomes in the primary species to the genome of the recent outgroup to infer the content of the ancestral post-duplication teleost chromosomes. This comparison reduces two pairs of modern chromosomes to a single, ancestral post-duplication pair.



**FIGURE 3.14:** Search for paralogous and orthologous chromosome segments. (A) Paralogous chromosomes are identified within the zebrafish or pufferfish. Pufferfish chromosomes 2 (Tni2) and 3 (Tni3) are drawn around the circumference of the top half of the circle. Green arcs represent paralogous genes on the two chromosomes. Similarly, zebrafish chromosomes 3 (Dre3) and 12 (Dre12) are drawn along the circumference of the bottom half of the circle and blue lines represent paralogous genes between them. (B) Orthologous chromosomes between zebrafish and pufferfish. The same pufferfish (Tni2, Tni3) and zebrafish (Dre3, Dre12) chromosomes are drawn around the circumference of the circle with arcs between the circles showing orthologs among the four chromosomes. Tni2 is strongly orthologous to Dre12 (green) and Tni3 is strongly orthologous to Dre3 (blue).

2. We next infer the content of the ancestral pre-duplication chromosome of a ray-fin (Actinopterygian) fish, which existed about 300 million years ago, by collapsing the post-duplication pair of chromosomes.
  
3. Finally, we compare the pre-duplication ray-fin fish chromosome to our ancient outgroup, the lobe-fin (Sarcopterygian) fish called *Homo sapiens*. This final comparison allows us to infer the content of the ancestral bony fish (Osteichthyes) chromosome that existed about 450 million years ago.

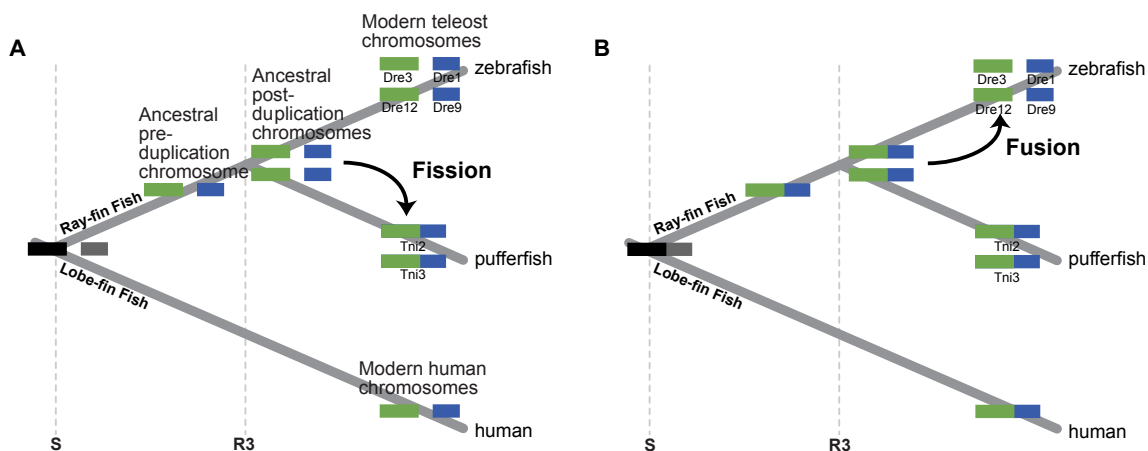
Our focus was to reconstruct the ancestral chromosome and gene orders for *Danio rerio* chromosome 3 (Dre3), one of the 25 zebrafish chromosomes. We examined the zebrafish/human pipeline results and identified paralogous genes within the *Danio rerio* genome to infer chromosome segments that constitute the most likely paralogon produced in the R3 duplication event. This analysis yielded *Danio rerio* chromosome 12 (Dre12) as the most likely Dre3 paralogon. We can visualize these results using a circle plot (Fig. 3.14) with the *Danio* chromosomes drawn as arcs around the circumference of a circle and with arcs between the chromosomes representing pairs of paralogous genes. The lower half of Figure 3.14A shows that genes distributed along the full length of Dre3 have duplicates distributed along the full length of Dre12, but that the order of paralogs is quite different in the two homeologous chromosomes, as evidenced by the crossing of lines that join paralogs. These types of differences in gene order would occur if many chromosome inversions occurred on both homeologous chromosomes since the R3 genome duplication event.

Next, we examined orthologs of genes from Dre3 and Dre12 in pufferfish, using the zebrafish/pufferfish pipeline results. This analysis yielded *Tetraodon nigroviridis* chromosome 2 (Tni2) as most closely related to Dre3, and Tni3 as most closely related to Dre12 (Fig. 3.14B). The principle of transitive homology ([109]) demands that the chromosome homeologous to Tni2 would be Tni3, and our data verified this prediction (Fig. 3.14A).

The distribution of orthologs revealed several features with implications regarding the mechanisms of chromosome evolution. First, zebrafish chromosomes appear to be stuffed into short regions on pufferfish chromosomes (Fig. 3.14B). This fits with the dramatic diminution of pufferfish genomes, a derived feature achieved by decreasing the length of introns and intergenic regions [31].

The second result apparent from the analysis is that gene order on Dre3 matches gene order on Tni3 far better than gene order on Dre3 matches gene order on Dre12. This result would be predicted by the hypothesis that fewer inversions occurred since the speciation event that produced the diverging zebrafish and pufferfish lineages (producing Dre3 and Tni3) than occurred since the genome duplication event that produced Dre3 and Dre12. If one assumes that the rate of the fixation of inversions in populations is roughly constant over time and between lineages, then these results suggest that the R3 genome duplication event was substantially earlier than the zebrafish/pufferfish speciation event.

Third, the analysis shows that nearly all pufferfish orthologs of Dre3 occupy only the lower portion of Tni3, and nearly all pufferfish orthologs of Dre12 reside only in the upper part of Tni2 (Fig. 3.14B). Two possible hypotheses can explain these distributions. According to the pufferfish fusion hypothesis, the last common ancestor of zebrafish and pufferfish had a chromosome like Dre3 (or Dre12), and that, in the pufferfish lineage, this chromosome became the lower part of Tni3 (or the upper part of Tni2), which joined an unrelated chromosome that became the top portion of



**FIGURE 3.15:** Two hypotheses for the reconstruction of ancestral chromosomes. (A) Pufferfish fusion hypothesis. (B) Zebrafish fission hypothesis.

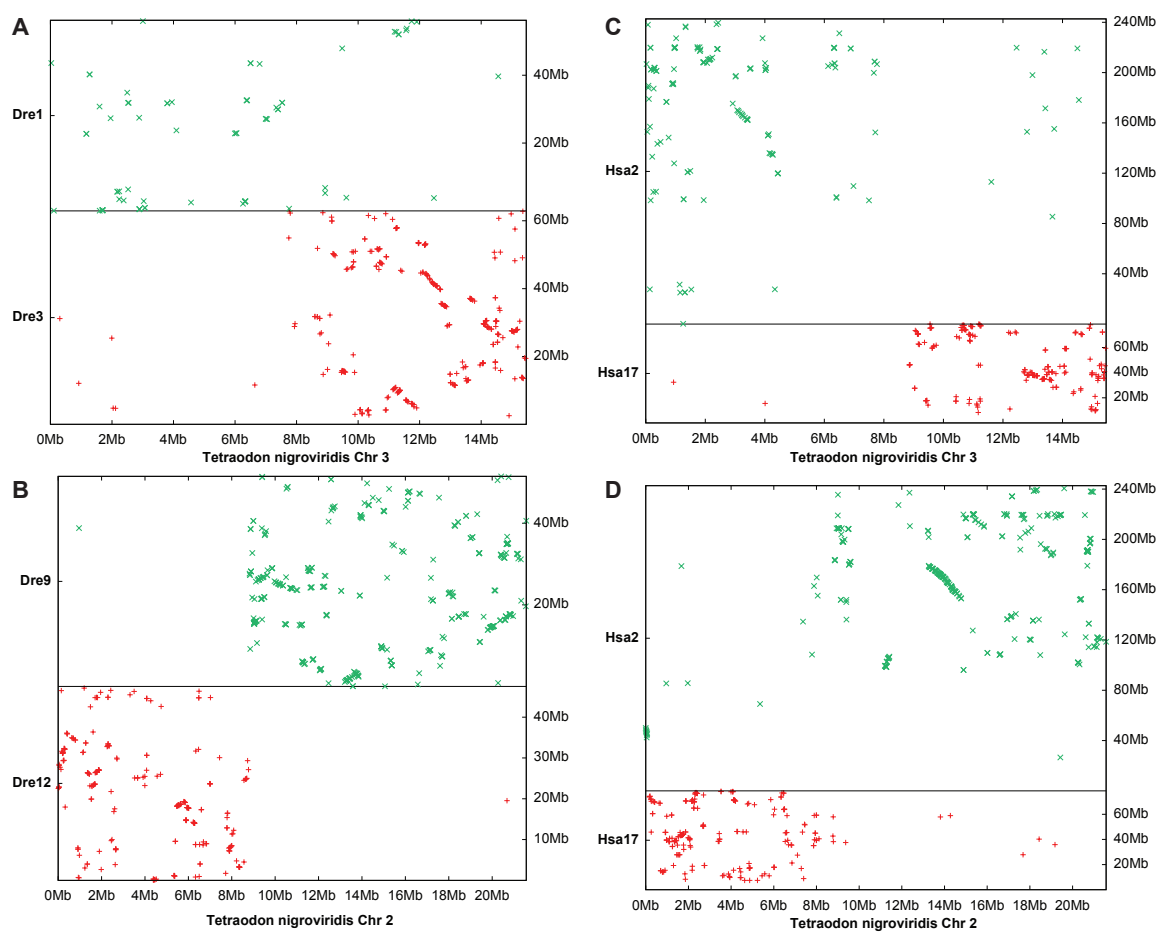
Tni3 (or lower part of Tni2) (Fig. 3.15A). The alternative hypothesis, the zebrafish fission hypothesis, is that the last common ancestor of zebrafish and pufferfish had a chromosome like Tni3 (or Tni2), and that in the zebrafish lineage, this chromosome broke roughly in half, yielding Dre3 from the lower half of Tni3, and Dre12 from the upper half of Tni2 (Fig. 3.15B).

The pufferfish fusion hypothesis and the zebrafish fission hypothesis make different predictions for the nature of the pufferfish chromosomes that are not related to Dre3 and Dre12. According to the pufferfish fusion hypothesis (Fig. 3.15A), the non-Dre3/12 portion of pufferfish chromosomes Tni2 and Tni3 (gray) would most likely be unrelated to each other because the fusion events that created Tni2 and Tni3 would have occurred independently of each other. Under the zebrafish fission hypothesis, however (Fig. 3.15B), the non-Dre3/12 portions of pufferfish chromosomes Tni2 and

Tni3 (gray) would be orthologous to the same portion of the human genome because they would have been part of the same ancestral pre-duplication chromosome.

We can visualize orthologs between pairs of chromosomes using a gene homology matrix [109], in which one of the chromosomes being compared is displayed along the X-axis of the plot, while the other is displayed along the Y-axis (or, multiple chromosomes can be stacked on the Y-axis). Then, orthologous genes are represented as a cross in the plot located at their physical coordinates on each chromosome. These visualizations of our pipeline data showed that the non-Dre3 portion of Tni3 (corresponding to Dre1, Fig. 3.16A), and the non-Dre12 portion of Tni2 (orthologous to Dre9, Fig. 3.16B) are both orthologous to the long arm of human chromosome two (Hsa2, Fig. 3.16C,D). 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 pre-duplication chromosome that was the ancestor to Dre3 consisted of a chromosome that was substantially similar to the sum of the genetic content of pufferfish chromosomes Tni2 and Tni3. This result is somewhat counterintuitive because *T. nigroviridis* has 21 chromosomes, while zebrafish and most other teleosts have 25 ([74]), which is expected if chromosome fusion occurred more frequently in the pufferfish lineage than in most teleosts. Thus, although the zebrafish fission hypothesis works best for this case, for other chromosomes, the answer is likely to be quite different.





**FIGURE 3.16:** Ancestral chromosome reconstruction. The portion of pufferfish chromosomes Tni2 and Tni3 that do not correspond to zebrafish chromosomes are orthologous to Dre1 and Dre9, respectively (A and B), but in both cases, are orthologous to much of human chromosome Hsa2 (C and D). This suggests that the ancestral chromosome state was the sum of the two pufferfish chromosomes.

The best way to finalize the inference of the ancestral chromosome would be to analyze the situation in closely related outgroups, including a post-R3 teleost outgroup and a pre-R3 non-teleost ray-fin outgroup. Although appropriate outgroup lineages exist, including for post-R3 the Anguilliformes (eels) and the Osteoglossiformes (butterfly fish and bonytongues), and the pre-R3 outgroups Amiiformes (bowfin) and Semionotiformes (gars) [49, 68], unfortunately none have available genomic resources necessary to resolve the issue.

Finally, the analysis reveals two special regions of pufferfish chromosome Tni2 that have extensive regions of conserved gene order, one at about 9 Mb and one at about 11 Mb. The corresponding regions in human occupy about 9 Mb and about 30 Mb of Hsa2, remarkably long conserved regions (at least 14 and 54 genes, respectively) preserved for a remarkably long time. Future challenges will be to understand the mechanisms for this preservation and to identify other similar regions on other chromosomes.

### **3.4 Summary**

In this chapter we described the Reciprocal Best Hit Analysis Pipeline, a high-throughput ortholog assignment algorithm that accounts for the effects of the R1, R2, and R3 whole-genome duplications in the vertebrates and features an effective paralog clustering method and a novel noise reduction algorithm. We ran the pipeline with a number of different datasets and described some general trends between genomes that

have experienced whole-genome duplications and genomes that have not. We then applied the resulting dataset to infer the gene content of a teleost/human ancestral chromosome. The data produced by the RBH Analysis Pipeline is very useful for determining the orthology or paralogy of individual genes and gene families and the aggregated data can be used to infer areas of conserved macro-synteny across different genomes. However, we want to look at conservation of synteny at a much finer scale, so that we can investigate the evolutionary history of individual gene families. To accomplish that goal we created the Synteny Database, which we describe in the next chapter.

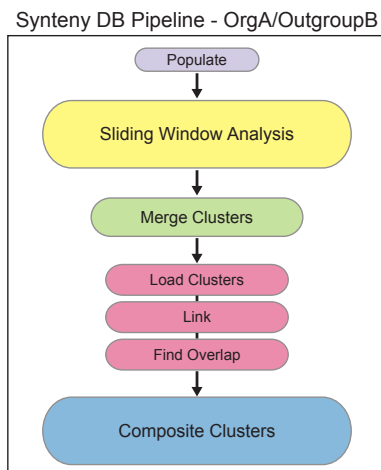
## CHAPTER IV

### THE SYNTENY DATABASE

In the previous chapter, we demonstrated how a reciprocal best hit algorithm applied to whole genomes could produce evidence of conserved synteny – the tendency of neighboring genes to retain their relative positions and orders on chromosomes over evolutionary time. As we described in the Introduction (Chapter I), in a WGD event, duplicated chromosomes (homeologs) initially have their gene orders intact. Between the time of duplication and speciation events, however, genes can be lost from one homeolog or the other, and inversions and other chromosome rearrangements can occur independently on the two duplicated homeologs. These events occurring in the chromosomal vicinity of a gene in question give an identity to all of the genes in the neighborhood. These neighborhoods can be compared between extant species and provide a source of additional evidence, independent of sequence identities or phylogenetic trees, to infer the evolutionary history of gene families.

We developed an automated system to identify conserved syntenic regions within a genome. The Synteny Database is able to cluster paralogous and orthologous genes

into syntenic regions by employing a sliding window analysis and relies on data generated by the RBH Analysis pipeline (Chapter III), which identifies paralogous gene groups in a primary genome and *anchors* those groups to their appropriate orthologous genes in an outgroup genome. The sliding window analysis identifies chromosomal segments within the primary genome and between 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 in the conserved regions. These syntenic clusters are checked to ensure that they are statistically significant through a permutation analysis and the results are presented to the researcher as a searchable, web-based database of conserved syntenic clusters. The system allows for the analysis of fully or partially assembled genomes [15], and is optimized for the investigation of individual gene families in multiple lineages. The Synteny Database is able to detect chromosome inversions and translocations and allows for the inference of ohnologs gone missing. After describing the implementation of the Synteny Database, we present two case studies to demonstrate the utility of the system: the evolution of the ARNTL and MSX gene families in the amphioxus, *Ciona intestinalis*, zebrafish, and human genomes. This work originally appeared in [20] and served as the primary investigative tool in [18].

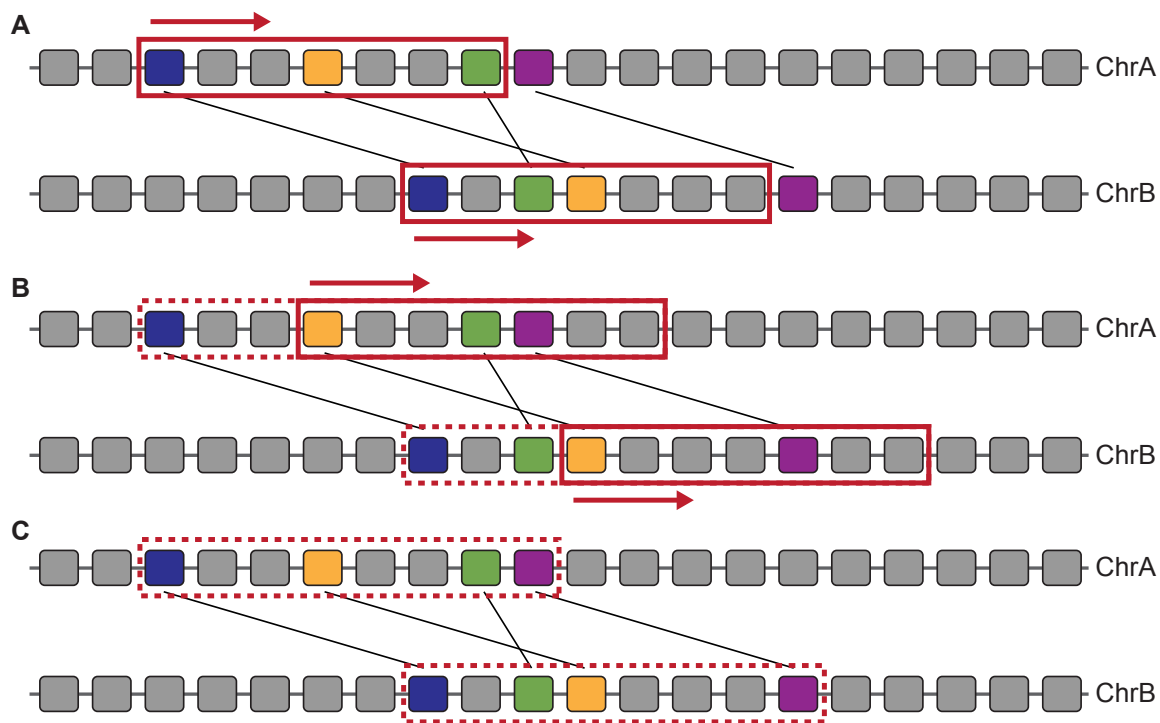


**FIGURE 4.1:** The PIP-based pipeline that populates the Synteny Database.

## 4.1 Methods

Given a set of paralogous gene groups in a primary genome with the members of each group co-orthologous to a single gene in an outgroup genome, we wish to look for regions of conserved synteny among paralogous chromosome segments within the primary genome and between the primary and outgroup genomes. Similar to the RBH Analysis Pipeline, the Synteny Database is populated using a PIP-based pipeline (see Section 3.1.1). The first stage of the pipeline (Fig. 4.1) populates the system with the gene groups built by the RBH Analysis Pipeline for a particular primary genome/outgroup genome data set. The second stage in the pipeline executes the sliding window analysis.

Given a pair of paralogous genes on chromosomes **A** and **B** in the primary genome, we want to locate other paralogs that are in the same neighborhood (with one near the paralog on **A** and the other near the paralog on **B**). We define the neighborhood



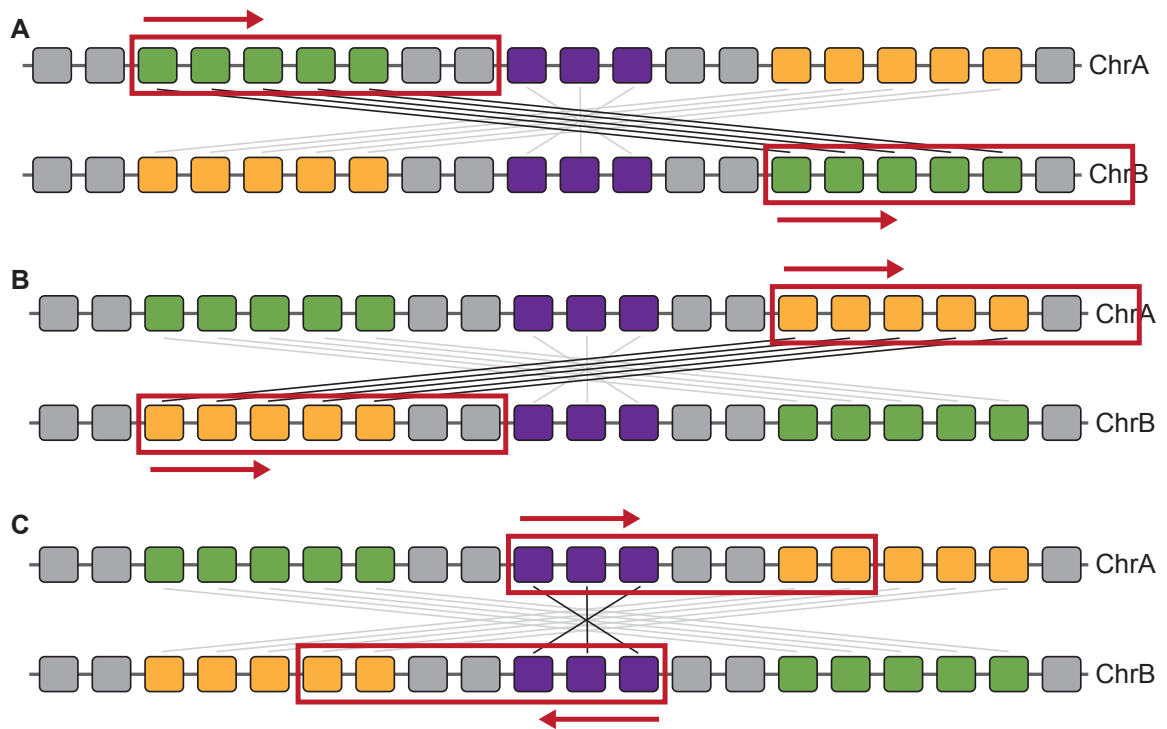
**FIGURE 4.2:** Sliding window analysis. (A) The algorithm begins by placing a pair of windows around a set of paralogs (blue genes) and it begins scanning forward for additional paralogs within the bounds of the two windows. (B) When an additional pair of paralogs are found (orange genes), the windows are advanced and the search continues. (C) If the search reaches the tail of either window without finding another pair of paralogs then the syntenic cluster is closed and recorded.

by placing a pair of *windows* of a particular size around our paralogs of interest, where the size of the window is measured in numbers of contiguous genes (Fig. 4.2). In detail, the algorithm starts by comparing the first and second chromosomes of the primary genome, which we refer to as chromosomes **A** and **B**, respectively. It places the first window on the first gene of chromosome **A** and moves this window until it finds a pair of genes, one on each of the two chromosomes, that are members of the same paralogy group. It then places the second window at the starting location of the gene on chromosome **B** and marks the start of a syntenic cluster (Fig. 4.2A). The software then continues to search for paralogous genes located within the space bounded by the two windows. If another pair is found, the windows are advanced to the starting positions of the new pair of paralogous genes and the search continues (Fig. 4.2B). 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 (Fig. 4.2C). The position of the first window is then reset to the first gene on chromosome **A** that was not part of the last syntenic cluster and the search is restarted. This gene may be located within the same genomic region as the previous syntenic cluster, although the corresponding paralogs on chromosome **B** will be located on a different genomic segment. The analysis pipeline continues this process until all paralogous genes on chromosomes **A** and **B** have been examined.



To identify conserved syntenic areas where the order of the genes has been inverted between two chromosomes (genes are ordered upstream on one segment and downstream on the corresponding segment), the pipeline restarts the search again and now runs the two windows in opposing directions, again recording found clusters. 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 paralogs.

Pseudocode for the sliding window analysis is available in Appendix C. As described above, a new syntenic cluster is always seeded with an initial pair of orthologs or paralogs (that mark the starting position of the sliding windows) and additional pairs of genes may be added to the cluster as the sliding windows advance. If we consider  $n$  to be the number of pairs of paralogs or orthologs, then the worst case execution time occurs when there is no conservation of synteny. In this case, for each pair of genes on the chromosomes being examined, the length of the window will be searched, and having found no additional syntenic genes, the window will reset to the first pair of genes to occur after the initial seeds of the cluster and the search will continue. So, the algorithm would search the length of the window (which has a maximum length of  $n$ ) for each of the  $n$  pairs of genes, giving an execution time on the order of  $O(n^2)$ . In practice, the algorithm executes below this limit as the window size is much smaller than  $n$  and the number of orthologs or paralogs to examine is no more than a few tens of thousand.



**FIGURE 4.3:** Syntenic cluster detection. (A) Detection of syntenically conserved (green) genes. (B) The orange syntenic genes will not be detected along with the green genes as they have been transposed on chromosome B. (C) Detecting inverted segments of genes requires the sliding windows to be run in opposite directions, therefore, the purple, syntenic genes will not be detected along with the green or orange genes.

The sliding window algorithm is able to detect three types of architectural features in the genome. First, it is able to detect genes simply syntenic to one another (Fig. 4.3A): the green genes on chromosome A are all paralogous to the green genes on chromosome B. As the algorithm searches forward in the sliding window, it will detect each additional pair of paralogous green genes and move the window forward. When the window reaches the first yellow gene, however, it will not add the yellow paralogs to the cluster since the corresponding genes on chromosome B fall before the start of the cluster – their positions have been transposed. This situation is remedied, however, after the cluster is closed and the algorithm resets the position of the window to the first yellow gene on chromosome A (Fig. 4.3B). Now, the corresponding window will be placed at the first yellow gene on chromosome B, allowing the detection of this transposition. The green and yellow sets of paralogs will be detected as two distinct clusters when the algorithm has completed examining all paralogs on chromosomes A and B.

A third type of feature the algorithm detects is an inversion of genes between two chromosomes (Fig. 4.3C). The algorithm detects these clusters by running the two windows in opposite directions on chromosomes A and B.

The next stage of the pipeline (Fig. 4.1, green) merges clusters detected in the previous stage that occupy areas on the chromosome within a sliding window's length of one another. Given the green, orange, and purple clusters in Figure 4.3, this stage of the pipeline would merge all three into a single syntenic cluster. The membership

of these *subclusters* is recorded by the pipeline and is later utilized by the web-based rendering routines when visualizing syntenic clusters (Sec. 4.1).

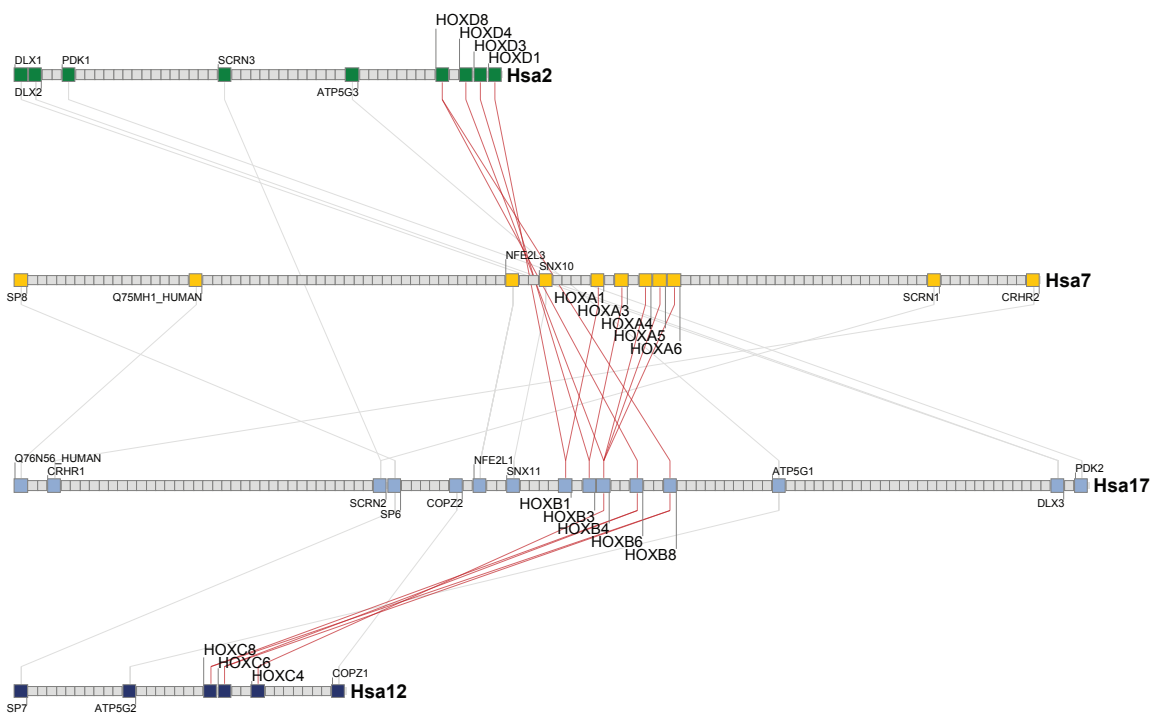
Following the merge operation, a number of housekeeping stages are executed that first, systematically store and index the clusters, the genes present on the clusters, and the paralogy links between the genes. Next, all of the detected clusters are compared to find clusters that overlap physically on the same chromosome and to find clusters that contain the same genes. While some of this data is utilized in the user interface, most of it is used in the next analysis stage in order to generate composite clusters.

### Composite Clusters

Due to the nature of the sliding window analysis, the pipeline discovers conserved syntenic regions in a pairwise fashion. One effect of this strategy is that two or more logical clusters can overlap in the same physical space on a single chromosome. For example, if in a hypothetical genome a single region of chromosome 2 has genes that are paralogous to genes on chromosome 10, and those same genes are also paralogous to genes on chromosome 12, the Synteny Database reports four rather than three clusters – one pair representing the conservation between chromosomes 2 and 10 and a second pair representing the conserved regions on chromosomes 2 and 12. We refer to the first cluster on chromosome 2 as **A** and its paralogous region on chromosome 10 as **a**; similarly, we refer to the second cluster on chromosome 2 as **B** and its paralogous partner on chromosome 12 as **b**. **A** and **B** occupy the same overlapping

physical space on chromosome 2 and contain some or all of the same gene members. This is in contrast to clusters that occupy the same space on a particular chromosome but have no overlapping gene members (and hence are not part of a larger conserved region).

While it is often useful to consider pairwise clusters, considering larger conserved regions can also be important. To accomplish this task, the analysis pipeline consolidates cluster pairs to create composite clusters (Fig. 4.1, blue). In the example above, we would like to consolidate the four regions **A**, **a**, **B**, and **b** into three regions: **A/B**, **a**, and **b**. This is accomplished by taking each cluster in the system and finding all other clusters that share at least one gene with it (i.e. that overlap on the same physical chromosome space). Once the system has assembled a list of clusters, it then tries each permutation of the clusters looking at the intersection of member genes. If, for example, cluster **A** shares a common gene with cluster **B**, and cluster **B** shares a common gene with cluster **C**, then the system will check to see if clusters **A**, **B**, and **C** all have at least one gene in common. **A**, **B**, and **C** all share the same physical space on a single chromosome and have paralogous partner regions, **a**, **b**, and **c** somewhere else in the genome. The system will continue to check for smaller numbers of paralogons next, examining if clusters **A** and **B**, **B** and **C**, or **A** and **C** have at least one gene in common. If it finds common genes, then the pipeline records a composite cluster. The human *HOX* cluster genes in Figure 4.4 are a nice example of a composite cluster formed by this process. In this example, three cluster pairs



**FIGURE 4.4:** The composite  $HOXB_4$  paralogue syntenic cluster showing paralogous regions on human chromosomes 2 (Hsa2), Hsa7, Hsa17 and Hsa12. This composite cluster was generated from three pairs of clusters: Hsa17/Hsa2, Hsa17/Hsa7, and Hsa17/Hsa12. Results were generated by the Synteny Database using a 50-gene sliding window and the visualization of the cluster was generated by the web-based user interface. Lines connecting paralogous HOX cluster genes are red.

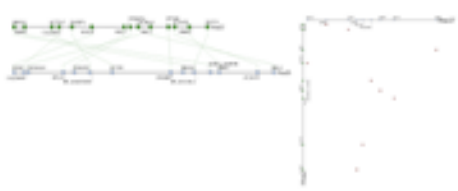
were merged with **A/a** representing the regions on Hsa17/Hsa2, **B/b** representing Hsa17/Hsa7, and **C/c** representing Hsa17/Hsa2.

To identify conserved syntenies between species, the system performs the entire analysis again, this time considering orthologs and comparing each chromosome of the primary genome to every chromosome of the outgroup genome. We experimented with four window sizes, 25, 50, 100, and 200 genes in length.

Synteny Database
Synteny Tools

[circle plots](#) | [dotplots](#) | [view rbb data](#)

› What is the Synteny Database?



Composite Clusters

- [arntl1a: #16049](#); 2 shared [genes](#)  
Dre25, Dre12, Dre18, Dre3
- [arntl1a: #16690](#); 1 shared [gene](#)  
Dre25, Dre18, Dre7

Paralogous Pairwise Clusters

- [arntl1a: #72760](#); 3 [gene pairs](#)  
Dre25 -- Dre18
- [arntl1a: #74838](#); 4 [gene pairs](#)  
Dre25 -- Dre7

Orthologous Pairwise Clusters

- [arntl1a: #138905](#); 11 [gene pairs](#)  
Dre25 -- Hsa11

**Enter a gene to search for:**

  
(e.g. 'aldh1a2', 'ENSG00000068793', or 'HOXD\*')

**Select additional options:**

Clustering algorithm  
sliding window size:  genes

**Source genome:**

**Outgroup:**

**Variant:**

FIGURE 4.5: Synteny Database Web Interface.

## User Interface

The data generated by the analysis pipeline is coupled with a web-based interface to provide a searchable set of conserved syntenic regions to the researcher (Fig. 4.5). The web-based interface allows the user to choose a primary and outgroup genome and submit a gene name; it then returns a list of paralogous, orthologous, and composite clusters. If the user chooses to view one of the clusters, the system will draw images of the cluster in a fully scaled view of the chromosome segment, in a scale-free view showing gene order, and as a gene homology matrix (defined in Chapter III). The code to draw these images is modular and efficient – first generating an abstracted, unit-length version of the image, which is cached as a binary object; then drawing all three types of images from the abstracted object whenever necessary in a user-specified scale and format (either raster or vector). The web-based system also exports gene membership lists for the clusters in Microsoft Excel format, allows the user to zoom in to subsets of the cluster, and to manipulate the size of the images among other features.

### 4.1.1 Verification

Figure 4.4 shows an example of the output of the Synteny Database pipeline. It displays the four paralogous regions in the human genome that contain HOX cluster genes (described in 1.2) and was generated using amphioxus as an outgroup. An



analysis of this syntenic HOX cluster reveals the strengths and limitations of the Synteny Database. First, the pipeline identified all four HOX clusters, including several additional neighboring paralogs, and was able to combine them into a single composite cluster. These results are consistent with the recent work by [102], including the identification of the syntenically conserved neighboring DLX and NFE2L3 gene families in Fig. 4.4, as well as the identification of the MPP, IGFBP, SLC4A, and UPP gene families in additional nearby clusters (not shown).

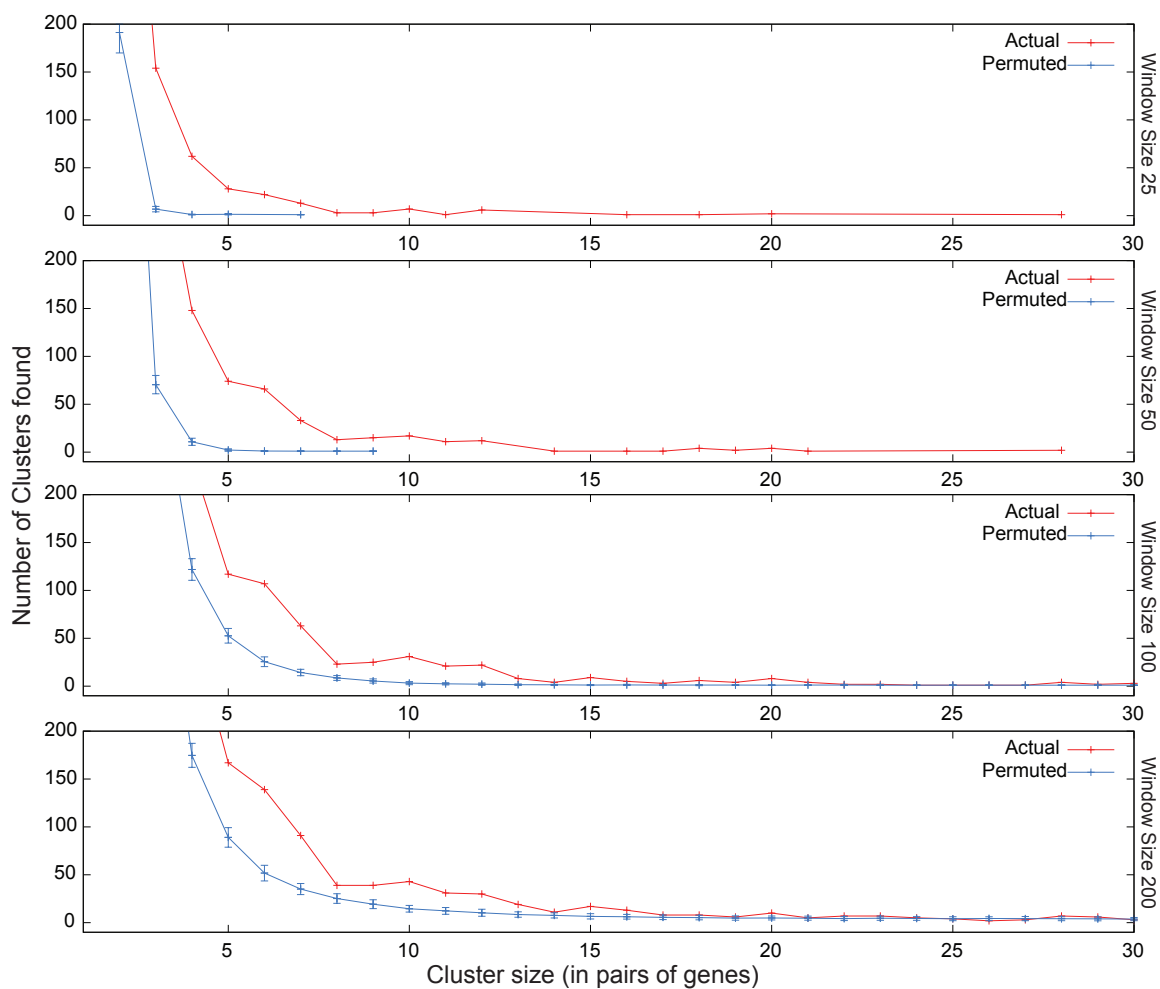
Not all of the HOX genes were identified by the Synteny Database, however, for two primary reasons. First, the choice of an outgroup genome strongly influences the composition of paralogy groups. If an ortholog of the members of a paralogy group has diverged significantly, or has been lost in the outgroup, then the analysis pipeline will not be able to anchor one or more members of the paralogy group making them unavailable for the Synteny Database to cluster into conserved regions. For example, all four paralogous HOX regions in Fig. 4.4 are missing genes posterior to *HOX8*. Although all the HOX genes are picked up by the BLAST search in the human primary genome, the *HOX9* through *HOX13* genes in amphioxus are highly divergent [6, 48] and they are not picked up by the outgroup BLAST analysis. If we instead consider using the urochordate *Ciona intestinalis* as outgroup for the human HOX cluster, the signal of syntenic conservation is even weaker since *Ciona* possesses only nine HOX genes located on two chromosomes, including several rearrangements of those genes [48]. In addition, recent work has show much weaker conservation

of synteny in *Ciona* relative to amphioxus [86]. The second major reason why the Synteny Database did not identify some HOX genes is due rate asymmetry in some of the HOX genes, which we previously described in Section 3.2.2. Often, the solution to this problem is to use a different, more closely related outgroup genome; unfortunately, not all desirable outgroup organisms have been fully sequenced. As analysis of the HOX clusters demonstrates, however, and as the case studies below will confirm, the Synteny Database does indeed detect a wide array of syntenic conservation, including paralogous regions within genomes, orthologous regions between genomes, chromosome inversions, and ohnologs gone missing.

#### 4.1.2 Permutation Analysis

It is important to question whether paralogons (segments of chromosomes conserved since the last WGD) defined by the Synteny Database are the result of a large-scale duplication event or are simply chance associations mistakenly detected by our sliding window analysis. To examine this question, we attempted to approximate the underlying distribution of syntenic clusters using permutation analysis – repeatedly randomizing the genomic locations of our paralogous genes and re-executing our clustering algorithm 100 times.

Figure 4.6 plots the results of the analysis for the human genome using amphioxus as an outgroup. For each sliding window length, we plotted with error bars the average number of clusters of a particular size that were detected after randomizing our data



**FIGURE 4.6:** A permutation analysis of all syntenic clusters that the Synteny Database found in the human genome using amphioxus as an 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.

(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. If the sliding window analysis was simply detecting chance associations between paralogs or orthologs, then we would expect the size and number of clusters detected by the algorithm in the permuted data to be roughly equivalent to the size and number of clusters in the actual data.

The results showed that with all window sizes, the vast majority of clusters found from the randomized data were small and contained few gene pairs. For a 25-gene window, using the randomized data, 97.9% of the clusters found had only one pair of genes. Likewise, for a 50-gene window, 95.5% of clusters had only one pair of genes; for a 100-gene window 97.8% of clusters contained two or less gene pairs; for a 200-gene window 96.9% of clusters contained three or fewer gene pairs. As the length of the gene window increased, the pipeline did generate larger clusters from the randomized data, and with a window size of 200 genes the simulation generated clusters from randomized data that were as large as any actual cluster produced in the original analysis. In all cases, larger clusters, and more of them, were found in our actual data compared to the permuted data.

We can then consider the question: are the size of the clusters found in the human genome (using amphioxus as an outgroup) *significantly* larger than those that would be found by chance alone? A t-test showed 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  $1.7 \times 10^{-126}$ ,  $1.0 \times 10^{-239}$ ,  $2.8 \times 10^{-207}$ , and  $8.6 \times 10^{-41}$  for window sizes of 25, 50, 100, and 200 genes, respectively) and we can reject the hypothesis that the clusters detected by the Synteny Database were chance occurrences. Based on our permutation analysis, we conclude that analyses should usually use the 50 or 100-gene windows for most reliable results.

### 4.1.3 Data Sources

For the following case studies, Ensembl [12, 55] 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/>).

To further evaluate the utility and efficacy of the Synteny Database, as well as the underlying analysis pipelines used to populate it, we used the Database to help determine the evolutionary history of two problematic gene families; results of these two analyses follow.

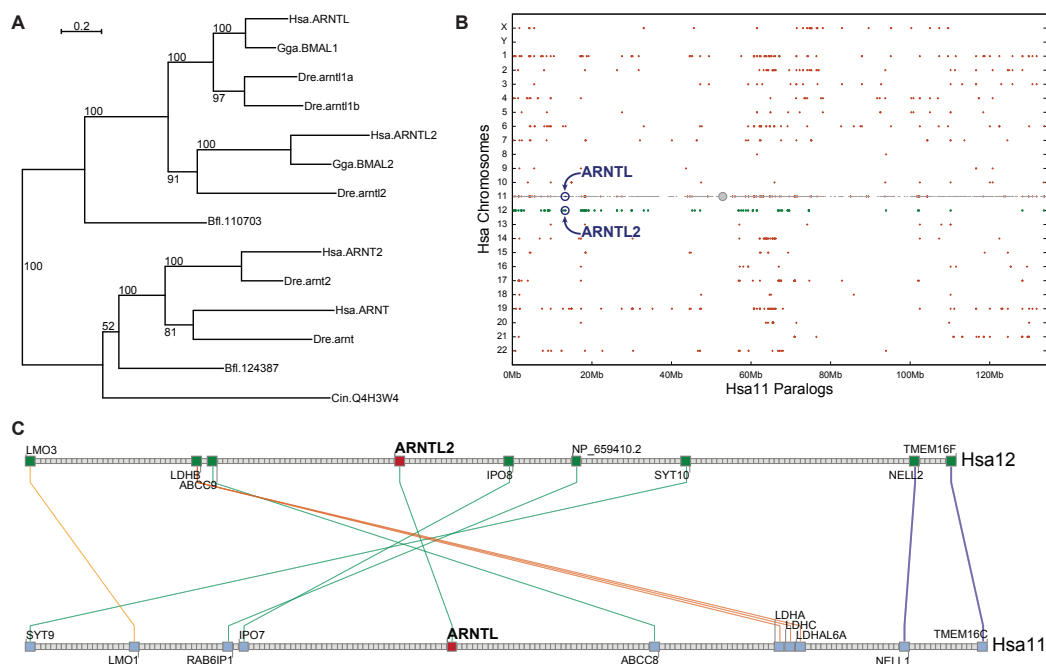
## 4.2 Case Study: The ARNTL Gene Family

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 widely conserved with homologs in protostomes and deuterostomes. ARNTL, working together with CLOCK, activates PER1 to regulate the circadian clock, a system that provides daily periodicity for biochemical, physiological, and behavioral activities [47, 37, 77]. We tested the ability of the RBH 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 searched 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 positions of the ARNTL genes, as well as the syntenic neighborhood around those genes, to reflect the existence of the R1 and R2 duplication events in the vertebrate lineages and the R3 duplication event in the teleost fish. We therefore identified ARNTL orthologs and paralogs in each of these species and use the Synteny Database in two steps to search for evidence of conserved synteny supporting the duplication events, first showing orthologous conservation between species for the

ARNTL genes, and second, showing paralogous conservation within a species. This evidence will allow us to confirm or reject our orthology and paralogy assignments.

### 4.2.1 ARNTL Paralogs in the Human Genome

We can examine the origins of ARNTL paralogs in three steps: the output from the RBH Analysis Pipeline, a comparison of those results to phylogenetic analysis, and inferences obtained from the Synteny Database. According to the results of the RBH Analysis Pipeline, *ARNTL*, located on human chromosome 11 (Hsa11), has a single paralog in the human genome, *ARNTL2*, on chromosome 12 (Hsa12) [42]. Because the genome assembly of *Ciona intestinalis* [92] does not contain an ARNTL ortholog, the RBH 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 [81, 13], has an ortholog of human *ARNT* and *ARNT2* as well as an ortholog of *ARNTL* and *ARNTL2* (Fig. 4.7A). This analysis emphasizes the problem illustrated by Figure 1.6: reciprocal BLAST procedures can assign false orthologies in the case of lost gene duplicates. Because the current genome assembly of *Ciona* lacks an ortholog of the ARNTL genes, we will use the amphioxus genome as an outgroup to search for syntenic conservation among the human ARNTL paralogs.



**FIGURE 4.7:** 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 Phylml [39] 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 paralogs 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).



### 4.2.2 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 gray disc shows the index chromosome 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. 4.7B) 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*,

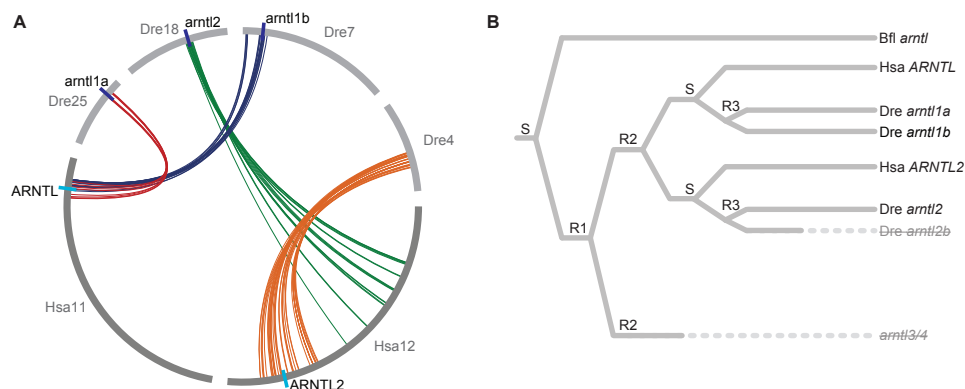
bordered on one side by *LMO3/LMO1* and on the other by *TMEM16F/TMEM16C* using a sliding window size of 50. To evaluate the relationship of window size and shared gene pairs, we performed a permutation analysis, described in the Materials and Methods section. In brief, with longer windows, the likelihood of finding a pair of orthologs that are syntenic in two species will increase solely by chance rather than being a true, evolutionarily conserved synteny. 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. The central portion of the cluster contains an elegant inversion of several pairs of genes, with the *ARNTL/ARNTL2* paralogs serving as the pivot (Fig. 4.7C). 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. The lines on the gene trace make chromosome rearrangements readily apparent.

### 4.2.3 ARNTL Paralogs in Teleost Fish

The hypothesis that teleost fish experienced a third genome duplication after splitting from the lineage that led to humans [7, 85, 104, 51, 74], 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 RBH Analysis Pipeline results to identify the zebrafish orthologs of human *ARNTL* and *ARNTL2* and then used the Synteny Database to search for conserved synteny in the regions surrounding those orthologs. The ortholog circle plot of Figure 4.8A 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 connecting those lines join orthologous gene pairs positioned relative to their location on the chromosome. The 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 RBH 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* [21, 50]. 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



**FIGURE 4.8:** Evolutionary relationships between 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) 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.

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 about these results is: why did the pipeline anchor zebrafish *arntl2* to a human ortholog different from the published conclusion? To answer this question, we must recall how the analysis pipeline works; it first searches for paralogous groups of genes within the primary organism, zebrafish in this case, and then tries to split the groups into different duplication events by anchoring them to their proper ortholog in the non-duplicated outgroup (in this case, human). In principle, we would expect all three zebrafish genes to fall into a single paralogous group that should in turn split into two groups after matching the zebrafish genes with their proper human orthologs. In this case, 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. The *arntl1a* and *arntl1b* genes both found human *ARNTL* as their top BLAST hit and a retro-BLAST of *ARNTL* found *arntl1a* and *arntl1b* as its top two hits, all highly significant alignments. On the other hand, an *arntl2* BLAST search hit *ARNTL* and *ARNTL2* with approximately the same magnitude – quite significant, but not significant enough to differentiate between the two human genes (the *ARNTL* hit has a length of 594 amino acids and 56% identity while the *ARNTL2* hit has a length of 560 amino acids and 53% identity). The pipeline therefore assigned zebrafish *arntl2* to the first human gene it hit causing *arntl2* to group with the wrong human gene, *ARNTL*.

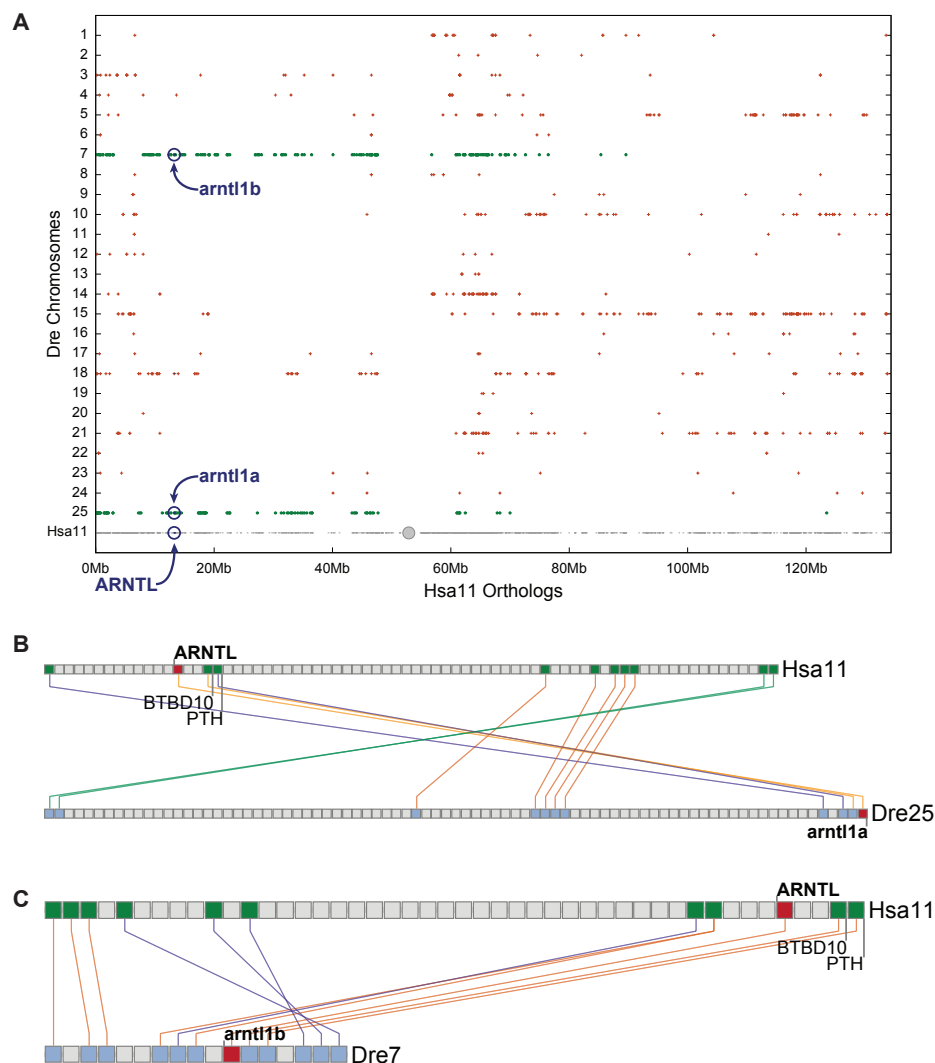
The *arntl2* example highlights an inherent limit to the power of an RBH-based approach. While an RBH analysis is highly desirable in many respects, if members of a paralogy group in the primary genome, or their ortholog in the outgroup have experienced significantly different rates of divergence, then the pipeline can assign a gene to the wrong paralogy group or to the wrong ortholog. 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. In fact, *ARNTL2* has diverged far enough that *ARNTL* is better conserved to zebrafish *arntl2* than is *ARNTL2*. A phylogenetic analysis (Fig. 4.7A)

confirmed the published results and led us to tentatively reject the assignment from the orthology pipeline.

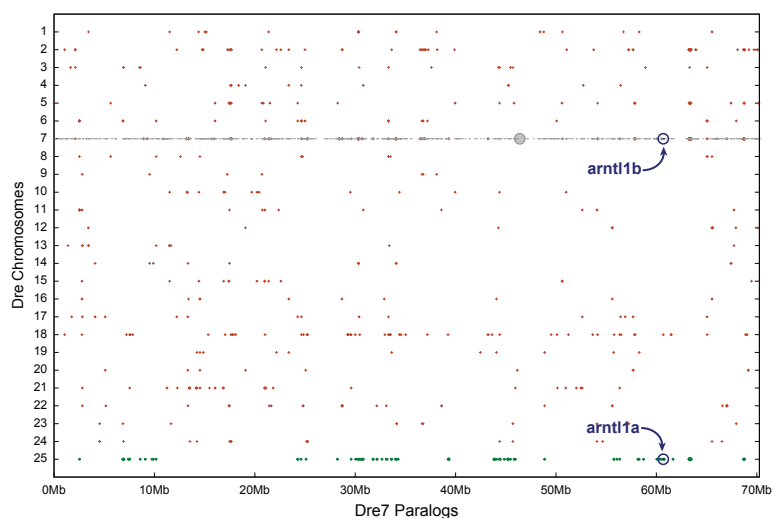
We next sought to use conserved synteny to provide an independent line of evidence not based on sequence similarities.

#### 4.2.4 Orthology and Paralogy of Zebrafish *arntl1* Chromosome Segments

The phylogeny showed that *arntl1a* and *arntl1b*, located on zebrafish chromosomes 25 (Dre25) and 7 (Dre7) respectively, are co-orthologous to the human gene *ARNTL*. An orthology dotplot for Hsa11 clearly showed strong conservation between genes on the short arm of Hsa11 and zebrafish genes on both chromosomes Dre7 and Dre25, with a weaker signal on Dre18 (Fig. 4.9A). The Synteny Database identified conserved regions between Hsa11 and both Dre25 and Dre7; a gene-by-gene comparison (Fig. 4.9B) showed ten pairs of orthologous genes surrounding the *ARNTL/arntl1a* orthologs, including human genes *BTBD10* and *PTH* as very-near neighbors to the *ARNTL* gene. Similarly, the orthologous syntenic cluster associated with zebrafish *arntl1b* has ten pairs of orthologs between Hsa11 and Dre7, once again including human genes *BTBD10* and *PTH* immediately adjacent to *ARNTL* (Fig. 4.9C). Finally, after using the Synteny Database to identify syntenically conserved regions between the zebrafish and human genomes, we could ask whether the indicated regions on



**FIGURE 4.9:** Conserved synteny in *ARNTL* evolution. (A) Dotplot showing the *Danio rerio* (*Dre*) orthologs of Hsa11 genes. Zebrafish orthologs of Hsa11 genes are plotted vertically above the corresponding grey dot on their respective *Dre* chromosome. The short arm of Hsa11 shows strong orthology with *Dre*7 and *Dre*25 (colored green) and weaker orthology with *Dre*18. (B) The *ARNTL* and *arntl1a* orthologous syntenic cluster showing the conserved region between Hsa11 and *Dre*25. The cluster shows eight pairs of orthologs surrounding the *ARNTL* genes (50-gene sliding window). (C) The *ARNTL* and *arntl1b* orthologous syntenic cluster showing ten pairs of orthologs surrounding *ARNTL* and *arntl1b* (50-gene sliding window).



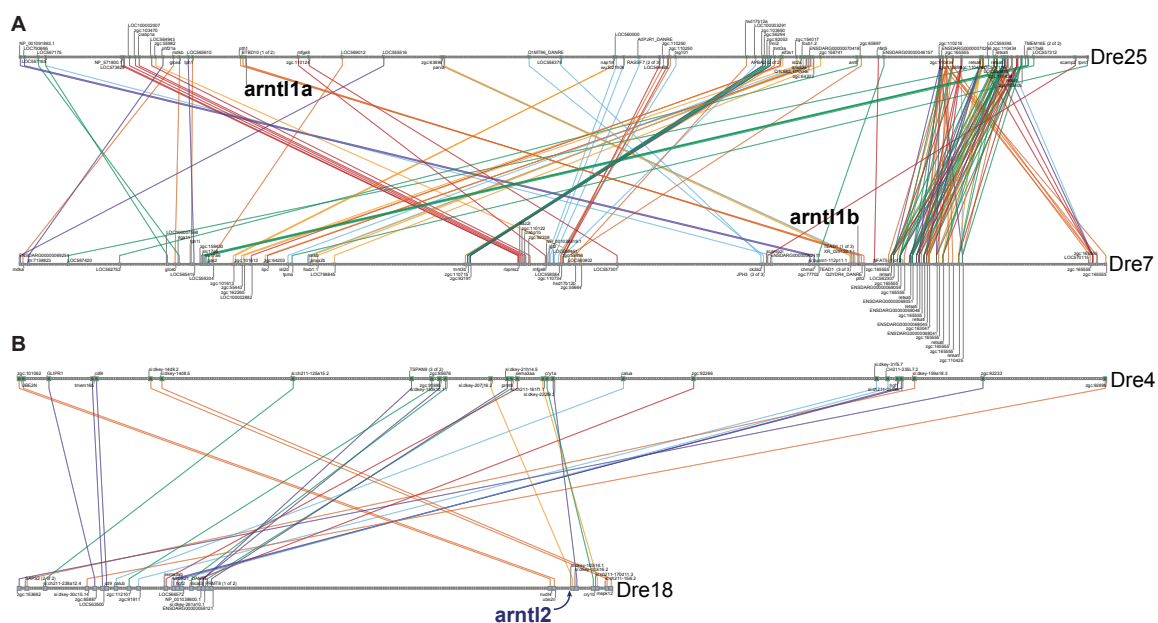
**FIGURE 4.10:** Dre7 paralogy dotplot showing the broadly conserved region between Dre7 (grey) and Dre25 (colored green) containing the *arntl1b* and *arntl1a* paralogs.

Dre7 and Dre25 are conserved as expected under the hypothesis that they are paralogs produced from a third full genome duplication. Examination of the paralogy dotplot for Dre7 (Fig. 4.10) showed conservation with Dre25 across the full length of the chromosome (see Fig. 4.11A for the Synteny Database gene trace). This cluster contains 78 gene pairs including *arntl1a* and *arntl1b*, as well as the directly adjacent paralogs *btbd10* and *pth*. These data provide strong syntenic support indicating that Dre7 and Dre25 are paralogs.

#### 4.2.5 Orthology of Zebrafish *arntl2* Chromosome Segments

The automated pipeline did a good job at finding conserved syntenic regions between the two zebrafish co-orthologs and *ARNTL* – with a pair of conserved genes





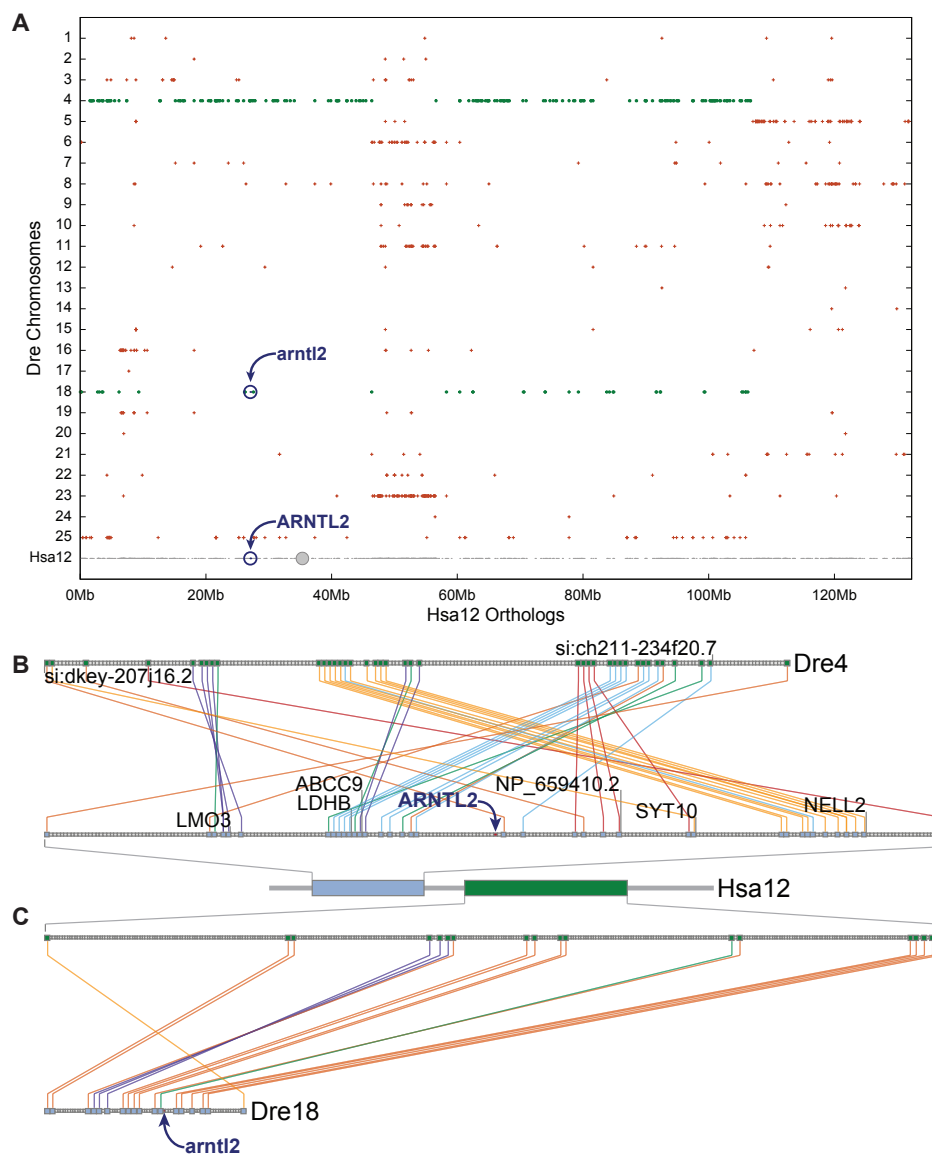
**FIGURE 4.11:** Conserved syntenies for zebrafish *arntl* paralogs. (A) The *arntl1a* and *arntl1b* paralogous syntenic cluster showing a conserved region between Dre7 and Dre25 and discovered using a 100-gene sliding window. (B) The *arntl2* paralogous syntenic cluster between Dre18 and Dre4 discovered using a 200-gene sliding window. The right and left gene groups on Dre18 correspond to regions (i) and (ii) in Figure 4.14. These two regions were discovered as separate clusters when using smaller window sizes.

in a pair of paralogs in the zebrafish that showed strong conservation to their human ortholog and the human chromosome region. For *ARNTL2*, the analysis started again with an orthology dotplot, this time for Hsa12; this automated analysis revealed strong conservation along more than 80% of the length of Dre4 (Fig. 4.12A), 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 paralogs of Hsa12 (Fig. 4.12B). Furthermore, the Synteny Database found a second region on Hsa12 that is 12Mb distant from *ARNTL2* that shows strong syntenic conservation with Dre18. The Dre18 half of the cluster tightly spans the region containing the zebrafish *arntl2* ortholog (Fig. 4.12C). The Dre4/Hsa12 conserved region contains 38 pairs of orthologous genes while the Dre18/Hsa12 cluster contains 18 orthologous gene pairs providing strong support. This set of gene traces from the Synteny Database poses the question: if Dre4 and Dre18 are paralogs 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. 4.7C), we find that several paralogs within that region of Hsa11 connect it to the Hsa12/Dre18

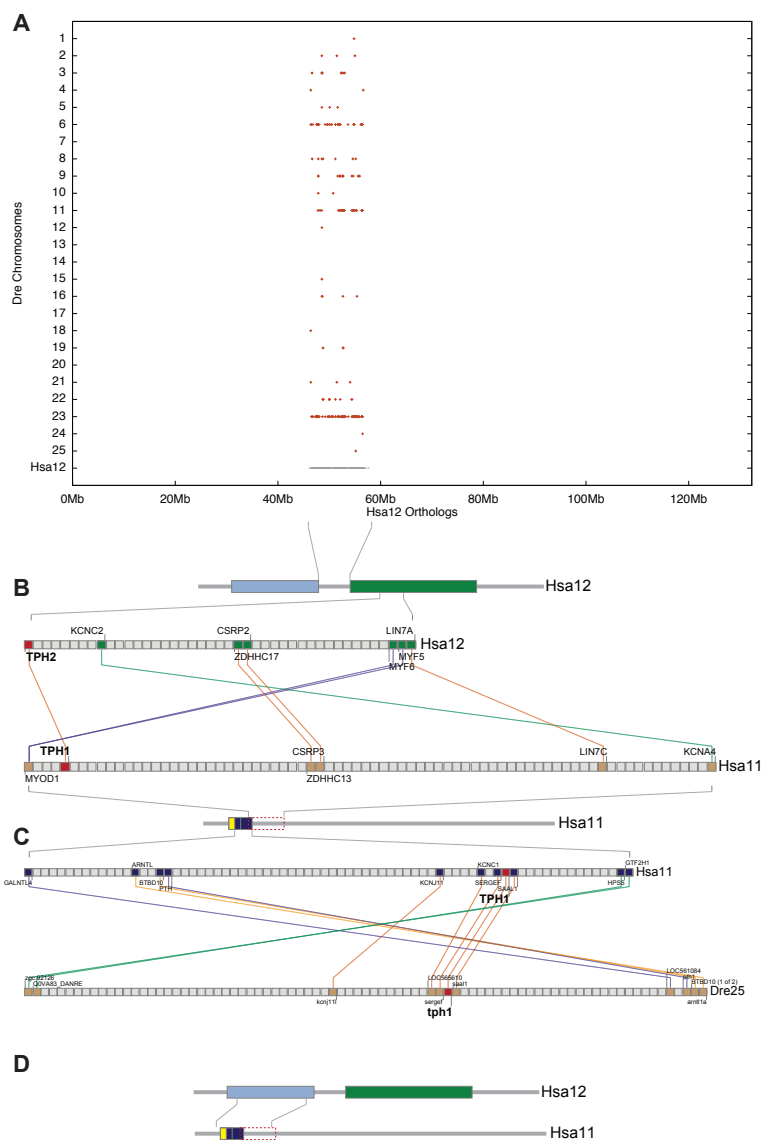
region, including *TPH1/TPH2*, and *CSRP3/CSRP2* on Hsa11 and Hsa12 respectively. Given two regions on Hsa12, one that is orthologous to Dre4 and the other orthologous to Dre18, with both of those regions on Hsa12 paralogous to Hsa11, the architecture suggests that an inversion on ancestral Hsa12 must have occurred that moved *ARNTL2* relative to other genes after the lineage leading to humans split from the lineage leading to zebrafish (see Fig. 4.13 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 [84]. 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. 4.12B) depending on the layout of the ancestral chromosome prior to the transposition event.

#### 4.2.6 Paralogy of Zebrafish *arntl2* Chromosome Segments

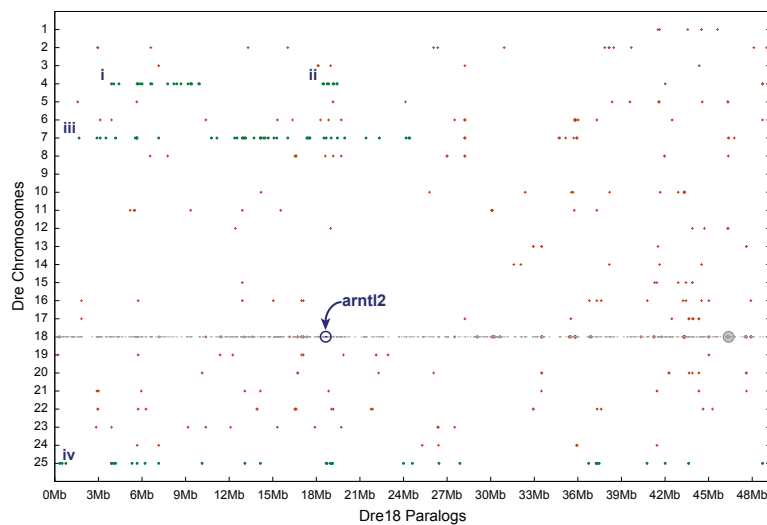
Having established good syntenic support showing co-orthologous regions between zebrafish chromosomes 4 and 18 and Hsa12, the last task is to test for paralogy of Dre4 and Dre18. Again, the regions corresponding to Hsa12 in the *ARNTL2* part of this case study are not as clear as those corresponding to Hsa11 in the *ARNTL*-related portion. The paralogy dotplot of Dre18 versus other zebrafish chromosomes shows only two tightly conserved regions containing paralogs on Dre4 (colored and marked *i* and *ii* in Fig. 4.14A), with several genes in region *ii* having paralogs quite close to *arntl2*. The gene trace in Fig. 4.11D shows these regions in greater detail.



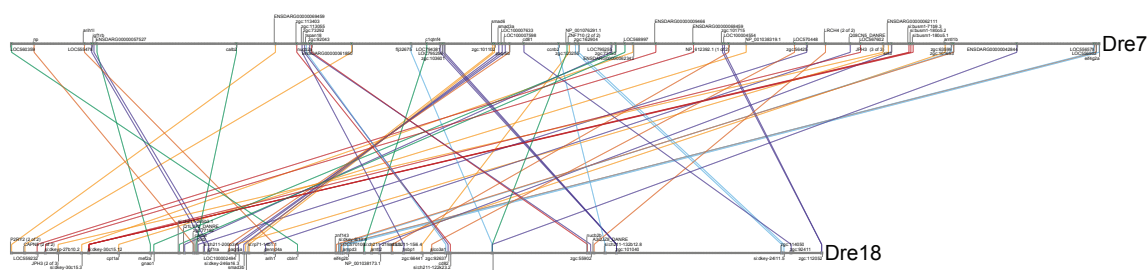
**FIGURE 4.12:** Conserved synteny for *ARNTL* genes. (A) Hsa12 orthology dot-plot against *Danio rerio*. Hsa12 shows orthology with Dre4 (green), Dre18 (green), and weakly with Dre25. (B) The *ARNTL2* orthologous syntenic cluster showing strong syntenic conservation between Hsa12 and Dre4. Several genes that are part of the original Hsa11/Hsa12 paralogous cluster (Fig. 4.7C) 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. 4.8A) due to an inversion occurring after the zebrafish and human lineages diverged.



**FIGURE 4.13:** Support for an inversion on Hsa12. (A) Dotplot showing orthologs from the region between the two Hsa12/Dre4 (Fig. 4.12B), Hsa12/Dre18 (Fig. 4.12C) clusters in zebrafish. This region is not related to Dre4 or Dre18. (B) A cluster on Hsa12 that overlaps the Hsa12/Dre4 cluster (green rectangle) and is paralogous to a region on Hsa11 (dotted red rectangle) that overlaps the Hsa11/Dre25 cluster (dark blue rectangle). (C) The Hsa11/Dre25 cluster (dark blue box, Fig. 4.9B) overlaps with the Hsa11 portion of the cluster from B. The Hsa11/Dre7 cluster (Fig. 4.9C) is shown as a yellow rectangle for reference. The *TPH1/TPH2* human paralogs as well as the zebrafish *tph1* ortholog (red genes) link the clusters from B and C. (D) The *ARNTL/ARNTL2* paralogous cluster (Fig. 4.7C) overlaps the Hsa11/Hsa12 cluster from B (dotted red), Hsa11/Dre7 cluster (yellow), the Hsa11/Dre25 cluster (dark blue), and the Hsa12/Dre4 cluster (light blue).

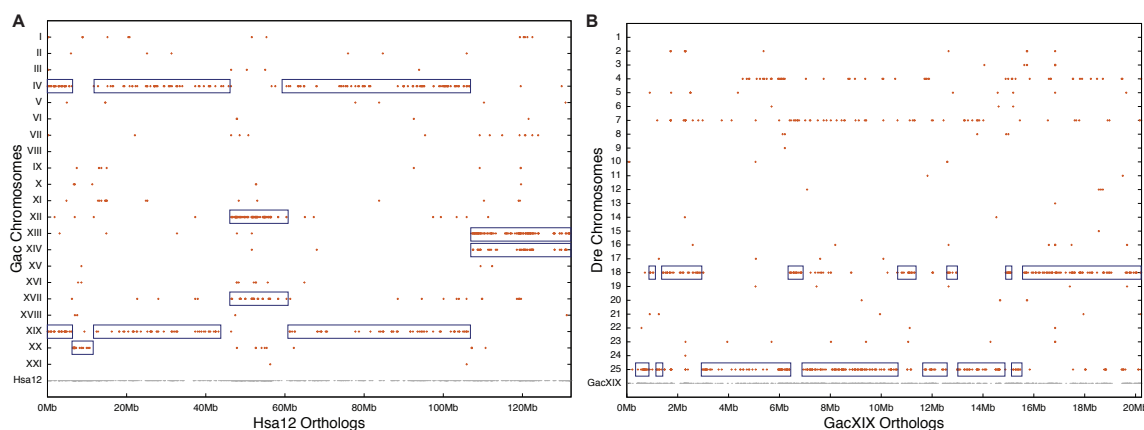


**FIGURE 4.14:** The Dre18 parity dotplot shows four major conserved regions: two on Dre4 marked (i) and (ii), an additional larger region on Dre7, marked (iii), and a region on Dre25 marked (iv). The conserved region on Dre7 represents ohnologs gone missing from human chromosome 12, the site of *ARNTL2*; the region on Dre25 represents a post-R3 translocation from Dre18.



**FIGURE 4.15:** A syntenic cluster between Dre18 and Dre7. The conserved region on Dre7 represents ohnologs gone missing from human chromosome 12, the site of *ARNTL2*. This cluster corresponds to region (iii) in Figure 4.14.

Another region of conserved synteny also appears in the Dre18 paralogy dotplot along Dre7 (colored and marked *iii* in Figure 4.14). Figure 4.15 shows this region in greater detail and it is hard to dismiss the 50 paralogous gene pairs in Dre7 as noise. At least two possible hypotheses might explain the Dre18 paralogs located on Dre7. First, R3 might have resulted in *arntl2* ohnologs on two ancestral teleost chromosomes that we will call AncA and AncB. A translocation event could have moved a portion of AncA onto another chromosome, AncC. The modern descendants of AncA, AncB, and AncC would then be Dre4, Dre18, and Dre7, respectively. If this were the case, the paralogous genes on Dre18, Dre4, and the translocated genes on Dre7 should all have orthologs on human chromosome 12. The orthology dotplot for Hsa12, however (Fig. 4.12A), shows few orthologs between Hsa12 and Dre7; these data make the possibility of an ancient translocation highly unlikely. An alternative hypothesis would explain cluster *iii* of Figure 4.14 as ohnologs gone missing in the human lineage. As discussed in the introduction, if the ancestral chromosome that became today's Hsa11 experienced a significant number of gene losses after splitting from the lineage that led to the zebrafish, then the pipeline would assign zebrafish genes that were orthologous to now lost genes on Hsa11 to their most closely related ancient paralogs on Hsa12. Therefore, zebrafish genes from Dre4 and Dre18 that were orthologous to genes now lost on Hsa12 might erroneously appear in the paralogy dotplot for Dre7. A similar situation exists for paralogs on Dre18 and Dre25 (marked *iv* in Fig. 4.14) and the same two hypotheses can explain the presence of



**FIGURE 4.16:** Conserved synteny in stickleback. (A) Human chromosome 12 (Hsa12) orthologous dotplot against the stickleback. Most of the length of Hsa12 is orthologous to *Gasterosteus aculeatus* (Gac) linkage groups IV and XIX. (B) Orthology dotplot of stickleback linkage group XIX against zebrafish. The dotplot shows that several portions of zebrafish chromosome 18 (Dre18) have been translocated to Dre25 (boxed regions) since the divergence of the zebrafish and stickleback lineages.

paralogs on Dre25. Figure 4.12A shows a number of Hsa12 orthologs located on Dre25, suggesting translocations between Dre18 and Dre25. Figure 4.16 shows how the Synteny Database can help resolve such questions: GacIV and GacXIX are the stickleback paralogs of Hsa12 (Fig. 4.16A) and GacXIX is paralogous to portions of both Dre18 and Dre25 in the zebrafish genome (Fig. 4.16B). These dotplots confirm a translocation between ancestral Dre18 and Dre25 followed by several inversions since the stickleback and zebrafish lineages diverged.

In summary, analysis using the Synteny Database suggests the following model for the origin of the zebrafish and mammalian *ARNTL*-related genes (Fig. 4.8B). 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 near either *si:dkey-207j16.2* or *si:ch211-234f20.7* on Dre4 as inferred from orthologies of neighboring genes. These results are consistent with the recent work by [115].

#### 4.2.7 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 us to make several observations. First, the RBH 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. This limit stems from the RBH algorithm's use of protein sequence alignments, via BLAST, to associate genes. Measuring evolutionary relatedness by the statistical significance of sequence alignments cannot account for large changes in rates of divergence; in this case, the tetrapod *ARNTL2* genes are diverging more rapidly than the *ARNTL* genes (Fig. 4.7A), which appears to have caused the pipeline to assign all three zebrafish

genes as orthologs of *ARNTL*. 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 RBH 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.

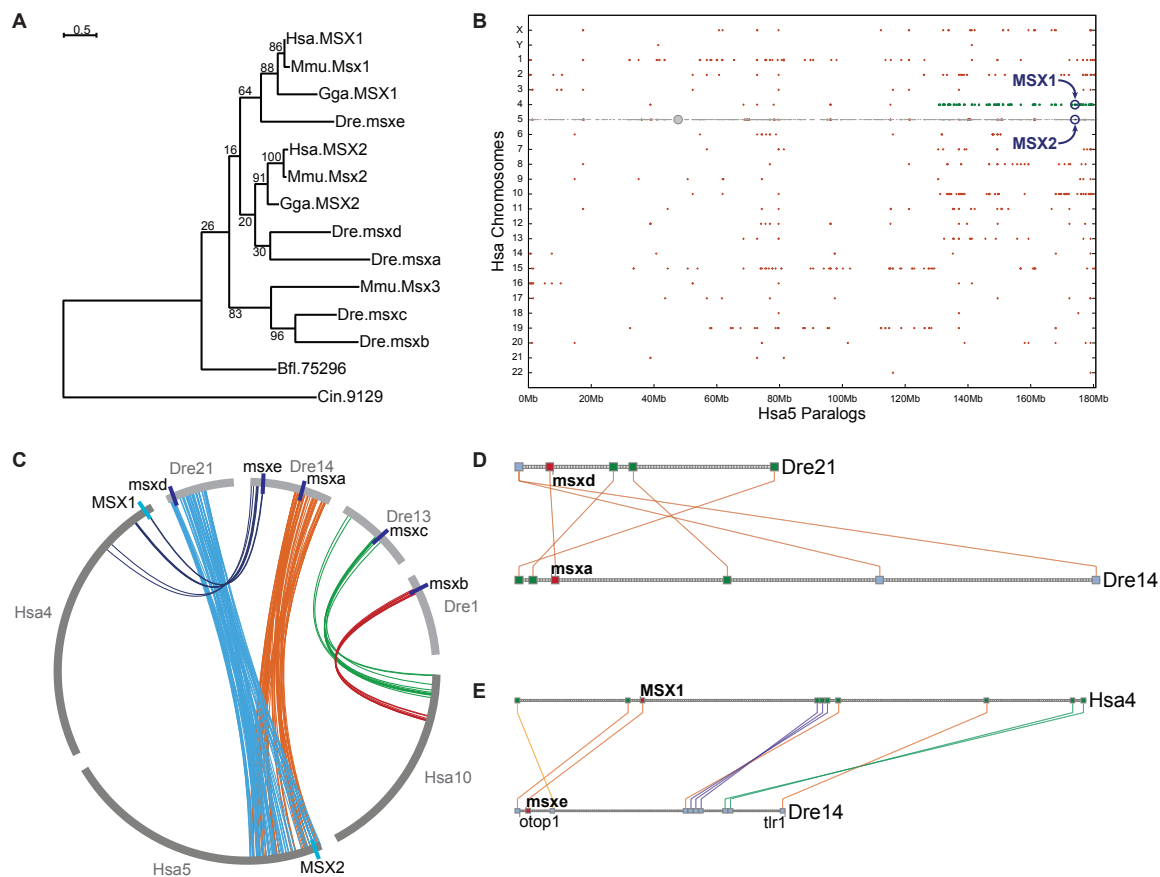
### 4.3 Case Study: The MSX Gene Family

The following section uses the Synteny Database to explore a particularly problematic gene family with difficult RBH orthology assignments and ambiguous phylogenetic trees that have led to controversial orthology assignments in the literature.

The vertebrate muscle segment homeobox (MSX) gene family members act as transcriptional repressors that help pattern limb and craniofacial development [24, 30, 88]. The MSX gene family is ancient, with homologs in *Drosophila* and other protostomes (e.g. insects, worms, molluscs) as well as in the radiata, including the sea anemone, *Nematostella* [91]. Stem chordates likely had a single MSX gene as do the genomes of the urochordate *Ciona intestinalis* and the cephalochordate amphioxus today (genes *ENSCING00000009129* and *75296*, respectively). Humans have two paralogs, *MSX1* and *MSX2*, and mouse has in addition a third copy, *Msx3* [96]. The zebrafish genome has five MSX paralogs called *msxa*, *msxb*, *msxc*, *msxd*, and *msxe*. The zebrafish paralogs were initially characterized by phylogenetic and functional analysis [30] and were re-examined manually for syntenic conservation of the regions surrounding the human, mouse, and zebrafish MSX genes [83]. Ekker's phylogenetic analysis found that the vertebrate *MSX1* and *MSX2* genes formed distinct monophyletic groups (the *MSX1* genes from the different species grouped into a single subtree indicating orthology among them, and similarly for *MSX2*) and in some analyses he found that *msxb* and *msxc* were most related to the *Msx3* gene in mouse.

Because the analysis was unable to determine orthology definitively for any of the zebrafish MSX genes, the nomenclature wisely used letters rather than numbers, which might prematurely suggest orthology where none exists. Likewise, while functional studies suggested that the expression patterns of *msxb* and *msxc* were again related to *Msx3* expression, overlapping expression patterns of various paralogs in mammals and in zebrafish made it difficult to assign orthologies with confidence. (Parenthetically, expression patterns are not usually useful to assign orthology, except as specifically identified characters in the framework of a careful phylogenetic analysis, because they are gained and lost rather readily, especially after gene duplication.) Manual examination of the regions surrounding the human, zebrafish, and mouse MSX genes indicated that zebrafish *msxa* and *msxd* were co-orthologous to human *MSX2*; zebrafish *msxb* and *msxe* were co-orthologous to human *MSX1*; and that zebrafish *msxc* was orthologous to mouse *Msx3*. In the remainder of this section we will use the Synteny Database and associated tools to re-examine published results. The question is: Does the Synteny Database provide more predictive power than a phylogenetic or manual syntenic analysis alone?

We performed a phylogenetic analysis on the MSX gene family using a maximum likelihood analysis (previous trees were built using Neighbor-joining and maximum parsimony methods [30, 83]). Using cDNA sequences from human, mouse, zebrafish, chicken, amphioxus and *Ciona intestinalis*, we generated a tree using amino acid data. Figure 4.17A shows the tree resulting from our analysis of the MSX protein sequences



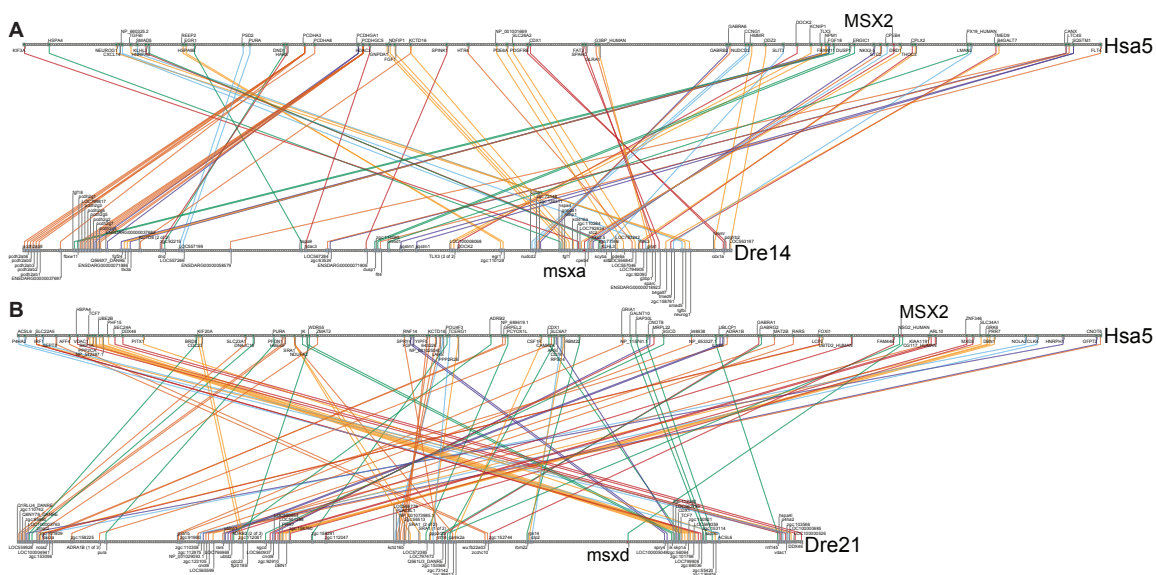
**FIGURE 4.17:** Analysis of the MSX gene family. (A) A phylogenetic tree of the MSX gene family. Generated with Phylml [39] using a maximum likelihood algorithm with a fixed amino acid model (JTT). Bootstrap values are reported on the internal nodes. Abbreviations: *Danio rerio* (Dre), human (Hsa), mouse (Mmu), chicken (Gga), amphioxus (Bfl), and *Ciona intestinalis* (Cin). (B) Hsa5 paralogy dotplot. The plot shows that a 50Mb region of chromosome 5 is paralogous to Hsa4 (green). The position of *MSX2* on Hsa5 is marked on the plot as is the position of *MSX2* on Hsa4. (C) A circle plot summarizing the human and zebrafish clusters used in the analysis of the MSX gene family. (D) The paralogous syntenic region surrounding the *msxa* and *msxd* genes on Dre14 and Dre21, respectively. This cluster was discovered by the Synteny Database using a 100-gene sliding window. (E) The *MSX1* and *msxe* orthologous syntenic cluster showing the conserved region between Hsa4 and Dre14. The cluster shows nine pairs of orthologs surrounding the MSX genes, as discovered by the synteny database using a 100-gene sliding window.

using maximum likelihood with a fixed amino acid model (JTT) [39, 53]. The tree shows three major clades (monophyletic groups) corresponding to the three mammalian *MSX* genes, with non-vertebrate chordates as outgroups. The tree grouped *Danio rerio* genes *msxa* and *msxd* as sisters in the *MSX2* clade with low bootstrap values, *msxe* into the *MSX1* clade, and *msxb* and *msxc* in the *Msx3* clade with good bootstrap values, and poor resolution among the three clades. Human, mouse, and chicken *MSX* genes grouped in the *MSX1* and *MSX2* clades with high bootstrap values (greater than 80%) but in both cases the zebrafish genes (*msxe* as well as *msxa* and *msxd*, respectively) grouped into clades without good bootstrap support. Finally, while the *Ciona* and amphioxus *MSX* genes fell basally on the tree, the *Ciona MSX* gene was quite divergent, falling as outgroup to all other genes despite the general modern consensus that urochordates, not cephalochordates, are the sister group of the vertebrates [75, 26, 112, 113, 27]. Despite using a more robust phylogenetic method than previous analyses, the numerous low bootstrap values reflect the uncertainty reported in earlier phylogenetic tree building attempts [30]. The weakly-supported hypothesis for the evolutionary history of the *MSX* gene family from the maximum likelihood tree (Fig. 4.17A) raises the question: Does automated analysis of conserved syntenies provide clarifying evidence for or against various hypotheses for *MSX* gene histories? The next section examines the results produced by the analysis pipeline and the Synteny Database.

### 4.3.1 Chordate MSX Genes

In concurrence with published data, the reciprocal best hit pipeline found two MSX paralogs in the human genome, *MSX1* on Hsa4 and *MSX2* on Hsa5. Both genes BLASTed to a single gene in amphioxus (75296) and *C. intestinalis* (*ENSC-ING00000009129*). In the case of *Ciona*, the pipeline filtered *MSX2* from the results because the BLAST hit covered only 37% of the length of the human and *Ciona* sequences – indicating that *MSX2* seems to be diverging at a faster rate than *MSX1* in the human genome and that the *Ciona* ortholog appears to be diverging much more rapidly than amphioxus *msx*. The full length of Hsa4 is paralogous to a portion of Hsa5, spanning approximately 65 megabases, as shown by the dotplot in Figure 4.17B, confirming previous work [25].

For the zebrafish genome, the RBH analysis pipeline identified all five *msx* paralogs, assigning *msxe* as the sole ortholog of human *MSX1*, consistent with the tree, but assigning all four remaining zebrafish *msx* genes (*msxa*, *msxb*, *msxc*, and *msxd*) as co-orthologs of human *MSX2*, likely reflecting the low bootstrap support of the tree. When the RBH pipeline probed the zebrafish genome using mouse as outgroup, the assignments mirrored the human MSX genes with, rather surprisingly, no orthologs assigned to mouse *Msx3*. In the case of *msxe*, the reciprocal best BLAST hit to *MSX1* was significantly stronger than the next best hit (to *MSX2*), confirming the phylogeny (Fig. 4.17A). When the analysis considered relationship strengths for the remaining four paralogs, it found that the four genes grouped as: ((*msxb*, *msxc*)(*msxa*,



**FIGURE 4.18:** Conserved synteny for *MSX2*-related genes. (A) The *MSX2* and *msxa* orthologous syntenic cluster showing the conserved region between human chromosome 5 (Hsa5) and zebrafish chromosome 14 (Dre14). The cluster shows a large number of orthologs surrounding the *MSX* genes, as discovered by the synteny database using a 100-gene sliding window. (B) The *MSX2* and *msxd* orthologous syntenic cluster showing the conserved region between Hsa5 and Dre21. The cluster shows a large number of orthologs surrounding the *MSX* genes, as discovered by the synteny database using a 100-gene sliding window.

*msxd*)). When it considered the corresponding orthologous BLAST search, however, the pipeline found that *msxa* and *msxd* both hit *MSX2* with a highly significant score, and *msxb* and *msxc* hit human *MSX1* and *MSX2* with about the same significance – making their automated assignment by the analysis pipeline ambiguous. BLAST results for the zebrafish *MSX* paralogs against the mouse were consistent with those for the human genome.

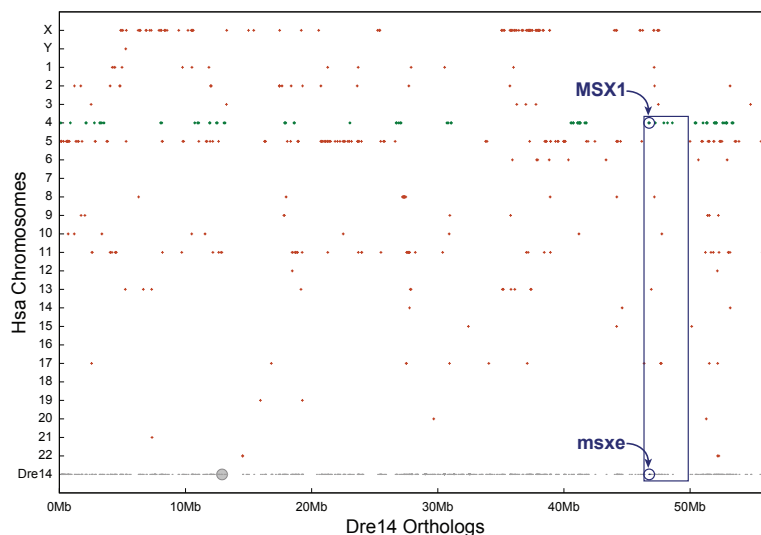


### 4.3.2 Conserved Syntenies for *MSX2* Paralogs

The chromosome region around the human *MSX2* gene on Hsa5 was highly conserved with the two corresponding regions in the zebrafish genome for *msxa* and *msxd* on Dre14 and Dre21, respectively. A circle plot (Fig. 4.17C) shows a 25Mb region on Dre14 and another 25Mb region on Dre21 that both correspond to the region surrounding *MSX2* on Hsa5 (see also Figs. 4.18A and B). Similarly, the Synteny Database detected a small cluster that contains both *msxa* and *msxd* as well as four additional related pairs of genes paralogous on Dre14 and Dre21 (Fig. 4.17D). Three of the pairs of genes in the cluster are co-orthologous to genes on Hsa5 including *hspa4/hspa4l*, *fgf1/LOC100005049*, and *cdx1a/CDX1* on Dre14 and Dre21, respectively. The sum of the results support the phylogenetic tree and are consistent with prior results with respect to *MSX2* [83]. We conclude that *msxa* should be called *msx2a* and *msxd* should be called *msx2b*.

### 4.3.3 Conserved Syntenies for *MSX1* Paralogs

We next considered the syntenic region surrounding the human *MSX1* gene. A strong reciprocal BLAST hit supported an orthologous relationship between *MSX1* and *msxe*, and conserved syntenies supported this conclusion. The Synteny Database found a nicely conserved region between Dre14 and Hsa4 (Fig. 4.17E). On Dre14, *otop1* orthologs border the region on one side and *tlr1* orthologs on the other. Between them lies the *MSX1/msxe* gene pair, a nice inversion containing four orthologous gene



**FIGURE 4.19:** Dre14 orthology dotplot against the human genome. The region surrounding the zebrafish *msxe* gene on Dre14 shows syntenic conservation with Hsa4, not Hsa5 (boxed area) which would be expected if *msxe* had been produced by a tandem duplication of *msxa* rather than as part of the R3 whole genome duplication.

pairs as well as three additional orthologous pairs. The *msxe* gene falls on Dre14, the same chromosome that contains *msxa*, although the two genes are separated by over twenty megabases. The location of two MSX genes on the same chromosome can be explained by one of at least two hypotheses. In the first scenario, *msxe* was created by a tandem duplication event from an ancestral *msxa/e* gene followed by chromosome inversions that later separated the resulting *msxa* and *msxe* genes. In the second scenario, *msxe* is a product of whole genome duplication events and sometime after R3 a translocation moved it onto the same chromosome as *msxa*. A translocation event is likely to move more than a single gene and so if *msxe* and *msxa* resulted from a tandem duplication we would expect to see conserved synteny with Hsa5, the location of *MSX2* (brown lines in Fig. 4.17C), not with Hsa4, the location of *MSX1* (dark blue

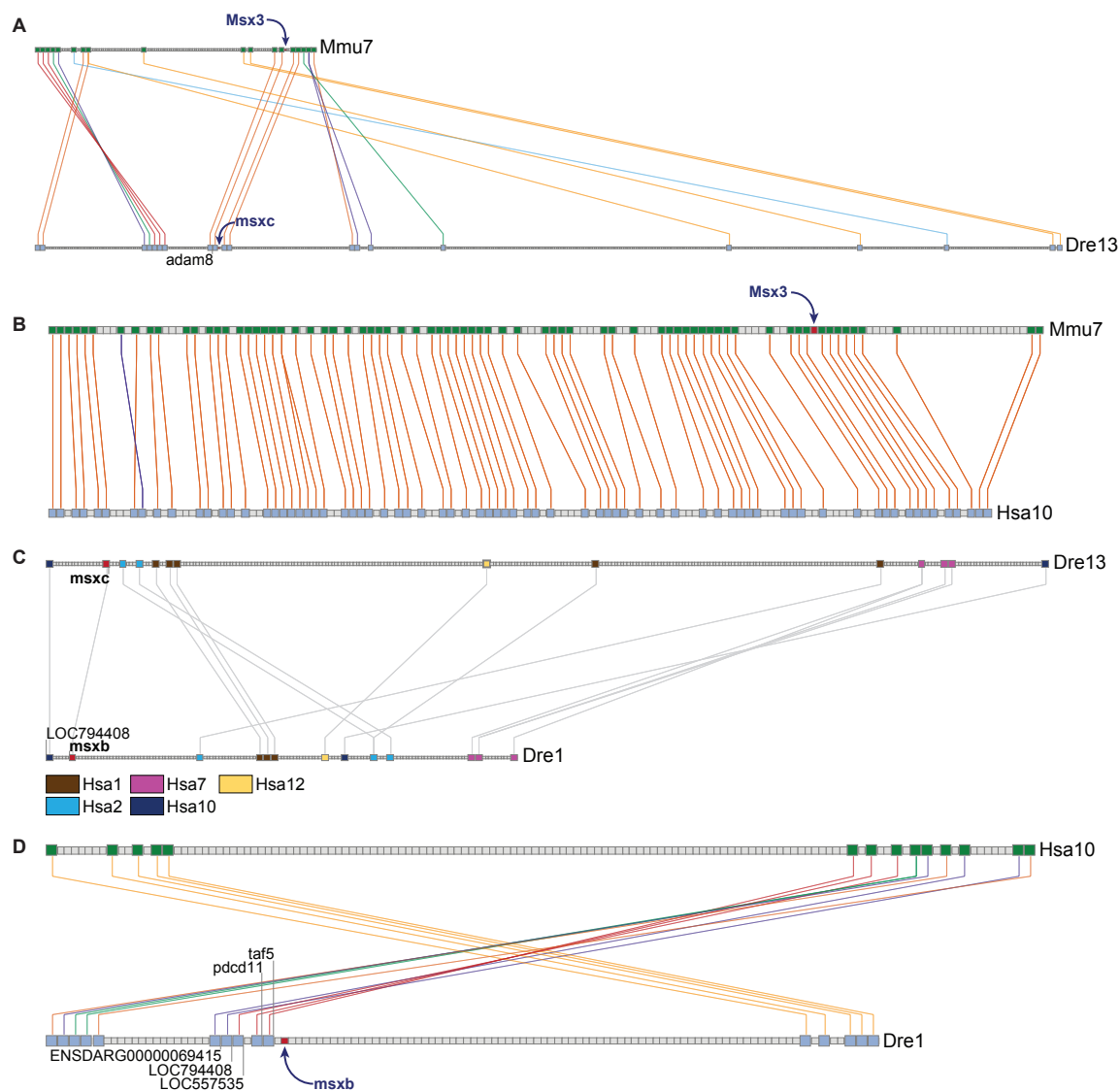
lines in Fig. 4.17C). The data, however, shows syntenic conservation between Dre14 and Hsa4, containing *msxe* and *MSX1*, respectively (Fig. 4.17E and Fig. 4.19). For this reason, a translocation of *msxe* after the R3 duplication is more parsimonious than a tandem duplication. This analysis supports previous work that suggests that zebrafish *msxe* is an ortholog of *MSX1* and should be renamed *msx1*.

#### 4.3.4 Conserved Syntenies for *msxb* and *msxc*

The RBH analysis and syntenic data provided evidence that paralogous zebrafish genes *msxa* and *msxd* are co-orthologous to human gene *MSX2* and that *msxe* is orthologous to *MSX1*. These data are consistent with our phylogenetic analysis as well as with prior results [83]. We now return to the remaining zebrafish genes. As discussed above, the analysis pipeline ambiguously assigned *msxb* and *msxc* to the human *MSX2* gene. The assignment is prone to error because at least two R1 and R2 ohnologs have gone missing in the human MSX gene family and if either *msxb* or *msxc* is orthologous to one of these missing genes, then the pipeline will assign it to the next closest, extant ortholog. This situation suggests two hypotheses for the evolutionary origin of the *msxb* and *msxc* genes. In the first hypothesis, *msxb* and *msxc* are both co-orthologous to one of the human ohnologs gone missing (we could call them *MSX3* and *MSX4*). In the second hypothesis, *msxb* is orthologous to one of the ohnologs gone missing and *msxc* is orthologous to the other. More complicated hypotheses, of course, are also possible. Because *msx3*, one of the human ohnologs gone missing, still

exists in the mouse, analysis starts there. Although the phylogenetic tree grouped *msxb* and *msxc* with mouse *Msx3* (Fig. 4.17A), the reciprocal best hit pipeline once again assigned both zebrafish paralogs to the *Msx2* gene. A BLAST search of the zebrafish genes against the mouse genome showed that both *msxb* and *msxc* had better scores to *Msx2* than *Msx3*. The reverse BLAST search, using *Msx2* as a query against the zebrafish genome, returned *msxd*, *msxa*, *msxc*, and *msxb* in that order. Using *Msx3* as a query against zebrafish returned *msxc*, and *msxb* as the top two hits, but the scores for these two hits were both lower than all four of the BLAST hits for *Msx2*. Therefore, the RBH analysis pipeline erroneously grouped *msxb* and *msxc* with mouse *Msx2*. The mouse *Msx3* gene has apparently diverged far enough from its zebrafish orthologs that the pipeline does not have the power to make the proper assignment.

The Synteny Database identified an orthologous cluster between Dre13 and mouse (*Mus musculus*) chromosome 7 (Mmu7). Although *msxc* is not part of this Dre13/Mmu7 cluster (because the analysis pipeline erroneously assigned it to the wrong paralog group), *adam8* which is the next nearest genomic neighbor to *msxc* is a member. Additionally, there are nineteen more Dre13/Mmu7 gene pairs surrounding *adam8* (Fig. 4.20A). It is important to note that the region on Mmu7 containing *Msx3* and orthologous to the *msxc*-containing portion of Dre13 is extremely well conserved to a portion of Hsa10 with what seems to be a surgical deletion of what would have been



**FIGURE 4.20:** Conserved synteny for *Msx3*. (A) The *Msx3* and *msxc* orthologous syntenic cluster showing the conserved region between mouse chromosome 7 (Mmu7) and Dre13. The cluster shows nineteen pairs of orthologs surrounding the *Msx3/msxc* genes, as discovered by the Synteny Database using a 200-gene sliding window. (B) The *Msx3* orthologous region between Mmu7 and Hsa10 containing 72 orthologous gene pairs as discovered with a 25-gene sliding window. Since the *MSX3* gene has been lost in the human lineage, this cluster, along with the previous cluster (Fig. 4.20A), imply orthology between zebrafish *msxc* and the human ohnolog gene missing, *MSX3*. (C) The *msxb* orthologous syntenic cluster showing the conserved region between Dre1 and Hsa10. The cluster contains 14 orthologous gene pairs and was generated from the Synteny Database with a 100-gene sliding window. It falls on Hsa10 approximately 14Mb from the Dre13/Mmu7/Hsa10 cluster.

*MSX3* if this ortholog had not gone missing in the human lineage (Fig. 4.20B). We conclude that *msxc* is an ortholog of *Msx3*, consistent with prior results.

Having established syntenic conservation between *msxc* and mouse *Msx3*, we asked: Does the region containing *msxc* have a paralogon in the zebrafish genome? The Synteny Database found a paralogous syntenic cluster between the portion of Dre13 containing *msxc* and a part of Dre1 containing *msxb* and twelve additional pairs of paralogs (Figure 4.20C). We annotated the diagram to show the chromosomal origin of human orthologs for each set of zebrafish paralogs. Unlike the cluster supporting the *msxa/msxd* paralogs (Fig. 4.20D), the members of this Dre1/Dre13 cluster have orthologs on a number of human chromosomes, including Hsa1, Hsa2, Hsa10, and Hsa12. This implies that after humans diverged from the lineage that led to teleost fish, a large number of translocations occurred for this ancient chromosome segment either in the human lineage or in the zebrafish lineage, or both; a comparison of gene orders with an outgroup that did not experience the R3 duplication event would show which model is correct.

Earlier analysis of the MSX gene family used the zebrafish meiotic linkage map [121] as a base to search for conserved synteny, which limited the analysis of an orthologous syntenic cluster for *msxb*. The sequence of the zebrafish genome, however, provides a more detailed view for the discovery of conserved syntenies. Are there other genes that were not available in the linkage map that might now provide additional evidence for the orthology of *msxb*? Of the seven downstream neighbors of

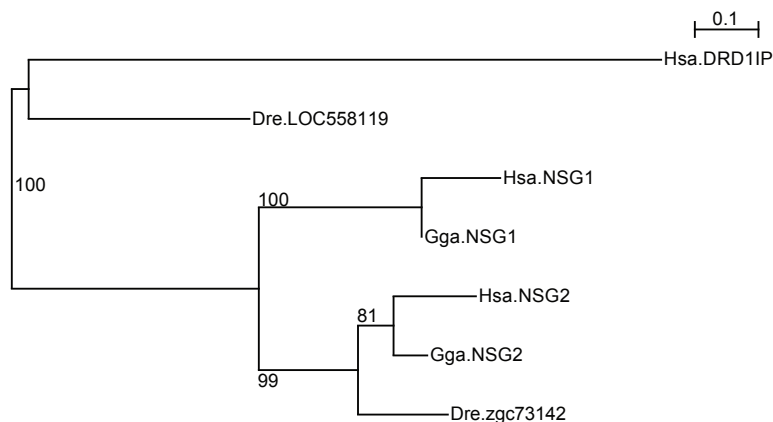
*msxb* on Dre1, five have human orthologs on chromosome 10 in a region approximately 15 megabases downstream from the presumed location of the lost *MSX3* gene (*taf5*, *pdcd11*, *LOC557535*, *LOC794408*, and *ENSDARG00000069415*) (red arcs in Fig. 4.17C). Of these five genes, two have paralogs on Dre13 (one of which is a member of the Dre1/Dre13 paralogous cluster shown in Fig. 4.20C: *LOC794408*). The syntenic orthologous cluster showing the Dre1/Hsa10 conservation can be seen in Figure 4.20D. This result would be expected under the hypothesis that *msxb* is an ortholog of the missing human *MSX3* gene.

### 4.3.5 Resolving the ambiguity of *msxb*

We have uncovered strong evidence that *msxc* is an ortholog of mouse gene *Msx3* and that the region surrounding *msxc* on Dre13 is conserved on both Mmu7 and Hsa10 (with the ancient *MSX3* gene now missing from the human genome, Fig. 4.20A, B). Additionally, we have a paralogous syntenic cluster associating the regions surrounding *msxc* and *msxb* (Fig. 4.20C), although the cluster is not orthologous to a single location in the human genome, and an orthologous cluster between the *msxb* region on Dre1 and near the *MSX3* region on Hsa10 (Fig. 4.17C and 4.17D). These results lead to the conclusion that *msxb* and *msxc* are both co-orthologs of *Msx3*. This assignment of orthology for *msxb* conflicts with the previous analysis, which had assigned *msxb* as a paralog of *msxe* (and orthologous to human *MSX1*). The results from the RBH analysis pipeline provided additional data to help resolve the history of *msxb*. Starting

at *msxb* on Dre1, the nearest upstream neighbor on the chromosome is *LOC558119*. The RBH pipeline reports that *LOC558119* is orthologous to human *NSG1* on Hsa4. *NSG1* itself has two paralogs, *NSG2* and *DRD1IP* on Hsa5 and Hsa10, respectively. *NSG1*, *NSG2*, and *DRD1IP* are either the direct neighbors of *MSX1*, *MSX2*, and the now lost *MSX3*, or the next-nearest neighbor. The positions of these genes are shown in a circle plot (Fig. 4.22A). Prior work reasoned that since *NSG1* is the direct neighbor of *MSX1*, and *LOC558119* is the direct neighbor of *msxb*, then *msxb* must be paralogous to *msxe* (the zebrafish ortholog of *MSX1*) [83]. Although, this conclusion is not strongly ruled out by phylogenetic analysis, the Dre1/Dre13 syntenic cluster described above conflicts with this scenario (Fig. 4.20C). If the assignment of orthology between human *NSG1* and *LOC558119* was incorrect, however, and instead *LOC558119* is orthologous to *DRD1IP* on Hsa10, then the Dre1/Dre13 syntenic cluster and the nearest-neighbor BLAST data would be in agreement. The two possible orthology assignments are outlined by red-dotted lines in Fig. 4.22A. A close examination of the BLAST results shows that *LOC558119* may be orthologous to *DRD1IP*. The *LOC558119* gene's top three BLAST hits in the human genome are *NSG1*, *NSG2*, and *DRD1IP* in that order. All three hits have approximately the same length and percent identity (167-172aa alignment length, 39-48% identity). Also, while *NSG1*'s top BLAST hit in zebrafish is *LOC558119*, *DRD1IP*'s top BLAST hit in zebrafish is also *LOC558119*. Rapid divergence in the *DRD1IP/NSG1* human genes, or in the

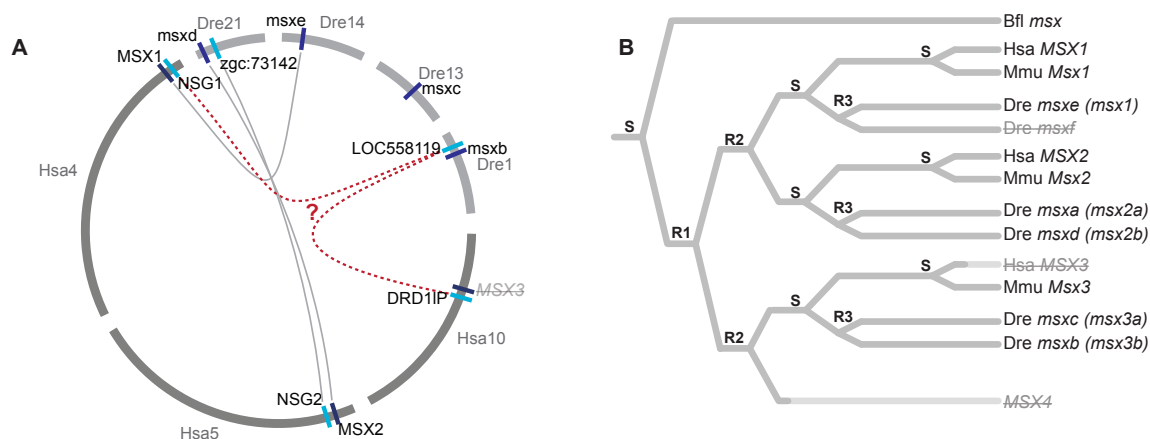




**FIGURE 4.21:** NSG gene family tree allowing inference of orthology between zebrafish gene *LOC558119* and human gene *DRD1IP*, not *NSG1*.

zebrafish *LOC558119* gene may be responsible for an incorrect assignment by the pipeline.

To further explore these results, we built a phylogenetic tree of the NSG genes along with their zebrafish and chicken orthologs. The results (Fig. 4.21) are consistent with an assignment of orthology between zebrafish *LOC558119* and human *DRD1IP* which is in agreement with the Dre1/Dre13 syntenic cluster and hence *msxb* as an ortholog of *Msx3*. An alternative possibility is that *msxb* is orthologous to the ohnolog gone missing, *MSX4*. If *msxb* is not the R3 paralog of *msxe*, given the proximity of NSG paralogs to their MSX neighbors, it may be that *LOC558119* BLASTs best to *NSG1* only because the true human ortholog of *LOC558119* has been lost. This possibility cannot be ruled out, but it would contradict the strong Dre1/Dre13 syntenic cluster and the Dre1/Hsa10 orthologous cluster and is therefore less parsimonious with an assignment of paralogy between *msxc* and *msxb*, and co-orthology of *msxc*



**FIGURE 4.22:** Evolutionary history of the MSX Gene Family. (A) A circle plot showing the positions of a subset of the MSX genes in human and zebrafish. The plot indicates the orthology assignments of the neighboring NSG gene family and shows the two possible orthology assignments for zebrafish gene *LOC558119*. The neighboring *LOC558119* gene is useful to help determine the proper orthology of zebrafish gene *msxb*. (B) A gene tree showing the evolutionary history of the chordate MSX gene family. *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 strike-through text have been lost.

and *msxb* to *MSX3*. We thus conclude that *msxb* is highly likely to be an ortholog of *Msx3*.

### 4.3.6 An MSX Family History

The automated analysis of orthologies and conserved synteny supports the following evolutionary history of the MSX family (Fig. 4.22B). The chordate MSX gene family arose from a single gene in stem chordates, represented by a single homeobox-containing gene in the basally diverging chordates amphioxus and *Ciona intestinalis* today. That gene was duplicated in the R1 and R2 duplication events to give four copies, of which *Msx1* and *Msx2* remain in mouse and human, *Msx3* remains in the

mouse, and *MSX4* apparently died a pauper's death with no descendants. After the lineage leading to teleost fish diverged from the lineage leading to humans, the R3 duplication event and subsequent gene losses resulted in five MSX genes in the zebrafish genome. Of those five genes, *msxe* is orthologous to *MSX1* (i.e. *msxe* could be called *msx1*), the paralogs *msxa* and *msxd* are co-orthologous to human *MSX2* and could be called *msx2a* and *msx2b*, respectively, and paralogs *msxc* and *msxb* are co-orthologous to mouse *Msx3* and could be called *msx3a* and *msx3b*, respectively. Thus, note that human has no orthologs of two zebrafish *msx* family genes and one mouse *Msx* family gene. This understanding has major implications for the connectivity of human and model system genomes when interpreting this important gene family. The MSX genes represent a difficult, although typical, case study for the Synteny Database and its associated tools.

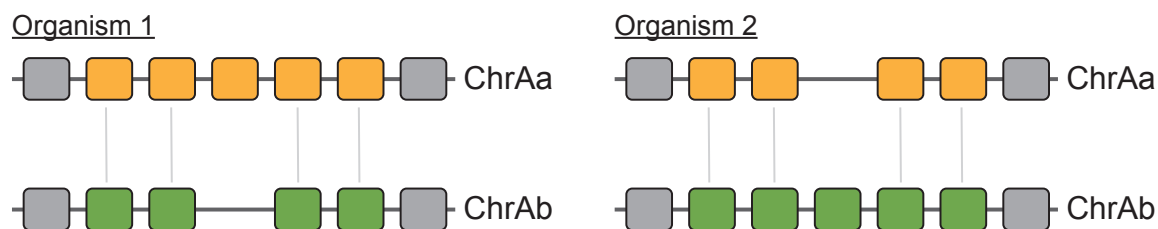
The Synteny Database and associated tools provided several advantages in characterizing the evolutionary history of the MSX gene family; it can perform analyses in multiple species and if a particular gene is missing or hard to identify due to sequence divergence, a neighboring gene can be used as a proxy. Despite difficult identification of the *msxc* gene in zebrafish, the system was able to associate its neighbor, *adam8*, with a region of conserved synteny in the mouse. Similarly, we were able to identify the *Msx3* gene in mouse and associate the syntenic area around it to an orthologous area in the human genome corresponding to the lost *MSX3* gene. Clusters produced

by the Synteny Database span large regions of the genome revealing syntenic conservation that would be tedious and time consuming to identify by hand, such as the Dre1/Hsa10 orthologous *msxb* cluster, while the depth of the data provided by the RBH analysis pipeline allows for the investigation of any individual result to establish confidence in the totality of the results, as was the case for the *msxb* gene.

## CHAPTER V

### IDENTIFYING OHNOLOGS GONE MISSING

One of the major consequences of a whole-genome duplication event is rapid gene loss; as time passes and speciation events occur, differential gene loss occurs in the resulting lineages. In fact, it has been hypothesized that the differential loss of these duplicated genes may contribute to speciation events and we presented evidence showing how this phenomenon may occur in *Arabidopsis* (Sec. 1.4). The teleost fish, with more species than any other vertebrates, should contain numerous examples ohnologs gone missing resulting from the R3 whole-genome duplication. In particular, the number of fully sequenced teleost genomes should allow for the detection of reciprocal gene loss (RGL) – when alternative paralogs are lost in different species (e.g. the *a* copy of an R3 gene duplicate is lost in one species and the *b* copy is lost in the other). We presented the work of Sémon and Wolfe [95] who studied this problem in Chapter II.



**FIGURE 5.1:** An example of reciprocal gene loss. Chromosomes Aa and Ab are paralogs resulting from a WGD previous to the speciation of Species 1 and Species 2. The yellow genes on chromosome Aa are orthologous as are the green genes on chromosome Ab. Grey lines connect syntenically conserved paralogs. The *b* copy (green) of the gene has been lost in Species 1, while the *a* copy (yellow) has been lost in Species 2.

One of the major challenges for an orthology assignment algorithm is accounting for ohnologs gone missing. In Chapter IV, we described a system to detect chromosomal segments within a genome, and between genomes, whose gene contents were syntenically conserved. In the application of that algorithm in two case studies, we were able to infer several ohnologs gone missing in the zebrafish and human genomes by manually comparing the syntenic neighborhoods of paralogous and orthologous genes. In this chapter, we combine the datasets of the RBH Analysis Pipeline and the Synteny Database in a pair of related algorithms to detect conserved syntenic neighborhoods across different species and use those neighborhoods to automatically infer ohnologs gone missing in teleost and human genomes. Identifying ohnologs gone missing in the teleosts allows us to investigate a number of architectural features unique to post-duplication genomes, such as reciprocal gene loss, while investigating ohnologs gone missing in the human genome will allow us to identify genes lost in the human lineage since the ancestral human and teleost lineages diverged.

Our strategy for these algorithms is novel and relies on the use of what we refer to as micro-synteny. Aggregating our paralog and ortholog mappings generated by the RBH Analysis Pipeline allowed us to investigate the conservation of gene orders in several mammalian and teleost genomes (Chapter III); this data indicated that although the R3 duplication signal was present, the teleost genomes had undergone significant architectural rearrangements since the divergence of the ancestral human and teleost lineages. We used the Synteny Database (Chapter IV) to cluster the conserved gene orders into syntenically conserved regions and demonstrated that a small sliding window size provided the most statistically significant regions of conservation. Further, in the study of the ARNTL and MSX gene families we showed that to confidently infer an ohnolog gene missing the most immediate syntenic neighborhood of any particular gene must be well conserved. Combining these results from our earlier analyses, our strategy for detecting ohnologs gene missing must rely on local syntenic conservation, or micro-synteny.

Consider one architectural feature of post-duplication genomes formed by a pair of ohnologs gene missing: reciprocal gene loss. Figure 5.1 shows a simplified example of RGL in two species, with the *b* copy of a set of R3 paralogs lost in Species 1 (green), and the *a* copy lost in Species 2 (yellow). The RBH Analysis Pipeline would incorrectly find that the extant ohnologs were co-orthologs; to correct the results and infer reciprocal gene loss, one could compare the immediate syntenic neighborhoods of the existing copies of the gene (the yellow and green genes). Once one had demonstrated

that the regions of yellow and green genes were orthologous between species 1 and 2 one could infer a reciprocal gene loss and correct the misassignment by the RBH Analysis Pipeline. Previous work investigating RGL [95] pursued this strategy using much larger regions of synteny – an approach prone to producing false positives (in fact, one of the author’s primary examples, inferring RGL for the *MATN3* gene in zebrafish and pufferfish was a false positive, matching the wrong conserved segments together erroneously implying an ohnolog gone missing when the gene was actually present on another, less-well conserved chromosomes).

Besides focusing on the use of micro-synteny, the second major component of our OGM detection strategy is the ability to aggregate data from multiple genomes – identifying the same genomic neighborhoods in a number of species. This makes it possible to provide multiple lines of evidence for ohnologs gone missing, accumulated from the comparisons of multiple teleost genomes. The most likely predictions for ohnologs gone missing will have supporting evidence from multiple species of fish. In the remainder of this chapter we will present the algorithm developed to detect ohnologs gone missing in the teleost lineages, based on their human orthologs, as well as a variant on that algorithm that uses multiple teleost orthologs to detect ohnologs gone missing in the human genome. We will then present the results and examine a number of different architectural genomic features our two algorithms are able to detect.

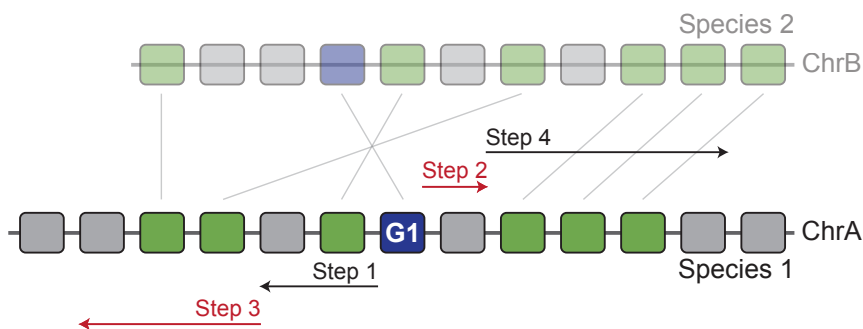


## 5.1 Methods

We present two PIP-based pipelines, the first, the Teleost OGM Pipeline, searches for ohnologs gone missing in teleost species based on human ortholog genes, and the second, the Human OGM Pipeline, operates in reverse, using conserved syntenic regions in multiple teleost species to identify ohnologs gone missing in the human genome. The kernel of these algorithms relies on the following idea. For any particular gene, we want to enumerate the micro-synteny around that gene; that micro-synteny will be provided by a cluster from the Synteny Database. Once we have established evidence of micro-synteny, we then want to look at the corresponding half of the synteny cluster in a second genome and investigate if our gene of interest has an ortholog in that region. We will apply this pattern in several different ways to associate orthologous regions of multiple genomes in order to infer ohnologs gone missing. Prior to describing the two pipelines in detail, we will first discuss our approach to detect micro-synteny and to reconcile BLAST results.

### 5.1.1 Micro-synteny Detection Algorithm

The heart of the two OGM pipelines lies in the micro-synteny detection algorithm. Given a particular gene in the genome, this algorithm seeks to determine if the immediate neighborhood of genes is syntenically conserved. To achieve this goal the algorithm queries the Synteny Database and constructs a list of orthologous clusters that overlap the gene of interest. The algorithm is diagrammed in Figure 5.2, where



**FIGURE 5.2:** Micro-synteny search algorithm. Given a syntenic cluster that spans a segment of chromosome A (ChrA) in Species 1, the algorithm searches the area around gene *G1*, in order to determine if there is a locally conserved syntenic neighborhood within the larger syntenic cluster. The algorithm alternates, searching upstream and downstream of *G1*, greedily counting the number of neighbors that are syntenically conserved before it encounters the maximum number of gaps allowed.

our gene of interest, *G1* resides on chromosome **A** of Species 1. The corresponding, orthologous half of the cluster is shown occupying chromosome **B** in a Species 2 while grey lines connect orthologous gene pairs. It is not necessary that *G1* itself is a member of the cluster, in fact, we often expect the gene not to be a member, since membership is based on orthology (and we are looking for genes in the corresponding orthologous half of the cluster that have been lost). Starting at *G1*, the algorithm will alternate searching upstream and downstream from *G1* in a greedy fashion. When the algorithm encounters a “gap”, or a gene that is not a member of the cluster, it is recorded and the algorithm switches directions; halting when the gap limit has been reached. If enough neighboring genes are found before the gap limit is reached, the micro-syntenic region is recorded.

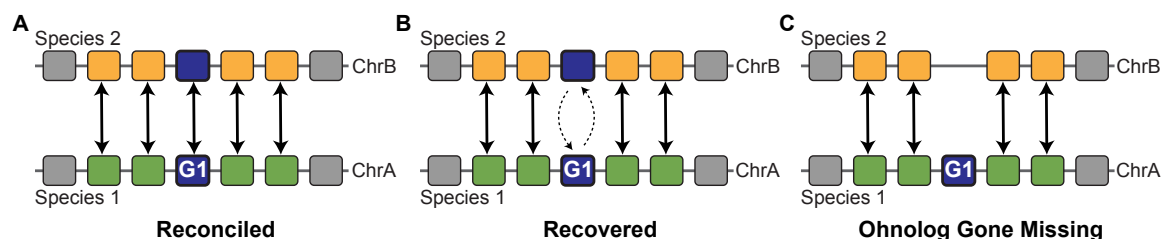
Pseudocode for the micro-synteny algorithm is available in Appendix D and executes in  $O(n^2)$  time. For each gene in the genome being examined ( $n$ ), at maximum,

the algorithm would visit every other gene on the same chromosome once, although in practice, a local syntenic neighborhood rarely extends more than a few tens of genes in either direction from the location of the gene of interest.

### 5.1.2 Reconciliation

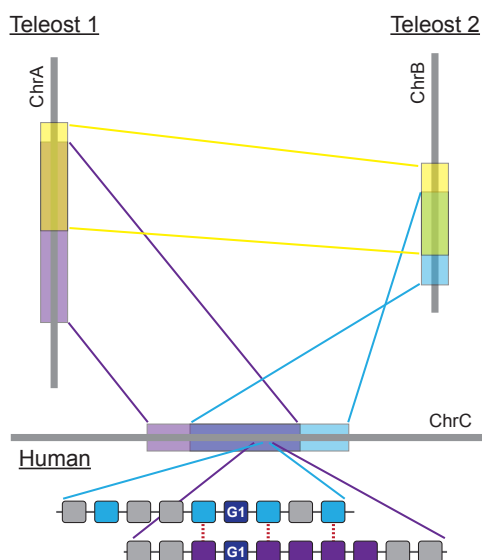
The micro-synteny search algorithm identifies genes in a genome that have a locally conserved syntenic neighborhood. This neighborhood consists of a set of genes that are orthologous to genes, similarly conserved, on a chromosome in a second species. However, the micro-synteny search algorithm makes no guarantee that the orthologous genes in the second species reside in a local neighborhood themselves (although their distance from one another is limited by the sliding window that defined the syntenic cluster). The micro-synteny algorithm defines a neighborhood of syntenically conserved genes around  $G1$  but says nothing about  $G1$  itself. Therefore, we need to investigate whether  $G1$  has an ortholog in the second species.

Investigating whether gene  $G1$  from Species 1 has an ortholog in Species 2 results in three possible outcomes (Fig. 5.3). First, in the vast majority of cases,  $G1$  will have an ortholog located in the corresponding half of the cluster in Species 2. In this case, we consider the orthology of  $G1$  to be **reconciled** with the syntenic cluster that defines its local neighborhood (Fig. 5.3A). If  $G1$  has an ortholog, but it is not located in the corresponding syntenic cluster in Species 2, we drop the gene from further consideration. In the second case,  $G1$  does not have an ortholog (as defined



**FIGURE 5.3:** Reconciliation. A segment of chromosome A (ChrA) from Species 1 and chromosome B (ChrB) in Species 2 are shown. Yellow and green genes are orthologous and define a local syntenic neighborhood around *G1*. Lines with two-way arrows represent a reciprocal best hit relationship. (A) In the majority of cases, *G1* will have an ortholog in Species 2 located within the local syntenic neighborhood and can be considered *reconciled*. (B) In some cases, *G1* does not have an ortholog in Species 2, but may have BLAST hits (dotted lines) that connect it to a gene within the syntenic neighborhood allowing the algorithm to *recover* the ortholog. (C) If *G1* has no significant BLAST hits to the orthologous genome, the algorithm records a tentative *ohnolog gene missing*.

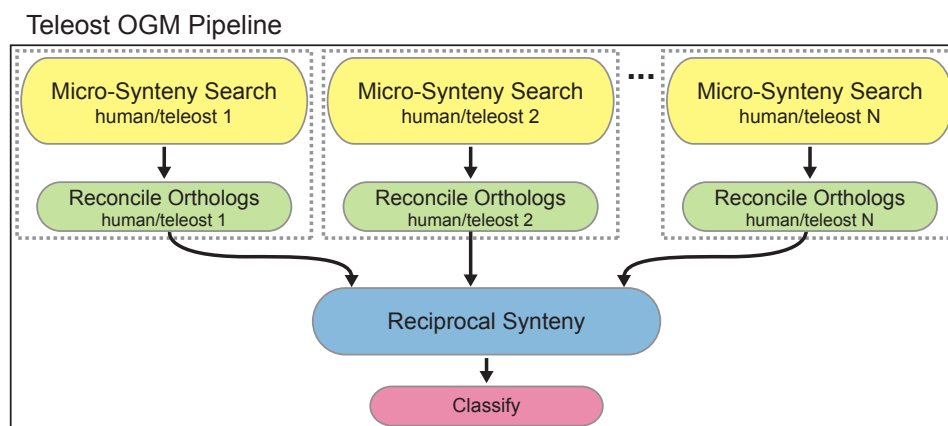
by the RBH Analysis Pipeline) (Fig. 5.3B). In this case, the algorithm looks up the forward and reverse BLAST results for *G1* and determines if there is a gene in the search results that is located in the proper syntenic neighborhood in Species 2. If it finds such a relationship it **recovers** the *G1* ortholog – in effect using conserved synteny to correct an error in the RBH Analysis Pipeline. This corrects situations, for example, when evolutionary rate asymmetry has prevented the pipeline from finding the correct ortholog (see Section 3.2.2 for an example). In the final case, not only does *G1* not have an ortholog, but it has no significant BLAST hits to any genes in the orthologous genome (Fig. 5.3C). In this situation, the algorithm records a tentative **ohnolog gene missing**, although that designation is not meaningful until corroborated by additional evidence. Two pipelines that integrate this data to provide corroborating evidence are described next.



**FIGURE 5.4:** Teleost OGM Schematic. The Teleost OGM Pipeline tries to find regions of locally conserved synteny in the human genome and links those regions to areas in two teleost genomes using clusters from the Synteny Database. It then searches for a teleost to teleost syntenic region to form triangles of reciprocal synteny between two teleost genomes and the human genome.

### 5.1.3 The Teleost OGM Pipeline

The Teleost OGM Pipeline examines human genes and uses conserved synteny to find corresponding regions in multiple teleost genomes. As shown in Figure 5.4, for a human gene *G1*, the pipeline attempts to find a locally conserved syntenic region in a teleost species (purple cluster), and a second, overlapping region in a second teleost species (blue cluster). Finally, the pipeline will search for a third, teleost-teleost syntenic cluster (yellow) to form a triangle of reciprocally conserved synteny, linking regions from human to teleost, from teleost to teleost, and from teleost to human. Based on the existence of teleost orthologs and on the strength of conserved synteny,

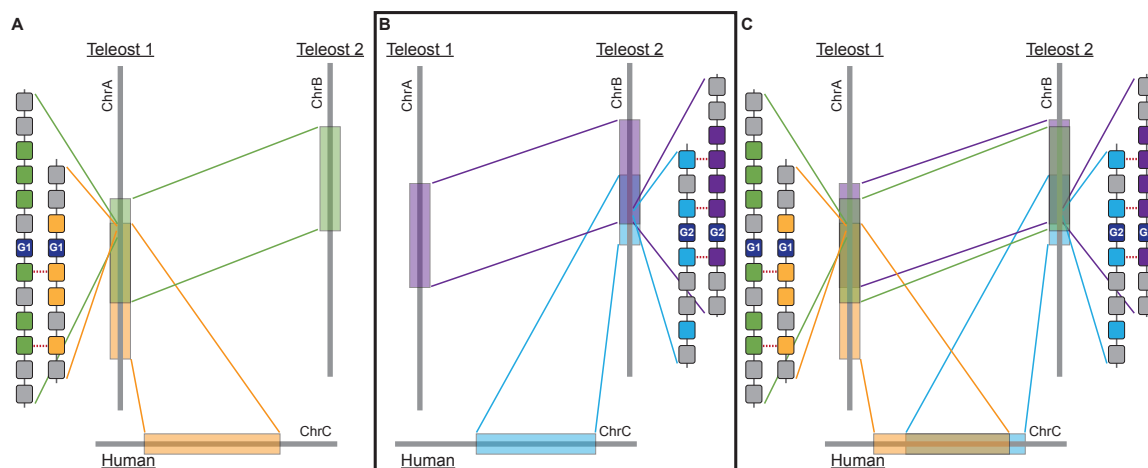


**FIGURE 5.5:** The Teleost OGM Pipeline repeatedly searches for locally conserved synteny between genes from the human genome and  $N$  teleost genomes. After reconciling human orthologs in the teleost genomes, regions are compared between pairs of teleost genomes to find regions of reciprocal synteny.

the pipeline can confirm existing orthologs, or infer an ortholog gone missing in one of the teleost genomes.

In more detail, the schematic for the PIP-based pipeline is shown in Figure 5.5. As described above, the first stage of the Teleost OGM Pipeline searches for local syntenic neighborhoods for each human gene  $G1$  with regard to the first teleost species, T1. The second stage reconciles orthologs for genes in which a locally syntenic neighborhood could be defined, recovering likely BLAST hits and inferring tentative orthologs gone missing. This series of steps is repeated for each teleost genome in the analysis, T2, T3..., TN. When successful, the pipeline will have found a set of clusters that span  $G1$ , the first linking the local human syntenic neighborhood around  $G1$  to teleost species T1, the second linking the same neighborhood to the teleost species T2, and so on. Often, some proportion of the human genes surrounding  $G1$  will

be members of multiple human syntenic neighborhoods (these genes are connected in Fig. 5.4 by red, dotted lines). These results are fed into the reciprocal synteny stage of the pipeline which considers results from the teleosts in pairs; given teleost species T1, T2, T3, and T4, the reciprocal synteny stage will examine regions of synteny between the human, T1, and T2 genomes, followed by regions between human, T3, and T4 genomes, and so on. Returning to our example of *G1*, the pipeline has two sets of clusters, the first connecting the human genome to T1, and the second connecting the human genome to T2. The pipeline will next query the Synteny Database and search for clusters that can link the two teleost regions (Fig. 5.4, yellow clusters). So, given *G1*, if the three overlapping neighborhood genes marked in Fig. 5.4 are *G2*, *G3*, and *G4*, the pipeline will check the orthologs of those three genes in both T1 and T2; giving us *G2T1*, *G3T1*, and *G4T1*, in the first teleost species, and *G2T2*, *G3T2*, and *G4T2*, in the second teleost species. If any of these teleost neighboring orthologs are members of the teleost to teleost syntenic region (yellow cluster), then having successfully linked the local syntenic neighborhood of the original human gene to regions in two teleost species the system will record a region of conserved reciprocal synteny. The reciprocal synteny analysis can be repeated with an arbitrary number of teleost genomes to provide independent lines of evidence for reciprocal synteny. In the final classification stage, the pipeline annotates which areas of reciprocal synteny actually contained orthologs gone missing and combines the results from multiple teleost species comparisons in



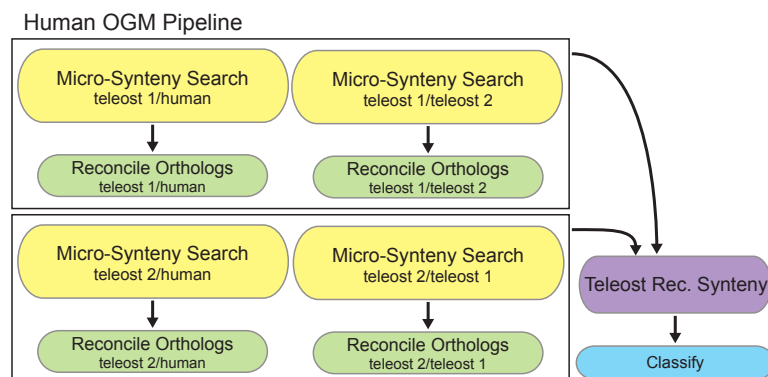
**FIGURE 5.6:** Human OGM Schematic

order to detect the presence of several additional features which we will discuss below in the results.

### 5.1.4 The Human OGM Pipeline

The Human OGM Pipeline shares much of its strategy with the Teleost OGM Pipeline, although the implementation yields more robust results. The idea underlying the analysis is to start with a gene in a teleost species and to search for locally conserved syntenic neighborhoods in a second teleost species as well as in the human genome (Fig. 5.6A). Then, this process is repeated starting with a gene in a second teleost species and searching into the first teleost species as well as into the human genome (Fig. 5.6B). Finally, these two sets of data are reconciled between the teleost species in order to define a area of conserved synteny in the human genome that is conserved in both teleost species (Fig. 5.6C).





**FIGURE 5.7:** Human OGM Pipeline

In more detail, the schematic for the PIP-based pipeline is shown in Figure 5.7. As with the Teleost OGM Pipeline, the first four stages of the pipeline define local syntenic neighborhoods between the three genomes and reconcile the orthologs associated with them (Fig. 5.6A, yellow and green clusters, and Fig. 5.6B, purple and blue clusters). At this point, we have a list of genes from teleost species T1 that have locally conserved synteny in teleost species T2 and in the human genome, and we have a list of genes from T2 that have locally conserved synteny in T1 and in the human genome. The teleost reciprocal synteny stage combines these two lists of genes based on the following criteria. First, given gene  $G1$  in T1, and  $G2$  in T2, the analysis stage identifies genes from the local neighborhood in T1 that are orthologous to genes from the local neighborhood in T2 creating syntenic support between teleost species T1 and T2 for genes  $G1$  and  $G2$  (Fig. 5.6,  $G1$  and  $G2$ ). Second, when teleost synteny can be established, the pipeline compares the corresponding human genes related to teleost genes  $G1$  and  $G2$  and verifies that both teleost regions implicate the same

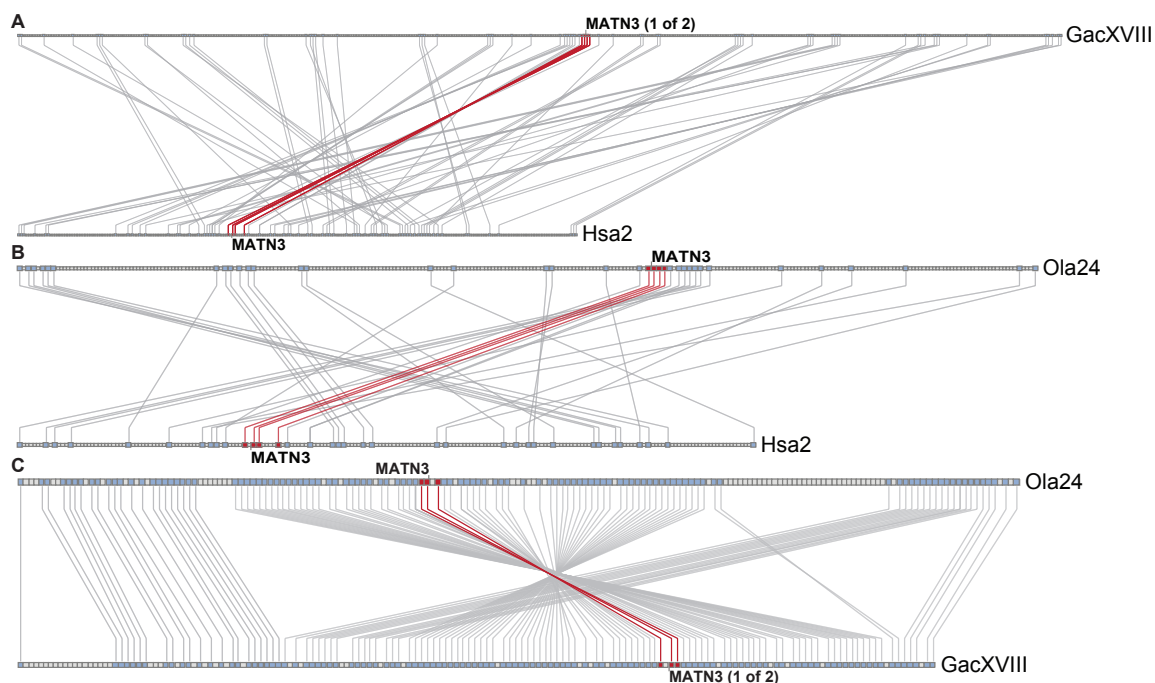
human gene and that local syntenic genes from both teleost species are orthologous to genes in the local human syntenic neighborhood. This entire analysis can then be repeated for additional pairs of teleost genomes. Finally, all of the results are fed into the classification stage where the teleost orthologs from the multiple analyses are chained together to group multiple lines of evidence and human orthologs gone missing are recorded.

## 5.2 Results

We executed the Teleost OGM Pipeline as well as the Human OGM pipeline with several teleost genomes, including zebrafish, stickleback, and medaka, against the human genome. The results of these two analyses follow.

### 5.2.1 The Teleost OGM Pipeline

We compared the human genome against three teleost genomes, in two analyses: human versus zebrafish and stickleback as well as human versus stickleback and medaka. The Teleost OGM Pipeline identified 5,760 unique cases of reciprocal conserved synteny; that is, for 5,760 human genes the pipeline was able to identify locally conserved regions around that gene linked to an ortholog in each of two teleost species, and, the pipeline was able to identify locally conserved synteny between the two teleost orthologs, creating a triangle of conserved synteny supporting the three



**FIGURE 5.8:** Reciprocal synteny of the human *MATN3* gene as identified by the Teleost OGM Pipeline. (A) Syntenic conservation between human chromosome 2 (Hsa2) and stickleback chromosome 18 (GacXVIII) as determined by the Synteny Database. Locally conserved synteny, as discovered by the micro-synteny algorithm by the Teleost OGM Pipeline is colored red. (B) Syntenic conservation between Hsa2 and medaka chromosome 24 (Ola24). (C) Syntenic conservation between teleost genomes, GacXVIII and Ola24.

orthologs. In more detail, the zebrafish/stickleback dataset produced 3,454 cases of reciprocal synteny while the stickleback/medaka dataset produced 4,709 cases. Of the 5,760 unique cases, 2,403 of them had support from both the zebrafish/stickleback and stickleback/medaka datasets.

Figure 5.8 provides a detailed account of one case of conserved reciprocal synteny. For the human *MATN3* gene, the pipeline identified regions of locally conserved synteny between the region surrounding *MATN3* on Hsa2 and in the stickleback genome on chromosome XVIII (Fig. 5.8A). The pipeline identified four of the surrounding nine

human genes as the locally conserved neighborhood within a Synteny Database cluster that spans 37 megabases of human chromosome 2. Repeating the operation with the medaka genome, the pipeline identified the exact same local neighborhood, syntenically conserved to medaka chromosome 24 (Fig. 5.8B), although the human/medaka cluster from the Synteny Database is smaller, only spanning 20 megabases of human chromosome 2. So, the locally conserved syntenic neighborhood surrounding *MATN3* produced four stickleback orthologs on GacXVIII as well as four medaka orthologs on Ola24. Next, the pipeline searched the Synteny Database for stickleback/medaka clusters and was able to identify a cluster that overlapped the regions on GacXVIII and Ola24 that also included three of the four human orthologs (Fig. 5.8C, red). Using the zebrafish/stickleback dataset, reciprocal synteny was also identified between Hsa2, GacXVIII, and Dre20.

We define an architectural feature of a genome as an emergent property created by the location of a set of genes within the genome. Reciprocal synteny, or the conservation of local syntenic neighborhoods across a set of genomes, is the simplest architectural feature the Teleost OGM pipeline can identify. By aggregating areas of reciprocal synteny, the classification stage of the pipeline is able to identify several other features as well. In 424 cases, for a particular human gene, the pipeline was able to identify both paralogous regions in a teleost produced in the R3 WGD event. So, given two teleost genomes, the pipeline is able to identify the duplicated region in teleost genome **A**, which includes paralogon **Aa** and paralogon **Ab**, and in the second

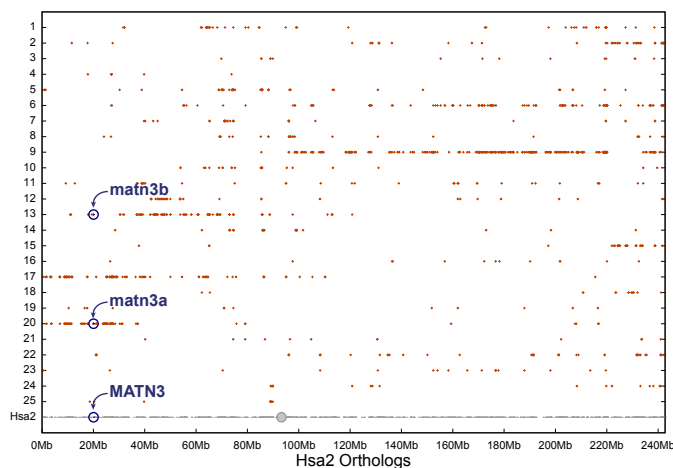
teleost genome, **B**, the pipeline is able to identify the two paralogs produced by R3, **Ba** and **Bb**. Finally, the pipeline is able to associate the orthologous regions between the genomes, associating **Aa** to **Ba** and **Ab** to **Bb** with conserved synteny. In 119 cases, the pipeline identified R3 paralogs in the zebrafish/stickleback dataset, and in 371 cases the pipeline identified R3 paralogs in the stickleback/medaka dataset. In 66 of the 424 cases, the pipeline was able to identify paralogous regions in both the zebrafish/stickleback and stickleback/medaka datasets showing orthology between the human genome and both duplicated regions in all three teleost genomes.

A third architectural feature the Teleost OGM pipeline can identify are R3 ohnologs gone missing. This feature is identified in the same way as the previous feature, associating **Aa** to **Ba** and **Ab** to **Bb**, but in this case, the **a** or **b** copy of the gene has been lost since the R3 WGD event in both teleost species. The Teleost OGM pipeline was able to identify R3 ohnologs gone missing in 150 cases, 38 cases in the zebrafish/stickleback dataset and 136 cases in the stickleback/medaka dataset. In 25 cases, the pipeline identified an R3 OGM in all three teleost genomes.

The final architectural feature the Teleost OGM Pipeline can identify is reciprocal gene loss, where the **a** copy of an R3 duplicate is lost in one teleost genome, but the **b** copy is lost in a second teleost genome. The pipeline was able to identify seven such cases, three in the zebrafish/stickleback dataset and four in the stickleback/medaka dataset. A full accounting of reciprocal gene loss cases is given in Table V.1.

Human Gene	Aa	Ba	Ab	Bb
<i>COL17A1</i>	ENSDARG00000069415	OGM	OGM	ENSGACG00000009340
Hsa10, 105.8M	Dre1, 46.5M	GacIX, 2.4M	Dre13, 24.3M	GacVI, 11.0M
<i>HELZ</i>	<i>zgc:77407</i>	OGM	OGM	<i>HELZ</i>
Hsa17, 62.5M	Dre3, 50.3M	GacXI, 11.3M	Dre6, 46.3M	GacIX, 12.9M
<i>C9orf90</i>	<i>LOC562755</i>	OGM	OGM	<i>C9orf90</i>
Hsa9, 129.9M	Dre8, 8.4M	GacXIII, 20.0M	Dre5, 23.5M	GacXIV, 14.4M
<i>UBL7</i>	<i>UBL7</i>	OGM	OGM	<i>UBL7</i>
Hsa15, 72.5M	GacII: 16.2M	Ola6, 11.6M	GacII, 4.3M	Ola3, 27.1M
<i>PPP1R1B</i>	<i>PPP1R1B</i>	OGM	OGM	<i>PPP1R1B</i>
Hsa17, 35.0M	GacV: 8.0M	Ola19, 8.7M	GacXI, 6.2M	Ola8, 5.8M
<i>MKX</i>	<i>MKX</i>	OGM	OGM	<i>MKX</i>
Hsa10, 28.0M	GacXXI, 0.7M	Ola20, 5.9M	GacIII, 10.3M	Ola17, 15.5M
<i>SCML1</i>	<i>SCML2 (2 of 2)</i>	OGM	<i>SCML2 (1 of 2)</i>	OGM
HsaX, 17.7M	GacII, 6.0M	Ola3, 12.6M	GacVIII, 18.2M	Ola4, 31.0M

**TABLE V.1:** Cases of reciprocal gene loss between human genes, teleost species **A**, and teleost species **B**, as discovered by the Teleost OGM Pipeline.



**FIGURE 5.9:** Hsa2 versus *Danio rerio* dotplot. The plot shows that the short arm of Hsa2 shows conservation to Dre20, Dre17, and Dre13. For human *MATN3* (marked), the zebrafish paralogs would be Dre20 and Dre13.

Identifying locally conserved synteny is the heart of the Teleost OGM algorithm and the key to finding cases of reciprocal gene loss in the teleosts. Returning to our earlier example using *MATN3*, for the human/stickleback analysis, the Synteny Database cluster that spanned the *MATN3* gene on Hsa2 stretches for 37 megabases – a large cluster compared to the size of most produced using a 50-gene sliding window, but one that still only covers approximately 15% of the total length of human chromosome 2. The human/zebrafish cluster between Hsa2 and Dre20 is just over 5 megabases in length, but still spans 52 genes. Besides Dre20, Hsa2 also has significant conserved synteny on Dre17 and also Dre13; if one wanted to identify the second area of conserved synteny in zebrafish, a broad measure of conserved synteny comparing Dre20 and Dre17, as well as Dre20 and Dre13 may indicate that Dre17 was the paralogon. However, as a dotplot makes very clear (Fig. 5.9, for the region

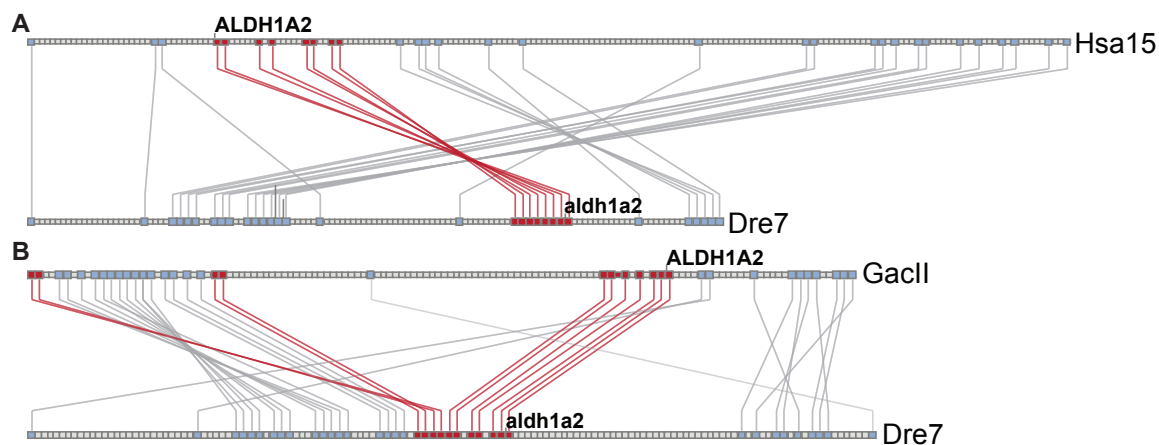
local to *MATN3* on Hsa2, the proper paralogs are Dre20 and Dre13, not Dre20 and Dre17. An algorithm that chooses the wrong paralogs due to a broad measure of synteny is likely to produce many false positives. This is exactly the error made for the *MATN3* gene in previous work [95], which incorrectly matched Dre20 and Dre17 as paralogs, inferring reciprocal gene loss where none occurred.

The key to the Teleost OGM Pipeline is effectively identifying the local syntenic neighborhood for a particular gene. Too strict of a measure will create false negatives, missing opportunities to identify the architectural features described above; too promiscuous and the algorithm will create false positives. We have erred on the strict side, but additional analyses of local neighborhoods may indicate an optimum measure of locality. Continuing to add additional teleost genomes to the analysis will provide additional information for human genes. In this analysis, we were able to make significantly more inferences from the stickleback/medaka dataset than from the zebrafish/stickleback dataset. Adding more closely related teleost species to the analysis may be the most productive route to enlarge our results. In the following section, we reverse our analysis, starting with genes in the teleosts and making inferences about ohnologs gone missing in the human genome.

### 5.2.2 The Human OGM Pipeline

We executed the Human OGM Pipeline with three teleost genomes arranged in two pairs, zebrafish/stickleback, and zebrafish/medaka. The pipeline was able to



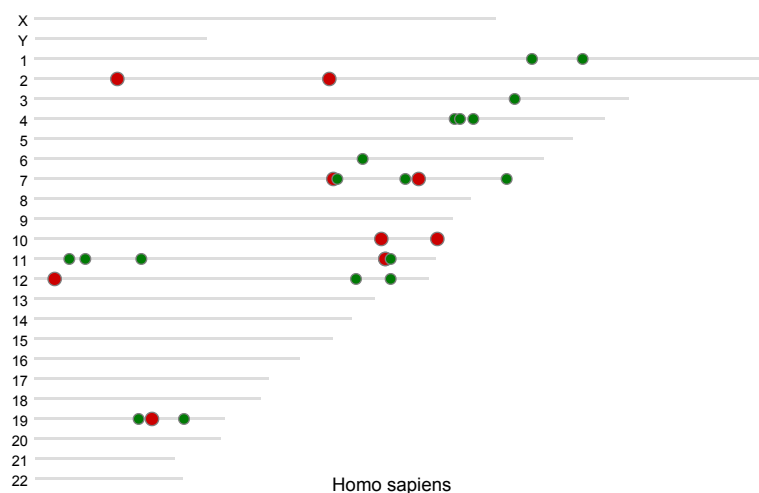


**FIGURE 5.10:** Reciprocal synteny of *ALDH1A2* as identified by the Human OGM Pipeline. (A) Syntenic conservation between Dre7 and Hsa15 as determined by the Synteny Database. Locally conserved synteny, as discovered by the micro-synteny algorithm is colored red. (B) Syntenic conservation between Dre7 and GacII.

detect 4,247 teleost orthologs that exhibited locally conserved synteny to the human genome and to at least one additional teleost genome. In 1,959 cases, conserved human synteny was supported by two teleost genomes, while in 2,288 cases support was provided by all three teleost genomes. Since we used the zebrafish in both of our pairwise comparisons, a conserved zebrafish region was involved in all found cases of reciprocal synteny. However, in 3,429 of the 4,247 cases conservation was found in the stickleback genome, and in 3,106 cases conservation was found in the medaka genome as well, consistent with the fact that zebrafish and stickleback are more closely related than zebrafish and medaka (see Sec. 3.2).

As an example we will review the results for the human *ALDH1A2* gene located on human chromosome 15. The Human OGM Pipeline was able to identify

locally conserved syntenic neighborhoods between *ALDH1A2* on Hsa15 and the orthologs of *ALDH1A2* in the zebrafish, stickleback and medaka genomes as well as establishing neighborhoods between the zebrafish/stickleback and zebrafish/medaka genomes. Starting with the zebrafish *aldh1a2* ortholog located on chromosome 7 (Dre7), the micro-synteny algorithm searched for locally conserved synteny within zebrafish/human syntenic clusters (Fig. 5.10A) as well as within zebrafish/stickleback syntenic clusters (Fig. 5.10B). The algorithm found eight locally conserved genes all directly downstream of *aldh1a2* with no gaps in a zebrafish cluster that spanned a total of 6.8 megabases along Dre7. The Dre7 cluster is linked to an 18 megabase cluster on Hsa15 and the eight locally conserved genes are inverted (Fig. 5.10A, red). Likewise, the micro-synteny algorithm found eleven locally conserved genes, with two interleaved gaps, directly downstream of *aldh1a2* when searching zebrafish/stickleback clusters (Fig. 5.10B). As the teleosts are much more closely related to each other than to human, we expect to be able to identify larger, local syntenic regions when comparing them. Six of the same zebrafish genes are involved in both the zebrafish/human and zebrafish/stickleback clusters which provided enough evidence for the pipeline to consider the human and two teleost ALDH1A regions conserved. Next, the analysis is repeated starting with the stickleback ortholog of *ALDH1A2* on chromosome II (GacII). In this case, 19 locally conserved neighboring genes, with two gaps, anchor the region directly upstream and downstream of *ALDH1A2* to the human genome, and five genes, directly downstream, anchors it to the zebrafish. The pipeline now



**FIGURE 5.11:** Ohnologs gone missing as identified by the Teleost OGM Pipeline. Green circles represent OGMs supported by evidence from two teleost species. Red circles represent OGMs supported by evidence from three teleost species.

assembles the two sets of synteny and finds they correspond to one another. Most importantly, of the human orthologs from zebrafish (eight genes) and the human orthologs from stickleback (19 genes), seven of the genes are in common to both analyses, and based on this fact, the pipeline declares that the three human, zebrafish, and stickleback regions are orthologous to one another. An independent analysis by the pipeline, following the same procedure also identifies a conserved region on medaka chromosome 3 containing the medaka *ALDH1A2* ortholog, adding a third orthologous region in the teleosts and bolstering confidence in the conserved synteny found in the human genome.

Having established confidence in our methodology, we identified 27 cases of ohnologs gone missing from the human genome (Fig. 5.11). Of those 27, nine were supported by the zebrafish, stickleback, and medaka genomes (red dots) while the remaining

18 were supported by either the zebrafish/stickleback or zebrafish/medaka genomes (green dots). The OGM do not appear to be distributed in any regular pattern, although five chromosomes contain at least three OGM. Additionally, larger chromosomes do not appear to be more prone to having OGM as the largest three human chromosomes have two or fewer OGM. Interestingly, Hsa7, Hsa10, and Hsa12 have OGM very close to the end of the chromosome leading one to wonder if these genes were lost when ancient chromosomes broke and rearranged themselves. Also, Hsa4, Hsa7, and Hsa11 contain multiple OGM in very close proximity to one another suggesting possible recombination hot-spots.

The small number of identified ohnologs gone missing is commensurate with the conservative nature of our algorithm design. The lost genes we have identified would have existed in the last common ancestor of the teleost and mammalian lineages, hundreds of millions of years ago. As opposed to the studies we reviewed in Section 2.3, we cannot rely on the existence of pseudogenes as the forces of genetic drift would have long ago destroyed any physical remnants of such genes on the chromosome. We therefore have to infer the former location of the gene based on its still existing neighbors. However, as the teleost genomes have undergone a large number of rearrangements since the R3 duplication event (Section 3.2), a permissive algorithm could create many false positives. To increase the number of OGM we can detect, the best approach is to continue to add additional teleost genomes to the pool of data. With multiple, independent lines of evidence we can have confidence in the

pipeline predictions. Second, we can make a general estimate of when a particular human gene was lost by investigating its existing teleost orthologs. For example, if a teleost ortholog of a human OGM exists in mouse, then we know the gene was lost very recently. To establish an estimate of when the gene was lost we can test for the existence of an ortholog to the teleost gene in increasingly distant species (relative to human), such as in chicken and in the chordate amphioxus.

### 5.3 Summary

In this chapter, we built on the dataset provided by the Synteny Database to investigate the effects of differential gene loss following whole genome duplication events. We built two pipelines that could identify a number of architectural features in teleost and human genomes, including R3 ohnologs gone missing and reciprocal gene loss in the teleosts, as well as ohnologs that have been lost in the human lineage since the R3 event. The small number of cases of reciprocal gene loss identified in the teleosts, while interesting, are not enough to make any inferences about the role RGL may play in speciation following a WGD. A careful study of our micro-synteny algorithm may improve our ability to identify locally conserved syntenic regions and the addition of more teleost data to our analysis will provide us with a richer dataset from which to make inferences. If a complete study of the teleosts does not provide significantly more cases of RGL then it would be difficult to argue that RGL is a major driver of speciation. Additional study of the distribution of human ohnologs

gone missing may provide information as to what factors preserve syntenic regions in a genome by indicating where gene loss has occurred; comparing gene loss hot-spots to their orthologous regions in other genomes may provide insights into whether architectural changes in the genome facilitate the loss of syntenic conservation.

## CHAPTER VI

### CONCLUSION

In this work we executed a series of analyses, each building on the previous, to investigate the nature of the genome, exploring evolutionary relationships between genes and within conserved segments of the genome. We applied massive computational resources, based on a series of novel algorithms, to generate over a dozen separate databases. The design of these algorithms focused on how two biological phenomena shape the data from which we wish to draw inferences. First, the evolution of life has been punctuated by whole genome duplication events, a determining force in the architecture of the genome and in the number and distribution of gene copies across the tree of life. Second, the differential loss of genes that follows a whole genome duplication event creates ohnologs gone missing, complicating processes involved in determining evolutionary conservation.

In our first major contribution, we designed and implemented the RBH Analysis Pipeline to assign orthology between genes. Given a primary and an outgroup genome, the pipeline employed a single-linkage clustering algorithm to first create

groups of paralogous genes and then anchor those groups to single genes in the outgroup genome. In addition, the pipeline utilized a novel noise-reduction algorithm free of arbitrary parameters governing its operation. We ran the pipeline against a number of teleost, mammalian, and chordate genomes identifying orthologs in a pattern consistent with the R1, R2, and R3 duplication events and in a number proportional to the evolutionary relatedness of the primary and outgroup genomes. We then used this data to infer the conserved gene order of an ancient human/teleost ancestor.

Building on these datasets, the Synteny Database aggregated paralogous and orthologous gene relationships to define regions of conserved synteny within genomes and between genomes. The Synteny Database is the first system to detect conserved synteny at a fine granularity, presenting the results in an intuitive, web-based interface to the researcher. As part of this second major contribution, we used the Synteny Database to study the evolutionary history of two gene families, using conserved synteny to verify and correct orthology assignments, and to identify instances of ohnologs gone missing.

Our final contribution involved the design of two novel algorithms, modularly built on top of the datasets generated by the RBH Analysis Pipeline and the Synteny Database. These algorithms are the first general methods to infer reciprocally conserved synteny, ohnologs gone missing, and reciprocal gene loss in an arbitrary



genome and we employed them against several teleost genomes as well as the human genome to identify these genomic features.

## 6.1 Future Work

Continuing work in this problem space should proceed along three different tracks. First, the existing pipelines should be run with additional genomes; adding the remaining teleost genomes to the OGM pipelines of Chapter V may increase the number of ohnologs gone missing we can identify. Generally, work to characterize the function and evolutionary history of genes in teleost fish is often driven by the human genome. Since much of the work in the teleosts is in the service of human disease, and since the human genome possesses the richest annotation, researchers tend to study teleost orthologs of interesting human genes. Identifying teleost genes for which there is no longer a human ortholog, because it has become an ohnolog gone missing, instantly creates a list of potentially novel genes in the teleosts for which there has likely been little research. Such a list would be very valuable to the wider research community. Expanding the number of cases of reciprocal gene loss we can identify, by analyzing additional teleost genomes, may provide additional evidence that reciprocal gene loss contributes to speciation. Our ability to investigate this question would be increased greatly if the genome of a much more closely related outgroup to the teleosts, such as the gar, became available.

In the implementation of our OGM pipelines, we introduced a novel algorithm to infer orthology between genes based on locally conserved synteny (Section 5.1.2). A second major track future work should follow is employing this algorithm on a wider scale to bootstrap the Synteny Database. We could use this algorithm to correct misassignments of orthology by the RBH Analysis Pipeline due to asymmetric evolutionary rates between gene duplicates (Section 3.2.2). With corrections made in orthology assignment, we could re-generate the Synteny Database and use the improved clusters of conserved synteny to search for additional misassignments, repeating the process until we see no more improvement. This dataset would contain a map of conserved synteny more accurate than most other algorithmic approaches. A series of synteny maps for a number of genomes could then be used to infer the ancestral architecture of the teleost genome at a higher resolution than has been possible previously.

A third and final track for future research would focus further on the asymmetry of evolutionary rates for gene duplicates. A recent application of the Synteny Database to the ALDH1A gene family [18] has suggested that evolutionary rate asymmetry may extend beyond individual pairs of duplicated genes and may be a more general architectural feature of the genome: one paralogon may be more well conserved than the other. By examining the evolutionary rate variation among pairs of duplicates that make up a well-conserved region of the genome we may be able to empirically

determine the extent to which paralogs have been conserved and if one paralogon is better conserved than another.

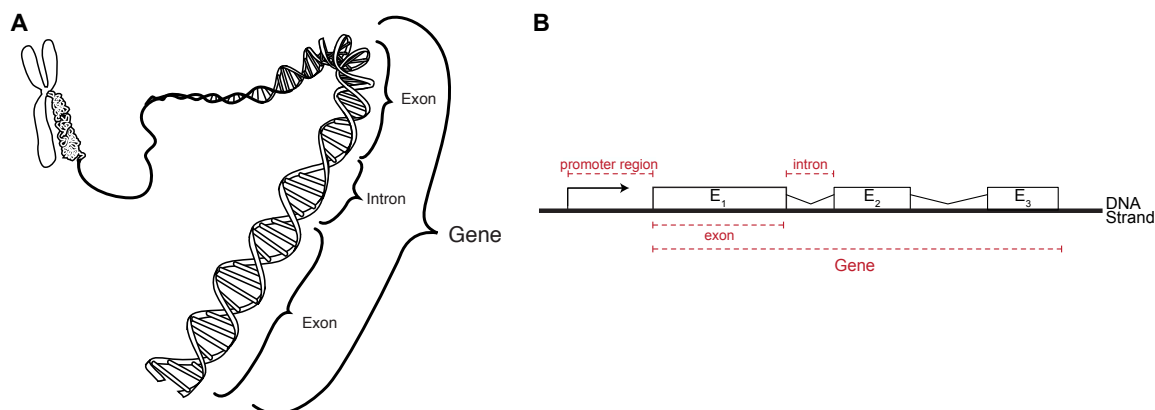
# APPENDIX A

## IMPORTANT BIOLOGICAL CONCEPTS

Although an extensive treatment of DNA and all of the processes involved in its transcription and translation is beyond the scope of this work (see [9], [65], and [69] for an introduction), we will briefly describe some biological concepts as they relate to the topics in this dissertation.

Every living organism contains a linearly arranged set of information that describes a series of genes [126]. These genes describe how to build and execute all the systems that make up the organism, from describing the organism's body plan to the regulation of the number of white blood cells for the immune system. This deoxyribonucleic acid (DNA) is present in every cell of every organism from single-celled bacteria to complex organisms with multiple, cooperating tissue types and internal organs such as mammals.

DNA is composed of four types of nucleotides, which are known by their bases adenine (A), cytosine (C), guanine (G), and thymine (T). These bases can be classified into two categories based on their chemistry, the purines and the pyrimidines, that



**FIGURE A.1:** Two illustrations of a gene. (A) This physical representation shows how the gene is arranged, along with its introns and exons, within the DNA double helix and where it is stored on the chromosome. Illustration from [61]. (B) This representation shows the basic layout of a gene on a strand of DNA along with its functional units. Pictured, from left to right is a promoter region, followed by three exons ( $E_1$ ,  $E_2$ , and  $E_3$ ), separated by two introns (angled lines).

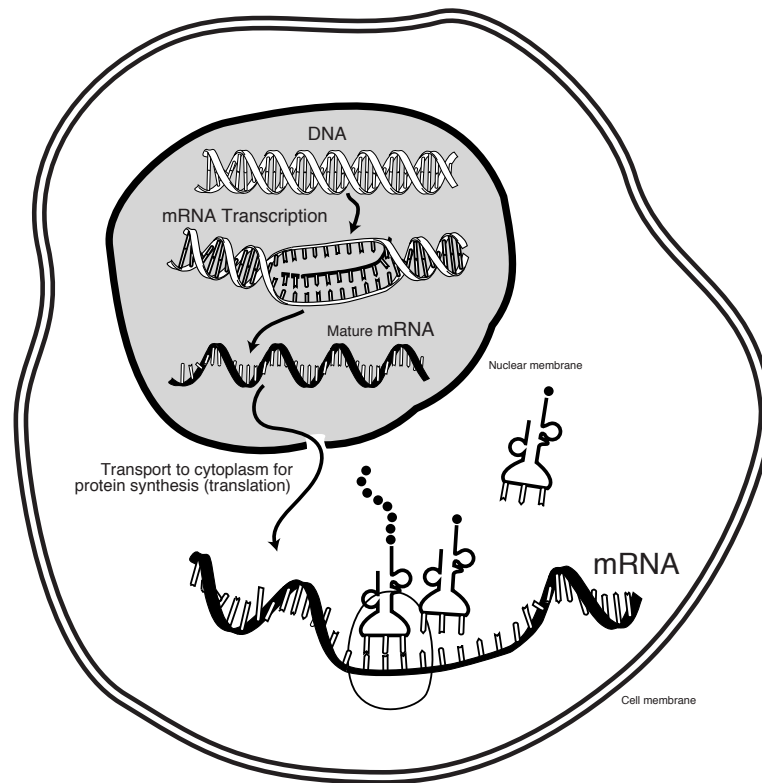
naturally pair with one another – the purine adenine with the pyrimidine thymine as well as the purine guanine with the pyrimidine cytosine. DNA is composed of two strands of these nucleotides, complementary to one another and arranged as a double-helix. Due to this complementary nature, if given one strand of the DNA, the other strand may be re-constructed from it.

The DNA strands encode a series of genes or functional units a portion of which are protein-coding genes. The beginning and ending of each gene is marked by a particular set of nucleotides and, internally, each gene contains one or more exons and introns. Exons and introns are differentiated by the fact that the code specified within an exon will become part of the final protein, whereas the code contained within an intron will be spliced out of the sequence before the final protein is completed. Introns are thought to serve a regulatory role in the production of the protein. Areas

immediately preceding a gene, known as promoter regions, regulate the circumstances under which that gene is read (Fig. A.1B). Interestingly, only a very small fraction of an organism's genome contains code for functional genes, for humans, it is only 3%. The remaining 97%, known as nongenic DNA, was popularly described as “junk” DNA for a time and is not fully understood. Some regions of nongenic DNA are known to be genes that have been rendered non-functional by mutations (commonly referred to as *pseudogenes*), other regions contain highly-repetitive stretches of nucleotides (satellite DNA) [9], while still other portions serve regulatory purposes for protein-coding genes [65]. Large regions of nongenic DNA are filled with self-replicating genetic elements (transposons). These elements can propagate themselves throughout the genome but do not generally serve a functional purpose for the genome's host [9].

In order to create a protein, the internal machinery of the cell first splits the DNA strands and reads the nucleotides belonging to a particular gene. This process is known as transcription and it produces a complementary strand of RNA called the primary transcript. After this initial reading, the primary transcript contains a faithful copy of the DNA including exons as well as introns. The primary transcript is further processed to splice out the introns making messenger RNA (mRNA). During this processing stage, select exons can also be spliced out of the transcript creating a number of *splice variants*, or alternate copies of the gene.

Within the exons of a gene, each group of three nucleotides, known as a codon, specifies a particular amino acid. A gene encodes for a series of amino acids that are



**FIGURE A.2:** An illustration of the transcription and translation process, from [62].

combined to create the protein product of the gene. Since there are four different nucleotide bases, it is possible to encode  $4^3 = 64$  amino acids. However, only twenty different types of amino acids are used in the formation of proteins and, therefore, multiple codons can specify a single amino acid. For this reason, the genetic code is referred to as degenerate [126, 65].

Once processing of the mRNA is complete, the mRNA moves from the cell nucleus into the cytoplasm where cellular ribosomes attach to it and begin the process of translation. During this process, the codons that make up the mRNA are read, and the corresponding amino acids are fetched and attached to one another creating

a chain of polypeptides. As this chain is assembled, the polypeptides fold into a final, three-dimensional protein that, when complete, is then utilized by the organism in some functional way. For example, the protein may act as a signaling protein triggering additional proteins to be synthesized or it may be involved in catalyzing a chemical reaction within the organism. Transcription and translation is illustrated in Figure A.2.

A gene is expressed (transcribed and translated) only at certain times and in certain locations within the organism. The expression of a gene is controlled by a variable number of regulatory regions physically located near the gene on the DNA strand (usually within the promoter region). These regions (referred to as enhancers) serve as binding sites for other proteins (transcription factors) to attach to the DNA and either promote or repress the expression of the regulated gene. Often, a combination of promoters and repressors work together to provide precise expression of a gene in time and space. Each set of distinct expression patterns represents one function of a particular gene, and often, genes have multiple functions. A mutation to an enhancer region upstream of a gene can disable the binding of a transcription factor and hence affect the expression of that gene for one or more functions. Besides expression, as mentioned above, a single gene can also produce multiple splice variants and the production of splice variants is controlled by the same regulatory elements (although the location of these regulatory regions controlling the expression of splice variants is often located within the introns of the gene). Over evolutionary time,



genes can acquire multiple functions, with the ability to produce multiple protein products, expressing those products in different processes occurring very precisely in time and space.

Organisms that have a relatively recent common ancestor share significant portions of their DNA including many protein-coding genes. Many times, a whole gene, portions of a gene, or even whole segments of a chromosome are conserved between organisms. However, because of mutations and other evolutionary changes, the code is rarely identical in different species, or even in different individuals of the same species. Enumerating these differences allows us to make many inferences about the organisms. Due to the degeneracy of the genetic code, comparison of segments of the genetic code translated into amino acids is often more forgiving than those performed with nucleotides since many nucleotide mutations do not alter the resulting amino acid. For this reason, amino acid translations are often used when comparing distantly related sequences.

## APPENDIX B

### SINGLE LINKAGE CLUSTERING ALGORITHM

```

SINGLE_LINKAGE_CLUSTERING(query_genes)
1  ▷ Load forward and reverse BLAST data for the
2  ▷ genes in the primary genome (query_genes)
3  POPULATE_BLAST_DATA(query_genes)
4  do
5  new_predictions ← 0
6  for each query_gene in query_genes
7      if query_gene.for_hits = 0
8          then continue;
9  for each for_hit in query_gene.for_hits
10     ▷ Check for Reciprocal Best Hit
11     result ← CHECK_FOR_RBH(query_gene, for_hit)
12 if result = TRUE
13     then
14         merges ← 0
15         if defined(gene_group_map[query_gene] = FALSE)&&
16         defined(gene_group_map[for_hit] = FALSE)
17             then
18                 NEW_GROUP(groups);
19                 merges +=
20                 MERGE_INTO_GROUP(gene_group_map, groups, query_gene, for_hit)
21         else
22             ▷ Create a unique list of groups where the
23             ▷ query or predicted gene can be found.
24             indexes ← UNIQUE_GROUPS(gene_group_map, query_gene, for_hit)
25             ▷ Merge this new prediction into its respective group.
26             merges +=
27             MERGE_INTO_GROUP(gene_group_map, groups, query_gene, for_hit)
28             ▷ Merge any groups this new prediction links together.
29             for each index in indexes
30                 MERGE_GROUPS(gene_group_map, groups, index)
31         if merges > 0
32             then new_predictions ++
33 while new_predictions > 0

```

CHECK\_FOR\_RBH(*query\_gene*, *for\_hit*)

```
1  rev_hits ← for_hit.rev_hits
2  if COUNT(rev_hits) = 0
3    then
4      return FALSE

5  ▷ Collapse our list of reverse hits to account for already detected paralogs.
6  REDUCE(for_hit, rev_hits)

7  ▷ See if our prediction (the top reverse hit) BLASTed back to the original gene.
8  rev_hit ← rev_hits[0]

9  if query_gene ≠ rev_hit
10   then
11     return FALSE
12 return TRUE
```

## APPENDIX C

### SLIDING WINDOW ALGORITHM

```

1  ▷ Examine each different sliding window size (25, 50, 100, 200)
2  for each sliding_window in sliding_windows

3      ▷ Retrieve an array of chromosomes for the primary (query) genome
4      query_chromosomes ← orgs[org_id]
5      ▷ Now, retrieve an array of chromosomes for the outgroup (pred) genome
6      pred_chromosomes ← orgs[outgroup_id]

7      ▷ Compare each query chromosome against the predicted chromosomes
8      while (COUNT(query_chromosomes) > 0)
9          query_chr ← SHIFT(query_chromosomes)
10         for each pred_chr in pred_chromosomes
11             preds ← query_chr.predictions
12             DEFINE_CLUSTER(preds, clusters, sliding_window, FORWARD)
13             DEFINE_CLUSTER(preds, clusters, sliding_window, INVERTED)

```

```

DEFINE_CLUSTER(predictions, clusters, sliding_window, direction)
1  do
2      num_predictions ← COUNT(predictions)
3
4      for i ← 0 to num_predictions
5          pred ← SHIFT(predictions);
6
7          ▷ Start a new cluster if necessary
8          if (defined(cluster) = FALSE)
9              then NEW_CLUSTER(pred, cluster, sliding_window)
10             continue
11
12         ▷ Check if the predicted gene falls before the start of this cluster.
13         if (pred.location < cluster.start_location)
14             then PUSH(cold_predictions, pred)
15             continue
16
17         ▷ Calculate the distance between the end of the
18         ▷ cluster and the next prediction
19         ▷ on both the query and prediction halves of the sliding window
20         q_dist = GENE_DISTANCE(pred.query, cluster)
21         p_dist = GENE_DISTANCE(pred.pred, cluster)
22
23         if (p_dist < sliding_window && q_dist < sliding_window)
24             then ADD_CLUSTER_MEMBER(pred, cluster)
25             else
26                 if q_dist ≥ sliding_window
27                     then
28                         ▷ We have exhausted the window, close the cluster.
29                         CLOSE_CLUSTER(clusters, cluster, outstanding)
30                         break
31                     else
32                         ▷ Room in the window, place the
33                         ▷ prediction aside for next round
34                         PUSH(cold_predictions, pred, outstanding)
35
36         CLOSE_CLUSTER(clusters, cluster, outstanding)
37
38     ▷ Re-sort the cold predictions and start searching the chain again.
39     PUSH(cold_predictions, predictions)
40     predictions ← SORT(cold_predictions)
41 while outstanding > 0

```

## APPENDIX D

### MICRO-SYNTENY ALGORITHM

```

MICRO-SYNTENY_DETECTION(gene, clusters, genome)
1  ▷ Examine every syntenic cluster that spans gene
2  syn_clusters ← gene.syn_clusters
3  for each syn_id in syn_clusters
4      ▷ Create a sorted index of all the genes in this cluster
5      sorted ← SORT(clusters[syn_id])
6      max_index ← COUNT(sorted) - 1
7      start_index ← gene.index_position
8      i ← start_index - 1, j ← start_index + 1
9      count ← 0, gaps ← 0
10     ▷ If this gene is a member of this cluster, count it as conserved.
11     if sorted[start_index].type = PRESENT
12         then count ++
13     do
14         if i >= 0
15             then
16                 ▷ Keep searching in the same direction until we hit a gap.
17                 while sorted[i].type = PRESENT && i >= 0
18                     count ++
19                     i --
20                 if i >= 0
21                     then i --
22                     gaps ++
23                 ▷ Now change directions and repeat the procedure.
24                 if j <= max_index
25                     then
26                         while sorted[j].type = PRESENT && j <= max_index
27                             count ++
28                             j ++
29                 if j <= max_index
30                     then gaps ++
31                     j ++
32                 while (i >= 0 || j <= max_index) && gaps < GAP_LIMIT
33                 if count < NEIGHBORS
34                     then continue
35                 o.syn_id ← syn_id, o.neighbors ← count
36                 PUSH(gene.ogm, o);

```

# BIBLIOGRAPHY

- [1] F. W. Allendorf and G. H. Thorgaard. Tetraploidy and the evolution of salmonid fishes. In B. Turner, editor, *The Evolutionary genetics of fishes*, pages 1–53. Plenum Publishing, New York, 1984.
- [2] J. Altschmied, J. Delfgaauw, B. Wildea, J. Duschla, L. Bouneaub, J. Volffa, and M. Scharl. Subfunctionalization of Duplicate *mitf* Genes Associated With Differential Degeneration of Alternative Exons in Fish. *Genetics*, 161:259–267, May 2002.
- [3] S. F. Altschul and W. Gish. Local Alignment Statistics. *Methods in Enzymology*, 266:460–480, 1996.
- [4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, May 1990.
- [5] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
- [6] C. T. Amemiya, S. J. Prohaska, A. Hill-Force, A. Cook, J. Wasserscheid, D. E. Ferrier, J. Pascual-Anaya, J. Garcia-Fernàndez, K. Dewar, and P. F. Stadler. The Amphioxus *Hox* Cluster: Characterization, Comparative Genomics, and Evolution. *Journal of Experimental Zoology (Mol Dev Evol)*, 310B, 2008.
- [7] A. Amores, A. Force, Y. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. Zebrafish *hox* Clusters and Vertebrate Genome Evolution. *Science*, 282(5394):1711–1714, November 1998.
- [8] A. Amores, T. Suzuki, Y. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J. H. Postlethwait. Developmental roles of pufferfish *Hox* clusters and genome evolution in ray-fin fish. *Genome Research*, 14(1):1–10, January 2004.
- [9] N. Barton, D. Briggs, J. Eisen, D. Goldstein, and N. Patel. *Evolution*. Cold Spring Harbor Laboratory Press, 2007.

- [10] A. Berglund, E. Sjölund, G. Östlund, and E. L. L. Sonnhammer. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, 38:D263–D266, 2008.
- [11] D. Bikard, D. Patel, C. L. Mett , V. Giorgi, C. Camill, M. J. Bennett, and O. Loudet. Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within *A. thalianac*. *Science*, 323:623–626, 2009.
- [12] E. Birney, T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyra, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard, and M. Clamp. An Overview of Ensembl. *Genome Research*, 14(5):925–928, May 2004.
- [13] J. E. Blair and S. B. Hedges. Molecular Phylogeny and Divergence Times of Deuterostome Animals. *Molecular Biology and Evolution*, 22(11):2275–2284, 2005.
- [14] I. Braasch, J. Volff, and M. Schartl. The Endothelin System: Evolution of Vertebrate-Specific Ligand-Receptor Interactions by Three Rounds of Genome Duplication. *Molecular Biology and Evolution*, 2009.
- [15] J. T. Bridgham, J. E. Brown, A. Rodr guez-Mar , J. M. Catchen, and J. W. Thornton. Evolution of a New Function by Degenerative Mutation in Cephalochordate Steroid Receptors. *PLoS Genetics*, 4(9), 2008.
- [16] K. P. Byrne and K. H. Wolfe. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, 15:1456–1461, 2005.
- [17] K. P. Byrne and K. H. Wolfe. Consistent Patterns of Rate Asymmetry and Gene Loss Indicate Widespread Neofunctionalization of Yeast Genes After Whole-Genome Duplication. *Genetics*, 175, 2007.
- [18] C. Ca estro, J. Catchen, A. Rodr guez-Mar , H. Yokoi, and J. Postlethwait. Consequences of Lineage-Specific Gene Loss on Functional Evolution of Surviving Paralogs: ALDH1A and Retinoic Acid Signaling in Vertebrate Genomes. *PLoS Genetics*, 5(5):e1000496, February 2009.
- [19] J. M. Catchen, J. S. Conery, and J. H. Postlethwait. Inferring Ancestral Gene Order. *Methods in Molecular Biology*, 452:365–383, 2008.



- [20] J. M. Catchen, J. S. Conery, and J. H. Postlethwait. Automated identification of conserved synteny after whole genome duplication. *Genome Research*, May 2009.
- [21] N. Cermakian, D. Whitmore, N. S. Foulkes, and P. Sassone-Corsi. 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, 2000.
- [22] J. S. Conery, J. M. Catchen, and M. Lynch. Rule-based workflow management for bioinformatics. *VLDB Journal*, 14(3):318–329, 2005.
- [23] L. David, S. Blum, M. W. Feldman, U. Lavi, and J. Hillel. 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, 2003.
- [24] D. Davidson. The function and evolution of *Msx* genes: pointers and paradoxes. *Trends in Genetics*, 11(10):405–411, 1995.
- [25] P. Dehal and J. L. Boore. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology*, 3(10):1700–1708, October 2005.
- [26] F. Delsuc, H. Brinkmann, D. Chourrout, and H. Philippe. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439:965–968, 2006.
- [27] F. Delsuc, G. Tsagkogeorga, N. Lartillot, and H. Philippe. Additional molecular support for the new chordate phylogeny. *Genesis*, 46(11):592–604, 2008.
- [28] J. Dufayard, L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perrière. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603, 2005.
- [29] S. Eddy. What is Bayesian Statistics? *Nature Biotechnology*, 22:1177–1178, 2004.

- [30] M. Ekker, M.-A. Akimenko, M. L. Allende, R. Smith, G. Drouin, R. M. Lungille, E. S. Weinberg, and M. Westerfield. Relationships Among *msx* Gene Structure and Function in Zebrafish and Other Vertebrates. *Molecular Biology and Evolution*, 14(10):1008–1022, 1997.
- [31] G. Elgar, M. Clark, A. Green, and R. Sandford. How good a model is the Fugu genome? *Nature*, 387(140), 1997.
- [32] A. Enright, S. V. Dongen, and C. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30:1575–1584, 2002.
- [33] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. Yan, , and J. Postlethwait. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, 151:1531–1545, 1999.
- [34] P. G. Foster. *The Idiot’s Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies, Unleashed*. 2001.
- [35] J. Garcia-Fernàndez. The genesis and evolution of homeobox gene clusters. *Nature Reviews Genetics*, 6:881–892, 2005.
- [36] J. Garcia-Fernàndez and P. W. H. Holland. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature*, 370:563–566, 1994.
- [37] N. Gekakis, D. Staknis, H. B. Nguyen, F. C. Davis, L. D. Wilsbacher, D. P. King, J. S. Takahashi, and C. J. Weitzcircadian. Role of the CLOCK protein in the mammalian circadian mechanism. *Science*, 280(5369):1564–1569, June 1998.
- [38] W. Gish. WU BLAST, 2003.
- [39] S. Guindon and O. Gascuel. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704, 2003.
- [40] M. W. Hahn, M. V. Han, and S. Han. Gene Family Evolution across 12 *Drosophila* Genomes. *PLoS Genetics*, 3(11):e197, 2007.
- [41] S. Hoegg, J. L. Boore, J. V. Kuehl, and A. Meyer. Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*. *BMC Genomics*, 8:317, 2007.

- [42] J. B. Hogenesch, Y. Gu, S. M. Moran, K. Shimomura, L. A. Radcliffe, J. S. Takahashi, and C. A. Bradfield. The Basic Helix-Loop-Helix-PAS Protein MOP9 Is a Brain-Specific Heterodimeric Partner of Circadian and Hypoxia Factors. *Journal of Neuroscience*, 20, 2000.
- [43] P. W. H. Holland and J. Garcia-Fernàndez. *Hox* genes and chordate evolution. *Developmental Biology*, 173:382–395, 1996.
- [44] J. P. Huelsenbeck and K. A. Crandall. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review Ecological Systems*, 28:437–466, 1997.
- [45] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, March 2001.
- [46] A. L. Hughes and R. Friedman. 2R or not 2R: testing hypotheses of genome duplication in early vertebrates. *Journal of Structural and Functional Genomics*, 3:85–93, 2003.
- [47] M. Ikeda and M. Nomura. 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, April 1997.
- [48] T. Ikuta and H. Saiga. Organization of Hox Genes in Ascidians: Present, Past, and Future. *Developmental Dynamics*, 233:382–389, 2005.
- [49] J. G. Inoue, M. Miyab, K. Tsukamotoa, and M. Nishida. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the “ancient fish”. *Molecular Phylogenetics and Evolution*, 26(1):110–120, 2003.
- [50] T. Ishikawa, J. Hirayama, Y. Kobayashi, and T. Todo. Zebrafish CRY represses transcription mediated by CLOCK-BMAL heterodimer without inhibiting its binding to DNA. *Genes to Cells*, 7:1073–1086, 2002.

- [51] O. Jaillon, J. Aury, F. Brunet, J. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biéumont, Z. Skalli, L. Cattolico, J. Poulain, V. de Berardinis, C. Cruaud, S. Duprat, P. Brottier, J. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J. Volff, R. Guigó, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quétier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. R. Crollius. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, October 2004.
- [52] O. Jarinova, G. Hatch, L. Poitras, C. Prudhomme, M. Grzyb, J. Aubin, F. Bérubé-Simard, L. Jeannotte, and M. Ekker. Functional resolution of duplicated *hoxb5* genes in teleosts. *Development*, 135:3543–3553, 2008.
- [53] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282, 1992.
- [54] R. Jovelin, X. He, A. Amores, Y. Yan, R. Shi, B. Qin, B. Roe, W. A. Cresko, and J. H. Postlethwait. Duplication and divergence of *fgf8* functions in teleost development and evolution. *Journal of Experimental Zoology*, 308B(6):730 – 743, 2007.
- [55] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Research*, 6(1):31, January 2004.
- [56] M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.

- [57] H. Kikuta, M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engström, D. Fredman, A. Akalin, M. Caccamo, I. Sealy, K. Howe, J. Ghislain, G. Pezeron, P. Mourrain, S. Ellingsen, A. C. Oates, C. Thisse, B. Thisse, I. Foucher, B. Adolf, A. Geling, B. Lenhard, and T. S. Becker. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, 17:545–555, 2007.
- [58] S. Kuraku, A. Meyer, and S. Kuratani. Timing of Genome Duplications Relative to the Origin of the Vertebrates: Did Cyclostomes Diverge before or after? *Molecular Biology and Evolution*, 26(1):47–59, 2009.
- [59] D. Larhammar and C. Risinger. Molecular Genetic Aspects of Tetraploidy in the Common Carp *Cyprinus carpio*. *Molecular Phylogenetics and Evolution*, 3(1):59–68, 1994.
- [60] J. Lehmann, P. F. Stadler, and S. J. Prohaska. SynBlast: Assisting the analysis of conserved synteny information. *BMC Bioinformatics*, 9:351, 2008.
- [61] D. Leja. Gene, National Human Genome Research Institute (<http://www.accessexcellence.org/rc/vl/gg/gene.html>).
- [62] D. Leja. mrna, National Human Genome Research Institute (<http://www.accessexcellence.org/RC/VL/GG/mRNA.html>).
- [63] L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13:2178–2189, 2003.
- [64] W. Li. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics*, 95(1):237–258, 1980.
- [65] W. Li. *Molecular Evolution*. Sinauer Associates, 1997.
- [66] W. Li, Z. Gu, A. R. O. Cavalcanti, and A. Nekrutenko. Detection of gene duplications and block duplications in eukaryotic genomes. *Journal of Structural and Functional Genomics*, 3:27–34, 2003.
- [67] M. Lynch and J. S. Conery. The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494):1151 – 1155, 2000.

- [68] M. Miya, H. Takeshima, H. Endo, N. B. Ishiguro, J. G. Inoue, T. Mukai, T. P. Satoh, M. Yamaguchi, A. Kawaguchi, K. Mabuchi, S. M. Shiraid, and M. Nishida. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, 26(1):121–138, 2003.
- [69] D. Mount. *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2001.
- [70] MPI Group. *MPICH2*. Argonne National Laboratory, February 2009.
- [71] S. Mungpakdee, H. Seo, A. R. Angotzi, X. Dong, A. Akalin, and D. Chourrout. Differential Evolution of the 13 Atlantic Salmon Hox Clusters. *Molecular Biology and Evolution*, 25(7):1333–1343, 2008.
- [72] S. Mungpakdee, H. Seo, and D. Chourrout. Spatio-temporal expression patterns of anterior Hox genes in Atlantic salmon (*Salmo salar*). *Gene Expression Patterns*, 8:508–514, 2008.
- [73] Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Research*, 17:1254–1265, 2007.
- [74] K. Naruse, M. Tanaka, K. Mita, A. Shima, J. Postlethwait, and H. Mitani. A Medaka Gene Map: The Trace of Ancestral Vertebrate Proto-Chromosomes Revealed by Comparative Gene Mapping. *Genome Research*, 14(5):820–828, May 2004.
- [75] H. Oda, H. Wada, K. Tagawa, Y. Akiyama-Oda, N. Satoh, T. Humphreys, S. Zhang, and S. Tsukita. A novel amphioxus cadherin that localizes to epithelial adherens junctions has an unusual domain organization with implications for chordate phylogeny. *Evolution and Development*, 4(6):426–434, 2002.
- [76] S. Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.
- [77] M. P. Pando and P. Sassone-Corsi. Unraveling the mechanisms of the vertebrate circadian clock: zebrafish may light the way. *BioEssays*, 24:419–426, 2002.
- [78] G. Panopoulou and A. J. Poustka. Timing and mechanism of ancient vertebrate genome duplications - the adventure of a hypothesis. *Trends in Genetics*, 21(10):559–567, 2005.
- [79] M. Pébusque, F. Coulier, D. Birnbaum, and P. Pontarotti. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Molecular Biology and Evolution*, 15(9):1145–1159, 1998.

- [80] G. Perrière, L. Duret, and M. Gouy. HOBACGEN: Database System for Comparative Genomics in Bacteria. *Genome Research*, 10:379–385, 2000.
- [81] H. Philippe, N. Lartillot, and H. Brinkmann. Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution*, 22(5):1246–1253, 2005.
- [82] J. Postlethwait, A. Amores, W. Cresko, A. Singer, and Y. Yan. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends in Genetics*, 20(10):481–490, 2004.
- [83] J. H. Postlethwait. The Zebrafish Genome: A Review and *msx* Gene Case Study. *Vertebrate Genomes*, 2:183–197, 2006.
- [84] J. H. Postlethwait. The zebrafish genome in context: ohnologs gone missing. *Journal of Experimental Zoology (Mol Dev Evol)*, 308B(5):563–577, October 2006.
- [85] J. H. Postlethwait, Y. Yan, M. A. Gates, S. Horne, A. Amores, A. Brownlie, A. Donovan, E. S. Egan, A. Force, Z. Gong, C. Goutel, A. Fritz, R. Kelsh, E. Knapik, E. Liao, B. Paw, D. Ransom, A. Singer, M. Thomson, T. S. Abduljabbar, P. Yelick, D. Beier, J. Joly, D. Larhammar, F. Rosa, M. Westerfield, L. I. Zon, S. L. Johnson, and W. S. Talbo. Vertebrate genome evolution and the zebrafish gene map. *Nature Genetics*, 18:345 – 349, 1998.
- [86] N. H. Putnam, T. Butts, D. E. K. Ferrier, R. F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. Terry, J. Yu, E. Benito-Gutiérrez, I. Dubchak, J. Garcia-Fernández, J. J. Gibson-Brown, I. V. Grigoriev, A. C. Horton, P. J. de Jong, J. Jurka, V. V. Kapitonov, Y. Kohara, Y. Kuroki, E. Lindquist, S. Lucas, K. Osoegawa, L. A. Pennacchio, A. A. Salamov, Y. Satou, T. Sauka-Spengler, J. Schmutz, T. Shin-I, A. Toyoda, M. Bronner-Fraser, A. Fujiyama, L. Z. Holland, P. W. H. Holland, N. Satoh, and D. S. Rokhsar. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453:1064–1071, 2008.
- [87] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.
- [88] C. Ramos and B. Robert. *msh/Msx* gene family in neural development. *Trends in Genetics*, 21(11):624–632, 2005.
- [89] M. Remm, C. E. V. Storm, and E. L. L. Sonnhammer. Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *Journal of Molecular Biology*, 314:1041–1052, 2001.

- [90] C. Risinger and D. Larhammar. Multiple loci for synapse protein SNAP-25 in the tetraploid goldfish. *Proceedings of the National Academy of Sciences of the USA*, 90:10598–10602, 1993.
- [91] J. F. Ryan, P. M. Burton, M. E. Mazza, G. K. Kwong, J. C. Mullikin, and J. R. Finnerty. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biology*, 7(7), 2006.
- [92] Y. Satou, K. S. Imai, M. Levine, Y. Kohara, D. Rokhsar, and N. Satoh. A genomewide survey of developmentally relevant genes in *Ciona intestinalis* I. Genes for bHLH transcription factors. *Development Genes and Evolution*, 213:213–221, 2003.
- [93] D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440:341–345, 2006.
- [94] M. Schmid and C. Steinlein. Chromosome banding in Amphibia. XVI. High-resolution replication banding patterns in *Xenopus laevis*. *Chromosoma*, 101(2):123–132, 1991.
- [95] M. Sémon and K. H. Wolfe. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends in Genetics*, 23(3):108–112, 2007.
- [96] S. M. Shimeld, I. J. McKay, and P. T. Sharpe. The murine homeobox gene *Msx-3* shows highly restricted expression in the developing neural tube. *Mechanisms of Development*, 55(2):210–210, April 1996.
- [97] C. Simillion, K. Janssens, L. Sterck, and Y. Van de Peer. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, 24(1):127–128, 2008.
- [98] C. Simillion, K. Vandepoele, Y. Saeys, and Y. Van de Peer. Building Genomic Profiles for Uncovering Segmental Homology in the Twilight Zone. *Genome Research*, 14:1095–1106, 2004.
- [99] N. G. Smith, R. Knight, and L. D. Hurst. Vertebrate genome evolution: a slow shuffle or a big bang? *Bioessays*, 21(8):697 – 703, 1999.
- [100] J. Spring. Vertebrate evolution by interspecific hybridisation – are we polyploid? *Federation of European Biochemical Societies*, 400(1):2–8, 1997.
- [101] R. Stallman, R. McGrath, and P. Smith. *GNU Make*. The Free Software Foundation, February 2009.



- [102] G. Sundström, T. A. Larsson, and D. Larhammar. Phylogenetic and chromosomal analyses of multiple gene families syntenic with vertebrate Hox clusters. *BMC Evolutionary Biology*, 8:254, 2008.
- [103] M. Suyama, E. Harrington, P. Bork, and D. Torrents. Identification and Analysis of Genes and Pseudogenes within Duplicated Regions in the Human and Mouse Genomes. *PLoS Computational Biology*, 2(6):e76, 2006.
- [104] J. S. Taylor, I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. *Genome Research*, 13:382–390, 2003.
- [105] J. S. Taylor, Y. Van de Peer, I. Braasch, and A. Meyer. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philosophical Transactions: Biological Sciences*, 356:1661–1679, 2001.
- [106] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423, 2001.
- [107] O. Turunen, R. Seelke, and J. Macosko. *In silico* evidence for functional specialization after genome duplication in yeast. *Federation of European Microbiological Societies*, 9:16–31, 2008.
- [108] T. Uyeno and G. R. Smith. Tetraploid Origin of the Karyotype of Catostomid Fishes. *Science*, 175(4022):644–646, 1972.
- [109] Y. Van de Peer. Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics*, 5:752–763, 2004.
- [110] A. van Hoof. Conserved Functions of Yeast Genes Support the Duplication, Degeneration and Complementation Model for Gene Duplication. *Genetics*, 171(4):1455–1461, 2005.
- [111] K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van de Peer. The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between *Arabidopsis* and Rice. *Genome Research*, 12:1792–1801, 2002.
- [112] A. Vienne and P. Pontarotti. Metaphylogeny of 82 gene families sheds a new light on chordate evolution. *Int J Biol Sci*, 2:32–37, 2006.
- [113] H. Wada, M. Okuyama, N. Satoh, and S. Zhang. Molecular evolution of fibrillar collagen in chordates, with implications for the evolution of vertebrate skeletons and chordate phylogeny. *Evolution and Development*, 8(4):370–377, 2006.

- [114] D. P. Wall, H. B. Fraser, and A. E. Hirsh. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711, 2003.
- [115] H. Wang. Comparative genomic analysis of teleost fish *bmal* genes. *Genetica*, 136(1):149–161, October 2009.
- [116] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61, 2007.
- [117] G. A. Watterson. On the Time for Gene Silencing at Duplicate Loci. *Genetics*, 105(3):745–766, 1983.
- [118] C. Winkler, M. Schäfer, J. Duschl, M. Schartl, and J.-N. Volf. Functional Divergence of Two Zebrafish Midkine Growth Factors Following Fish-Specific Gene Duplication. *Genome Research*, 13:1067–1081, 2003.
- [119] K. Wolfe. Robustness – it’s not where you think it is. *Nature Genetics*, 25:3–4, May 2000.
- [120] J. M. Woltering and A. J. Durston. The zebrafish *hoxDb* cluster has been reduced to a single microRNA. *Nature Genetics*, 38:601–602, 2006.
- [121] I. G. Woods, C. Wilson, B. Friedlander, P. Chang, D. K. Reyes, R. Nix, P. D. Kelly, F. Chu, J. H. Postlethwait, and W. S. Talbot. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Research*, 15:1307–1314, August 2005.
- [122] Z. Yang. *Mathematics of Evolution and Phylogeny*, chapter Bayesian Inference in Molecular Phylogenetics, pages 63–90. Oxford University Press, 2005.
- [123] J. Zhang. Parallel Functional Changes in the Digestive RNases of Ruminants and Colobines by Divergent Amino Acid Substitutions. *Molecular Biology and Evolution*, 20(8):1310–1317, 2003.
- [124] J. Zhang, Y. ping Zhang, and H. F. Rosenberg. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics*, 30:411–415, March 2002.
- [125] J. Zhu, J. Z. Sanborn, M. Diekhans, C. B. Lowe, T. H. Pringle, and D. Haussler. Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage. *PLoS Computational Biology*, 3(12):2498–2509, 2007.
- [126] E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8:357–366, 1965.