A Broad Analysis of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat

Valia Shoja

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Application

Liqing Zhang, Ph.D. Chair
Lenwood S. Heath, Ph.D.
Zhijian Tu, Ph.D.

November 10, 2006
Blacksburg, Virginia

Keywords: Comparative Genomics, Gene Duplication,
Tandemly Arrayed Genes, Gene Expression.

A Broad Analysis of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat

Valia Shoja

(**ABSTRACT**)

Tandemly arrayed genes (TAG) play an important functional and physiological role in the genome. Most previous studies have focused on individual TAG families in a few species, yet a broad characterization of TAGs is not available. We identified all the TAGs in the genomes of human, chimp, mouse, and rat and performed a comprehensive analysis of TAG distribution, TAG sizes, TAG gene orientations and intergenic distances, and TAG gene functions. TAGs account for about 14-17% of all the genomic genes and nearly one third of all the duplicated genes in the four genomes, highlighting the predominant role that tandem duplication plays in gene duplication. For all species, TAG distribution is highly heterogeneous along chromosomes and some chromosomes are enriched with TAG forests while others are enriched with TAG deserts. The majority of TAGs are of size two for all genomes, similar to the previous findings in *C. elegans*, *A. thaliana*, and *O. sativa*, suggesting that it is a rather general phenomenon in eukaryotes.

The comparison with the genome patterns shows that TAG members have a significantly higher proportion of parallel gene orientation in all species, corroborating Graham's claim that parallel orientation is the preferred form of orientation in TAGs. Moreover, TAG members with parallel orientation tend to be closer to each other than all neighboring genes with parallel orientation in the genome. The analysis of GO function indicate that genes with receptor or binding activities are significantly over-represented by TAGs. Simulation reveals that random gene rearrangements have little effect on the statistics of TAGs for all genomes. It is noteworthy to mention that gene family sizes are significantly correlated with the extent of tandem duplication, suggesting that tandem duplication is a preferred form of duplication, especially in large families.

There has not been any systematic study of TAG genes' expression patterns in the genome. Taking advantage of recent large-scale microarray data, we were able to study expression divergence of some of the TAGs of size two in human and mouse for which the expression data is available and examine the effect of sequence divergence, gene orientation, and physical proximity on the divergence of gene expression patterns. Our results show that there is a weak negative correlation between sequence divergence and expression similarity between the two members of a TAG, and also a weak negative correlation between physical proximity of two genes and their expression similarity. No significant relationship was detected between gene orientation and expression similarity. Moreover, we compared the expression breadth of upstream and downstream duplicate copies and found that downstream duplicate does not show significantly narrower expression breadth. We also compared TAG gene pairs with their neighboring non-TAG pairs for both physical proximity and expression similarity. Our results show that TAG gene pairs do not show any distinct differences in the two aspects

from their neighboring gene pairs, suggesting that sufficient divergence has occurred to these duplicated genes during evolution and their original similarity conferred by duplication has decayed to a level that is comparable to their surrounding regions.

فلک بمردم نادان دهد زمام مراد تو اهل دانش و فضلی همین گناهت بس

Heavens fulfill wishes of the ignorant,
You seek wisdom and grace, this sin shall suffice you.
(Hafez, 14th AD)

"Does the evolutionary doctrine clash with religious faith? It does not. It is a blunder to mistake the Holy Scriptures for elementary textbooks of astronomy, geology, biology, and anthropology. Only if symbols are construed to mean what they are not intended to mean can there arise imaginary, insoluble conflicts. ...the blunder leads to blasphemy: the Creator is accused of systematic deceitfulness."
(Theodosius Dobzhansky [20])

Acknowledgments

I would like to express my gratitude toward my advisor, Dr. Liqing Zhang for her continuous support and supervision. I am also thankful to Dr. Lenwood S Heath and Dr. Zhijian Tu for serving in my advisory committee. Special thanks go to my parents, Mahin and Ali Shoja and my brothers, Amir Hossein and Amir Ali, for their untiring care and support over the years. And finally, I would like to thank my husband, Dr. Shahriar Setoodeh for his love and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

"The challenge of the Human Genome Project will be to go from ordering the letters of the DNA language to understanding the words, phrases, sentences, paragraphs, and finally the story of the genome" [51].

1.1 Gene Duplication

Genomes evolve by gain, loss, modification, and rearrangement of genetic material. Understanding the forces that affect the genome is essential in our understanding of the origin, survival, and adaptation of species [23]. The ultimate consequence of genome evolution is speciation. Chromosomal rearrangements, and divergent resolution of duplicated genes will lead to reproductive isolation [95]. Eukaryotic genomes are well-supplied with genetic duplication. Plant genomes particularly have gained their duplication via an extensive history of polyploidy events [7, 8, 9]. For example, more than 28% of existing *Arabidopsis thaliana* genes were duplicated during ancient whole-genome duplication events [45].

The importance of duplicated genes in providing raw materials for genetic innovation has been recognized since the 1930s and is highlighted in Ohno's 1970 book Evolution by Gene Duplication [69]; yet only in the past few years, has the availability of numerous genomic sequences made it possible to develop a quantitative measure of the proportion of genes in a genome that are risen from genome duplication (e.g., Table 1.1, taken from [108]). Divergence of many duplicated genes to a great extent such that their common origin can no longer be recognized indecates that the current estimates of the extent of gene duplications may be low.

Known mechanisms that can create duplicated genes include unequal crossover (i.e., tandem duplication), retrotransposition, replicable translocation, and chromosomal (or genome) duplication. Unequal crossover consists of chromosomal mispairing followed by the exchange of

Table 1.1: Prevalence of gene duplication in all three domains of life

	Total number of genes	Number of duplicated genes (% of duplicate genes)
Bacteria		
Mycoplasma pneumoniae	677	298 (44)
Helicobacter pylori	1590	266 (17)
Haemophilus influenzae	1709	284 (17)
Archaea		
Archaeoglobus fulgidus	2436	719 (30)
Saccharomyces cerevisiae	6241	1858 (30)
Caenorhabditis elegans	18424	8971 (49)
Eukarya		
Drosophila melanogaster	13601	5536 (41)
Arabidopsis thaliana	25498	16574 (65)
Homo sapiens	40580	15343 (38)

DNA between nonhomologous regions resulting in either gene duplication or gene deletion. Retrotransposition consists of reverse transcription of the mRNA transcript of a gene into double stranded DNA followed by insertion of the double stranded DNA into a location typically distant from the original gene. Many recent studied have focused on large-scale computational identification and characterization of retrotransposed duplicated genes with respect to their location and dynamics in species such as human, mouse, and arabidopsis [25, 61, 99, 112, 113]. However, the study of duplicated genes generated through unequal crossover has been restricted, for the most part, to contexts of duplication of individual genes (i.e., [2, 32, 68, 97, 111]).

1.2 Tandemly arrayed genes

Tandemly arrayed genes (TAGs) were discovered in the 1960's, with ribosomal RNA genes and immunoglobulin genes among the first instances of TAGs identified [12, 76, 86]. These TAGs exhibit two distinct patterns of sequence divergence. Either, TAG members retain identical or nearly identical sequences (e.g., ribosomal RNA genes); or, array members exhibit significant sequence divergence (e.g., the immunoglobulin and MHC genes).

Followings are some of the most fundamental reasons for investigating duplicated genes rising from unequal crossover: First, unequal crossover produces tandemly arrayed genes (TAGs) - duplicated genes that are neighbors on a chromosome. These genes constitute a large pro-

portion of several eukaryotic genomes. For example, at least 10% of the genes in the genomes of worm, arabidopsis, human, and mouse are TAGs [82, 83, 109]. Second, TAGs play an essential role in functions of organisms. Some of the most studied examples of TAGs are the ribosomal RNA genes (the central component of the ribosome), the major histocompatibility complex (MHC, which plays an important role in the immune system, autoimmunity, and reproductive success of vertebrates), Homeobox genes (which play a key role in animal body plan development), MADS-Boxgenes (which play a central role in flower development in plants), and nucleotide-binding site plus leucine-rich repeat genes (NBS-LRR, which represent the major class of disease resistance genes in flowering plants). In short, TAGs promote genomic diversity to enhance disease resistance, satisfy the requirement for a large amount of gene product, and contribute to the fine-tuning of developmental stages and physiological functions [56, 69]. Moreover, TAG generation is a common mechanism which organisms resort to when experiencing stress or selective pressure [53]. Third, mutations in TAGs and tandem duplications of some genes contribute to a number of human diseases, including breast cancer [41], neurofibromatosis [5], Thalassemia [1, 28], and Pelizaeus-Merzbacher disease [24, 46, 72].

1.3 Gene Expression

As a result of advance large-scale gene expression profiling, many important questions on the evolution of gene function have been emerging: i.e. what are the patterns of gene expression in duplicated genes? How exactly has duplication contributed to the functional diversity? Many evolutionary biologists have address similar questions by employing large-scale gene expression analysis and their studies have already shed lights on evolutionary divergence of gene expression and function of duplicated genes (e.g., [36, 60, 43, 59].) However, there have not been any such studies on the effect of chromosome location on expression divergence of different types of duplicated genes: comparing expression divergence among tandemly or dispersed duplicated genes. The expression of tandemly arrayed genes has been commonly assumed to be correlated [17, 102]. In most of these studies, it has been assumed that TAGs are co-expressed and have correlated expression and therefore are discarded from further analysis in order to minimize the correlations contributed by these genes. However, the validity of the assumption has never been tested.

Some TAGs may undergo frequent gene conversion in their upstream regions and consequently have highly identical sequences among copies in their upstream regions (Zhang, submitted). Therefore, the homogenization in upstream regions hinders divergence in sequences and could lead to a synchronization of gene expression among TAG genes. Also, it has been shown that duplicated genes that are not in proximity of each other have limited gene conversion [64]. The upstream regions of these genes might be more diverged from each other than those among TAGs. Therefore, TAGs unlike dispersed duplicated genes might have highly similar expression profiles not only due to the location and linkage effect, but

also due to the homogenization effect contributed by gene conversion.

It has been observed that linked neighboring genes show higher similarity in gene expression profiles than non-linked genes in many different organisms such as in human [16, 55], *C. elegans* [54], *A. thaliana* [74], *D. melanogaster* [19, 10], and *S.cerevisiae* [19]. Given the strong empirical evidence of highly synchronized gene expression among neighboring genes, it is possible to conclude that tandemly arrayed genes share higher similarities in gene expression than those duplicated genes which are not TAGs. Therefore, it is relevant and essential to examine its impact on the divergence of tandem duplicates.

1.4 Thesis Layout

In the next chapter, a specific method of gene duplication, tandem duplication, is reviewed among three mammalian genomes. Tandemly arrayed genes are identified in the genomes of human, mouse, and rat. A possibility of random arrangement of duplicated genes in form of tandem arrays is analyzed by simulation. Also, results of a comprehensive analysis of TAG distribution, sizes, orientations, intergenic regions, and GO categories are given . Notion of TAG "forest" and "desert" are introduced and enrichment and depletion of them are tested among all chromosomes in three genomes. The correlation of tandem duplication and gene family size is examined.

In chapter 3 gene expression of tandemly arrayed genes is studied. Due to the fact that large-scale gene expression data is only available for human and mouse genomes, gene expression analysis only covers these two genomes. Since the linkage effect only makes sence when a pair of genes is studied, TAGs of size 2 are chosen.

Finally, chapter 4 summarizes what has been accomplished along with future research directions.

Chapter 2

A Roadmap of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat

2.1 Introduction

DNA duplication is the principle process by which the genetic raw material is provided for the origin of evolutionary novelties such as new gene function and expression patterns and is important in adaptive evolution [103]. Possible duplication mechanisms include unequal crossover (i.e., tandem duplication), retrotransposition, and replicative translocation [69, 65]. Duplication events may vary in content and frequency: while whole genome duplications appear to be limited, small tandem duplications appear to be quite common. In fact, localized duplication of genomic segments and rearrangement of chromosomal segments have been proposed to be two major factors in eukaryotic genome evolution [23].

The availability of complete genomic sequences makes it possible to investigate how genomes are structured by different mechanisms of gene duplication. Tandem duplication of related genes has been shown to act as the driving evolutionary force in the origin and maintenance of gene families [73] and has been a common mechanism of genetic adaptation to environmental challenges in organisms such as bacteria [4, 39, 78], yeast [11], mosquitoes [53], plants [38, 52, 84], and humans and other mammals [90].

Specifically, we identified all tandemly arrayed genes (TAGs) in the genomes of human, mouse, and rat and addressed the following issues: First, since duplicated genes can be arranged in tandem or dispersed on different chromosomes, we want to determine how many duplicated genes are in tandem arrays. This will shed light on the contribution of tandem duplication to gene duplication in the three mammalian genomes. Second, we are interested in examining the chromosomal distribution of TAGs to see whether there is significant

clustering of TAGs on some chromosomal regions. Third, about 70% of the TAGs in the *A. thaliana* genome have only two members in the array [109], do the three mammalian genomes show a similar pattern? Fourth, is there any non-random association between gene function defined by Gene Ontology (GO) categories and TAGs? We expect that genes with certain functions may prefer tandem arrangement over other types of more dispersed spatial arrangements, as tandem arrangement can either entail high probability of generating more duplicated copies or promote a desired degree of diversity or homogeneity via concerted evolution [69]. Fifth, it has been hypothesized that the preferred orientation of TAGs is parallel since locating on different strands is detrimental to the stability of the array [33]. We thus examined the orientations of array members and compared them to the genome pattern. Finally, is tandem duplication a preferred mechanism of duplication for large gene families? In other words, do we observe more TAGs in larger families than smaller ones?

2.2 Material and Methods

The peptide sequences for human, mouse, and rat (version 35: October 2005) were obtained from Ensembl Genome Browser: http://www.ensembl.org. In this version, there are 33869 genes in the human genome, 36471 genes in the mouse genome, and 32543 genes in the rat genome. Sequences annotated as unknown, random, and mitochondrial were removed and only genes with known chromosome location were retained. For genes that have overlapping chromosomal locations, we discarded all shorter genes and kept the longest ones. Similar methods have been used in previous studies [29, 62]. After removal of these genes, we retained 19727 genes for human, 21305 genes for mouse, and 18468 for rat (Table 2.2). These genes were used for the all-against-all BLASTP [3] search with the BLOSUM62 matrix and the SEG filter, which masks regions of low compositional complexity [104].

Next, we applied TribeMCL with the default parameters to cluster genes into putative gene families. TribeMCL uses the Markov clustering algorithm (MCL) for the assignment of proteins into families based on the similarity matrix generated from the all-against-all BLASTP comparison of sequences [26]. Human genes were clustered into 9278 families, of which 6746 are singletons (i.e., only 1 gene in the family). Mouse genes were clustered into 9790 families, of which 7262 are singletons. Rat genes were clustered into 8168 families, of which 6002 are singletons. These singleton families were removed from the dataset. For the rest of the families with more than two members, we examined the chromosomal locations of the family members to decide whether they are TAG genes.

TAGs are usually defined as genes that are duplicated tandemly on chromosomes. During evolution, mutations such as insertion of genes that are unrelated to the TAG members (i.e., not through duplication) can disrupt the tandem spatial arrangement of the original TAGs and thus make the TAG imperfect. To be as broad as possible, we examined the effect of these inserted genes (identified hereafter as spacers) on the quantities of TAGs. We defined spacers as genes that have a BLAST E-value higher than 10^{-10} with the other members

Table 2.1: Statistics of TAGs as a function of the number of spacers allowed in the array.

	Human			Mouse			Rat		
Spacer	TAGs	TAG genes	% of TAG genes	TAGs	TAG genes	% of TAG genes	TAGs	TAG genes	% of TAG genes
0	784	2150	10.9	820	2832	13.3	727	2762	15.0
1	902	2727	13.8	939	3465	16.3	778	3163	17.1
2	959	2996	15.2	1001	3748	17.6	818	3380	18.3
3	1000	3157	16.0	1032	3912	18.4	834	3510	19.0
4	1033	3275	16.6	1053	4011	18.8	852	3620	19.6
5	1051	3354	17.0	1072	4098	19.2	865	3692	20.0
6	1063	3412	17.3	1089	4179	19.6	884	3758	20.3
7	1082	3471	17.6	1095	4224	19.8	899	3813	20.6
8	1107	3539	17.9	1103	4274	20.1	913	3859	20.9
9	1125	3595	18.2	1108	4310	20.2	917	3901	21.1
10	1143	3639	18.4	1117	4357	20.5	923	3931	21.3

in the array and TAGs as duplicated genes with less than 0-10 spacers in between and calculated the numbers of TAGs that satisfy the criteria [109]. Similar to the observation in *A. thaliana* [109], we found that, for all three species, the counts of TAGs increase as more spacers are allowed in the array and most dramatically when only one spacer is allowed in the array (Figure 2.1). Therefore, for the remainder of the study, we focused on TAGs with at most one spacer. Using this criterion, we do not consider tandem duplications that contain multiple spacer genes.

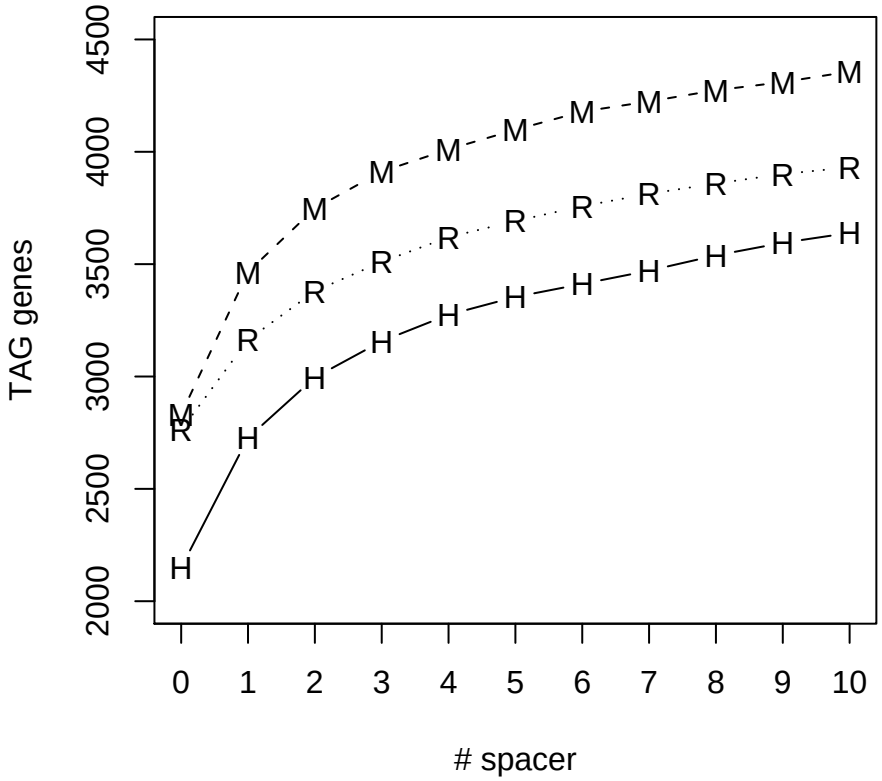

Figure 2.1: Count of TAG genes as a function of spacers

It is possible that due to various genome rearrangements, some duplicated genes that were not originated by tandem duplication or the mechanism of unequal crossover appear together as TAGs. To examine whether the amount of non-real TAGs can adversely affect the statistics of TAGs, we evaluated the likelihood of random arrangement of duplicated genes happening to be TAGs. We numbered all the genes in each genome, randomly picked the locations of all the duplicated genes, and computed the proportion of duplicated genes that appear to be TAGs. We repeated this process 10000 times and obtained the distribution of the proportion

of randomly duplicated genes that belong to TAGs (Figure 2.2).

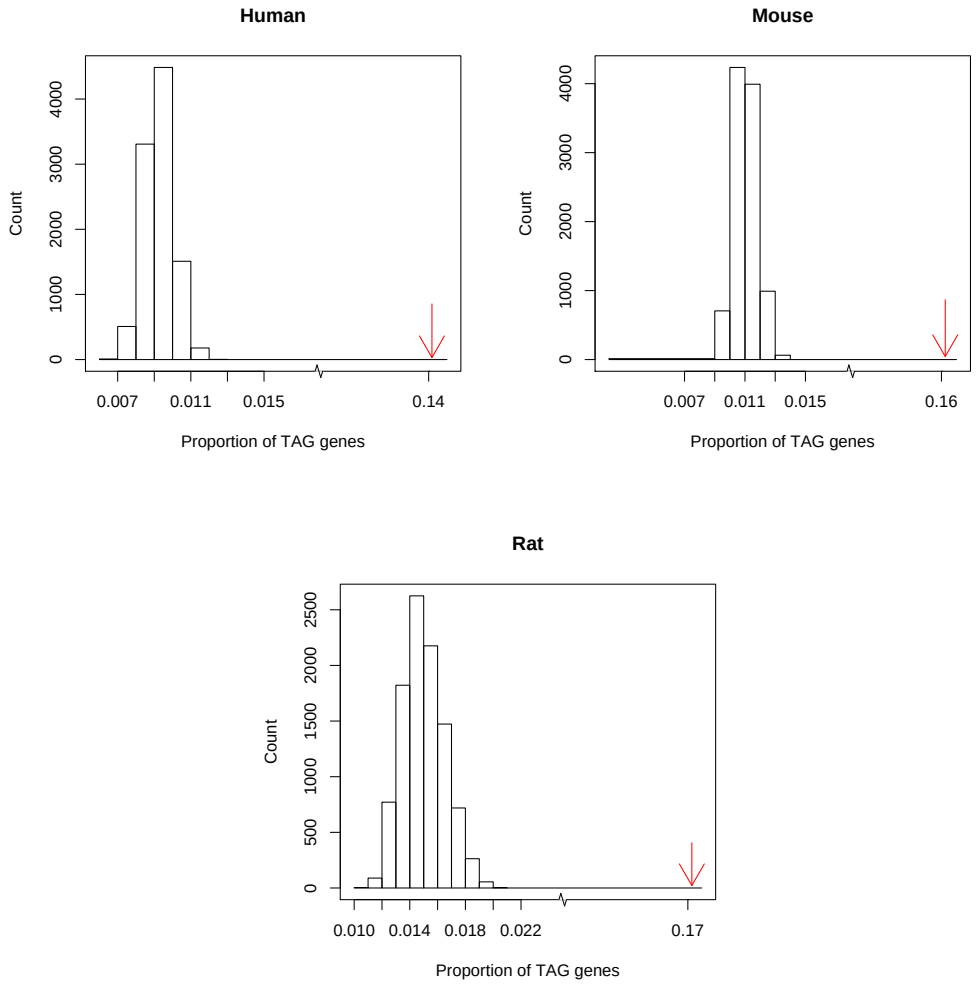

Figure 2.2: Distributions of the proportions of TAG genes in the 10000 simulated samples in human, mouse, and rat genomes. The arrow marks the observed values of the proportions of TAG genes in the genome.

To formally investigate the heterogeneous distribution of TAGs along the chromosomes in the three species, we partitioned every chromosome into 5MB-long blocks and calculated the proportion of genes that are TAGs in each block and identified the blocks that are TAG "deserts" and "forests". TAG deserts are defined as blocks where there are no TAG genes. TAG forests are regions with TAG proportions in the upper 10% of the distribution of TAG proportions. We asked the question: are there chromosomes that are either enriched or depleted with TAG forests or deserts? We applied the hypergeometric test for the enrichment

or depletion of TAG forests or deserts in each chromosome. We also checked pericentromeric and subtelomeric regions for enrichment of TAG deserts or forests. This analysis was limited to the human genome since it is the only species with information on location of these regions. We used the same definition of pericentromeric and subtelomeric regions as initially suggested by [6].

To examine what GO functions are most likely to be over-represented by TAGs across the three species, we used Onto Express [21]. Hypergeometric tests were performed with the Bonferroni correction for multiple testing [88].

2.3 Results

There are 783 perfect TAG clusters (i.e., having zero spacers) containing 2150 genes in the human genome, 820 perfect TAGs with 2832 genes in the mouse genome, and 727 perfect TAGs with 2762 genes in the rat genome (Table 2.1). The perfect TAG genes account for up to 15% of the non-overlapping genes in the three genomes. When one spacer gene is allowed, TAG genes account for 14%, 16%, and 17% of the total genes in the human, mouse, and rat genomes, respectively, suggesting that tandemly duplicated genes are a major feature of the mammalian genomes (Figure 2.1). Although tandem duplication has been known as one of the mechanisms of gene duplication for two decades, we still do not know quantitatively how much tandem duplication has contributed to gene duplication in the genome. Here we calculated the proportions of TAG genes after removing single-member clusters from the non-overlapping gene dataset for all three genomes (Table 2.2). Approximately 13000 human genes, 14043 mouse genes, and 12466 rat genes are likely products of gene duplication. Of these duplicated genes, more than 21%, 25%, and 25% in the human, mouse, and rat genomes are TAGs (Table 2.2), respectively, thus suggesting that tandem duplication is a predominant mechanism of gene duplication in these mammalian genomes.

Various genome rearrangements can create fortuitous TAGs (non-real TAGs) from the existing dispersed members of duplicated genes. Our simulation shows that the effect of non-real TAGs on real TAG statistics is negligible (Figure 2.2): the maximum proportion of non-real TAGs among the 10000 simulated samples is only 1.3% in human, 1.4% in mouse, and 2.1% in rat. The observed proportions of TAGs in the genomes are much higher than the values in the simulated samples (p-value=0), suggesting that the occurrence of TAGs is unlikely due to genome rearrangement and random distributions.

TAG size refers to the number of TAG genes in the array. For all species, most of the TAGs are of size two. There are altogether 902 TAG clusters in the human genome and TAGs of size two (616) account for more than 68% of the TAGs. The mouse and rat genomes have 939 and 778 TAG clusters respectively, of which $\sim 60\%$ belong to TAGs of size 2. The distribution of TAG sizes shows similar patterns across the three genomes with the majority of TAGs having only two members and far fewer larger TAGs (Figure 2.3). The mouse and

Table 2.2: Counts and proportions of duplicated genes that are TAGs.

Species	Non-overlapping genes	duplicated genes	TAG genes	Percentage of TAG genes in duplicated genes	Percentage of TAG genes in the genome
Human	19727	12981	2727	21.0%	13.8%
Mouse	21305	14043	3465	24.7%	16.3%
Rat	18468	12466	3163	25.4%	17.1%

rat genomes appear to have more large TAGs than the human genome. Consistent with this observation, the average numbers of genes in TAGs are about 3.7 and 4.2 in mouse and rat versus. 3.1 in human, suggesting differences in duplication activities and capabilities among these genomes.

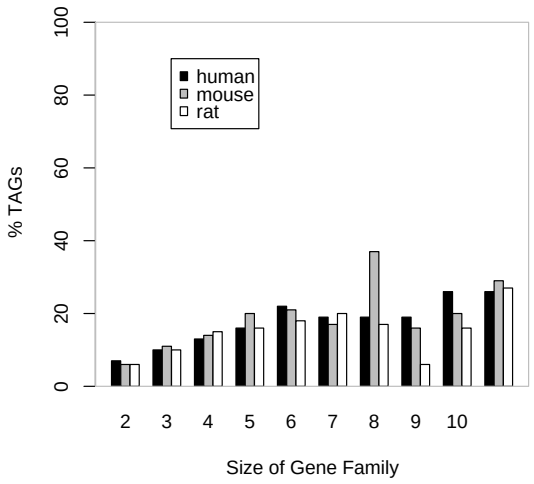

Figure 2.3: Distributions of TAG sizes in human, mouse and rat genomes.

The physical locations of all TAGs in the human, mouse, and rat genomes are shown in Figure 2.4. In all the genomes, there is a great heterogeneity in the TAG distribution along the chromosomes. Hypergeometric tests show that in the human genome, there is significant depletion of TAG deserts in chromosomes 17 and 22, enrichment of TAG deserts in chromosomes 8 and 13, depletion of TAG forests in chromosome 10, and enrichment of TAG forests in chromosomes 9 and 19 (Table 2.3). Furthermore, we observed that at least one subtelomeric region in twelve chromosomes are TAG deserts and the remaining subtelomeric regions have either no genes or low TAG densities, suggesting that subtelomeric regions are not the preferred locations for TAGs. In contrast, pericentromeric regions appear to be

concentrated with TAGs. The proportions of genes that are TAGs in these regions are high and in fact, thirteen chromosomes have at least one pericentromeric region which is TAG forest.

Table 2.3: Human: Analysis of Enrichment/Depletion of TAG desert/forest.

CHR	I	oD	eD	pValue(D)	oF	eF	pValue(F)
1	46	12	16	0.1286	7	5	0.2117
2	48	20	17	0.1863	3	5	0.2168
3	39	16	14	0.2474	4	4	0.5907
4	38	16	13	0.2088	3	4	0.4003
5	36	17	13	0.077	3	4	0.4459
6	34	14	12	0.2638	3	4	0.4939
7	31	9	11	0.315	3	3	0.5694
8	29	17	10	0.0061	1	3	0.1598
9*	25	6	9	0.1738	6	3	0.0412
10*	27	8	10	0.3631	0	3	0.0433
11	26	8	9	0.4174	5	3	0.1348
12	26	8	9	0.4174	2	3	0.4574
13*	19	11	7	0.0308	2	2	0.6664
14	18	7	6	0.4426	1	2	0.4064
15	17	4	6	0.2371	1	2	0.4388
16	15	4	5	0.3565	2	2	0.4914
17*	15	0	5	0.0015	1	2	0.5086
18	15	5	5	0.572	1	2	0.5086
19*	11	1	4	0.0606	4	1	0.0223
20	12	3	4	0.3507	1	1	0.625
21	8	4	3	0.2877	2	1	0.2081
22*	7	0	2	0.0492	1	1	0.5501
X	30	9	11	0.363	4	3	0.4048
Y	6	2	2	0.6521	2	1	0.128

In the mouse genome, there is significant depletion of TAG deserts in chromosomes 6 and 11 and enrichment of TAG deserts in chromosome 15 (Table 2.4). In the rat genome, there is significant depletion of TAG deserts in chromosomes 1, 4, and 10, enrichment of TAG deserts in chromosomes 2 and 5, depletion of TAG forests in chromosome 9, and enrichment of TAG forests in chromosomes 1 and 4 (Table 2.5). We note that first, no tests remain significant after the Bonferroni correction for multiple testing [88]; second, excluding chromosome Y from the dataset produced similar results for all species.

Table 2.6 shows the chromosome orientation of TAG members in the three genomes. The proportions of TAG gene pairs with parallel orientation ($\Rightarrow\Rightarrow$ or $\Leftarrow\Leftarrow$) in human, mouse,

Table 2.4: Mouse: Analysis of Enrichment/Depletion of TAG desert/forest.

CHR	I	oD	eD	pValue(D)	oF	eF	pValue(F)
1	39	10	12	0.3378	3	5	0.207
2	36	7	11	0.1038	4	5	0.461
3	31	13	9	0.0999	4	4	0.6013
4	30	13	9	0.0781	3	4	0.4139
5	30	9	9	0.5903	4	4	0.6301
6*	30	4	9	0.0266	7	4	0.089
7	27	5	8	0.1291	6	4	0.1357
8	25	9	8	0.3198	2	4	0.3275
9	24	6	7	0.3849	4	3	0.4008
10	26	12	8	0.0555	2	4	0.3005
11*	24	2	7	0.0104	5	3	0.2043
12	23	9	7	0.2245	1	3	0.1624
13	22	10	7	0.0869	6	3	0.0588
14	23	7	7	0.5626	4	3	0.368
15*	20	11	6	0.0153	1	3	0.2271
16	19	5	6	0.4723	3	3	0.4765
17	18	5	5	0.534	3	3	0.439
18	18	7	5	0.2754	1	3	0.2816
19	12	1	4	0.0825	2	2	0.4909
X	32	8	10	0.338	3	4	0.3595
Y	1	0	0	0.7	0	0	0.8667

Table 2.5: Rat: Analysis of Enrichment/Depletion of TAG desert/forest.

CHR	I	oD	eD	pValue(D)	oF	eF	pValue(F)
1*	53	13	20	0.0232	12	7	0.0346
2*	51	26	19	0.031	7	7	0.5311
3	34	11	13	0.3132	6	4	0.2904
4*	37	8	14	0.0238	9	5	0.0434
5	34	13	13	0.5473	6	4	0.2904
6*	29	17	11	0.0159	1	4	0.0811
7	28	15	11	0.0605	3	4	0.4751
8	25	6	10	0.1037	5	3	0.2285
9*	22	9	8	0.4616	0	3	0.0405
10*	22	3	8	0.0115	6	3	0.0583
11	17	10	6	0.0614	1	2	0.3145
12	9	3	3	0.5374	0	1	0.274
13	22	5	8	0.1001	2	3	0.4208
14	22	8	8	0.5384	2	3	0.4208
15	21	9	8	0.3933	3	3	0.5456
16	18	9	7	0.2002	1	2	0.2834
17	19	9	7	0.2605	4	2	0.2376
18	17	6	6	0.5219	1	2	0.3145
19	11	4	4	0.5931	0	1	0.2049
20	11	5	4	0.4069	1	1	0.5581
X*	32	13	12	0.4358	1	4	0.0561

and rat are 68%, 76%, and 72%, respectively. Therefore, majority of neighboring members in the TAGs are on the same strand. Interestingly, the proportion of gene pairs of convergent orientation ($\Rightarrow\Leftarrow$ is roughly the same as that of divergent orientation $\Leftarrow\Rightarrow$ in all species. We also calculated the genome proportions of gene pairs for the three types of orientations and compared them with the observed orientation in TAGs. For all species, the proportion of gene pairs with parallel orientation is much higher in TAGs than in the genome and the percentages of gene pairs with convergent or divergent orientations in TAGs are only about half of that in the genome. The chi-square "goodness of fit" test [87] shows that the distribution of different types of orientations in TAGs is significantly different from that of all genes in the genome for all species (df=2, p-value= $2.2e^{-16}$).

The pattern of the distribution of the TAG gene orientations is distinctly different from that of all genes in the genome, so do the physical distances between TAG genes also show a different pattern from that of all genes in the genome? The cumulative distributions of the intergenic distances for both TAG genes and all genes in the genome are shown in Figure 2.5. In the human, mouse, and rat genomes, the gene pairs with convergent orientation tend to have shorter intergenic distances than those with parallel orientation which in turn tend to have shorter distances than those with divergent orientation.

Table 2.6: Observed number of parallel, convergent, and divergent orientation among gene pairs, with their respective proportions in parentheses.

	Human		Mouse		Rat	
Parallel	10068	1246	11329	1913	9520	1722
	(51.1%)	(68.3%)	(53.2%)	(75.7%)	(51.7%)	(72.2%)
Convergent	4811	285	4975	305	4448	320
	(24.4%)	(15.6%)	(23.4%)	(12.1%)	(24.2%)	(13.4%)
Divergent	4815	294	4975	308	4448	343
	(24.4%)	(16.1%)	(23.4%)	(12.2%)	(24.2%)	(14.4%)

We compared GO categories for molecular function, biological process, and cellular components for TAG genes. Since GO terms are hierarchical and there are many possible levels one can use to test for functional enrichment, we chose for simplicity only the top ten most represented GO categories with known molecular function to examine functional associations in TAGs (Table 2.7). The top ten GO categories are similar among human, mouse, and rat, except that the ranking of each category differs (Tables 2.7, 2.8, 2.9). For example, olfactory receptor activity is the most represented molecular function in rat, and it ranks third in human and forth in mouse. The results demonstrate that genes with the molecular function of either binding or receptor activity tend to be TAGs in these mammalian genomes. The analysis of biological process and cellular component also show a similar pattern. Interestingly, for all three species, duplicated genes that are not TAGs show a ranking of GO categories very similar to TAGs (Tables 2.10, 2.11, 2.12).

Table 2.7: Top ten most represented molecular functions in TAG genes of the human, mouse, and rat genomes.

Species	Rank	Molecular Function	Total	Corrected p-value
human	1	Receptor activity	428	0.0
	2	Metal ion binding	336	0.0
	3	Olfactory receptor activity	303	1.0
	4	Zinc ion binding	301	0.0
	5	Nucleic acid binding	171	0.0
	6	Protein binding	133	$7.0E^{-5}$
	7	DNA binding	129	0.28
	8	Transcription factor activity	126	0.50
	9	Calcium ion binding	118	$1.3E^{-4}$
	10	Structural molecule activity	69	0.0
Mouse	1	Receptor activity	1075	1.0
	2	Signal transducer activity	881	1.0
	3	G-protein coupled receptor activity	861	1.0
	4	Olfactory receptor activity	790	1.0
	5	DNA binding	157	0.0
	6	Hydrolase activity	152	0.0
	7	Protein binding	130	0.0
	8	Rhodopsin-like receptor activity	127	0.0
	9	Oxidoreductase activity	97	1.0
	10	Peptidase activity	94	0.58
Rat	1	Olfactory receptor activity	1021	1.0
	2	Hydrolase activity	86	0.0
	3	Protein binding	83	0.0
	4	DNA binding	78	0.0
	5	Receptor activity	61	0.0
	6	Oxidoreductase activity	53	0.0
	7	Transcription factor activity	52	$1.8E^{-4}$
	8	Zinc ion binding	51	1.0
	9	G-protein coupled receptor activity	47	0.24
	10	Calcium ion binding	46	0.0

Table 2.8: Top ten most represented cellular components in TAG genes of the human, mouse, and rat genomes.

Species	Rank	Cellular Component	Total	Corrected p-value
Human	1	Integral to membrane	567	0.0
	2	Nucleus	400	1.0
	3	Membrane	235	1.0
	4	Integral to plasma membrane	175	0.50
	5	Extracellular space	126	0.0
	6	Extracellular region	86	0.0
	7	Cytoplasm	70	0.0
	8	Plasma membrane	65	1.0
	9	Intermediate filament	62	0.0
	10	Membrane fraction	46	0.12
Mouse	1	Integral to membrane	1190	0.0
	2	Extracellular space	386	0.03
	3	Nucleus	278	0.0
	4	Membrane	271	0.01
	5	Integral to plasma membrane	183	0.0
	6	Extracellular region	97	0.08
	7	Endoplasmic reticulum	62	0.17
	8	Transcription factor complex	61	1.0
	9	Intermediate filament	49	0.0
	10	Intracellular	44	0.0
Rat	1	Integral to membrane	1249	0.0
	2	Extracellular space	175	0.0
	3	Membrane	141	0.0
	4	Nucleus	102	0.0
	5	Integral to plasma membrane	87	0.0
	6	Extracellular region	63	0.01
	7	Cytoplasm	45	0.0
	8	Endoplasmic reticulum	37	$6.0E^{-5}$
	9	Membrane fraction	34	0.0
	10	Microsome	34	1.0

Table 2.9: Top ten most represented biological process in TAG genes of the human, mouse, and rat genomes.

Species	Rank	Biological Process	Total	Corrected p-value
Human	1	Signal transduction	432	0.0
	2	G-protein coupled receptor protein signaling pathway	385	0.0
	3	Regulation of transcription, DNA-dependent	303	0.0
	4	Secondary perception	275	1.0
	5	Perception of smell	216	1.0
	6	Transcription	214	0.0
	7	Transport	85	0.13
	8	Proteolysis and peptidolysis	83	$1.0E^{-5}$
	9	Immune response	75	$2.0E^{-5}$
	10	Cell adhesion	66	0.15
Mouse	1	Signal transduction	899	1.0
	2	G-protein coupled receptor protein signaling pathway	881	1.0
	3	Perception of smell	792	1.0
	4	Regulation of transcription, DNA-dependent	177	0.0
	5	Transport	158	0.0
	6	Transcription	135	0.0
	7	Proteolysis and peptidolysis	81	1.0
	8	Development	79	$8.3E^{-4}$
	9	Immune response	68	0.02
	10	Response to pheromone	68	1.0
Rat	1	G-protein coupled receptor protein signaling pathway	1082	1.0
	2	Perception of smell	1013	1.0
	3	Transport	86	0.0
	4	Regulation of transcription, DNA-dependent	76	$1.0E^{-5}$
	5	Proteolysis and peptidolysis	57	1.0
	6	Signal transduction	42	0.0
	7	Development	40	0.39
	8	Ion transport	38	0.0
	9	Electron transport	36	$3.7E^{-4}$
	10	Cell adhesion	33	$2.0E^{-5}$

Table 2.10: Top ten most represented molecular functions of non-TAG duplicated genes in the human, mouse, and rat genomes.

Species	Rank	Molecular Function	Total	Corrected p-value
Human	1	Metal ion binding	848	0.5
	2	Protein binding	800	0.44
	3	ATP binding	752	$1.0E^{-5}$
	4	Zinc ion binding	614	$6.0E^{-5}$
	5	Transferase activity	549	6.39
	6	Receptor activity	524	1.0
	7	Transcription factor activity	476	1.0
	8	DNA binding	442	1.0
	9	Hydrolase activity	369	0.44
	10	Calcium ion binding	368	0.023
Mouse	1	Protein binding	982	0.0
	2	DNA binding	770	0.0
	3	Receptor activity	595	0.0
	4	ATP binding	586	0.0
	5	Transferase activity	560	0.0
	6	Hydrolase activity	491	0.05
	7	Kinase activity	410	0.0
	8	Transcription factor activity	408	0.0
	9	Signal transducer activity	318	0.0
	10	Nucleic acid binding	300	$7.0E^{-5}$
Rat	1	Protein binding	698	1.0
	2	DNA binding	468	0.0
	3	ATP binding	430	0.05
	4	Transferase activity	395	0.07
	5	Hydrolase activity	327	0.003
	6	Receptor activity	319	0.0
	7	Transcription factor activity	303	0.0
	8	Zinc ion binding	235	1.0
	9	Kinase activity	220	0.01
	10	Calcium ion binding	203	0.38

Table 2.11: Top ten most represented biological processes of non-TAG duplicated genes in the genomes of human, mouse, and rat genomes.

Species	Rank	Biological Process	Total	Corrected p-value
Human	1	Regulation of transcription, DNA-dependent	776	1.0
	2	Signal transduction	681	1.0
	3	Transcription	519	0.05
	4	Protein amino acid phosphorylation	366	0.0
	5	Transport	281	1.0
	6	Development	263	0.003
	7	G-protein coupled receptor protein signaling pathway	223	$2.8E^{-4}$
	8	Intracellular signaling cascade	212	0.11
	9	Proteolysis and peptidolysis	201	1.0
	10	Cell adhesion	192	1.0
Mouse	1	Regulation of transcription, DNA-dependent	698	0.0
	2	Transport	630	0.0
	3	Transcription	511	$7.6E^{-4}$
	4	Signal transduction	444	0.0
	5	Development	296	0.002
	6	Biological process unknown	286	0.0
	7	Protein amino acid phosphorylation	276	0.0
	8	G-protein coupled receptor protein signaling pathway	268	1.0
	9	Intracellular signaling cascade	212	0.0
	10	Ion Transport	186	0.0
Rat	1	Regulation of transcription, DNA-dependent	470	1.0
	2	Transport	379	0.0
	3	Signal transduction	325	$6.6E^{-4}$
	4	G-protein coupled receptor protein signaling pathway	219	1.0
	5	Protein amino acid phosphorylation	212	0.05
	6	Development	183	0.0
	7	Intracellular signaling cascade	173	0.0
	8	Ion transport	127	0.0
	9	Neurogenesis	124	$7.6E^{-}4$
	10	Cell adhesion	123	$3.0E^{-5}$

Table 2.12: Top ten most represented cellular components of non-TAG duplicated genes in the genomes of human, mouse, and rat.

Species	Rank	Cellular Component	Total	Corrected p-value
Human	1	Nucleus	1532	1.0
	2	Integral to membrane	1072	0.005
	3	Membrane	1003	0.0
	4	Integral to plasma membrane	526	0.002
	5	Cytoplasm	452	0.004
	6	Plasma membrane	244	0.01
	7	Membrane fraction	239	0.26
	8	Mitochondrion	188	$1.1E^{-4}$
	9	Cytoskeleton	185	1.0
	10	Golgi stack	185	1.0
Mouse	1	Integral to membrane	1486	0.0
	2	Nucleus	1329	0.0
	3	Membrane	948	0.0
	4	Extracellular space	791	0.0
	5	Cytoplasm	399	0.12
	6	Cellular component unknown	286	0.0
	7	Intracellular	283	0.0
	8	Mitochondrion	256	0.02
	9	Integral to plasma membrane	199	0.37
	10	Plasma membrane	189	0.45
Rat	1	Integral to membrane	964	0.0
	2	Nucleus	830	0.0
	3	Membrane	623	0.0
	4	Extracellular space	507	0.0
	5	Integral to plasma membrane	361	0.0
	6	Cytoplasm	351	0.003
	7	Intracellular	201	$2.0E^{-5}$
	8	Plasma membrane	176	0.0
	9	Mitochondrion	170	0.0
	10	Membrane fraction	167	0.0

2.4 Discussion

2.4.1 Significance of tandem duplication

Previous and current studies all suggest that TAGs are a major component of the genome. The percentages of TAG genes in different genomes of plants and animals span a narrow range of 10-17% (10% for *C. elegan* [82]; 17% for *A. thaliana* [109]; 14% for *O. sativa* [107]; 14-17% for human, mouse, and rat (Table 2.2). Moreover, about 21-25% of the duplicated genes in the three mammalian genomes are TAG genes (Table 2.2), suggesting that tandem duplication is a major method of gene duplication in many genomes.

Studies on recent duplication in several mammalian genomes show that intrachromosomal duplications are more common than interchromosomal duplications [23, 18, 30, 110]. Intrachromosomal duplication may include one or more genes and depending on the locations and the mechanisms of duplication, it can be tandem duplication. In fact, the number of intrachromosomal duplicated genes is significantly correlated with the number of TAG genes for the three species (p-value$\ll 1e^{-5}$, Table 2.13).

2.4.2 Contribution of tandem duplication to different sized gene families

The research on TAGs has been largely confined to individual families of TAGs that serve important physiological functions, such as ribosomal RNA genes, histone genes, immunoglobulin genes, and MHC genes. In these large gene families, most of the members arose through tandem duplication. The unanswered question remains as to whether tandem duplication is a more favored duplication mechanism in large gene families than small ones. It is expected that the larger the family is, the more likely the family resorts to tandem duplication as an efficient way of creating more duplicated genes [69].

To examine this issue, we calculated the average proportion of TAG genes in gene families of different sizes for all three genomes. The average proportion of TAG genes in gene families of different sizes ranges from 15% to 33% in three genomes and appear to be higher in large families than small families (Table 2.14). However, although the average proportions of TAG genes and family sizes are positively correlated (i.e., large families tend to have on average more TAG genes than smaller ones) in human (Spearman's rank correlation coefficient $\rho = 0.91$, p-value = 0.00047) and mouse ($\rho = 0.71$, p-value = 0.0275), but not in rat ($\rho = 0.53$, p-value=0.1133), the proportion of TAGs in all individual families and family sizes do not show significant correlation (p-value> 0.05). The observation that large families tend to have on average higher percentages of TAGs than small ones could be due to the possibility that large families have a high likelihood of being tandem through random arrangement. However, our simulation (Figure 2.2) shows that this is unlikely because

Table 2.13: Number of duplicated genes, tag genes, and the proportion of duplicated genes that are TAGs for each chromosome in the human, mouse, and rat genomes.

	Human			Mouse			Rat		
Chromosome	Duplicated	TAG	Prop (%)	Duplicated	TAG	Prop (%)	Duplicated	TAG	Prop (%)
1	1306	319	24.4	864	173	20.0	1755	582	33.2
2	874	122	14.0	1047	256	24.5	811	133	16.4
3	670	98	14.6	731	144	19.7	1079	391	36.2
4	489	91	18.6	755	135	17.9	874	306	35.0
5	560	79	14.1	762	139	18.2	762	178	23.4
6	652	134	20.6	757	261	34.5	521	84	16.1
7	665	132	19.8	1186	450	37.9	756	177	23.4
8	442	53	12.0	668	107	16.0	722	234	32.4
9	511	109	21.3	795	240	30.2	380	49	12.9
10	489	68	13.9	673	148	22.0	889	270	30.4
11	844	302	35.8	1000	288	28.8	229	33	14.4
12	700	155	22.1	482	93	19.3	287	39	13.6
13	219	23	10.5	591	165	27.9	419	86	20.5
14	381	88	23.1	527	138	26.2	426	96	22.5
15	419	52	12.4	546	109	20.0	438	133	30.4
16	476	91	19.1	414	103	24.9	307	39	12.7
17	680	158	23.2	634	183	28.9	395	87	22.0
18	198	38	19.2	349	74	21.2	308	69	22.4
19	900	343	38.1	451	126	27.9	286	26	9.1
20	339	45	13.3	-	-	-	337	101	30.0
21	127	34	26.8	-	-	-	-	-	-
22	308	47	15.3	-	-	-	-	-	-
X	647	126	19.5	769	129	16.8	475	50	10.5
Y	81	20	24.7	17	4	23.5	-	-	-

random permutations of all the duplicated genes yield a tiny amount of non-real TAGs, and for large gene families, which is a much smaller subset of all duplicated genes, the proportions of non-real TAGs should be even smaller.

2.4.3 Distribution of TAGs on the chromosomes

In all species, TAG distribution shows great heterogeneity along the chromosomes with some chromosomes enriched with TAGs and some depleted of TAGs (Figure 2.4). Using the definitions of TAG deserts and forests, we studied TAG enrichment and depletion with respect to individual chromosomes. Interestingly, the chromosomes that have greater than expected numbers of TAG forests tend to have less than expected numbers of TAG deserts (Figure 2.6 and Tables 2.3, 2.4, 2.5), suggesting that TAGs have preferences for chromosomes. Furthermore, using the information on subtelomeric and pericentromeric regions in human, we found that TAGs tend to be enriched in pericentromeric regions and thus have preference for specific locations as well. In this regard, it is worth noting that it has been shown that pericentromeric regions are enriched with recent segmental duplications in human [6, 110]. As aforementioned, some of the recent segmental duplications in human might be in fact tandem duplication. Therefore, the common preference for pericentromeric regions by tandem duplication and recent segmental duplication might not be a coincidence.

An interesting question relevant to the TAG distribution is whether the regions that are enriched with TAGs are also rich in other non-TAG duplicates. Our analysis shows that in human, the two regions that are statistically enriched in TAGs also show enrichment of other non-TAG duplicates, whereas in rat, the regions that are rich in TAGs show depletion of non-TAG duplicates.

2.4.4 Distribution of TAG sizes

It has been observed that the majority of TAGs have only two members in the array in many genomes such as *A. thaliana* [109], *C. elegans* [82], and rice [107], suggesting that it might be a rather general phenomenon in eukaryotes (Figure 2.3). The distribution of TAG sizes can be described by a power law distribution, a common type of distribution that appear in various biological quantities such as the distribution of gene family sizes in different eukaryotes [26] and prokaryotes [44].

Because most TAGs have only two members in the array, we expect that large families contain many small TAGs in order to achieve the large requirement of gene copies. Consistent with the expectation, we observed that large gene families tend to have many small TAGs located on different chromosomes instead of a handful of large tandem arrays. For example, the largest gene family in human is a class that contains zinc finger- containing transcription factors. More than 300 genes of this family are members of 90 TAGs (with 2-18 genes in

the arrays) located on eighteen chromosomes. Of the 90 TAGs, 45 arrays are of size 2, 16 of size 3, 12 of size 4, and 17 of size ≥ 5. Similarly, the largest gene family in mouse is a class that contains olfactory receptor genes. Almost 800 genes in this family are products of tandem duplication and are located on seventeen chromosomes. The largest TAG in this family has 55 members and is located on chromosome 10. Of the 68 TAGs in this family, 12 arrays are of size 2, eight of size 3, one of size 4, and 47 of size ≥ 5. Our hypothesis is that a large tandemly arrayed cluster is evolutionarily unstable. It is easy to imagine that once the array becomes large, various genome rearrangements such as insertions of transposable elements, inversions, and translocations can interrupt the array and reduce its size. Moreover, as the array size increases, the rate of unequal crossover might increase as well. Consequently, the rate of fluctuation in copy number will increase and so will the instability of the array. The instability of the large array may become deleterious at certain threshold and be acted against by natural selection. Ohno discussed the possible deleterious effect that TAGs could generate due to the fluctuations in array size by unequal crossover and pointed out that genes in tandem array are not stable and have to be able to cope with the fluctuation in gene dosage [69]. Unfortunately, this scenario is speculative since there have been no empirical studies on how or whether the rate of unequal crossover is affected by array size.

Another possible explanation is that large arrays are not the preferred form of arrangement that can satisfy the requirement of highly differentiated functions among array members. For example, a few gene families such as histone genes and ribosomal RNA genes in human are in large TAGs due to high gene dosage requirements. However, for the majority of genes, diversity in function might be more preferred than quantity, as clearly demonstrated by genes with binding, receptor activities and disease resistance functions (Table 2.7).

2.4.5 TAG gene orientations and intergenic distances

Graham defined "tandem arrays" as arrays in which a DNA segment is repeated head-to-tail, with all copies in the same orientation, and suggested that unequal recombination homogenizes head-to-tail tandem arrays, but would cause arrays with oppositely-oriented repeats to undergo disastrous duplication-deletion events which results in these arrays being rare [33]. The definition of TAGs in the current study is somewhat different from that of Graham's. Nevertheless, the results show that compared with the genome, parallel orientation in TAG genes appears to be more favored than divergent or convergent orientations (Table 2.6), corroborating Graham's conjecture, at least in the three genomes. Furthermore, consistent with the great disparity between the proportion of parallel orientations in TAGs and that in the genomes, the intergenic distances of genes in parallel orientations also show the greatest disparity between TAGs and the genomes compared to the distances among genes in convergent and divergent orientations. This raises the question of the evolutionary significance of parallel orientations in TAGs: why TAGs with parallel orientation show distinct patterns from that in the genome. Is there any adaptive significance with parallel orientation or does

Table 2.14: Average proportion of TAG genes in gene families of different size with 95% confidence intervals (CI).

	Human		Mouse		Rat	
Family Size	%TAG genes	CI	%TAG genes	CI	%TAG genes	CI
2	0.07	0.05 - 0.08	0.06	0.05 - 0.08	0.06	0.04 - 0.07
3	0.10	0.08 - 0.12	0.11	0.09 - 0.13	0.10	0.07 - 0.12
4	0.13	0.10 - 0.17	0.14	0.10 - 0.18	0.15	0.11 - 0.19
5	0.16	0.11 - 0.21	0.20	0.14 - 0.26	0.16	0.11 - 0.21
6	0.22	0.16 - 0.29	0.21	0.13 - 0.28	0.19	0.11 - 0.26
7	0.19	0.12 - 0.26	0.17	0.09 - 0.25	0.20	0.11 - 0.29
8	0.19	0.08 - 0.29	0.37	0.26 - 0.47	0.17	0.08 - 0.25
9	0.19	0.09 - 0.28	0.16	0.05 - 0.25	0.06	0.00 - 0.12
10	0.26	0.12 - 0.38	0.20	0.06 - 0.33	0.16	0.01 - 0.28
> 10	0.32	0.22 - 0.29	0.37	0.25 - 0.34	0.39	0.22 - 0.31

the observed pattern simply reflect the pattern of unequal crossover? To our knowledge, there have been no previous studies examining the effect of parallel vs. other types of orientations in TAGs other than Graham's hypothesis. More studies are needed to investigate the underlying mechanism and the nature of this phenomenon.

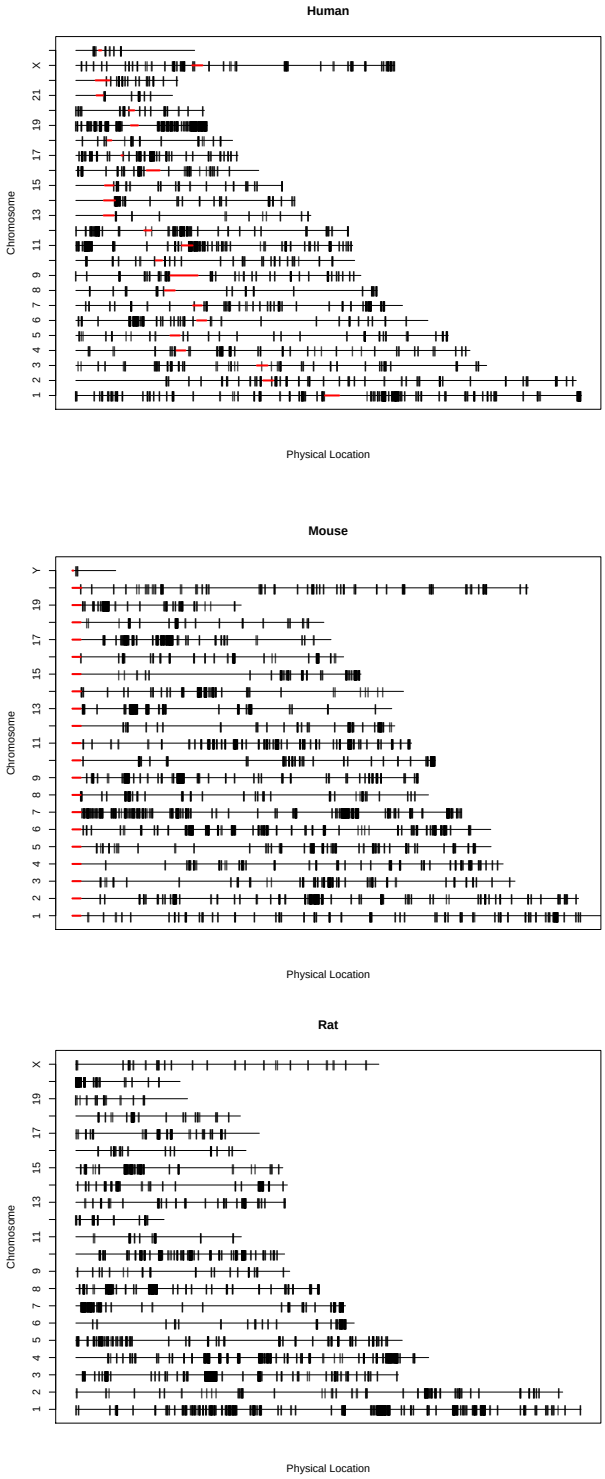

Figure 2.4: Distribution of TAGs on chromosomes in the human, mouse, and rat genomes. The red segments mark the positions of the centromeres when the positions are known.

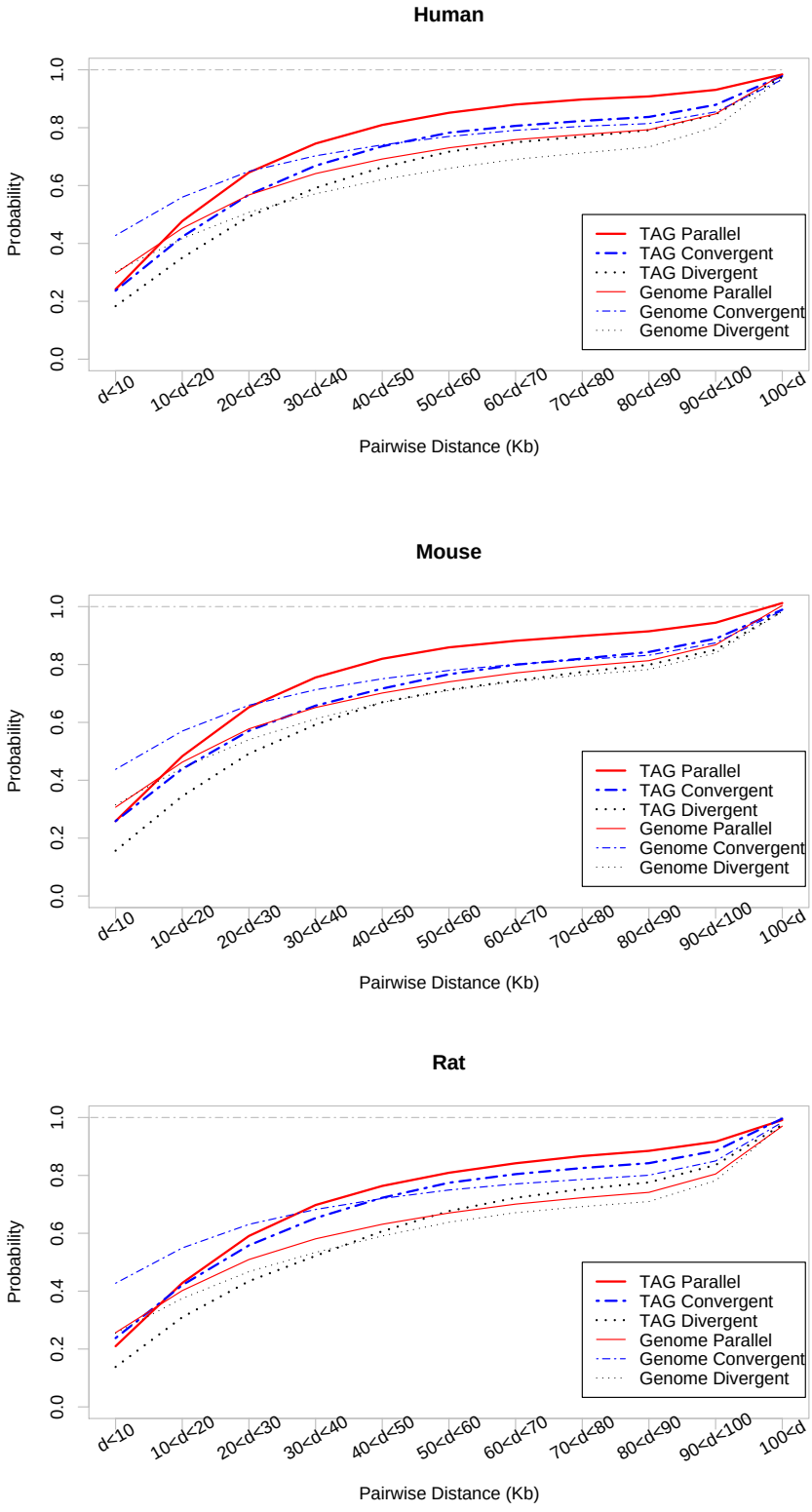

Figure 2.5: Cumulative distribution of the intergenic distances of gene pairs with different orientations in TAGs and the whole genome.

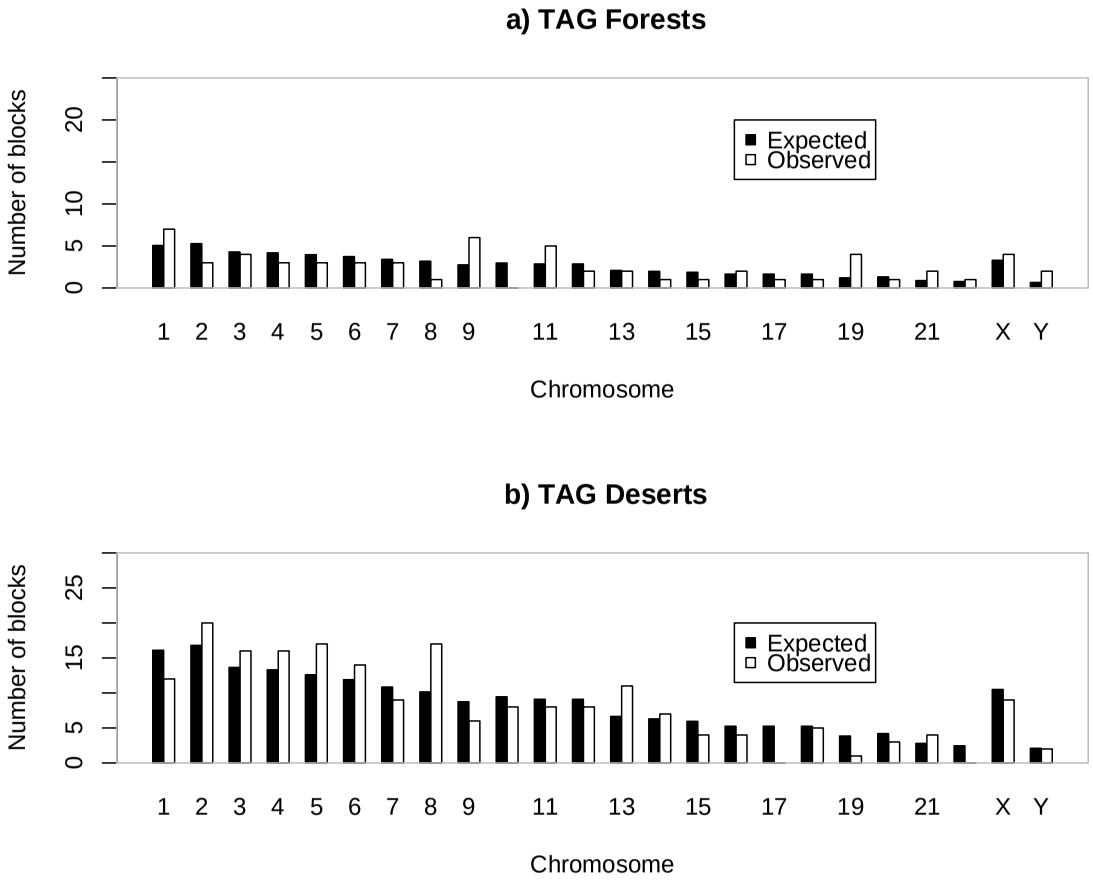

Figure 2.6: Observed and expected number of TAG forests and deserts for each human chromosome. The expected number of TAG forests/deserts on a chromosome is calculated by the total number of TAG forests/deserts in the genome times the proportion of blocks that the chromosome contains.

Chapter 3

Gene Expression of TAGs in Human and Mouse Genomes

3.1 Introduction

Gene expression is an important indicator of gene function. Although detailed gene function is hard to decipher without many biochemical and physiological experiments, it is now much easier to study gene function in terms of gene expression. This is mainly due to the increasing availability of large-scale gene expression profiling. Consequently, many important questions on the evolution of gene function started to be addressed from a gene expression perspective. One of the important questions that has benefited from large-scale gene expression data is regarding the evolutionary divergence of gene expression in duplicated genes. To date, two pictures seem to be emerging from the studies on expression divergence of duplicated genes. First, divergence of gene expression appears to follow the duplication-degeneration-complementation model [27], i.e., after duplication, duplicated genes tend to be expressed in different set of tissues, and the sum of their expressed tissues is that of ancestral single copy gene (e.g., [40, 43, 98, 101]). Second, duplicated genes tend to diverge in expression pattern quickly after duplication (e.g., [35, 36, 43, 60]).

However, little is known about evolutionary divergence of gene expression in tandemly arrayed duplicated genes (TAGs). These genes are duplicated genes that are neighboring to each other on one chromosome and account for nearly one third of all duplicated genes in several completed eukaryotic genomes such as human, mouse, rat, worm, arabidopsis, and rice [13, 83, 107, 109]. Thus studying expression divergence in these genes will provide insights into the function divergence of a large proportion of duplication.

To understand the evolution of gene expression in TAGs, we compiled a list of TAGs in mouse and human. We studied their evolutionary divergence of expression patterns, and addressed how expression divergence is determined or affected by sequence divergence, physical

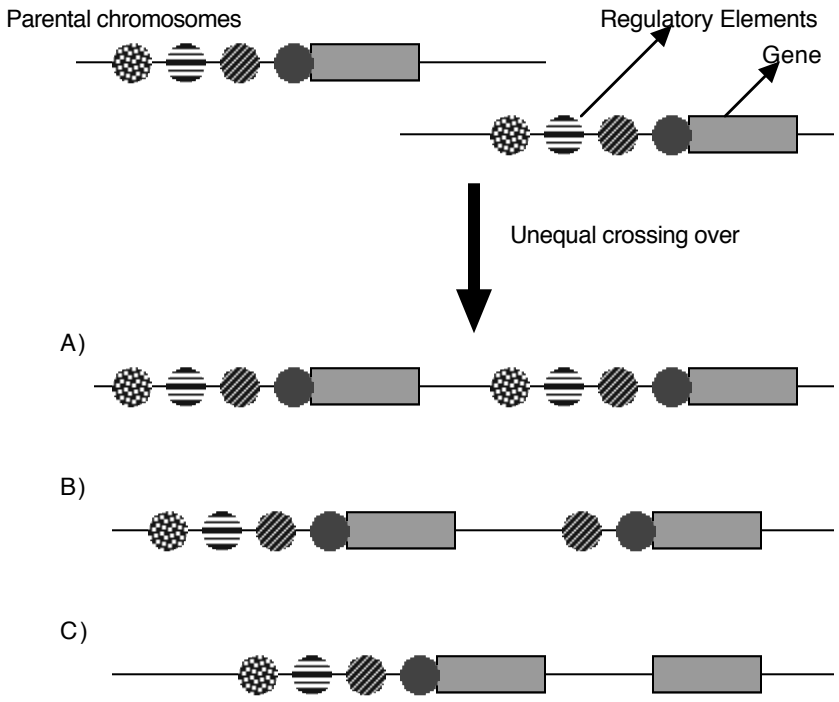

Figure 3.1: Possible scenarios of unequal crossover

closeness, and relative gene orientation. Moreover, TAGs were generated through unequal crossover, depending on the location of the crossover point, as illustrated in Figure 3.1, the downstream duplicated copy can get either the full set of regulatory elements in one extreme situation or partial or no regulatory elements at all in the other extreme situation. In the latter case, the downstream copy is born with "defects". To be functional again, it has to capture and obtain upstream signals for expression. Thus, we expect that the downstream copy should have a narrower expression breadth than its upstream copy. Finally, it has been shown that neighboring genes have highly correlated gene expression patterns in diverse organisms such as human, worm, fly, yeast, and arabidopsis [19, 54, 55, 89, 102]. One of the causes for the higher-than-random expression association between neighboring linked genes was thought to be due to tandem duplication since removing TAGs appears to reduce the overall degree of expression correlation [54]. However, it has not been formally examined how the degree of expression correlation of TAGs, as a special case of neighboring gene pairs, compared to their surrounding neighboring gene pairs. Therefore, here we contrasted the degree of expression association between two members in a TAG with their corresponding neighboring gene pairs to examine the interplay of duplication and physical linkage.

3.2 Materials and Methods

Gene family information and protein sequences for human and mouse were retrieved from Ensembl (version 39, June 06 http://www.ensembl.org). Because the information regarding chromosome location is needed to determine TAGs, only genes with known chromosome locations were kept for further analysis. We used the same method to identify TAGs as in Shoja and Zhang [83]. For the purpose of this study, we considered TAGs with at the most one spacer, as we have previously shown that this definition seems to be a good tradeoff between the most stringent definition which does not allow any spacers and less stringent definitions (see Fig. 1 in [83]). Moreover, because patterns of crossover can be very complex for TAGs with more than two members which in turn can complicate the interpretation of gene orientations, we limited our study to TAGs of size 2. This filtration did not reduce the number of arrays we studied greatly as we have shown previously that most of TAGs have two members in the array for both mouse and human [83]. Altogether, we obtained 1348 and 1618 TAGs in human and mouse, respectively.

Human and mouse gene expression data were from the second version of the Gene Expression Atlas (http:// symatlas.gnf.org), which is a collection of gene expression experiments that surveyed expression patterns of the human and mouse transcriptomes in a panel of 79 human and 61 mouse tissues [92]. This study used Affymetrix HG-U133A array in addition to two custom-made arrays: GNF1H for human and GNF1M for mouse, designed according to human and mouse genome sequences. The results presented here were based on data generated from applying the MAS5 condensation algorithm to the Affymetrix data. The algorithm reports an average difference (AD) value for each gene, which is an estimate of the expression level in a tissue sample [42, 91]. Details of sample annotation and preparation are given in the Su et al. (2004) and at GNF (http://wombat.gnf.org/). Two experimental replicates (samples) for each tissue were obtained in each species. Therefore, we used the average of the two samples for each tissue.

We used the annotation available in Ensembl and GNF to link TAGs with their probe sets. Probe sets containing probes with higher likelihood of cross hybridization between genes (Affymetrix IDs indicated by a suffix of "_x_at" or "_s_at") are considered "suboptimal" reporters of gene expression [43]. For genes with more than one probe sets, if the higher confidence probe sets are available, the lower confidence reporters were discarded and the average of the remaining probe sets were taken. We found that most of TAGs have either only one gene mapped to probe set, or none of the two genes linked to probe sets. Altogether, we were able to obtain a total of 361 and 212 TAGs for human and mouse, respectively.

Two measurements of tissue specificity are employed. One is expression breadth, defined as the number of tissues that the gene has an AD value of greater than 200 [91]. This number corresponds to $\approx 3-5$ copies per cell [91]. The other is tissue specificity index, τ, introduced

by Yanai et al. [105]. The τ of a specific gene i is

$$\tau_i = \frac{\Sigma_{j=1}^{n}\left(1 - \frac{logS(i,j)}{logS(i,max)}\right)}{n-1},$$

where n is the total number of either human or mouse tissues, $S(i,j)$ is gene i's expression in tissue j, and $S(i,max)$ is the highest expression signal of gene i across n tissues. To minimize the influence of noise from low intensities, we let $S(i,j)$ be 100 if it is lower than 100 [57]. The τ value ranges from 0 to 1, with higher values indicating higher tissue specificities. If a gene is equally expressed in all tissues, $\tau = 0$. On the other hand, if a gene is only expressed in a few tissues, τ approaches 1.

We used two measures to quantify similarity in expression profiles between the AD values of two TAG members: Pearson correlation coefficient (r), and the Jaccard index (also known as Jaccard similarity coefficient). Jaccard index evaluates the degree of overlap in the types of tissues that two genes are expressed in and is computed using set relations: $J(A,B) = \frac{A \cap B}{A \cup B}$, where the numerator corresponds to the number of tissues in which both members of a TAG are expressed and the denominator corresponds to the number of tissues in which at least one member is expressed.

The nucleotide sequences of TAG genes were aligned based on the alignments of corresponding protein sequences using the suite of programs in EMBOSS [75]. The number of synonymous substitutions per synonymous site (K_S) and the number of nonsynonymous substitutions per nonsynonymous site (K_A) were calculated by a maximum likelihood approach using PAML [106].

Simulation tests were performed for different properties using the following bootstrap permutation method. Since the parallel orientation has a larger sample size than that of the convergent or divergent orientations, we randomly sample from parallel TAGs when comparing with either of the other two orientations. This method allows comparing a property among groups of same size. The sampling was repeated 10,000 times. Distribution of the average of sampled parallel TAGs was plotted and compared with the average of either of the other two orientations.

3.3 Results

3.3.1 TAG Statistics

Table 3.1 shows counts of TAGs of size two in different orientations. In human, altogether we identified 361 TAGs of size two, with 247 in parallel, 59 in convergent, and 55 in divergent orientations. In mouse, there are 212 TAGs of size two, with 150 in parallel, 28 in convergent, and 34 in divergent orientations.

Table 3.1: Numbers of TAGs of size two in different orientations.

	Orientation			
Species	Parallel	Convergent	Divergent	Total
Human	247	59	55	361
Mouse	150	28	34	212

3.3.2 Expression Divergence

Figure 3.2 shows the distribution of the two measurements of expression similarity or divergence between TAG members for all the TAG genes in human and mouse. Both Pearson's r and Jaccard index J show that the majority of human and mouse TAG genes appear to have diverged in expression: 78% of genes in human have $\rho < 0.5$ and 82% of genes in mouse have $\rho < 0.5$; 31% of genes in human have $J < 0.1$ and 52% of genes in mouse have $J < 0.1$. Both indices show that mouse seems to have more genes that are diverged in their expression.

3.3.3 Expression Divergence and Gene Orientation

The expression patterns were studied by comparing the correlation coefficients of both members of an array and their relative orientations. Table 3.2 shows the ranges and medians of correlation coefficients for TAGs in different orientations. For both human and mouse, the medians and ranges of correlation coefficients between TAG members do not differ greatly among different orientations. The bootstrap permutation test as well as the Wilcoxon signed-rank test show that the orientation of a TAG has no effect on the expression correlation of its two members (p-values range from 0.18 to 0.91).

Table 3.2: Expression correlation in different orientations.

Orientation	Human			Mouse		
	Lower quartile	Median	Upper quartile	Lower quartile	Median	Upper quartile
Parallel	0.02	0.17	0.45	-0.03	0.11	0.36
Convergent	0.05	0.19	0.60	-0.03	0.09	0.35
Divergent	0.06	0.19	0.58	0.05	0.10	0.24
All	0.01	0.18	0.46	-0.02	0.11	0.35

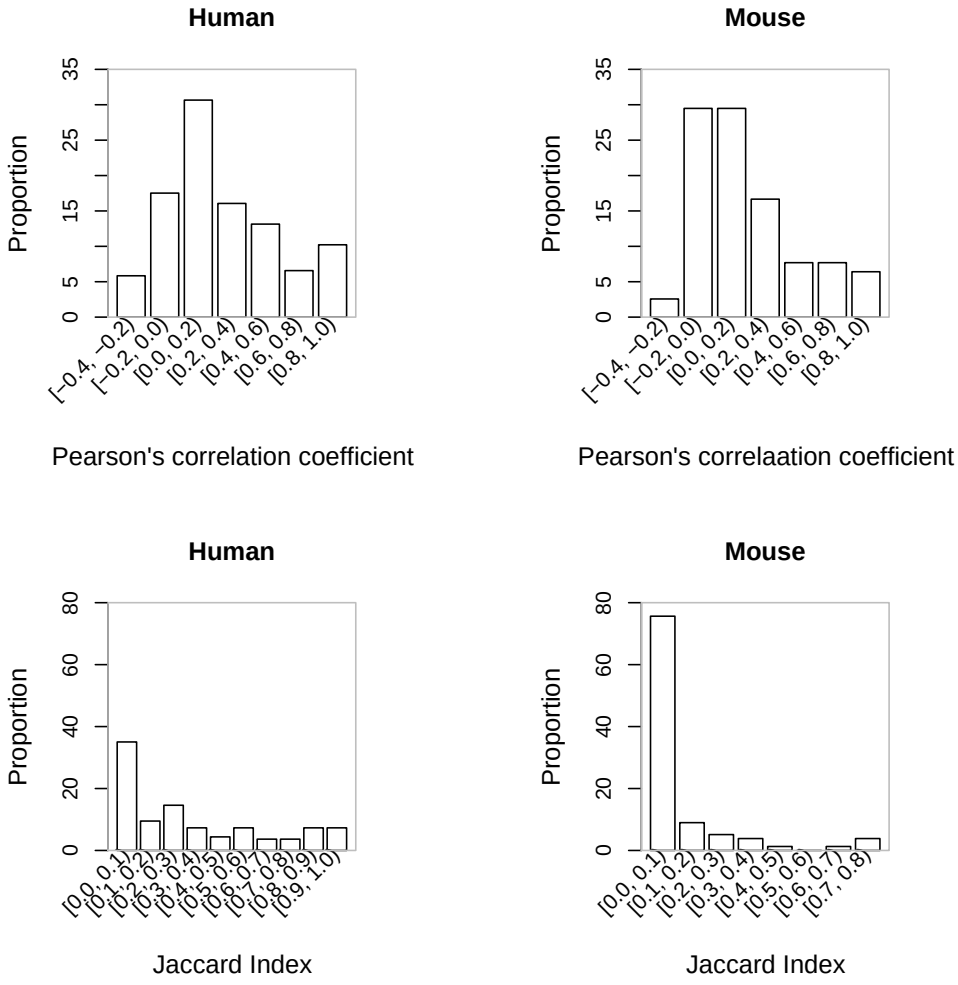

Figure 3.2: Measurements of expression similarities for TAG genes: Pearson Correlation Coefficient and Jaccard Index

3.3.4 Expression Divergence and Sequence Divergence

The basic statistics of synonymous (K_S) and nonsynonymous (K_A) distances are shown in Table 3.3. It is clear that most of the TAGs are very diverged in their coding sequences as more than 81% of the TAGs in human and 83% of the TAGs in mouse have $K_S > 1$.

Table 3.3: Sequence divergence (K_S and K_A) in different orientations.

Divergence	Orientation	Human			Mouse		
		Lower quartile	Median	Upper quartile	Lower quartile	Median	Upper quartile
K_S	Parallel	1.41	5.69	64.40	1.58	8.25	63.84
	Convergent	1.18	3.81	65.94	2.02	41.67	67.32
	Divergent	1.85	32.35	71.33	2.23	11.81	60.47
	All	1.37	7.35	64.81	1.54	8.31	64.47
K_A	Parallel	0.29	0.44	0.61	0.28	0.46	0.63
	Convergent	0.19	0.39	0.54	0.25	0.45	0.59
	Divergent	0.26	0.48	0.63	0.44	0.51	0.66
	All	0.27	0.44	0.61	0.29	0.48	0.63

The correlation between K_S and r of expression similarity is negative but not significant (Spearman's $\rho = -0.08$, $p-value = 0.13$ for human; $\rho = -0.06$, $p-value = 0.34$ for mouse). Since saturation at synonymous sites might affect estimates of K_S, we limited $K_S < 1$ and still found no significant correlation for mouse(human: $\rho = -0.26$, $p - value = 0.03$ mouse: $\rho = 0.002$, $p - value = 0.99$). The correlation between K_A and r of expression similarity is negative but not significant(human: $\rho = -0.06$, $p - value = 0.23$, mouse: $\rho = -0.006$, $p - value = 0.93$).

Table 3.3 also shows sequence divergence for different orientations. We applied both bootstrap permutation tests and Wilcoxon signed-rank tests to examine whether relative gene orientation in TAGs has any effect on sequence divergence (K_S and K_A). In both species, all of the pairwise comparisons of sequence divergence among different orientations showed no significance (p-value ranges from 0.33 to 0.88).

3.3.5 Expression Divergence and Intergenic Distances

We examined the effect of intergenic distances on the expression divergence of TAGs. Table 3.4 shows the ranges and medians of intergenic distances for all TAGs and TAGs with different gene orientations separately. When considering all TAGs, a negative correlation between intergenic distances and expression correlation is observed in human ($\rho = -0.15$, p-value=0.004) but not in mouse ($\rho = 0.06$, p-value=0.37). When separating TAGs into groups

of different orientations, a correlation between expression correlation and intergenic distances is observed only for TAGs with parallel orientation in human ($\rho = -0.14$, p-value=0.03).

Table 3.4: Intergenic Distances (Kb) in different orientations.

Orientation	Human			Mouse		
	Lower quartile	Median	Upper quartile	Lower quartile	Median	Upper quartile
All	9.35	23.21	51.60	8.70	21.14	47.35
Parallel	7.99	18.61	39.88	6.48	15.00	32.21
Convergent	8.61	19.02	31.45	5.85	20.40	42.66
Divergent	7.12	23.09	79.32	17.00	27.29	48.99

We also examined effects of spacers on expression divergence of TAGs, as spacers are effectively increasing the intergenic distances of two neighboring TAG genes. Spacers were defined as genes that have a BLASTP e-value higher than 10^{-10} with other members in the array. Because the sample size of TAGs with 1 spacer is very small for both human and mouse, we performed bootstrap resampling tests and found that TAGs with one spacer do not show significant higher divergence in expression patterns than ones without spacers (human: p-value=0.07 and mouse: p-value =0.9).

3.3.6 TAGs versus Neighboring Linked Genes

To examine the effect of physical linkage on expression divergence of TAGs, we identified neighboring non-TAG gene pairs either to the immediate left or right side of TAGs, and compared their expression divergences with those of TAGs. We were able to identify 105 neighboring non-TAGs pairs in human and 62 in mouse. For these pairs, we calculated the expression correlation r for these genes, and then applied paired t-tests to compare expression correlation difference for the group of neighboring non-TAG gene pairs with the group of their corresponding neighboring TAG gene pairs. Results of the tests show that the two paired groups are not significantly different from each other in both human ($p - value = 0.17$) and mouse ($p - value = 0.24$).

We are also interested in whether TAGs with parallel orientation have shorter intergenic distances than their neighboring non-TAG gene pairs. In both species, the average intergenic distance of corresponding neighboring pairs were greater than that of the parallel orientated TAGs. However, paired t-tests show that the difference between these paired groups is not significant for both species (Human: p-value =0.1; Mouse: p-value=0.3)

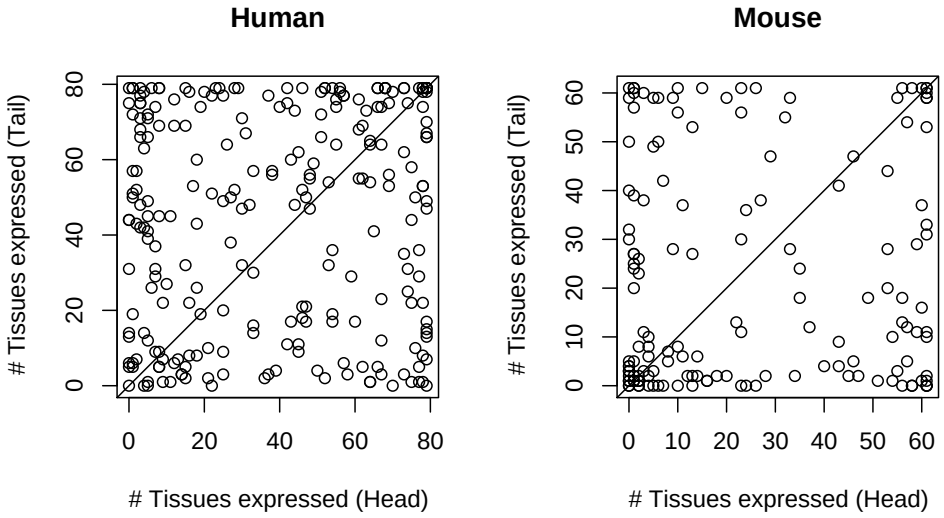

Figure 3.3: Dotplot of number of tissues for head and tail genes in human and mouse genomes. The line refers to the equal numbers of expressed tissues in upstream and downstream genes.

3.3.7 Expression Patterns of Upstream and Downstream Genes

The expression patterns of upstream and downstream genes in parallel orientation were compared both in terms of the number of tissues that the two genes are expressed and their tissue specificity. Figure 3.3 shows the numbers of tissues that upstream and downstream copies are expressed, respectively. In human, there are 96 TAGs with upstream genes more widely expressed than downstream genes, whereas 139 TAGs with the opposite pattern, and 12 upstream and downstream genes equally widely expressed. In mouse, there are 77 TAGs with upstream genes more widely expressed than downstream genes, whereas 65 TAGs with the opposite pattern, and 8 upstream and downstream genes equally widely expressed. In terms of tissue specificity, in human, there are 103 TAGs with upstream genes less specific than downstream genes, whereas 137 TAGs with the opposite pattern, and 7 upstream and downstream genes with the same tissue specificity. In mouse, there are 76 TAGs with upstream genes less specific than downstream genes, whereas 72 TAGs with the opposite pattern, and 2 upstream and downstream genes with the same tissue specificity.

3.4 Discussions

Gleaning indications on possible divergence of gene functions using expression data becomes a routine practice in understanding the evolution of duplicated genes (e.g., [35, 36, 43, 60, 100]).

For instance, Gu et al. [36] examined 400 duplicate gene pairs in yeast for their expression divergence using microarray data and found that more than 40% of the gene pairs in the study show diverged expression pattern even when $K_S < 0.1$ and more than 80% for $K_S > 1.5$. Similarly, Makova and Li showed that of the 1404 duplicate gene pairs that they studied in human, more than 73% show diverged expression in at least one tissue when $Ks < 0.064$ and the percentage increases to 90% as $K_S > 1.2$ [60]. Therefore, both studies suggest that expression patterns of duplicate genes diverge rapidly after duplication. Furthermore, both studies show that expression similarity measured by Pearson's r is significantly negatively correlated with synonymous sequence divergence measured by K_S.

In addition, Gu et al. [36] found that there is a weak correlation between expression similarity and K_A when $K_A < 0.7$. This negative correlation becomes much higher for $K_A < 0.3$. They noted that the 0.3 selection is arbitrary and used two other values ($K_A < 0.25$ and $K_A < 0.35$) and found a similar negative correlation. Consistent with Gu et al's observation in yeast, Makova and Li also found a weak but significant negative correlation between expression similarity and K_A for $K_A < 0.7$ in the human data, and the negative correlation becomes stronger when limiting the dataset to gene pairs with $K_A < 0.2$. Taken together, the two studies in yeast and human suggest that expression divergence and protein sequence divergence is coupled during the early stage of post gene duplication.

Contrary to the findings of Li and collaborators in an earlier study, Wagner [100] found no significant correlation between expression divergence and protein sequence divergence in 144 yeast duplicated genes and concluded that expression divergence of duplicated genes is decoupled with the divergence of protein sequences. The data Wagner used was based on the expression of duplicated genes measured at multiple time points in 4 physiological processes in yeast. The data analyzed by Gu et al. contained a total of 400 gene pairs using microarray data from 14 difference processes in the same species. It seems most likely that the dataset used by Wagner was too small to detect any statistical significance. In fact, Wagner's study seems to serve a good analogy to our study as we also did not find any significant correlation between expression divergence and sequence divergence in TAGs, unlike what has been found in the study of human duplicated genes by Makova and Li [60].

We examined our data further in terms of both the number of data points for correlation analysis and different K_S and K_A values to make a close comparison with the study of Makova and Li. One difference is that we used the microarray data of Su et al's 2004 [92], and the study of Makova and Li used an earlier data produced by the same research group. However, this is unlikely the main reason for the discrepancy between the two studies.

Another major difference is that Makova and Li limited their gene pairs to $K_S < 1.4$ and $K_A < 0.7$. Using the criteria, they kept 1230 gene pairs for correlation analysis. In our study, we have 361 TAGs and the majority of them have very diverged sequences (more than 80% have $K_S > 1$, Table 3.3). Applying the same criteria to the 361 human TAGs, we are left with only 75 TAGs. Analysis of the 75 TAGs in human shows that there is a weak negative correlation between expression similarity r and K_S ($r = -0.19$, $p - value = 0.096$);

a weak nonsignificant correlation between r and K_A ($r = -0.18$, $p - value = 0.127$); and the correlation becomes much higher when $K_A < 0.2$ ($r = -0.30$, $p - value = 0.042$). Therefore, our results seem to be largely consistent with what other research groups found before with regards to the correlation between expression similarity and sequence divergence in duplicated genes, except that most of the correlations in our study are not statistically significant, unlike previous studies, which is most likely due to our smaller sample sizes.

Using the same criteria of $K_S < 1.4$ and $K_A < 0.7$, we got 35 TAGs in mouse. No significant correlation was detected between r and K_S ($r = -0.15$, $p - value = 0.396$), and between r and K_A when $K_A < 0.2$ ($r = -0.42$, $p - value = 0.086$), although it is noted that the negative correlation coefficients shown by these 35 TAGs are much higher than the ones calculated based on the entire dataset, suggesting that expression similarity of duplicated genes and their sequence divergence are more strongly coupled at early stages of duplication.

The inequality between the original copy and newly arose one at the onset of the birth of the new gene was noted by Katju and Lynch [47, 48]. They pointed out that previous standard model of gene duplication assumes an exact duplication of its original gene, whereas in reality, partial duplication of the original gene combined with events such as exon shuffling and gene fusion can be also common so as to affect the ultimate fate of newly arising duplicate. They compared the exon-intron structure of duplicated genes that have $K_S < 0.1$ and discovered that about 60% of the duplicated copies exhibit structural divergence, and more than 50% of the duplicated copies that have $K_S = 0$ display structural difference between copies [47, 48]. These observations show that it is common that the newly arising gene was born without the full set of the exons that its ancestral copy does. The actual proportion of incomplete duplications could be even higher than the current estimate as only exon-intron structures were compared between duplicated genes in their study.

Considering the complexity of a gene structure, it is not difficult to imagine that incomplete duplication can also happen to regulatory regions of a gene, in which case, only some portions (0-100%) of the promoter elements of ancestral copy get duplicated and inherited by the newly arising copy. When considering this partial duplication scenario with respect to tandemly arrayed genes, we can see that this kind of incomplete duplication can be achieved mechanistically through unequal crossover.

As illustrated in Figure 3.1, if crossover occurs anywhere upstream of the set of regulatory elements, the downstream gene gets the entirety of regulatory regions and is identical to the upstream copy and their ancestral single copy. Both genes may diverge subsequently following the DDC model either to achieve complementary functions or to obtain new functions that are not present in their ancestral copy. However, if crossover occurs somewhere in the middle of promoter regions, the downstream gene may only get part or none of the regulatory elements that the upstream copy has and is thus born to be "crippled" in terms of gene expression. In the extreme case of born without any regulatory elements, the downstream copy has to capture promoter elements from anywhere upstream of its coding region. Therefore, the initial expression capacity of the downstream copy depends on how many

regulatory elements it gets. Assuming null mutation occurs equally likely to regulatory elements of upstream and downstream copies, it is expected that the downstream copy will have on average a narrower expression breadth and thus higher tissue specificity than its upstream copy.

What we observe in the current study does not support this prediction of more tissue-specific downstream than upstream genes using both expression breadth measurement and tissue specificity index (Figure 3.3). There are several explanations. One is that noise in microarray experiments reduces the reliability of the data and compromise our results. However, we do not expect that the noise in the microarray data will affect one gene copy more heavily than the other; therefore, upstream and downstream genes will be equally likely influenced and will not create any specific patterns that cause us to fail the null hypothesis.

An important factor that can influence our results substantially is the age of the TAGs. If these genes were duplicated a long time ago, which could be the case of current study as the majority of TAGs show a sequence divergence of $K_S > 1$ (Table 3.3). Even if downstream genes did not inherit any regulatory elements at the onset of duplication, they have plenty of time to capture upstream signals and obtain regulatory elements during evolution and be expressed in different tissues. In fact, capturing upstream signal for expression has been reported in a number of cases such as retrotransposed genes. When first inserted into a different chromosomal location than its parental genes, the newly born genes have no regulatory regions. After millions of years of evolution, some retrogenes have been found to obtain regulatory elements from its upstream regions and become expressed (e.g., [58]).

In order to examine closely whether age has an effect on our prediction, we grouped the TAGs into low, medium, and high divergences based on their K_S. We considered only TAGs with divergence of $K_S <= 1.3$ in both human and mouse. There are altogether 47 TAGs in human and 27 TAGs in mouse that satisfy this criterion. The low, medium, and high divergence correspond to K_S intervals of (0,0.3], (0.3,0.6], and (0.6,1.3], respectively. These bins were selected to maintain roughly equal sized groups of genes in each K_S interval. Altogether, the low, medium, and high groups contain 16, 16, and 15 TAGs in human, respectively; and 7, 8, and 12 TAGs in mouse, respectively. We then calculated the proportion of TAGs that upstream gene has wider expression breadth than downstream copy in each group. If age does have a bearing on the proposition, we expect to see that TAGs in the low K_S category should have higher proportion of TAGs that upstream gene is more widely expressed than downstream copy. Figure 3.4 shows the result for both species. The proportions of TAGs that have upstream genes more widely expressed than downstream copy for the three K_S intervals are 62.5%, 25.0%, 73.3% in human, and 57.1%, 50.0%, 50.0% in mouse using the number of tissues genes are expressed; and are 62.5% 56.3%, 46.7% in human, and 57.1%, 37.5%, 50.0% in mouse using tissue specificity index. Therefore, it seems that in recent duplicated TAGs, there is a higher proportion of TAGs that bear our prediction. As sequences are more diverged, new regulatory elements can be captured by the downstream gene and thus evolve to have a new expression pattern. As the distribution of Jaccard Index shows in Figure 3.2, most of the two members in TAGs share little overlap in the tissues they

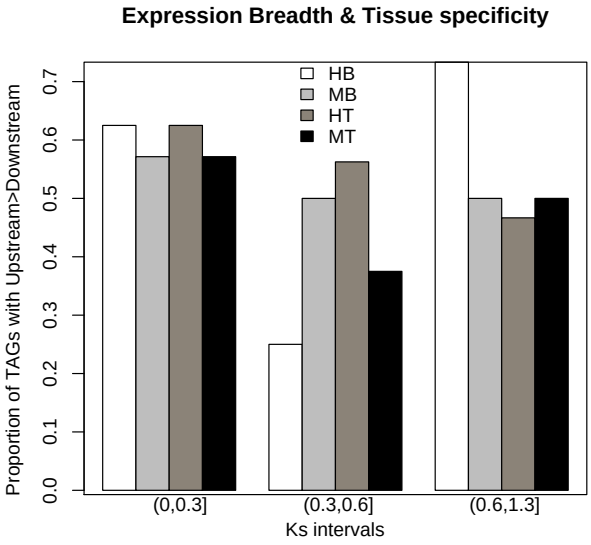

Figure 3.4: Expression breadth and tissue specificity among TAG genes in the human and mouse genomes.

are expressed, suggesting the possibility that some downstream genes might have obtained different regulatory motifs from its upstream region after duplication.

An important assumption that is embedded in the inequality duplication model is that duplication does not contain the complete regulatory elements. However, if most or all of tandem duplication include upstream motifs, then we expect that downstream genes should be equally likely to be expressed in more tissues than upstream copy, and we expect no particular patterns. The intergenic distances between TAG members are large: the distances range from 47bp - 4.3Mbp in human, and 160bp - 0.9Mbp in mouse with median of 23Kbp and 21Kbp in the two species respectively (Table 3.4). This suggests that most of the tandem duplications giving rise to these TAGs might have included the complete set of regulatory elements of their ancestral copy, in which case the model will no longer hold. As many kinds of mutations such as capture and gain of regulatory motifs to both upstream and downstream copies given sufficient time of evolution after duplication, testing the inequality duplication model, especially in the regulatory regions, might require very delicate situations and focusing on young duplicated genes should be a better test for the hypothesis, although at the same time, the data that we currently have are not sufficiently large to reach any concrete conclusion. More studies should be conducted as appropriate data become available.

It has been shown that gene expression is highly correlated between neighboring genes on a

chromosome in organisms such as human [55], *C. elegans* [17, 54], yeast [19], fly [89], and *A. thaliana* [102]. However, the correlation appears to have different causes. For example, in the *C. elegans* genome, tandem duplication seem to be especially common, removing duplicated gene pairs will reduce the degree of expression correlation in the genome. However, in other species such as yeast, the coexpression of neighboring genes seems to be determined by higher order structures such as chromosomal domain level controlled expression activity [19].

The effect of intergenic distance and gene orientation on expression correlation between neighboring genes have been explored previously in different organisms (e.g., [17, 19, 54, 102]). TAGs should be a special case in this perspective as they not only are neighboring genes, but also share certain degree of sequence similarity due to common origin. It seems that neighboring genes in divergent orientation tend to have higher expression correlation than other two types of gene orientations. But the effect of gene orientation does not seem to have very strong effect on degree of expression correlation as the difference in expression correlation for different orientation appears to be small [102]. For example, in *C. elegans*, expression patterns of neighboring gene pairs that are closer together and transcribed either in the same direction (parallel orientation) or divergent direction are found to have strong positive correlation [17]. It has been shown that neighboring genes in the genome of *A. thaliana* are coexpressed [102] suggesting that the reason for coexpression among these genes is mainly due to the orientation of gene pairs. Recently, another study showed that neighboring genes experience continuous concerted expression during evolution, leading to the formation of coexpressed gene clusters [81]. Our current study examined expression similarity in tandemly arrayed duplicated genes. Our results showed that the orientation of a TAG gene pair has no effect on its expression correlation. However, as we discussed above, it is likely that we do not have enough data to detect any statistical significance. More data are needed to examine more closely whether gene orientation in TAGs have any detectable effect on expression divergence.

However, Cohen et al. [19] have shown that although gene pairs that are in divergent orientation show highest expression correlation among the three types of gene orientations, when controlling intergenic distances to be similar among all different orientations, there is no significant difference between distribution of expression correlation among different gene orientations. This is consistent with our results, because the intergenic distances for TAGs with different orientations are not statistically different from each other, so is their expression correlations among different orientations.

Chapter 4

Conclusions

4.1 Summary

Tandemly arrayed genes (TAG) play an important functional and physiological role in the genome. Most previous studies have focused on individual TAG families in a few species, yet a broad characterization of TAGs is not available. As part of this study, we identified all TAGs in the genomes of human, mouse, and rat and performed a comprehensive analysis of TAG distribution, TAG sizes, TAG gene orientations and intergenic distances, and TAG gene functions. TAGs account for about 14-17% of all genes in the human, mouse, and rat genomes and nearly one third of all duplicated genes, highlighting the predominant role that tandem duplication plays in gene duplication. For all species, TAG distribution is highly heterogeneous along chromosomes and some chromosomes are enriched with TAG forests (where more than 30% of the genes are TAGs) while others are enriched with TAG deserts (regions with no TAG genes). Majority of TAGs are of size two for all genomes, similar to the previous findings in *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Oryza sativa*, suggesting that it is a rather general phenomenon in eukaryotes. TAG members have a significantly higher proportion of parallel gene orientation than other non-TAG genes in all species, suggesting that parallel orientation is the preferred form of orientation in TAGs; moreover, TAG members with parallel orientation tend to be closer to each other than the overall neighboring non-TAG genes with parallel orientation. Our analyses of Gene Ontology indicates that genes with receptor or binding activities are significantly overrepresented by TAGs. Interestingly, our Monte Carlo simulation reveals that random genome rearrangements have little effect on TAG quantities, i.e., the probability of finding duplicated gens appearing as TAGs by random chance is negligible. Finally, it is noteworthy to mention that gene family sizes are significantly correlated with extent of tandem duplication, suggesting that tandem duplication is a preferred form of duplication especially in large families.

4.2 Future Research

4.2.1 Identification of All TAGs in the Related Mammalian Genomes

Identification and analysis of TAGs can be extended to other mammalian genomes such as dog, chicken, opossum, and macaque in order to gain a more comprehensive and clear picture of this class of genes. These genomes are completely or nearly completely sequenced and annotated and are subjects of biological and biomedical research due to their important evolutionary niche and application. Moreover, these species along with human, mouse, and rat genomes exhibit a nicely spaced species phylogeny (Figure 4.1). Since chicken is quite distant in evolution from mammals, its inclusion may effect the accuracy of evolutionary inferences. Since the available macaque genome is in its preliminary stages, its inclusion in the TAG study may also effect the accuracy of the TAG analysis and should be awaited for better-quality and more improved assemblies. Opossum splits off from other mammals approximately 130-180 million years ago [34].

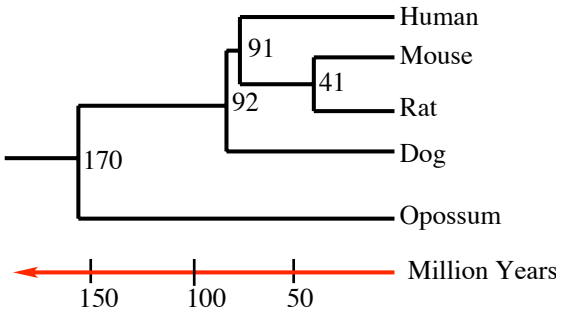

Figure 4.1: Phylogeny of five mammals with estimates of time of speciation.

4.2.2 Determination of Syntenic Blocks among Related Genomes

Synteny is the preserved order of genes between related species. Due to rearrangements, chromosomes vary in size and number among species, but the content of the chromosomes remains much the same, even between distantly related species like humans and mice. Sizes and numbers of syntenies identified among species vary with evolutionary distances separating species; the more closely related species are, the more likely that they share larger syntenic regions. Thus, syntenic regions are better indicators of evolutionary closeness of species than similarities between sequences, for rates of genome rearrangements that may disrupt syntenic regions tend to be lower than rates of nucleotide mutations [49, 70, 71, 80, 79].

Combining synteny information with sequences can be an extremely powerful method for assigning ortholog or paralog relationships to duplicated genes [14].

4.2.3 Assignment of Homology Relationships to TAG Members

A clear distinction between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes [50].

Figure 4.2 depicts ortholog and paralog relationships; while, Figure 4.3 illustrate two possible scenarios complicating the inference of ortholog or paralog relationships. Assuming that B1 is the duplicate of gene A1 in a species and A1 and A2 are ortholog genes in two species. In the gene conversion scenario, gene A1 converts to gene B1. If only sequence identities are used to determine the homology relationships, then A1 and B1 will be misinterpreted as products of a recent gene duplication and both will be considered orthologous to A2. In the gene loss case, loss of A2 and B1 causes A1 and B2 to appear as orthologs; although, they are truly paralogs. Therefore, relying on sequence similarity alone to infer ortholog or paralog relationships can be misleading. Combining sequence comparison, phylogeny, and synteny can be extremely powerful and informative in determining evolutionary relationships [14].

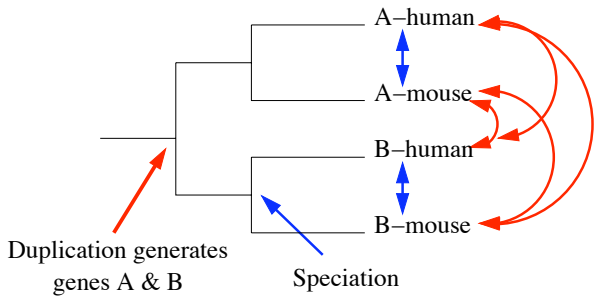

Figure 4.2: Ortholog or paralog relationships among duplicated genes. $\longleftrightarrow$ denotes paralog relationship while $\longleftrightarrow$ corresponds to an ortholog relationship.

In order to determine the ortholog or paralog relationships, the coding regions of TAGs in different species should be aligned based on their protein sequence alignments, using MUSCLE [22]. Also the neighbor-joining algorithm in PAUP [94] and Bayesian analysis in MrBayes [77] should be used to construct phylogenies. Combining phylogeny, synteny,and ortholog or paralog relationships information among gene members in an array can be determined with high confidence. The synonymous distance between genes (K_S) for all pairwise comparisons of putative orthologs may also be computed and compared to distances obtained through comparison of orthologous, single-copy genes to identify abnormalities.

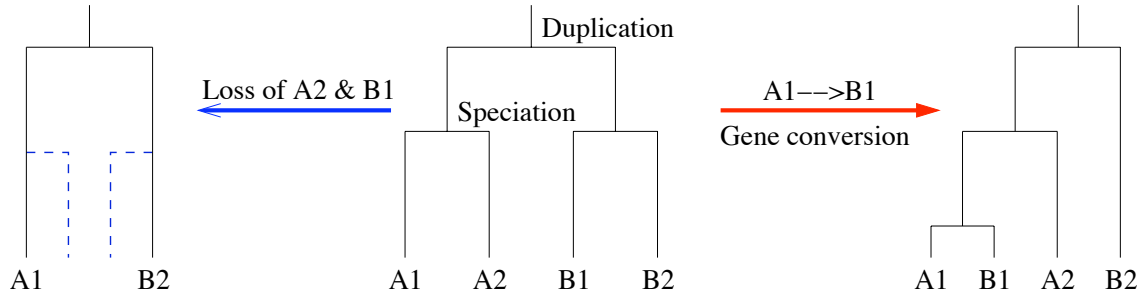

Figure 4.3: Two Scenarios exemplify the complication of the ortholog or paralog relationships: gene conversion and gene loss.

4.2.4 Gene Gain and Loss Evaluation in TAGs

The evolutionary dynamics of some TAG families have been modeled as birth and death processes that model individual gains and losses of array members [66, 63, 15, 67]. Also, a recent study employed a birth and death model to estimate rates of gene duplication and loss in gene families [37].

Both methodologies may be used to determine loss or gain of TAGs and further show whether there has been a complete loss or gain in some species by examining the phylogenetic trees of the corresponding coding sequences. By using ortholog and paralog relationships, gains and losses may be identified. Absence of some array members in a species may be due to a complete loss relative to the other species which has the corresponding gene (it could also be gain of the gene in the other species after speciation). Another possible case is gene movement to a different location via genome rearrangement.

4.2.5 Characterization of Patterns of Concerted Evolution in TAGs

The term "concerted evolution" refers to the tendency of duplicated genes to evolve in unison maintaining nearly identical sequences [114]. There are two types of concerted evolution: unequal crossover and gene conversion [56].

Repeated unequal crossovers can lead to changes in both the number and the sequences of genes in an array (Figure 4.4). The simulations show that repeats can develop from nonreptitious DNA as

a result of the random accumulation of random mutations and random homology-dependent unequal crossovers. These simulations indicate that, under an assumption of neutral evolution, repeated unequal crossovers tends to homogenize the sequences in an array [85]. Therefore, through unequal crossover, the genes in TAGs can evolve synergistically and can have highly identical sequences.

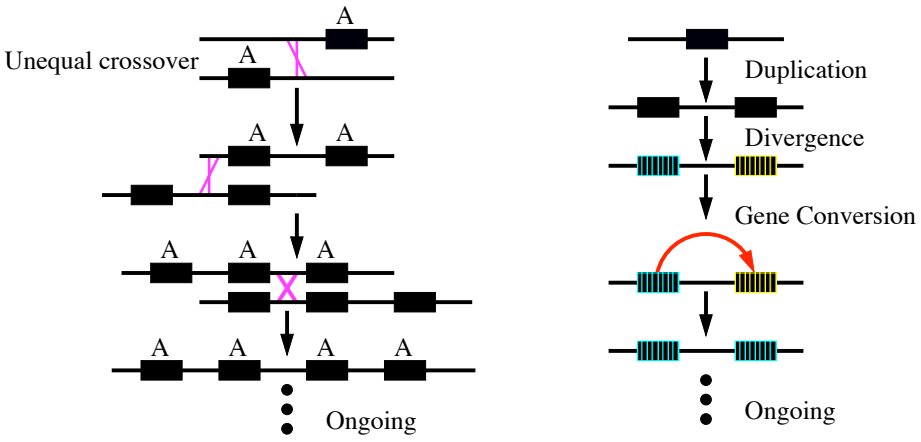

Figure 4.4: Two types of conceted evolution: unequal crossover, and gene conversion

Gene conversion is a nonreciprocal transfer of genetic information during meiotic division (Figure 4.4). After gene conversion, DNA sequence is transferred from one DNA helix (which remains unchanged) to another DNA helix, whoes sequence is altered. As with unequal crossover, gene conversion tends to homogenize TAGs. It is believed that different levels of sequence homogeneity or diversity result from differences in rates of concerted evolution and rates of DNA mutations [31, 93, 96].

The main features associated with concerted evolution, including the locality of gene conversion within a gene, the lengths of converted tracks, and the donor or receptor relationships should be examined in great detail to identify patterns of concerted evolution.

Bibliography

[1] S. Agarwal, S. Sarwai, S. Agarwal, U.R. Gupta, and S. Phadke. Thalassemia intermedia: heterozygous beta-thalassemia and co-inheritance of an a gene triplication. *Hemoglobin*, 26:321–3, Aug 2002.

[2] G. Aguileta, J.P. Bielawski, and Z.H. Yang. Gene conversion and functional divergence in the beta-globin gene family. *J Mol Evol*, 59:177–89, Aug 2004.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–402, Sep 1997.

[4] R.P. Anderson and J.R. Roth. Tandem genetic duplications in phage and bacteria. *Annu Rev Microbiol*, 31:473–505, 1977.

[5] M. Bahuau, I. Laurendeau, A. Pelet, B. Assouline, T. Lamireau, L. Taine, B. Bail, P. Vergnes, S. Gallet, M. Vidaud, S. Lyonnet, D. Lacombe, and D. Vidaud. Tandem duplication within the neurofibromatosis type 1 gene (nf1) and reciprocal t(15;16)(q26.3;q12.1) translocation in familial association of nf1 with intestinal neuronal dysplasia type b (ind b). *J Med Genet*, 37:146–50, Feb 2000.

[6] J.A. Bailey, A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res*, 11:1005–17, Jun 2001.

[7] G. Blanc, A. Barakat, R. Guyot, R. Cooke, and M. Delseny. Extensive duplication and reshuffling in the arabidopsis genome. *Plant Cell*, 12:1093–101, Jul 2000.

[8] G. Blanc, K. Hokamp, and K.H. Wolfe. A recent poly- ploidy superimposed on older large-scale duplications in the arabidopsis genome. *Genome Res*, 13:137–44, Feb 2003.

[9] G. Blanc and K.H. Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16:1667–78, Jul 2004.

[10] A.M. Boutanaev, A.I. Kalmykova, Y.Y. Shevelyov, and D.I. Nurminsky. Large clusters of co-expressed genes in the drosophila genome. *Nature*, 420:666–9, Dec 2002.

[11] C.J. Brown, K. Todd, and R.F. Rosenzweig. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol*, 15:931–42, Aug 1998.

[12] D.D. Brown and I.B. Dawid. Specific gene amplification in oocytes. oocyte nuclei contain extrachromosomal replicas of the genes for ribosomal rna. *Science*, 160:272–80, Apr 1968.

[13] B.S. Gaut C. Rizzon, L. Ponger. Striking similarities in the genomic distribution of tandemly arrayed genes in arabidopsis and rice. *PLoS Comput Biol*, Sep 2006.

[14] S.B. Cannon and N.D. Young. Orthoparamap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*, 4:35. Epub, Sep 2003.

[15] S.B. Cannon, H.Y. Zhu, A.M. Baumgarten, R. Spangler, G. May, D.R. Cook, and N.D. Young. Diversity, distribution, and ancient taxonomic relationships within the tir and non-tir nbs-lrr resistance gene subfamilies. *J Mol Evol*, 54:548–62, Apr 2002.

[16] H. Caron, M. Peter, P. van Sluis, F. Speleman, J. de Kraker, G. Laureys, J. Michon, L. Brugieres, P.A. Voute, A. Westerveld, and et al. Evidence for two tumour suppressor loci on chromosomal bands 1p35-36 involved in neuroblastoma: one probably imprinted, another associated with n-myc amplification. *Hum Mol Genet*, 4:535–9, Apr 1995.

[17] N. Chen and L.D. Stein. Conservation and functional significance of gene topology in the genome of caenorhabditis elegans. *Genome Res*, 16:606–17, May 2006.

[18] J. Cheung, M.D. Wilson, J. Zhang, R. Khaja, J.R. MacDonald, H.H. Heng, B.F. Koop, and S.W. Scherer. Recent segmental and gene duplications in the mouse genome. *Genome Biol*, 4:R47, Jul 2003.

[19] B.A. Cohen, R.D. Mitra, J.D. Hughes, and G.M. Church. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*, 26:183–6, Oct 2000.

[20] T. Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35:125–9, Mar 1973.

[21] S. Draghici, P. Khatri, P. Bhavsar, A. Shah, S.A. Krawetz, and M.A. Tainsky. Onto-tools, the toolkit of the modern biologist: Onto-express, onto-compare, onto-design and onto-translate. *Nucleic Acids Res*, 31:3775–81, Jul 2003.

[22] R.C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, Aug 2004.

[23] E. Eichler and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793–7, Aug 2003.

[24] D. Ellis and S. Malcolm. Proteolipid protein gene dosage effect in pelizaeus-merzbacher disease. *Nat Genet*, 6:333–4, Apr 1994.

[25] J.J. Emerson, H. Kaessmann, E. Betran, and M.Y. Long. Extensive gene traffic on the mammalian x chromosome. *Science*, 303:537–40, Jan 2004.

[26] A.J. Enright, S. Van Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30:1575–84, Apr 2002.

[27] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–45, Apr 1999.

[28] P. Fortina, T. Parrella, M. Sartore, E. Gottardi, V. Gabutti, K. Delgrosso, E. Mansfield, E. Rappaport, E. Schwartz, C. Camaschella, and S. Surrey. Interaction of rare illegitimate recombination event and a poly a addition site mutation resulting in a severe form of alpha thalassemia. *Blood*, 83:3356–62, Jun 1994.

[29] R. Friedman and A.L. Hughes. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol*, 20:154–61, Jan 2003.

[30] R. Friedman and A.L. Hughes. Two patterns of genome organization in mammals: the chromosomal distribution of duplicate genes in human and mouse. *Mol Biol Evol*, 21:1008–13, Jun 2004.

[31] L.Z. Gao and H. Innan. Very low gene duplication rate in the yeast genome. *Science*, 306:1367–70, Nov 2004.

[32] H.M. Goldstone and J.J. Stegeman. A revised evolutionary history of the cyp1a subfamily: Gene duplication, gene conversion, and positive selection. *J Mol Evol*, 62:708–17, 2006.

[33] G.J. Graham. Tandem genes and clustered genes. *J Theor Biol*, 175:71–87, Jul 1995.

[34] J.A. Graves and M. Westerman. Marsupial genetics and genomics. *Trends Genet*, 18:517–21, Oct 2002.

[35] X. Gu, Z. Zhang, and W. Huang. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci USA*, 102:707–12, Jan 2005.

[36] Z. Gu, D. Nicolae, H.H. Lu, and W.H. Li. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*, 18:609–13, Dec 2002.

[37] M.W. Hahn, T. De Bie, J.E. Stajich, C. Nguyen, and N. Cristianini. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15:1153–1160, Aug 2005.

[38] C.T. Harms, S.L. Armour, J.J. DiMaio, L.A. Middlesteadt, D. Murray, D.V. Negrotto, H. Thompson-Taylor, K. Weymann, A.L. Montoya, R.D. Shillito, and G.C. Jen. Herbicide resistance due to amplification of a mutant acetohydroxyacid synthase gene. *Mol Gen Genet*, 233:427–35, Jun 1992.

[39] P.J. Hastings, H.J. Bull, J.R. Klump, and S.M. Rosenberg. Adaptive amplification: An inducible chromosomal instability mechanism. *Cell*, 103:723–31, Nov 2000.

[40] X.L. He and J.Z. Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169:1157–64, Feb 2005.

[41] C.W. Heizmann, G. Fritz, and B.W. Schafer. S100 proteins: Structure, functions and pathology. *Front Biosci*, 7:d1356–68, May 2002.

[42] E. Hubbell, W.M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18:1585–92, Dec 2002.

[43] L. Huminiecki and K.H. Wolfe. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res*, 14:1870–9, Oct 2004.

[44] M.A. Huynen and E. van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol*, 15:583–9, May 1998.

[45] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408:796–815, Dec 2000.

[46] K. Inoue, H. Osaka, N. Sugiyama, C. Kawanishi, H. Onishi, A. Nezu, K. Kimura, S. Kimura, Y. Yamada, and K. Kosaka. A duplicated plp gene causing pelizaeus-merzbacher disease detected by comparative multiplex pcr. *Am J Hum Genet*, 59:32–9, Jul 1996.

[47] V Katju and M. Lynch. The structure and early evolution of recently arisen gene duplicates in the caenorhabditis elegans genome. *Genetics*, 165:1793–803, Dec 2003.

[48] V Katju and M. Lynch. On the formation of novel genes by duplication in the caenorhabditis elegans genome. *Mol Biol Evol*, 23:1056–67, May 2006.

[49] W.J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolutions cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 100:11484–9, Sep 2003.

[50] E.V. Koonin. An apology for orthologs - or brave new memes. *Genome Biol*, 2:COMMENT1005, Apr 2001.

[51] B.F. Koop. Human and rodent dna sequence comparisons: a mosaic model of genomic evolution. *Trends in Genetics*, 11:367–71, 1995.

[52] D. Leister. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet*, 20:116–22, Mar 2004.

[53] T. Lenormand, T. Guillemaud, D. Bourguet, and M. Raymond. Appearance and sweep of a gene duplication: adaptive response and potential for a new function in the mosquito culex pipiens. *Evolution*, 52:1705–12, 1998.

[54] M.J. Lercher, T. Blumenthal, and L.D. Hurst. Coexpression of neighboring genes in caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Res*, 13:238–43, Feb 2003.

[55] M.J. Lercher, A.O. Urrutia, and L.D. Hurst. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet*, 31:180–3, Jun 2002.

[56] W.H. Li. *Molecular Evolution*. Sinauer Associates, Sunderland, Mass., 1997.

[57] B.Y. Liao and J. Zhang. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol*, 23:1119–28, Jun 2006.

[58] M. Long and C.H. Langley. Natural selection and the origin of jingwei, a chimerical processed functional gene in drosophila. *Science*, 260:91–95, 1993.

[59] M. Lynch and V. Katju. The altered evolutionary trajectories of gene duplicates. *Trends Genet*, 20:544–9, Nov 2004.

[60] K.D. Makova and W.H. Li. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res*, 13:1638–45, Jul 2003.

[61] A.C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*, 3:e357. Epub, Oct 2005.

[62] A. McLysaght, K. Hokamp, and K.H. Wolfe. Extensive genomic duplication during early chordate evolution. *Nat Genet*, 31:200–4, Jun 2002.

[63] R.W. Michelmore and B.C. Meyers. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res*, 8:1113–30, Nov 1998.

[64] M. Mondragon-Palomino and B.S. Gaut. Gene conversion and the evolution of three leucine-rich repeat gene families in arabidopsis thaliana. *Mol Biol Evol*, 22:2444–56, Dec 2005.

[65] M. Nei. *Molecular Evolutionary Genetics*. Columbia University Press, New York, 1987.

[66] M. Nei, X. Gu, and T. Sitnikova. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci U S A*, 94:7799–806, Jul 1997.

[67] M. Nei and A.P. Rooney. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, 39:121–52, 2005.

[68] J.P. Noonan, J. Grimwood, J. Schmutz, M. Dickson, and R.M. Myers. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res*, 14:354–66, Mar 2004.

[69] S. Ohno. *Evolution by Gene Duplication*. Springer-Verlag, New York, 1970.

[70] P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Res*, 13:37–45, Jan 2003.

[71] P. Pevzner and G. Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, 100:7672–7, Jun 2003.

[72] W.H. Raskind, C.A. Williams, L.D. Hudson, and T.D. Bird. Complete deletion of the proteolipid protein gene (plp) in a family with x-linked pelizaeus-merzbacher disease. *Am J Hum Genet*, 49:1355–60, Dec 1991.

[73] A.B. Reams and E.L. Neidle. Selection for gene clustering by tandem duplication. *Annu Rev Microbiol*, 58:119–42, 2004.

[74] X.Y. Ren, M.W. Fiers, W.J. Stiekema, and J.P. Nap. Local coexpression domains of two to four genes in the genome of arabidopsis. *Plant Physiol*, 138:923–34, Jun 2005.

[75] P. Rice, I. Longden, and A. Bleasby. Emboss: The european molecular biology open software suite. *Trends in Genetics*, 16:276–7, Jun 2000.

[76] F.M. Ritossa and S. Spiegelman. Localization of dna complementary to ribosomal rna in the nucleolus organizer region of drosophila melanogaster. *Proc Natl Acad Sci U S A*, 53:737–45, Apr 1965.

[77] F. Ronquist and J.P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–4, Aug 2003.

[78] J.R. Roth, N. Benson, T. Galitski, K. Haack, J.G. Lawrence, and L. Miesel. Rearrangements of the bacterial chromosome: Formation and applications. *Escherichia coli and Salmonella Cellular and Molecular Biology*, 2:2256–2276, 1996.

[79] D. Sankoff. Rearrangements and chromosomal evolution. *Curr Opin Genet Dev*, 13:583–7, Dec 2003.

[80] D. Sankoff and J.H. Nadeau. Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proc Natl Acad Sci U S A*, 100:111889, Sep 2003.

[81] M. Semon and L. Duret. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol*, 23:1715–23, Sep 2006.

[82] C. Semple and K.H. Wolfe. Gene duplication and gene conversion in the caenorhabditis elegans genome. *J Mol Evol*, 48:555–64, May 1999.

[83] V. Shoja and L. Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Mol. Biol. Evol.*, 23:2134–41, Nov 2006.

[84] Y.Y. Shyr, A.G. Hepburn, and J.M. Widholm. Glyphosate selected amplification of the 5-enolpyruvylshikimate-3-phosphate synthase gene in cultured carrot cells. *Mol Gen Genet*, 232:377–82, Apr 1992.

[85] G.P. Smith. Evolution of repeated dna sequences by unequal crossover. *Science*, 191:528–35, Feb 1976.

[86] G.P. Smith, L. Hood, and W.M. Fitch. Antibody diversity. *Annu Rev Biochem*, 40:969–1012, 1971.

[87] G.W. Snedecor and W.G. Cochran. *Statistical Methods*. Iowa State University Press, 1989.

[88] R.R. Sokal and F.J. Rohlf. *Biometry*. W. H. Freeman, 1994.

[89] P.T. Spellman and G.M. Rubin. Evidence for large domains of similarly expressed genes in the drosophila genome. *J Biol*, 1:5.1–5.8, Jun 2002.

[90] G.R. Stark. Regulation and mechanisms of mammalian gene amplification. *Adv Cancer Res*, 61:87–113, 1993.

[91] A.I. Su, MP. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A. P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P. G. Schultz, and J.B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA*, 99:4465–70, Apr 2002.

[92] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*, 101:6062–7, Apr 2004.

[93] R.P. Sugino and H. Innan. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. *Genetics*, 171:63–9, Sep 2005.

[94] D. Swofford. Paup*. phylogenetic analysis using parsimony (and other methods) version 4. 2003.

[95] J.S. Taylor, Y. Van de Peer, and A. Meyer. Genome duplication, divergent resolution and speciation. *Trends in Genetics*, 17:299–301, Jun 2001.

[96] K.M. Teshima and H. Innan. The effect of gene conversion on the divergence between duplicated genes. *Genetics*, 166:1553–60, Mar 2004.

[97] J.H. Thomas. Concerted evolution of two novel protein families in caenorhabditis species. *Genetics*, 172:2269–81, Apr 2006.

[98] A. van Hoof. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics*, 171:1455–61, Dec 2005.

[99] N. Vinckenbosch, I. Dupanloup, and H. Kaessmann. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*, 103:3220–5, Feb 2006.

[100] A. Wagner. Decoupled evolution of coding region and mrna expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *PNAS*, 97:6579–84, Jun 2000.

[101] W. Wang, H. Yu, and M. Long. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in drosophila species. *Nat Genet*, 36:523–7, May 2004.

[102] E.J. Williams and D.J. Bowles. Coexpression of neighboring genes in the genome of arabidopsis thaliana. *Genome Res*, 14:1060–7, Jun 2004.

[103] K.H. Wolfe and W.H. Li. Molecular evolution meets the genomics revolution. *Nat Genet*, 33:255–65, Mar 2003.

[104] J. Wootton and S. Federhen. Statistics of local complexity in amino acid sequences and sequence database. *Comput Chem*, 17:149–63, 1993.

[105] I. Yanai, D. Graur, and R. Ophir. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS*, 8:15–24, 1 2004.

[106] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13:555–6, Oct 1997.

[107] J. Yu, J. Wang, W. Lin, S. Li, H. Li, J. Zhou, and P. Ni et al. The genomes of oryza sativa: A history of duplications. *PLoS Biol*, 3:e38, Feb 2005.

[108] J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18:292–8, Jun 2003.

[109] L. Zhang and B.S. Gaut. Does recombination shape the distribution and evolution of tandemly arrayed genes (tags) in arabidopsis thaliana? *Genome Res*, 13:2533–40, Dec 2003.

[110] L. Zhang, H.H. Lu, W.Y. Chung, J. Yang, and J. Li. Patterns of segmental duplication in the human genome. *Mol Biol Evol*, 22:135–41, Jan 2005.

[111] L.Q. Zhang, S.K. Pond, and B.S. Gaut. A survey of the molecular evolutionary dynamics of twenty-five multigene families from four grass taxa. *J Mol Evol*, 52:144–56, Feb 2001.

[112] Y.J. Zhang, Y.R. Wu, Y.L. Liu, and B. Han. Computational identification of 69 retroposons in arabidopsis. *Plant Physiol*, 138:935–48, Jun 2005.

[113] Z.L. Zhang, P.M. Harrison, Y. Liu, and M. Gerstein. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res*, 13:2541–58, Dec 2003.

[114] E.A. Zimmer, S.L. Martin, S.M. Beverley, Y.W. Kan, and A.C. Wilson. Rapid duplication and loss of genes-coding for the alpha-chains of hemoglobin. *Proc Natl Acad Sci U S A*, 77:2158–62, Apr 1980.