

Origin and evolution of novel sequences by gene duplication

by

Raquel Assis

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2011

Doctoral Committee:

Professor Alexey S. Kondrashov, Chair
Professor Brian Athey
Associate Professor Noah A. Rosenberg
Assistant Professor John Kim
Assistant Professor Patricia J. Witkopp

© Raquel Assis 2011

All Rights Reserved

To my parents, Avi and Debbie Assis

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Alexey Kondrashov, for his unwavering academic and emotional support throughout the last few years. Working with Alex has been one of the most rewarding and unique experiences of my life. Unlike most advisors, Alex is always willing to discuss science, whether in a typical office setting, as my bird is perched on his shoulder, or while building a wall outside his house. Without him, I would never have discovered the crying room in LSI, laughed even half as many times, or fallen in love with a turtle and fuzzy yellow caterpillar named Fred. Though I will soon lose him as an advisor, I am grateful to have gained him as a colleague and friend.

I would also like to thank Fyodor Kondrashov, who, though a Ph.D. student himself, advised me on the first project of my thesis. My transition from experimental to computational biology was difficult, but seamless, under Fedya's guidance. He has also become a great friend and has continued to be supportive throughout my graduate career. Even though he is now working in Barcelona, my emailed questions never go unanswered for more than a day.

Next, I would like to thank my committee—Brian Athey, John Kim, Noah Rosenberg, and Patricia Wittkopp—for their helpful comments and discussions. I would like to extend a special thanks to Noah Rosenberg and Trisha Wittkopp for their support with fellowship and postdoctoral applications. While not a committee member, I would also like to thank Shamil Sunyaev for his helpful discussions on gene conversion.

Though I am now the only member of the Kondrashov lab, I would like to acknowledge all of the past members for their support and guidance. In particular, I would like to thank my officemate, Olga Grushko, who has become a great friend over the years. With her support, I have made it through some difficult problems. I would also like to thank Anya Gerasimova, a great friend with whom I enjoyed sharing many annoying programming quirks. Though he only spent a short time in our lab, I would like to acknowledge Chris Duchesneau for his friendship and helpful conversations during my first year in the Kondrashov lab. Additionally, I would like to thank an unofficial member of our lab, Yegor Bazykin, who has become another great friend and colleague. Though he lives in Russia, he is always quick to answer my questions by email or Skype.

I would like to extend a special thanks to the Rosenberg lab, which has become like a second lab to me. Their friendship and support has helped me through many obstacles during the last few years. Specifically, I want to thank Lucy Huang for helping me with a programming problem and with some statistical analyses, and Trevor Pemberton for constantly cooking delicious food and baking me birthday cakes every year.

I would next like to acknowledge those in my program and building who have made my daily life at Michigan more pleasant. In Bioinformatics, I would like to thank Dan Burns for the helpful discussions and coffee, Jeff de Wet for his support and interesting conversations, and Sandy Hall for her friendship and abundant supply of candy. In LSI, I would like to thank the IT administrator, John Herlocher, for letting me remotely access some computers to run Bridges on, and my custodian, Oli Hudson, for always being there to talk to and make me smile.

I would also like to acknowledge the Wang lab at the University of Florida for their influence in my decision to obtain a Ph.D. In particular, I would like to thank Kevin Wang for giving me the opportunity to perform research at the undergraduate level,

Stephen Larner for guiding me through my research project, Meghan O'Donoghue for her help with various techniques, and Shankar Sadasivan, who specifically encouraged me to come to Michigan.

My remaining acknowledgments go to those who played important roles in my life and helped to shape me into who I am today. First and foremost, I would like to thank my parents, Avi and Debbie Assis, for their unconditional love and support. Though my father passed away several years ago, he is always in my thoughts and has a major influence on all of my academic decisions. My mother, who faced the difficult task of raising three kids by herself, is an inspiration to me. She has always believed in me and, as a result, I strive to make her proud through academic success.

Next, I would like to thank my two sisters, Joy and Jackie Assis, for always being there for me. As a child, I fought with them all the time, but as adults, they have become my closest friends. Also in this category is my cousin, Nicole Zaldeti, who has always been the big sister I never had.

I would also like extend my gratitude to my grandparents, Chaim and Ruth Assis from my father's side, and Philip and Olga Kubie from my mother's side. Though my grandfathers passed away several years ago, their influence on me lives on. My grandmothers have been extremely supportive throughout my life, and especially during the last few years, even though they have no idea what it is I do.

I would next like to thank my aunts, uncles, and cousins for their love and support over the years. Specifically, I would like to thank my aunt Lisa for driving me to school and supporting me with my career goals, my uncle Norman for teaching me how to drive and making me laugh with his corny jokes, my aunt Debbie for her friendship and support, and my uncle Jaime, who, despite never answering his phone, has always been there for me.

My next acknowledgment goes to my boyfriend, Michael DeGiorgio, who stood by me through a number of academic and personal difficulties. He supported me as I

first learned how to program a few years ago and has helped me countless times with programming problems since then. No matter what he is doing, he is always willing to drop everything when I need help or just someone to talk to. Mike is also a great person and friend, and I am very grateful to have him in my life.

I am also deeply indebted to Mike's parents, Meryl and Sal DeGiorgio, who took me in as their own and are like second parents to me. They have always been there for me and, like parents, provided me with unconditional love and support over the years. Here, I would also like to acknowledge some of Mike's extended family. Specifically, I want to thank Carmen Rodriguez, Doreen and Frank Charon, Andy Rodriguez, Crystal Rodriguez, Jodie Staton, and Tony Valle, for their love and support.

I would also like to acknowledge all of my friends for their support over the years. In particular, I would like to thank my childhood best friend, Deb Gober, and another great friend of mine, Joel Marino. Deb is one of the most caring and optimistic people I know and, without her support, I never would have made it through some of the biggest obstacles I encountered in my life. Joel has also been there for me over the years and has played a particularly important role in some of my career choices.

Last, but definitely not least, I would like to thank my green-cheeked conure, Doomies. She is a great listener and always knows how to cheer me up with kisses.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	xi
CHAPTER	
I. Introduction	1
II. Nested genes and increasing organizational complexity of metazoan genomes	13
2.1 Introduction	13
2.2 Evolutionary dynamics of nested gene structures	14
2.3 Acquisition of nested gene structures	15
2.4 No functional significance of nested gene structures	16
2.5 Predicting the course of genome structure evolution	17
2.6 Conclusions and perspective	18
2.7 Methods	19
2.7.1 Identification and quality control of nested gene pairs	19
2.7.2 Comparative genomic analysis of nested gene structures	20
2.7.3 Analysis of gene expression	21
2.7.4 Estimating the rate of nested gene evolution	21
2.7.5 Genomes used in this study	22
III. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution	42
3.1 Introduction	42
3.2 Results	43

3.2.1	Recent acquisition of many piRNA clusters	43
3.2.2	Ectopic recombination as a mechanism of piRNA cluster origin	43
3.2.3	Two distinct subpopulations of clusters	45
3.2.4	Unremarkable small-scale evolution of piRNA clusters	46
3.3	Discussion	46
3.4	Materials and methods	48
3.4.1	Classification of rodent piRNA cluster	48
3.4.2	Phylogenetic analysis	48
3.4.3	Identification of paralogs	48
3.4.4	Measurement of small-scale evolution	49
3.5	Acknowledgments	49
IV. Bridges: a tool for identifying local similarities in long sequences		88
4.1	Introduction	88
4.2	Implementation	89
4.2.1	Filtering input sequences	90
4.2.2	Identifying local similarities	90
4.2.3	Post-processing local similarities	91
4.3	Discussion	92
4.4	Acknowledgements	93
V. A strong deletion bias in nonallelic gene conversion		94
5.1	Introduction	94
5.2	Results and Discussion	95
5.3	Methods	97
5.3.1	Estimation of the proportion of gene conversion-consistent indels attributed to ordinary mutation	98
5.3.2	Ascertainment of the effect of sequencing errors on our observations	100
VI. Conclusion		106
BIBLIOGRAPHY		111

LIST OF FIGURES

<u>Figure</u>		
1.1	Mechanisms of novel gene acquisition.	9
1.2	Life cycle of a mutation within a pair of paralogs that undergo gene conversion.	10
1.3	Phenotypic outcomes of duplicate genes.	11
1.4	Nested gene structure.	12
2.1	Phylogenetic analysis of gains and losses of nested gene structures.	36
2.2	Dynamics of gain and loss of nested gene structures.	37
2.3	Scenarios for the origin of a nested gene structure.	38
2.4	Distributions of total coding sequence lengths of external, internal, and non-nested genes in a) <i>H. sapiens</i> , b) <i>D. melanogaster</i> , and c) <i>C. elegans</i> genomes.	39
2.5	Distributions of numbers of exons in external, internal, and non-nested genes in a) <i>H. sapiens</i> , b) <i>D. melanogaster</i> , and c) <i>C. elegans</i> genomes.	40
2.6	Distributions of total gene lengths of external, internal, and non-nested genes in a) <i>H. sapiens</i> , b) <i>D. melanogaster</i> , and c) <i>C. elegans</i> genomes.	41
3.1	Acquisition of a cluster-harboring sequence.	86
3.2	Schematic used to identify ectopic recombination as the mechanism of an insertion.	87

5.1	A phylogenetic approach for detecting insertions and deletions in non-allelic gene conversion.	102
5.2	Properties of paralogs.	103
5.3	Indels consistent with gene conversion.	104
5.4	A phylogenetic approach for detecting fixed indels.	105

LIST OF TABLES

Table

2.1	Mechanisms of origin of human internal genes	23
2.2	Human recently evolved nested gene pairs	24
2.3	<i>Drosophila melanogaster</i> recently evolved nested gene pairs	33
2.4	<i>Caenorhabditis elegans</i> recently evolved nested gene pairs	35
3.1	Rat and mouse piRNA clusters	50
3.2	piRNA clusters acquired after rat-mouse divergence	82

CHAPTER I

Introduction

The field of evolutionary genetics is centered around a single question: How do genetic changes give rise to novel phenotypes in evolution? Long before the molecular nature of genes had been uncovered, a number of geneticists hypothesized that new functions could emerge by copying and modifying preexisting genes, a process now known as gene duplication (Figure 1.1A) (*Bridges*, 1918; *Haldane*, 1933; *Muller*, 1935). However, it was not until 1970, when Susumu Ohno published his classic book “Evolution by Gene Duplication” (*Ohno*, 1970), that gene duplication was widely recognized by the scientific community. Ohno stated that, while mutations in existing genes can account for within-species differentiation or adaptive radiation from an immediate ancestor, they cannot cause large changes in evolution, because these changes occur via the acquisition of new genes with previously nonexistent functions (*Ohno*, 1970). Furthermore, he argued that all new genes must arise via gene duplication (*Ohno*, 1970). Though the mapping from genotype to phenotype is much more complex than imagined at that time, and likely also involves changes in noncoding DNA, molecular genetic studies have confirmed that gene acquisition is a key process in the evolution of novel phenotypes (*Zhang*, 2003). Moreover, almost all new genes can be traced back to ancestral genes (*Lynch*, 2002b; *Nei and Rooney*, 2005; *Zhou et al.*, 2008), revealing gene duplication to be the primary mechanism of novel gene

acquisition.

Another mechanism of novel gene acquisition is retrotransposition, in which a spliced mRNA from an existing gene is reverse transcribed and inserted into a different genomic location (Figure 1.1*B*). Because it results in a new gene copy, or “retrogene”, that lacks introns, retrotransposition is often considered a subtype of gene duplication. While most retrogenes are initially not functional due to the absence of upstream regulatory elements (*Brosius, 1991*), some can gain function by recruiting regulatory sequences from nearby genes (*Drouin and Dover, 1990; Long and Langley, 1993; Martignetti and Brosius, 1993; Charlesworth et al., 1998; Courseaux and Nahon, 2001; Wang et al., 2002*). Such retrogenes often also recruit flanking noncoding DNA and coding regions of nearby genes, forming chimeric structures (*Marques et al., 2005; Wang et al., 2006; Bai et al., 2007*). Thus, retrotransposition can lead to the creation of structurally and functionally unique genes and is believed to be the second most common mechanism in the acquisition of novel genes (*Kaessmann et al., 2009*).

A third mechanism of gene acquisition is horizontal gene transfer (HGT), also called lateral gene transfer (LGT), in which genes are directly transferred between species or between organelles and nuclei within an organism (Figure 1.1*C*). Though HGT occurs frequently between prokaryotes and between organelles and nuclei, it appears to be a rare phenomenon in eukaryotes (*Koonin et al., 2001; Keeling and Palmer, 2008*). It has been hypothesized that this is because eukaryotes typically have a highly segregated germ line that is protected from foreign DNA (*de Koning et al., 2001; Anderson et al., 2001; Keeling and Palmer, 2008*). Though this segregation likely inhibits HGT to some extent, it does not stop it completely, as there are several strong cases for HGT in eukaryotes (*Anderson et al., 2005; Berriman, 2005; Derelle, 2006; Hotopp et al., 2007*). Nevertheless, HGT plays only a minor role in the acquisition of novel genes in eukaryotes.

The last known mechanism of gene acquisition is *de novo* evolution, in which

random sequence changes within noncoding DNA result in the creation of a functional gene (Figure 1.1D). Though this mechanism was once believed to be either absent (Ohno, 1970) or extremely rare (Long *et al.*, 2003), genomic studies have recently unveiled several cases of *de novo* evolution (Levin *et al.*, 2006; Begun *et al.*, 2007a; Chen *et al.*, 2007; Cai *et al.*, 2008; Zhou *et al.*, 2008; Li *et al.*, 2010). Many of these gene acquisitions were mediated by repetitive elements, in one case via the extensive lineage-specific expansion of short tandem repeats (Zhou *et al.*, 2008). It is believed that such genes may first evolve the ability to be transcribed before becoming protein-coding genes (Cai *et al.*, 2008). However, despite the evolutionary importance of *de novo* acquisition of novel genes, it is still quite rare in comparison to other mechanisms.

In some cases, multiple mechanisms act together via the process of exon shuffling, in which a novel gene is constructed from the exons of two or more ancestral genes (Gilbert, 1978). An interesting case of exon shuffling is that of the *Jingwei* (*jgw*) gene in *Drosophila* (Wang *et al.*, 2000; Zhang *et al.*, 2000). In 1982, a portion of *jgw* closely resembling the alcohol dehydrogenase *Adh* gene was identified in *Drosophila yakuba* and *Drosophila teisseiri*. Comparison of this copy to the ancestral *Adh* gene revealed that the copy lacked introns and was thus created by retrotransposition of *Adh* (Jeffs and Ashburner, 1991). Further studies of *jgw* showed that, after the *Adh* retrogene was inserted, it recruited exons and introns from nearby genes, forming a chimeric gene with novel functions (Wang *et al.*, 2000; Zhang *et al.*, 2000).

This thesis focuses on gene duplication, the primary source of observed gene acquisitions. Gene duplication produces two or more identical copies of a gene, which are termed paralogs. Due to their redundancy, paralogs are thought to be under relaxed constraint, or negative selection, immediately following gene duplication (Ohno, 1970; Lynch and Force, 2000). While this gives paralogs the freedom to evolve new functions via neutral and beneficial mutations, it also leaves them vulnerable to dele-

terious mutations. Consequently, it is hypothesized that paralogs will be inactivated, or pseudogenized, within a few million generations via the accumulation of deleterious mutations (*Lynch and Force, 2000*). The prevalence of pseudogenes in many genomes, particularly in those with low genomic deletion rates, supports this idea (*Harrison and Gerstein, 2002*). However, most sequenced genomes also contain numerous functional paralogs, many of which are hundreds of millions of generations old (*Ferris and Whitt, 1979; Lundin, 1993; Sidow, 1996; Brookfield, 1997; Nadeau and Sankoff, 1997; Postlethwait et al., 1998*). Thus, it is likely that some combination of evolutionary forces acts to preserve paralogs over long evolutionary times (*Lynch and Force, 2000; Kondrashov and Kondrashov, 2006*).

A number of evolutionary forces act on paralogs following gene duplication. Mutation produces random changes among paralogs, leading to their divergence. If present, negative selection counteracts this divergence by removing deleterious mutations from the population. Positive selection has the opposite effect, facilitating the divergence of paralogs by increasing the frequency of beneficial mutations in the population. If the population size is small, there may also be random fluctuations in the frequency of mutations in the population, or genetic drift. Additionally, due to their sequence similarity, paralogs may be subject to gene conversion, or the unidirectional transfer of genetic information between similar genomic segments. This process inhibits the divergence of paralogs, irrespective of the direction of transfer. To illustrate this, let us consider the life cycle of a mutation, m , within a pair of identical paralogs, A and B (Figure 1.2). We will assume that neither positive nor negative selection is acting on these paralogs and, hence, the mutation's effect on fitness is not relevant. In this example, m is introduced at a particular locus, ℓ , in A, causing the divergence of A and B. Gene conversion between A and B at ℓ can lead to one of two outcomes. If A transfers its sequence at ℓ to B, m will be "fixed" in both copies. On the other hand, if B transfers its sequence at ℓ to A, m will be lost, and both copies will have

the ancestral sequence at ℓ . In both of these cases, the result is that A and B again have identical sequences, thus inhibiting the divergence process.

The evolutionary fate of paralogs is determined by the interaction among mutation, selection, gene conversion, and genetic drift. Though most paralogs are ultimately pseudogenized, those that remain functional do so via one of the following mechanisms (Figure 1.3): 1) acquisition of a novel function in one copy (neofunctionalization), 2) division of ancestral functions among copies (subfunctionalization), or 3) preservation of ancestral functions in all copies (conservation).

In neofunctionalization, one copy of a gene acquires a novel function(s) while the other retains the ancestral functions. Two models are commonly used to describe the mechanism of neofunctionalization: Dykhuizen-Hartl (*Dykhuizen and Hartl*, 1980; *Hartl and Dykhuizen*, 1981) and adaptation (*Hahn*, 2009). According to the Dykhuizen-Hartl model, positive selection does not play a role in the evolution of a new function. Rather, neutral mutations accumulate due to genetic drift, and the new function only becomes advantageous when the genetic background changes. In contrast, the adaptation model predicts that neofunctionalization is attained through the fixation of adaptive mutations by positive selection.

Subfunctionalization is the division of ancestral functions among duplicate genes, so that each copy performs a subset of the functions of the ancestral gene. This process is often explained by the duplication-degeneration-complementation (DDC) model (*Force et al.*, 1999; *Stoltzfus*, 1999), which was derived from the observation that many genes, particularly those involved in development, have multiple independent subfunctions (*Bender et al.*, 1983; *Slusarski et al.*, 1995). In the DDC model, deleterious mutations occur in both copies of a gene, damaging different functions of each. Thus, to preserve all functions of the ancestral gene, both copies are fixed by positive selection.

In contrast to neofunctionalization and subfunctionalization, conservation occurs

when duplicate genes are under strong evolutionary constraint, leading to the preservation of ancestral functions in both copies. This process is best explained by the dosage model (*Ohno*, 1970), in which there is a selective advantage to producing more of a particular gene, leading to the fixation of both copies by positive selection. Conservation is an interesting evolutionary phenomenon, because it is one of two paths utilized to upregulate a specific gene, with the other being to modify regulatory regions of the ancestral gene so that it is expressed at a higher level.

In this dissertation, I explore two aspects of evolution by gene duplication. First, I study the origin of functional DNA sequences by gene duplication, and then I investigate the evolution of paralogs by gene conversion. A common theme of this dissertation is the direct phylogenetic approach taken, which allows me to ascertain when and how specific evolutionary events occurred. In Chapters II and III, I use this approach to identify recent evolutionary gains and losses of functional DNA segments and investigate the molecular mechanisms and selective forces responsible for these events. Then, in Chapter V, I utilize the same approach again to analyze insertions and deletions (indels) produced by gene conversion between ancient pairs of paralogs, which are located using a method described in Chapter IV. In this last chapter, I also explore the effect of gene conversion on the evolutionary fate of paralogs.

In Chapter II, I investigate the origin of nested genes, which are protein-coding genes located in the introns of other protein-coding “host” genes (Figure 1.4). I utilize phylogenetic relationships within vertebrate, *Drosophila*, and *Caenorhabditis* genomes to study three aspects of nested gene acquisition. First, I examine their evolutionary dynamics, or gains and losses during the course of evolution. Specifically, I am interested in the relative rates of gains and losses and how this affects the organizational complexity of animal genomes. Next, I study the formation of nested gene structures by determining which sequences were acquired in evolution and then attempting to trace these sequences back to ancestral sequences. Last, I infer which evolutionary

forces are responsible for the long-term maintenance of nested gene structures by comparing the sequence and functional data of genes in nested structures to those of un-nested genes.

In Chapter III, I study the origin of long transcripts, or “clusters”, containing Piwi-interacting RNAs (piRNAs), which are small RNAs that are expressed primarily in germline cells in a wide range of plants and animals. piRNAs bind to highly conserved P-element induced wimpy testis (Piwi) proteins, and the resulting complex is involved in silencing transposable elements (*Aravin et al.*, 2007). For this analysis, I utilize clusters annotated in rat and mouse genomes. To determine when clusters arose, I compare the cluster-containing regions from a particular rodent to orthologous regions, or those derived from the same ancestral sequence, in the other rodent and human genomes. I next attempt to locate the ancestral sequences from which clusters arose to identify the molecular mechanisms of cluster acquisition. Using these results, I hypothesize about the evolutionary forces leading to the emergence and maintenance of piRNA clusters in mammalian genomes.

In Chapter IV, I describe Bridges, which is a method for identifying similar genomic segments within and between genomes. Though there are other algorithms that are commonly used for this purpose, the advantage of Bridges is that it contains 20 parameters that enable the user to tailor a search to a particular goal. In Chapter V, I use Bridges to locate unique pairs of paralogs in *Drosophila* and primate genomes. Then, I investigate insertions and deletions (indels) produced by gene conversion between ancient pairs of paralogs. Though many studies have examined nucleotide substitutions produced by gene conversion, some of which discovered AT→GC biases (*Marias*, 2003; *Mancera et al.*, 2008; *Liu and Li*, 2008; *Berglund et al.*, 2009), little is known about length difference mutations in gene conversion. Thus, I first ascertain insertions and deletions to determine whether a similar bias exists for length difference mutations. Next, I calculate the rate of gene conversion relative to that

of ordinary mutation. Finally, combining the relative indel frequencies and rate, I explore the effect of gene conversion on the evolutionary fates of paralogs.

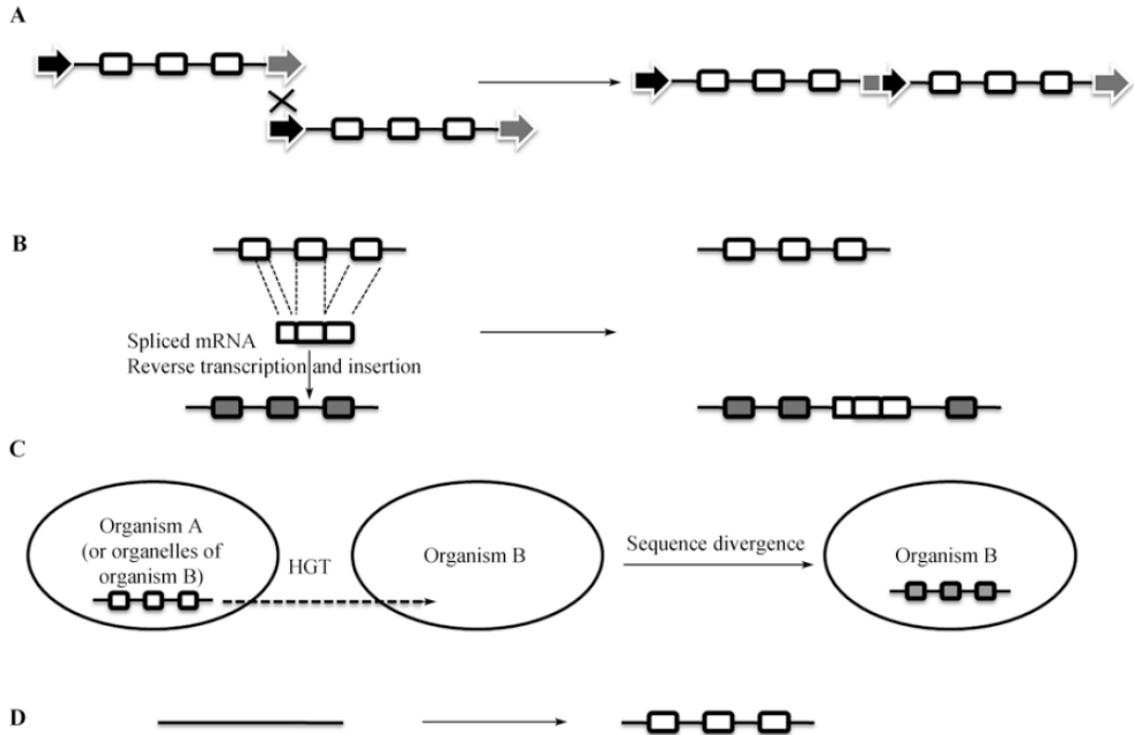


Figure 1.1: Mechanisms of novel gene acquisition. Boxes represent exons, while lines represent noncoding regions. A) Ectopic recombination. Recombination occurs between repetitive elements (black and gray arrows), resulting in tandem copies of a gene. B) Retrotransposition. A spliced mRNA is reverse-transcribed and inserted into another genomic location, producing a new gene copy lacking introns. C) Horizontal gene transfer. A gene is directly transferred from organism A to organism B. D) *De novo* origination. A new gene evolves from small-scale evolution of a noncoding sequence. Figure adapted from (Zhou and Wang, 2008).

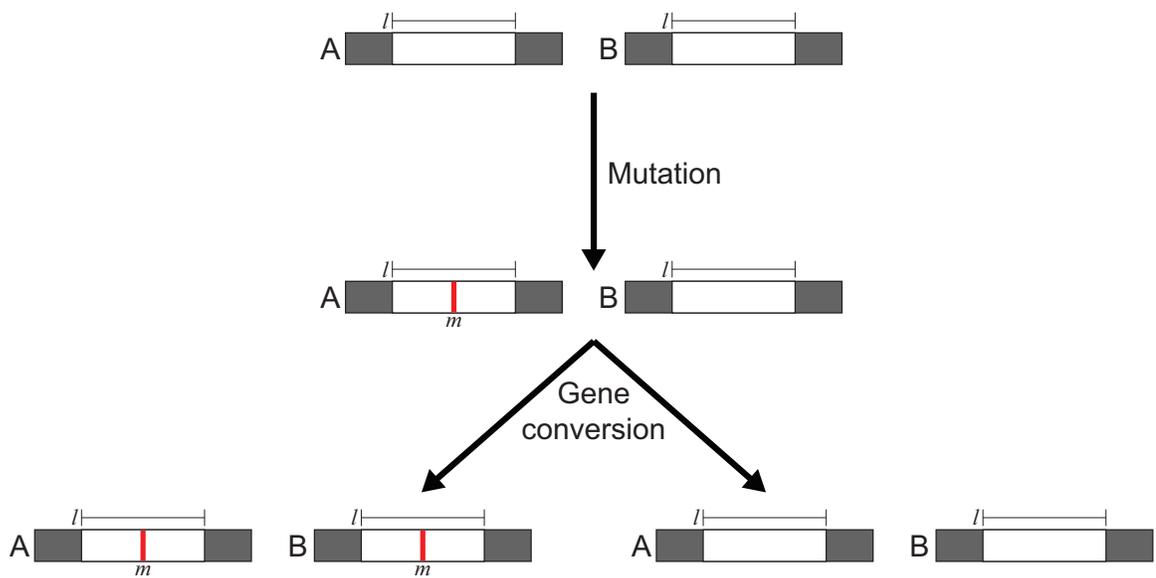


Figure 1.2: Life cycle of a mutation within a pair of paralogs that undergo gene conversion. A pair of paralogs, A and B, are identical immediately following gene duplication. Then a mutation, m (red), occurs at locus l (white), in A. Gene conversion between A and B results in one of two outcomes. In the first, depicted on the left, A transfers its sequence at l to B, leading to the fixation of m in both paralogs. In the second, depicted on the right, B transfers its sequence at l to A, leading to the loss of m .

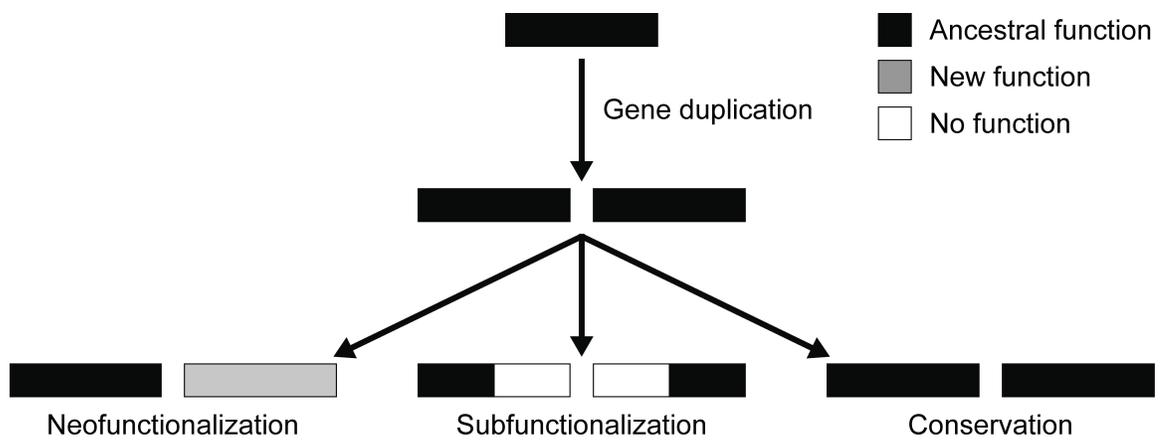


Figure 1.3: Phenotypic outcomes of duplicate genes. Gene duplication produces two identical copies of a gene, which can both be functionally maintained in the genome via the evolution of a new function in one copy (neofunctionalization), division of ancestral functions between copies (subfunctionalization), or preservation of ancestral functions in both copies (conservation).

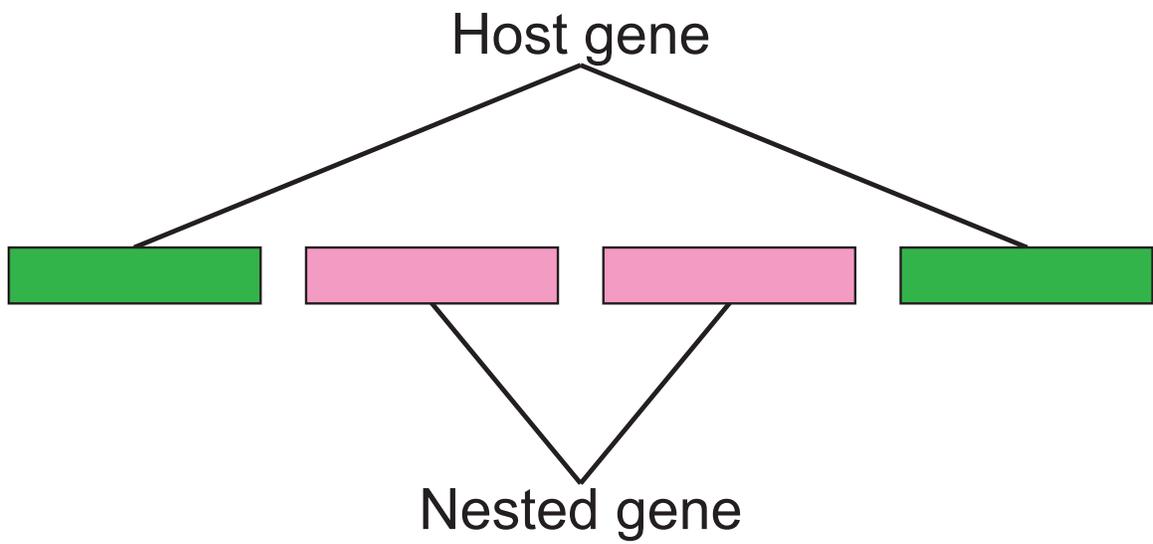


Figure 1.4: Nested gene structure.

CHAPTER II

Nested genes and increasing organizational complexity of metazoan genomes

2.1 Introduction

Eukaryotes are typically more complex than prokaryotes on the molecular, systems, and phenotypic scales of biological organization. In particular, genomes of multicellular eukaryotes possess a complex architecture that involves substantial overlapping of their transcribed regions (*Mironov et al.*, 1999; *Makalowska et al.*, 2005; *Willingham et al.*, 2006; *Kapranov et al.*, 2007) and protein-coding genes (*Misra et al.*, 2002; *Veeramachaneni et al.*, 2004; *Yu et al.*, 2005), forming an interleaving mosaic of exon and intron sequences. Although it is clear that such complex genome organization is made possible by the presence of introns, the rates and mechanisms of evolutionary events leading to gains and losses of overlapping gene arrangements have not been studied previously.

Previous studies of the evolution of genome complexity have primarily relied on correlations between the abundances of various genomic elements (introns, transposons, gene size, etc.) and the product of the effective population size and mutation rate (*Lynch*, 2002a; *Lynch and Conery*, 2003; *Lynch*, 2006; *Yi*, 2006; *Lynch*, 2007a,b). However, claims of causality based on such correlative analyses are always

inconclusive, because other potentially important factors can never be excluded. In an attempt to circumvent some of the limitations of the correlative approach, we explored the evolution of genomic complexity in a more direct manner, by tracing the evolutionary dynamics of nested pairs of protein-coding genes in animals. This study covers only one, perhaps not even the most common, class of interleaved gene arrangements because we left out the numerous intron-contained small RNA genes (*Mattick and Makunin, 2005*). Nevertheless, even this limited analysis clearly reveals the ongoing increase of the organizational complexity of animal genomes and suggests that this process occurs via a nonselective route.

2.2 Evolutionary dynamics of nested gene structures

The most common form of overlap between protein-coding genes in eukaryotes is a nested gene structure, and in a majority of such structures, the internal gene lies entirely within one intron of the external gene (*Misra et al., 2002; Yu et al., 2005*). Thus, we investigated the evolution of this class of nested gene structures in vertebrates, *Drosophila* and *Caenorhabditis*. A search of NCBI annotation records yielded 428, 815, 440 and 608 nested gene pairs in *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. briggsae* genomes, respectively. After eliminating gene pairs that might have been misannotated (see Methods), we arrived at sets of 128, 792, 429 and 233 nested gene pairs, respectively. Only a small minority of the protein sequences encoded by internal genes from each of these three major taxa show significant sequence similarity to internal genes products in the other two taxa (data not shown), suggesting that either these structures emerged independently and relatively late during evolution or that they were extensively and repeatedly lost.

By examining gene annotations and constructing sequence alignments, we identified the closest species with a completely sequenced genome in which each nested gene structure was absent. Absence of the nested structure in an appropriate outgroup

species indicates its emergence (gain) in the respective lineage, whereas presence of the nested structure in the outgroup indicates its loss (Figure 2.1). Gains were found in all three taxa, with the emergence of 55 internal genes in at least 40 independent events in vertebrates, 52 internal genes in at least 48 events in *Drosophila* and 22 internal genes in as many events in *Caenorhabditis*. The rate of these acquisitions was approximately uniform throughout the course of evolution (Figure 2.2). By contrast, losses of nested gene structures were much rarer, with none detected in vertebrates, 17 in *Drosophila* and 2 in *Caenorhabditis*.

2.3 Acquisition of nested gene structures

At least four scenarios are plausible for the formation of a nested gene structure: (i) an internal gene can evolve by insertion of a DNA sequence into an intron of a pre-existing gene, (ii) an internal gene can evolve *de novo* from an intronic sequence of a pre-existing gene, (iii) a gene can become internal after an adjacent gene acquires an additional exon(s) or (iv) a gene can become internal after fusion of two genes that flank it from the opposite sides (Figure 2.3).

By comparing the gene structures and encoded protein sequences of internal and external genes to complete gene sets from the respective species, we deduced the mechanisms of formation of nested gene structures in vertebrates (Table 2.1). Nearly all nested gene structures seem to have emerged by insertion of a DNA sequence, which arose by gene duplication or retrotransposition, into an intron of a pre-existing gene. The origin of an internal gene was classified as a retrotransposition when it was intronless in a given species, whereas its non-nested ortholog in a sister species contained introns. A duplication at the DNA level was inferred when both the internal gene and a non-nested ortholog in a sister species had introns. In cases where the internal gene and a non-nested ortholog were both intronless, retrotransposition and duplication at the DNA level could not be discriminated. Five internal genes in hu-

mans are candidates for *de novo* origin from intron sequences (see Methods), including one gene with no sequence similarity beyond apes [*placenta-specific 4 (PLAC4)*] and another with no similarity beyond old world monkeys [*saitohin (STH)*] (Table 2.2). Analysis of the 12 recently sequenced *Drosophila* genomes showed that the majority of *de novo* genes originate in introns (*Drosophila 12 Genomes Consortium*, 2007). Consistent with this observation, we found 11 internal genes in *D. melanogaster* with no sequence similarity to any genes in the genome of the closely related *D. yakuba*. We did not identify any nested gene structures that evolved via the remaining two scenarios.

2.4 No functional significance of nested gene structures

At least three hypotheses could explain the parallel accumulation of nested gene structures in different taxa. First, a nested structure might confer a selective advantage because of a functional or co-regulatory relationship between its members (*Henikoff and Eghtedarzadeh*, 1987; *Habib et al.*, 1998; *Jaworski et al.*, 2007; *Furia et al.*, 1993; *Davies et al.*, 2004). Second, according to the transcriptional collision model, members of a nested gene structure could interfere with each others transcription (*Da Lage et al.*, 2003; *Crampton et al.*, 2006; *Osato et al.*, 2007), resulting in alternative expression of these genes in different tissues or during different times in development. Finally, acquisition of a nested gene structure could be a neutral process (*Lynch*, 2002a; *Lynch and Conery*, 2003; *Lynch*, 2006; *Yi*, 2006; *Lynch*, 2007a,b; *Habib et al.*, 1998), driven by the presence of numerous long introns that provide niches for insertion of genes. Each of these hypotheses leads to a distinct prediction about the relationship between the expression of internal and external genes in a nested pair. The functional co-regulation hypothesis predicts a positive correlation between levels of their expression in similar tissues, the transcriptional collision hypothesis predicts a negative correlation and the neutral hypothesis predicts no correlation.

To discriminate between these three hypotheses, we analyzed gene expression data from human and *D. melanogaster* genomes (see Methods). We compared correlations of gene expression in 109 and 752 nested gene pairs in humans and *D. melanogaster*, respectively, to 1000 random sets of 109 and 752 adjacent gene pairs from corresponding genomes. There was no significant difference in mean correlation coefficients of gene expression levels between nested and adjacent genes in either human (0.33 ± 0.03 for nested and 0.33 ± 0.0008 for adjacent pairs) or *D. melanogaster* (0.041 ± 0.014 for nested and 0.030 ± 0.00046 for adjacent gene pairs), which is consistent with the neutral hypothesis. The observation that external genes have substantially more and longer introns than average in the respective species (*Yu et al.* (2005) and Figures 2.4-2.6) is also compatible with the neutral hypothesis. Furthermore, examination of the available functional information for nested gene pairs (Table 2.2) did not reveal any obvious connections (*Yu et al.*, 2005). Fixation of originally neutral or even slightly deleterious sequence segments, such as introns and transposable elements, through genetic drift acting in relatively small populations is a common phenomenon in eukaryotic evolution that might be partially responsible for the evolution of complex phenotypes (*Lynch*, 2002a; *Lynch and Conery*, 2003; *Lynch*, 2006; *Yi*, 2006; *Lynch*, 2007a,b). The increase in organizational complexity of intron-rich genomes via emergence of nested gene structures seems to be another facet of this process.

2.5 Predicting the course of genome structure evolution

The neutral hypothesis implies that the preferential evolutionary gain of nested gene structures is caused by metazoan genomes being far from neutral equilibrium with respect to birth and death of intron-contained genes (*Lynch and Conery*, 2003). We estimated the rate of acquisition of nested gene structures as ~ 0.4 , 0.9 and 0.2 events per million years in the *H. sapiens*, *D. melanogaster*, and *C. elegans* lineages,

respectively (see Methods). Because animal genomes currently contain ~ 500800 nested gene pairs, these rates indicate that nested gene structures began to emerge ~ 1 billion years ago, perhaps concurrent with the substantial intron gain that apparently occurred at the onset of metazoan evolution (*Carmel et al.*, 2007). These results suggest that metazoan introns are still far from saturation by internal genes and that the organizational complexity of metazoan genomes will continue to increase for many millions of years via the emergence of new nested gene structures. By the time metazoan genomes reach organizational complexity equilibrium, the overlap of functional elements is expected to be much greater than what we observe in extant taxa and will probably include numerous Russian doll-like nested structures. This process has already begun in fruit flies, with the *D. melanogaster* genome containing six cases where a nested gene structure is nested in another gene.

2.6 Conclusions and perspective

We have shown that the evolution of metazoan genomes is accompanied by a steady rise in the prevalence of nested arrangements of protein-coding genes, leading to increasingly complex genome architectures. In addition to nested protein-coding genes, animal genomes contain numerous complex arrangements, including incomplete overlaps of protein coding regions and their untranslated regions and various RNA genes (*Misra et al.*, 2002; *Veeramachaneni et al.*, 2004; *Makalowska et al.*, 2005; *Willingham et al.*, 2006; *ENCODE Project Consortium*, 2007; *Kapranov et al.*, 2007). In particular, a substantial fraction of microRNA and small nucleolar RNA genes are either fully contained within introns of protein-coding genes or overlap with protein-coding exons (*Mattick and Makunin*, 2005). It will be of major interest to determine whether the trend of increasingly complex genome organization reported here applies to RNA genes or incompletely overlapping gene structures.

2.7 Methods

2.7.1 Identification and quality control of nested gene pairs

Sequences and annotations for *H. sapiens*, *D. melanogaster*, *C. elegans*, and *C. briggsae* genomes were downloaded from the NCBI GenBank (*Benson et al.*, 2009) database at <ftp://ftp.ncbi.nih.gov/genomes/>. After selecting the longest isoform of each gene, we identified 428, 815, 440, and 608 nested gene pairs in each genome, respectively. Several measures were taken to exclude erroneously annotated nested genes. For the *H. sapiens*, *D. melanogaster*, and *C. elegans* genomes, we retained only RefSeq genes (*Pruitt et al.*, 2007). We also excluded all human genes with the labels “hypothetical” or “predicted” in the define of the GenBank-derived fasta file. For the *D. melanogaster* and *C. elegans* genomes, we kept only those genes that showed >95% sequence identity over >90% of the length of the best nucleotide BLAST (*Altschul et al.*, 1997) hit with complete mRNAs sequenced from the same species. We were more stringent with the quality control of human gene annotations because of the substantially longer intron and spacer sequences found in the human genome compared to *D. melanogaster* and *C. elegans*. The mRNAs were obtained from GenBank with the Entrez retrieval system (*Benson et al.*, 2009), using the species names and “complete” as key words and setting the limits option to mRNA molecules. Because annotation of the *C. briggsae* genome was the least reliable, we required that all *C. briggsae* genes have significant BLAST hits to protein sequences from the final set of *C. elegans* genes. In addition, all cases of nested gene evolution involving *C. briggsae* gene annotations were checked manually against *C. elegans* annotations using the BLAT program (*Kent*, 2002) on the UCSC genome browser (*Karolchik et al.*, 2008).

2.7.2 Comparative genomic analysis of nested gene structures

Genes that passed the above inclusion criteria were compared to the genomes of sister species and outgroups. We used the protein BLAT alignment tool on the UCSC genome browser, as well as the TBLASTN program (*Altschul et al.*, 1997), to compare protein sequences of internal and external genes to complete genomes. If an ortholog for an internal gene was not identified using either of these two methods, a TBLASN search was performed against the orthologous intron from the external gene. Thus, in order to classify a nested gene structure as having been gained or lost in evolution, we required that both the internal and external genes be found in the sister species and an outgroup. It is easier to find an internal gene within the orthologous intron of an external gene in an outgroup, which was our expectation for an evolutionary loss, than it is to find it in the entire genome of the outgroup, which was the requirement for an evolutionary gain (Figure 2.1). Thus, our approach was conservative and could have slightly biased the results in favor of discovery of evolutionary losses. Also, the requirement of finding both genes in both genomes prevented us from misidentifying as evolutionary events genes that are absent due to incomplete genome sequences. For vertebrates, an additional method was employed to analyze the evolution of nested gene structures. Alignments of regions in the sister and outgroup species orthologous to the nested gene pair were constructed using OWEN (*Ogurtsov et al.*, 2002). We began all alignments with a strict requirement of 16 successive matches and $p < 10^{-8}$ and progressively relaxed these parameters to 8 successive matches and $p < 0.01$, using the greedy algorithm to resolve any conflicts. Presence or absence of an internal gene in the orthologous external gene was judged based on the quality of the alignment. A gap in the alignment opposite the entire span of an internal gene in human indicated the absence of the internal gene in that genome. Both methods yielded the same results, with the exception of 5 cases, which are candidates for *de novo* gene creation. Candidate *de novo* genes were identified

when both TBLASTN and BLAT revealed no sequence similarity of an internal gene in a sister species. We did not apply the latter method to invertebrate genomes due to the higher degree of their divergence, which also prevented us from performing a systematic analysis of the modes of internal gene evolution in invertebrates.

2.7.3 Analysis of gene expression

Human gene expression data were obtained from (*Su et al.*, 2004), which included 73 healthy human tissues measured on the HG-U133A Affymetrix array. We computed the correlation of mean levels of expression of internal and external genes for 109 nested genes in humans. We next identified all adjacent pairs of RefSeq annotated genes in the human genome and randomly selected 109 such pairs 1000 times. We then compared the correlation coefficient of the 109 nested genes to the average correlation coefficient of the 1000 trials of 109 adjacent pairs. We employed the same statistical approach for *D. melanogaster* gene expression analysis. Gene expression data was obtained from (*Chintapalli et al.*, 2007), which included 11 different tissues measured on the GeneChip *Drosophila* Genome 2.0 Affymetrix array. We then compared the correlation of mean levels of gene expression of 668 *D. melanogaster* nested gene pairs and 1000 random samples of 752 adjacent gene pairs.

2.7.4 Estimating the rate of nested gene evolution

Of the 128 definite nested gene structures in the human lineage, we identified 55 that emerged after the divergence of human and zebrafish lineages ~ 450 million years ago (*Kumar and Hedges*, 1998). Assuming that these 128 nested gene structures are representative of the overall 428 annotations in the human genome, the observed number of internal gene gains give an estimate of ~ 0.4 gains per million years for all nested genes in the human genome ($55/128 \times 428/450$). In the *D. melanogaster* lineage, 48 internal genes were gained since the divergence of *D. melanogaster* and *D.*

pseudoobscura ~55 million years ago (Tamura et al., 2003), indicating a rate of ~0.9 gains per million years. Our analysis of the *C. elegans* genome was more restricted due to large distances between the *C. elegans*, *C. briggsae*, and *Pristionchus pacificus* genomes. Because we never considered cases where sequence similarity was not high enough to determine orthology, we described only a handful of cases of nested gene evolution. Nevertheless, an approximation was still possible due to the total number of nested genes showing a high enough sequence similarity between *C. elegans*, *C. briggsae*, and *P. pacificus* genomes. Of the 440 total *C. elegans* internal genes, exactly one half (220) were found in *C. briggsae* and *P. pacificus*, 11 of which were gains. Thus, the overall rate of nested gene gain was 22 per ~100 million years of evolution separating *C. elegans* and *C. briggsae* (Stein et al., 2003), or ~0.2 per million years.

2.7.5 Genomes used in this study

- Vertebrates: *Homo sapiens*, *Mus musculus*, *Monodelphis domestica*, *Gallus gallus*, *Danio rerio*
- *Drosophila*: *Drosophila melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis*
- Nematodes: *Caenorhabditis elegans*, *C. remanei*, *C. brenneri*, *Pristionchus pacificus*

Table 2.1: Mechanisms of origin of human internal genes

	After human-mouse	After human-opossum	After human-chicken	Total
Duplication	2	2	4	8
Retrotransposition	3	11	6 (5)	21 (20) ^a
Duplication or retrotransposition	2	3 (2)	16 (2)	21 (6)
<i>De novo</i> candidates	2	1	2	5

Independent events are shown in parentheses.

^aOne retrotransposition event was not dated to the degree of accuracy as other cases.

Table 2.2: Human recently evolved nested gene pairs

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
62122867	chr415 synaptotagmin-derived protein (Sytdep)	46409446	transmembrane protease, serine 11F	Gained after human-mouse divergence	Retrotransposition
50400081	Small ubiquitin-like protein (SUMO) 4	14149669	mitogen-activated protein kinase kinase 7 interacting protein 2	Gained after human-mouse divergence	Retrotransposition
10835093	pituitary transforming protein 2 (PTTG2); sister-chromatid separation inhibitor (securin)	50658061	TBC1 (tre-2/USP6, BUB2, cdc16) domain family, member 1	Gained after human-mouse divergence	Retrotransposition
28274703	Synaptonemal complex protein 3 (Family with sequence similarity 9, member C)	7382482	Rho GTPase activating protein 6 isoform 1	Gained after human-mouse divergence	DNA duplication
23199998	tRNA methylase, NOL1/NOP2/Sun domain family, member 5 isoform 1	17149849	FK506-binding protein 6	Gained after human-mouse divergence	DNA duplication
13929212	olfactory receptor, family 2, subfamily A, member 4	4826896	ectonucleotide pyrophosphatase/phosphodiesterase 3	Gained after human-mouse divergence	DNA duplication or retrotransposition
51510895	TSPY-like 6	20149322	muscle-type acylphosphatase 2	Gained after human-mouse divergence	DNA duplication or retrotransposition
56090269	saitohin (STH)	82534351	microtubule-associated protein tau isoform 1	Gained after human-mouse divergence	<i>De novo</i> candidate

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
33457306	placenta-specific (PLAC4)	4	19923395	beta-site APP-cleaving enzyme 2 isoform A preproprotein	Gained after human-mouse divergence <i>De novo</i> candidate
15082234	lactate dehydrogenase (LDH) A-like 6B		55956916	myosin IE	Gained after human-opossum divergence Retrotransposition
14149675	cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kDa, tau variant		10835242	protein kinase, cGMP-dependent, type I isoform 2	Gained after human-opossum divergence Retrotransposition
31542306	chromatin modifying protein 1B		33695153	guanine nucleotide binding protein (G protein), alpha activating activity polypeptide, olfactory type isoform 1	Gained after human-opossum divergence Retrotransposition
53832014	choroideremia-like escort protein 2		71999131	Rab opsin 3	Gained after human-opossum divergence Retrotransposition
56790947	zinc finger, BED domain containing 2		38683840	CD96 antigen isoform 1 precursor	Gained after human-opossum divergence Retrotransposition
48717485	DDI1, DNA-damage inducible 1, homolog 1		15451921	platelet derived growth factor D isoform 2 precursor	Gained after human-opossum divergence Retrotransposition
24371268	nucleosome assembly protein 1-like 5		7657152	hect domain and RLD 3	Gained after human-opossum divergence Retrotransposition
11545841	3-oxoacid CoA transferase 2		29571106	bone morphogenetic protein 8B preproprotein	Gained after human-opossum divergence Retrotransposition
9558731	replication protein 34kDa		5803003	diaphanous 2 isoform 156	Gained after human-opossum divergence Retrotransposition

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
22212896	protein phosphatase 3 regulatory subunit B, beta isoform	20143964	glutamate receptor, ionotropic, N-methyl-D-aspartate 3A	Gained after human- opossum divergence	Retrotransposition
21553335	DnaJ (Hsp40) homolog, subfamily B, member 7	11559925	X-prolyl aminopeptidase (aminopeptidase P) 3, putative	Gained after human- opossum divergence	Retrotransposition
13775180	testis expressed sequence 13A	11225607	interleukin 1 receptor accessory protein-like 2	Gained after human- opossum divergence	DNA duplication
23097246	pyrin domain containing 1	56605981	tripartite motifcontaining 72	Gained after human- opossum divergence	DNA duplication
7656938	CREBBP/EP300 inhibitor 1	44680159	rai-like protein Gained after human- opossum divergence	DNA duplication or retrotransposition	
12056463	coagulation factor VIII-associated protein	4503647	coagulation factor VIII isoform a precursor	Gained after human- opossum divergence	DNA duplication or retrotransposition
63029941	H2A histone family, member B1	4503647	coagulation factor VIII isoform a precursor	Gained after human- opossum divergence	DNA duplication or retrotransposition
38455398	NK inhibitory receptor precursor	28376635	RAB37, member RAS oncogene family isoform 3	Gained after human- opossum divergence	<i>De novo</i> candidate
46048458	YY2 transcription factor	7706693	membrane-bound transcription factor peptidase, site 2	Gained after human- chicken divergence	Retrotransposition
21040269	ring finger protein 133	57546902	Ca ²⁺ -dependent activator protein for secretion 2 isoform b	Gained after human- chicken divergence	Retrotransposition

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
37675277	ring finger protein 148	57546902	Ca ²⁺ -dependent activator protein for secretion 2 isoform b	Gained after human-chicken divergence	Retrotransposition
57232750	HMG2 like isoform 1	92110053	CUB and Sushi multiple domains 2	Gained after human-chicken divergence	Retrotransposition
14149651	nasopharyngeal carcinoma associated gene	77404397	staphylococcal nuclease domain containing 1	Gained after human-chicken divergence	Unclear
56790919	protein-8 haspin	6007851	integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)	Gained after human-chicken divergence	Retrotransposition
46275822	solute carrier family 5 (inositol transporters), member 3	16554616	mitochondrial ribosomal protein S6	Gained after human-chicken divergence	Retrotransposition
33356556	amelogenin (X chromosome) isoform 3 precursor	7382482	Rho GTPase activating protein 6 isoform 1	Gained after human-chicken divergence	DNA duplication
15147234	CaM-KII inhibitory protein	41281433	endothelin converting enzyme 2 isoform A	Gained after human-chicken divergence	DNA duplication
5730106	G protein-coupled receptor TYMSTR	13470092	FYVE and coiled-coil domain containing 1	Gained after human-chicken divergence	DNA duplication
4757898	calbindin 3	28559083	cytidine triphosphate synthase II	Gained after human-chicken divergence	DNA duplication
63053512	keratin associated protein 10-1	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
46250735	keratin associated protein 10-2	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
63053515	keratin associated protein 10-3	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38490568	keratin associated protein 10-4	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
63053517	keratin associated protein 10-5	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38490575	keratin associated protein 10-6	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38490571	keratin associated protein 10-7	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38490583	keratin associated protein 10-8	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38490577	keratin associated protein 10-9	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
32140182	keratin associated protein 10-10	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38490579	keratin associated protein 10-11	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
38490587	keratin associated protein 10-12	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
32140180	keratin associated protein 12-1	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
32140184	keratin associated protein 12-2	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38371737	keratin associated protein 12-3	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
38371744	keratin associated protein 12-4	22001420	Epilepsy associated, EPTP domain-containing protein	Gained after human-chicken divergence	DNA duplication or retrotransposition
40255098	mucin 15	13899227	transmembrane protein 16C	Gained after human-chicken divergence	<i>De novo</i> candidate
5901938	fasting-induced protein	23510360	Ras association domain family 4	Gained after human-chicken divergence	<i>De novo</i> candidate
13375885	chromosome 14 open reading frame 169	42558270	HEAT repeat containing 4	After human-zebrafish and before human-opossum divergence	Retrotransposition
14161692	calpain small subunit 2	47086907	lysophosphatidylcholin acyltransferase 2	After human-zebrafish divergence	-
47458036	hereditary sensory neuropathy, type II	12711660	WNK lysine deficient protein kinase 1	After human-zebrafish divergence	-

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
19923939	KTI12 homolog, chromatin associated	7705696	endoplasmic reticulum thioredoxin superfamily member, 18 kDa	After zebrafish divergence	human- zebrafish divergence
51511749	ecotropic viral integration site 2A isoform 1	4557793	neurofibromin isoform 2	After zebrafish divergence	human- zebrafish divergence
18201870	G protein-coupled receptor 82	4502567	calcium/calmodulin-dependent serine protein kinase (MAGUK family)	After zebrafish divergence	human- zebrafish divergence
4885301	G protein-coupled receptor 17	40804748	LIM and senescent cell antigen-like domains 2	After zebrafish divergence	human- zebrafish divergence
15029528	G protein-coupled receptor 18	29366838	UBA domain containing 2	After zebrafish divergence	human- zebrafish divergence
41281535	purinergic receptor P2Y, G-protein coupled, 14	93277088	mediator of RNA polymerase II transcription, subunit 12 homolog (<i>S. cerevisiae</i>)-like	After zebrafish divergence	human- zebrafish divergence
12232483	purinergic receptor P2Y12	93277088	mediator of RNA polymerase II transcription, subunit 12 homolog (<i>S. cerevisiae</i>)-like	After zebrafish divergence	human- zebrafish divergence
31377772	G protein-coupled receptor 171	93277088	mediator of RNA polymerase II transcription, subunit 12 homolog (<i>S. cerevisiae</i>)-like	After zebrafish divergence	human- zebrafish divergence
29171721	purinergic receptor P2Y, G-protein coupled, 13	93277088	mediator of RNA polymerase II transcription, subunit 12 homolog (<i>S. cerevisiae</i>)-like	After zebrafish divergence	human- zebrafish divergence

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
31542858	G protein-coupled receptor 87	93277088	mediator of RNA polymerase II transcription, subunit 12 homolog (<i>S. cerevisiae</i>)-like	After zebrafish divergence	human- zebrafish divergence
49258196	syntaxin 19	33598956	ADP-ribosylation factor-like 2-like 1 isoform 1	After zebrafish divergence	human- zebrafish divergence
21359933	latexin	18390331	G elongation factor, mitochondrial 1	After zebrafish divergence	human- zebrafish divergence
27734715	potassium inwardly-rectifying channel J13	42476299	trinucleotide repeat containing 15 isoform b	After zebrafish divergence	human- zebrafish divergence
12597641	leucine rich repeat containing 19	13376669	coiled-coil domain containing 2	After zebrafish divergence	human- zebrafish divergence
34452725	polypeptide N-acetylgalactosaminyltransferase 4	26665869	WD repeat domain 51B	After zebrafish divergence	human- zebrafish divergence
38257144	potassium channel tetramerisation domain containing 4	4758488	general transcription factor IIF, polypeptide 2, 30kDa	After zebrafish divergence	human- zebrafish divergence
93004094	immunoglobulin superfamily, member 6	19923449	methyltransferase like 9 After human-zebrafish divergence	-	-
4504051	guanine nucleotide binding protein, alpha z polypeptide	7657530	rhabdoid tumor deletion region protein 1	After zebrafish divergence	human- zebrafish divergence
72534676	leucine rich repeat containing 17 isoform 1	24432072	F-box and leucine-rich repeat protein 13	After zebrafish divergence	human- zebrafish divergence
88702793	slit-like 2	88702975	coronin 7 After human-zebrafish divergence	-	-

Internal gi	Internal gene function	External gi	External gene function	Time of origin	Mechanism of origin
4506333	pentraxin-related gene, rapidly induced by IL-1 beta	22003858	ventricular zone domain PH domain	After zebrafish divergence	human-
			pressed homolog 1		divergence

Table 2.3: *Drosophila melanogaster* recently evolved nested gene pairs

Internal gi	External gi	Time of origin
24641019	24641017	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
45552435	45551006	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
62862012	62862008	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
62862354	62862344	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
62862402	62862398	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24641150	116007148	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24641152	116007148	Gained after <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24584521	17647189	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24648314	24648308	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24649164	24649162	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
22024114	24652747	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24666061	24666053	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24646310	45550746	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24654876	85725044	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24583886	116007320	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
17737845	116007976	Gained before <i>D. melanogaster</i> - <i>D. ananassae</i> divergence
24584831	24584828	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24642002	17530937	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24665587	17136434	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
17864626	17137184	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
85726402	17137184	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24659396	17736971	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
28571228	18859667	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
78706724	21355909	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24664494	21357533	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
21357663	21357661	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
28574168	24584296	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24584380	24584378	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
17137204	24584665	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
17137228	24584665	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24584671	24584665	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
19921460	24584828	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24642571	24642569	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
18860537	24643238	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
21358627	24646843	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24646914	24646908	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
21356137	24648503	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
78706662	24666053	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
18859643	28571135	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24663496	45551553	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
45551029	45552495	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence

Internal gi	External gi	Time of origin
24660410	45552989	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24586638	62471689	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
28574577	62484462	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
62862256	62862246	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24762390	78707567	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
45550944	116007292	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
21356559	116008042	Gained before <i>D. melanogaster</i> - <i>D. yakuba</i> divergence
24650750	21357731	Lost in <i>D. ananassae</i>
18859959	24642282	Lost in <i>D. ananassae</i>
24650883	28572031	Lost in <i>D. ananassae</i>
24582046	45550133	Lost in <i>D. ananassae</i>
24641008	18857949	Lost in <i>D. pseudoobscura</i>
24668676	21356469	Lost in <i>D. pseudoobscura</i>
19921722	24586174	Lost in <i>D. pseudoobscura</i>
24586184	24586174	Lost in <i>D. pseudoobscura</i>
24586186	24586174	Lost in <i>D. pseudoobscura</i>
28573282	24586174	Lost in <i>D. pseudoobscura</i>
24656107	24656102	Lost in <i>D. pseudoobscura</i>
20130261	24658511	Lost in <i>D. pseudoobscura</i>
24762757	24762755	Lost in <i>D. pseudoobscura</i>
45551164	24762755	Lost in <i>D. pseudoobscura</i>
45550836	85725262	Lost in <i>D. pseudoobscura</i>

Table 2.4: *Caenorhabditis elegans* recently evolved nested gene pairs

Internal gi	External gi	Time of origin
17543776	115533052	Gained in <i>C. elegans</i>
25141286	25141274	Gained in <i>C. elegans</i>
71994600	17510041	Gained in <i>C. elegans</i>
71986750	32565120	Gained in <i>C. elegans</i>
17535289	32563849	Gained in <i>C. elegans</i>
71985903	133930931	Gained in <i>C. elegans</i>
86575243	71999521	Gained in <i>C. elegans</i>
17554756	71988666	Gained in <i>C. elegans</i>
17538450	71982042	Gained in <i>C. elegans</i>
17551914	71996779	Gained in <i>C. elegans</i>
17535035	71992322	Gained in <i>C. elegans</i>
157773237	157773219	Lost in <i>C. brenneri</i>
157747773	157747769	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157748993	157748983	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157759447	157759445	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157752690	157752686	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157752688	157752686	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157755321	157755317	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157755319	157755317	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157755323	157755317	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157763540	157763538	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157766214	157766208	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157765346	157765344	Gained after <i>C. briggsae</i> - <i>C. remanei</i> divergence
157748649	157748647	Gained before <i>C. briggsae</i> - <i>C. remanei</i> divergence
157764794	157764792	Gained after <i>C. briggsae</i> - <i>C. elegans</i> divergence
157773237	157773219	Lost in <i>C. remanei</i>
157773239	157773219	Lost in <i>C. remanei</i>
157773233	157773219	Lost in <i>C. remanei</i>
157773241	157773219	Lost in <i>C. remanei</i>

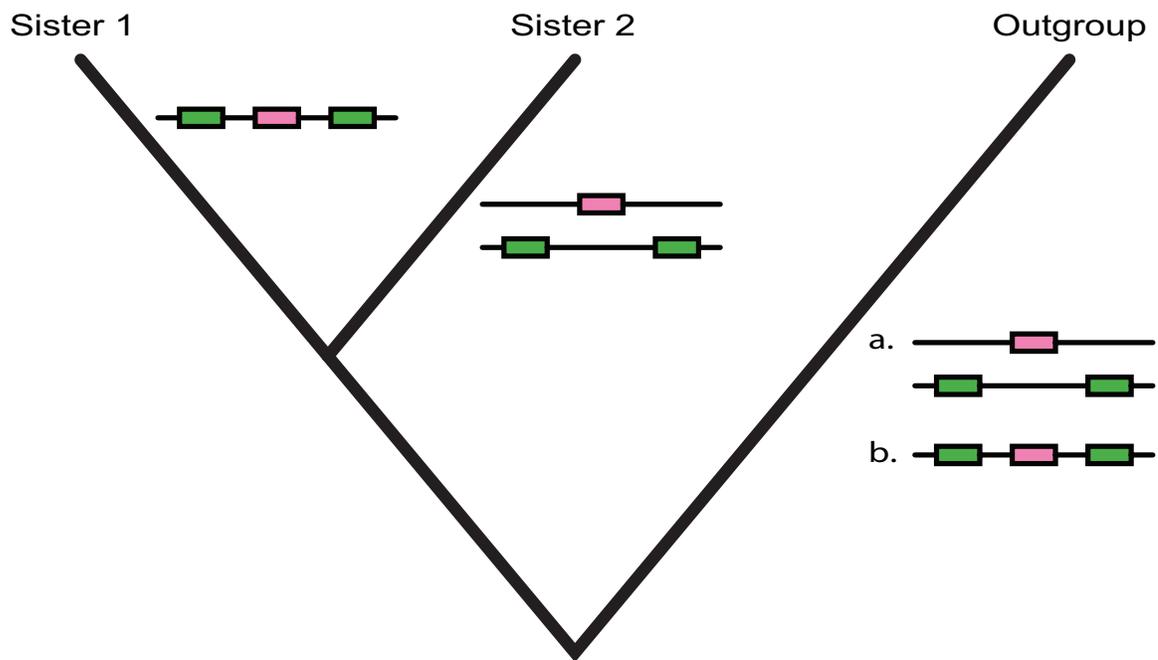


Figure 2.1: Phylogenetic analysis of gains and losses of nested gene structures. Gain or loss of a nested gene structure must have occurred if, within a pair of sister species, the structure is present in one but absent in the other. (a) Absence of the nested structure in the outgroup indicates its gain in sister 1. (b) Presence in the outgroup indicates its loss in sister 2.

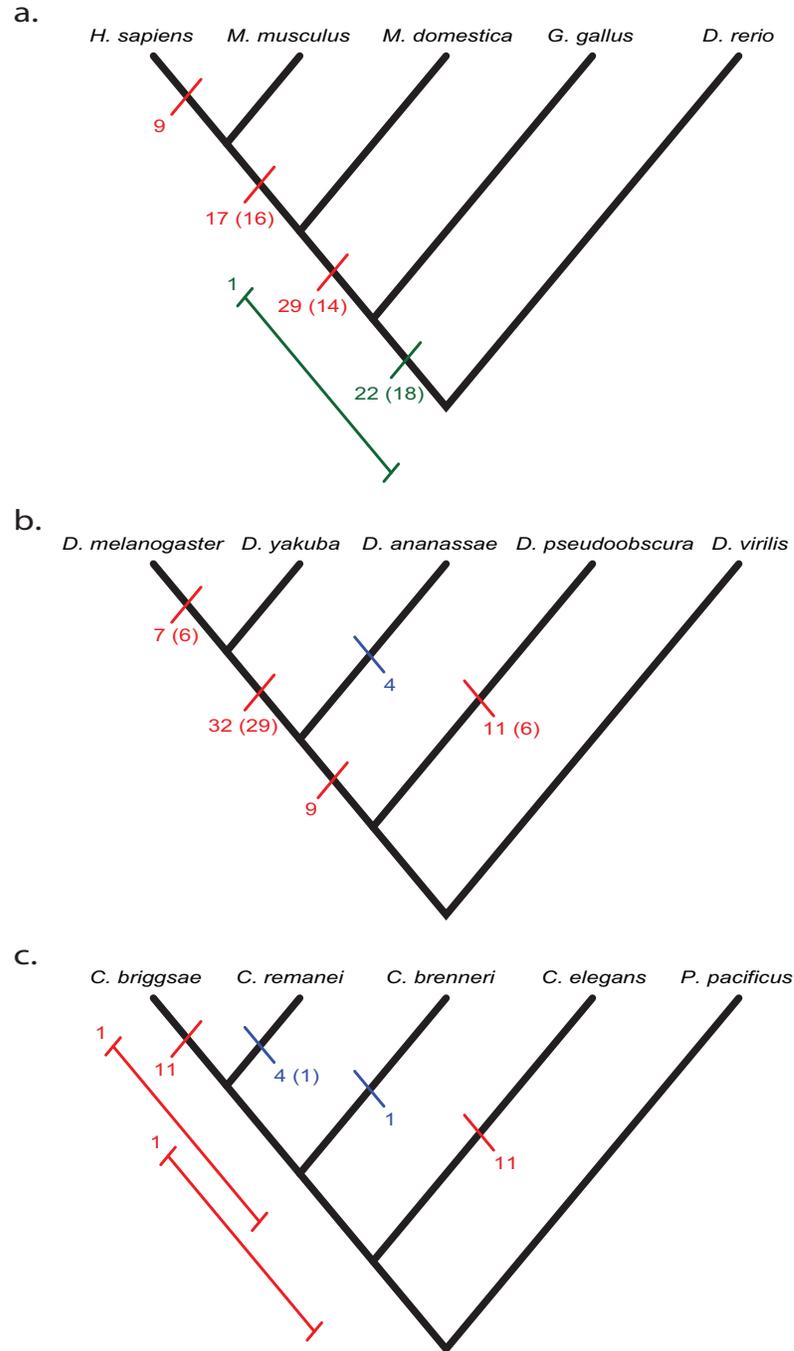


Figure 2.2: Dynamics of gain and loss of nested gene structures. Gains and losses of internal genes are labeled on the (a) vertebrate, (b) *Drosophila*, and (c) nematode phylogenies in red and blue, respectively. Nested gene structures that have a different nested state in the most distant outgroup, and therefore cannot be resolved between gains or losses, are shown in green. Independent events, or those that occur in different introns, are shown in parentheses. Events that could not be timed with a high enough resolution are shown on the side of each phylogeny.

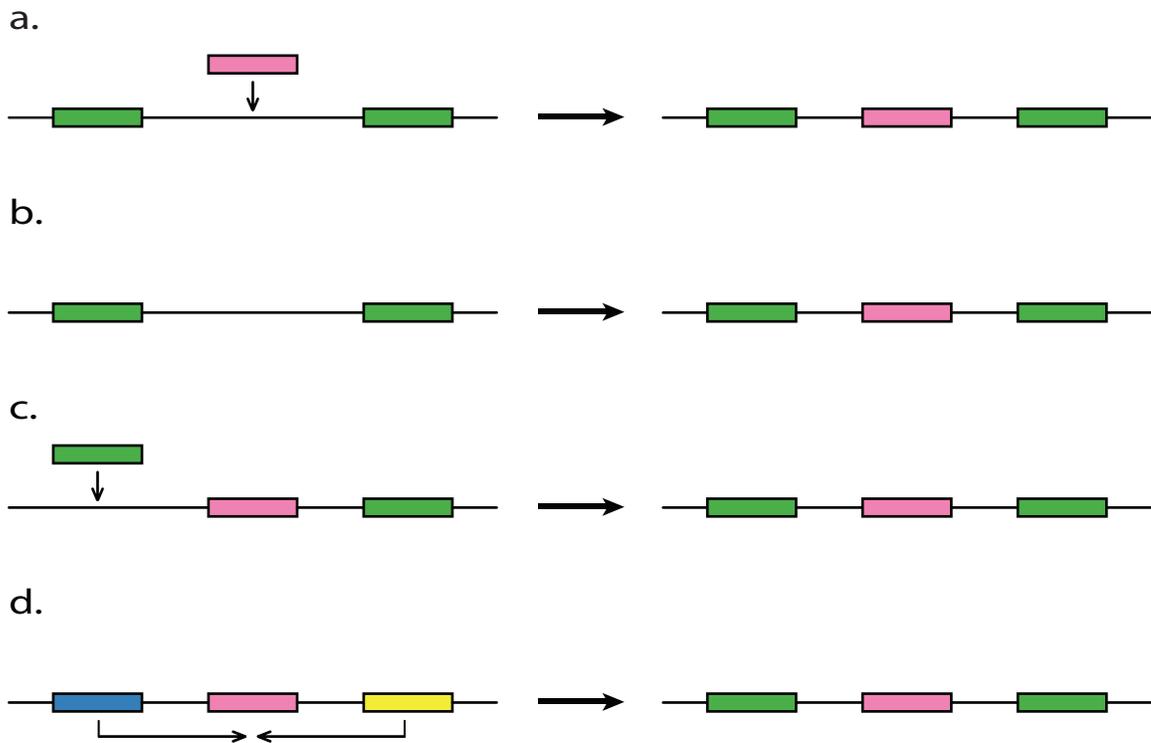


Figure 2.3: Scenarios for the origin of a nested gene structure. (a) Evolution of an internal gene by insertion of a DNA sequence into an intron of a pre-existing gene. (b) *De novo* evolution of a gene from an intronic sequence of a pre-existing gene. (c) Internalization of a gene after exon(s) acquisition of an adjacent gene. (d) Internalization of a gene via fusion of two flanking genes. Color key: pink, internal gene; green, exons of the external gene; blue and yellow, flanking genes.

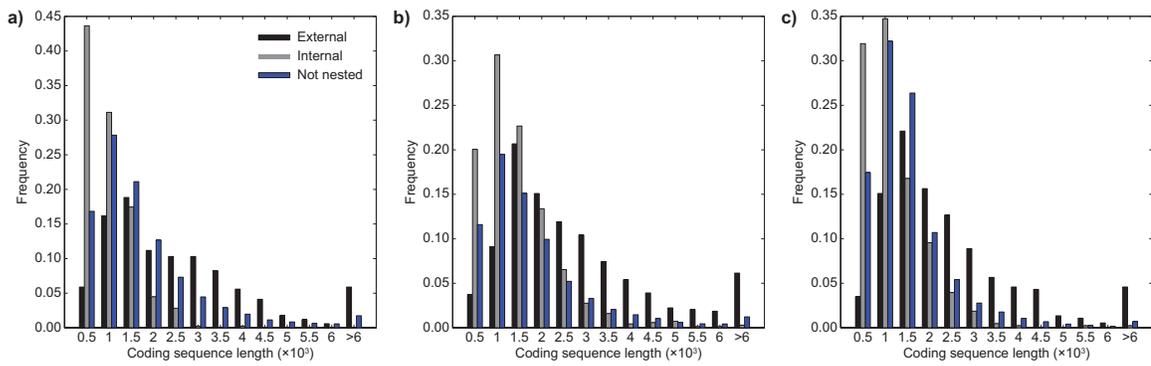


Figure 2.4: Distributions of total coding sequence lengths of external, internal, and non-nested genes in a) *H. sapiens*, b) *D. melanogaster*, and c) *C. elegans* genomes.

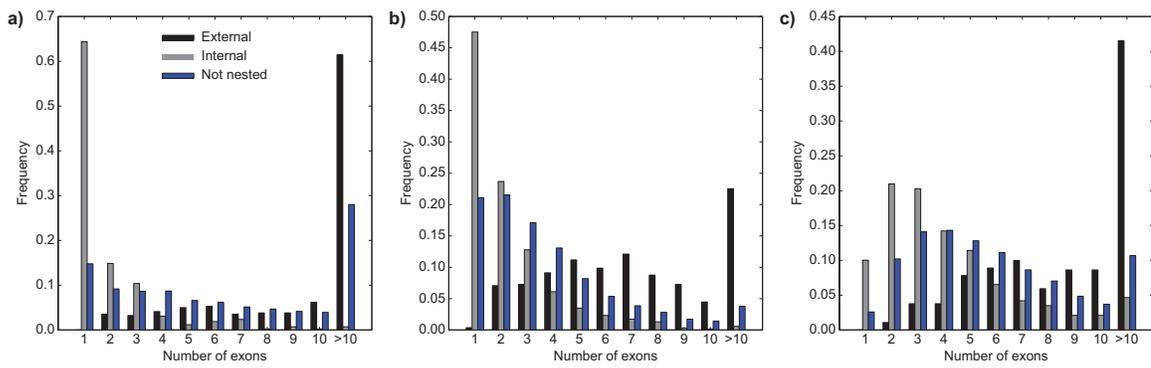


Figure 2.5: Distributions of numbers of exons in external, internal, and non-nested genes in a) *H. sapiens*, b) *D. melanogaster*, and c) *C. elegans* genomes.

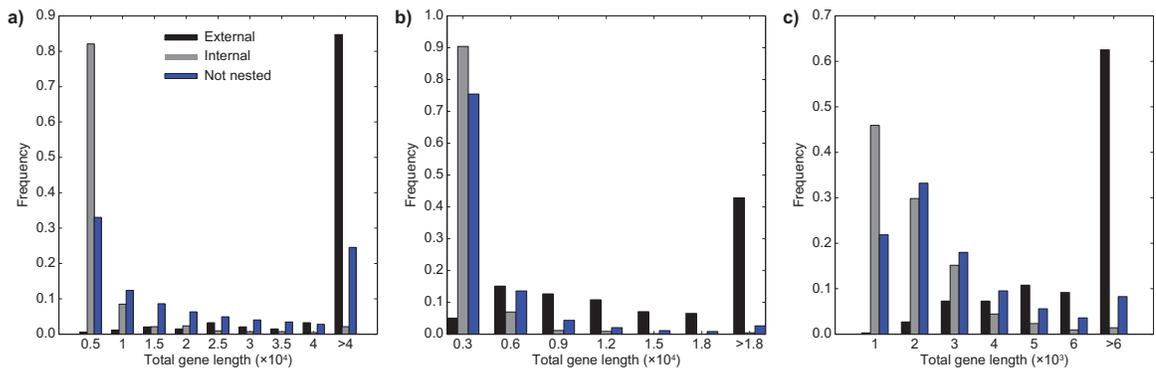


Figure 2.6: Distributions of total gene lengths of external, internal, and non-nested genes in a) *H. sapiens*, b) *D. melanogaster*, and c) *C. elegans* genomes.

CHAPTER III

Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution

3.1 Introduction

Eukaryotic genomes contain a variety of small noncoding RNAs, including microRNAs (miRNAs), repeat-associated small interfering RNAs (rasiRNAs), small interfering RNAs (siRNAs), and Piwi-interacting RNAs (piRNAs). miRNAs regulate the expression of protein-coding genes, rasiRNAs are involved in transposon silencing, and siRNAs play a dual role in silencing genes and transposons (*Sontheimer and Carthew, 2005*). Because of a number of similarities between rasiRNAs in *Drosophila* and piRNAs in mammals, rasiRNAs are considered to be a subclass of piRNAs. Thus, mammalian piRNAs are hypothesized to also be involved in transposon silencing, although they may perform other functions as well (*Aravin et al., 2007*). Some noncoding RNAs, in particular miRNAs, evolve very slowly (*Shabalina and Koonin, 2008*). In contrast, small-scale evolution of piRNA sequences proceeds at a rate typical of nonfunctional genomic regions (*Lau et al., 2006*). Here, we consider the large-scale evolution of mammalian piRNA clusters.

3.2 Results

3.2.1 Recent acquisition of many piRNA clusters

Thus far, 140 rat and mouse piRNA clusters have been described, each of which is most likely transcribed as a unit and subsequently processed into mature piRNAs (Aravin *et al.*, 2006; Girard *et al.*, 2006; Grivna *et al.*, 2006; Lau *et al.*, 2006; Watanabe *et al.*, 2006; Aravin *et al.*, 2007). We studied the evolution of each of these clusters within their genomic contexts (Table 3.1). For this purpose, we obtained regions that included 2 flanking protein-coding genes on either side of a cluster and constructed pairwise alignments of orthologous rat, mouse, human, dog, and cow regions. Thirty-seven clusters overlap protein-coding genes, often spanning several exons and introns. All of these clusters are ancestral, being present in rat, mouse, and human, which is not surprising because protein-coding genes are generally conserved. Among the remaining 103 clusters, each of which is contained within an intergenic region, only 43 are ancestral. The other 60 intergenic clusters were acquired recently. Fourteen were acquired after rat-mouse divergence, being present in 1 rodent (sister 1) and absent (aligned reliably against a gap) in the other rodent (sister 2) and in human (outgroup) (Fig. 3.1). Another 44 were acquired between rodent-primate and rat-mouse divergences, being present in rat and mouse and absent in human and in dog and/or cow. Evolution of 2 clusters is obscure because of a lack of reliable rodent-human alignments of the intergenic regions harboring them.

3.2.2 Ectopic recombination as a mechanism of piRNA cluster origin

Close similarity between rat and mouse genomes made it possible to reconstruct the course of events that led to the acquisition of 9 of the 14 rat- or mouse-specific clusters (Table 3.2). All 9 clusters arose via insertions of long DNA segments. Paralogs from which these sequences most likely originated (source paralogs) were identified

for 7 insertions along with several more distant paralogs in 6 cases. Six source paralogs are located on the same chromosome, between 192 and 259,000 Kb from the site of insertion. In these cases, the source paralog is similar not only to the inserted sequence, but also to segments upstream and/or downstream of the site of insertion, indicating that the insertion was mediated by ectopic (nonallelic homologous) recombination of these REs (*Lynch, 2007a*) (Fig. 3.2). In addition to flanking-acquired clusters and source paralogs, several REs are present in other locations. Often, all copies are confined to a single chromosome and sometimes also to 1 (either rat or mouse) genome. In the single case where the source paralog is located on a different chromosome, a copy of a L1 transposable element in the inserted sequence probably mediated the insertion. Out of 7 identified source paralogs, 5 are known to harbor clusters and, because not all rodent clusters are known (*Betel et al., 2007*), others may as well. Although source paralogs could not be found for 2 insertions, perhaps because some regions of rodent genomes are still not sequenced, the presence of low copy-number REs flanking these insertions suggests that ectopic recombination was the mechanism of these insertions as well.

The remaining 5 clusters acquired after rat-mouse divergence most likely arose via 3 independent events, because 2 pairs of related nearby clusters were probably acquired together. All of these clusters have several paralogs, including other known clusters. Their mechanisms of acquisition remain unclear because of unusually high rat-mouse divergence of their genomic regions, which prevents identification of the exact coordinates of cluster-harboring insertions and source paralogs. However, because these clusters are surrounded by REs, and their paralogs are mostly confined to the same chromosomes, their acquisitions were probably also because of ectopic recombination.

It is likely that the same mechanisms led to the 44 more distant acquisitions, although similarity between rodent and human genomes was insufficient to identify

the precise locations of cluster-harboring insertions. At least 1 rodent paralog was found for 35 of these clusters, and multiple paralogs are present in most cases, many of which are known clusters. The absence of paralogs in 9 cases could be because of either incomplete rodent genome sequences or longer times since cluster origin, which may have allowed some of them to diverge beyond recognition. In contrast to the pattern observed with acquired clusters, only 14 out of 80 ancestral clusters have an identifiable paralog, including 4 with multiple paralogs.

3.2.3 Two distinct subpopulations of clusters

Because of the presence of REs, genomic contexts of 60 acquired clusters are remarkably unstable. Only 13 are located within genomic regions that were preserved after acquisition of the cluster. The remaining 47 are within regions that underwent major rearrangements, including insertions, deletions, and inversions of genes and large (>100 Kb) segments of DNA. The pattern is very different for ancestral clusters. The genomic context was preserved in all 3 species for 66 ancestral cluster regions and was disrupted by nearby rearrangements for the other 14. Thus, clusters can be divided into 2 rather distinct subpopulations: stable and expanding.

The high rate of cluster acquisition and large-scale evolution of their genomic regions is unusual for mammalian genomes (*Lynch, 2007a*). To quantify this contrast, we randomly chose 103 intergenic cluster-like segments in rat or mouse as controls. With the exception of 7 control sequences for which no reliable alignments with human segments could be obtained, all control segments are ancestral. Thus, no clear cases of acquisition or loss were encountered. Furthermore, genomic regions harboring control segments are generally stable, as only 8 of them underwent major rearrangements.

3.2.4 Unremarkable small-scale evolution of piRNA clusters

In contrast to their rapid large-scale evolution, small-scale evolution of clusters proceeds at rates typical for mammalian genomes (*Lau et al.*, 2006). For 38 ancestral intergenic clusters within collinear genomic contexts, the mean cluster conservations are 0.59 in mouse-rat and 0.11 in rodent-human comparisons, respectively. The corresponding mean conservations are 0.54 and 0.12 for intergenic sequences surrounding these clusters, 0.53 and 0.13 for intergenic sequences between flanking genes, and 0.56 and 0.13 for control segments.

3.3 Discussion

Expansion of piRNA clusters, which are in effect noncoding genes, closely parallels the expansion of protein-coding genes by gene duplication. A variety of mechanisms are responsible for the duplication of protein-coding genes (*Cusack and Wolfe*, 2006), including ectopic recombination (*Lupski and Stankiewicz*, 2005; *Kwan-Wood and Jeffreys*, 2007; *Lynch*, 2007a; *Yang et al.*, 2008). Like piRNA clusters, protein-coding genes arising by ectopic recombination are often confined to the same chromosomes as their ancestral genes for 2 reasons because they are also flanked by mostly chromosome-specific REs (*Yang et al.*, 2008), and because the rate of intrachromosomal ectopic recombination is higher (*Lichten and Haber*, 1989). Thus, the tendency of clusters to reside on a small number of chromosomes (*Aravin et al.*, 2006; *Girard et al.*, 2006; *Grivna et al.*, 2006; *Lau et al.*, 2006; *Watanabe et al.*, 2006) is likely because of their mechanism of origin.

Approximately 43% (60/140) of all rodent piRNA clusters arose after rodent-primate divergence, and this fraction increases to 58% if clusters that overlap protein-coding genes are excluded. This exceeds the highest known expansion rate for a family of mammalian genes, that of olfactory receptors, 33% of which were acquired in mouse

after rodent-primate divergence (*Nimura and Nei, 2005*). Furthermore, gene losses are common for all large families of genes, including olfactory receptors (*Nimura and Nei, 2005*) and miRNAs, which are lost at the same rate with which they are acquired (*Lu et al., 2008*). However, not a single cluster loss was observed, although our method of analysis could readily detect such events.

Rapid expansion of piRNA clusters during the course of mammalian evolution is most likely driven by positive selection. Although the presence of REs increases the rates of both insertions and deletions (*Lupski, 2007*), deletions usually occur at a much higher rate than insertions (Table 8.1 in *Grivna et al. (2006)*). Thus, 60 cluster acquisitions without a single loss cannot be because of mutational pressure. More generally, long insertions are unlikely to be selectively neutral, and only beneficial ones can be fixed (*Kondrashov and Kondrashov, 2006*). Data on copy-number variants (CNVs) overlapping clusters within rat and mouse populations can be used to investigate selection on cluster acquisitions. Because positive selection increases the rate of evolution, but does not induce any long-lasting polymorphisms, the McDonald-Kreitman test (*Smith and Eyre-Walker, 2002*) would indicate positive selection if such CNVs are rare. Although currently available data on rat (*Guryev et al., 2008*) and mouse (*Graubert et al., 2007; She et al., 2008*) seem to be consistent with this, analysis of a larger number of wild-type rat and mouse genotypes is necessary. If piRNAs are indeed involved in transposon silencing, it is natural to assume that selection for cluster acquisitions is caused by an arms race between expanding families of mammalian transposons and piRNA clusters.

3.4 Materials and methods

3.4.1 Classification of rodent piRNA cluster

Genomic locations of 100 rat and 94 mouse piRNA clusters were obtained from ref. 4. Clusters labeled as rat-mouse orthologs were checked to ensure that they were located between orthologous flanking genes; if not, they were analyzed as distinct clusters. Two clusters were not analyzed because of the poor quality of available sequences in their genomic regions.

3.4.2 Phylogenetic analysis

Sequences of clusters, along with the 4 closest flanking genes, were downloaded from GenBank (*Benson et al.*, 2009) at <ftp://ftp.ncbi.nih.gov/genomes/>. Orthologous segments from the other rodent, human, dog, and cow genomes were identified by applying BLASTP (*Altschul et al.*, 1997) to flanking genes. Pairwise alignments of these regions between the cluster-containing rodent(s) and all other species were constructed with OWEN (*Ogurtsov et al.*, 2002). Parameters were initially strict, with a requirement of 16 successive matches and $P < 10^{-8}$, and were progressively relaxed to 8 successive matches and $P < 0.001$. A cluster was considered to be conserved within a pair of species when it was part of an unambiguous alignment.

3.4.3 Identification of paralogs

Insertion sites of clusters acquired after rat-mouse divergence corresponded to alignment gaps. Paralogs for inserted sequences were identified using BLASTN (24) and BLAT (*Karolchik et al.*, 2008). All paralogs were aligned against the inserted sequence with OWEN, and the paralog with the best alignment was assumed to be the source. Paralogs were also found in the same way for REs and all remaining clusters, with the requirement that BLAT alignments covered $>50\%$ of the query sequence.

3.4.4 Measurement of small-scale evolution

Rat-mouse and rodent-human divergences for ancestral cluster regions were measured for clusters, surrounding intergenic segments, and intergenic sequences between flanking genes. Divergences between cluster-containing sequences acquired after rat-mouse divergence and their source paralogs were also calculated for clusters and surrounding inserted segments. Conservation scores were computed by dividing the number of matching bases by the length of the regions of interest.

3.5 Acknowledgments

We are deeply indebted to John Kim for proposing that we study the evolution of piRNAs. We also thank Michael Lynch, Fyodor Kondrashov, Yegor Bazykin, and David Ginsburg for their helpful comments and suggestions.

Table 3.1: Rat and mouse piRNA clusters

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 38	chr9: 105120000 – 105160000	40001	Intergenic	Acquired after rat-mouse divergence	>2	>2	Rat 28 / Mouse 16, Rat 37, Rat 74	Chromosomal rearrangements
Rat 99	chr17: 20354000 – 20367000	13001	Intergenic	Acquired after rat-mouse divergence	0	0		Inversion of cluster region
Rat 39	chr1: 47839000 – 47873000	34001	Intergenic	Acquired after rat-mouse divergence	>2	>2	Rat 49, Rat 56, Rat 86	Inversion of rodent-specific predicted gene flanking cluster
Rat 56	chr1: 47729000 – 47753000	24001	Intergenic	Acquired after rat-mouse divergence	>2	>2	Rat 39, Rat 49, Rat 86	Inversion of rodent-specific predicted gene flanking cluster
Rat 65	chr15: 84157000 – 84160000	3001	Intergenic	Acquired after rat-mouse divergence	1	0	Rat 81	None
Mouse 84	chr8: 35860000 – 35873000	13001	Intergenic	Acquired after rat-mouse divergence	0	>2	Mouse 70	None
Mouse 35	chr10: 85297000 – 85334000	37001	Intergenic	Acquired after rat-mouse divergence	>2	>2	Mouse 42, Mouse 54	None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 42	chr5: 25405000 - 25430000	25001	Intergenic	Acquired after rat-mouse divergence	>2	>2	Rat 59	Complex rearrangements
Rat 52	chr3: 140578000 - 140591000	13001	Intergenic	Acquired after rat-mouse divergence	>2	1	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Complex rearrangements
Rat 59	chr5: 25621000 - 25636000	15001	Intergenic	Acquired after rat-mouse divergence	>2	>2	Rat 42	Complex rearrangements

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 72	chr3: 140378000 - 140397000	19001	Intergenic	Acquired after rat-mouse divergence	>2	0	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Complex rearrangements
Rat 53	chr3: 140041000 - 140077000	36001	Intergenic	Acquired after rat-mouse divergence	>2	0	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Inversion near cluster
Rat 71	chrX: 124261000 - 124262000	1001	Intergenic	Acquired after rat-mouse divergence	>2	0	Rat 91	Complex rearrangements

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Mouse 50	chr8: 91999000 - 92013000	14001	Intergenic	Acquired after rat-mouse divergence	0	0		Inter-chromosomal insertion of flanking gene and insertion of intergenic DNA near cluster
Mouse 34	chr6: 128190000 - 128224000	34001	Intergenic	Acquired after rodent-primate divergence	1	1		Large insertion of intergenic DNA near cluster
Rat / Mouse 20	chr10: 92629000 - 92654000 / chr11: 103533000 - 103559000	25001 / 26001	Intergenic	Acquired after rodent-primate divergence	0	0		Inversion of flanking gene

Cluster(s) of cluster(s)*	Coordinates of cluster(s)*	Length of cluster(s) ter(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 24 / Mouse 41	chr4: 119527000 - 119539000 / chr6: 85140000 - 85156000	12001 / 16001	Intergenic	Acquired rodent-primate divergence	>2	>2	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	None
Rat 32	chr3: 140968000 - 140994000	26001	Intergenic	Acquired rodent-primate divergence	>2	0	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41 Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Complex rear- rangements

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 35 / Mouse 27	chr4: 12001 / 116384000 - 116396000 / chr6: 81912000 - 81933000	21001	Intergenic Acquired	after rodent-primate divergence	0	0	0	None
Rat 48 / Mouse 24	chr1: 11001 / 184495000 - 184506000 / chr7: 121421000 - 121461000	40001	Intergenic Acquired	after rodent-primate divergence	0	0	0	None
Rat 67	chr11:82952000 - 82970000	18001	Intergenic Acquired	after rodent-primate divergence	>2	>2	>2	None
Rat 76 / Mouse 64	chr5: 8001 / 143156000 - 143164000 / chr4: 123278000 - 123287000	9001	Intergenic Acquired	after rodent-primate divergence	0	0	0	None
Rat 86	chr1: 51025000 - 51039000	14001	Intergenic Acquired	after rodent-primate divergence	>2	>2	Rat 39, Rat 49, Rat 56	Insertion of intergenic DNA on both sides of cluster

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 90	chr4: 165034000 – 165054000	20001	Intergenic	Acquired rodent-primate divergence	>2	>2		Complex rearrangements
Mouse 63	chr14: 22077000 – 22113000	36001	Intergenic	Acquired rodent-primate divergence	1	>2	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51	Chromosomal fission near cluster region
Mouse 48	chr5: 142790000 – 142801000	11001	Intergenic	Acquired rodent-primate divergence	0	0		None
Mouse 85	chr13: 24271000 – 24273000	2001	Intergenic	Acquired rodent-primate divergence	0	0		None
Mouse 92	chr4: 123336000 – 123342000	6001	Intergenic	Acquired rodent-primate divergence	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 17 / Mouse 21	chr5: 79245000 - 79257000 / chr4: 61654000 - 61673000	12001 / 19001	Intergenic	Acquired after rodent-primate divergence	1	1		Inversion of cluster region
Rat 37	chr9: 105449000 - 105469000	20001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Rat 28 / Mouse 16, Rat 38, Rat 74	Inter-chromosomal insertion of flanking genes and inversion of intergenic DNA near cluster
Rat 43	chr17: 17515000 - 17556000	41001	Intergenic	Acquired after rodent-primate divergence	2	1	Mouse 60	Inter-chromosomal insertion of flanking genes and inversions of flanking genes on both sides of cluster

Cluster(s) of cluster(s)*	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of cluster(s)	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 46 / Mouse 42	chr7: 18786000 - 18819000 / chr10: 86113000 - 86163000	33001 / 50001	Intergenic	Acquired after rodent-primate divergence	>2	>2		Inter- chromosomal insertion and inversion of flanking genes
Rat 54	chr1: 102090000 - 102111000	21001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Rat 58, Rat 61 / Mouse 45	Inter- chromosomal insertion of flanking genes and inversion of cluster region
Rat 58	chr1: 103000000 - 103019000	19001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Rat 54, Rat 61 / Mouse 45	Inter- chromosomal insertion of flanking genes and insertion of intergenic DNA

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 61 / Mouse 45	chr1: 18001 / 28001 chr7: 102845000 / 49122000 - 49150000	18001 / 28001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Rat 54, Rat 58	Inter-chromosomal insertion of flanking genes and inversion of intergenic DNA near cluster
Rat 74	chr9: 105273000 - 105298000	25001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Rat 28 / Mouse 16, Rat 37, Rat 38	Inter-chromosomal insertion of flanking genes and inversion of cluster region
Rat 80	chr5: 79201000 - 79204000	3001	Intergenic	Acquired after rodent-primate divergence	0	0		Inversion of cluster region
Rat 93	chr10: 90649000 - 90661000	12001	Intergenic	Acquired after rodent-primate divergence	0	0		Inversion of flanking gene

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat	chr2: 40,57 / 105696000	32001, 37001 /	Intergenic	Acquired after rodent-primate divergence	>2	>2		None
Mouse	chr2: 91 - 105728000, 105657000 - 105694000	25001						
	chr3: 20186000 - 20211000							
Mouse	chr10: 54 86171000 - 86183000	12001	Intergenic	Acquired after rodent-primate divergence	>2	>2		Inter-chromosomal insertion and inversion of flanking genes
Mouse	chr14: 39 17284000 - 17311000	27001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Mouse 26	Chromosomal rearrangements
Mouse	chr13: 60 52433000 - 52440000	7001	Intergenic	Acquired after rodent-primate divergence	1	1	Rat 43	Inter-chromosomal insertion of flanking genes and inversion of flanking genes

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs after	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Mouse 70	chr8: 26050000 - 26065000	15001	Intergenic	Acquired rodent-primate divergence	after 0	0	1	Mouse 84	2 large deletions at boundaries of clusters in rodents
Mouse 31	chr2: 151143000 - 151162000	19001	Intergenic	Acquired rodent-primate divergence	after >2	>2	>2	Mouse 36	Inversion of flanking genes
Rat 23 / Mouse 30	chr3: 141656000 - 141670000 / chr2: 151055000 - 151073000	14001 / 18001	Intergenic	Acquired rodent-primate divergence	after >2	>2	>2	Rat 30 / Mouse 29	Inversion of flanking genes
Rat 29	chr11: 84973000 - 85028000	55001	Intergenic	Acquired rodent-primate divergence	after 0	0	0		Inversion of flanking genes
Rat 30 / Mouse 29	chr3: 141591000 - 141601000 / chr2: 150989000 - 151008000	10001 / 19001	Intergenic	Acquired rodent-primate divergence	after >2	>2	>2	Rat 23 / Mouse 30	Inversion of flanking genes

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 47 / Mouse 61	chr4: 120368000 - 120385000 / chr6: 86062000 - 86079000	17001 / 17001	Intergenic	Acquired after rodent-primate divergence	0	0		Inversion of flanking genes
Rat 81	chr10: 36520000 - 36523000	3001	Intergenic	Acquired after rodent-primate divergence	1	0	Rat 65 Inversion of flanking genes	
Mouse 26	chr14: 40245000 - 40263000	18001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Mouse 39	Inversion of flanking genes
Mouse 36	chr2: 150882000 - 150899000	17001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Mouse 31	Inversion of flanking genes
Mouse 65	chr13: 49938000 - 49944000	6001	Intergenic	Acquired after rodent-primate divergence	>2	>2	Mouse 86	Inversion of flanking genes

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Mouse 81	chr14: 40079000 – 40097000	18001	Intergenic	Acquired rodent-primate divergence	>2	>2	>2		Inversion of flanking genes
Rat 33	chr12: 42905000 – 42939000	34001	Intergenic	Acquired rodent-primate divergence	>2	1			Inversion of cluster region
Rat 73	chr2: 105598000 – 105623000	25001	Intergenic	Acquired rodent-primate divergence	>2	>2	>2	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Inversion of flanking genes
Mouse 38	chr15: 83436000 – 83450000	14001	Intergenic	Acquired rodent-primate divergence	0	0	0	Rat 60	Inversion of cluster and flanking gene
Mouse 28	chr10: 66270000 – 66293000	23001	Intergenic	Acquired rodent-primate divergence	0	0	>2		None

Cluster(s) of cluster(s)*	Coordinates	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s) after rodent-primate divergence	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 60	chr7: 121600000 - 121604000	4001	Intergenic	Unknown	0	1	Mouse 38	Inversion of flanking genes
Rat 63 / Mouse 33	chr9: 92473000 - 92479000 / chr1: 92900000 - 92905000	6001 / 5001	Intergenic	Unknown	0	0		None
Mouse 67	chr13: 51154000 - 51161000	7001	Intergenic	Unknown	0	0		None
Rat 49	chrUn: 42595000 - 42641000	46001	Intergenic	Not analyzed	>2	>2	Rat 39, Rat 56, Rat 86	n/a
86	chr17: 20779000 - 20785000	6001	Intergenic	Not analyzed	>2	>2	Mouse 65	n/a
Rat 31	chr17: 20742000 - 20810000	68001	Intergenic	Ancestral	0	0		Inversion of flanking genes

Cluster(s) of cluster(s)*	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 85 / Mouse 68,89	chr 17:20188000 - 20198000 / chr13: 49820000 - 49833000, 49373000 - 49377000	10001 / 13001, 4001	Intergenic	Ancestral	>2	>2	>2	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Inversion of flanking genes
Rat 16 / Mouse 15	chr7: 112772000 - 112797000 / chr15: 74701000 - 74752000	25001 / 51001	Intergenic	Ancestral	0	0	0		None
Rat 36 / Mouse 32	chr5: 75292000 - 75303000 / chr4: 57398000 - 57406000	11001 / 8001	Intergenic	Ancestral	0	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 79 / Mouse 66	chr9: 90835000 / chr1: 90839000	4001 / 4001	Intergenic	Ancestral	0	0		None
Rat 97	chr12: 91310000 - 91314000	8001	Intergenic	Ancestral	0	0		None
Rat 41,51 / Mouse 44	chr1: 13931000 - 13959000, chr10: 13901000 - 13927000	28001, 26001 / 35001	Intergenic	Ancestral	0	0		None
Rat 68 / Mouse 72	chr13: 37654000 - chr1: 37658000	4001 / 3001	Intergenic	Ancestral	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 1 / Mouse 1	chr20: 5480000 - 5581000	101001 / 95001	Intergenic	Ancestral	0	0		None
	/ chr17: 25458000 - 25553000							
Rat 2 / Mouse 5	chr4: 164469000 - 164546000	77001 / 80001	Intergenic	Ancestral	0	0		None
	/ chr6: 127799000 - 127879000							
Rat 5 / Mouse 6	chr1: 129070000 - 129164000	94001 / 79001	Intergenic	Ancestral	0	0	Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	None
	/ chr7: 69622000 - 69701000							

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 6 / Mouse 13	chr8: 57725000 - 57777000 / chr9: 54236000 - 54298000	52001 / 62001	Intergenic	Ancestral	0	0		None
Rat 7 / Mouse 2	chr8: 72069000 - 72129000 / chr9: 67715000 - 67803000	60001 / 88001	Intergenic	Ancestral	0	0		None
Rat 8	chr12: 5597000 - 5633000	36001	Intergenic	Ancestral	0	0		None
Rat 9 / Mouse 14	chr7: 68946000 - 69002000	60001 / 56001	Intergenic	Ancestral	0	0		None
Rat 10 / Mouse 17	chr5: 114521000 - 114575000 / chr4: 93669000 - 93725000	54001 / 56001	Intergenic	Ancestral	0	0		None

Cluster(s) of cluster(s)*	Coordinates of cluster(s)*	Length of cluster(s) ter(s)	Genomic context of cluster(s) ter(s)	Evolution of cluster(s) ter(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 12 / Mouse 9	chr7: 96057000 - 96116000 / chr15: 59313000 - 59370000	59001 / 57001	Intergenic	Ancestral	0	0		Inversion of flanking gene
Rat 18	chr15: 218000 - 262000	44001	Intergenic	Ancestral	0	1	Mouse 12	None
Rat 19 / Mouse 37	chr19: 40002000 - 40017000 / chr8: 108718000 - 108731000	15001 / 13001	Intergenic	Ancestral	0	0		None
Rat 25	chr19: 15907000 - 15945000	38001	Intergenic	Ancestral	0	0	Mouse 25	None
Rat 28 / Mouse 16	chr9: 105522000 - 105551000 / chr17: 64393000 - 64458000	29001 / 65001	Intergenic	Ancestral	>2	>2	Rat 37, Rat 38, Rat 74	Inversion of cluster region

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 34 / Mouse 43	chr5: 154408000 - 154412000 / chr4: 134996000 - 135009000	4001 / 13001	Intergenic	Ancestral	0	0	0		None
Rat 44 / Mouse 75	chr12: 20765000 - 20783000 / chr5: 135888000 - 135908000	18001 / 20001	Intergenic	Ancestral	0	0	0		None
Rat 45	chr4: 119575000 - 119601000	26001	Intergenic	Ancestral	>2	>2	>2	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 50 / Mouse 59	chr1: 241697000 - 241724000 / chr19: 36956000 - 36962000	27001 / 6001	Intergenic	Ancestral	0	0		None
Rat 62	chr14: 41998000 - 42011000	13001	Intergenic	Ancestral	0	0		None
Rat 64	chr9: 7094000 - 7096000	2001	Intergenic	Ancestral	0	0		None
Rat 69	chr17: 18976000 - 19001000	25001	Intergenic	Ancestral	0	0		None
Rat 84	chr6: 93598000 - 93616000	18001	Intergenic	Ancestral	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 87 / Mouse 77	chr7: 31878000 - 31887000 / chr10: 94234000 - 94238000	9001 / 4001	Intergenic	Ancestral	0	0		None
Rat 91	chrX: 123931000 - 123946000	15001	Intergenic	Ancestral	>2	0	Rat 71	None
Rat 92 / Mouse 79	chrX: 152895000 - 152918000 / chrX: 62174000 - 62206000	23001 / 32001	Intergenic	Ancestral	0	0		None
Rat 94	chr2: 44320000 - 44347000	27001	Intergenic	Ancestral	0	0		None
Rat 95 / Mouse 47	chr2: 121139000 - 121146000 / chr3: 34676000 - 34687000	7001 / 11001	Intergenic	Ancestral	0	0	None	

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of cluster(s)	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Mouse 25	chr8: 90782000 – 90795000	13001	Intergenic	Ancestral	1	0	0	Rat 25	None
Mouse 82	chr1: 57602000 – 57611000	9001	Intergenic	Ancestral	0	0	0		None
Mouse 90	chr13: 20869000 – 20871000	2001	Intergenic	Ancestral	0	0	0		None
Mouse 93	chr10: 84029000 – 84037000	8001	Intergenic	Ancestral	0	0	0		None
Mouse 7	chr5: 112507000 – 112564000	57001	Intergenic	Ancestral	0	0	0		Inter-chromosomal movement of flanking genes
Mouse 62	chr17: 47253000 – 47256000	3001	Intergenic	Ancestral	0	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 11 / Mouse 11	chr1: 124891000 - 124959000 / chr7: 65734000 - 65810000	68001 / 76001	Intergenic	Ancestral	0	0		None
Rat 89	chr20: 4001	4001	Intergenic	Ancestral	0	0		None
Mouse 87	chr8: 115820000 - 115829000	9001	Intergenic	Ancestral	0	0		None
Rat 66 / Mouse 55	chr4:165701000 - 165712000 / chr6: 128462000 - 128474000	11001 / 12001	Overlapping gene(s)	Ancestral	0	0		Inversion of flanking genes
Rat 3 / Mouse 4	chr3: 76817000 - 76892000 / chr2: 92374000 - 92462000	75001 / 88001	Overlapping gene(s)	Ancestral	0	0		None

Cluster(s) of cluster(s)*	Coordinates	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of rat paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 4 / Mouse 8	chr6: 12254000 - 12262400 / chr12: 95876000 - 95950000	84001 / 74001	Over- lapping gene(s)	Ancestral	0	0	0		None
Rat 13 / Mouse 18	chr12: 44134000 - 44189000 / chr5: 114033000 - 114076000	55001 / 43001	Over- lapping gene(s)	Ancestral	0	0	0		Inversion of flanking genes
Rat 14 / Mouse 3	chr18: 63622000 - 63658000 / chr18: 67419000 - 67485000	36001 / 66001	Over- lapping gene(s)	Ancestral	0	0	0		Multiple inver- sions of flank- ing genes
Rat 15	chr20: 13186000 - 13223000	37001	Over- lapping gene(s)	Ancestral	0	1	1	Mouse 19	Inversion of flanking genes

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 21 / Mouse 46	chr3: 115118000 - 115142000 / chr2: 127432000 - 127445000	24001 / 13001	Overlapping gene(s)	Ancestral	1	1	Rat 5 / Mouse 6, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	Inversion of flanking genes
Rat 22 / Mouse 40	chr4: 122403000 - 122440000 / chr6: 88042000 - 88069000	37001 / 27001	Overlapping gene(s)	Ancestral	0	0		Inversion of flanking genes
Rat 26 / Mouse 23	chr20: 29911000 - 29965000 / chr10: 62171000 - 62224000	54001 / 53001	Overlapping gene(s)	Ancestral	0	0		None

Cluster(s) of cluster(s)*	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 27 / Mouse 22	chr7: 116628000 - 116678000 / chr15: 78725000 - 78767000	50001 / 42001	Over- lapping gene(s)	Ancestral	0	0		None
Rat 55 / Mouse 58	chr2: 221181000 - 221209000 / chr3: 124237000 - 124263000	28001 / 26001	Over- lapping gene(s)	Ancestral	1	1	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 51, Mouse 63	None
Rat 70	chr13: 73630000 - 73667000	37001	Over- lapping gene(s)	Ancestral	0	0		None
Rat 75	chrUn: 15394000 - 15406000	12001	Over- lapping gene(s)	Ancestral	0	0		None
Rat 77	chr12: 27425000 - 27443000	18001	Over- lapping gene(s)	Ancestral	0	1	Mouse 80	None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 78	chr7: 94716000 - 94745000	29001	Over-lapping gene(s)	Ancestral	0	0		None
Rat 82	chr1: 92677000 - 92707000	30001	Over-lapping gene(s)	Ancestral	0	0		None
Rat 83 / Mouse 71	chr9: 90967000 - 90984000 / chr1: 91445000 - 91467000	17001 / 22001	Over-lapping gene(s)	Ancestral	0	0		None
Rat 88 / Mouse 49	chr8: 47138000 - 47143000 / chr9: 44078000 - 44086000	5001 / 8001	Over-lapping gene(s)	Ancestral	0	0		None
Rat 96	chr15: 187000 - 217000	30001	Over-lapping gene(s)	Ancestral	0	0		None
Rat 98	chr5: 63093000 - 63106000	13001	Over-lapping gene(s)	Ancestral	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Rat 100 / Mouse 94	chr10: 84482000 - 84488000 / chr11: 95874000 - 95880000	6001 / 6001	Overlapping gene(s)	Ancestral	0	0		Inter-chromosomal movement of flanking gene
Mouse10	chr5: 148721000 - 148766000	45001	Overlapping gene(s)	Ancestral	0	0		None
Mouse 12	chr14: 21743000 - 21796000	53001	Overlapping gene(s)	Ancestral	1	0	Rat 18	Inter-chromosomal movement of flanking genes
Mouse 19	chr10: 75491000 - 75537000	46001	Overlapping gene(s)	Ancestral	1	0	Rat 15	None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs	Number of paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Mouse 51	chr14: 22009000 – 22038000	29001	Overlapping gene(s)	Ancestral	1	1	Rat 5 / Mouse 6, Rat 21 / Mouse 46, Rat 24 / Mouse 41, Rat 32, Rat 45, Rat 52, Rat 53, Rat 55 / Mouse 58, Rat 72, Rat 73, Rat 85 / Mouse 68/89, Mouse 63	None
Mouse 52	chr10: 127189000 – 127202000	13001	Overlapping gene(s)	Ancestral	0	0		None
Mouse 53	chr7: 27076000 – 27084000	8001	Overlapping gene(s)	Ancestral	0	0		None
Mouse 56	chr11: 108102000 – 108107000	5001	Overlapping gene(s)	Ancestral	0	0		Large insertion into intron of overlapping gene
Mouse 57	chr5: 134131000 – 134139000	8001	Overlapping gene(s)	Ancestral	0	0		None

Cluster(s)	Coordinates of cluster(s)*	Length of cluster(s)	Genomic context of cluster(s)	Evolution of cluster(s)	Number of paralogs of rat	Number of mouse paralogs	Paralogous cluster(s)	Rearrangements occurring after acquisition of cluster(s)
Mouse 69	chr9: 51827000 - 51831000	4001	Over-lapping gene(s)	Ancestral	0	0		None
Mouse 73	chr15: 73176000 - 73188000	12001	Over-lapping gene(s)	Ancestral	0	0		None
Mouse 74	chr5: 101021000 - 101040000	19001	Over-lapping gene(s)	Ancestral	0	0		None
Mouse 76	chr15: 80000000 - 80039000	39001	Over-lapping gene(s)	Ancestral	0	0		None
Mouse 78	chr8: 109603000 - 109621000	18001	Over-lapping gene(s)	Ancestral	0	0		None
Mouse 80	chr5: 129378000 - 129393000	15001	Over-lapping gene(s)	Ancestral	1	0	Rat 77	None
Mouse 83	chr10: 60946000 - 60952000	6001	Over-lapping gene(s)	Ancestral	0	0		None
Mouse 88	chr6: 92248000 - 92270000	22001	Over-lapping gene(s)	Ancestral	0	0		None

*All coordinates are given according to the UCSC Genome Browser (mouse build mm7, rat build rn3).

Table 3.2: piRNA clusters acquired after rat-mouse divergence

Cluster	Coordinates of cluster*	Coordinates of insertion	Coordinates of source paralog	Known cluster(s) in source paralog	Coordinates of upstream RE†	Copies of downstream stream RE	Copies of upstream RE†	Coordinates of stream RE	Copies of downstream stream RE	Insertion mechanism
Mouse 35	chr10: 85201534 - 85151751 - 85322289 (37001)	chr10: 85151751 - 85972115 - 86210281 (238167)	chr10: 85972115 - 85148628 - 85151750 (3123)	Mouse 42, Mouse 54	chr10: 85148628 - 85151750 (3123)	None	4, 4, 4	None	None	Ectopic recombination
Mouse 50	chr8: 95933856 - 95947856 (14001)	chr8: 95925416 - 95964942 (39527)	chr8: 95925416 - 95964942 (39527)	n/a	n/a	n/a	n/a	n/a	n/a	Ectopic recombination
Mouse 84	chr8: 3815988 - 38168988 (13001)	chr8: 38151734 - 38187761 (36028)	chr8: 38151734 - 38187761 (36028)	None	chr8: 38140985 - 38151733 (10749)	chr8: 38140985 - 38151733 (10749)	0, 0, 0	chr8: 38187762 - 38211000 (23239)	0, 0, 0	Ectopic recombination
Rat 38	chr9: 104910520 - 104950520 (40001)	chr9: 104901373 - 105002947 (101575)	chr9: 105195330 - 105281154 (85825)	Rat 37	chr9: 104893529 - 104901372 (7844)	chr9: 104893529 - 104901372 (7844)	4, 0, 0	chr9: 105002948 - 105011956 (9009)	3, 0, 0	Ectopic recombination
Rat 39	chr1: 47836055 - 47870055 (34001)	chr1: 47801096 - 47896357 (95,262)	chr1: 48366238 - 48721121 (354884)	None	chr1: 47726051 - 47801095 (75,045)	chr1: 47726051 - 47801095 (75,045)	>10, 0, 1	chr1: 47896358 - 48018741 (122384)	6, 0, 0	Ectopic recombination

Cluster	Coordinates of cluster*	Coordinates of insertion	Coordinates of source paralog	Known cluster(s) in source paralog	Coordinates of upstream RE	Coordinates of downstream RE	Copies of upstream RE†	Coordinates of downstream RE	Copies of downstream RE	Insertion mechanism
Rat 42	chr5: 25405000 - 25430000 (25001)	Unknown	Not identified	n/a	n/a	n/a	n/a	n/a	n/a	Ectopic recombination (most likely acquired with Rat 59)
Rat 52	chr3: 140672427 - 140685427 (13001)	Unknown	Not identified	n/a	n/a	n/a	n/a	n/a	n/a	Ectopic recombination (most likely acquired with Rat 72)
Rat 53	chr3: 140135427 - 140171427 (36001)	chr3: 140062045 - 140187651 (125607)	chr3: <140618130 - 140689393 (>71264)	Rat 52	None	chr3: 140187652 - 140223334 (35683)	None	chr3: 140187652 - 140223334 (35683)	0, 0, 0	Ectopic recombination
Rat 56	chr1: 47726055 - 47750055 (24001)	chr1: 47726051 - 47755189 (29139)	chrUn: 42574523 - 42621624 (47102)	Rat 49 (extends downstream beyond paralog)	None	chr1: 47755190 - 47769541 (14352)	None	chr1: 47755190 - 47769541 (14352)	>10, 0, 0	Ectopic recombination

Cluster	Coordinates of cluster*	Coordinates of insertion	Coordinates of source paralog	Known cluster(s) in source paralog	Coordinates of upstream stream RE	Copies of upstream RE [†]	Coordinates of downstream stream RE	Copies of downstream stream RE	Insertion mechanism
Rat 59	chr5: 25621000 - 25636000 (15001)	Unknown	Not identified	n/a	n/a	n/a	n/a	n/a	Ectopic recombination (most likely acquired with Rat 42)
Rat 65	chr15: 84141220 - 84144220 (3001)	chr15: 84135414 - 84144834 (9421)	chr10: 36518413 - 36960911 (442499)	Rat 81	None	None	None	None	Retrotransposed by L1Pa3 TE
Rat 71	chrX: 124187567	Unknown	Not identified	n/a	n/a	n/a	n/a	n/a	Ectopic recombination
Rat 72	chr3: 140472427	Unknown	Not identified	n/a	n/a	n/a	n/a	n/a	Ectopic recombination (most likely acquired with Rat 52)

Cluster	Coordinates of cluster*	Coordinates of insertion	Coordinates of source paralog	Coordinates of source paralog	Known cluster(s) in source paralog	Coordinates of stream RE	Coordinates of up-stream RE [†]	Copies of upstream RE [†]	Coordinates of stream RE	Copies of downstream stream RE	Insertion mechanism
Rat 99	chr17: 20354000 - 20367000 (13001)	chr17: 20352501 - 20372926 (20426)	Not identified	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Ectopic recombination

* All coordinates are given according to the UCSC Genome Browser (mouse build mm9, rat build rn4). Lengths of the described segments are shown in parentheses.

[†]The total number of paralogs in the same genome, the number of these that reside on different chromosomes, and the number of paralogs in the other rodent genome are shown. These figures do not include repeats flanking the source paralog.

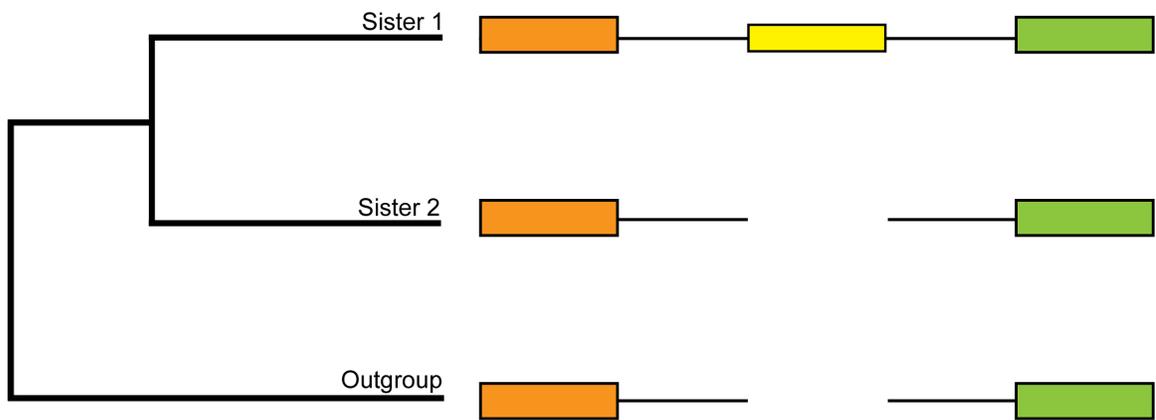


Figure 3.1: Acquisition of a cluster-harboring sequence. Alignment of a cluster-harboring segment (yellow) in sister 1 to a gap in sister 2 and in an outgroup indicates that this segment was acquired in the lineage of sister 1 after it diverged from the lineage of sister 2. Flanking protein-coding genes are depicted in orange and green.

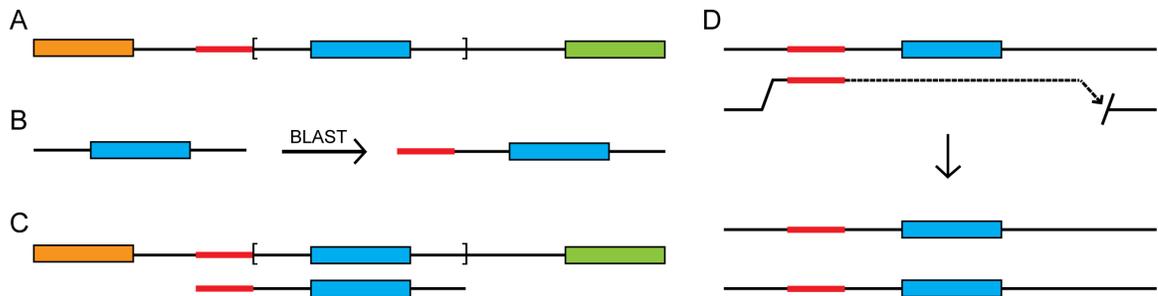


Figure 3.2: Schematic used to identify ectopic recombination as the mechanism of an insertion. (A) Architecture of a typical cluster-harboring genomic region. Two protein-coding genes (orange and green) flank an intergenic region containing an acquired cluster (blue) that is preceded by a RE (red). The inserted segment is depicted within brackets. (B) The inserted segment (Left) is scanned against the genome to locate the source paralog, which is depicted within brackets in its genomic context. The source paralog harbors a paralogous cluster (blue) and is preceded by a RE (Right). (C) Alignment of the cluster- and source paralog-harboring regions indicates that similarity between the 2 sequences includes the REs preceding paralogous clusters. (D) The most likely insertion mechanism is ectopic recombination. Following a double-stranded break (Bottom strand), recombination occurs between homologous REs preceding the source paralog (Top strand) and the site of the double-stranded break. Extension and reannealing of the broken strand generates a new cluster-harboring segment.

CHAPTER IV

Bridges: a tool for identifying local similarities in long sequences

4.1 Introduction

Identifying homologous genomic segments is fundamental to tackling a number of biological problems, including mapping functional elements, predicting protein structures, quantifying molecular evolutionary dynamics and establishing phylogenetic relationships. Homologous segments can be located with high accuracy by employing the SmithWaterman approach, a dynamic programming algorithm (*Smith and Waterman*, 1981). However, because it entails examining every possible alignment, the SmithWaterman approach is computationally intensive and time-consuming, rendering its use unrealistic for many large-scale projects (*Altschul et al.*, 1990).

In the past quarter century, several heuristic tools have been developed to rapidly locate homologous segments (*Pearson and Lipman*, 1988; *Altschul et al.*, 1990; *Kent*, 2002). Rather than traversing sequences base-by-base, such programs limit their focus to regions with short exact word matches. Though this means that sensitivity is lower when searching for distantly related similarities, heuristic approaches are orders of magnitude faster than the SmithWaterman approach and have low computational costs associated with them (*Altschul et al.*, 1990). For these reasons, such tools have

become an invaluable resource for biologists and form the backbone of bioinformatics.

Recently, we were faced with the task of identifying pairs of unique paralogous segments in the *Drosophila melanogaster* genome. We encountered a number of obstacles when attempting to use the entire genome as both a query and database with currently available search tools. Thus, we developed Bridges, which can perform rapid memory-efficient heuristic searches on genome-scale datasets. Another asset of Bridges is that it is highly flexible, with 20 parameters that enable the user to tailor a search to his or her particular goals.

4.2 Implementation

Bridges requires two files as input: a database sequence file and a query sequence file. The query file can contain either a single query or a list of queries in FASTA format. Additionally, the user can modify 27 parameters, 20 of which influence the results produced by the program. The output file lists parameters used, coordinates and alignment scores of similarities and, optionally, corresponding sequences.

When multiple queries are specified, each query is individually compared to the database. Thus, similar to other heuristic programs, the user can compare several sequences to a database in a single run. The user can also choose to look for similarities on the direct strand, reverse-complemented strand or both. Further, there is the option to ignore similarities residing on the diagonal of the alignment matrix, which is useful when the same sequence is being used as the query and database.

Bridges can be used in the same capacity as other heuristic search tools, such as FASTA (*Pearson and Lipman, 1988*) and BLAST (*Altschul et al., 1990*). Its algorithm is similar to those employed by such programs and can be split into the following three stages, the third of which is optional:

1. filtering input sequences;

2. identifying local similarities; and
3. post-processing local similarities.

4.2.1 Filtering input sequences

In this stage, Bridges masks low-complexity regions of the database and query sequences. Strictness of filtering can be adjusted via four parameters, though it is important to note that lax parameters may increase the runtime of the next stage if input sequences are highly repetitive. All *Ns* are automatically masked, but the user can decide whether to filter lowercase letters already present in sequences from previous masking. The user can also choose the word size used for masking, as well as the maximum frequency of a word in both query and database sequences.

Filtering is accomplished in two steps. First, a lookup table of all words and their frequencies in a sequence is constructed. Then, all words that occur more than allowed are masked. The output file specifies the fraction of each sequence that was masked. Bridges also separately outputs filtered sequences, with masked characters in lowercase.

4.2.2 Identifying local similarities

As with other heuristic search tools, this stage is performed by examining regions containing exact word matches between the database and query sequences. Word length is given as a parameter, allowing the user to control the sensitivity of a search. However, choice of word length is also critical to runtime and memory usage. While decreasing word length increases search sensitivity, it also significantly increases runtime and memory requirements. Thus, one should only use short words (<10 nt) when looking for weak similarities. Additional parameters are maximum distance between words, mismatch and gap penalties, and the minimum score for a local similarity.

Identification of local similarities begins with the construction of a lookup table of

all words in the database sequence. Next, the query sequence is scanned, and positions for all words it has in common with the database are recorded. Consecutive word matches are then linked, forming long chains of exact matches. Bridges compares all pairs of chains, temporarily linking them if the distance between them is less than or equal to the maximum distance specified. The alignment score is calculated by subtracting the multiple of the gap length and the gap penalty from the number of exact matches. All possible linked and unlinked similarities are scored, and the highest scoring configurations are kept. Resulting local similarities with scores greater than or equal to the minimum score undergo post-processing if the user elects this option. Otherwise, these similarities are sent to the output file.

4.2.3 Post-processing local similarities

Though optional, post-processing includes two unique features that can be exploited for specialized project goals. One is the removal of local similarities that occur at a lower or higher copy number than desired by the user. For example, one may want to look for similarities that occur at least three times and a maximum of five times. In our case, since we sought only pairs of paralogs, we set both the minimum and maximum number of similarities to two. Other heuristic search tools report similarities of all copy numbers, which would have required us to filter the results accordingly.

The second feature is merging neighboring local similarities. Here, the user specifies the maximum distance between similarities for merging. This is useful when one is looking for long regions of homology or similarities that may include large insertions or deletions. For example, this feature was valuable to us since our goal was to obtain full-length paralogs, and would be similarly advantageous to someone attempting to locate orthologs.

The first step of post-processing is calculating the coverages of each sequence.

Similarities residing within or within a fraction of (another parameter) low- or high-coverage regions designated by the user are removed. Next, each pair of remaining similarities is evaluated. If members of a pair reside along the same diagonal (*i.e.*, there are no gaps within their alignment), they are linked if their distance is less than or equal to the maximum distance. Otherwise, gap length, a penalty chosen by the user, and maximum distance are used together to determine whether they should be linked. Once all comparisons are completed, finished similarities are sent to the output file.

4.3 Discussion

Bridges is a fast and efficient search tool for identifying homologous segments between long sequences. In a single run of Bridges on a Linux machine, we were able to compare the entire *D. melanogaster* genome to itself and specifically locate paralogous pairs. This took <2 h and used a maximum of 1.4 GB of memory during the entire run. Some important parameters used were a masking word size of 13, a searching word size of 12 and a minimum score of 100. While BLAST took approximately the same amount of time and memory to run, it produced shorter similarities, including those along the diagonal of the alignment. BLAT, with default parameters, ran for over 2 weeks before reaching the upper limit of 32 GB of memory and crashing. Thus, using either of these programs would have required us to filter and stitch together local similarities or to split up the genome substantially.

An added strength of Bridges lies in its ability to be guided by the user via an array of parameters. This flexibility allows the user to control the sensitivity and specificity of a search. For our purposes, having a parameter-rich program to work with was invaluable in that we were able to specifically locate the types of similarities we were interested in. Such flexibility can also be problematic if the user does not know what parameter values will produce the desired output. However, because Bridges produces

output rapidly, the user is free to experiment with several sets of parameters. In fact, we found it helpful to examine different types of output to better understand how modification of certain parameters affected our results. For example, increasing the word size or the minimum score resulted in fewer, but stronger, similarities. Thus, the ability of the user to experiment with parameters is, in itself, yet another strong asset to Bridges.

4.4 Acknowledgements

Bridges is named after Calvin B. Bridges, who described the first pair of paralogous genomic segments (*Bridges*, 1936). This work was supported by a University of Michigan Rackham Merit Fellowship.

CHAPTER V

A strong deletion bias in nonallelic gene conversion

5.1 Introduction

Every genome contains similar DNA segments. In diploids, such segments can be classified as orthologs or paralogs. Orthologs, or allelic segments, are paired copies located at the same genomic loci on maternal and paternal chromosomes. In contrast, paralogs, or nonallelic segments, are found at different genomic loci and can have any copy number, in which each copy is derived from an ancestral sequence via gene duplication (*Koonin, 2005*).

Related sequence segments can diverge from one another via ordinary mutation or converge via gene conversion. Ordinary mutation is generally AT-biased for nucleotide substitutions (*Gojobori et al., 1982; Alvarez-Valin et al., 2002; Echols et al., 2002*) and deletion-biased for length difference mutations (*Petrov, 2002*). A number of studies have examined nucleotide substitutions caused by allelic and nonallelic gene conversion, many of which have uncovered a GC-bias (*Marias, 2003; Mancera et al., 2008; Liu and Li, 2008; Berglund et al., 2009*). Here, we explore length difference mutations produced by nonallelic gene conversion.

In contrast to orthologs, paralogs have their own independent long-term phylogenies, making it possible to apply a direct phylogenetic approach to study their coevolution by gene conversion (Figure 5.1). For this approach, we utilized multiple

alignments of pairs of paralogs in three closely-related species: two sisters and an outgroup. First, we ascertained all cases in which, at a particular alignment position, there was an ancestral length difference between the paralogs, *i.e.*, the difference was present in one sister and in the outgroup. We then examined orthologous positions in the other sister and identified those cases for which there was no length difference between paralogs. Elimination of a length difference was due to an insertion in the lineage of that sister if one paralog acquired an additional nucleotide(s) at that position, and was due to a deletion if it lost a nucleotide(s) at that position. If the event resulted in the paralogs having identical states at the affected position, it was consistent with gene conversion.

5.2 Results and Discussion

Since our approach required that paralogs be present in triplets of closely-related species, we chose to analyze gene conversion events in *Drosophila* and primate lineages, for which whole-genome sequences of multiple close species are available. For *Drosophila*, we used *D. melanogaster* and *D. simulans* as sister species and *D. yakuba* as an outgroup and, for primates, we used human and chimpanzee as sisters and orangutan as an outgroup. We obtained 338 and 10,449 pairs of paralogs that were present in all three species of *Drosophila* and primates, respectively (Figure 5.2).

Within these paralogs, we identified 179 insertions and 614 deletions consistent with gene conversion in *Drosophila*, and 132 insertions and 455 deletions consistent with gene conversion in primates (Figure 5.3a). Thus, there were 3.4 times as many deletions as insertions in both lineages, which was highly significant ($p < 0.0001$). The deletion bias was similar for intra- and inter-chromosomal paralogs (data not shown). Moreover, in primates, we found that this deletion bias was considerably stronger for large indels (Figure 5.3b). Because only a very small fraction of these indels could be due to either ordinary mutation or sequencing errors (see Methods),

these biases were primarily due to nonallelic gene conversion.

To determine how this deletion bias affects paralog evolution, we estimated the rate of nonallelic gene conversion in each lineage. For primates, we performed a simple calculation. There were 28,701 sites for which there was an ancestral length difference between paralogs. Conversion-consistent indels occurred at 587 of these sites, resulting in 0.02 indels per site. Because the K_s between human and chimp is ~ 0.0123 (*Chen and Li, 2001*), the length of each sister lineage is $\sim 0.00615 K_s$ units. Thus, the per-site rate of gene conversion in primates is ~ 3.4 times higher than that of ordinary substitution mutation.

For *Drosophila*, a more complex estimate was needed. There were 793 conversion-consistent indels that occurred at 960 possible sites, resulting in 0.83 indels per site. Due to this high proportion, it was necessary to correct for multiple conversion events per site. If we assume that gene conversion is a Poisson process, like mutation, the mean number of events per site is $-\ln(1 - 0.83)$, or ~ 1.8 . The K_s between *D. melanogaster* and *D. simulans* is ~ 0.12 (*Heger and Ponting, 2007*), which implies that the per-site gene conversion rate is 30 times higher than that of ordinary substitution mutation in *Drosophila*.

Thus, nonallelic gene conversion is a rapid deletion-biased force acting on *Drosophila* and primate paralogs. Ordinary mutation is also deletion-biased for small indels (< 400 nt) in both of these lineages, with a 9:1 ratio of deletions to insertions in *Drosophila* and a 5:1 ratio in primates. In the absence of selection, this mutation bias leads to a rapid reduction in genome size, which can only be counterbalanced by large insertions (*Petrov, 2002*).

Deletion-biased gene conversion has an analogous, but distinct, effect on genome size evolution. To illustrate this, let us consider the life cycle of a length difference mutation within two paralogs. First, ordinary mutation introduces an insertion or deletion in one paralog. Then, deletion-biased gene conversion occurs between the

paralogs. If the initial mutation was an insertion, it is removed. Otherwise, the deletion is transmitted to the second paralog, *i.e.*, fixed within the pair of paralogs. In the absence of selection, this results in cooperative shrinkage of these paralogous sequence segments.

Cooperative shrinkage of paralogs can be quantified by phylogenetic detection of fixed conversion-induced indels (Figure 5.4). To do this, we ascertained all cases for which, ancestrally, two paralogs had identical lengths at a particular site and, in one sister, they acquired matching indels at that position. This condition implies that, in the ancestral lineage of the sister, ordinary mutation produced an indel in one paralog, and that this indel was later copied to the other paralog (fixed) by gene conversion. In *Drosophila*, we detected 74 fixed insertions, with a total inserted sequence length of 391 nt, and 176 fixed deletions, with a total deleted sequence length of 1,660 nt. In primates, we found four fixed insertions, with a total inserted sequence length of 4 nt, and 24 fixed deletions, with a total deleted sequence length of 438 nt. Thus, in both lineages, fixed deletions were much longer and more frequent than fixed insertions. Subtracting total insertion lengths from total deletion lengths, we arrived at effective deletion lengths of 1,269 nt in *Drosophila* and 434 nt in primates. The total sequence length of all paralogs was 208,956 nt in *Drosophila* and 5,003,429 nt in primates. Therefore, the shrinkage rate of paralogs by gene conversion was ~ 0.103 per K_s unit in *Drosophila* and ~ 0.014 per K_s unit in primates. This implies that, in the absence of selection, these paralogs would shrink exponentially and disappear in $\sim 138 K_s$ units in *Drosophila* and $\sim 1,096 K_s$ units in primates.

5.3 Methods

Whole-genome sequences of *Drosophila melanogaster*, *Drosophila simulans*, *Drosophila yakuba*, *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), and *Pongo pygmaeus* (orangutan) were downloaded from the UCSC Genome Bioinformatics site

at <http://genome.ucsc.edu>. We used Mega BLAST (*Zhang et al.*, 2000) (default parameters) and Bridges (*Kondrashov and Assis*, 2010) (KM = 13, FilterDBase = 20, FilterQuery = 20, KS = 12, CoeffMis = 0.01, CoeffGap = 0.05, FlatGap = 10, MaxDist = 50, MinWeight = 100, CoeffMisPost = 0.1, MaxDistPost = 1000) to locate pairs of similar sequence segments in the genomes of *D. melanogaster* and *H. sapiens*. To avoid short repeats, we required that each sequence was at least 100 nt long. After examining the output from these methods, we set a cutoff of 78% sequence identity between paralogs. If paralogs were located on the same chromosome, we required that they were separated by at least 100 nt to avoid sequencing or genome mapping errors. We used the BLASTN (*Altschul et al.*, 1990) (default parameters) and Mega BLAST (default parameters) algorithms to locate orthologs for each paralog in the sister and outgroup species, attaching 500 nt flanks to the ends of each paralog to ensure that they were assigned correctly (*i.e.*, within the correct genomic context). If both paralogs were identified in the three species of a lineage, we performed a multiple alignment of all sequences with MUSCLE (*Edgar*, 2004).

To assess the statistical significance of each deletion bias, we used a binomial sign test, which is an exact probability test that assumes that two categories are equally likely to occur. In our case, the two categories were insertions and deletions, and the number of trials (n) was the total number of indels observed. For each test, we used $\alpha = 0.05$ and reported two-tailed p values.

5.3.1 Estimation of the proportion of gene conversion-consistent indels attributed to ordinary mutation

It was important to assess the likelihood that some fraction of indels consistent with gene conversion were actually produced by ordinary mutations. To estimate the proportion of these false positives in our data, we calculated the probability of observing ordinary mutations resembling gene conversion indels in each lineage. We

considered the simplest case, in which there was an existing indel at a particular site in one paralog, and a second, identical, indel is produced by ordinary mutation at the same site in the second paralog. For simplicity, we assumed that both indels had a length of 1 nt. We denoted the probability that ordinary mutation causes the second indel as p_{indel} , the probability that the second indel occurs at the same site as the first as p_{site} , and the probability that the second indel is of the same type (insertion or deletion) as the first as p_{type} . Given that the first indel was already present, the probability of arriving at this configuration is $p_{conf} = p_{indel} \times p_{site} \times p_{type}$.

Assuming that insertions and deletions do not occur at the ends of sequences, as in our empirical analysis, and denoting the length of the second paralog as ℓ , the probability that the second indel occurs at the same site as the first is $p_{site} = 1/(\ell - 1)$ if the first indel was an insertion and $p_{site} = 1/(\ell - 2)$ if it was a deletion.

We used $p_{indel} = 0.012$ for *Drosophila* (Begun *et al.*, 2007b), and $p_{indel} = 0.00196$ for primates (*The Chimpanzee Sequencing and Analysis Consortium*, 2005). For both lineages, we used $\ell = 100$ nt, which was the minimum paralog length in our dataset. There were 18 insertions and 184 deletions consistent with ordinary mutation in *Drosophila*, and 533 insertions and 562 consistent with ordinary mutation in primates. Thus, we split p_{type} into p_i for insertions and p_d for deletions and used $p_i = 0.09$ and $p_d = 0.91$ for *Drosophila* and $p_i = 0.49$ and $p_d = 0.51$ for primates.

In *Drosophila*, p_{conf} is $\sim 1.09 \times 10^{-5}$ if the first indel was an insertion, and p_{conf} is $\sim 1.11 \times 10^{-4}$ if it was a deletion. In primates, p_{conf} is $\sim 9.7 \times 10^{-6}$ if the first indel was an insertion and p_{conf} is $\sim 1.02 \times 10^{-5}$ if it was a deletion. To estimate the number of such configurations in our *Drosophila* and primate datasets, we split each into $N = T/\ell$ segments of length ℓ , where T is the length of the target sequence on which the second indel mutations can occur (half of the total length of paralogs). Assuming that pairs of paralogs are independent from other pairs of paralogs, the probability that there are m mutations that fit the desired configuration in our dataset is binomially

distributed. Thus, the expected number of such mutations is Np_{conf} . Applying this to our *Drosophila* dataset ($T = 104,478$), we expect to observe ~ 0.01 cases in which the first indel was an insertion and ~ 0.12 cases in which it was a deletion. In primates ($T = 2,458,412$), these expectations were ~ 0.24 and ~ 0.25 when the first indel was an insertion or a deletion, respectively. Given that the probability that we do not observe any such configurations in a dataset is $(1 - p_{conf})^N$, the probability that there is at least one is $1 - (1 - p_{conf})^N$. In *Drosophila*, this probability is ~ 0.01 for cases in which the first indel was an insertion and ~ 0.11 when it was a deletion and, in primates, it is ~ 0.21 for cases in which the first indel was an insertion and ~ 0.22 when it was a deletion. Thus, for this simple scenario, it is highly unlikely that any conversion-consistent configurations were produced by ordinary mutation in either dataset. Moreover, in both datasets, a substantial proportion of paralogs (Figure 5.2a) and indel events (Figure 5.3a) were larger than the assumptions made here. Hence, the probabilities of such configurations are even smaller than these estimates in our empirical datasets.

5.3.2 Ascertainment of the effect of sequencing errors on our observations

In addition to ordinary mutation, some proportion of gene conversion-consistent indels may be artifacts of sequencing errors. To ensure that sequencing errors were not responsible for observed deletion biases, we calculated the proportions of insertions and deletions consistent with sequencing errors for each lineage. Because all observed indels can be due to sequencing errors, we conservatively assumed that the number of sequencing errors of each type (insertion or deletion) was the sum of conversion-consistent and ordinary mutation-consistent events of that type.

In *Drosophila*, we observed 197 insertions (179 conversions and 18 ordinary mutations) and 798 deletions (614 conversions and 184 ordinary mutations) that could be attributed to sequencing errors. In primates, there were 665 insertions (132 con-

versions and 533 ordinary mutations) and 1,017 deletions (455 conversions and 562 ordinary mutations) that could be due to sequencing errors. Because sequencing errors, like ordinary mutations, can occur at any site, the target size for such errors was 104,478 nt for *Drosophila* and 2,458,418 nt for primates (see above).

Dividing the sum of each type of indel by its target size yielded sequencing error-consistent insertion proportions of $\sim 1.89 \times 10^{-3}$ for *Drosophila* and $\sim 2.7 \times 10^{-4}$ for primates, and deletion proportions of $\sim 7.64 \times 10^{-3}$ for *Drosophila* and $\sim 4.14 \times 10^{-4}$ for primates. These small proportions imply that length difference sequencing errors are rare in our datasets.

For comparison, we also calculated the proportions of insertions and deletions consistent with gene conversion in each lineage. Because indels produced by gene conversion can only occur at sites in which there is an ancestral sequence difference, the target sizes for such events were 960 nt for *Drosophila* and 28,701 nt for primates. Given these targets, conversion-consistent insertion proportions are ~ 0.19 in *Drosophila* and $\sim 4.6 \times 10^{-3}$ in primates, and deletion proportions are ~ 0.64 for *Drosophila* and ~ 0.02 for primates. These proportions are all much higher than those of sequencing errors; therefore, only a very small proportion of observed indels can be artifacts of sequencing errors.

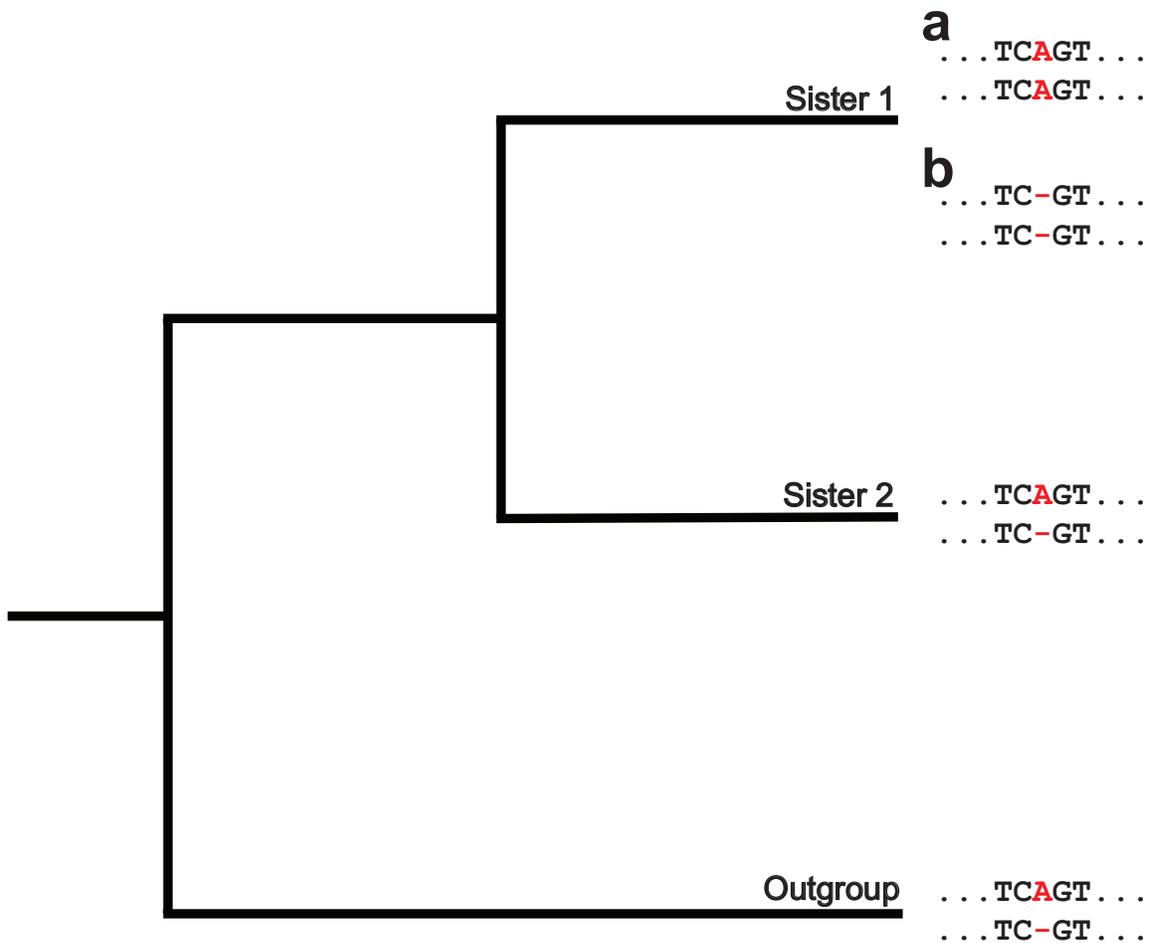


Figure 5.1: A phylogenetic approach for detecting insertions and deletions in nonallelic gene conversion. Depicted is a hypothetical multiple alignment for pairs of paralogs in two sisters and an outgroup. The two sequences for each species represent a pair of paralogs, and the position of interest is colored in red. At this position, there is a length difference (A/-) between the paralogs in sister 2 and the outgroup (ancestral state). In the lineage of sister 1, an (a) insertion or (b) deletion of a nucleotide occurs at this position in one paralog. Because either the insertion or deletion event results in the paralogs having matching states at this position (A/A or -/-), they are both consistent with gene conversion.

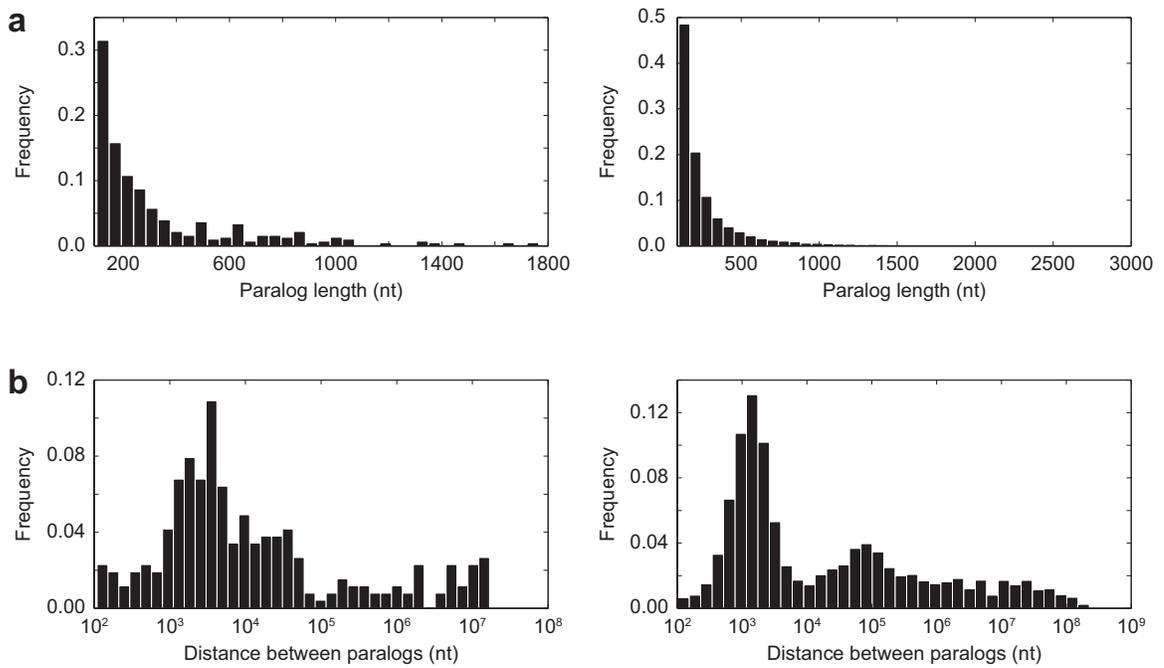


Figure 5.2: Properties of paralogs. (a) Distribution of paralog sequence lengths in *Drosophila* (left) and primates (right). (b) Distribution of distances between paralogs located on the same chromosome in *Drosophila* (left; represents 79% of paralogs) and primates (right; represents 59% of paralogs). Distances are plotted on a log scale.

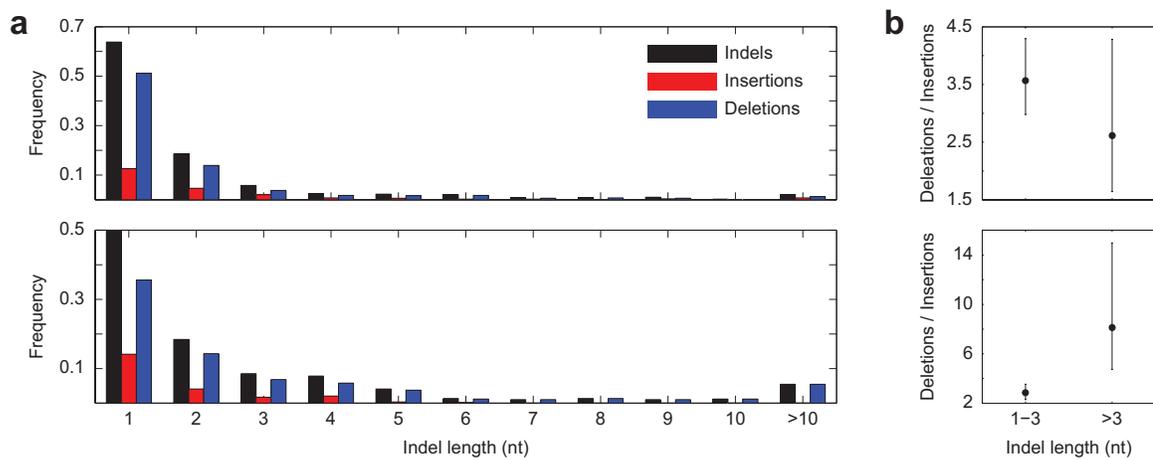


Figure 5.3: Indels consistent with gene conversion. (a) Length distributions of all indels, insertions, and deletions in *Drosophila* (top) and primates (bottom). (b) Strength of deletion bias as a function of indel length in *Drosophila* (top) and primates (bottom). Error bars represent confidence limits from binomial sign tests (see Methods).

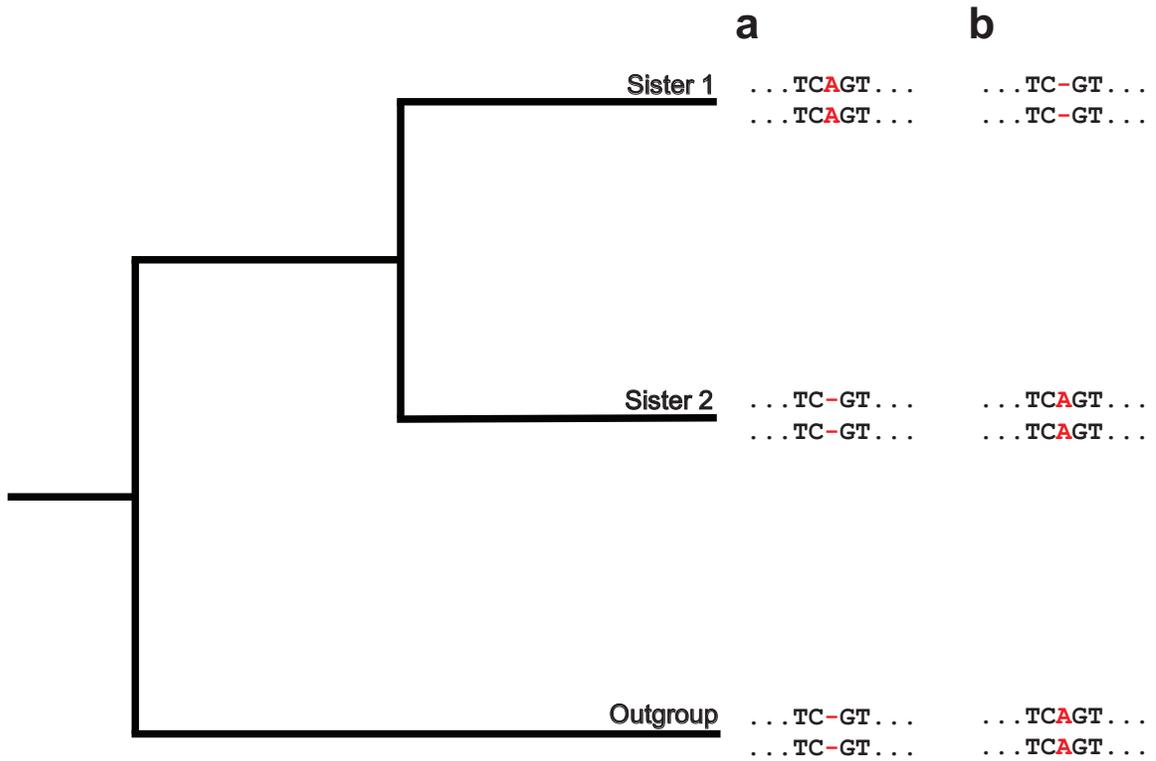


Figure 5.4: A phylogenetic approach for detecting fixed indels. Depicted are hypothetical multiple alignments for pairs of paralogs in two sisters and an outgroup. The two sequences for each species represent a pair of paralogs, and the position of interest is colored in red. (a and b) At this position, both paralogs have identical lengths in sister 2 and the outgroup (ancestral state). In the lineage of sister 1, matching (a) insertions or (b) deletions occur at this position in the paralogs. Each of these situations corresponds to an ordinary mutation producing an indel in one paralog, and this indel being transferred to the other paralog, or fixed, by gene conversion.

CHAPTER VI

Conclusion

In this dissertation, I applied a direct phylogenetic approach to investigate various questions pertaining to the origin and evolution of novel DNA sequences. My studies of the origins of nested genes and piRNAs demonstrated the dominant role of gene duplication in acquisition of novel functional sequences. Then, by exploring the evolution of paralogs after gene duplication, I found that gene conversion has a strong influence on the evolutionary fates of paralogs.

In Chapter II, I discovered that nested genes were gained more frequently than they were lost in animal genomes. I elucidated the primary mechanism of formation of most nested gene structures as the insertion of nested genes into the introns of their host genes. Most nested genes had ancestral copies in other genomic regions, indicating that they arose via some form of gene duplication. Sequence analyses of nested and ancestral genes revealed that most nested genes arose via either standard gene duplication or retrotransposition. Moreover, comparisons of tissue-specific expression correlations between nested and un-nested pairs of genes uncovered no evidence of positive selection favoring nested gene structures. Examination of functional annotations and sequence characteristics of nested, host, and un-nested genes supported this finding: Nested genes tend to be short and intronless, host genes tend to be long and have multiple introns, and un-nested genes tend to be intermediate in both

length and number of introns. Thus, I concluded that nested genes arose via a neutral process caused by the presence of many long unsaturated introns in animal genomes. This implies that nested gene structures are far from equilibrium and will continue to arise at a high rate, leading to an increase in the organizational complexity of animal genomes over evolutionary time.

In Chapter III, I found that mammalian piRNA clusters arose at a higher rate than any known gene family, with not a single loss of a cluster observed. This rapid acquisition of piRNA clusters occurred by duplication and insertion of long DNA segments. I was able to locate the ancestral sequences of many recently acquired clusters, most of which are clusters themselves. Interestingly, I also discovered that many of these clusters are part of large paralogous families. Often, ancestral and derived clusters are located on the same chromosomes and flanked by identical chromosome-specific repetitive elements (REs), which caused genomic instability and large rearrangements in these regions. Upon further investigation, I uncovered the primary molecular mechanism of cluster acquisition to be ectopic recombination between flanking REs, which led to duplications and insertions of adjacent piRNA clusters. Thus, it appears that new clusters propagated from old clusters on the same chromosomes at an extremely high rate caused by the presence of flanking REs. This high rate of cluster acquisition and lack of losses suggests that the expansion of piRNA cluster families was likely due to positive selection driven by an arms race between piRNA clusters and the rapidly growing families of transposable elements they silence.

Chapter IV described Bridges, a heuristic tool developed to locate similar sequence segments within and between genomes. Bridges proceeds in three stages, the third of which is optional: filtering input sequences, identifying local similarities, and post-processing of local similarities. The main advantage of this tool over other heuristic tools, such as BLAST (*Altschul et al.*, 1990) and BLAT (*Kent*, 2002), is that it is highly customizable due to the abundance and flexibility of parameters. For example,

in Chapter V, I used Bridges to locate paralogs that were unique, longer than 100 nt, and only present in pairs. In comparison, BLAST returned much shorter sequences that had to be stitched together and filtered for unique pairs, and BLAT ran out of memory even when searching *Drosophila* genomes. Due to its flexibility, Bridges can be applied to locate many different types of similar sequences. Thus, researchers interested in locating similar sequence segments with specific characteristics will benefit from this tool.

In Chapter V, I used a direct phylogenetic approach to ascertain insertions and deletions produced by gene conversion between pairs of paralogs in *Drosophila* and primate lineages. Interestingly, I discovered that gene conversion between paralogs is strongly deletion-biased in both lineages. Calculation of the per-site rates of gene conversion revealed that gene conversion occurs at a much higher rate than ordinary mutation in both lineages, with the relative rate being ten times higher in *Drosophila* than in primate genomes. Further investigation revealed that this high rate, coupled with the observed deletion bias, causes the rapid fixation of deletion mutations within a pair of paralogs, leading to the cooperative shrinkage and eventual disappearance of pairs of neutrally evolving paralogs. Hence, this study showed that evolution of paralogs by gene conversion alone results in their rapid decay and removal from genomes. Therefore, maintenance of paralogs over long evolutionary time periods is likely due to the action of other evolutionary forces. Moreover, if the overall rate of gene acquisition is comparable to those observed in nested genes and piRNA clusters, the disappearance of current paralogs will be counteracted by the rapid emergence of new paralogs via gene duplication.

This dissertation investigated two interrelated topics: the origin of novel sequences, and the evolution of these sequences after their emergence. Chapters II and III explored the origin of novel sequences, with Chapter II focusing specifically on nested genes, and Chapter III on piRNA clusters. Strikingly, there were many

similarities between the origins of nested genes and piRNA clusters. Both classes of novel sequences were acquired at a high rate, mostly derived by gene duplication, and rarely or never lost during evolution. The main difference between these two sequence classes was that nested genes likely arose and were maintained via a neutral evolutionary process, whereas piRNA clusters may have expanded because of strong positive selection. Using a tool developed in Chapter IV, Chapter V examined the evolution of novel sequences after their emergence. The focus of this analysis was on the evolution of ancient paralogs by gene conversion-induced indels. Results from this study demonstrated the powerful role that gene conversion plays in paralog evolution and provide insight into the forces driving the long-term maintenance of paralogs.

Findings from this dissertation advance the field of evolutionary genetics by providing insight into how new genes arise in the genome, as well as how they evolve long after their emergence. In addition, my results raise several novel questions pertaining to gene duplication. For example, how does the evolution of noncoding nested genes, as well as of partially overlapping genes, compare to that of protein-coding nested genes? Also, is there a general pattern of rapid gains in long functional sequences in the genome? More generally, what evolutionary forces act to maintain paralogs in the genome? This dissertation demonstrated a role for both neutral processes and positive selection, and it would be of great interest to determine how frequent each is, as well as the factors that contribute to each. Nested genes may be maintained via neutral processes simply because they are located in introns. Do sequences inserted in other noncoding locations also typically evolve neutrally? If not, the maintenance of such sequences, even if they are not protein-coding, may indicate their functional importance in the genome. Moreover, it would be interesting to determine whether nonallelic gene conversion is deletion-biased in all eukaryotes, as well as to compare the strength of gene conversion to that of selection and explore how all evolutionary forces act together to determine the fate of paralogs after duplication.

Beyond the scope of this dissertation, many questions about gene origin remain unanswered. In particular, do genes created by different mechanisms take distinct evolutionary paths? Because the probability of a functional gene arising through random sequence changes of a nonfunctional DNA sequence is almost zero (*Jacob, 1977*), it would be particularly intriguing to study the evolutionary paths of *de novo* genes. For example, how are coding and regulatory regions constructed? Does positive selection play a role in this process? If so, how early must it act and how strong must it be to create a functional gene out of a noncoding sequence? With respect to evolution after gene duplication, this dissertation only focused on gene conversion, which is one of several forces that act on paralogs. Though many studies have examined the different forces acting on paralogs, only a few have investigated the interaction among these forces and their strengths and influences on evolutionary outcomes of paralogs (*Moore and Purugganan, 2003; Kondrashov and Kondrashov, 2006; Teshima and Innan, 2008*). Furthermore, it would be of great interest to determine the frequencies of different phenotypic outcomes and elucidate the evolutionary paths of each.

Studying the origin and evolution of duplicate genes is crucial to the understanding of the evolution of phenotypic novelty and divergence of species. My dissertation research focused on a few out of potentially millions of exciting questions about gene duplication. Here I have highlighted what I believe are some of the most relevant to the advancement of the field of evolutionary genetics. Perhaps the most interesting of these questions pertain to the correlation between sequence and functional changes of paralogs. The recent availability of many inter- and intra-species genome sequences and development of genome-scale experimental methods will enable researchers to bridge this gap between genomic and phenotypic evolution, shedding light on the origin of novel phenotypes and driving the field of evolutionary genetics forward.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990), Basic local alignment search tool, *J. Mol. Biol.*, *215*, 403–410.
- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, *25*, 3389–3402.
- Alvarez-Valin, F., G. Lamollea, and G. Bernardi (2002), Isochores, GC3 and mutation biases in the human genome, *Gene*, *300*, 161–168.
- Anderson, J. O., W. F. Doolittle, and C. L. Nesbo (2001), Genomics: Are there bugs in our genome?, *Science*, *292*, 1848–1850.
- Anderson, J. O., S. W. Sarchfield, and A. J. Roger (2005), Gene transfers from nanoarchaeota to an ancestor of diplomonads and parabasalids, *Mol. Biol. Evol.*, *22*, 85–90.
- Aravin, A., et al. (2006), A novel class of small RNAs bind to MILI protein in mouse testes, *Nature*, *442*, 203–207.
- Aravin, A. A., G. J. Hannon, and J. Brennecke (2007), The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race, *Science*, *318*, 761–764.
- Bai, Y., C. Casola, C. Feschotte, and E. Betran (2007), Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*, *Genome Biol.*, *8*, R11.
- Begun, D. J., H. A. Lindfors, A. D. Kern, and C. D. Jones (2007a), evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade, *Genetics*, *176*, 1131–1137.
- Begun, D. J., et al. (2007b), Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*, *PLoS Biol.*, *5*, e310.
- Bender, W., M. Akam, F. Karch, P. A. Beachy, and M. t. Peifer (1983), Molecular genetics of the Bithorax complex in *Drosophila melanogaster*, *Science*, *221*, 23–29.
- Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler (2009), GenBank, *Nucleic Acids Res.*, *36*, D25–D30.

- Berglund, J., K. S. Pollard, and M. T. Webster (2009), Hotspots of biased nucleotide substitutions in human genes, *PLoS Biol.*, *7*, e1000,026.
- Berriman, M. e. a. (2005), The genome of the African trypanosome *Trypanosoma brucei*, *Science*, *309*, 416–422.
- Betel, D., R. Sheridan, D. S. Marks, and C. Sander (2007), Computational analysis of mouse piRNA sequence and biogenesis, *PLoS Comp. Biol.*, *3*, 2219–2227.
- Bridges, C. B. (1918), Maroon: A recurrent mutation in *Drosophila*, *Proc. Natl. Acad. Sci. U.S.A.*, *4*, 316–318.
- Bridges, C. B. (1936), The bar ‘gene’ a duplication, *Science*, *83*, 210–211.
- Brookfield, J. F. Y. (1997), Genetic redundancy, *Adv. Genet.*, *36*, 137–155.
- Brosius, J. (1991), Retroposons—seeds of evolution, *Science*, *251*, 753.
- Cai, J., R. Zhao, H. Jiang, and W. Wang (2008), *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*, *Genetics*, *179*, 487–496.
- Carmel, L., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin (2007), Patterns of intron gain and conservation in eukaryotic enes, *BMC Evol. Biol.*, *7*, 192.
- Charlesworth, D., F. L. Liu, and L. Zhang (1998), The evolution of the alcohol dehydrogenase gene family by loss of introns in the plants of the genus *Leavenworthia* (*Brassicaceae*), *Mol. Biol. Evol.*, *15*, 552–559.
- Chen, F. C., and W. H. Li (2001), Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of human and chimpanzees, *Am. J. Hum. Genet.*, *68*, 444–456.
- Chen, S. T., H. C. Cheng, D. A. Barbash, and H. P. Yang (2007), Evolution of hydra, a recently evolved testis-expressed gene with nin alternative first exons in *Drosophila melanogaster*, *PLoS Genet.*, *3*, e107.
- Chintapalli, V. R., J. Wang, and J. A. Dow (2007), Using FlyAtlas to identify *Drosophila melanogaster* models of human disease, *Nat. Genet.*, *39*, 715–720.
- Courseaux, A., and J. L. Nahon (2001), Birth of two chimeric genes in the Hominidae lineage, *Science*, *291*, 1293–1297.
- Crampton, N., W. A. Bonass, J. Kirkham, C. Rivetti, and N. H. Thomson (2006), Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy, *Nucleic Acides Res.*, *34*, 5416–5425.
- Cusack, B. P., and K. H. Wolfe (2006), Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates, *Mol. Biol. Evol.*, *24*, 679–686.

- Da Lage, J.-L., C. Maisonhaute, F. Maczkowiak, and M.-L. Cariou (2003), A nested alpha-amylase gene in *Drosophila ananassae*, *J. Mol. Evol.*, *57*, 355–362.
- Davies, W., R. J. Smith, G. Kelsey, and L. S. Wilkinson (2004), Expression patterns of the novel imprinted genes Nap115 and Peg13 and their non-imprinted host genes in the adult mouse brain, *Gene Expr. Patterns*, *4*, 741–747.
- de Koning, A. P., F. S. Brinkman, S. J. Jones, and P. J. Keeling (2001), Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*, *Mol. Biol. Evol.*, *17*, 1769–1773.
- Derelle, E. t. (2006), Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features, *Proc. Natl. Acad. Sci. USA*, *103*, 11,647–11,652.
- Drosophila* 12 Genomes Consortium (2007), Evolution of genes and genomes on the *Drosophila* phylogeny, *Nature*, *450*, 203–218.
- Drouin, G., and G. A. Dover (1990), Independent gene evolution in the potato actin gene family demonstrated by phylogenetic procedures for resolving gene conversions and phylogeny of angiosperm actin genes, *J. Mol. Evol.*, *31*, 132–150.
- Dykhuizen, D. E., and D. L. Hartl (1980), Selective neutrality of 6pgd allozymes in *Escherichia coli* and the effects of genetic background, *Genetics*, *96*, 801–817.
- Echols, N., P. Harrison, S. Balasubramanian, N. M. Luscombe, P. Bertone, Z. Zhang, and M. Gerstein (2002), Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes, *Nucleic Acids Res.*, *30*, 2515–2523.
- Edgar, R. C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, *32*, 1792–1797.
- ENCODE Project Consortium (2007), Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, *447*, 799–816.
- Ferris, S. D., and G. S. Whitt (1979), Evolution of the differential regulation of duplicate genes after polyploidization, *J. Mol. Evol.*, *12*, 267–317.
- Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait (1999), Preservation of duplicate genes by complementary, degenerative mutations, *Genetics*, *151*, 1531–1545.
- Furia, M., P. P. D’Avino, S. Crispi, D. Atiaco, and L. C. Polito (1993), Dense cluster of genes is located at the ecdysone-regulated 3C puff of *Drosophila melanogaster*, *J. Mol. Biol.*, *231*, 531–538.
- Gilbert, W. (1978), Why gene in pieces?

- Girard, A., R. Sachidanandam, G. J. Hannon, and M. A. Carmell (2006), A germline-specific class of small RNAs binds mammalian Piwi proteins, *Nature*, *442*, 199–202.
- Gojbori, T., W. H. Li, and D. Graur (1982), Patterns of nucleotide substitution in pseudogenes and functional genes, *J. Mol. Evol.*, *18*, 360–369.
- Graubert, T. A., et al. (2007), A high-resolution map of segmental DNA copy number variation in the house mouse, *PLoS Genet.*, *3*, 21–29.
- Grivna, S. T., E. Beyret, Z. Wang, and H. Lin (2006), A novel class of small RNAs in mouse spermatogenic cells, *Genes Dev.*, *20*, 1709–1714.
- Guryev, V., et al. (2008), Distribution and functional impact of DNA copy number variation in rat, *Nat. Genet.*, *40*, 538–545.
- Habib, A. A., J. R. Gulcher, T. Högnason, L. Zheng, and K. Stefánsson (1998), The OMgp gene, a second growth suppressor within the NF1 gene, *Oncogene*, *16*, 1525–1531.
- Hahn, M. W. (2009), Distinguishing among evolutionary models for the maintenance of gene duplicates, *J. Hered.*, *100*, 605–617.
- Haldane, J. B. S. (1933), The part played by recurrent mutation in evolution, *Am. Nat.*, *67*, 5–19.
- Harrison, P. M., and M. Gerstein (2002), Studying genomes through the aeons: protein families, pseudogenes and proteome evolution, *J. Mol. Biol.*, *318*, 1155–1174.
- Hartl, D. L., and D. E. Dykhuizen (1981), Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli*, *Proc. Natl. Acad. Sci. USA*, *78*, 6344–6348.
- Heger, A., and C. Ponting (2007), Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes, *Genome Res.*, *17*, 1837–1849.
- Henikoff, S., and M. K. Eghtedarzadeh (1987), Conserved arrangement of rested genes at the *Drosophila* Gart locus, *Genetics*, *117*, 711–725.
- Hotopp, J. C. D., et al. (2007), Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes, *Science*, *317*, 1753–1756.
- Jacob, F. (1977), Evolution and tinkering, *Science*, *196*, 1161–1166.
- Jaworski, D. M., M. Beem-Miller, G. Lluri, and R. Barrantes-Reynolds (2007), Potential regulatory relationship between the nested gene DDC8 and its host gene tissue inhibitor of metalloproteinase-2, *Physiol. Genomics*, *28*, 168–178.
- Jeffs, P., and M. Ashburner (1991), Processed pseudogenes in *Drosophila*, *Proc. R. Soc. Lond.*, *244*, 151–159.

- Kaessmann, H., N. Vinckenbosch, and M. Long (2009), RNA-based gene duplication: mechanistic and evolutionary insights, *Nat. Rev. Genet.*, *10*, 19–31.
- Kapranov, P., A. T. Willingham, and T. R. Gingeras (2007), Genome-wide transcription and the implications for genomic organization, *Nat. Rev. Genet.*, *8*, 413–423.
- Karolchik, D., et al. (2008), The UCSC Genome Browser Database: update, *Nucleic Acids Res.*, *36*, D773–D779.
- Keeling, P. J., and J. D. Palmer (2008), Horizontal gene transfer in eukaryotic evolution, *Nat. Rev. Genet.*, *9*, 605–618.
- Kent, W. J. (2002), BLAT—the BLAST-like alignment tool, *Genome Res.*, *12*, 656–664.
- Kondrashov, A. S., and R. Assis (2010), Bridges: a tool for identifying local similarities in long sequences, *Bioinformatics*, *26*, 2055–2056.
- Kondrashov, F. A., and A. S. Kondrashov (2006), Role of selection in fixation of gene duplications, *J. Theor. Biol.*, *239*, 141–151.
- Koonin, E. V. (2005), Orthologs, paralogs, and evolutionary genomics, *Annu. Rev. Genet.*, *39*, 309–338.
- Koonin, E. V., K. S. Makarova, and L. Aravind (2001), Horizontal gene transfer in prokaryotes, *Annu. Rev. Microbiol.*, *55*, 709–742.
- Kumar, S., and S. B. Hedges (1998), A molecular timescale for vertebrate evolution, *Nature*, *392*, 917–920.
- Kwan-Wood, G. L., and A. J. Jeffreys (2007), Processes of *de novo* duplication of human α -globin genes, *Proc. Natl. Acad. Sci. U.S.A.*, *104*, 10,950–10,955.
- Lau, N. C., A. G. Seto, J. Kim, S. Kuramochi-Miyagawa, T. Nakano, D. P. Bartel, and R. E. Kingston (2006), Characterization of the piRNA complex from rat testes, *Science*, *313*, 363–367.
- Levin, M. T., C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun (2006), Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression, *Proc. Natl. Acad. Sci. U.S.A.*, *103*, 9935–9939.
- Li, C. Y., et al. (2010), A human-specific *de novo* protein-coding gene associated with human brain functions, *PLoS Comput. Biol.*, *6*, e1000,734.
- Lichten, M., and J. E. Haber (1989), Position effects in ectopic and allelic mitotic recombination in *Saccharomyces cerevisiae*, *Genetics*, *123*, 261–268.
- Liu, G., and H. Li (2008), The correlation between recombination rate and dinucleotide bias in *Drosophila melanogaster*, *J. Mol. Evol.*, *67*, 358–367.

- Long, M., and C. H. Langley (1993), Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*, *Science*, *260*, 91–95.
- Long, M., E. Betran, K. Thornton, and W. Wang (2003), The origin of new genes: Glimpses from the young and old, *Nat. Rev. Genet.*, *4*, 865–875.
- Lu, J., Y. Shen, Q. Wu, S. Kumar, B. He, S. Shi, R. W. Carthew, S. M. Wang, and C. I. Wu (2008), The birth and death of microRNA genes in *Drosophila*, *Nat. Genet.*, *40*, 351–355.
- Lundin, L. (1993), Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse, *Genomics*, *16*, 1–9.
- Lupski, J. R. (2007), Genomic rearrangements and sporadic disease, *Nat. Genet.*, *39*, S43–S47.
- Lupski, J. R., and P. Stankiewicz (2005), Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes, *PLoS Genet.*, *1*, 627–633.
- Lynch, M. (2002a), Intron evolution as a population-genetic process, *Proc. Natl. Acad. Sci. U. S. A.*, *99*, 6118–6123.
- Lynch, M. (2002b), Genomics, gene duplication and evolution, *Science*, *297*, 945–947.
- Lynch, M. (2006), The origins of eukaryotic gene structure, *Mol. Biol. Evol.*, *23*, 450–468.
- Lynch, M. (2007a), *The origins of genome architecture*, Sinauer Associates, Inc.
- Lynch, M. (2007b), The frailty of adaptive hypotheses for the origins of organismal complexity, *Proc. Natl. Acad. Sci. U. S. A.*, *104*, 8597–8604.
- Lynch, M., and J. S. Conery (2003), The origins of genome complexity, *Science*, *302*, 1401–1404.
- Lynch, M., and A. Force (2000), The probability of duplicate gene preservation by subfunctionalization, *Genetics*, *154*, 459–473.
- Makalowska, I., C.-F. Lin, and W. Makalowski (2005), Overlapping genes in vertebrate genomes, *Comput. Biol. Chem.*, *29*, 1–12.
- Mancera, E., R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz (2008), High-resolution mapping of meiotic crossovers and non-crossovers in yeast, *Nature*, *454*, 479–485.
- Marias, G. (2003), Biased gene conversion: implications for genome and sex evolution, *Trends Genet.*, *19*, 330–338.
- Marques, A. C., I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann (2005), Emergence of young human genes after a burst of retroposition in primates, *PLoS Biol.*, *3*, e357.

- Martignetti, J. A., and J. Brosius (1993), *BC200* RNA: a neural RNA polymerase III product encoded by a monomeric Alu element, *Proc. Natl. Acad. Sci. USA*, *90*, 11,563–11,567.
- Mattick, J. S., and I. V. Makunin (2005), Small regulatory RNAs in mammals, *Hum. Mol. Genet.*, *14*, R121–R132.
- Mironov, A. A., J. W. Frickett, and M. S. Gelfand (1999), Frequent alternative splicing of human genes, *Genome Res.*, *9*, 1288–1293.
- Misra, S., et al. (2002), Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review, *Genome Biol.*, *3*, 1–22.
- Moore, R. C., and M. D. Purugganan (2003), The early stages of duplicate gene evolution, *Proc. Natl. Acad. Sci. USA*, *100*, 15,682–15,687.
- Muller, H. J. (1935), The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere, *Genetica*, *17*, 237–252.
- Nadeau, J. H., and D. Sankoff (1997), Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution, *Genetics*, *147*, 1259–1266.
- Nei, M., and A. P. Rooney (2005), Concerted and birth-and-death evolution of multi-gene families, *Annu. Rev. Genet.*, *39*, 121–152.
- Nimura, Y., and M. Nei (2005), Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages, *Gene*, *346*, 23–28.
- Ogurtsov, A. Y., M. A. Roytberg, S. A. Shabalina, and A. S. Kondrashov (2002), OWEN: Aligning long collinear regions of genomes, *Bioinformatics*, *18*, 1703–1704.
- Ohno, S. (1970), *Evolution by gene duplication*, Springer-Verlag, Berlin, Germany.
- Osato, N., Y. Suzuki, K. Ikeo, and T. Gojobori (2007), Transcriptional interferences *in cis* natural antisense transcripts of humans and mice, *Genetics*, *176*, 1299–1306.
- Pearson, W. R., and D. J. Lipman (1988), Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. U.S.A.*, *85*, 2444–2448.
- Petrov, D. A. (2002), Mutational equilibrium model of genome size evolution, *Theor. Popul. Biol.*, *61*, 531–544.
- Postlethwait, J. H., et al. (1998), Vertebrate genome evolution and the zebrafish gene map, *Nat. Genet.*, *18*, 345–349.
- Pruitt, K. D., T. Tatusova, and D. R. Maglott (2007), Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, *35*, D61–D65.

- Shabalina, S. A., and E. V. Koonin (2008), Origins and evolution of eukaryotic RNA interference, *Trends Ecol. Evol.*, *23*, 578–587.
- She, X., Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler (2008), Mouse segmental duplication and copy number variation, *Nat. Genet.*, *40*, 909–914.
- Sidow, A. (1996), Gen(om)e duplications in the evolution of early vertebrates, *Curr. Opin. Genet. Dev.*, *6*, 715–722.
- Slusarski, D. C., C. K. Motsny, and R. Holmgren (1995), Mutations that alter the timing and pattern of *cubitus interruptus* gene expression in *Drosophila melanogaster*.
- Smith, N. G. C., and A. Eyre-Walker (2002), Adaptive protein evolution in *Drosophila*, *Nature*, *415*, 1022–1024.
- Smith, T. F., and M. S. Waterman (1981), Identification of common molecular subsequences, *J. Mol. Biol.*, *147*, 195–197.
- Sontheimer, E. J., and R. W. Carthew (2005), Silence from within: Endogenous siRNAs and miRNAs, *Cell*, *122*, 9–12.
- Stein, L. D., et al. (2003), The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics, *PLoS Biol.*, *1*, E45.
- Stoltzfus, A. (1999), On the possibility of constructive neutral evolution, *J. Mol. Evol.*, *49*, 169–181.
- Su, A. I., et al. (2004), A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc. Natl. Acad. Sci. U.S.A.*, *101*, 6062–6067.
- Tamura, K., S. Subramanian, and S. Kumar (2003), Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks, *Mol. Biol. Evol.*, *21*, 36–44.
- Teshima, K. M., and H. Innan (2008), Neofunctionalization of duplicated genes under the pressure of gene conversion, *Genetics*, *178*, 1385–1398.
- The Chimpanzee Sequencing and Analysis Consortium (2005), Initial sequence of the chimpanzee genome and comparison with the human genome, *Nature*, *437*, 69–87.
- Veeramachaneni, V., W. Makalowski, M. Galdzicki, R. Sood, and I. Makalowska (2004), Mammalian overlapping genes: the comparative perspective, *Genome Res.*, *14*, 280–286.
- Wang, W., J. Zhang, C. Alvarez, A. Llopart, and M. Long (2000), The origin of the *Jingwei* gene in the complex modular structure of its parental gene, *Yellow Emperor* in *Drosophila melanogaster*, *17*, 1294–1301.
- Wang, W., F. G. Brunet, E. Nevo, and M. Long (2002), Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*, *Proc. Natl. Acad. Sci. USA*, *99*, 4448–4453.

- Wang, W., et al. (2006), High rate of chimeric gene origination by retroposition in plant genomes, *Plant Cell*, *18*, 1791–1802.
- Watanabe, T., A. Takeda, T. Tsukiyama, K. Mise, T. Okuno, H. Sasaki, N. Minami, and H. Imai (2006), Identification and characterization of two novel classes of small RNAs in the mouse germline: Retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes, *Genes Dev.*, *20*, 1732–1743.
- Willingham, A. T., et al. (2006), Transcriptional landscape of the human and fly genomes: nonlinear and multifunctional modular model of transcriptomes, *Cold Spring Harb. Symp. Quant. Biol.*, *71*, 101–110.
- Yang, S., et al. (2008), Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*, *PLoS Genet.*, *4*, 78–87.
- Yi, S. V. (2006), Non-adaptive evolution of genome complexity, *Bioessays*, *28*, 979–982.
- Yu, P., D. Ma, and M. Xu (2005), Nested genes in the human genome, *Genomics*, *86*, 414–422.
- Zhang, J. (2003), Evolution by gene duplication: an update, *Trends Ecol. Evol.*, *18*, 292–298.
- Zhang, Z., S. Schwartz, L. Wagner, and W. Miller (2000), A greedy algorithm for aligning DNA sequences, *J. Comput. Biol.*, *7*, 203–214.
- Zhou, Q., and W. Wang (2008), On the origin and evolution of new genes—a genomic and experimental perspective, *J. Genet. Genom.*, *35*, 639–648.
- Zhou, Q., et al. (2008), On the origin of new genes in *Drosophila*, *Genome Res.*, *18*, 1446–1455.