Reconstruction and Analysis of Gene Family Evolution in Mammals

Jin Jun
University of Connecticut, 2010

Gene duplication and loss is a dynamic and ongoing process during evolution and both play a significant role in the rise of variable size gene families originating from a single ancestral gene. Re-creating the evolutionary history of these gene families is an important goal of contemporary comparative genomics as understanding gene family histories can reveal many of the evolutionary forces acting on the lineage in question.

In order to accurately unravel gene family histories, the precise relationship between genes in the gene family must be determined. Relationship can be described in two terms: orthologous and paralagous. Orthologous genes are corresponding copies of an ancestral gene in two descendent genomes. Paralagous genes are multiple copy genes in a given genome related by gene duplication events. Many methods have been proposed for the determination and identification of these relationships between members of large gene families. To date the vast majority of these methods rely upon protein sequence information to determine orthology. However, as the protein sequence is under strong selective constraints for function the relationship between the protein sequences of two members of a gene family would be a reflection of both the function of those genes and the ancestral relationship between them.

In an attempt to separate these two confounding factors this thesis proposes several methods for utilizing non-coding sequence information to determine ancestral relationship between members of a gene family. This approach has several advantages including independence from selective influences on the protein coding region, freedom from computationally intensive multiple alignment methods, and the ability to incorporate pseudogenes as explicit markers of gene loss in gene family histories.

Reconstruction and Analysis of Gene Family Evolution in Mammals

Jin Jun

B.S., Korea University, Seoul, Korea, 1992
Ph.D., Korea University, Seoul, Korea, 1999

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2010

APPROVAL PAGE

Doctor of Philosophy Dissertation

Reconstruction and Analysis of Gene Family Evolution in Mammals

Presented by

Jin Jun

Major Advisor: _____

Craig E. Nelson

Major Advisor: _____

Ion I. Măndoiu

Associate Advisor: _____

Saguthevar Rajasekaran

Associate Advisor: _____

Peter Gogarten

Associate Advisor: _____

Yufeng Wu

University of Connecticut

2010

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the course of evolutionary time, gene duplication and loss has led to the evolution of highly complex genomes with thousands of genes and hundreds of gene families (Lynch, 2007). The rate of gene duplication and loss is sufficiently high that variation in gene copy number is not only revealed between species but is also an important component of individual variation within populations (Fortna et al., 2004). Changes in gene copy number can impact fitness in dramatic ways including in a number of human health conditions (McCarroll & Altshuler, 2007; Redon et al., 2006). For this reason, the impact of changes in gene copy number on evolution and human health are under increasing scrutiny.

In order to reconstruct gene family evolution we must infer a probable evolutionary history of a set of homologous genes using existing characters of those genes. The inferred relationships between the members of this group represents the history of gene family and can be represented as a tree of genes with duplication and loss events on each branch of the tree. If we understand when and what types of evolutionary events happened in the gene family, it will reveal the evolutionary driving forces and constraints on the gene family, for example, the lineage specific gene duplication and deletion rates.

Many discrete molecular mechanisms can lead to duplication or loss. Four primary mechanisms of gene duplication have been described in consensus (Hahn, 2009; Li, 1997): whole genome duplication or polyploidization (Hurley et al., 2005; Skrabanek, 1998; Van de Peer et al., 2003), unequal crossing over or tandem duplication (Shoja & Zhang, 2006), duplicative transposition (Bailey et al., 2003; Friedman & Hughes, 2004; Samonte & Eichler, 2002), and retrotransposition (Harrison et al., 2005; Zhang et al., 2002). The retention rates of gene duplicates by different mechanisms have extremely varied (Lynch, 2007, Ch. 8). Of these, whole genome duplication, tandem duplication, and duplicative transposition are DNA-mediated duplication events (abbreviated here as DD for DNA Duplication), while retrotransposition is RNA-mediated (abbreviated here as RD for RNA Duplication). Whole genome duplication has been important to the evolution of many lineages (Jaillon et al., 2004; Van de Peer et al., 2003), but it is a relatively rare event (Panopoulou & Poustka, 2005). Unlike whole genome duplication events, tandem duplication and duplicative transposition (sometimes collectively referred to as segmental duplications) occur continuously and have contributed significantly to the divergence of gene content between mammalian genomes. Duplication by retrotransposition also occurs quite frequently, but because retrotransposed gene copies are duplicated by an RNA-mediated mechanism they lack the promoter and other flanking regulatory sequences of the parental gene. For this reason, retroduplication events have long been believed to give rise primarily to non-functional pseudogenes (Petrov & Hartl, 1999; Shemesh et al., 2006). Recent studies however, have indicated the presence of many apparently functional retrocopies in various mammalian genomes, challenging traditional perspectives on the relevance of this event to genome evolution (Kaessmann et al., 2009; Sakai et al., 2007; Svensson et al., 2006; Vinckenbosch et al., 2006). Very recently retrotransposition has also been shown to contribute siRNA genes to the genome (Tam et al., 2008; Watanabe

et al., 2008).

Once a gene or genes have been duplicated the retention of those new gene copies in the genome is subject to evolutionary forces. Under neutral conditions population genetics indicates that the vast majority of newly duplicated genes will be lost. The fact that many duplicates are retained has led to the generation of many models attempting to account further retention of gene duplicates . Several of these models have been tested in computational simulations (e.g. Innan, 2009), and in molecular genetic experiments on small model organisms (e.g. Hendrickson et al., 2002). In general these models predict four major fates of newly duplicated gene copies (Hahn, 2009): 1) conservation, 2) subfunctionalization, 3) neofunctionalization, and 4) nonfunctionalization or gene deletion. Although a handful of examples for each of these outcomes have been described, most gene families cannot be explained by just one model, or sometimes by none of them (Dharia et al., 2010).

An essential step in reconstructing the evolutionary history of a gene family is determining the precise relationship between homologous members of that family. Two relationships are particularly important in this process: orthology and paralogy. Orthologs are the genes in different species that derive from a common ancestor, and paralogs are homologous genes that have diverged by gene duplication (Fitch, 2000; Koonin, 2005).

Selective pressure on the coding region of the genes and/or various mutation rates on specific lineages/timeframes, and the lack of diversification of very recent duplicates (Fitch, 2000) can make the precise determination of orthology and paralogy very difficult. Many approaches to detect true orthologs within a group of homologous genes have been attempted. Reciprocal best hit (RBH) based methods and their clustering variants are one type of approach in this field. For example, Inparanoid (Berglund et al., 2008) is quite successful at pinpointing the true or-

thologs from homologous genes by using protein sequence similarity. However, these cluster-based approaches do not use a tree to define the clusters, thus leaving the detailed orthology unclear. In contrast, phylogenetic tree based approaches can accurately identify the evolutionary history of a gene family (Arvestad, 2003), but their heavy computational load often makes these methods impractical.

One of the difficulties in detecting orthology and reconstructing gene family evolution arises from gene duplication and loss events. For example, with two duplicates on each genome of the species being compared, many combinations of gene loss events can lead orthology definition problem to a wrong orthology inference (Figure 1.1A). This pattern becomes problematic for a gene family evolutionary tree reconstruction due to the possible combination of wrong orthology relationships. Although there has been a significant amount of work on phylogenetic gene tree building to deal with the duplication and loss events, much of this work is confounded by gene loss events. I address this problem by including in our study pseudogenes as explicit evidence of gene loss. Pseudogenes are non-functional copies of duplicated genes that arise from gene duplications followed by disablements (frame shift or in-frame stop codon). Even with the limited power of detecting older fossils of lost genes, pseudogenes are still very useful information and, at least for young loss events, it is quite accurate to revive the gene loss events on mammalian lineages (Schrider et al., 2009).

It is worth noting that the definitions of orthology might be different in various contexts. *Functional orthologs* (Bandyopadhyay et al., 2006; Fang et al., 2010) are orthologs that play the same biological role in different species. On the other hand, we are interested in *ancestral orthologs*, which are the direct descendants of the ancestral genes. Sankoff (1999) called such genes the *true exemplars*, namely, the ones that best reflect the original position of the ancestral gene in the ancestral genome. Ancestral orthologs might not be functional orthologs when the current

role of functional orthologs was affected by evolutionary forces, for example, convergent evolution. As shown in Figure 1.1A, gene deletion events can introduce incongruence between ancestral orthology and functional orthology. In this case, pseudogenes can help us define the ancestral orthologs. Being the remnants of lost genes, pseudogenes can act as placeholders along the reconstruction process of gene family evolutionary history.

Furthermore, when there is not significant difference between two young gene duplicates, especially between inparalogs, the duplicates after the last speciation, (Figure 1.1B), functional or ancestral orthology definition can be indecisive on the newly duplicated genes. Frequent and continuous duplication events in mammalian genomes make the orthology identification problem more challenging due to many new duplicates. However, if we incorporate the duplication mechanisms in defining orthology, it might be possible to distinguish the ancestral orthologs from inparalogs simply because a retrotransposed copy cannot be an ancestral ortholog, as shown in Figure 1.1B.

A more important issue in orthology detection and tree reconstruction approaches is that the vast majority of these methods use protein sequences for measuring evolutionary distance or divergence. Orthology relationships or gene trees inferred from these measures can be confounded by functional constraints of the protein coding sequence and therefore might not reveal the ancestral relationship correctly. A number of molecular genetic events including gene displacement, concerted evolution, and functional convergence can lead to the convergence of protein coding sequence that might not reflect the ancestral relationship between genes.

To address this issue, I use non-coding sequence information to infer ancestral orthology. Since the gene order and the gene structure are less sensitive to the selective pressure on the genes of interest, I use local synteny and intron orthology

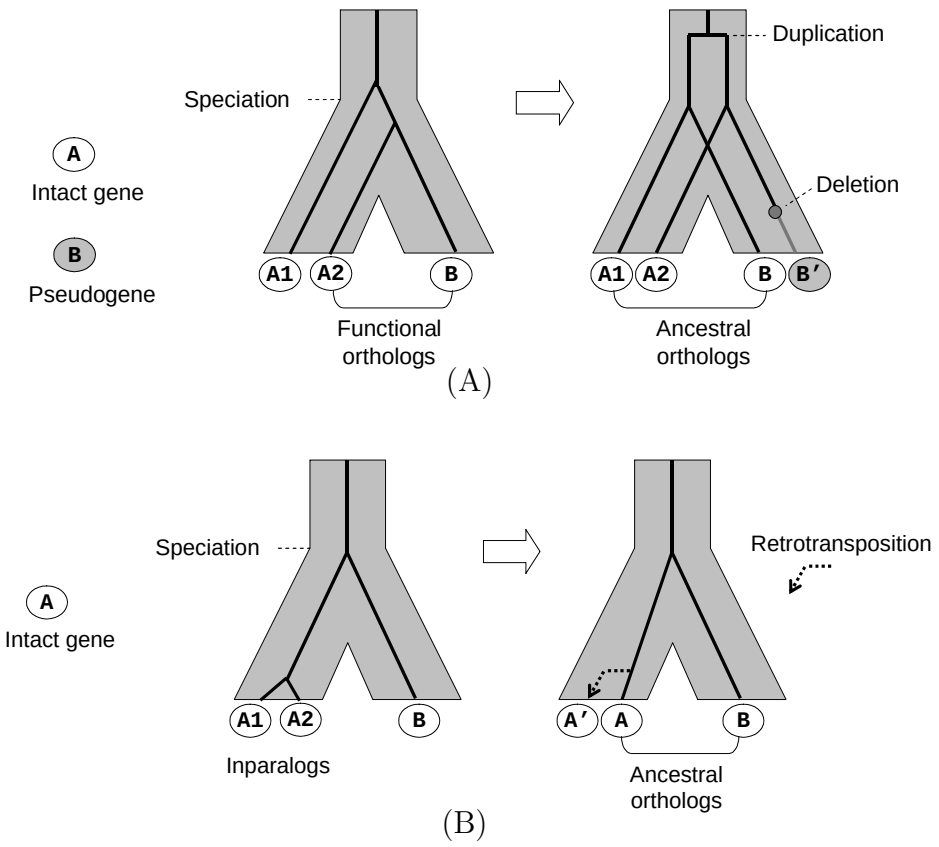


Figure 1.1: Challenges of orthology definition. (A) Gene deletion events can lead to a wrong orthology inference. Without the pseudogenes, A2 and B intact genes can be inferred as orthologs. However, (A2 intact gene, B' pseudogene) and (A1 intact gene, B intact gene) on the right are the ancestral ortholog pairs. (B) Retrotransposed copy, A', is not an ancestral ortholog of B gene.

between genes to infer their orthology and gene family evolution history. Using non-coding sequence information for this purpose is not novel. Gene order information (Fu et al., 2007; Wapinski et al., 2007a; He & Goldwasser, 2008) and gene structure information (Csurös et al., 2007; Csurös, 2008; Pavesi et al., 2008) have been used to define orthology and to reconstruct gene family evolutionary history. However, as far as we know, using local synteny and gene structure together in one framework has not been tried. Additionally, by using non-coding sequences, we can distinguish duplicate genes arising from different duplication mechanisms, e.g. retrotranspositions and DNA-mediated duplications, which is helpful to find

the ancestral orthologs (Figure 1.1B). Once we understand the gene family evolutionary history, we can pursue detailed studies on 1) relative contribution of new functional genes by different duplication mechanisms, 2) lineage and family specific preference of duplication mechanisms, and 3) the fates of gene duplicates and their retention models.

The rest of this thesis is organized as follows. I present local synteny driven orthology in Chapter 2 of this thesis. In Chapter 3, I show how to use this idea to identify gene duplication events of two different duplication mechanisms: DNA-mediated and RAN-mediated. Using this method, in Chapter 4, I analyze the evolutionary history of mammalian gene families, and try to answer the above three questions. Finally, I summarize the current status of this work together with possible future work in Chapter 5.

Chapter 2

Local Synteny Driven Orthology Definition

The accurate determination of orthology is central to comparative genomics. Pinpointing the origin of new genes, understanding the evolution of new gene families, and assessing the impact of gene and genome duplication events all require the accurate assignment of orthology between genes in distinct genomes. In complex genomes with large gene families this task requires differentiating between genes that have diverged through a speciation event (orthologs) and those derived through duplication events (paralogs). Determination of orthology and paralogy is especially challenging in mammalian species. Very large gene families, high rates of gene duplication and loss, multiple mechanisms of gene duplication, and high rates of retrotransposition all combine to make the determination of orthology between mammalian genes difficult.

Given the importance of accurate orthology assignment, many methods have been developed to identify orthologous genes. Most of these methods rely upon analysis of the inferred protein sequence of the genes in question by clustering the

[1]The results presented in this chapter are published in Jun et al. (2009a)

results of protein sequence comparisons in the classification of putative orthologs. Examples of this approach include reciprocal best BLAST (Altschul, 1997) hits and more inclusive BLAST based clustering methods. Splitting of these clusters based on relative similarity can distinguish between older and newer duplication events and is implemented in the widely used Inparanoid algorithm (Berglund et al., 2008) and related approaches (e.g. Li et al., 2003) While these methods are robust and easily implemented, they rely upon a single character, the protein sequence, for classifying genes into orthologous groups.

Recently, methods have been proposed that use genomic context in addition to protein sequence to improve orthology assignment. These methods have been most successfully implemented in fungal genomes (Kellis et al., 2004; Wapinski et al., 2007b) and in prokaryotic genomes (Lemoine et al., 2007, 2008), where gene order is far less variable than in eukaryotes. An interesting implementation of this approach is found in the SOAR and MSOAR algorithms (Fu et al., 2006, 2007), which seek to assign orthology by minimizing the recombination distance between two genomes. In most of these approaches, synteny blocks covering some percentage of the genome are used hierarchically with protein coding information to assign orthology between similar genes. Approaches that exploit synteny information can be particularly useful in resolving ambiguous sequence based matches between putative orthologs. Recently, Han & Hahn (2009) used local synteny information to identify parent-daughter relationships among duplicated genes. However, it is worth noting that "phylogenetic shadowing" (Boffelli et al., 2003) approaches used in genome assembly might lead to a lack of independence between sequence and synteny information.

In this study, I evaluate a simple method of using gene order (local synteny) in the identification of mammalian orthologs (2.2). I explicitly compare the relative performance of local synteny with other well-known orthology definition methods,

including Inparanoid, in terms of relative accuracy (2.4). I also analyze the cases of discordance between local synteny and Inparanoid to examine the false detection rates of each method (2.5). Finally, the local synteny and gene structure information are applied to resolve ambiguous many-to-many orthology relationships into one-to-one ortholog pairs (2.6). I start by introducing the terminology and informal definition of the problem followed by the previous approaches (2.1). I conclude by presenting experimental results showing that local synteny only can determine the orthology and sometimes it can improve the orthology while the traditional methods cannot.

2.1 Problem Definition and Previous Approaches

Given the sets of homologous genes on multiple genomes, we want to find counterpart pairs (or clusters) of homologous genes with various objective functions. The objective function in a typical orthology definition problem is to minimize the sum of pairwise distances or to minimize the evolutionary events to explain the data being analyzed. Since most methods assume that there is a strong correlation between DNA or proteins sequences and biological functions, a typical orthology definition problem can be considered as functional orthology definition problem. Fang et al. (2010) classified the (functional) orthology definition methods into two classes: 1) clustering pairs of the same biological functional genes, and 2) identifying the evolutionary events using phylogenetic tree, thus defining orthologs.

The most straightfoward and intuitive way to find the same biological functional genes is RBH (reciprocal best hits) on protein sequences, under the assumption that one gene does the same role in two genomes and both are present. Inparanoid algorithm (Berglund et al., 2008) is one of widely accepted methods in this category. However, not only this simple assumption makes the RBHs far from the true

orthology sometimes, also the transitive issues of using RBH are well understood. For handling more than two genomes, a transitivity rule should be in place. For example, MultiParanoid (Alexeyenko et al., 2006) uses single-linkage clustering on the multiple pairwise Inparanoid outputs. However, Johnson (2007) shows that the statistical proof that RBHs with a simple transitivity rule are neither sufficient nor necessary condition for orthology. To avoid this transitive issue, COG uses a three-way RBH to define COGs (Clusters of Orthologous Groups of protein) and combines these COGs with stringent conditions (Tatusov, 1997). However, due to the much denser RBH graph in eukaryote genomes, this approach can introduce higher false positive errors compared to prokaryotes. To overcome this issue, more comprehensive approaches have been developed. OrthoMCL uses the adjusted p-values of protein alignment in order to normalize the biased gene distances followed by Markov clustering algorithm (Li et al., 2003). Another approach to handle the transitivity issue of RBH is OMA (Roth et al., 2005) and Roundup (Deluca et al., 2006). Both use the global sequence alignment instead of local sequence alignment in RBH identification to minimize a false positive RBH owing to sharing common protein domain. Roundup uses RSD (reciprocal smallest distance (Wall et al., 2003)) which relies on global sequence alignment and maximum likelihood estimation of evolutionary distances to detect orthologs between two genomes.

Compared to this clustering based approach, as Fitch (2000) claimed that all the evolutionary process in principle could be uncovered by a phylogenetic tree, phylogenetic tree based approach is promising to minimize the transitive issue. The tree based approach uses gene/species tree reconciliation for orthology detection, i.e. the incongruence between species and gene tree must be explained by the evolutionary events, such as gene duplication, deletions, and horizontal gene transfer. Main challenge in tree based orthology inference methods is from 'tree reconciliation' step. Although the reconciliation step itself is very intuitive, some

combination of evolutionary events on the genomes and genes can lead to a wrong gene tree for an input for the reconciliation step. Since the reconciliation defines the evolutionary events on the tree, most tree reconciliation methods also provide orthology relationship. SYNERGY (Wapinski et al., 2007a), TreeBEST (Li et al., 2006), NOTUNG (Chen et al., 2000) and PrIME-GSR (Akerborg et al., 2009) are in this category. Some of these methods use parsimonious approach for reconciliation (NOTUNG and SYNERGY), and the rest use likelihood based approach (PrIME-GSR). TreeBEST is a hybrid one using a stochastic context free grammar approach to minimize duplication and loss events over the multiple gene trees, by various tree building algorithms including maximum likelihood one. This approach is being used for the orthology definition in Ensembl Compara database (Vilella et al., 2009).

Another distinct approach is to minimize the evolutionary events. MSOAR/ MultiMSOAR (Fu et al., 2007; Fu & Jiang, 2008) and gene team model (He & Goldwasser, 2008) are designed to find a set of orthologs (MSOAR) and conserved gene clusters (gene team model) by minimizing the reversal events. However, not only these methods need excessive computational power, they might give the user wrong orthology information if gene conversion or convergent evolution occurred, as they are primarily designed to use coding sequences.

Even with multiple methods for the orthology definition problem, we still have the following issues: 1) circular usage of coding sequences, both for distance measure and evolutionary fates, 2) paralogs being too similar to identify the original copy, when they are young duplicates, and 3) impractical computational burden for some methods.

I use local gene order information to define orthologous genes, similar to MSOAR and gene team model use the gene order information. However, in order to minimize the impact of gene deletion issue, impractical computational time, and micro-

rearrangements inducing the noise to global optimization algorithm, I use local synteny information. I propose a simple method for the measurement of gene order (local synteny) for the identification of mammalian orthologs in the following section.

2.2 Measures of Local Synteny

Several orthology inference methods, such as Inparanoid (Berglund et al., 2008) and OrthoMCL (Li et al., 2003), use coding sequence similarity (for example Blastp score (Altschul, 1997), Protdist (Felsenstein, 2004)) as primary orthology signal. Instead, I used local synteny information to determine orthology. I define the local synteny of two genes as the maximum number of unique homologous matches between their six neighboring genes (three upstream and three downstream immediate neighbors for each gene (Figure 2.1). Homology between two neighboring genes is defined as Blastp E-value<1e-5. To validate the use of local synteny for inferring orthology, I evaluated the correlation between Protdist and local synteny using a dataset derived from the Pfam protein family database. Pfam families are highly accurate protein families based on protein domains (Finn et al., 2006). I randomly selected 1,000 cross-species homologous protein pairs (homologs belonging to a given Pfam family) from five mammalian genomes: *Homo sapiens* (human), *Pan troglodytes* (chimp), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Canis familiaris* (dog). To avoid protein family-specific bias in this analysis, I chose one homologous pair from each Pfam family. For each pair, I computed Protdist and the degree of local synteny between the two genes. Figure 2.2A shows that there is negative correlation between Protdist and the local synteny of these samples ($r = -0.67$ with p-value<0.0001). This is not surprising as gene order is conserved between DNA segments resulting from speciation or large-scale segmen-

tal duplication events. However, as local synteny is not directly computed based on the coding sequence, we can use local synteny to test hypotheses of orthology between two genes independent of their coding sequence.

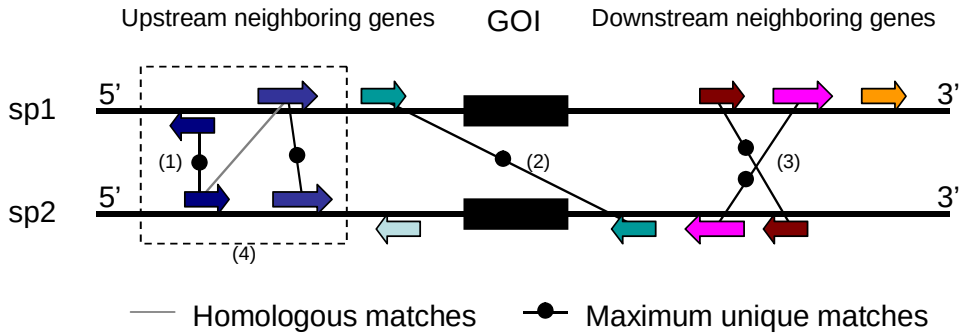


Figure 2.1: Diagram illustrating the computation of the maximum number of unique homologous matches. I counted the homologous matches between 3 neighboring genes (shown as filled arrows with corresponding gene orientations) on each side of the two genes of interest (GOI, shown as two black boxes). Homology between neighboring genes (shown as line between genes) is defined as Blastp E-value<1e-5. The homologous matches do not need to be between the genes with the same orientations (1) or on the same strand (2). Also they do not need to be co-linear (3). When there are many-to-many homologous matches, I choose the maximum unique matches (4). The number of maximum unique homologous matches in this case is 5.

Theoretically two non-orthologous genes should not share homologous neighboring genes. However, there is a small probability of homology matches occurring by chance. Moreover, rearrangements, insertions, and deletions will lead to loss of local synteny between orthologous genes. In order to account for these events, I wanted to determine an optimal window size and match percentage that could reliably identify orthologs based on local synteny. In Figure 2.2B, I test the impact of the number of matches between six neighboring genes to determine local synteny. Student's t-tests comparing Protdist means illustrate that 0, 1, and >1 matches have significantly different Protdist means (p-value ≤ 0.05). In order to choose the threshold of homologous matches, I calculate the false positive rate

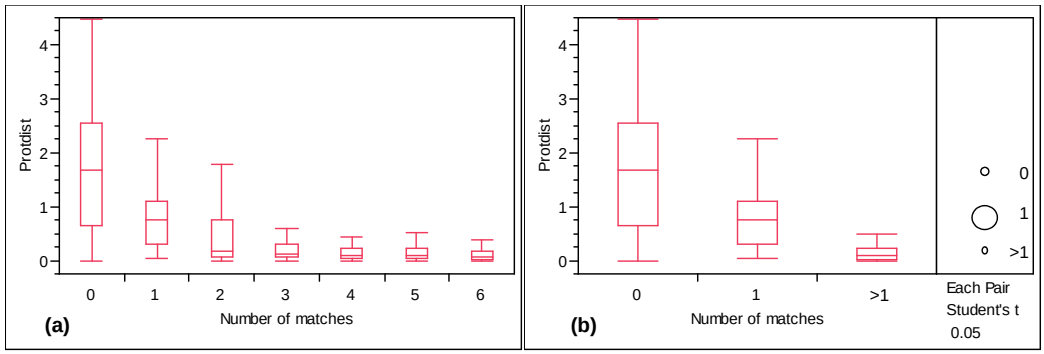


Figure 2.2: The box plot of Protdist in each level of local synteny. Local synteny is measured by the maximum number of unique homologous matches between 6 neighboring genes. (a) It shows a negative relationship between Protdist and the number of matches (r = –0.67 with p-value<0.0001). (b) No match, one match and more than one match have significantly different Protdist means (Student's t-test, p-value< 0.05). Protdist and the numbers of matches are calculated from the randomly sampled 1,000 cross-species homologous protein pairs (defined as belonging to the same Pfam families).

(FP) and false negative rate (FN) to Inparanoid orthologs and Ensembl orthologs then choose the threshold that minimizes the sum of FP and FN events. For six neighbors, the pairs with more than one homologous match minimizes FP and FN rates (0.152 to Inparanoid orthologs, and 0.151 to Ensembl orthologs (Table 2.1). Increasing the window size to 10 or 20 flanking genes does not show a significant difference in detecting orthology. Based on these results I define orthology by local synteny when the number of maximum unique homologous matches between the six neighboring genes is greater than one. I will refer to those pairs as syntenic from now on.

No. neighbors	Threshold	To Inparanoid orthologs FP(%)	FN(%)	Sum(%)	To Ensembl orthologs FP(%)	FN(%)	Sum(%)
6	>0	16.6	3.0	19.6	16.1	3.9	20.0
6	**>1**	10.3	4.9	**15.2**	9.4	5.7	**15.1**
6	>2	8.0	9.2	17.2	7.2	10.1	17.3
10	>2	10.1	4.7	14.8	9.1	5.5	14.6
10	**>3**	8.7	5.9	**14.6**	7.7	6.7	**14.4**
10	>4	8.2	10.5	18.7	7.4	11.3	18.7
20	>3	10.8	4.0	14.8	9.8	4.8	14.6
20	**>4**	9.8	4.4	**14.2**	8.9	5.1	**14.0**
20	>5	9.6	5.2	14.8	8.6	6.0	14.6

Table 2.1: False positive (FP) and false negative rates (FN) of local synteny measures to the Inparanoid orthologs and Ensembl orthologs, with using different number of neighbors (6, 10 and 20) and different thresholds to be syntenic.

2.3 Method and Datasets

2.3.1 Datasets for ortholog definitions

Five species analyzed (human, chimp, mouse, rat and dog) were obtained from Ensembl release 48 (Ensembl, 2007). I only used protein coding genes in the Ensembl database. For genes with multiple alternative transcripts I used the longest transcripts as the representative ones. I used Pfam families (Finn et al., 2006) to choose ortholog candidates. Since there are more than hundreds of millions possible protein pairs among five genomes in Pfam families, I sampled our datasets in the following way. First, I randomly selected 1,000 families from 3,418 Pfam families which have at least two representative proteins from different genomes. One cross-species protein pair was selected from each Pfam family in order to avoid a bias from big families. I used 10 sample datasets for the LCA experiment and one of them was used in the discordance analysis.

2.3.2 Local synteny based and other orthology definitions

Local synteny is measured by homology between the neighboring genes of two genes of interest. In this study the maximum unique homology matches between two sets of six neighboring genes (three upstream and three downstream neighbors) was used. The matches do not need to be co-linear or between genes on the same strand/orientation either (see Figure 2.1), which allows for genome micro-rearrangements as well as gene losses and insertions in the flanking region. The homology between neighboring genes is decided by pre-calculated Blastp (Altschul, 1997) results in the Ensembl Compara database (Flicek et al., 2008). To avoid having high local synteny due to proximate tandem array genes, I considered the tandem array genes as one neighboring gene. Within a tandem array, each gene was counted separately.

I used five orthology detection methods and Ensembl orthology in comparison with our local synteny based orthology. For each of the five orthology detection methods – namely Inparanoid (Berglund et al., 2008), OrthoMCL (Li et al., 2003), SBH (single or one-way best hits), RBH (reciprocal best hits) and BLASTP (one-way Blastp hits with the threshold) – I used the pre-computed Blastp outputs in Ensembl database as input data. Parameters and thresholds used for each method are as follows:

1. BLASTP: homology detection using E-value cutoff (= 1e-5).

2. SBH: Single-way or One-way Best Hit. 'Best-hit' is defined as the hit (or multiple hits tied) with the highest E-value (E-value cutoff = 1e-5).

3. RBH: Reciprocal Best Hit. 'Best-hit' is defined as same as SBH (E-value cutoff = 1e-5).

4. Inparanoid (v2.0): bit score cutoff = 50 bits and sequence overlap cutoff =

0.5.

5. OrthoMCL (v1.4): E-value cutoff = 1e-5 and MCL inflation index = 1.5; MCL package (v02-063) was used.

2.3.3 Latent Class Analysis (LCA)

The accurate determination of orthology is critically important to comparative genomics. However it has been a challenge to compare the various orthology determination methods without a reliable gold-standard orthology dataset (Hui & Zhou, 1998; Chen et al., 2007). The statistical technique of Latent Class Analysis allows estimates of false positive and false negative rates from data based on agreement and disagreement between various ortholog definitions.

The frequency table of agreements and disagreements between orthology detection methods was calculated and used for LCA. LCA was performed using the LEM package (Vermunt, 1997) with default parameters to estimate the false positive and false negative rates. I used a basic latent model to produce Figure 2.3A assuming independence between various orthology detection methods. However, all methods I considered are solely or partially based on protein sequences. In order to account for these dependencies, I applied another latent model with an extra latent variable. With such a model, called latent class model with random effects or a continuous factor (CFactor) model, the responses of different tests are assumed to be independent (Qu et al., 1996; Chen et al., 2007). Although the estimated error rates from the CFactor model (Figure 2.3B) are less tightly distributed than ones from the basic model, the relative values are not changed significantly compared with Figure 2.3A. The looser distribution of values in the CFactor model is likely due to a lack of convergence in these runs. I discarded the LCA runs with poorly converged error rates, which stopped at the local optima.

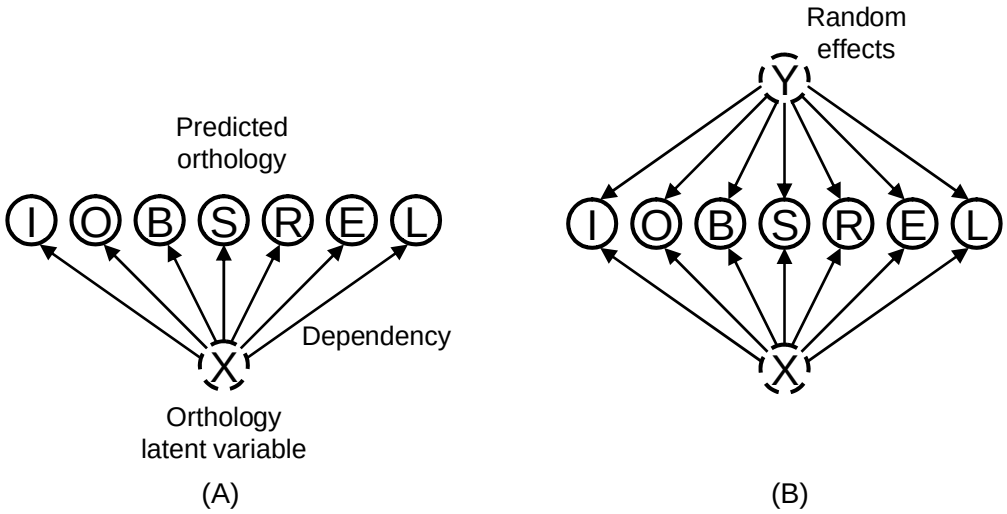


Figure 2.3: Latent class models. (A) A basic model with one latent variable (X) to estimate the error rates of orthology detection methods. (B) A CFactor model with another a continuous factor (Y) to account for the dependencies between orthology detection methods. I: Inparanoid, O: OrthoMCL, B: BLASTP, S: SBH, R: RBH, E: Ensembl, and L: Local synteny.

2.3.4 Intron Conservation Ratio (ICR)

Gene structure similarity is measured by the intron conservation ratio (ICR) between two intron-bearing genes (Rogozin et al., 2003). For genes with multiple alternative transcripts we developed a collapsed gene model that incorporates all potential exons of that gene. Resulting exon coordinates were used to obtain the protein alignments and also to align the positions of introns. ICR between two homologous genes was calculated as the ratio of the number of positional homologous introns divided by the total number of intron positions from the protein/intron alignment, similar to the method in Rogozin et al. (2003). Introns with less than 40BP were ignored in ICR calculation.

2.3.5 Case analysis

For the discordant cases between Inparanoid, local synteny, and/or ICR based orthology (Section 2.5) I investigated: 1) any significant Blastp hits other than the

sampled pairs with low ICR values for detecting false positive cases of Inparanoid orthology, 2) the other Inparanoid orthologs in all 5 species in order to confirm rearrangement history in those families, and 3) Inparanoid orthologous counterparts of non-Inparanoid sampled pairs in Figure to confirm these sampled pairs from distant paralogs.

Mouse-rat many-to-many Inparanoid ortholog groups were collected and analyzed to reconstruct their evolutionary history by considering the genomic location and intron content of their member genes (Section 2.6). Two example ortholog groups were chosen due to their unambiguous evolutionary history: for most groups it is difficult to unambiguously reconstruct the full evolution due to the presence of intermingled events.

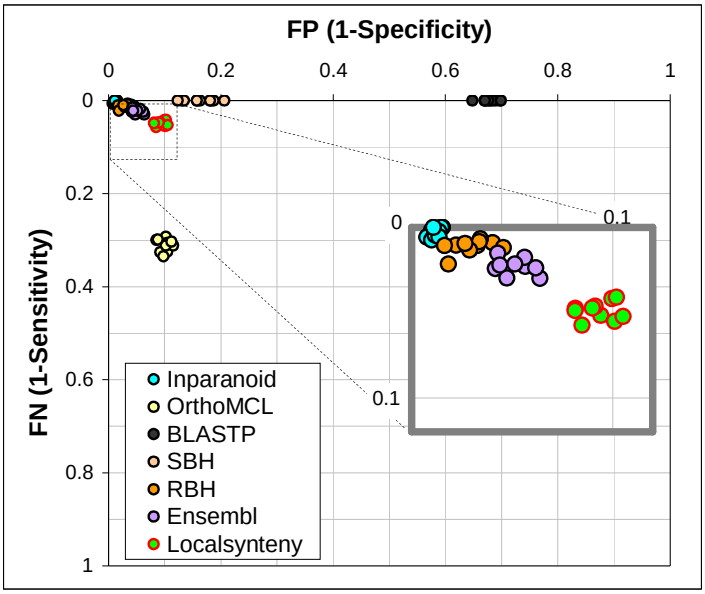
2.4 False Positive and False Negative Rates Estimated by LCA

Since many orthology detection methods use more complicated algorithms than just coding sequence similarity, a high correlation between Protdist and local synteny (Figure 2.2) is not sufficient evidence that local synteny captures true orthology. For a more rigorous analysis I compared local synteny based orthology to the orthology relationships inferred by six well-known orthology detection methods: Inparanoid (Berglund et al., 2008), OrthoMCL (Li et al., 2003), RBH (Reciprocal Best Hit), SBH (Single-way or One-way Best Hit), BLASTP, and orthology data from Ensembl (Flicek et al., 2008). Since there is no gold standard of orthology, I performed Latent Class Analysis (LCA) (Chen et al., 2007; Hui & Zhou, 1998) to estimate the accuracy (sensitivity and specificity) in the absence of a reliable standard. LCA estimates false positive (FP) and false negative rates (FN) based on agreement and disagreement between various ortholog definitions. To mini-

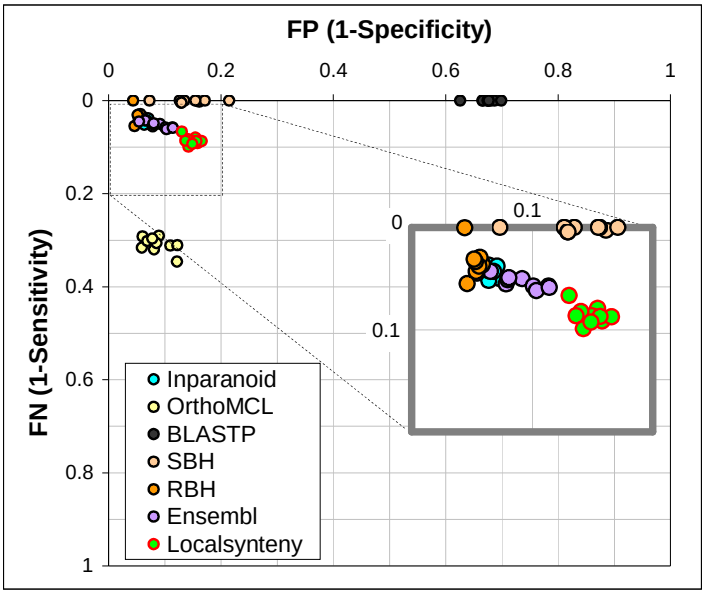
mize sampling bias 10 LCA's were performed on random samples with a size of 1,000 orthologous genes from five mammalian genomes using the same sampling method described in Section 2.2. For one sample 1,000 Pfam families were randomly selected, then one cross-species protein pair was selected from each Pfam family for analysis by the seven compared orthology inference methods. For all methods based on coding sequence similarity, the longest proteins were used as the representative proteins of the genes.

Figure 2.4A shows that orthology inference based on local synteny yields a lower FP rate than SBH/BLASTP and a lower FN rate than OrthoMCL, reinforcing the interpretation that orthology can be accurately inferred without coding sequence information. However, local synteny has a slightly higher FN rate than the four orthology methods based on coding sequence similarity (BLASTP, SBH, RBH and Inparanoid). This is partially due to the fact that these coding sequence based methods cannot distinguish retrotransposed genes from the original copies unless retrotransposed genes are sufficiently diverse. This might lead to incorrect orthology assignments (retrotransposed copies replace the original ortholog genes) or ambiguous orthology assignments (one-to-many or many-to-many ortholog groups including retrotransposed copies as their members) by these methods. Local synteny also has a slightly higher FP rate than Inparanoid and RBH. This is likely due to the fact that local synteny cannot distinguish DNA-mediated duplicates from the original copies. I analyze these discordances in more detail in the following section.

Because error rates estimated in this way may be affected by which methods are included in the analysis, we must consider the FP and FN rates estimated here as relative error rates. Furthermore, these error rates might not reflect rates obtained from a genome-wide implementation of these methods. There is an ongoing effort on standardizing protein datasets for benchmarking orthology determination



(A)



(B)

Figure 2.4: Estimated false positive (FP) and false negative rates (FN) for seven orthology detection methods with (A) a basic model and (B) a CFactor latent model. FP and FN rates are estimated for each method by using LCA from 10 sampling replicates. Inset figures show the FP and FN rates of orthology detection methods having lowest FP and FN rates.

methods (Gabaldón et al., 2009) which would help resolve this issue. In Figure 2.4, Inparanoid and RBH agree more closely than any other pair of orthology definitions. This is in accord with the fact that Inparanoid uses the reciprocal best

hit to start core ortholog idenfitication. SBH and BLASTP have zero FN rates and higher FP rates than Inparanoid and RBH due to less stringent conditions used for ortholog detection in these methods. The order of FP rates (BLASTP > SBH > RBH) is concordant with the stringency of each method. The FP rate of local synteny based orthology falls between SBH and RBH (Figure 2.4A). This is reasonable considering the fact that SBH and local synteny based orthology cannot distinguish close paralogs from ortholog pairs, but local synteny can separate retrotransposed paralog copies from orthologs. The FN rate of local synteny based orthology is higher than those of Blastp based orthologies (Inparanoid, RBH, SBH, and BLASTP). This may be due to distant paralogs retaining flanking genes, but diverging in their coding sequences enough to be distinguishable by Blastp. Retrocopies miscalls by Inparanoid (e.g. Figure 2.6) are also likely to have contributed to the relative FN rate of local synteny based orthology. Due to the fact that OrthoMCL detected the smallest number of orthologs in any sample (data not shown) the estimated FN rates from OrthoMCL are the highest in this experiment (approximately 0.3). This is opposite from the result of Chen et al. (2007). where OrthoMCL and Inparanoid were shown to have lowest estimated FP and FN rates and OrthoMCL has lower FN rates than Inparanoid. The apparent disparity between these results could be explained by the fact that LCA is designed to estimate consensus FP and FN rates without any guarantee that the estimated rates are close to absolute values, and difference in the species sets used in the two experiments: our dataset of five mammalian genomes and the comparatively distant seven eukaryotic genomes used in Chen et al. (2007).

Figure 2.4B shows the estimated error rates by using the CFactor model (Figure 2.3B) for seven orthology detection methods. Not like Figure 2.4A, which is from a basic model (Figure 2.3A), the error rates of Inparanoid and RBH are overlapped each other and the estimated rates from the CFactor model are less tightly

distributed, e.g. standard errors of the estimated FP and FN of Inparanoid with a basic model are 0.0008 and 0.0008 respectively, while ones from the CFactor model are 0.0013 and 0.0016. However, the relative positions of the estimated error rates from two models are very similar. In fact, the averaged distances between the error rates of Inparanoid and synteny-based method are very close; 0.0934 from a basic model and 0.0927 from the CFactor model.

2.5 Discordance Between Inparanoid Orthologs and Local Synteny Driven Orthologs

Because Inparanoid is one of the most widely used ortholog definition methods (Bandyopadhyay et al., 2006; Su et al., 2006; Ho Sui et al., 2007) and is purely based on the coding sequence information, I decided to do a more thorough comparison of Inparanoid and local synteny based orthology. Figure 2.5 shows the agreement and disagreement between these two orthology prediction methods. The majority of these samples are concordant between two ortholog predictions (syntenic/Inparanoid (55.1%) and non-syntenic/non-Inparanoid (37.9%)), which agree with the LCA results (Figure 2.4). However, 2.5% of Inparanoid orthologs are non-syntenic, and 4.5% of gene pairs are syntenic but not Inparanoid orthologs. In order to identify the source of this discordance I employed intron-based evidence, as described in the Method section 2.3.

Figure 2.5 shows ICR histograms for pairs of genes falling in each of the four classes of agreement/disagreement between Inparanoid and local synteny. In both concordance cases, namely for syntenic Inparanoid orthologs (Figure 2.5A) and non-syntenic non-Inparanoid orthologs (Figure 2.5D), ICR is in strong agreement with the orthology assignments made by the two methods. Indeed, most of the syntenic Inparanoid orthologs have ICR of 1 (Figure 2.5A), and the majority of

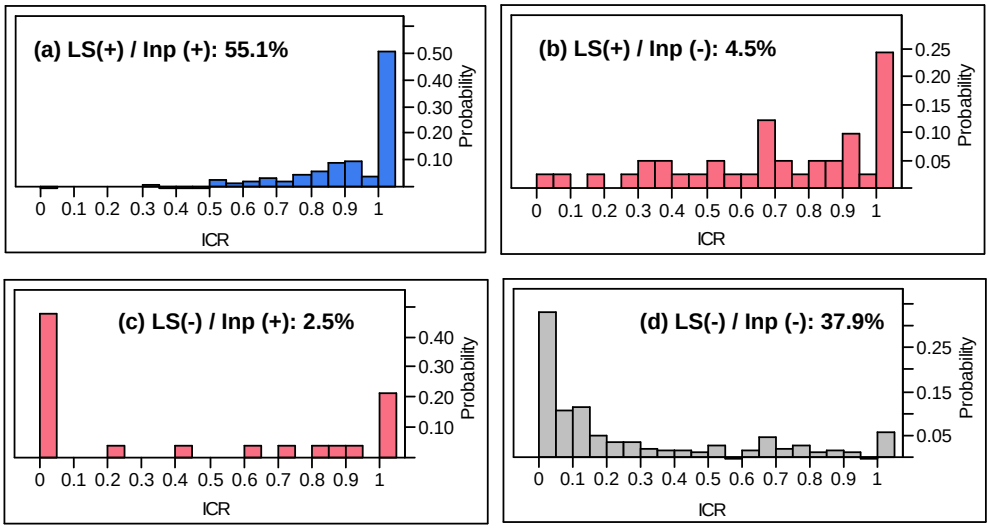


Figure 2.5: Intron conservation ratio (ICR) histograms in four concordant and discordant cases between Inparanoid orthology and local synteny based orthology. 7% of disagreement between Inparanoid orthology and local synteny based orthology and the majority (93%) of sample pairs are concordant between two orthology methods. Most of the pairs concordant between local synteny and Inparanoid, (A) syntenic Inparanoid ortholog (denoted as LS(+)/Inp(+)) and (D) non-syntenic non-Inparanoid pairs (LS(-)/Inp(-)), are also concordant with ICR: orthologs have a high ICR and non-orthologs have a low ICR. However in discordant cases, (B) syntenic non-Inparanoid (LS(+)/Inp(-)) and (C) non-syntenic Inparanoid orthologs (LS(-)/Inp(+)), ICR histograms show a partial agreement with two orthology definitions. Also there are small numbers of non-syntenic non-Inparanoid pairs (LS(-)/Inp(-)) having perfect ICR in panel (D).

non-syntenic non-Inparanoid orthologs have ICR < 0.5 (Figure 2.5D). In the two discordant cases (7% of the evaluated gene pairs) intron evidence can be used to resolve the conflicting assignments made by Inparanoid and local synteny. About 3/4 of syntenic non-Inparanoid orthologs (Figure 2.5B) have ICR > 0.5 and half of non-syntenic Inparanoid orthologs (Figure 2.5C) have ICR $= 0$. This suggests that in these cases local synteny based orthology assignments are more often concordant with gene structure evidence (ICR) than those based on coding sequence similarity. However, the ICR histogram of non-syntenic Inparanoid orthologs (Figure 2.5C) has a bimodal distribution, which might arise from a mixture of FNs from local

synteny (pairs with high ICR) and FPs from Inparanoid (pairs with low ICR). I further investigate these cases in the following subsections.

2.5.1 Non-syntenic Inparanoid orthologs with zero ICR: Retrotransposed copies

All the non-syntenic Inparanoid ortholog pairs with zero ICR contain one intronless copy and one intron-bearing member. Based on local synteny information and intron conservation ratio, these intronless copies are probably retrotransposed copies of the original orthologs. Of all the samples, 1.2% are non-syntenic 0-ICR Inparanoid orthologs. Inparanoid likely included retrotransposed copies in these orthologous groups because the retrotransposed copies have not diverged sufficiently to be distinguished from their parent gene. In 0.2% of samples Inparanoid chose retrotransposed copies even though there were other syntenic high-ICR copies. Figure 2.6 shows one of these pairs. In this case, the Blastp score of the dog gene to the retrotransposed rat gene (shown as Rat A) is smaller than to the syntenic high-ICR rat gene (Rat B), which caused an Inparanoid miscall, assigning the retrotransposed paralog as the ortholog.

The rest of non-syntenic 0-ICR Inparanoid orthologs (1.0%) of all samples are classified as one-to-many orthologs by Inparanoid. In many of these one-to-many orthologs, syntenic high-ICR pairs are chosen as "core orthologs" (based on the Inparanoid scores) with additional retrotransposed copies added to the orthology group because their protein sequences are still close to those of the original copies. By using local synteny, the retrotransposed copies in one-to-many orthologs would be distinguished from the other members. Thus, local synteny based orthology separates retrotransposed copies from those generated by speciation or other duplication mechanisms and can be more informative in recovering the evolutionary

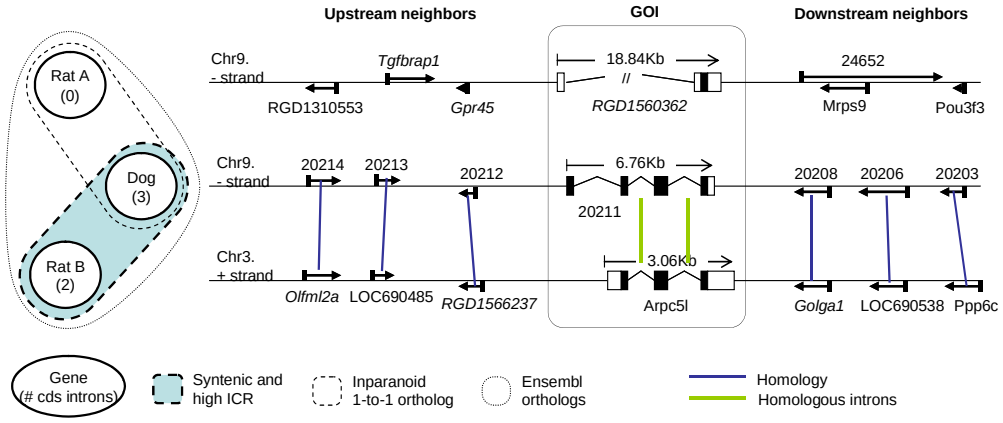


Figure 2.6: One example of the retrotransposed copy miscall cases by Inparanoid which is confirmed with local synteny and ICR. The Ensembl IDs of GOIs are ENSRNOG00000016444 (Rat A), ENSRNOG00000014317 (Rat B) and ENSCAFG00000020211 (Dog). Genes are shown corresponding to the strand of GOI in order to show the homology between neighboring genes. Five digits gene IDs are the last five digits of Ensembl gene IDs. IDs in italic typeset are predicted ones. Homology between two neighboring genes are defined by Blastp E-value<1e-5.

history of a gene family.

2.5.2 Non-syntenic Inparanoid orthologs with non-zero ICR: Loss of local synteny

All non-syntenic Inparanoid orthologs with an ICR > 0 (1.3%) are likely to result from the loss of local synteny. They each have one homologous match between neighboring genes (lower than the threshold of being syntenic) and are selected from distant species pairs (i.e. not human-chimp or mouse-rat). Each of these distant pairs is part of larger orthology groups (5 species), in which the counterparts in closer species pairs have higher local synteny (more than 1 match). This loss of local synteny likely results from rearrangements and gene insertions or gene losses in more distant species.

2.5.3 Syntenic non-Inparanoid orthologs: Distant paralogs

There are 2.3% syntenic non-Inparanoid orthologs with an ICR ≥ 0.5 (Figure 2.5B). The majority (2.1/2.3%) of syntenic non-Inparanoid orthologs with high ICR are likely distant paralogs: both genes are in different syntenic Inparanoid orthology groups with high ICR. This may result from old DNA-mediated duplication events followed by speciation events without significant local genome rearrangements. Local synteny cannot distinguish between orthologs created by large-scale segmental duplications and by polyploidy events. Moreover, the syntenic non-Inparanoid pairs with an ICR < 0.5 (1.1% of all tested pairs) are also likely distant paralogs, by old tandem duplications with different gene structures resulting in lower ICR.

2.5.4 Non-syntenic non-Inparanoid orthologs with ICR of 1: More distant paralogs

In the non-syntenic non-Inparanoid case (Figure 2.5D), where the most of these pairs have low ICR, we still find 2.2% of pairs with ICR equal to 1. All these pairs are likely distant paralogs. Table 2.2 contains the summary of all of these cases.

2.6 Local Synteny Breaks the Tie

Since local synteny is able to differentiate some speciation and duplication events, a phylogenetic tree (or duplication-speciation history) of ambiguous many-to-many orthologs may be determined using local synteny information. In mouse-to-rat ortholog definitions from Inparanoid, there are 131 many-to-many ortholog groups. The majority ($\sim$75%) of many-to-many groups are comprised of DNA-mediated duplicated copies (usually from tandem duplications combined with one or two

Orthology	ICR	Explanations
Non-syntenic Inparanoid orthologs	ICR = 0	Retrotransposed copy miscalls (0.2%) Part of 1-to-many (1.0%)
	ICR > 0	Loss of local synteny (1.3%)
Syntenic non-Inparanoid orthologs	ICR $\geq$ 0.5	Retrotransposed copy miscalls (0.2%) Distant paralogs (2.1%)
	ICR < 0.5	Distant paralogs (1.1%)
Non-syntenic non-Inparanoid orthologs	ICR = 1	Distant paralogs (2.2%)

Table 2.2: Summary of disagreement among three measures: Inparanoid orthology, local synteny based orthology, and intron conservation ratio (ICR).

distant segmental duplication events) while ∼20% have true orthologs (confirmed by local synteny and ICR) as well as non-syntenic intronless copies (probably from retrotransposition events). I present two many-to-many Inparanoid ortholog groups where the local synteny determines the order of evolutionary events in the gene family.

One of them is an example of orthologs from distant DNA-mediated duplication event(s) followed by possible rearrangements or gene gains/losses, and then speciation event (Figure 2.7). In this ortholog group, there are three mouse gene members, ENSMUSG00000001175 (MGI symbol: `Calm1`), ENSMUSG00000019370 (`Calm3`), ENSMUSG00000036438 (`Calm3`, which referred to `Calm3x` in the figure to avoid confusion) and two rat gene members, ENSRNOG00000004060 (`Calm1`), ENSRNOG00000016770 (`Calm3`). All the genes are on different chromosomes and the transcripts of these (protein coding) genes are known. Since all the cross-species pairwise sequence similarity measures are equal, Inparanoid could not pick distinct orthologs nor could phylogenetic tree-building programs determine the tree. Neither could ICR break the tie due to a high conservation of gene structure. Finally, no Ensembl ortholog prediction is made between these genes. However, in the fig-

ure, two Calm1 genes (ENSMUSG00000001175 and ENSRNOG00000004060) and two Calm3 genes (ENSMUSG00000019370 and ENSRNOG00000016770) have high local synteny (4 and 5 matches, respectively) and any other local synteny is either 0 or 1. In this specific case, local synteny helps break the tie in sequence based similarity. Using this information I can infer an old DNA-mediated duplication (DD) event before mouse-rat speciation giving rise to the Calm1 and Calm3 ancestors followed by rearrangements reducing the local synteny between these two mouse and rat ortholog pairs (Figure 2.7). The third mouse gene (Calm3x) has just one match with mouse Calm1 gene and rat Calm1 gene, and high levels of intron conservation, indicating a DNA-mediated duplication event, but I do not have enough local synteny information to tell precisely when the duplication occurred. Also since the mouse Calm3x gene has only one match with the rat Calm1 gene, local synteny does not find this apparent ortholog.

Another case where local synteny clarifies the history of closely related groups of duplicates can be seen in a mouse-rat many-to-many Inparanoid ortholog group including retrotransposed or RNA-mediated duplicated (RD) copies (Figure 2.8). This example contains two members from each species, ENSMUSG00000013701 (MGI symbol: Timm23 referred as Mm1 in the figure), ENSMUSG00000069622 (Timm23 as Mm2) and ENSRNOG00000019811 (Timm23 as Rn1), ENSRNOG000000-32900 (TIM23_RAT as Rn2), where each has one known transcript on different chromosomes. Again, neither Inparanoid nor pairwise Protdist analysis could discriminate orthologs due to identical cross-species Blastp measures. Ensembl has a bigger ortholog group including these four genes, but no better information about which are orthologs or RD copies. However two intron bearing genes have perfect local synteny (= 6) and an ICR = 1, and two intronless copies do not have any local syntenic match to the two intron bearing genes. Therefore, I can infer that two intron bearing genes are the main orthologs and two intronless genes are RD

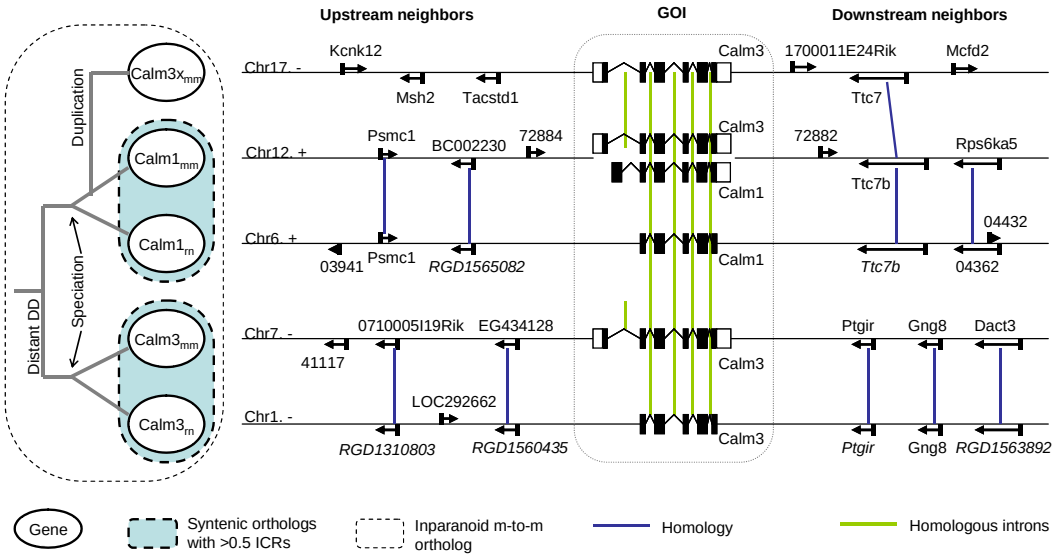


Figure 2.7: One example of many-to-many Inparanoid ortholog groups where a DD event proceeded mouse-rat speciation. The Ensembl gene IDs of GOI are ENSMUSG00000001175 (`Calm1mm`), ENSMUSG00000019370 (`Calm3mm`), ENSMUSG00000036438 (`Calm3xmm`), ENSRNOG00000004060 (`Calm1rn`), ENSRNOG00000016770 (`Calm3rn`). The gene structures and neighboring gene orders of five Calm1 and Calm3 genes in mouse and rat genomes are shown. The event tree on the left side is predicted based on the local synteny. DD: DNA-mediated duplication. Genes are shown corresponding to the strand of GOI in order to show the homology between neighboring genes. Five digits gene IDs are the last five digits of Ensembl gene IDs. IDs in italic typeset are predicted ones. Homology between two neighboring genes are defined by Blastp E-value<1e-5.

copies of these orthologs. Furthermore, since intronless genes are not syntenic to each other, I infer that the two intronless genes are the result of two separate RD events on each species lineage.

2.7 Discussion and Conclusions

2.7.1 Gene order as a measure of conservation

Synteny information has been used to cluster synteny blocks between two related genomes in order to detect orthologous gene pairs (Fu et al., 2007; Zheng et al.,

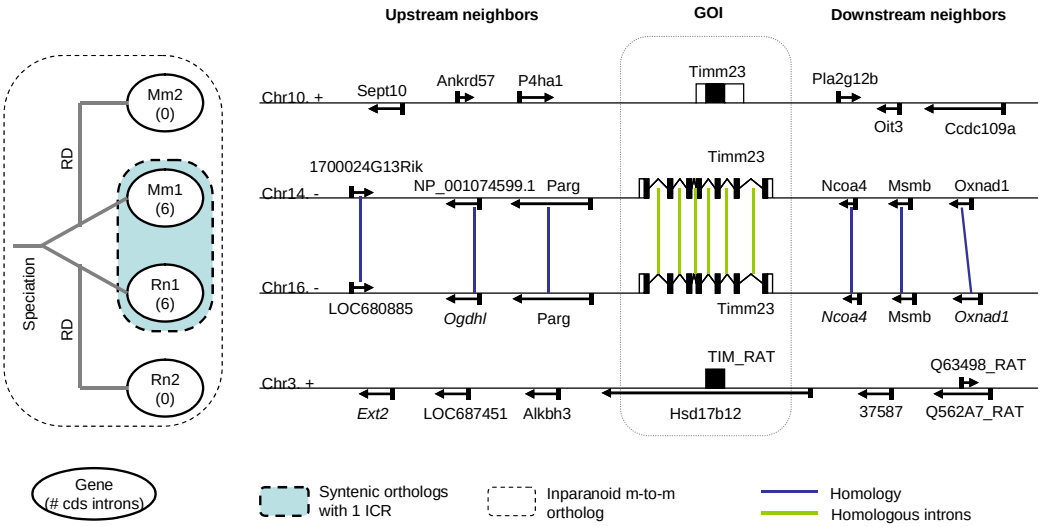


Figure 2.8: One example of many-to-many Inparanoid ortholog groups where RD events followed the mouse-rat speciation. The Ensembl gene IDs of GOI are ENSMUSG00000013701 (Mm1), ENSMUSG00000069622 (Mm2) and ENSRNOG00000019811 (Rn1), ENSRNOG00000032900 (Rn2). The gene structures and neighboring gene orders of four `Timm23` genes in mouse and rat genomes are shown. The event tree on the left side is predicted based on the local synteny. RD: retrotransposition. Genes are shown corresponding to the strand of GOI in order to show the homology between neighboring genes. Homology between two neighboring genes are defined by Blastp E-value<1e-5.

2005) and to reconstruct phylogenetic trees (Wapinski et al., 2007a). These synteny blocks are generally used as "genomic anchors" (Altschul, 1997) or to place gene loss/deletion events on phylogenetic trees (Chen et al., 2000; Poptsova & Gogarten, 2007), not as a definitive measure to distinguish close paralogs from distant paralogs. In Zheng et al. (2005), one of three methods to define orthologs between human and mouse used a genomic anchor approach. They identified synteny anchors and synteny blocks (Waterston et al., 2002; Mural et al., 2002) then introduced a local synteny approach for the anchor-poor regions by accepting the pairs of genes flanked by previously identified ortholog pairs. With this approach they found 11% more orthologs than by RBH alone. Another similar approach to local synteny was used to reconstruct phylogenetic gene trees from coding se-

quence similarity and local gene order (Wapinski et al., 2007a). In the algorithm SYNERGY, Wapinski et al. measure a synteny similarity score for a pair of genes by counting the neighboring genes in syntenic blocks. However, SYNERGY has not been tested in mammalian genomes. Finally, MSOAR (Fu et al., 2007) uses combinatorial optimization on global gene order to identify orthologs based on minimal rearrangement scenarios. However, Fu et al. point out that the global optimization might lead to false ortholog links in some scenarios. Because our local synteny exploits proximate synteny information I expect lower false positive ratios than those obtained with MSOAR.

Moreover, local synteny information has been widely used to confirm the orthologous genes: distinguishing parents and daughters (Han et al., 2009), confirming unitary pseudogenes (Zhang et al., 2010) and determining the orthology of LTR (Kijima & Innan, 2010).

Most importantly, when the orthology is defined by codon sequence similarity, testing any hypothesis of selective pressures on orthologous gene presents tautological challenges. Since orthology detection by local synteny is not based on the comparison of coding sequence information between candidate orthologs, testing the selective pressure between orthologs becomes a valid comparison of largely independent variables. However, gene order also degrades over evolutionary time. How well synteny will be able to effectively identify ancient orthologs remains to be seen.

2.7.2 Gene duplication mechanisms and orthologs

Gene duplications are a major force in genome evolution (Ohno, 1970). Genes are duplicated through two main duplication mechanisms; DNA-mediated and RNA-mediated (Ohno, 1970; Zhang, 2003). DNA-mediated duplications (DD) can include multiple genes and associated intergenic sequences and introns. On the other

hand, RNA-mediated duplication (RD), or retrotransposition, only copies coding sequences in the duplication event. Retrotransposed genes had been considered mostly "dead on arrival", but recent studies (Sakai et al., 2007; Vinckenbosch et al., 2006) show that there are many functional RD copies in the human and mouse genomes. Sometimes the difference between coding sequences of parental genes and RD copies is not large enough to distinguish the RD copy from the parent. RD copies, however, do not share introns or flanking genes with the parental paralog. Therefore, local synteny and gene structure can often separate the RD copy from the original gene. However, when neither has an intron, only local synteny will distinguish parental and RD copies. Based on our random sampling, approximately 8∼10% of Pfam orthologs are intronless gene pairs.

2.7.3 Gene order helps illuminate gene family evolution

Reconstructing phylogenetic trees informs our understanding of the evolutionary history of gene families. Using tree reconciliation between a species tree and gene tree we can identify duplication and lost events on the tree. However by distinguishing two duplication mechanisms, DNA-mediated and RNA-mediated, not just identifying duplication events, we can sometimes place duplication events in an appropriate phylogenetic context. For example, as Figure 2.7 and Figure 2.8 show, when coding sequence does not distinguish paralogs, local synteny can determine whether DNA or RNA-mediated duplication occurred first. Local synteny can also help place RD duplication events before or after speciation (Figure 2.8). Even when the coding sequence of RD duplicates drifts apart, pre-speciation RD genes often retain local synteny. Conversely, when RD duplicates are young enough to be indistinguishable by coding sequence comparison, synteny can discriminate between pre-speciation duplications, and independent RD duplication events in parallel lineages. Finally, local synteny information can often resolve the order of

iterative DNA-mediated duplication events in large gene families (see also Han & Hahn (2009)).

2.7.4 Conclusions

In this chapter, I show that local synteny alone is sufficient to identify orthologs within this five-mammal clade. Latent class analysis reveals that false positive and false negative rates of local synteny based orthology are comparable to those of coding sequence based methods. Also I investigate the reasons for concordance and a discordance between coding sequence based orthology (Inparanoid) and local synteny based orthology. In the five mammalian genomes studied, 93% of the sampled inter-species pairs were found to be concordant between the two orthology methods, illustrating that local synteny can largely substitute for coding sequence in identifying orthologs. However, 7% of pairs were found to be discordant. Discordance is often associated with evolutionary events like retrotransposition, iterative DNA-mediated duplication, and genome rearrangement. Analysis of discordant cases between local synteny and Inparanoid shows that local synteny can differentiate between true orthologs and recent retrogenes, can split ambiguous many-to-many orthology groups into more precise one-to-one ortholog pairs, and, when employed in a genome-wide screen, might help in highlighting possible cases of non-orthologous gene displacement by retrocopied paralogs in mammalian genomes.

Although our local synteny based method is reliable enough to define orthology and can easily distinguish RNA-mediated copies from source copies, local synteny information tends to have less discriminating power for successive DNA-mediated duplication events. This makes tandemly duplicated copies and inparalogs from DNA-mediated duplication hard to handle. Another challenge arises from the nature of local synteny in that local synteny depends entirely on the quality of

genome assembly. Situations of low coverage and low quality intergenic sequence can make local synteny information unreliable. As more advanced sequencing technologies and corresponding assembly algorithms are available, however, we expect more reliable genome assemblies and, thus more reliable local synteny based orthology detection.

As possible future work, we want to answer the following questions: (1) Can local synteny be used to define orthology between distant species and, if so, how many neighboring genes do we need to use? (2) Can we apply a simple transitivity rule for more than two genomes? The answer to these questions may require more accurate rearrangement models than simple counting of homologous matches. With these improvements we might be able to detect non-orthologous gene displacement (Koonin et al., 1996) and convergent evolution in wider range of mammalian genomes.

Chapter 3

A Method for Gene Duplication Events Detection

In order to reconstruct the evolutionary history of a gene family, we need to precisely locate the gene duplication and loss events on the phylogenetic tree. With a known species tree, the location of gene duplication and loss events is identical to the reconstruction of gene family evolution history. Previous gene family evolution reconstruction approaches attempt to minimize the gene duplication/loss (parsimony approach), e.g. Notung (Chen et al., 2000), or to find the most probable evolutionary history by maximum likelihood approach or Bayesian approach, e.g. PrIME-GSR (Akerborg et al., 2009). As discussed in earlier chapters, using coding sequence similarity and ignoring pseudogene information can lead these reconstruction approaches to incorrect conclusions. To overcome this problem, I developed a method to detect the gene duplication events on phylogenetic trees by using: 1) non-coding sequence information including local synteny and gene structure information, and 2) pseudogene information (Jun et al., 2008). By including pseudogenes into the data process pipeline, we minimize distortion associated with

[1]The results presented in this chapter are published in Jun et al. (2009b)

gene loss events.

In addition, by avoiding the use of coding sequences to define orthologs and build a tree, we can measure evolutionary forces on ancestral relationships independent of the tautologies born from defining relationships based on the characters of evolutionary interest (the coding sequence). For this reason I believe that local synteny may be a better method for estimating lineage specific selective pressure on the fate of gene duplicates. Furthermore, there are additional advantages to using non-coding sequence information to differentiate duplication mechanisms. RNA-mediated duplicated (RD) genes cannot bring the flanking sequences and intron sequences; whereas gene copies by DNA-mediated duplication (DD) mechanisms may keep the neighboring genes and similar gene structure. Our method, therefore, can detect and differentiate duplication events arising from these two different duplication mechanisms.

In this chapter, the problem definition and our contributions in this problem are introduced first (3.1). Then the detailed steps for gene duplication events detection method are illustrated (3.2) followed by some discussions and possible future work (3.3).

3.1 Problem definition

Given a set of homologous genes in various genomes (a gene family) with the corresponding species tree, we want to reconstruct an evolutionary history of that gene family. Available information for solving this problem includes coding sequences, gene structures and gene order information. Most previous approaches use coding sequences for building trees. The sequences can be used as a string of characteristics, e.g. DNA sequences or amino acid sequences, and then parsimony or statistical approaches can be applied to it. ML (Addario-Berry et al., 2003)

or Bayesian methods (Arvestad, 2003) can be easily applied on these strings of characteristics. In addition to coding sequence information, some methods use non-coding sequence information to build a tree. SYNERGY (Wapinski et al., 2007a) incorporated coding sequences and gene order information together for reconstructing an evolutionary events history.

As the ortholog definition problem suffers from the gene duplication/deletion issues, reconstruction problem of gene family evolution history also has been struggling with them. Especially, gene deletion events make this problem very hard, since a parsimonious solution is not always accurate to understand deletion events. To overcome this gene deletion problem, I use the pseudogene information as placeholders for gene deletion events. I start by augmenting the pseudogenes of every member of a gene family, which enable us to capture more duplication events and help us minimize the wrong orthology assignment due to gene deletion events.

3.2 Methods

3.2.1 Family definition

I used Ensembl protein families (release 37) (Hubbard et al., 2005) as the initial families. Ensembl protein family database is defined by running the Markov Clustering (MCL) algorithm (Enright, 2002). For genes with multiple alternative transcripts we developed a collapsed gene model that incorporates all predicted exons of the gene. Resulting exon coordinates were used to obtain a representative protein sequence that was used for subsequent homology assignment and dN/dS computations.

Pseudogenes were identified using PseudoPipe (Zhang et al., 2006) seeded with known transcripts from Ensembl. Each pseudogene was added to the corresponding Ensembl gene families. This process resulted in super-families consisting of both

protein coding genes and their related pseudogenes.

3.2.2 Identification of duplication events

Within each super-family a local synteny level was computed for all pairwise combinations of super-family members. Local synteny is defined as homology of upstream and downstream neighboring genes, as described in Chapter 2. For each pair, I checked homology between the 3 nearest up- and downstream neighboring Ensembl annotated genes. Homology between neighbors was defined by a Blastp (Altschul, 1997) E-value<1e-5. After this analysis, for every pair (g_i, g_j) of family members I obtained two numbers $0 \leq n_u^{ij}, n_d^{ij} \leq 3$ representing the homology upstream and downstream neighbors. A synteny level $s_{i,j}$ of **2** was assigned to every pair of genes or pseudogenes that had homologous neighbors on both sides, up and down (i.e., whenever $n_u^{ij}, n_d^{ij} \geq 1$). When one side lacked homologous neighbors, I assigned a synteny level $s_{i,j}$ of **1** only if the other side had at least two homologous neighbors; otherwise, I assigned a synteny level $s_{i,j}$ of **0**.

Local synteny levels were used in a two-stage clustering algorithm (Figure 3.1 to identify syntenic ortholog/paralog clusters. In our algorithm, for a set X of genes and pseudogenes, $Sp(X)$ denotes the set of species represented in X. For a set S of species, $LCA(S)$ denotes the last common ancestor in the phylogenetic tree. In the first stage, I used a single-linkage clustering algorithm to obtain core clusters by merging pairs of genes and pseudogenes with local synteny level of **2**, predicted to be either orthologs or paralogs resulting from DD events which preserve up and downstream neighbors. In the second stage, I merged pairs of core clusters if every member of one cluster had synteny level of **1** to every member of the other cluster (also predicted to be orthologs or paralogs from DD events). Any two non-overlapping clusters from this two-stage clustering algorithm are mutually non-syntenic. Second stage clusters spanning a phylogenetically contiguous subset

Input: Family of genes and pseudogenes, $F = \{g_1, g_2, \ldots, g_N\}$ with species information and pairwise synteny levels $s_{i,j}$

Initialize:

$C \leftarrow \emptyset$

$U \leftarrow \{g_1, g_2, ..., g_N\}$

(Stage 1) Single-linkage clustering with synteny level 2:

While $U \neq \emptyset$ **do**

Select an arbitrary member g_i of U
$U \leftarrow U \setminus \{g_i\}$; $C_{open} \leftarrow \{g_i\}$
While there exists $g_i \in U$ with synteny 2 to a member of C_{open}, **do**
$U \leftarrow U \setminus \{g_i\}$; $C_{open} \leftarrow C_{open} \cup \{g_j\}$
$C \leftarrow C \cup C_{open}$

(Stage 2) Merging of clusters with high average pairwise synteny:

While there is a pair (C_i, C_m) where **SYNTENIC_TEST**(C_i, C_m) is true, **do**

$C \leftarrow C \setminus \{C_i, C_m\}$
$C \leftarrow C \cup \{C_i \cup C_m\}$

Return C

SYNTENIC_TEST (A, B)

If $Sp(A)$ and $Sp(B)$ are subsets of different lineages, i.e.
$LCA(Sp(A)) \neq LCA(Sp(A \cup B))$ and $LCA(Sp(B)) \neq LCA(Sp(A \cup B))$, **then**

If $S_{i,j} = 1$ for every pair $g_i \in A$, $g_j \in B$ **then, return true**

Else, if $LCA(Sp(A)) = LCA(Sp(A \cup B))$, **then**

$A' \leftarrow$ set of genes and pseudogenes of A of species descending from $LCA(Sp(B))$

If $s_{i,j} = 1$ for every pair $g_i \in A'$, $g_j \in B$ **then return true**

Else, return false

Figure 3.1: Two-stage clustering algorithm. For a set X of genes and pseudogenes, $Sp(X)$ denotes the set of species represented in X. For a set S of species, $LCA(S)$ denotes the last common ancestor in the phylogenetic tree.

of the species represented in larger clusters from the same super-family represent putative descendants of RD events or DD events that have lost local synteny. Since retrotransposed gene copies generally lack introns due to their RNA-intermediate nature, I distinguish between these possibilities using intron content conservation scores as described below.

Within each cluster produced by the above clustering algorithm, there may be successive DD events. I used UPGMA (Unweighted Pair Group Method with Arithmetic mean) (Sokal & Sneath, 1973) to find these DD events. For input to UPGMA I compute the distance between two members g_i and g_j as the Pearson's correlation coefficient between the two vectors, $(n_u^{ik} + n_d^{ik})_k$ and $(n_u^{jk} + n_d^{jk})_k$, i.e., sums of upstream and downstream homologous neighbors with remaining genes g_k in the cluster. Given the UPGMA gene trees, I counted the inner nodes as DD events when two subtrees from such an inner node are in a species-subset relationship. If two subtrees from an inner node had disjoint species sets, this node was considered as a speciation event (Figure 3.2).

I distinguish between putative descendants of RD events or DD events that have lost local synteny using intron conservation scores between descendant genes and pseudogenes. The intron conservation rate between two paralogous genes was calculated as the ratio of the number of shared introns divided by the total number of intron positions from the protein/intron alignment between two genes (based upon the method of Rogozin et al. (2003)). An event was identified as an RD duplication if the average intron conservation rate to paralogs outside the cluster was below 1/3.

For RD events that do not have Ensembl gene models we lack Ensembl intron predictions. Accordingly, for these RD duplicates I used PseudoPipe intron predictions (Zhang et al., 2006). I filtered out RD events for which more than half of the descendant PseudoPipe predicted pseudogenes are annotated as probable

intron bearing copies.

3.2.3 Event assignment to tree branches

I used parsimony to assign each inferred duplication event to a specific branch of the 5-species tree. I assigned each event to the tree branch corresponding to the exact set of species spanned by the descendant genes of the detected duplication event, which I refer to as assigned events. Intact events are defined as those duplication events that have no apparent disruption (e.g., in-frame stop codons) of the protein coding sequence and an Ensembl annotated gene in each of the species spanned by the cluster.

3.3 Discussions

There is a significant number of gene families composed of tandem duplications, and many earlier gene family evolution reconstructions have been focused on these families (e.g. β-globin family (Patel et al., 2008)). Although using local synteny information to infer the duplication events is effective to discriminate the RD events from DD events, it is largely ineffective inside tandem arrays. The problem of inferring tandem duplication history is hard to solve (Tang et al., 2002; Bertrand & Gascuel, 2005) and many attempts to solve this problem use coding sequence information (e.g. Lajoie et al., 2010); hence the history inferred by these methods might not reflect the ancestral history of the genes within the array.

Other possible considerations include the use of UPGMA hierarchical clustering to identify successive DD events within syntenic clusters. UPGMA assumes a molecular clock but rearrangement events do not happen at the same rates in the different lineages. Therefore, approaches allowing various evolution rates on each lineage, like the neighbor-joining algorithm, might be better suited for inferring

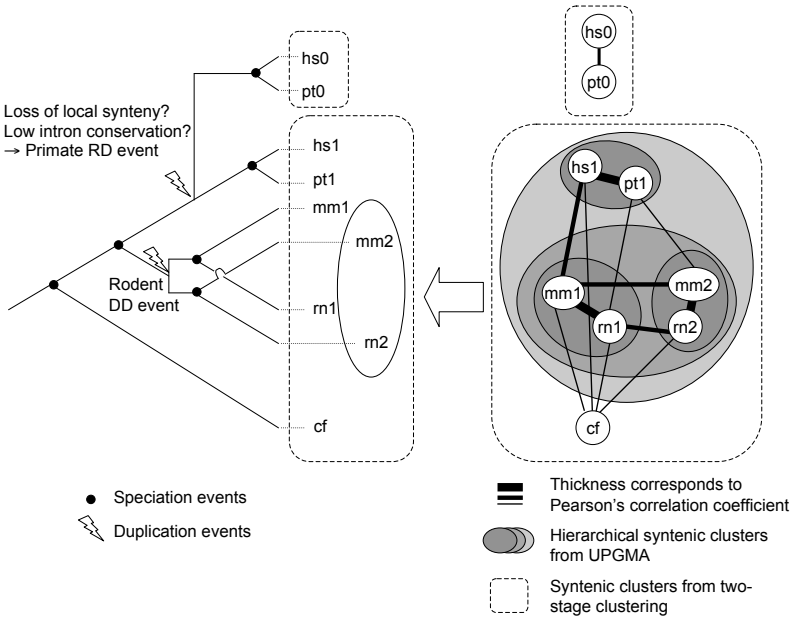


Figure 3.2: Inferring DD and RD events using local synteny and hierarchical clustering. This example shows how DD and RD events are inferred from a superfamily having nine members: two members per each species except for dog, from the results of our clustering algorithms (right) to corresponding events (left). By using two-stage clustering algorithm, two syntenic clusters are formed, shown as hollow rounded rectangles. Loss of introns in one cluster suggests that the loss of synteny was due to an RD event. UPGMA builds hierarchical clusters within each syntenic cluster and speciation and DD events are inferred based on species sets.

DD events. Also instead of our deterministic algorithms, using a probabilistic approach for defining duplication events might give us more information including the reliability of event inference.

Chapter 4

Application to Mammalian Gene Families

In an attempt to understand gene family evolutionary history, many gene families have been manually or semi-automatically curated. For example, mammalian β-globin gene family has been updated with evidence revealed by more advanced methods and new sequence data (Cooper et al., 2005; Patel et al., 2008; Wheeler et al., 2004, 2001). Our goal in this work is to develop a comprehensive method for inferring gene family evolutionary histories from genome sequence data.

In this chapter I use non-coding sequence information and pseudogene information to infer the evolutionary histories of gene families in five mammalian genomes: human, chimp, mouse, rat, and dog. After inferring these histories I compared the rates of DNA- and RNA-mediated duplication events (4.2 and 4.3), and the preservation and functionalization events (4.4 and 4.7) within each tree.

I also explored several hypotheses regarding the likely genetic events governing the fate of newly duplicated genes. In order to test if the RNA-mediated copies could coopt the existing regulatory elements I measured the distance to the nearest

[1]The results presented in this chapter are published in Jun et al. (2009b)

functional genes (4.5). Similarly, for DNA-mediated copies, I measured the degree of selective constraint on duplicates with disrupted or intact flanking sequences (4.6). Finally, I examined functional biases within gene families generated by RNA or DNA mediated duplication events (4.8 and 4.9).

4.1 Methods

4.1.1 Event detection, assignment and evidence of function

I applied the duplication event detection and assignment method, described in Chapter 3, to five mammalian genomes: *Homo sapiens* (human), *Pan troglodytes* (chimp), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Canis familiaris* (dog). Ensembl protein family annotations served as a starting point for our analysis, obtained from Ensembl (release 37) (Hubbard et al., 2005).

Functional events are defined by the clusters of putative protein coding genes with average dN/dS ratio below 0.5 over all pairs of genes within the cluster. In order to avoid the disruption by subsequent DD events, I computed average dN, dS, and dN/dS measures for the descendants of a duplication event assigned to branch b considering only the pairs of genes and pseudogenes coming from different lineages rooted at b. Pairwise dN and dS measures were counted using the `YN00` program of `PAML` (Yang, 1997) with correcting for multiple substitutions at the same site (Yang & Nielsen, 2000).

4.1.2 Determining RD integration site relative to genes and IPSs

For RD genes and processed pseudogenes located in intergenic regions, the distances from the RD copies to the nearest upstream/downstream Ensembl genes

on the same/opposite strand were measured. The distance was defined by the difference between the starting point of predicted coding sequence (intact or interrupted) and the starting point of the coding sequence of neighboring Ensembl genes.

Also, I compared the numbers of RD genes and processed pseudogenes located on genes and on IPS (Indel Purified Sequence) data (Lunter et al., 2006).

4.1.3 Determining rate asymmetry

The normalized rate asymmetry between two genes was defined as: $R = \frac{|(p_1 - p_2)|}{(p_1 + p_2)}$, where p_1 and p_2 are the dS, dN, and dN/dS values associated with each of the duplicates.

4.1.4 Detection of disrupted flanking regions

In order to locate genes with disrupted flanking regions, I examined the syntenic relationship between duplicates and their outgroup gene. I define paralogs as having direct synteny if the gene immediately adjacent to each paralog is orthologous to the outgroup gene's neighbor (using the same Blastp criteria as for local synteny). If both genes share direct synteny with the outgroup gene, then conservation of direct synteny is inferred. If one gene shares direct synteny and the other does not, then a disruption is inferred. I excluded cases where neither gene has synteny with the outgroup.

4.1.5 Gene ontology analysis

The correlation between the duplication mechanism and gene families was measured using the `GOstat` web tool (Beissbarth & Speed, 2004) with default parameters. I used the sets of genes from the 10 families most abundant in RD and

DD events, and reported the overrepresented GO terms with a p-value below 0.1. For families with one or two duplication events I selected mid-size families having between 7 and 17 Ensembl annotated genes and RD-only or DD-only events (594 RD-only families and 250 DD-only families). I performed the GOstat analysis for biological process terms on this list, using a minimal GO path length of 5.

4.2 Lineage distribution of duplication events

Over all five species, there were 17,341 Ensembl families comprising 113,543 genes. Excluding families with members on unassembled contigs (no reliable synteny information) and families with more than 50 Ensembl genes (due to the excessive computation time required to generate multiple alignments) resulted in 8,872 gene families containing 53,733 genes. By using PseudoPipe (Zhang et al., 2006), 17,226 pseudogenes (14,189 processed pseudogenes and 3,037 non-processed pseudogenes) were detected and augmented to the corresponding families.

By applying two-stage clustering algorithm (Figure 3.1), 3,018 clusters out of 27,869 clusters were ambiguous and were not considered as RD events. To detect DD events, UPGMA clustering algorithm is used on the Pearson's correlation coefficient between the vectors of local synteny, as described in the previous chapter (Figure 3.2). Remaining inner nodes are ambiguous between speciation and DD events followed by gene loss and were disregarded. In total, 39,673 inner tree nodes were classified as speciation events, 2,035 as DD events, and 1,642 were ambiguous.

Events giving rise to clusters of genes with no conservation of synteny relative to "parental" genes and low inter-cluster intron conservation rates were classified as RD events, while events giving rise to clusters of genes with high local synteny to parental genes were classified as DD events. Events corresponding to gene clusters with indeterminate intron conservation or local synteny to parental genes

were classified as ambiguous. This analysis resulted in the classification of a total of 2,035 DD events, 12,507 RD events, and 2,742 ambiguous events. Using parsimony to assign non-ambiguous events to branches of the species tree resulted in 52 DD and 45 RD events on the branch leading to primates and rodents (the in-group), 161 DD and 1,782 RD events on the primate branch leading to humans and chimps, and 88 DD and 522 RD events on the rodent branch leading to mice and rats (Figure 4.1). Gene duplication events for the root and terminal branches of the tree were also counted, but not used for further analysis due to the difficulty in estimating the degree of purifying selection on very recent duplication on the terminal branches, and the age of duplications at the root. A number of 386 DD and 429 RD events could not be reliably assigned to specific branches of the tree using parsimony and were also omitted from further analysis. Duplication event counts on the three internal branches of the tree reveal an excess of RD events over DD events along all but the deepest branches of the tree, suggesting an average rate of RD copy formation 3–10 times higher than that of DD copy formation (Figure 4.1). Deviation from this ratio along the in-group branch may be the result of a period of relative inactivity of retrotransposition compounded with the difficulty of detecting the products of old RD events not under purifying selective pressure (Marques et al., 2005).

4.3 Rates of duplication

Rates of retrotransposition vary significantly over time and bursts of retrotransposition have been reported in several mammalian lineages (Marques et al., 2005; Zhang et al., 2004). The synonymous substitution rate (dS) profiles of the duplicates identified in this study (Figure 4.2) are shaped by the rate of generation of new duplicates, the mutation rates along each lineage, the age of the genes identi-

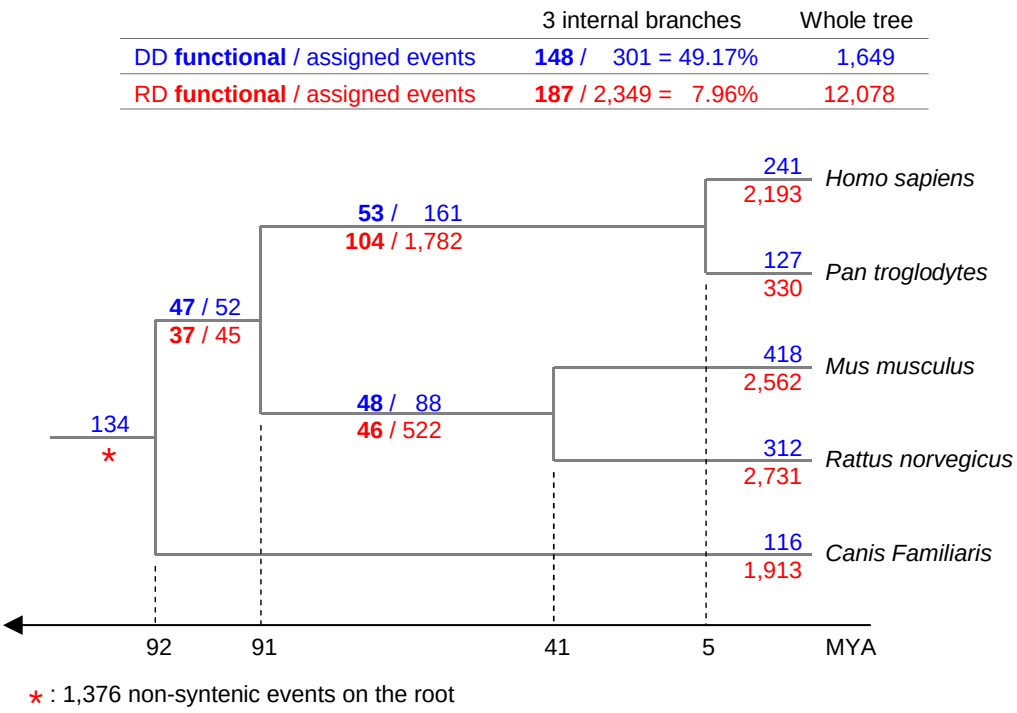


Figure 4.1: Numbers of gene duplication events from DNA-mediated duplication (blue numbers above the line) and RNA-mediated duplication (red numbers below the line). Numbers represent the assigned DD or RD events on each branch. Numbers typeset in bold on three internal branches are counts of functional events, defined in this study as intact events that yield clusters with average dN/dS ratio below 0.5 over pairs of homologous Ensembl genes. For three internal branches, fractions of the functional events over the total assigned events are shown, e.g., **53**/161 for DD events on primate branch. Evolutionary ages are based on (Ureta-Vidal et al., 2003).

fied in each interval, and our ability to identify genes uniformly along each lineage. Pseudogenes, for instance, become increasingly difficult to identify as they get older and diverge from their original sequence. RD events in all three internal branches show clear peaks in dS (Figure 4.2A). For duplications occurring on the primate branch this peak occurs around dS=0.1, while in rodents it occurs around dS=0.3 and in in-groups around dS=0.6∼0.8. This pattern is consistent with bursts of retrotransposition in each of these lineages, a high mutation rate in the rodent lineage, and the 36–Myr gap between the speciation events leading to rodent and primate lineages. Duplications occurring prior to the rodent/primate split display

a dS distribution significantly shifted toward higher dS values, consistent with the greater age of these duplicates.

DD events show a similar distribution in dS but a more uniform distribution of dS values than RD duplicates (Figure 4.2B and C), suggesting that DNA-mediated duplication is a more uniform process that occurs at less variable rates than retrotransposition. It is interesting to note that the inferred age distribution of DD events is more uniform than that of the RD duplicates but is not perfectly flat, suggesting that there may be some variation in the rate of DD events over evolutionary time. It has been suggested and accepted that non-allelic homologous recombination (NAHR) are mediated by pre-existing repeats, such as Alu elements (Bailey et al., 2003; Kim et al., 2008).

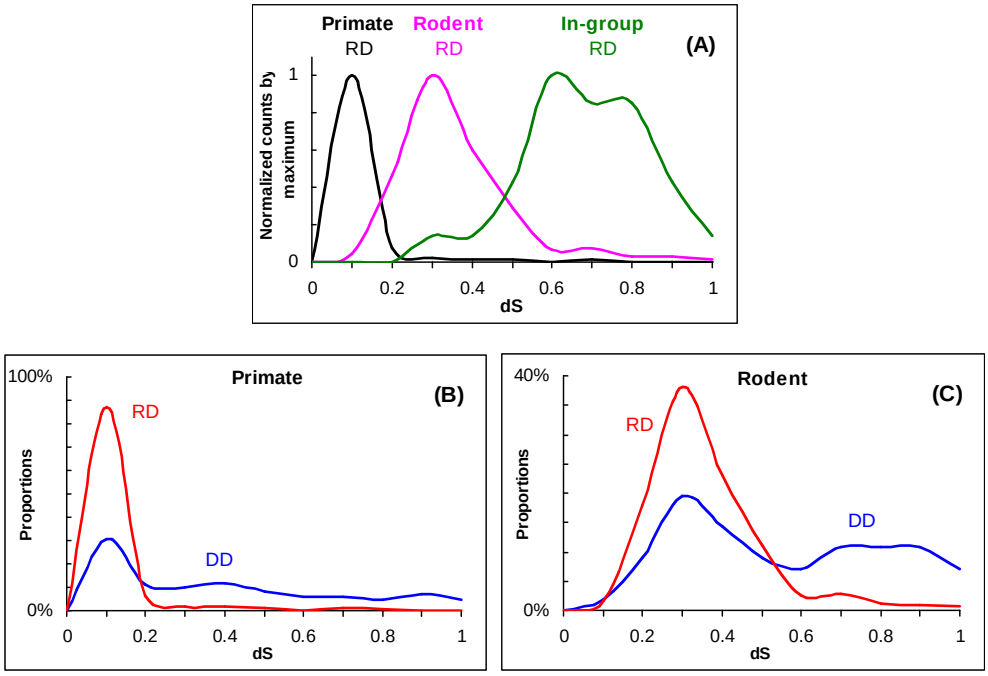


Figure 4.2: Histograms of average dS over pairs of Ensembl genes and pseudogenes. (A) For clusters resulting from RD events on the primate, rodent, and the in-group branch leading to primates and rodents. (B) For clusters resulting from DD events and RD events on the primate lineages. (C) On the rodent lineages.

4.4 Preservation rates of functional duplicate copies

The progressive influence of purifying selection over time is readily observed in dN vs. dS plots of DD and RD events from the three internal lineages. In Figure 4.3, RD events on the primate branch are compressed near the origin, consistent with a recent burst of retrotransposition in this lineage, while DD events display a more even age distribution. Non-intact events are generally interspersed with intact ones, except for the rodent lineage in which the effect of prolonged purifying selection results in the separation of inactivated RD events from the intact RD copies. In the in-group there are very few inactivated duplication events due to the difficulty in finding very old pseudogenes. However, some inactivated events remain interspersed among intact ones, suggesting that the gene duplicates resulting from these events were under purifying selection and may have only recently suffered interruption of their protein coding regions. Alternatively, these genes may encode partial protein products that remain under purifying selection. It is probable that young duplicate genes may escape inactivation for some time despite lacking any apparent function. Since Ensembl gene predictions rely upon the presence of an intact coding region rather than any evidence of selection pressure upon the sequence, the gene clusters resulting from intact duplication events should be comprised of both functional genes and duplicates that are not functional but have escaped inactivation. Evidence of purifying selection is often used as evidence for function, and the ratio of synonymous to non-synonymous changes (dN/dS) in the protein-coding region of a gene is a convenient way of estimating this selective pressure (Nekrutenko et al., 2002). For example, dN/dS ratio$<$0.5 has been used as stringent functionality criteria between retrotransposed genes and their parental genes (Emerson et al., 2004). Also Torrents et al. (2003) showed that there is a clear discrimination between dN/dS ratios of pseudogenes and those of functional genes,

supporting the use of dN/dS ratios as evidence of function. Here I compute dN/dS ratios between all pairs of descendants from each duplication event. This pairwise approach is computationally rapid, is independent of precise reconstruction of the entire gene tree, and allows for the detection of functionalized descendant clusters of a duplication event that are not constrained relative to the parental genes.

Analysis of the dN/dS ratios of clusters derived from duplication events is quite revealing. Figure 4.4A compares clusters of RD duplication event descendants with intact protein coding reading frames (intact) and clusters of RD duplicates with inactivated reading frames (inactivated). Aggregate dN/dS values of a significant portion of intact clusters overlap with the dN/dS values of inactivated clusters in the region of the graph where dN/dS is greater than $\sim$0.5. Assuming that the vast majority of inactivated clusters (clusters whose members have inactivating mutations in their protein coding regions) are not under purifying selection for protein coding function, those intact clusters that fall into this range are unlikely to encode functional proteins, despite lacking any clearly inactivating mutation. By inference, those clusters that display significantly lower aggregate dN/dS values ($<$0.5) are likely to be under stabilizing selection for protein coding function.

Panels B–D of Figure 4.4 compare dN/dS values of duplicate clusters derived from RD and DD events on each of the three internal branches of the mammalian tree. In the oldest internal branch of the tree (ingroup) very few clusters generated by either duplication mechanism can be detected that are not under some degree of purifying selection pressure. This is probably due to the difficulty in identifying very old non-functional sequences. Such sequences are expected to drift away from their parental sequence making identification increasingly difficult with advanced age. Clusters derived from duplication events along the rodent branch have a bi-modal distribution of dN/dS ratio resulting from RD and DD events that gave rise to putatively functional gene copies (aggregate dN/dS values$<$0.5), and clusters

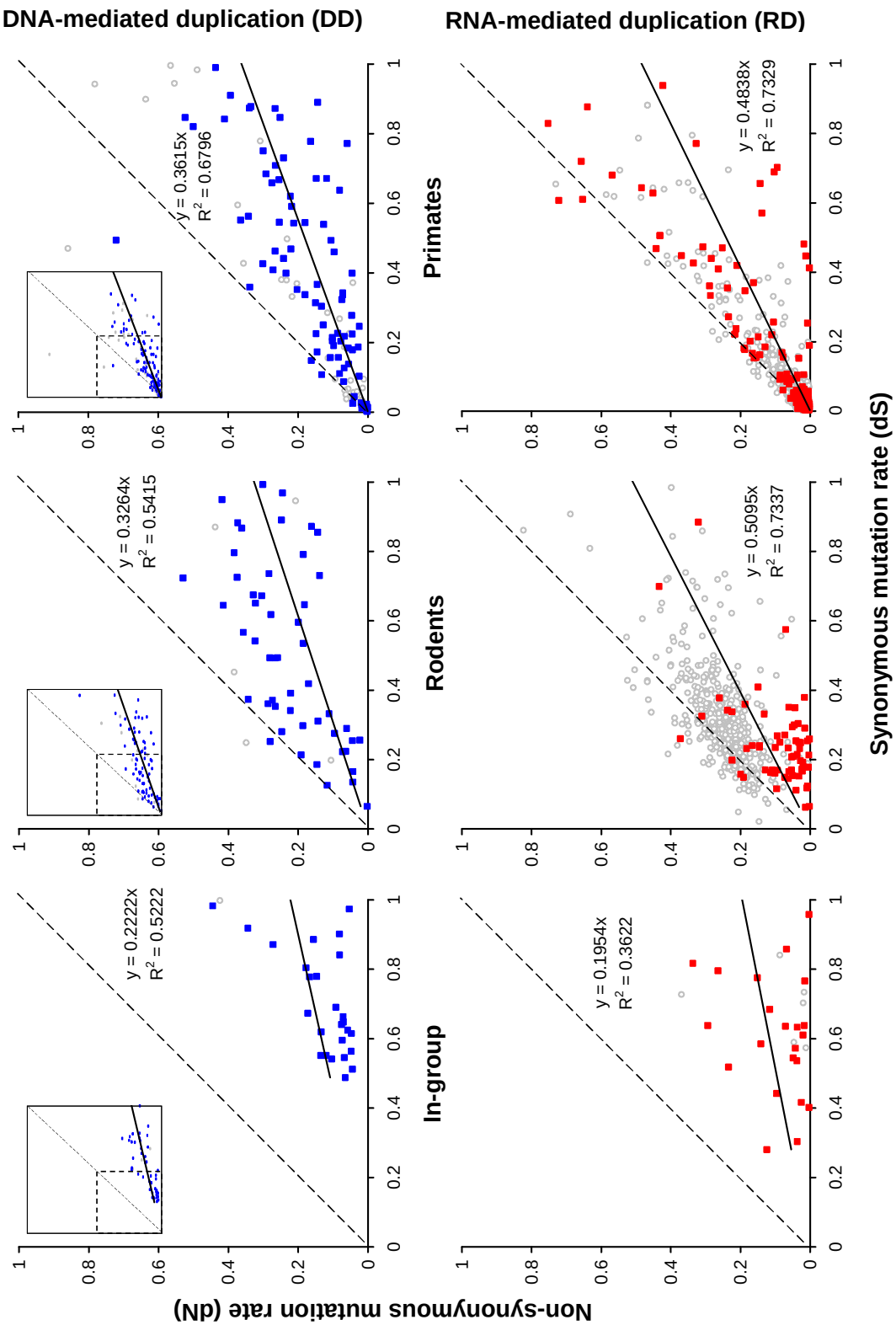


Figure 4.3: Dot-plots of averaged dN and dS over all the pairs of genes and pseudogenes within clusters resulting from DD intact events (blue dots in the upper three graphs), DD inactivated events (gray dots in the upper three graphs), RD intact events (red dots in the lower three graphs), and RD inactivated events (gray dots in the lower three graphs) for three internal branches: in-group, primate and rodent branches. The dashed lines on the dot-plots are the line with dN=dS (or dN/dS=1) and the solid lines are the trend lines of the intact events per each panel.

with no clear evidence of stabilizing selective pressure. Duplication events along the primate branch gave rise to clusters with more uniformly distributed aggregate dN/dS values spanning the entire range of measurements. This is likely to be a reflection of the relatively short period of time these new genes have been under purifying selection and is consistent with the relatively low dS values of duplicates detected along this branch (Figure 4.2B).

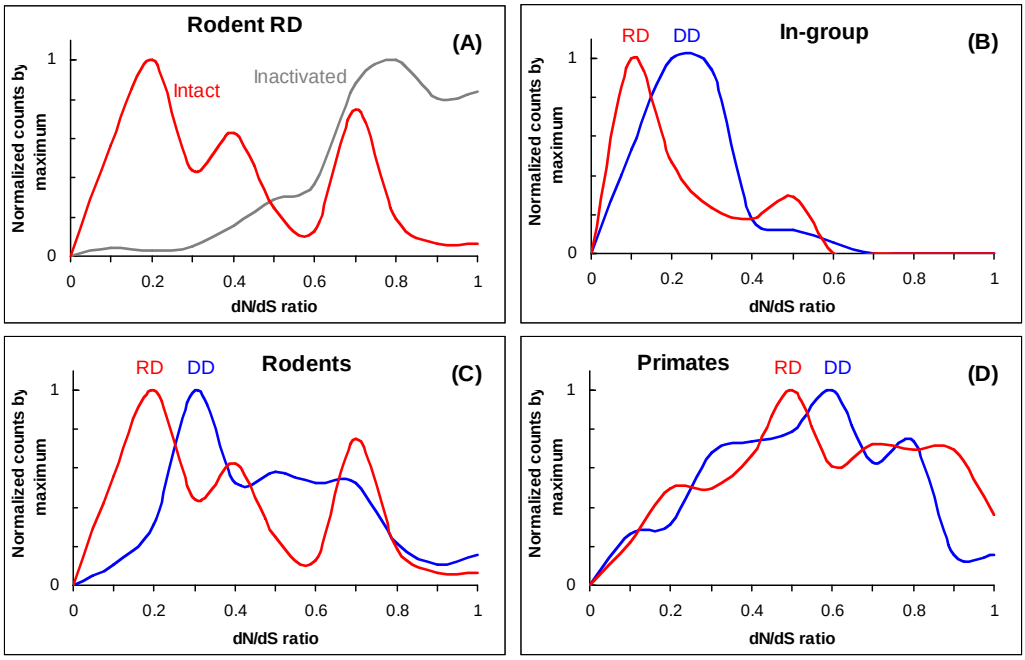


Figure 4.4: (A) Histograms of average dN/dS ratio over pairs of Ensembl genes for clusters resulting from intact RD events and average dN/dS ratio over pairs of genes and pseudogenes for clusters resulting from inactivated RD events on the rodent lineage. Histograms of average dN/dS ratio over pairs of Ensembl genes for clusters resulting from intact DD events and RD events on the in-group branch leading to primates and rodents (B), rodent (C), and primate (D).

4.5 Relative position of RD copies to the other genes

In order to address the question of why some RD events are under the stabilizing selective pressure (dN/dS ratio ≤ 0.5), I analyzed the location of RD copies. I categorized RD copies as either genic (mostly intronic) or intergenic, and measured the distance to the nearest Ensembl genes. Three classes of RD copies were defined: intact RD genes with dN/dS ≤ 0.5 (SS-RD), intact RD genes with dN/dS>0.5 (NI-RD) and processed pseudogenes (PP-RD).

Table 4.1 shows the number and percentage of RT copies on genic and intergenic sequences and corresponding p-values (χ^2 test by using the proportion of genic area in the whole genome). Although all the three classes of RD copies were less likely to be found inside other genes (all p-values<1e-06), processed pseudogenes (PP-RD) were found inside other genes two or three times more often than intact RD copies (SS-RD and NI-RD).

Numbers (%)		**SS-RD**	**NI-RD**	**PP-RD**
Strand dependent	Genic	23 (4.3%)	23 (4.5%)	432 (11.2%)
	intergenic	510 (95.7%)	485 (95.5%)	3,411 (88.8%)
	p-value$*$	5.90e-08	3.23e-07	2.58e-07
Strand independent	Genic	70 (13.1%)	73 (14.4%)	996 (25.9%)
	intergenic	463 (86.9%)	435 (85.6%)	2,847 (74.1%)
	p-value$*$	1.14e-11	7.28e-10	8.55e-07

Table 4.1: Numbers of retrotransposed insertions on genic versus intergenic sequence by RD events on three internal branches: in-group, primates, and rodents. Three classes of RD copies were used: intact RD genes with dN/dS ≤ 0.5 (SS-RD), intact RD genes with dN/dS>0.5 (NI-RD) and processed pseudogenes (PP-RD). $*$: χ^2 test p-value using the proportion of genic area in the whole genome.

For RD genes and processed pseudogenes located in intergenic regions, the distances from the RD copies to the nearest upstream/downstream Ensembl genes on

the same/opposite strand were measured. For intergenic RD copies SS-RD were consistently closer to neighboring genes than PP-RD regardless of strand or orientation (up- or down-stream, all p-values<0.001, Wilcoxon/Kruskal-Wallis test and Student's t-test, Figure 4.5). This is consistent with retrotransposed duplicates co-opting preexisting regulatory elements and the founding in Vinckenbosch et al. (2006). Not surprisingly, distances to upstream neighboring genes on the opposite strand (in a head-to-head configuration) have the shortest median distance to intact RD copies (95.8KB for SS-RD and 138.7KB for NI-RD; Figure 4.5C). This results are also somewhat consistent with previous works by Trinklein et al. (2004), Michalak (2008) and others. They showed that coexpressed gene clusters are formed through local sharing regulatory elements such as transcription factors, promoters, and enhancers. For example, in the human genome, more than 10% of genes form head-to-head pairs that may be subject to bidirectional expression mediated by common promoter sequences.

4.6 Disruptions in flanking regions are associated with greater asymmetry in dN and relaxed selective constraint[1]

For this experiment, we use a different method to define the duplicate events and the triples, two paralogs and ortholog as outgroup. See Jun et al. (2009c); Ryvkin et al. (2008) for the details.

We test the hypothesis that disruptions in the intergenic DNA surrounding DD duplicates would correspond to changes in the course of that duplicate's protein evolution. We define paralogs as having direct upstream or downstream synteny if

[1]The data presented here is obtained by separate methods published in Jun et al. (2009c); Ryvkin et al. (2008)

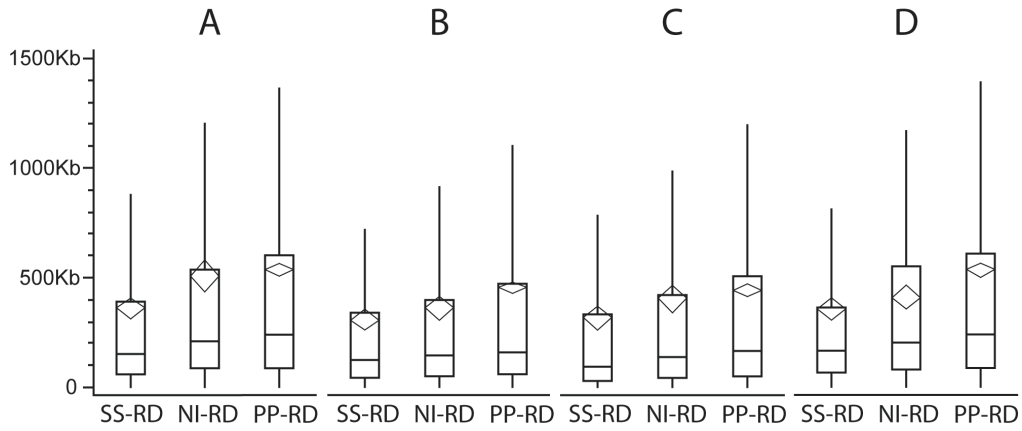


Figure 4.5: Relative location of RD events to Ensembl genes. Three classes of RD copies were used: intact RD genes with dN/dS $\leq$ 0.5 (SS-RD), intact RD genes with dN/dS>0.5 (NI-RD) and processed pseudogenes (PP-RD). (A) Distance to the nearest upstream gene on the same strand. (B) Distance to the nearest downstream gene on the same strand. (C) Distance to the nearest upstream gene on the opposite strand (head-to-head). (D) Distance to the nearest downstream gene on the opposite strand (tail-to-tail). Boxes represent interquartile range with the horizontal line being the median; diamonds span a 95% confidence interval around the mean assuming normality. The vertical lines span the extents of 95% of a normal distribution fit to the data. Three classes have significant different distance measures (p-value< 0.0001 for A,C,D; p-value<0.001 for B; Wilcoxon/Kruskal-Wallis Test). SS-RD and PP-RD have significantly different mean (p-value<0.0001 for A,D; p-value<0.001 for B,C, Student's t-test).

the gene immediately adjacent to each paralog is orthologous. If both genes share direct synteny with the outgroup gene, then conservation of synteny is inferred. If one gene shares synteny and the other does not, then a disruption is inferred. If neither gene has synteny with the outgroup, then there are two possible scenarios: either both duplicates experienced rearrangement resulting in a disruption of synteny, or the original gene experienced a disruption prior to the duplication event. We do not attempt to distinguish between these two scenarios, and accordingly we excluded the duplications where both genes lack any direct synteny with the outgroup.

Table 4.2 shows that distant DNA duplicates are more likely to be evolving asymmetrically at the protein and DNA levels when one of the genes has lost

direct synteny with the outgroup. This association is significant regardless of whether the disruption occurred upstream or downstream of the duplicate. However, the probability of observing asymmetric constraint (dN/dS), is only significantly higher for duplicates that experience upstream disruptions (p-value=0.021 vs. p-value=0.113). This suggests that changes in the 5' flanking DNA of a gene may have a greater impact on that gene's functional importance than changes in the 3' flanking DNA.

		Frequency of asymmetry		
Duplication type	N	dN	dS	dN/dS
Distant, 5' syntenic	22	50%	50%	9%
Distant, 5' disrupted	56	79%*	73%*	34%*
Distant, 3' syntenic	21	47%	38%	14%
Distant, 3' disrupted	65	78%**	80%***	$31\%^{n}$
Tandem, 5' syntenic	37	54%	43%	8%
Tandem, 5' disrupted	71	$65\%^{n}$	$49\%^{n}$	$11\%^{n}$
Tandem, 3' syntenic	35	54%	43%	9%
Tandem, 5' disrupted	78	$68\%^{n}$	$51\%^{n}$	$12\%^{n}$

Table 4.2: Frequencies of asymmetry in non-synonymous and synonymous substitution rates (dN and dS) and selective constraint (dN/dS) on DD copies by disruption of direct synteny. Significance was established via Fisher's exact test: *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$; ^{n}, not significant.

4.7 Distribution of duplication events within the mammalian tree

The total number of RD and DD duplication events detected in this study is illustrated in Figure 4.1. Along each branch the number of events giving rise to clusters with evidence of purifying selective pressure on their protein coding regions is in bold typeset, while the total number of events detected is in denominators, showing a fraction of the functional events over the total assigned events, e.g.,

53/161 for DD events on primate branch. From these numbers, it is clear that that I detect far more RD events than DD events, but that far fewer of these events give rise to functional protein coding genes than their DD counterparts. Analysis of the internal branches individually reveals small possible differences in the relative probability of these events giving rise to functional genes in different lineages. In the most basal branch shared by rodents and primates, there is a slight excess of functional DD events over functional RD events, while the two mechanisms appear to contribute equal numbers of functional events in the rodent lineage. The primate and rodent branches show similar rates of assigned DD events, but in primates fewer of these events give rise to functional descendants (Table 4.3). A decreased rate of functionalization is also apparent in the RD events on the primate lineage. Despite an RD event rate nearly twice that seen in rodents, the number of functional RD events in primates is only ∼25% greater than that in rodents.

	DD events			**RD events**		
	Assigned	Intact	Functional	Assigned	Intact	Functional
Rodents	1.76	1.56	0.96	10.4	1.42	0.92
Primates	1.87	1.31	0.62	20.7	3.41	1.21

Table 4.3: Rates of duplication events (per million years) for rodent and primate lineages

4.8 Distribution of functional events in gene families

Of 8,872 Ensembl families, 262 families have at least one functional duplication event. The distribution of functional events within gene families of different sizes

(Figure 4.6) reveals that there is an apparent lack of mixing of the two types of duplication events within families.

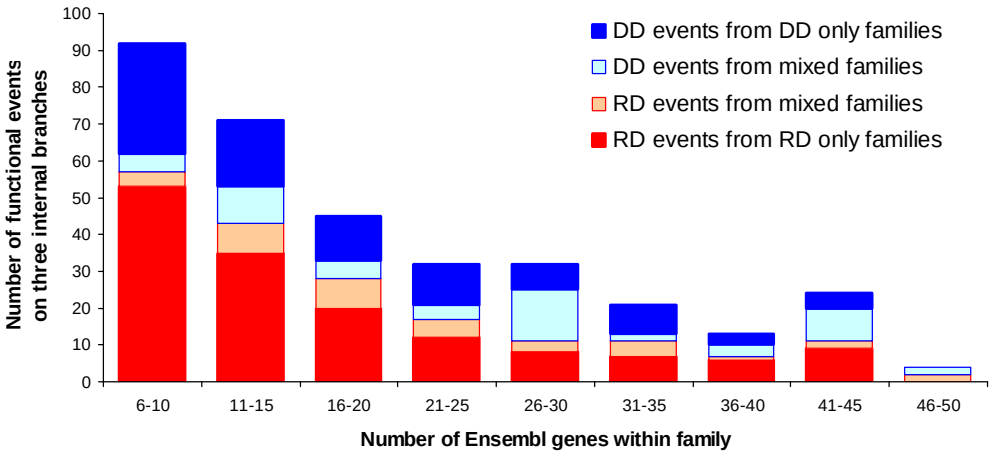


Figure 4.6: Distribution of functional events categorized by RD events in RD-only families, DD events in DD-only families, and RD events and DD events in the families including both events binned by the number of Ensembl genes within family.

Of the 262 families containing at least one functional event, 151 (57.6%) families have exclusively functional RD events and 101 (38.5%) have exclusively functional DD events, suggesting that gene families primarily evolve either by DNA-mediated duplication, or by RNA-mediated duplication (retrotransposition), but rarely by both mechanisms. I tested this idea by looking more closely at the distribution of functional RD and DD events in families with at most 4 functional events. I counted families that contained only DD or only RD functional events and compared these observations to the numbers of families expected based on a binomial distribution of the of two types of events, where the probability of DD functional events is $p = 148/335$ (total number of functional DD events/total number of functional events) and the probability of RD functional events is $q = 1 - p$ (Table 4.4). Chi-squared tests reveal that the observed mutual exclusivity of these two mechanisms is statistically highly significant.

	No. of events			
	2	**3**	**4**	**>4**
DD only families	14	4	3	1
RD only families	16	1	2	0
Total families	37	7	6	1
P-values	<0.01	<0.001	<1e-6	N/A

Table 4.4: Observed numbers of families having only RD functional events and only DD functional events and χ^2 p-values.

4.9 DNA- and RNA-mediated duplications give rise to different types of gene families

As the mechanisms that give rise to DNA- and RNA-mediated duplications have very different consequences for the properties of the duplicate copy, it might be reasonable to expect that each mechanism might be biased in the type of new functional genes it creates. I examined the types of genes created by each duplication mechanism by identifying Gene Ontology terms that are overrepresented in families derived from each mechanism. As it might be expected, the largest gene families generated through RD duplications are dominated by ribosomal proteins (Table 4.5. The very high levels of expression of these genes lead to a large number of intact RD events and a correspondingly large number of new gene copies. In contrast, the largest families of DD generated genes include a variety of functional categories including immune function (lipocalins, chemokines, and defensins), and large families of diverse molecules such as the olfactory receptors (Table 4.6).

Analysis of the types of genes born through each mechanism in the mid-size families (7–17 members) show a similar trend. Overrepresented GO terms for RD dominated families again include a number of RNA related categories. However, when the contribution of RD genes is analyzed based on individual families, a much

Ensembl family	Family description	No. of Ensembl genes	No. of intact RD events	No. of assigned RD events	GO	GO name
ENSF00000001265	60S RIBOSOMAL L27A	42	30	112	GO:0005842, GO:0005830, GO:0015934	Cytosolic large ribosomal subunit (sensu Eukaryota)
ENSF00000001824	60S RIBOSOMAL L36	39	27	104	GO:0005840, GO:0030529, GO:0006412	Ribosome
ENSF00000001192	40S RIBOSOMAL SA P40 34/67 KDA LAMININ	44	24	193	GO:0015935, GO:0005055	Small ribosomal subunit, laminin receptor activity
ENSF00000001687	60S RIBOSOMAL L13	30	23	85	GO:0005830, GO:0044445	Cytosolic ribosome (sensu Eukaryota)
ENSF00000001631	40S RIBOSOMAL S10	37	21	90	GO:0016283, GO:0005843	Eukaryotic 48S initiation complex
ENSF00000002263	40S RIBOSOMAL S12	34	20	95	GO:0016283, GO:0005843	Eukaryotic 48S initiation complex
ENSF00000001209	60S ACIDIC RIBOSOMAL P0	29	19	53	GO:0005842, GO:0006414, GO:0005830, GO:0042254, GO:0043284, GO:0015934	Cytosolic large ribosomal subunit (sensu Eukaryota)
ENSF00000001521	60S RIBOSOMAL L12	33	19	97	GO:0005842, GO:0005830, GO:0015934	Cytosolic large ribosomal subunit (sensu Eukaryota)
ENSF00000002778	60S RIBOSOMAL L35A	25	18	169	GO:0000049	tRNA binding
ENSF00000001964	40S RIBOSOMAL S7	30	18	76	GO:0016283, GO:0005843	Eukaryotic 48S initiation complex

Table 4.5: Top-10 RD-abundant families ordered in decreasing order of the number of intact RD events. All families here have no DD events and at least 18 RD events.

Ensembl family	Family description	No. of Ensembl genes	No. of intact DD events	No. of assigned DD events	GO	GO name
ENSF00000001153	Lipocalin-related protein and Bos/Can/Equ allergen	49	34	44	N/A	N/A
ENSF00000000477	Trace amine associated receptor/tar	42	22	28	GO:0001584, GO:0008227	Rhodopsin-like receptor activity, amine receptor activity
ENSF00000000443	Carboxylesterase precursor EC_3.1.1.1	45	20	28	GO:0004759, GO:0004091	Carboxylesterase activity; serine esterase activity
ENSF00000000958	Small chemokine	34	14	16	GO:0008009, GO:0042379	Chemokine receptor binding; chemokine activity
ENSF00000002170	Proteinase inhibitor I25, cystatin	27	14	15	GO:0004869, GO:0004866	Endopeptidase inhibitor activity; cysteine protease inhibitor activity
ENSF00000000691	Sialic acid binding IG lectin precursor siglec	41	13	22	GO:0016021, GO:0031224	Membrane part; intrinsic to membrane
ENSF00000001948	Apolipoprotein	25	13	18	GO:0042157, GO:0006869	Lipoprotein metabolic process, lipid transport
ENSF00000001951	Cystatin precursor	28	12	16	GO:0004869, GO:0004866	Endopeptidase inhibitor activity; cysteine protease inhibitor activity
ENSF00000002079	Olfactory receptor 8S1	27	12	14	GO:0007608, GO:0007606, GO:0004984	Sensory perception of chemical stimulus; sensory perception of smell, olfactory receptor activity
ENSF00000001990	Defensin related cryptdin precursor	19	11	15	GO:0042742, GO:0009617	Response to bacterium; defense response to bacterium

Table 4.6: Top-10 DD-abundant families ordered in decreasing order of the number of intact DD event. All families here have more than 10 DD events.

greater diversity of functions is revealed including members of the nuclear pore complex, topoisomerase, DNA binding proteins, cell cycle regulation, apoptosis, and energy metabolism. DD dominated mid-size families are involved in a variety of processes requiring more complex regulation of gene expression including the regulation of development, immune processes and odorant perception.

The RD/DD abundant families are from the long tail of the distribution, but the majority of RD/DD events are in the mid sizes of families. I listed the over-represented (p-value<0.1) biological process GO terms (Top Tens) from RD event only families with sizes of [7–17] (594 families, 6,052 genes, 736 annotated Hs gene names, 767 annotated Mm gene names), with minimal length of GO path as 5. Most of the overrepresented GO terms are overlapped between human and mouse, while all GO terms are ribosomal or RNA related ones (see Table 4.7). Over-represented (p-value<0.1) biological process GO terms (Top 10s) from DD event only families with sizes of [7–17] (250 families, 2,668 genes, 340 annotated Hs gene names,379 annotated Mm gene names), with minimal length of GO path as 5, show different results: there is no overlapping between two GO term lists from human genes and mouse genes. The overrepresented GO terms from human genes are mostly development related ones, while mouse genes are associated with sensory ones (see Table 4.8).

4.10 Application to Ribosomal protein families

Since local synteny is more effective to distinguish RD events from DD events, I applied the same method to the 79 ribosomal protein families where RD events are absolutely abundant. Eight mammalian and one outgroup genomes are considered: human, chimp, macaque, mouse, rat, dog, cow, opossum, and chicken as an outgroup. I found that one of riboprotein families (RPS28) does not include

GO term	p-value (Hs)	p-value (Mm)	GO name
GO:0016071	7.87e-41	4.4e-20	mRNA metabolic process
GO:0006397	8.57e-40	1.74e-21	mRNA processing
GO:0022613	7.24e-38	2.08e-11	Ribonucleoprotein complex biogenesis and assembly
GO:0006396	1e-37	3.31e-19	RNA processing
GO:0006605	3.17e-21		Protein targeting
GO:0006139	3.82e-18	7.26e-14	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process
GO:0051246	6.12e-18		Regulation of protein metabolic process
GO:0043170	5.39e-16	3.37e-14	Macromolecule metabolic process
GO:0008380	7.37e-16	1.39e-23	RNA splicing
GO:0044249	1.05e-14		Cellular biosynthetic process
GO:0000377		2.91e-08	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0000398		2.91e-08	Nuclear mRNA splicing, via spliceosome
GO:0000375		2.91e-08	RNA splicing, via transesterification reactions

Table 4.7: Top-10 GO terms from GOstat analysis on the human and mouse genes from RD-only families with size between 7 and 17.

human member in Ensembl rel. 50, thus I used cross-species Pseudopipe to detect it including other possible unitary pseudogenes similar method used in Zhang et al. (2010).

To reconstruct gene family history we need to build a gene tree based on detected duplication events. Since the duplication event detection method described in Chapter 3 produces the DD event-induced gene trees separated by RD events, we need to put these trees together into one tree. Considering possible multiple DD events on the same branches, there might be combinatorial possibilities to join these syntenic trees into one tree. Since local synteny or intron conservation information are defaulted between these trees, i.e. no conservation between any member across trees, I have to use the sequence similarity to deal with the problem.

GO term	p-value (Hs)	p-value (Mm)	GO name
GO:0006954	0.00042		Inflammatory response
GO:0048731	0.000424		System development
GO:0009611	0.000654		Response to wounding
GO:0048513	0.000772		Organ development
GO:0048518	0.00473		Positive regulation of biological process
GO:0016998	0.00473		Cell wall catabolic process
GO:0045682	0.0095		Regulation of epidermis development
GO:0000165	0.0104		MAPKKK cascade
GO:0008366	0.0128		Axon ensheathment
GO:0007272	0.0128		Ensheathment of neurons
GO:0009617		0.00204	Response to bacterium
GO:0050906		0.00204	Detection of stimulus during sensory perception
GO:0001580		0.00209	Detection of chemical stimulus during sensory perception of bitter taste
GO:0050912		0.00209	Detection of chemical stimulus during sensory perception of taste
GO:0050907		0.00244	Detection of chemical stimulus during sensory perception
GO:0050913		0.00276	Sensory perception of bitter taste
GO:0042742		0.00276	Defense response to bacterium
GO:0000188		0.00378	Inactivation of MAPK activity
GO:0009593		0.00576	Detection of chemical stimulus
GO:0009435		0.0119	NAD biosynthetic process

Table 4.8: Top-10 GO terms from GOstat analysis on the human and mouse genes from DD-only families with size between 7 and 17.

The best averaged similarity measure is used to place the root node of smaller tree to a proper branch of a bigger tree. Blastp scores are used between intact genes, and pseudogenes are assigned by their seed genes.

Figure 4.7 is gene family history of RPL 36A family generated by duplication events detection and assignment methods. Interestingly, even RD events seem uniformly distributed over the species and timeframe, the oldest RD event placed

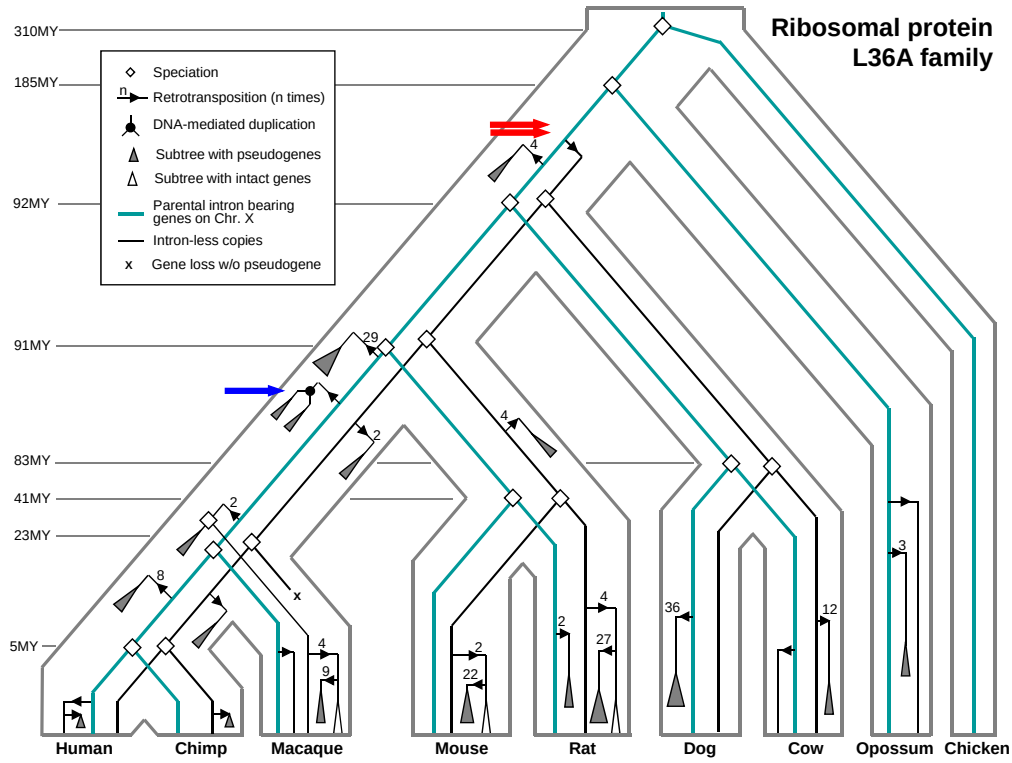


Figure 4.7: Evolution history of ribosomal protein L36A family including pseudogenes and duplication mechanism. Local synteny infers that nine ENSEMBL annotated ribosomal genes (one in each species) evolved from one ancestral intron-bearing gene. This ancestral gene gave rise to multiple retrotransposition events at various evolutionary time points as indicated by black lines with arrowheads. Some of these retro-copies were subject to subsequent DD events (blue arrow on primate lineage), while others were pseudogenized (grey deltas). A retrocopy generated from one of these events (at the base of the mammalian lineage on the branch between the LCA with opossum and the other mammals, double red arrow) is conserved in all descendant species except for macaque (where a different RD copy arising in the primate LCA, red arrow, is preserved). Recent, lineage-specific retrocopies often retain intact ORFs simply because they are relative young (open deltas.)

on the branch between opossum speciation and the other mammalian species is only duplication event having extant copies in all species, except for some young lineage-specific RD copies remaining intact probably due to the lack of time for pseudogenization. A simple dosage model is not adequate for explaining why

this old RD event can survive, because other RD copies formed by comparatively newer RD events become pseudogenes. Dosage balance cases are mostly triggered by WGD (DeLuna et al., 2008) but no WGD event reported near that branch. The intronless copies from this RD event have been experimentally proven to have narrower expression pattern than the original copies (Uechi, 2002), which discards the neo- or sub-functionalization from the possible retention model. I expect this dilemma is not unique to ribosomal protein families. There are only a handful of well-suited examples for each duplicate retention model and the majority of gene families are not explained well (Dharia et al., 2010). Thus a new retention model for gene duplicates is required.

4.11 Conclusions

In this chapter, I compared the rates of new gene formation by DNA-mediated duplication and RNA-mediated duplication in five mammalian genomes. I found that RNA-mediated duplication occurs at a much higher and more variable rate than DNA-mediated duplication, and gives rise to many more duplicated sequences over time (see 4.3). I showed that while the chance of RNA-mediated duplicates becoming functional is much lower than that of their DNA-mediated counterparts, the higher rate of retrotransposition leads to nearly equal contributions of new genes by each mechanism (see 4.2). I also found that functional RNA-mediated duplicates are closer to neighboring genes than non-functional RNA-mediated copies, consistent with cooption of regulatory elements at the site of insertion (see 4.5). Overall new genes derived from DNA and RNA-mediated duplication mechanisms are under similar levels of purifying selective pressure (see 4.4), but have broadly different functions (see 4.9). DNA-mediated duplicates are mainly affected by disruptions on flanking sequences (see 4.6) while the insertion sites of RD events have

an impact on the fates of RNA-mediated copies (see 4.5). RNA-mediated duplication gives rise to a diversity of genes but is dominated by the highly expressed genes of RNA metabolic pathways. DNA-mediated duplication can copy regulatory material along with the protein coding region of the gene and often gives rise to classes of genes whose function are dependent on complex regulatory information. This mechanistic difference may in part explain why I found that mammalian protein families tend to evolve by either one mechanism or the other, but rarely by both (see 4.8).

Although I have had success detecting gene duplication events in five mammalian genomes and building the gene family evolution history of riboprotein families, several challenges remain for future work. First, I have no conclusive way to place RD events on specific branches in the phylogenetic tree. In the case of riboprotein families (see 4.10) where the RD events are dominant, I used protein sequence similarity to place retro duplication events on likely branches and left the precise order of events unresolved. It is possible that we could improve this imperfect solution using synonymous mutation rates and the assumption of the molecular clock. For the pseudogenes from RD events, we can use the method devised by (Chou et al., 2002). It assumes that non-synonymous mutations are selected against until the gene is inactivated; thereafter, mutations at both synonymous and non-synonymous sites accumulate at the neutral mutation rate.

Another complication in our analysis concerns the definition of intron orthology. Intron orthology was defined by relative position on aligned protein sequences. However, unreliable results from multiple sequence alignment methods can lead to incorrect intron orthology determinations. For example, around the exon-intron junction area, the multiple sequence alignment method was unable to exhibit accurate alignment and thus needed a manual adjustment (Csurös et al., 2007, Fig. 1). To avoid this problem, I used a buffer length within which two intron posi-

tions are merged if the distance between two intron positions is smaller than the given threshold (Jun et al., 2008). However, because of insufficient knowledge of exon-intron junction evolution, determining this threshold is not trivial. Another problem with positional intron orthology is the representative transcript issue. I used either the longest transcript or the collapsed gene model as the representative transcript. However, the longest one transcript model cannot deal with alternative splice isoforms, and the collapsed one can have trouble when two genomes have different levels of sequencing, assembly coverage, and gene annotations.

To improve intron orthology, we need to consider a flanking sequence based intron orthology definition. The idea of this approach is that the short flanking coding sequences of each intron are enough to find conserved regions and small enough to avoid any circular effect of using coding sequences. This approach has the advantages of no need of multiple alignments, and produces more reliable intron orthology with alternative splice forms.

Finally, the gene family history reconstruction can be improved with possible future works including a probabilistic approach like PrIME-GSR (Akerborg et al., 2009). Although the current local synteny and gene structure information might not be enough to predict a reliable evolutionary history simply because of insufficient characters, it might be applicable and useful if we combine it with coding sequence information. Also since we know that different types of duplication mechanisms occur at different times with varying rates, we would be able to incorporate this information into the reconstruction method. For instance, WGDs occur early in the evolutionary time scale and very rarely, and it is extremely hard to capture the lost genes after WGDs. Tandem duplicates are comparatively younger events but it is very difficult to reconstruct the duplication history, so we might need to apply other tandem array evolution models (Bertrand & Gascuel, 2005; Tang et al., 2002) to detect the duplication/deletion events. Detection of non-tandem

DD genes is very sensitive to the size of the duplication block, whereas sometime ended up with gene fissions and gene fusions . Moreover, different duplicative transposition mechanisms can help to improve detection power. Another extreme case is retrotransposition, which is believed to be a burst of events on recent lineages. As I have shown that some RD genes landed in an intronic region, it might be interesting to check if any RD genes invoked gene fusions and to see if they can be considered in a reconstruction method.

Chapter 5

Conclusions

The duplication and divergence of single copy ancestral genes into large gene families has been an essential driving force in the evolution of organismal complexity. Over the last decade the emergence of new sequencing technologies has enabled the rapid and cost-effective sequencing of many mammalian genomes. A key challenge in understanding the evolution of gene families in these species is the accurate reconstruction of the evolutionary history of each gene family including duplication mechanism, and a destination of the fates of each newly born duplicates including diversification, stabilization or loss. The goal of this thesis has been to develop and evaluate new tools for the accurate reconstruction of gene family histories and to apply these tools to understanding the evolution of gene families in mammalian genomes.

In the first chapter I discuss the importance of determining ancestral orthology between genes in distinct genomes and contrast the definition of ancestral orthology with functional orthology. I also point out that methods utilizing the coding sequence of the series of genes to determine orthology may be confounded by the convergent evolution of those genes for common function and thereby highlight the value of methods that seek to define orthology using non-coding characters of

the genes in question. I also discuss the difficulty that gene loss events present for the reconstruction of gene family histories and propose that incorporating pseudogenes into gene family history reconstruction algorithms can help to minimize the impact of gene loss events on the reconstruction problem. Finally I suggest that a more complete picture of gene family evolution may be garnered by methods that incorporates duplication mechanism in the reconstruction scenario.

In the second chapter represents a new approach for the identification of ancestral Orthologs utilizing non-coding characters of the constituent genes. Specifically, I use local synteny information and intron content to identify orthologous genes in five mammalian genomes. I show that this approach is computationally simple but powerful enough to provide accuracy comparable to widely utilized methods exploiting coding sequence to determine orthology. In order to investigate the strengths and weaknesses of this approach I enumerate and explore cases of concordance and discordance between my orthology detection method and Inparanoid. This analysis shows that local synteny can distinguish retrotransposed duplicates from ancestral orthologs in cases where Inparanoid fails to do so. I also show that my method can dissect ambiguous clusters of homologous genes into distinct orthologous relationships.

In Chapter 3, I show how local synteny information can be used to determine orthology, identify the duplication mechanism, and infer a gene family history. Two clustering algorithms are presented in this chapter. The first is used to form synthetic clusters and separate retro-duplication events from DNA-mediated duplication events. The second is used to break syntenic clusters into sub-clusters defined by successive DNA duplication events. In this way I generate a map of duplication events within the species tree in question. These events can then be placed upon particular branches of the tree using parsimony.

In Chapter 4, I apply these methods to five mammalian genomes including:

human, chimpanzee, mouse, rat and dog. The results of this analysis are striking in that they indicate that roughly equal numbers of new genes are contributed to mammalian genomes by RNA- and DNA-mediated duplication events. While I find that the DNA-mediated duplications are far more likely to give rise to functional genes, RNA-mediated duplications have been at a sufficiently higher rate to bring their net contributions to similar levels. This analysis is the first analysis of mammalian gene family evolution utilizing strictly non-coding characters, discriminating between duplication mechanisms, and incorporating pseudogenes to generate more complete gene family histories. I also show in this chapter that mammalian gene families tend to evolve through one duplication mechanism or the other but rarely both. As might be expected families of genes that are expressed at very high levels contain large numbers of retro-duplicates, while gene families that rely upon complex regulatory information, or that exist in large tandem arrays, tend to evolve through DNA-mediated duplication events. Finally I show that the location of the newly duplicated gene plays a large role in determining the fate of that duplicate. DNA-mediated duplicates with disrupted flanking regions are more likely to diverge from their original state than are duplicates with intact flanking regions. RNA-mediated duplicates are strongly affected by their insertion sites. Intact retro-duplicates are found closer to other functional genes than are retro-duplicated pseudogenes. This result suggests that a prominent load of functionalization for retro-duplicates may include co-option of existing regulatory elements. Taken together these findings reinforce the importance of duplication mechanism to understanding the evolution of gene families.

Bibliography

Addario-Berry, L., Hallett, M., & Lagergren, J. (2003). Towards identifying lateral gene transfer events. *Pacific Symposium on Biocomputing*, (pp. 279–290).

Akerborg, O., Sennblad, B., Arvestad, L., & Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *PNAS*, *106*(14), 5714–5719.

Alexeyenko, A., Tamas, I., Liu, G., & Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22*(14), e9–15.

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Arvestad, L. (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, *19*(90001), 7i–15.

Bailey, J., Liu, G., & Eichler, E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics*, *73*(4), 823–834.

Bandyopadhyay, S., Sharan, R., & Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison. *Genome research*, *16*(3), 428–435.

Beissbarth, T., & Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, *20*(9), 1464–1465.

Berglund, A.-C., Sjölund, E., Ostlund, G., & Sonnhammer, E. L. L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic acids research*, *36*(Database issue), D263–6.

Bertrand, D., & Gascuel, O. (2005). Topological rearrangements and local search method for tandem duplication trees. *IEEE/ACM transactions on computational biology and bioinformatics*, *2*(1), 15–28.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., & Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, *299*(5611), 1391–1394.

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, *2*(4), e383.

Chen, K., Durand, D., & Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of computational biology*, *7*(3-4), 429–447.

Chou, H.-H., Hayakawa, T., Diaz, S., Krings, M., Indriati, E., Leakey, M., Paabo, S., Satta, Y., Takahata, N., & Varki, A. (2002). Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *PNAS*, *99*(18), 11736–11741.

Cooper, S. J. B., Wheeler, D., Hope, R. M., Dolman, G., Saint, K. M., Gooley, A. A., & Holland, R. A. B. (2005). The alpha-globin gene family of an Australian marsupial, Macropus eugenii: the long evolutionary history of the theta-globin

gene and its functional status in mammals. *Journal of molecular evolution*, *60*(5), 653–664.

Csurös, M. (2008). Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics*, *24*(13), 1538–1539.

Csurös, M., Holey, J. A., & Rogozin, I. B. (2007). In search of lost introns. *Bioinformatics*, *23*(13), i87–96.

Deluca, T. F., Wu, I.-H., Pu, J., Monaghan, T., Peshkin, L., Singh, S., & Wall, D. P. (2006). Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, *22*(16), 2044–2046.

DeLuna, A., Vetsigian, K., Shoresh, N., Hegreness, M., Colón-González, M., Chao, S., & Kishony, R. (2008). Exposing the fitness contribution of duplicated genes. *Nature genetics*, *40*(5), 676–681.

Dharia, A., Jun, J., & Nelson, C. E. (2010). Dynamics of Genome Architecture and the contribution of Gene Duplications. Unpublished.

Emerson, J. J., Kaessmann, H., Betrán, E., & Long, M. (2004). Extensive gene traffic on the mammalian X chromosome. *Science*, *303*(5657), 537–540.

Enright, A. J. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584.

Ensembl (2007). Ensembl Release 48.
URL `http://dec2007.archive.ensembl.org/index.html`

Fang, G., Bhardwaj, N., Robilotto, R., & Gerstein, M. B. (2010). Getting started in gene orthology and functional analysis. *PLoS Computational Biology*, *6*(3), e1000703.

Felsenstein, J. (2004). PHYLIP (Phylogeny Inference Package) version 3.6. URL `http://evolution.genetics.washington.edu/phylip.html`

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., & Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic acids research*, *34*(Database issue), D247–251.

Fitch, W. (2000). Homology a personal view on some of the problems. *Trends in genetics*, *16*(5), 227–231.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., & Searle, S. (2008). Ensembl 2008. *Nucleic acids research*, *36*(Database issue), D707–714.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., Karimpour-Fard, A., Glueck, D., McGavran, L., Berry, R., Pollack, J., & Sikela, J. M. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS biology*, *2*(7), E207.

Friedman, R., & Hughes, A. L. (2004). Two patterns of genome organization

in mammals: the chromosomal distribution of duplicate genes in human and mouse. *Molecular biology and evolution*, *21*(6), 1008–1013.

Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y., & Jiang, T. (2006). *A parsimony approach to genome-wide ortholog assignment*, vol. 3909 of *LNBI*, (pp. 578–594).

Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y., & Jiang, T. (2007). MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. *Journal of computational biology*, *14*(9), 1160–1175.

Fu, Z., & Jiang, T. (2008). Clustering of main orthologs for multiple genomes. *Journal of bioinformatics and computational biology*, *6*(3), 573–584.

Gabaldón, T., Dessimoz, C., Huxley-Jones, J., Vilella, A. J., Sonnhammer, E. L., & Lewis, S. (2009). Joining forces in the quest for orthologs. *Genome biology*, *10*(9), 403.

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *The Journal of heredity*, *100*(5), 605–617.

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., & Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome research*, *19*(5), 859–867.

Han, M. V., & Hahn, M. W. (2009). Identifying parent-daughter relationships among duplicated genes. *Pacific Symposium on Biocomputing*, *125*, 114–125.

Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N., & Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic acids research*, *33*(8), 2374–2383.

He, X., & Goldwasser, M. H. (2008). Identifying conserved gene clusters in the presence of homology families. *Journal of computational biology*, *12*(6), 638–656.

Hendrickson, H., Slechta, E. S., Bergthorsson, U., Andersson, D. I., & Roth, J. R. (2002). Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. *PNAS*, *99*(4), 2164–2169.

Ho Sui, S. J., Fulton, D. L., Arenillas, D. J., Kwon, A. T., & Wasserman, W. W. (2007). oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic acids research*, *35*(Web Server issue), W245–252.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., & Birney, E. (2005). Ensembl 2005. *Nucleic acids research*, *33*(Database issue), D447–53.

Hui, S. L., & Zhou, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical methods in medical research*, *7*(4), 354–370.

Hurley, I., Hale, M. E., & Prince, V. E. (2005). Duplication events and the evolution of segmental identity. *Evolution & development*, *7*(6), 556–567.

Innan, H. (2009). Population genetic models of duplicated genes. *Genetica*, *137*(1), 19–37.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., & Roest Crollius, H. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, *431*(7011), 946–957.

Johnson, T. (2007). Reciprocal best hits are not a logically sufficient condition for orthology. *Quantitative biology*, (pp. 1–8).

Jun, J., Mandoiu, I. I., & Nelson, C. E. (2009a). Identification of mammalian orthologs using local synteny. *BMC genomics*, *10*(1), 630.

Jun, J., Ryvkin, P., Hemphill, E., Mandoiu, I., & Nelson, C. (2009b). The birth of new genes by RNA- and DNA-mediated duplication during mammalian evolution. *Journal of computational biology*, *16*(10), 1429–1444.

Jun, J., Ryvkin, P., Hemphill, E., Mndoiu, I., & Nelson, C. (2008). Estimating the relative contributions of new genes from retrotransposition and segmental duplication events during mammalian evolution. In *6th RECOMB Comparative Genomics Satellite Workshop*, (pp. 40–54).

Jun, J., Ryvkin, P., Hemphill, E., & Nelson, C. (2009c). Duplication mechanism

and disruptions in flanking regions determine the fate of Mammalian gene duplicates. *Journal of computational biology*, *16*(9), 1253–1266.

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, *10*(1), 19–31.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, *428*(6983), 617–624.

Kijima, T. E., & Innan, H. (2010). On the estimation of the insertion time of LTR retrotransposable elements. *Molecular biology and evolution*, *27*(4), 896–904.

Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., & Gerstein, M. B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome research*, *18*(12), 1865–1874.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, *39*, 309–338.

Koonin, E. V., Mushegian, A. R., & Bork, P. (1996). Non-orthologous gene displacement. *Trends in genetics*, *12*(9), 334–336.

Lajoie, M., Bertrand, D., & El-Mabrouk, N. (2010). Inferring the evolutionary history of gene clusters from phylogenetic and gene order data. *Molecular biology and evolution*, *27*(4), 761–772.

Lemoine, F., Labedan, B., & Lespinet, O. (2008). SynteBase/SynteView: a tool to

visualize gene order conservation in prokaryotic genomes. *BMC bioinformatics*, *9*(1), 536.

Lemoine, F., Lespinet, O., & Labedan, B. (2007). Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC evolutionary biology*, *7*(1), 237.

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Hériché, J.-K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., Wong, G. K.-S., Zheng, W., Dehal, P., Wang, J., & Durbin, R. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research*, *34*(Database issue), D572–580.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, *13*(9), 2178–2189.

Li, W.-H. (1997). *Molecular Evolution*. Sinauer Associates.

Lunter, G., Ponting, C. P., & Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS computational biology*, *2*(1), e5.

Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates Inc.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., & Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. *PLoS biology*, *3*(11), e357.

McCarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, *39*(7 Suppl), S37–42.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, *91*(3), 243–248.

Mural, R. J., Adams, M. D., Myers, E. W., Smith, H. O., Miklos, G. L. G., Wides, R., Halpern, A., Li, P. W., Sutton, G. G., Nadeau, J., Salzberg, S. L., Holt, R. A., Kodira, C. D., Lu, F., Chen, L., Deng, Z., Evangelista, C. C., Gan, W., Heiman, T. J., Li, J., Li, Z., Merkulov, G. V., Milshina, N. V., Naik, A. K., Qi, R., Shue, B. C., Wang, A., Wang, J., Wang, X., Yan, X., Ye, J., Yooseph, S., Zhao, Q., Zheng, L., Zhu, S. C., Biddick, K., Bolanos, R., Delcher, A. L., Dew, I. M., Fasulo, D., Flanigan, M. J., Huson, D. H., Kravitz, S. A., Miller, J. R., Mobarry, C. M., Reinert, K., Remington, K. A., Zhang, Q., Zheng, X. H., Nusskern, D. R., Lai, Z., Lei, Y., Zhong, W., Yao, A., Guan, P., Ji, R.-R., Gu, Z., Wang, Z.-Y., Zhong, F., Xiao, C., Chiang, C.-C., Yandell, M., Wortman, J. R., Amanatides, P. G., Hladun, S. L., Pratts, E. C., Johnson, J. E., Dodson, K. L., Woodford, K. J., Evans, C. A., Gropman, B., Rusch, D. B., Venter, E., Wang, M., Smith, T. J., Houck, J. T., Tompkins, D. E., Haynes, C., Jacob, D., Chin, S. H., Allen, D. R., Dahlke, C. E., Sanders, R., Li, K., Liu, X., Levitsky, A. A., Majoros, W. H., Chen, Q., Xia, A. C., Lopez, J. R., Donnelly, M. T., Newman, M. H., Glodek, A., Kraft, C. L., Nodell, M., Ali, F., An, H.-J., Baldwin-Pitts, D., Beeson, K. Y., Cai, S., Carnes, M., Carver, A., Caulk, P. M., Center, A., Chen, Y.-H., Cheng, M.-L., Coyne, M. D., Crowder, M., Danaher, S., Davenport, L. B., Desilets, R., Dietz, S. M., Doup, L., Dullaghan, P., Ferriera, S., Fosler, C. R., Gire, H. C., Gluecksmann, A., Gocayne, J. D., Gray, J., Hart, B., Haynes, J., Hoover, J., Howland, T., Ibegwam, C., Jalali, M., Johns, D., Kline, L., Ma, D. S., MacCawley, S., Magoon, A., Mann, F., May, D., McIntosh, T. C., Mehta, S., Moy, L., Moy, M. C., Murphy, B. J., Murphy, S. D., Nelson, K. A., Nuri, Z., Parker, K. A., Prudhomme, A. C., Puri, V. N., Qureshi, H., Raley, J. C., Reardon, M. S., Regier, M. A., Rogers, Y.-H. C., Romblad, D. L., Schutz, J., Scott, J. L., Scott, R., Sitter, C. D., Smallwood, M., Sprague, A. C., Stewart, E., Strong, R. V., Suh, E., Sylvester, K., Thomas, R., Tint, N. N.,

Tsonis, C., Wang, G., Wang, G., Williams, M. S., Williams, S. M., Windsor, S. M., Wolfe, K., Wu, M. M., Zaveri, J., Chaturvedi, K., Gabrielian, A. E., Ke, Z., Sun, J., Subramanian, G., Venter, J. C., Pfannkoch, C. M., Barnstead, M., & Stephenson, L. D. (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, *296*(5573), 1661–71.

Nekrutenko, A., Makova, K. D., & Li, W.-H. (2002). The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome research*, *12*(1), 198–202.

Ohno, S. (1970). *Evolution by Gene Duplication*. London, United Kingdom: Allen and Unwin.

Panopoulou, G., & Poustka, A. J. (2005). Timing and mechanism of ancient vertebrate genome duplications – the adventure of a hypothesis. *Trends in genetics*, *21*(10), 559–567.

Patel, V. S., Cooper, S. J. B., Deakin, J. E., Fulton, B., Graves, T., Warren, W. C., Wilson, R. K., & Graves, J. A. M. (2008). Platypus globin genes and flanking loci suggest a new insertional model for beta-globin evolution in birds and mammals. *BMC biology*, *6*, 34.

Pavesi, G., Zambelli, F., Caggese, C., & Pesole, G. (2008). Exalign: a new method for comparative analysis of exon-intron gene structures. *Nucleic acids research*, *36*(8), e47.

Petrov, D. A., & Hartl, D. L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. *PNAS*, *96*(4), 1475–1479.

Poptsova, M. S., & Gogarten, J. P. (2007). BranchClust: a phylogenetic algorithm for selecting gene families. *BMC bioinformatics*, *8*, 120.

Qu, Y., Tan, M., & Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, *52*(3), 797–810.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., & Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454.

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., & Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology*, *13*(17), 1512–1517.

Roth, A., Dessimoz, C., Cannarozzi, G., Margadant, D., Gil, M., Gonnet, G., & Schneider, A. (2005). OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements. In *3rd RECOMB Comparative Genomics Satellite Workshop*, (pp. 61–72).

Ryvkin, P., Jun, J., Hemphill, E., & Nelson, C. (2008). Duplication Mechanism and Disruptions in Flanking Regions Influence the Fate of Mammalian Gene Duplicates. In *6th RECOMB Comparative Genomics Satellite Workshop*, (pp. 26–39).

Sakai, H., Koyanagi, K. O., Imanishi, T., Itoh, T., & Gojobori, T. (2007). Frequent

emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene*, *389*(2), 196–203.

Samonte, R. V., & Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. *Nature reviews. Genetics*, *3*(1), 65–72.

Sankoff, D. (1999). Genome rearrangement with gene families. *Bioinformatics (Oxford, England)*, *15*(11), 909–17.

Schrider, D. R., Costello, J. C., & Hahn, M. W. (2009). All human-specific gene losses are present in the genome as pseudogenes. *Journal of computational biology*, *16*(10), 1419–1427.

Shemesh, R., Novik, A., Edelheit, S., & Sorek, R. (2006). Genomic fossils as a snapshot of the human transcriptome. *PNAS*, *103*(5), 1364–1369.

Shoja, V., & Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Molecular biology and evolution*, *23*(11), 2134–2141.

Skrabanek, L. (1998). Eukaryote genome duplication - where's the evidence? *Current Opinion in Genetics & Development*, *8*(6), 694–700.

Sokal, R. R., & Sneath, P. H. A. (1973). *Numerical Taxonomy: The Principles and Practice of Numberical Classification*. San Francisco: W.H. Freeman and Company.

Su, Z., Wang, J., Yu, J., Huang, X., & Gu, X. (2006). Evolution of alternative splicing after gene duplication. *Genome research*, *16*(2), 182–189.

Svensson, O., Arvestad, L., & Lagergren, J. (2006). Genome-wide survey for biologically functional pseudogenes. *PLoS computational biology*, *2*(5), e46.

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., & Hannon, G. J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, *453*(7194), 534–538.

Tang, M., Waterman, M., & Yooseph, S. (2002). Zinc finger gene clusters and tandem gene duplication. *Journal of computational biology*, *9*(2), 429–446.

Tatusov, R. L. (1997). A Genomic Perspective on Protein Families. *Science*, *278*(5338), 631–637.

Torrents, D., Suyama, M., Zdobnov, E., & Bork, P. (2003). A genome-wide survey of human pseudogenes. *Genome research*, *13*(12), 2559–67.

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. *Genome research*, *14*(1), 62–66.

Uechi, T. (2002). Functional second genes generated by retrotransposition of the X-linked ribosomal protein genes. *Nucleic Acids Research*, *30*(24), 5369–5375.

Ureta-Vidal, A., Ettwiller, L., & Birney, E. (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature reviews. Genetics*, *4*(4), 251–262.

Van de Peer, Y., Taylor, J. S., & Meyer, A. (2003). Are all fishes ancient polyploids? *Journal of structural and functional genomics*, *3*(1-4), 65–73.

Vermunt, J. K. (1997). LEM: A General Program for the Analysis of Categorical Data.

URL `http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html`

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, *19*(2), 327–335.

Vinckenbosch, N., Dupanloup, I., & Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. *PNAS*, *103*(9), 3220–3225.

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

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007a). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, *23*(13), i549–558.

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007b). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, *449*(7158), 54–61.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M. A., Sakaki, Y., & Sasaki, H. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, *453*(7194), 539–543.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M.,

Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams,

S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., & Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562.

Wheeler, D., Hope, R., Cooper, S. B., Dolman, G., Webb, G. C., Bottema, C. D., Gooley, A. A., Goodman, M., & Holland, R. A. (2001). An orphaned mammalian beta-globin gene of ancient evolutionary origin. *PNAS*, *98*(3), 1101–1106.

Wheeler, D., Hope, R. M., Cooper, S. J. B., Gooley, A. A., & Holland, R. A. B. (2004). Linkage of the beta-like omega-globin gene to alpha-like globin genes in an Australian marsupial supports the chromosome duplication model for separation of globin gene clusters. *Journal of molecular evolution*, *58*(6), 642–652.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS*, *13*(5), 555–556.

Yang, Z., & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution*, *17*(1), 32–43.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298.

Zhang, Z., Carriero, N., & Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in genetics*, *20*(2), 62–67.

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P. M., & Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, *22*(12), 1437–1439.

Zhang, Z., Harrison, P., & Gerstein, M. (2002). Identification and analysis of over

2000 ribosomal protein pseudogenes in the human genome. *Genome research*, *12*(10), 1466–1482.

Zhang, Z. D., Frankish, A., Hunt, T., Harrow, J., & Gerstein, M. (2010). Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome biology*, *11*(3), R26.

Zheng, X. H., Lu, F., Wang, Z.-Y., Zhong, F., Hoover, J., & Mural, R. (2005). Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics*, *21*(6), 703–710.