7-1-2015

COMPARATIVE STUDY OF GENOMIC FEATURES OF EVOLUTIONARILY YOUNG GENE DUPLICATES

Lijing Bu

Lijing Bu

Candidate

Biology

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Vaishali Katju, Chairperson

Ulfar Bergthorsson

Jeffrey Long

Donald O. Natvig

COMPARATIVE STUDY OF GENOMIC FEATURES OF EVOLUTIONARILY YOUNG GENE DUPLICATES

by

LIJING BU

M.B., Clinical Medicine, Taishan Medical College, 2004
M.S., Medical Genetics, Wenzhou Medical College, 2009

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biology

The University of New Mexico
Albuquerque, New Mexico

July, 2015

DEDICATION

To the youth that is eventually gone

致终将逝去的青春

The love that embraces all life

所有生命的热爱

The curiosity and exploration that never ends

无尽的好奇心和探索

ACKNOWLEDGMENTS

I heartily acknowledge my advisor, Dr. Vaishali Katju, and my committee member, Dr. Ulfar Bergthorsson, for helping me through the academic exploration with their guidance and valuable opinions. Their curiosity, persistence and professional style will continue to inspire me.

I also thank my committee members, Dr. Jeffrey Long, Dr. Donald Natvig, Dr. Thomas Turner and Dr. Steve Poe, for their valuable suggestions to this study and to my professional development. Gratitude is extended to the financial support during graduate school, National Science Foundation Fellowship in Biological Informatics grant (DBI 05327535) award to Dr. Vaishali Katju, NSF grant (DEB-0952342) to Dr. Ulfar Bergthorsson and Dr. Vaishali Katju, and University of New Mexico Department of Biology Graduate/Teaching Assistantships to me. Also thanks to University of New Mexico Center for Advanced Research Computing, for providing access to their high-performance computing system.

Thanks to Dr. Jinyu Wu, my mentor and friend who showed me the magic of Bioinformatics for the first time. Thanks to George Rosenberg, manager of the Molecular Biology Facility, for sharing the computational resources and the fun of software exploration, thank you. Thanks to Postdoc fellow in Dr. Katju's lab, Dr. Anke Konrad, for discussing interesting Bioinformatics questions and editing this dissertation. Gratitude is extended to lab mates, Master's candidate Lucy Packer and Ph. D. student

Brian White in Dr. Katju's lab, Ph. D. candidate James Farslow and Ph. D. student Julie Spencer from Dr. Bergthorsson's Lab.

Thanks to my parents, Guixiang Chen and Jianjun Bu for raising me up with open minds and wise words, and giving me a happy carefree childhood. Thanks to my little sister Dandan Bu, for always believing in me to do the best.

Thanks to Mr. Google, for providing easier ways of accessing knowledge.

Finally to my wife, Lijun Lu, your love and faithful company is the greatest gift of all.

COMPARATIVE STUDY OF GENOMIC FEATURES OF EVOLUTIONARILY YOUNG GENE DUPLICATES

by

LIJING BU

M.B., CLINICAL MEDICINE, TAISHAN MEDICAL COLLEGE, 2004
M.S., MEDICAL GENETICS, WENZHOU MEDICAL COLLEGE, 2009
PH. D., BIOLOGY, UNIVERSITY OF NEW MEXICO, 2015

ABSTRACT

Gene duplication is considered a major contributor to genome evolution and functional diversity. Differences in genomic features (such as structural resemblance, transcriptional orientation, and genomic location) between members of a gene duplicate pair may indicate the possible duplication mechanisms, as well as the evolutionary fates the paralogs may experience. In addition to these genomic features, molecular genetic features, such as differences in codon usage and expression levels may provide further insight into functional changes between paralogs. In this dissertation, multiple genomic analyses were conducted in order to evaluate the differences in genomic and genetic properties between duplicate copies in order to understand the effect duplication mechanisms may have on the divergence of duplicate pairs.

Chapter Two focuses on differing patterns of sequence asymmetry, codon usage, and gene expression levels between the members of gene duplicate pairs belonging to two different populations of paralogs in *Saccharomyces cerevisiae*: ohnologs, which arose via a whole genome duplication (WGD), and small segmental duplication (SSD)

paralogs. It is shown that ohnologs have more highly conserved gene order (synteny) relative to SSD paralogs, despite their greater evolutionary age. Within SSD pairs, the derived paralog (the copy with lower synteny) seems to evolve faster, simultaneously exhibiting a lower CIA value and lower expression levels relative to the ancestral copy. While synteny and evolutionary rate differences were not coupled in ohnolog pairs, the relationship between evolutionary rate asymmetry, CAI, and expression levels was similar to that observed in SSD pairs. These results indicate that codon usage contributes to rate asymmetry in the evolution of gene duplicates in both, ohnologs and SSD paralogs, while differences in synteny (as experienced by SSD pairs, but not very young ohnologs) only affects rate asymmetry in SSD pairs. This may imply relaxed selection on codon usage and the expression of derived copies, potentially leading to the acquisition of novel functions over time.

Chapters Three and Four focus on the effects of structural resemblance and other genomic features on young gene duplicate pairs within the *Homo sapiens* (human) and *Pan troglodytes* (chimpanzee) genomes. The results imply that the majority of gene duplicates in both species are structurally *complete* duplications, encompassing the entire coding region of a gene. The chimpanzee genome additionally contains a large fraction (46%) of retrotransposed young gene duplicates relative to the human genome (13%) which may be due to differences in genome architecture, such as mobile element content between the two genomes. While RNA-mediated processes lead to a majority of inter-chromosomal paralogs, DNA-mediated paralogs reside largely on the same chromosome, in which case inter-paralog distance does not increase over time. These results in

conjunction with results of previous studies in nematodes, yeast, and flies, suggest that the structural resemblance types and location of duplicates are closely linked to the duplication mechanism by which paralog pairs arise. This is also true for closely related species, as illustrated by the comparison of the human and chimpanzee genomes.

The above studies illustrate the relationship duplication span (as illustrated in Chapter Two) and mechanisms (illustrated in Chapters Three and Four) have on the location, synteny, structural resemblance types, and functionality of gene duplicates in different genomes. The findings imply that differences in mechanisms between species can have significant effects on the genome evolution and divergence between even closely related taxa.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

Gene Duplication and Its Evolutionary Role

The process of gene duplication results in additional copies of pre-existing genes in the genome. The scale of gene duplications range from large whole genome duplication (WGD) resulting from polyploidization, to small scale duplication (SSD), generated through DNA-mediated mechanisms (double strand break and repair) or RNA-mediated mechanisms (retrotransposition) (Katju 2012). Since the formal proposal of gene duplication as an important source of genome evolution and functional diversity (Bridges 1936; Muller 1936; Ohno 1970), recent studies during the genomic era based on an abundance of sequence data have continuously revealed details about the trajectory of evolution by gene duplication. With an empirically estimated high rate of 10^{-7} to 10^{-3} per gene per generation (Katju and Bergthorsson 2013; Lipinski et al. 2011), gene duplication constantly introduces new endogenous genomic content into the genome, most of which will become silenced (*nonfunctionalized*) through mutation (Fisher 1935; Haldane 1933) but may be kept as potential material for novel functions during evolution. Given the high rate of origin, even the small proportion of surviving duplicates are abundant. These retained duplicates diverge and eventually follow one of four evolutionary fates: (i) retention of the redundant copy (Clark 1994) if higher expression was selected for (Bergthorsson et al. 2007); (ii) retention of two complementary partial copies through *subfunctionalization* (Force et al. 1999); (iii) *neofunctionalization* (Long et al. 2003)

through mutations (Ohno 1970) or exon shuffling (Gilbert 1978), leading to a shift in function or an acquisition of a new function; (iv) new spatial expression patterns (Gokcumen et al. 2013; Makova and Li 2003) by inheriting new regulatory elements. The relationship between mechanisms of gene duplication and their evolutionary fate is still vague, and is made more complex when one incorporates gene duplicates of all ages (Katju 2012). This is because the early genomic features of gene duplication will experience erosion brought about by later genome recombination events, which may mask their initial evolutionary patterns. Gene conversion is one of the most problematic mechanisms leading to gene duplicate pairs appearing younger (more similar) than they actually are, due to non-reciprocal exchange of homologous sequences (Jeffreys 1979). Pseudogenes, while mostly functionally silent, have been shown to occasionally gain new regulatory elements and, hence, to regain activity (Zheng and Gerstein 2007). Recent studies report that pseudogenes can be transcribed and act as silent RNA that regulates the original gene function (Guo et al. 2009; Pink et al. 2011). A conversion between a gene and its pseudocopy could quickly silence the functional copy and cause gene dysfunction (Chen et al. 2007).

The systematic analysis of young gene duplicates in their early stages of evolution provides a comprehensive understanding of their evolutionary trajectory, and it can help to identify the most influential factors that affect their fate. Projects in this dissertation take advantage of current available genomic information for model organisms and examine the early evolutionary dynamics and genomic features of paralogs in major model organisms under a stringent evolutionary framework that has restrictions for age

and family size. In Chapter Two, the rates of sequence evolution are estimated and compared against codon usage and expression levels for duplicates with two different mechanisms of origin (WGD and SSD) in *Saccharomyces cerevisiae*. Chapters Three and Four focus on the genomic features, particularly the structural types of duplications (*complete*, *partial*, *chimeric* and *retroposed*) within evolutionarily young gene duplicates in the human and chimpanzee genomes.

Gene Duplication and Sequence Asymmetry

The *nonfunctionalization*, *neofunctionalization* and *subfunctionalization* models have predicted asymmetric evolutionary rates for gene duplicates (Cusack and Wolfe 2007). Unless selection is acting against the retention of the ancestral function, one copy has to maintain the original ancestral function under purifying selection. While either the derived or the ancestral copy of a gene duplicate pair can undergo a functional shift, for the purpose of this dissertation, the ancestral copy will refer to the one that retains ancestral location. Hence the copy that inserts into a new location in the genome will be referred to as the derived copy. For *nonfunctionalization*, the derived copy is silenced, and is assumed to be free of selection. Under the *neofunctionalization* model (Ohno 1970), the derived copy develops a novel function and experiences a shift in its functionality, or it may be beneficial to fitness by assuming novel gene dosage under the influence of positive selection or relaxed purifying selection (Hughes 1994; Lynch and Conery 2000; Zhang et al. 1998). If the ancestral copy and the derived copy experience complementary silencing of regulatory or coding elements (*subfunctionalization*),

selection will act differently on different parts of the gene sequences of two copies, but will drive the retention of both copies in order to retain all functionality of the ancestral copy (Force et al. 1999; Lynch and Force 2000). Two copies of subfunctionalized gene duplicates will experience asymmetry in evolutionary rate, as they will have different proportions of nonfunctional and functional regions with the former being rendered free to accumulate mutations. Although early studies detected no asymmetry in the rates of sequence divergence of paralogs (Cronn et al. 1999; Hughes and Hughes 1993; Kondrashov et al. 2002; Robinson-Rechavi and Laudet 2001; Zhang et al. 2002), this was likely due to the inclusion of aged paralogs. Subsequent studies have observed that duplicates with asymmetric rates of sequence evolution could account for up to 17% ~ 30% of all evolutionarily recent gene duplicates (Conant and Wagner 2003; Kellis et al. 2004; Kim and Yi 2006; Nembaware et al. 2002; Van de Peer et al. 2001). The movement of gene duplicates to a new genomic location distant from the ancestral copy often results in the loss of ancestral regulatory elements and the potential acquisition of a novel expression environment for the derived copy (Cusack and Wolfe 2007; Han et al. 2009; Katju and Lynch 2003, 2006; Lynch and Force 2000).

The movement of gene duplicates to a new genomic location distant from the ancestral copy often results in the loss of ancestral regulatory elements and the potential acquisition of a novel expression environment for the derived copy (Cusack and Wolfe 2007; Han et al. 2009; Katju and Lynch 2003, 2006; Lynch and Force 2000). Theoretically, the derived copy should have a faster evolutionary rate than its ancestor due to relaxed selection. From a practical standpoint, it is challenging to determine the

identities of the ancestral versus derived copy for the purpose of measuring their respective rates of molecular evolution. Furthermore, the methods to determine the ancestor or derived status for the two paralogs differ between DNA-mediated and RNA-mediated duplication events. For RNA-mediated duplicates, the derived copy is easy to identify given that it lacks introns and possesses a poly-A tail. For DNA-mediated duplicates, the ancestral and derived copy can be distinguished by determining the extent of conservation of flanking gene order (synteny) compared to an ortholog in the closest outgroup species with a single-copy ortholog (Cusack and Wolfe 2007; Han et al. 2009). Studies have shown that the derived copies have a faster rate of sequence evolution in mammal species including human, macaque, mouse and rat (Cusack and Wolfe 2007; Han et al. 2009).

Chapter Two reports an analysis of gene duplicates with low synonymous sequence divergence in *S. cerevisiae*. Yeast contains a large set of paralogs which were generated during an ancient polyploidization event (WGD). The paralogs generated from this whole-genome duplication event are referred to as ohnologs. The respective ancestral and derived copies within these ohnolog pairs show little to no sequence divergence and serve well as a control group to contrast with the study of rate asymmetry among duplicates originating from small scale duplications (SSD). The comparative analysis between the ancestral and derived copies of ohnologs and small scale duplications will provide further evidence for the reduction in selective constraints and its impact on functional novelty. Additionally, we further tested the potential correlation of

sequence asymmetry and the differences in codon usage and gene expression between two copies, in order to specify possible subjects that selection may have acted on.

Gene Duplicates and Structural Resemblance

In addition to the sequence asymmetry that could develop among paralogs during their evolutionary history, the initial mechanisms of gene duplicate formation may create derived copies with varying degrees of structural resemblance to the ancestral copy which in turn may influence their evolutionary fate (Katju 2012). The different structural classes of duplicates are defined here as (i) *complete* if the region of duplication covers the canonical coding region of the gene, from the start to the stop codon; (ii) *partial* if the region of duplication only covers part of the ancestral gene's coding region; (iii) *chimeric* if the region of duplication covers part of the gene's coding region and the derived copy fuses with neighboring sequences to form new coding regions; (iv) *retroposed* if the derived copy was generated through retrotransposition, during which it loses all introns and gains a poly A tail. In order to obtain a novel function or shift in function, the duplicates with *complete* structural resemblance to ancestral genes often have to wait for the accumulation of neofunctionalizing mutations introduced by relaxed purifying selection or positive selection (Bergthorsson et al. 2007; Ohno 1970), while the *heterogeneous* gene duplicates (*partial*, *chimeric*, and *retrotransposed*) may have no or a shorter "waiting period" because the gain of novel coding regions or cis-regulatory elements rapidly confer novel function (Courseaux and Nahon 2001; Long et al. 2003; Wang et al. 2006; Zhou et al. 2008), or lead to faster rates of subfunctionalization or

neofunctionalization. Several systematic evolutionary studies suggest that these heterogeneous duplicates exist and could account for large proportions of recent gene duplicates in eukaryotic genomes (Katju and Lynch 2003, 2006; Katju et al. 2009; Meisel 2009; Zhou et al. 2008). It has been revealed that the structurally heterogeneous gene duplicates (partial/chimeric) are most prevalent in the worm genome (Katju and Lynch 2003), which likely originate due to duplication events with smaller duplication spans (1.4 kb) relative to the average gene length (2.5 kb). In contrast, the majority of young gene duplicates in the yeast genome are *complete* duplicates, which may be due to (i) on average, large duplication spans (2.5 kb) which are more likely to extend across the complete region of a gene (median length 1.1 kb), and/or (ii) selection against partial/chimeric duplicates with slightly deleterious fitness effects due to increased efficiency of selection in yeast owing to a large effective population size (1×10^{10}) (Katju 2012; Katju et al. 2009).

In order to further explore the patterns of duplication and investigate the similarities and differences between duplication events in various primate genomes, Chapters Three and Four follow the same proposed evolutionary framework for young gene duplicates in the human and chimpanzee genomes as has previously been applied to worm and yeast (Katju and Lynch 2003; Katju et al. 2009). The emerging patterns of structural categories delineated in human and chimpanzee were compared to results from previous studies on *C. elegans*, *S. cerevisiae* and *Drosophila* in order to reveal generalized and unique patterns for the evolutionary and genomic features of young gene duplicates.

CHAPTER TWO

Local Synteny and Codon Usage Contribute to Asymmetric Sequence Divergence of *Saccharomyces cerevisiae* Gene Duplicates

Lijing Bu[1], Ulfar Bergthorsson[1], Vaishali Katju[1, §]

[1]Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA.

[§] Corresponding author

Manuscript published in:

BMC Evolutionary Biology 2011, 11:279 DOI: 10.1186/1471-2148-11-279.

Abstract

Duplicated genes frequently experience asymmetric rates of sequence evolution. Relaxed selective constraints and positive selection have both been invoked to explain the observation that one paralog within a gene-duplicate pair exhibits an accelerated rate of sequence evolution. In the majority of studies where asymmetric divergence has been established, there is no indication as to which gene copy, ancestral or derived, is evolving more rapidly. In this study we investigated the effect of local synteny (gene-neighborhood conservation) and codon usage on the sequence evolution of gene duplicates in the *S. cerevisiae* genome. We further distinguish the gene duplicates into those that originated from a whole-genome duplication (WGD) event (ohnologs) versus small-scale duplications (SSD) to determine if there exist any differences in their patterns of sequence evolution. For SSD pairs, the derived copy evolves faster than the ancestral copy. However, there is no relationship between rate asymmetry and synteny conservation (ancestral-like *versus* derived-like) in ohnologs. mRNA abundance and optimal codon usage as measured by the CAI is lower in the derived SSD copies relative to ancestral paralogs. Moreover, in the case of ohnologs, the faster-evolving copy has lower CAI and lowered expression. Together, these results suggest that relaxation of selection for codon usage and gene expression contribute to rate asymmetry in the evolution of duplicated genes and that in SSD pairs, the relaxation of selection stems from the loss of ancestral regulatory information in the derived copy.

Introduction

The appearance of novel biochemical traits contributing to phenotypic diversity is inextricably linked with the constant input of new genetic fodder via gene and genome duplication. However, a mere duplication of an ancestral locus far from guarantees the origin of a novel gene product and the majority of gene duplicates end up being silenced following a brief evolutionary existence (Haldane 1933; Ohno 1970). For those paralogs that emerge unscathed by deleterious mutations, the first clues as to how paralogs are able to forge an independent evolutionary trajectory may be provided by studying their patterns of expression divergence and relative rates of molecular evolution.

Early studies of DNA sequence divergence between paralogs suggested there was little or no difference between duplicate gene-copies in their rates of evolution (Cronn et al. 1999; Hughes and Hughes 1993; Kondrashov et al. 2002; Robinson-Rechavi and Laudet 2001; Zhang et al. 2002). These results were used to argue against the hypothesis proposed by Ohno that following gene duplication, one copy is under relaxed selection and begins to accumulate previously 'forbidden' mutations (Ohno 1970). However, these analyses may have had limited power to detect differences in evolutionary rates, or rate asymmetry, because they analyzed old duplicates, while an increase in the evolutionary rate is easiest to detect in young gene duplicates (Lynch and Katju 2004). Subsequent studies have demonstrated relatively large rate asymmetry between duplicate genes (Conant and Wagner 2003; Kellis et al. 2004; Kim and Yi 2006; Nembaware et al. 2002; Van de Peer et al. 2001). For instance, 20%–30% of paralogous gene in *Saccharomyces*

cerevisiae displayed significant differences in evolutionary rate (Conant and Wagner 2003) and one or both paralog(s) exhibited accelerated evolution in 17% of the cases (Kellis et al. 2004).

The phrase "gene duplication" appears to imply that all functionally relevant features of an ancestral gene are duplicated and therefore the two resulting gene copies ought to be functionally equivalent. In fact, there may be numerous differences between the two "copies". The derived copy often does not retain the full regulatory element repertoire of the ancestral copy or has some structural or genomic location differences relative to the ancestral gene (Cusack and Wolfe 2007; Katju and Lynch 2003, 2006; Lynch and Katju 2004). These differences suggest that the derived copy might be expected to evolve under divergent constraints relative to the progenitor gene, either due to relaxation of natural selection or due to selection for novel attributes. In the majority of studies where asymmetric divergence has been established, there is no indication as to which gene copy, ancestral or derived, is evolving more rapidly. 'Derived' and 'ancestral' in the context of this study refer to the location of the paralogs in the genome rather than function. Recently, a study of gene duplicates in the mouse genome found that relocated gene copies following duplication, and in particular retrotransposed copies, evolved faster than paralogs in their ancestral location (Cusack and Wolfe 2007). Similarly, a study in four mammalian genomes found that genes that came to reside in a different location following gene duplication were more likely to display evidence of adaptive evolution relative to gene copies that did not relocate (Han et al. 2009).

In the case of a new gene-copy originating from a small-scale duplication (SSD) event and relocating some genomic distance from the ancestral copy, the identity of the ancestral and derived copies can be established by conservation of synteny flanking the paralogs or chromosomal location in comparison to a single-copy ortholog in an outgroup genome (Cusack and Wolfe 2007; Katju and Lynch 2006). Distinguishing the ancestral from the derived copy becomes problematic in the case of whole-genome duplication (WGD henceforth). For example, in the instance of a genome resulting from allopolyploidy where duplicate gene-copies result from hybridization rather than gene duplication, naming ancestral and derived genes has no biological relevance.

Here we examine paralogs with low synonymous divergence in the *S. cerevisiae* genome to determine if it is the derived copy that evolves faster than the ancestral copy following gene duplication. Most duplicates in yeast originated from a WGD event (Kellis et al. 2004; Wolfe and Shields 1997) and for reasons mentioned in the preceding paragraph, it is inappropriate to assign ancestral and derived status to gene copies in the same manner as duplicates arising from SSD events. Gene duplicates that were previously identified as resulting from the WGD event are henceforth referred to as 'ohnologs' and were analysed separately from those resulting from SSD events to test if these two pools of duplicated genes behaved differently with respect to their rates of molecular evolution.

Methods

Identification of Gene Duplicates in S. cerevisiae with Low Synonymous Divergence

We initially selected gene families in the *S. cerevisiae* genome identified in a preceding study (Katju et al. 2009) that comprised only two members and synonymous divergence (K_S) ≤ 0.35. This set had been extracted via the Genome History program (Connant and Wagner 2002) using the following parameters: (i) minimum translated ORF length of 100 aa, (ii) minimum number of aligned residues to accept pair being 100 aa, and (iii) using the BLAST matrix BLOSUM62 and acceptance of all BLAST hits with e ≤ 1e-07. The majority of gene duplicates within this initial sample were identified as '*ohnologs*' (Wolfe 2000) or duplicates originating from a WGD event (Byrne and Wolfe 2005; Dietrich et al. 2004; Gordon et al. 2009; Kellis et al. 2004; Wong et al. 2002). To further increase representation of gene duplicate pairs originating from small-scale duplication (SSD) events, we raised the K_S cut-off to 1.0 for two-member families and additionally included three-member gene families with K_S cut-off equal to 0.35. Ohnologs and SSD pairs in *S. cerevisiae* were distinguished by consulting Byrne and Wolfe's reconciled ohnolog list from recent comparative genomics studies (Byrne and Wolfe 2005). The initial dataset after this first set of filtering procedures comprised 47 ohnologs and 31 SSD pairs.

Determination of the Extent of Synteny Preservation with Outgroup Genomes

Synteny blocks (regions of conserved gene order) were retrieved on the YGOB database (http://wolfe.gen.tcd.ie/ygob/). For ohnologs, the single-copy ortholog within the reconstructed ancestor chromosome that is hypothesized to exist immediately before the occurrence of the WGD event 100–200 mya (Gordon et al. 2009) was used as a reference outgroup. For SSD-originating paralogs, the sequence of the most recent ancestor of the paralogs was inferred based on related genes in seven post-WGD yeast species (*Saccharomyces paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, *Candida glabrata*, and *Kluyveromyces polyspora*) using the codeml program of PAML by the setting the RateAncestor = 1 (Koshi and Goldstein 1996; Yang 2006; Yang et al. 1995). Tajima's Relative Rate test was then performed using DNA and protein sequences in triplets containing the two focal *S. cerevisiae* paralogs and their inferred ancestral sequence. In addition, duplications involving more than one gene locus, also referred to as 'linked sets' (Katju et al. 2009) were treated as a single duplication.

We used two measures to quantify the extent of gene-neighborhood conservation of each *S. cerevisiae* paralog in its upstream and downstream flanking regions. The first measure tallied the number of continuously shared genes with the outgroup genome in both the upstream and downstream directions. The second measure tallied the total number of genes shared with the outgroup genome within a block comprising 20 loci in both the upstream and downstream flanking regions. After excluding duplicate pairs with neither synteny nor outgroup information, the sample size of our study comprised 43 and 15 pairs of ohnologs and SSD-originated duplicates, respectively (Supplemental Tables S1 and S2).

Determining the Degree of Asymmetry among Paralogs

Tajima's Relative Rate test (Tajima 1993), as implemented in MEGA version 4.0 (Tamura et al. 2007) was used to determine if one of the paralogs was evolving faster. For SSD pairs, the designated outgroup sequence was a single-copy ortholog in an outgroup genome closely-related to *S. cerevisiae*. In the event that multiple outgroup species possessed a single-copy ortholog corresponding to *S. cerevisae*'s paralogs, we selected as outgroup the ortholog in the most closely-related outgroup genome. With respect to three-member gene families, the Tajima's test was only performed for the two most closely-related gene copies. For ohnologs, the outgroup was the phylogenetically closest species that contained a single-copy ortholog to the *S. cerevisiae* duplicate pair and diverged from the *Saccharomyces sensu stricto* group prior to the WGD event.

Genome and protein sequences of 11 fully sequenced yeast species were downloaded from the YGOB (http://wolfe.gen.tcd.ie/ygob/) and KEGG (http://www.genome.jp/kegg/catalog/org_list.html) databases. Outgroup identification was performed using DNA and protein sequences of the paralogs as queries in BLASTN and BLASTP searches against the genomic and protein sequences of the 11 yeast species. The BLAST outputs were filtered and organized using a Perl script. Gene duplicate pairs and their associated outgroup sequences were first aligned with ClustalW 2.0 and then manually checked and improved, when necessary, before the analysis.

The Wilcoxon signed-rank test was used to test if, collectively speaking, the ancestral and derived copies of a gene duplicate pair are evolving at the same rate. Since the ohnolog copies could not be classified as ancestral or derived, this tests if the rate of evolution is associated with the conservation of flanking synteny. Five pairs of ohnologs with equal number of unique sites were excluded from the Wilcoxon signed-rank test to yield a final sample of 38 ohnolog pairs. For SSD pairs, the paralog with the greater upstream synteny compared to the outgroup is taken to be the ancestral copy. In the event that both paralogs have equal continuous synteny, the total synteny gene number within 20 gene loci was further included as a measure of synteny conservation. If the information above was insufficient for distinguishing the ancestral and the derived copies, the total synteny within 20 upstream and downstream gene loci was utilized.

Relationship between Codon Usage, mRNA Abundance and Rate Asymmetry

The Codon Adaptation Index (CAI) was calculated using the JCat tool (www.jcat.de) (Grote et al. 2005; Sharp and Li 1987). The JCat tool uses the method of Carbone and colleagues (Carbone et al. 2003) to select a set of reference genes with optimal codon usage. In order to determine if differences in the rates of evolution are related to changes in optimal codon usage, we tested for correlation between the difference in number of unique sites (number of unique sites at the ancestral locus – number of unique sites at the derived locus) and the difference in CAI between paralogs (CAI of ancestral locus – CAI of derived locus).

An association between CAI and rate asymmetry between paralogs would suggest that gene expression is imposing differential constraints on the paralogs. As a proxy for gene expression, we obtained mRNA abundance data for all the paralogs in this study from a dataset consisting of transcript counts using single-molecule sequencing (Lipson et al. 2009). This data was used to test for an association between mRNA abundance and nucleotide rate asymmetry for both SSD pairs (FIGURE 3) and ohnologs.

Results

Greater Conservation of Synteny in Ohnologs

We initially commenced the analysis with 43 pairs of ohnologs and 15 SSD-derived gene duplicate pairs. These only included gene pairs that could be unambiguously assigned a single ortholog in an outgroup genome and the identification of local synteny conservation. Despite massive gene loss and genomic rearrangements in the evolutionary period subsequent to the WGD event, ohnologs have more extensive tracts of synteny relative to SSD-originated gene duplicates (Table 1). For instance, the average total upstream and downstream number of syntenic genes in the flanking regions for ohnologs versus SSD pairs is 19.87 and 4.67, respectively. Additionally, Wilcoxon signed-ranks tests revealed no significant difference in the extent of syntenic tracts in the upstream and downstream flanking regions within each population of yeast paralogs (ohnologs and SSD pairs).

Rate of Molecular Evolution of Ohnologs is decoupled from Synteny Conservation

Nine and zero of 43 ohnolog pairs displayed significant asymmetry based on Tajima's Relative Rate test (uncorrected for multiple comparisons) using DNA (Additional File 1, Table S1) and amino acid sequences (Additional File 2, Table S2), respectively. Of these nine pairs of ohnologs, the faster evolving copy was associated with less synteny conservation in seven instances. This would indicate that the rate of evolution for paralogs formed via polyploidization might be influenced by the degree of

preserved synteny. However, a nonparametric rank correlation test testing for association between synteny (sum of upstream and downstream continuous synteny) and the number of unique nucleotide sites was nonsignificant (*Kendall's tau* = 0.0132; $p = 0.91$). Likewise, we found no significant association between synteny preservation and the number of unique sites at the amino acid level (*Kendall's tau* = 0.0086; $p = 0.94$).

Derived Gene Copies Originating from SSD Events Exhibit Accelerated Rates of Molecular Evolution

Seven of 15 SSD pairs showed significant asymmetry using a Tajima's Relative Rate test at the nucleotide and amino acid level, respectively (Additional File 3, Table S3 and Additional File 4, Table S4). Six of these seven SSD pairs exhibited rate asymmetry both at the nucleotide and amino acid level. In all seven instances of significant rate asymmetry between paralogs at the nucleotide level, the derived copy exhibited accelerated rates of molecular evolution. In six of the seven instances of significant rate asymmetry at the amino acid level, the derived copy was the faster-evolving paralog. A Wilcoxon signed-ranks test of all 15 SSD pairs showed that collectively, the derived copies tend to possess a greater number of unique sites, suggesting accelerated molecular evolution at the nucleotide level ($T = -25.0$; $p = 0.024$) as well as the amino acid level ($T = -21.0$; $p = 0.029$).

CAI Results

Codon adaptation index (CAI) is a measure of optimal codon usage and it is positively correlated with levels of gene expression (Sharp and Li 1987). Following gene or genome duplication, there may be a period of relaxed selection resulting in lower CAI. If relaxation of selection does not apply equally to both paralogs, we may observe greater reduction in the use of optimal codons and CAI in one of the paralogs. We tested for the degree of association between the difference in CAI values between the two paralogs and the degree of rate asymmetry at the nucleotide level (difference in unique sites between the two paralogs generated from the Tajima's Relative Rate test) for both pools of gene duplicates in the *S. cerevisiae* genome. For SSD pairs, the derived paralogs have a significantly lower CAI than the ancestral paralogs (*Wilcoxon signed-ranks test*: $T = 39.5$; $p = 0.011$). However, we did not find a significant association between nucleotide rate asymmetry and change in CAI (*Kendall's tau* = 0.226; $p = 0.25$) (FIGURE 1). That is, faster-evolving paralogs did not have lower CAI values than slowly-evolving paralogs for SSD pairs. In contrast, we find a strong negative correlation between rate asymmetry and a difference in CAI values among ohnologs (*Kendall's tau* = -0.453; $p < 0.0001$) (FIGURE 2). Here, the faster-evolving paralogs resulting from the whole genome duplication event also have lower optimal codon preference.

Ohnologs and SSD duplicate pairs also differ with respect to their CAI values. The median CAI value for ohnologs and SSD pairs are 0.70 and 0.11, respectively. Indeed, CAI values averaged across both paralogs were determined to be significantly greater for ohnologs relative to SSD pairs (*Wilcoxon two-sample test*: Z = -4.723; $p < 0.0001$).

Faster-Evolving Paralogs Have Lower mRNA Abundance

The preceding CAI results suggest that relaxed selective constraints due to reduced expression of the derived paralog may contribute significantly to rate asymmetry between ancestral and derived paralogs. We find that ancestral paralogs are expressed at significantly higher levels (greater mRNA abundance) than derived paralogs for SSD pairs (*Wilcoxon signed-ranks test*: $T = 37.5$; $p < 0.017$). In contrast, ancestral-like ohnologs with greater syntenic preservation do not differ significantly in their expression levels compared to derived-like ohnologs with lower syntenic preservation (*Wilcoxon signed-ranks test*: $T = 52$; $p = 0.54$).

We additionally tested if there is a relationship between transcription levels of paralogs and their degree of rate asymmetry at the nucleotide level. FIGURE 3 shows a significant correlation between the ratio of paralog-specific RNA and the ratio of unique sites in derived and ancestral copies of SSD pairs ($r = 0.87$, *Kendall's tau* = 0.74, $p < 0.0002$). Likewise, we find a significant association between the ratio of paralog-specific RNA and the ratio of unique sites in derived and ancestral copies for ohnologs ($r = 0.38$, *Kendall's tau* = 0.225, $p = 0.0343$).

Discussion

Duplicated genes frequently experience an initial increase in their rate of evolution and nonsynonymous substitutions relative to synonymous substitutions. Moreover, recent analyses of young gene duplicates in several eukaryotic genomes indicate that paralogs exhibit asymmetric rates of sequence divergence in the evolutionary period soon after duplication (Conant and Wagner 2003; Cusack and Wolfe 2007; Kondrashov et al. 2002; Panchin et al. 2010; Scannell and Wolfe 2008; Wagner 2002; Zhang et al. 2003). Together, these observations indicate that initial relaxation of selection, or adaptive evolution, after duplication is limited to one of the paralogs, and that the slower-evolving paralog is more constrained by its ancestral function (Conant and Wagner 2003; Zhang et al. 2003). The majority of past studies did not distinguish between the ancestral and derived copies within a gene-duplicate pair, which in turn has precluded an unambiguous assessment of which copy is under stringent versus relaxed selective constraints.

There is some evidence that derived paralogs evolve faster than their counterparts residing at ancestral locations. In their study of evolutionarily young rodent gene duplicates, Cusack and Wolfe (Cusack and Wolfe 2007) assigned ancestral versus derived states to paralogs and demonstrated that genomic relocation of one paralog by retrotransposition engenders rate asymmetry in the sequence evolution of paralogs, commonly manifested as an accelerated rate of sequence evolution in the relocated paralog. Likewise, in bacterial genomes, the majority of paralogs that appear to have

moved away from their ancestral gene neighborhood evolved faster than static paralogs (Notebaart et al. 2005). Furthermore, a study of gene duplicates in four mammalian genomes determined that signatures of positive selection were more frequent in the derived copies than genes at their ancestral locations (Han et al. 2009).

In this study, we analysed the rate of evolution in yeast paralogs for which an ancestral versus derived status could be assigned by analyzing synteny as manifested in gene-neighborhood conservation. There was significantly greater gene-neighborhood conservation in ohnologs relative to SSD pairs. Although ohnologs originated from an ancient polyploidization event and rampant genome-wide deletions have since restored functional normal ploidy in these *Saccharomyces* species (Cliften et al. 2006; Scannell et al. 2006), it is noteworthy that this extensive gene-neighborhood conservation has persisted. There is no difference in the extent of gene-neighborhood conservation in the upstream and downstream regions of the paralogs for both populations of duplicates (ohnologs and SSD), suggesting, on average, equal rates of preservation/loss of upstream and downstream neighboring genes.

The majority of gene duplicates with low sequence divergence in *S. cerevisiae* stem from an ancient WGD event rather than segmental duplications. Subsequent to the WGD event, there has been extensive loss of genetic material with an estimated 10% of the original ohnologs remaining (Kellis et al. 2004). Deletions of genetic material within a WGD-derived homology block have the potential to remove or rearrange regulatory sequences for the remaining genes in the block. Therefore, the DNA sequence of a

paralog associated with more extensive gene-neighborhood conservation (i.e. local synteny) might be under stronger purifying selection than a paralog residing in regions that have endured more gene loss and rearrangements. While it is problematic to assign ancestral versus derived states to gene duplicates originating from WGD events, we reasoned that a paralog within an ohnolog pair could be characterized as being ancestral-like or derived-like based on the extent of gene-neighborhood conservation it shared with a single-copy ortholog in an outgroup genome. We then sought to test the hypothesis that ancestral-like gene-copies within ohnolog pairs are more likely to maintain ancestral gene function and therefore exhibit lower rates of sequence evolution. In contrast, gene-copies displaying a reduction in the extent of local synteny relative to the ortholog may be predisposed to accelerated rates of sequence evolution and the resultant fates of neofunctionalization or nonfunctionalization. However, we find no evidence of an association between rate asymmetry in ohnologs and local gene-neighborhood conservation. In other words, for ohnologs, a decline in local gene-neighborhood conservation (derived-like) does not engender accelerated rates of sequence evolution either at the nucleotide or amino acid level. This is in contrast to a study of vertebrate genomes that found a significant correlation between synteny preservation and sequence conservation (Abi-Rached et al. 2002). We speculate that the greater number of regulatory sites in vertebrate genomes might engender greater sensitivity to syntenic changes relative to yeast. However, ohnologs in yeast do exhibit a strong significant relationship between rate asymmetry and CAI such that the faster-evolving paralogs have lower CAI. The rate asymmetry in ohnologs also seems to be to some degree caused by relaxation of selection for codon usage in one copy.

Among the SSD pairs in our sample, it is the derived copy that evolves faster on average, both at the nucleotide and the amino acid level. This lends credence to Ohno's original hypothesis that duplication enables redundancy, enabling one copy to explore new evolutionary space by accumulating mutations (Ohno 1970). It is likely that segmental duplications frequently do not capture the full repertoire of regulatory sequences (Lynch and Katju 2004) associated with the ancestral genes and/or result in the insertion of the derived copy into a region of the genome with different chromatin structure and potentially under the influence of different regulatory elements. Under these conditions, mutations that interfere with the ancestral gene's original function would still be selected against, whereas the derived copy could be under relaxed or positive selection. For SSD pairs, the rate asymmetry at the nucleotide level is likely due to a regime of relaxed selective constraints as there is a significant reduction in the CAI of the derived paralogs within SSD pairs. The CAI compares the codon usage of a gene to codon usage in highly expressed genes; hence, the reduction in the CAI values of derived paralogs suggests that selection for optimal codon usage has been relaxed in the derived copy. Puzzlingly, we failed to detect any correlation between nucleotide sequence asymmetry of SSD paralogs and changes in their CAI values. This may stem from limited power given the small sample size of available SSD duplicates in the yeast genome.

If the rate asymmetry in paralogs is largely a consequence of relaxation of selection in the derived paralog, it should also be manifested as different levels of

expression among the two copies. Previous work has shown that the evolutionary rate in yeast is strongly influenced by gene expression (Drummond et al. 2005, 2006). In both the yeast ohnologs and SSD pairs studied here, mRNA abundance is correlated with the rate of evolution. Moreover, within SSD pairs, it is the derived paralogs that have lowered mRNA abundance relative to the ancestral loci. Both the CAI and mRNA abundance suggest that selective constraints on gene expression is a significant driver of evolutionary rate asymmetry in paralogs.

Conclusions

Following gene duplication, there is a general increase in the rate of evolution, and this increase is frequently asymmetric in that one paralog evolves at an accelerated pace. Asymmetry in the rate of molecular evolution after duplication has been variously associated with the evolution of novel functions, change in the number of interactions, and relaxation of selection. Here we address the related question if certain factors predispose one paralog to evolve faster. For instance, segmental duplications may translocate the derived copy to a different regulatory environment where it may evolve under different or reduced constraints (Lynch and Katju 2004). Despite a limited sample of gene-duplicate pairs originating from recent small-scale duplications in *S. cerevisiae*, we find that the derived copy tends to evolve faster and is under reduced selection for codon usage. Accelerated rates in ohnologs are also associated with reduced selection for codon usage. Moreover, the rate of evolution is negatively correlated with mRNA abundance for ohnologs as well as SSD pairs. This adds to the evidence from mammals

(Han et al. 2009) that genes are not born equal and that the duplication process predisposes the derived copy to an evolutionary trajectory of initially reduced selective constraints and one that is perhaps more conducive to the evolution of new functions.

Tables

Table 1: Averaged measures of synteny preservation for 43 pairs of ohnologs *versus* 15 SSD pairs in the *S. cerevisiae* genome.

For all measures of synteny (upstream continuous, downstream continuous, upstream total, and downstream total), the extent of synteny preservation is significantly greater in ohnologs relative to SSD pairs based on Wilcoxon tests.

Synteny Measure	*Ohnologs*	*SSD pairs*	*p-value*
Upstream continuous	1.41	0.47	0.0002
Downstream continuous	1.50	0.20	<0.0001
Upstream continuous + Downstream continuous	2.91	0.67	
Upstream total	10.08	3.00	<0.0001
Downstream total	9.79	1.67	<0.0001
Upstream total + Downstream total	19.87	4.67	

Table S1. Tajima's Relative Rate Test for Ohnolog DNA sequences.

	Ancestral Paralog (A)	*Derived Paralog (B)*	*Outgroup (C)*	χ^2	*p-value*	*Unique Sites*		
						A	**B**	**C**
1	YBL027W	YBR084C-A	kla:KLLA0E12463g	0.03	0.85746	15	16	54
2	YBL072C	YER102W	kla:KLLA0E20559g	3.6	0.05778	2	8	72
3	YBR031W	YDR012W	kla:KLLA0B07139g	0.11	0.73888	5	4	155
4	YBR048W	YDR025W	kla:KLLA0A10483g	5.76	0.01638	5	16	41
5	YDL131W	YDL182W	kla:KLLA0F05489g	5.59	0.0181	20	38	176
6	YDL191W	YDL136W	kla:KLLA0F05247g	0.33	0.5637	1	2	39
7	YDR342C	YHR092C	kla:KLLA0D13310g	1	0.31731	84	90	210
8	YDR447C	YML024W	kla:KLLA0B01474g	0.22	0.63735	8	10	41
9	YEL034W	YJR047C	kla:KLLA0E22286g	0.02	0.8759	21	20	37
10	YER074W	YIL069C	kla:KLLA0C07755g	0.33	0.5637	5	7	33
11	YFR031C-A	YIL018W	kla:KLLA0D16027g	1.81	0.17793	10	17	55
12	YGL031C	YGR148C	kla:KLLA0E10857g	0.5	0.4795	14	18	44
13	YGR034W	YLR344W	kla:KLLA0B05742g	13.5	0.00024	3	21	34
14	YGR118W	YPR132W	kla:KLLA0B11231g	5.4	0.02014	12	3	28
15	YGR138C	YPR156C	kla:KLLA0E03729g	0.4	0.52454	64	57	444
16	YGR192C	YJR009C	ago:AGOS_AER031C	0.93	0.33592	11	16	169
17	YHL033C	YLL045C	kla:KLLA0E00506g	0.5	0.4795	18	14	88
18	YHR066W	YDR312W	kla:KLLA0C14586g	0.64	0.42503	35	42	349
19	YHR141C	YNL162W	kla:KLLA0D07832g	0	1	3	3	26
20	YHR203C	YJR145C	kla:KLLA0B03652g	0.07	0.79625	7	8	79
21	YKL006W	YHL001W	kla:KLLA0B13409g	0.29	0.59298	8	6	53
22	YKR059W	YJL138C	kla:KLLA0A05731g	0.2	0.65472	2	3	188
23	YLR333C	YGR027C	kla:KLLA0B06193g	2.13	0.1444	15	8	37
24	YML026C	YDR450W	kla:KLLA0B01562g	0.2	0.65472	11	9	25
25	YML063W	YLR441C	kla:KLLA0B05060g	6.12	0.01338	20	39	59
26	YML073C	YLR448W	kla:KLLA0B04686g	0.38	0.53709	19	23	67
27	YMR121C	YLR029C	kla:KLLA0F17633g	18.69	0.00002	33	6	35
28	YMR142C	YDL082W	kla:KLLA0E22099g	16.03	0.00006	7	32	50
29	YMR143W	YDL083C	kla:KLLA0E22077g	0.14	0.70546	15	13	27
30	YMR186W	YPL240C	kla:KLLA0D12958g	0.03	0.86853	74	72	255
31	YMR230W	YOR293W	kla:KLLA0B08173g	2.58	0.10829	13	6	44
32	YNL209W	YDL229W	kla:KLLA0D19041g	0.1	0.75762	22	20	189
33	YOL120C	YNL301C	kla:KLLA0A07227g	3.85	0.04986	8	18	54
34	YOL121C	YNL302C	kla:KLLA0A07194g	0	1	10	10	39
35	YOR133W	YDR385W	kla:KLLA0E02926g	3	0.08326	0	3	227
36	YOR182C	YLR287C-A	kla:KLLA0C04809g	4.5	0.03389	1	7	21
37	YOR312C	YMR242C	kla:KLLA0F08657g	0.62	0.43277	15	11	47
38	YPL079W	YBR191W	kla:KLLA0E23727g	0.14	0.70546	15	13	32
39	YPL090C	YBR181C	kla:KLLA0E24090g	0	1	3	3	73
40	YPL198W	YGL076C	kla:KLLA0D03410g	3.2	0.07364	14	6	100
41	YPL220W	YGL135W	kla:KLLA0B02002g	0	1	0	0	69
42	YPR080W	YBR118W	kla:KLLA0B08998g	0	1	1	1	83
43	YPR102C	YGR085C	kla:KLLA0F08261g	1.67	0.19671	5	10	48

Table S2. Tajima's Relative Rate Test for Ohnolog amino acid sequences.

	Ancestral Paralog (A)	*Derived Paralog (B)*	*Outgroup (C)*	χ^2	*p-value*	*Unique Sites*		
						A	**B**	**C**
1	YBL027W	YBR084C-A	kla:KLLA0E12463g	0	1	0	0	24
2	YBL072C	YER102W	kla:KLLA0E20559g	0	1	0	0	24
3	YBR031W	YDR012W	kla:KLLA0B07139g	0	1	0	0	50
4	YBR048W	YDR025W	kla:KLLA0A10483g	0	1	0	0	15
5	YDL131W	YDL182W	kla:KLLA0F05489g	2.13	0.1444	8	15	13
6	YDL191W	YDL136W	kla:KLLA0F05247g	0	1	0	0	13
7	YDR342C	YHR092C	kla:KLLA0D13310g	1.98	0.1599	16	25	108
8	YDR447C	YML024W	kla:KLLA0B01474g	1	0.3173	0	1	15
9	YEL034W	YJR047C	kla:KLLA0E22286g	0.08	0.7815	7	6	11
10	YER074W	YIL069C	kla:KLLA0C07755g	0	1	0	0	11
11	YFR031C-A	YIL018W	kla:KLLA0D16027g	0	1	0	0	15
12	YGL031C	YGR148C	kla:KLLA0E10857g	0	1	2	2	19
13	YGR034W	YLR344W	kla:KLLA0B05742g	1	0.3173	0	1	10
14	YGR118W	YPR132W	kla:KLLA0B11231g	0	1	0	0	3
15	YGR138C	YPR156C	kla:KLLA0E03729g	0.18	0.6698	12	10	104
16	YGR192C	YJR009C	ago:AGOS_AER031C	1.6	0.2059	3	7	42
17	YHL033C	YLL045C	kla:KLLA0E00506g	1	0.3173	3	1	37
18	YHR066W	YDR312W	kla:KLLA0C14586g	0.82	0.3657	4	7	104
19	YHR141C	YNL162W	kla:KLLA0D07832g	0	1	0	0	10
20	YHR203C	YJR145C	kla:KLLA0B03652g	0	1	0	0	19
21	YKL006W	YHL001W	kla:KLLA0B13409g	1	0.3173	1	0	18
22	YKR059W	YJL138C	kla:KLLA0A05731g	0	1	0	0	61
23	YLR333C	YGR027C	kla:KLLA0B06193g	1	0.3173	1	0	14
24	YML026C	YDR450W	kla:KLLA0B01562g	0	1	0	0	10
25	YML063W	YLR441C	kla:KLLA0B05060g	1.29	0.2568	2	5	14
26	YML073C	YLR448W	kla:KLLA0B04686g	0.5	0.4795	5	3	27
27	YMR121C	YLR029C	kla:KLLA0F17633g	0	1	1	1	8
28	YMR142C	YDL082W	kla:KLLA0E22099g	0	1	0	0	23
29	YMR143W	YDL083C	kla:KLLA0E22077g	0	1	0	0	6
30	YMR186W	YPL240C	kla:KLLA0D12958g	2.57	0.1088	4	10	65
31	YMR230W	YOR293W	kla:KLLA0B08173g	2	0.1573	2	0	19
32	YNL209W	YDL229W	kla:KLLA0D19041g	0.33	0.5637	2	1	54
33	YOL120C	YNL301C	kla:KLLA0A07227g	0	1	0	0	18
34	YOL121C	YNL302C	kla:KLLA0A07194g	1	0.3173	0	1	17
35	YOR133W	YDR385W	kla:KLLA0E02926g	0	1	0	0	60
36	YOR182C	YLR287C-A	kla:KLLA0C04809g	0	1	0	0	7
37	YOR312C	YMR242C	kla:KLLA0F08657g	0	1	0	0	15
38	YPL079W	YBR191W	kla:KLLA0E23727g	2	0.1573	2	0	9
39	YPL090C	YBR181C	kla:KLLA0E24090g	0	1	0	0	29
40	YPL198W	YGL076C	kla:KLLA0D03410g	2	0.1573	2	0	27
41	YPL220W	YGL135W	kla:KLLA0B02002g	0.08	0.7815	0	0	19
42	YPR080W	YBR118W	kla:KLLA0B08998g	0	1	0	0	17
43	YPR102C	YGR085C	kla:KLLA0F08261g	1	0.3173	0	1	16

Table S3. Tajima's Relative Rate Test for DNA sequences of SSD pairs using a maximum-likelihood generated ancestral sequence as outgroup.

	Ancestral Paralog (A)	*Derived Paralog (B)*	χ^2	*p-value*	*Unique Sites*		
					A	**B**	**C (ancestral sequence)**
1	YDL075W	YLR406C	2.88	0.0896	5	12	2
2	YDR039C	YDR038C	0	1	0	0	15
3	YDR533C	YOR391C	6.74	0.0094	11	27	7
4	YFL009W	YER066W	49.5	0	11	77	12
5	YFL058W	YNL332W	0	1	2	2	10
6	YGL258W	YOR387C	5.76	0.0164	5	16	17
7	YHR055C	YHR053C	0	1	0	0	8
8	YHR056C	YHR054C	0	1	0	0	108
9	YLR044C	YLR134W	160.17	0	15	201	5
10	YNL067W	YGL147C	1	0.3173	21	15	8
11	YOL055C	YPL258C	0.97	0.3258	124	109	82
12	YOL086C	YMR303C	97.85	0	5	112	0
13	YOR388C	YPL276W_275W	19.59	0	2	25	56
14	YOR389W	YPL277C_278C	28.58	0	20	71	99
15	YPL279C	YOR390W	2	0.1573	6	2	0

Note: Cells containing two gene IDs comprise cases where the exon-intron structure of the original locus has been altered to comprise two genes.

Table S4. Tajima's Relative Rate Test for amino acid sequences of SSD pairs using a maximum-likelihood generated ancestral sequence as outgroup.

	Ancestral Paralog (A)	*Derived Paralog (B)*	χ^2	*p-value*	*Unique Sites*		
					A	**B**	**C (ancestral sequence)**
1	YDL075W	YLR406C	1	0.3173	0	1	0
2	YDR039C	YDR038C	0	1	0	0	0
3	YDR533C	YOR391C	13	0.0003	0	13	0
4	YFL009W	YER066W	37	0	0	37	0
5	YFL058W	YNL332W	0	1	0	0	0
6	YGL258W	YOR387C	3.57	0.0588	1	6	2
7	YHR055C	YHR053C	0	1	0	0	3
8	YHR056C	YHR054C	0	1	0	0	42
9	YLR044C	YLR134W	59.24	0	2	65	0
10	YNL067W	YGL147C	0	1	2	2	2
11	YOL055C	YPL258C	36.94	0	36	25	27
12	YOL086C	YMR303C	16.67	0	2	22	0
13	YOR388C	YPL276W_275W	11	0.0009	0	11	7
14	YOR389W	YPL277C_278C	10.31	0.0013	8	27	26
15	YPL279C	YOR390W	3	0.0833	3	0	0

Note: Cells containing two gene IDs comprise cases where the exon-intron structure of the original locus has been altered to comprise two genes.

Figures

FIGURE 1. Nucleotide sequence asymmetry and codon adaptation index (CAI) for 15 SSD pairs in the *S. cerevisiae* genome.

The sequence asymmetry measure on the *x* axis was calculated as the difference between unique nucleotide sites at the ancestral copy and the derived copy. The *y* axis represents the difference in CAI values between the ancestral copy and the derived copy for the same SSD pair. There was no significant association between differences in rate asymmetry and CAI values for SSD pairs (*Kendall's tau* = 0.226; $p = 0.25$).

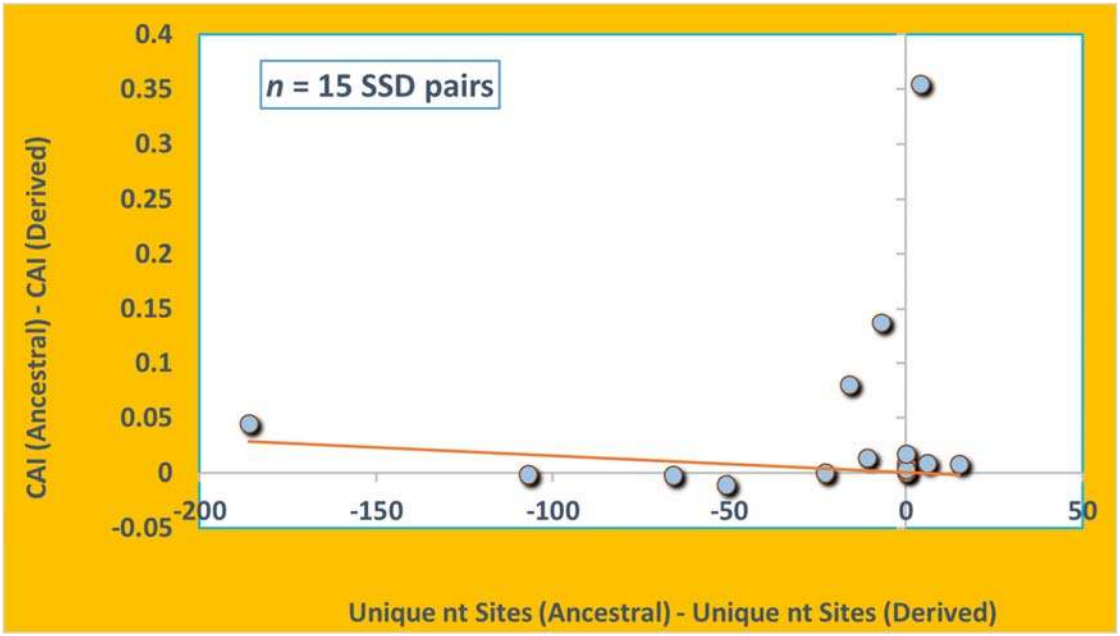


FIGURE 2. Negative relationship between nucleotide sequence asymmetry and codon adaptation index (CAI) for 43 pairs of ohnologs in the *S. cerevisiae* genome.

The sequence asymmetry measure on the *x* axis was calculated as the difference between unique nucleotide sites at the ancestral-like copy and the derived-like copy within an ohnolog pair. The *y* axis represents the difference in CAI values between the ancestral-like copy and the derived-like copy for the same ohnolog pair. There was a significant negative correlation between differences in rate asymmetry and CIA values for ohnologs (*Kendall's tau* = -0.453; $p < 0.0001$).

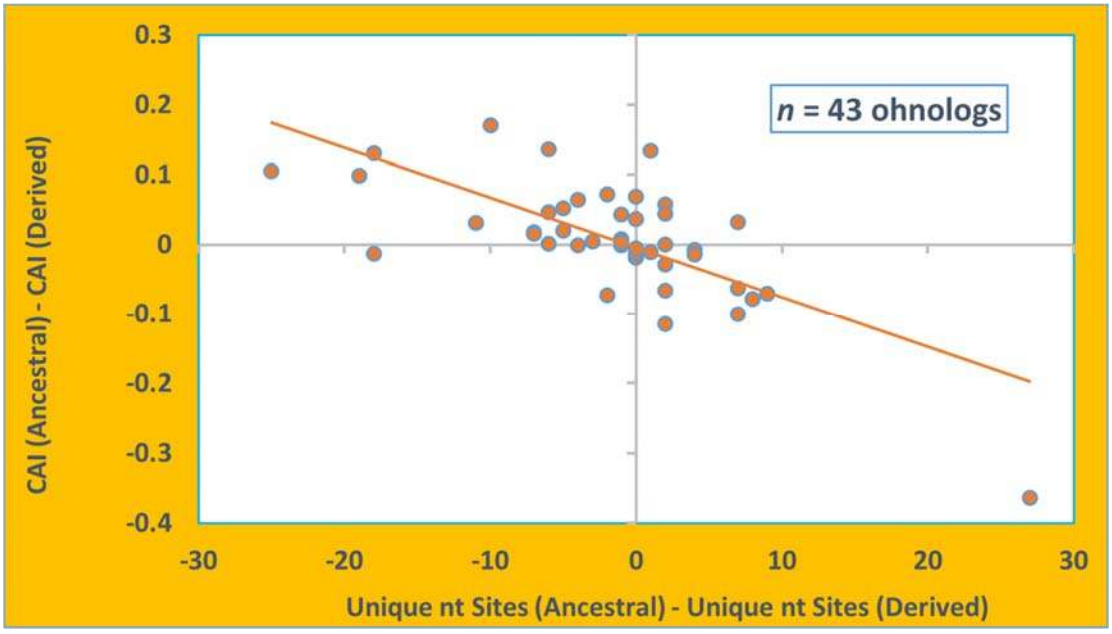
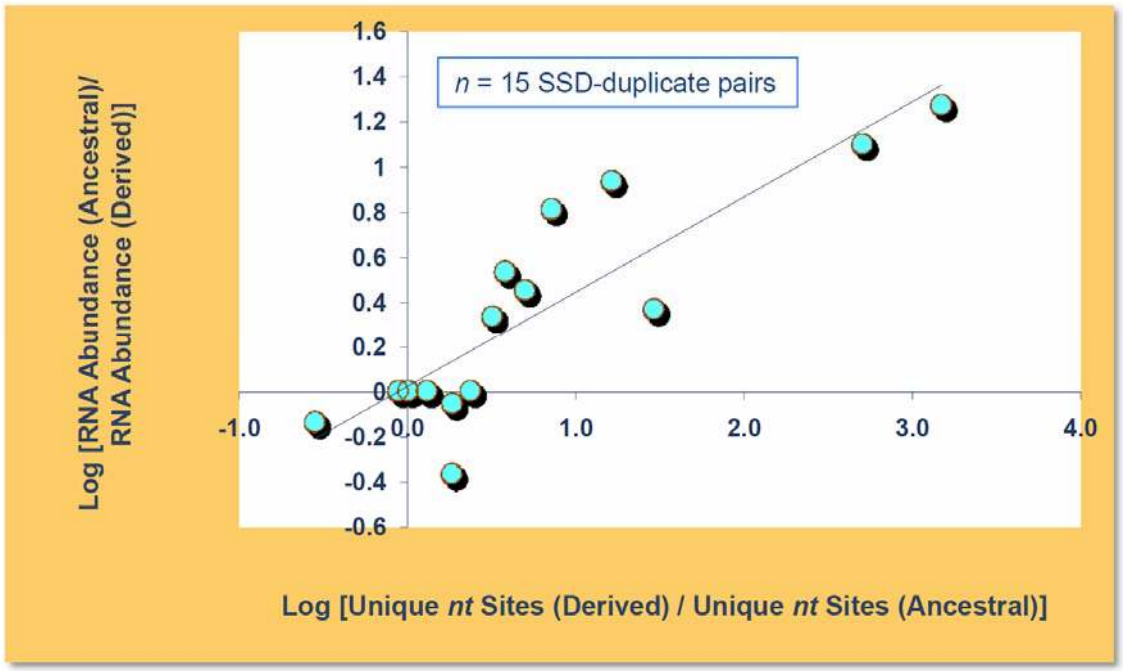


FIGURE 3. Nucleotide sequence asymmetry and mRNA abundance for 15 SSD pairs in the *S. cerevisiae* genome.

The sequence asymmetry at the nucleotide level is expressed as the $\log_{10}$ (unique sites in the derived paralog/unique sites in the ancestral paralog) and relative RNA abundance is expressed as the $\log_{10}$ (RNA count for ancestral paralog/RNA count for derived paralog). There is a significant correlation between divergence between paralogs at the sequence level and divergence in their expression profiles (as represented by mRNA abundance) (*Kendall's tau* = 0.74; $p < 0.0002$).



CHAPTER THREE

Early evolutionary history and genomic features of gene duplicates in the human genome

Submitted in 2015.

Abstract

Human gene duplicates have been the focus of intense research since the development of array-based and targeted next-generation sequencing approaches in the last decade. These studies have primarily concentrated on determining the extant copy-number variation from a population-genomic perspective but lack a robust evolutionary framework to elucidate the early structural and genomic characteristics of gene duplicates at emergence and their subsequent evolution with increasing age. We analyzed 184 gene duplicate pairs comprising small gene families in the draft human genome with ≤ 10% synonymous sequence divergence. Human gene duplicates primarily originate from DNA-mediated events, taking up genomic residence as *intrachromosomal* copies in direct or inverse orientation. The distribution of paralogs on autosomes follows random expectations in contrast to their significant enrichment on the sex chromosomes. Furthermore, human gene duplicates exhibit a skewed gradient of distribution along the chromosomal length with significant clustering in pericentromeric regions. Surprisingly, despite the large average length of human genes, the majority of extant duplicates (83%) are *complete* duplicates, wherein the entire ORF of the ancestral copy was duplicated. The preponderance of *complete* duplicates is in accord with an extremely large median duplication span of 36 kb for our data set, which enhances the probability of capturing ancestral ORFs in their entirety. With increasing evolutionary age, human paralogs exhibit (i) a decline in the frequency of *intrachromosomal* paralogs, and (ii) a decline in the proportion of *complete* duplicates. These changes may reflect lower survival rates of certain classes of duplicates and/or the role of purifying selection. Duplications arising

from RNA-mediated events comprise a small fraction (11.4%) of all human paralogs and are more numerous in older evolutionary cohorts of duplicates.

Introduction

The recent genomic era has established gene duplication as a dominant contributor to the origin of new genes and novel traits, which in turn fuels adaptation, niche diversification and increase in biocomplexity. Two characteristics of gene duplicates lend to their primacy in effecting evolutionary change, namely (i) their role in the creation of genetic redundancy or novel genes, and (ii) their high rate of spontaneous origin. The high supply rate of genetically and functionally redundant gene copies might be especially advantageous when the environment imposes immediate selection for increased gene dosage and gene expression (Bergthorsson et al. 2007). The promiscuity of the gene duplication process leading to the duplication of DNA segments across gene boundaries, often in conjunction with the inclusion of noncoding DNA sequence to yield a novel open reading frame, can additionally yield new genes with distinctly novel functions (Katju 2012; Katju and Lynch 2006). Notable examples of the fashioning of novel genes from the incomplete duplication of ancestral gene sequences account for the origin of antifreeze glycoproteins in Antarctic fish (Chen et al. 1997; Deng et al. 2010) and the evolution of hermaphroditism in *Caeonorhabditis elegans* from an obligately outcrossing ancestor (Katju et al. 2008). The second salient characteristic of gene duplicates is their astoundingly high rates of spontaneous origin. Empirical estimates of locus-specific or genome-wide spontaneous rates of gene duplication range from 10^{-3} to 10^{-7} per gene per generation (Katju and Bergthorsson 2013; Lipinski et al. 2011). These high rates of gene duplication directly contribute to the high frequency of copy-number variants (CNVs) being uncovered in population-genomic studies (Maydan et al. 2010; Nair et al. 2008; Redon et al. 2006).

Classical models of gene duplication make the key assumption that duplicated genes originate structurally and functionally redundant to the ancestral copy. An evolutionary trajectory leading to the origin of a hitherto novel function is thought to occur under a regime of relaxed selective constraints due to gradual accumulation of previously 'forbidden' deleterious mutations (Ohno 1970). However, unbiased studies of entire age-cohorts of evolutionarily young gene duplicates in a few species have demonstrated the existence of gene copies bearing structural heterogeneity (*partial* or *chimeric* gene duplicates) due to incomplete duplication across ORFs and/or recruitment of novel noncoding sequences (Katju and Lynch 2003; Katju et al. 2009; Meisel 2009; Zhou et al. 2008). With respect to small segmental duplication (SSD) events, the frequency of *complete* gene duplicates (entire duplication of an ancestral ORF) can be highly variable; 39% in *C. elegans* (Katju and Lynch 2003), 41-44% in *Drosophila* species (Zhou et al. 2008) and 89% in *Saccharomyces cerevisiae* (Zhou et al. 2008). Additionally, gene duplication via retrotransposition, which results in the insertion of the duplicate copy in a random location in the genome, likely engenders acquisition of novel regulatory elements and altered gene expression patterns. These *heterogeneous* gene duplicates (*partial*, *chimeric*, and *retrotransposed*) are more likely to be nonfunctionalized but also have the potential to gain immediate novel functions (Katju 2012). The diverse structural classes of gene duplicates, if identified in their early evolutionary existence, can provide insights into the mutational mechanisms underlying their origin as well as the sequence alterations that facilitate molecular innovations (Katju 2012). To date, we have a limited understanding of the population dynamics and

selective constraints influencing different structural classes of gene duplicates. A comparative study of gene duplicates with low synonymous divergence in the *C. elegans* and *S. cerevisiae* genomes implied that both species –specific differences in mutational input and strength of natural selection molded the distribution of gene duplicates in these two genomes (Katju et al. 2009).

Investigating the interplay between evolutionary forces and mutation in patterning the distribution of gene duplicates in the human genome might be of particular interest for several reasons. First, there has been a spate of population-genomic studies establishing widespread copy-number variation in humans and other hominoid and primate species (Bailey et al. 2003; Fortna et al. 2004; Gokcumen et al. 2013; Redon et al. 2006). Second, segmental gene duplications (one form of CNVs) have demonstrated a signature of expansion in early hominoid evolution (Samonte and Eichler 2002). Whereas a large fraction of the chromosomal rearrangements created by segmental duplications in humans are implicated in Mendelian and complex genetic disease (Botstein and Risch 2003; Emanuel and Shaikh 2001; Inoue and Lupski 2002; Sebat et al. 2007), they additionally serve as important substrates for the origin of evolutionary innovations. Although the most common fate of gene duplicates may be immediate pseudogenization upon arrival, the extraordinary high rates of spontaneous gene duplication likely have a substantial influence on the trajectory of evolution by enabling the origin of discernible numbers of gene substrates for neofunctionalization (Katju and Bergthorsson 2013). In the context of human evolution, there is substantial interest in delineating the genetic changes that account for the emergence of human-specific morphological and behavioural changes since their divergence from other primates.

Given the role of gene duplication in the emergence of evolutionary novelties and their high spontaneous rates of origin, human-specific gene duplicates would appear to be a promising avenue for investigation. Two notable examples of adaptive copy-number changes in humans involve the *AMY1* (Perry et al. 2007) and *SRGAP2C* (Charrier et al. 2012; Dennis et al. 2012) genes.

To date, there has been no systematic study in a strict evolutionary context that comprehensively characterizes the structural and genomic features of a large, unbiased population of evolutionarily young gene duplicates in the human genome. Such a study would provide a rich natural history perspective on the mutational origins of human gene duplicates, the degree of structural resemblance between paralogs, and the patterns of genomic traffic in the early stages of their evolution. In addition, it would enable future comparative genomic research investigating differences in the genomic architecture of human- and chimpanzee-specific gene duplicates. Structural and genomic features of novel paralogs at inception can greatly influence their evolution and ultimate fate. In order to test the importance of structural features on the evolution of young gene duplicates, we performed a genome-wide survey of the entire population of evolutionarily young paralogs belonging to small gene-families in the human genome. Because subsequent mutational events in the evolutionary life of gene duplicates can rapidly erode their key characteristics at inception, we limited our analyses to putative evolutionarily young gene duplicates (synonymous divergence per synonymous site $K_S \leq 0.1$) in the current human genome assembly with the similarity search cutoff capable of capturing paralogs with differing levels of structural resemblance. To our knowledge, this study is

the first to delineate the relative fractions of *complete*, *partial*, and *chimeric* paralogs within an unbiased population of gene duplicates in the human genome.

Methods

Similarity Based Grouping and Estimation of Evolutionary Divergence

Genome sequences and annotated genome features for the human genome assembly GRCh37 were downloaded from Ensembl release version 72 (Flicek et al. 2013). To minimize the inclusion of splice variants during the similarity search, we selected the longest transcript for each coding gene as the canonical transcript using in-house Perl scripts. Protein sequences and coding sequences of 20,214 canonical transcripts were downloaded from the BioMart interface of the Ensembl site. Similarity search was performed using an all-against-all BLASTP with a cutoff E-value of $\leq 10^{-10}$ and an amino acid identity $\geq 40\%$. To ensure that evolutionarily young but structurally heterogeneous gene duplicates (e.g. *partial* or *chimeric* duplicates) were not excluded from the initial sequence filtration steps, we did not use the high identity cutoff of 90%, which is widely used in other studies of this nature. Genes with higher levels of similarity than the cutoff value were clustered into one family. Multiple genes were pooled into one gene family based on the single-link principle. For example, if protein A hits proteins B and C with BLASTP E-values $\leq 10^{-10}$ and identity $\geq 40\%$, then A, B, and C were included in the same family, regardless of the similarity for the comparison of B and C. Linked duplicate sets, which comprised the duplication of multiple open reading frames via a single duplication event, were treated as a single gene duplicate. The K_S values of all members within a linked set were averaged to yield a single K_S value.

For each gene duplicate pair, a protein sequence alignment was generated by the CLUSTALW2 program (Larkin et al. 2007). Thereafter, the nucleotide sequences were aligned based on the protein sequence alignment profile using PAL2NAL (Suyama et al. 2006). The measure of synonymous sequence divergence in coding regions (K_S) for gene paralogs was recalculated using the pairwise model (runmode = -2) of the *codeml* program in the PAML package (Yang 2007). Putative evolutionarily young gene duplicate pairs ($K_S \leq 0.1$) were retained for further analysis.

Investigating the Frequency of Ectopic Gene Conversion between Paralogous Sequences

For each of the 184 duplicate pairs within our dataset, protein sequences of both human paralogs were used as queries in the BLASTP program to search and identify, where possible, the best hit in the chimpanzee protein database. The coding sequences of the human paralogs and their best-hit chimpanzee ortholog(s) were input and aligned in a single sequence file using the CLUSTALW2 program (Larkin et al. 2007). A statistical test for gene conversion was implemented in the GENECONV program, version 1.81a (Sawyer 1989) with default settings and additional option (/lp) to detect both global and pairwise inner fragments supporting gene conversion. Significance of gene conversion was determined by a permutation test correcting for multiple comparisons.

Visualization of Duplication Breakpoints and Determination of the Degree of Structural Resemblance between Paralogs

To locate the duplication breakpoints for large human gene pairs, sequences within 200 kb flanking region (800 kb for few pairs) of each gene were aligned using the pairwise alignment tool LASTZ (Harris 2007). The LASTZ program uses a seeded pattern-matching method to find out local similarities for large genomic DNA sequences. To obtain a graphic view for all identified young gene duplicates, the LASTZ alignment results in conjunction with the genome features were imported into the Generic Synteny Browser, GBrowse_syn (McKay et al. 2010). With the aid of an interactive alignment of the two focal paralogous sequences, we further identified the duplication break points, duplication span, and the degree of structural resemblance between paralogs (Katju and Lynch 2003).

We further filtered out *same-location* pairs and *shadow/redundancy* pairs for gene families comprising three to five members. The *same-location* pairs shared the same chromosomal coordinates while being assigned different gene names. This was taken to reflect annotation errors rather than true gene duplication events. We also removed *shadow* pairs within multiple-member gene families, which were representative of sequence similarity rather than true duplication events. For example, a five-member gene family could have been generated through four gene duplication events, although BLASTP would yield ten gene duplicates pairs based on pairwise comparisons of sequence similarity. In this hypothetical example, only four gene duplicate pairs representing the true duplication events were retained, while removing the six additional duplicate pairs displaying sequence similarity. The representative four gene duplicate

pairs were selected for inclusion based on a UPGMA tree generated from their pairwise K_S values.

The initial genome-wide search identified 286 gene duplicates pairs with low synonymous divergence in the human genome based on DNA (or protein) sequence similarity. The putative gene duplicates were subsequently filtered with respect to evolutionary age ($K_S \leq 0.1$) and family size (≤ 5 members). During the visualization check, 24 *same-location* pairs and 57 *shadow* pairs were removed, and 64 gene pairs were merged into 42 linked sets. Finally, we identified 184 duplication events, comprising 142 non-linked duplications and 42 linked sets.

Statistical Tests

Statistical tests were performed using the R program package version 3.01(R Core Team 2014). All duplicate pairs were initially classified into three age-cohorts ($K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$). If the latter two of the three cohorts showed no significant statistical difference with respect to the focal characteristic, comparisons were then performed between two cohorts ($K_S = 0$, and $0 < K_S \leq 0.1$).

Chromosomal Location

The frequency distribution of duplications between and within chromosomes was analyzed with a goodness-of-fit *G*-test. The number of gene duplicates per chromosome was compared to the number of protein-coding genes per chromosome. Each gene

duplicate pair with both paralogs residing on the same chromosome was counted as a single duplication event. In instances where the two paralogs were located on different chromosomes, each paralog was counted as a half event. This was done because both paralogs resulted from a single duplication event and the identity and location of the ancestral paralog could not be determined. A goodness-of-fit test was also performed on the distance of intrachromosomal paralogs from the centromeres. The chromosomes were divided into 10 Mb bins and the number of duplicates compared to the number of genes per bin. In the events that the two paralogs comprising a duplicate pair were located in different bins, each paralog was counted as half.

Results

We identified 184 human gene duplicate pairs belonging to small gene families (≤ 5 members) with low synonymous sequence divergence of 10% or less ($K_S \leq 0.1$) (Supplemental Table S1). Because the evolutionary dynamics of paralogs in large multigene families may differ markedly from those of paralogs comprising small gene families, we restricted our analyses to human paralogs belonging to families comprising five or less paralogs. The chromosomal location was confirmed for both paralogs belonging to 172 pairs. The remaining 12 pairs comprised at least one paralog located on a supercontig with an unassigned chromosomal location. Supplemental Table 1 lists the identification numbers of all paralogs comprising the 184 human gene duplicate pairs in conjunction with other relevant information such as synonymous divergence between paralogs, chromosomal location of the two paralogs, the assigned category of structural resemblance, transcriptional orientation of paralogs, duplication span (bp) and physical distance between paralogs located on the same chromosome.

Assessment and Controlling for the Role of Ectopic Gene Conversion in Confounding Evolutionary Age Estimates of Paralogous Sequences

We tested all 184 duplication events in our study for signatures of gene conversion using a chimpanzee ortholog as an outgroup sequence. We found evidence for gene conversion in the coding sequences of 26 of the 184 duplicate pairs tested, comprising 18 single-locus duplications and eight linked sets representing the duplication of more than one protein-coding ORF during a single duplication event. We conducted

all subsequent statistical analyses of the genomic and structural features of human paralogs on two separate data sets: (i) all 184 duplicate pairs including the 26 sets that exhibited a positive signature of gene conversion, and (ii) 158 duplicate pairs by excluding 26 sets showing evidence of gene conversion. The exclusion of the 26 duplicate sets showing evidence of gene conversion did not qualitatively alter our results. For each subsequent analyses that involves K_S as a parameter, we report the significance values of statistical tests with and without inclusion of the 26 duplicate sets exhibiting evidence of gene conversion.

L-shaped Frequency Distribution of Human Gene Duplicates

Assuming that the synonymous sequence divergence between paralogs is an adequate proxy for evolutionary time, the K_S values between paralogs were used to generate a relative age-distribution of the focal 184 human gene duplicate pairs (Fig.1). The distribution of putative evolutionarily young human gene duplicates is strongly L-shaped with the highest density of gene duplicates occurring in the youngest age cohorts and a strong decline in gene duplicate frequencies with increasing synonymous divergence. The youngest age-cohort of human gene duplicates ($K_S = 0$), which we refer to as the 'newborn' cohort, notably comprise more than 40% of all duplicate pairs within our data set. Moreover, it appears that >50% of the young gene duplicates identified have their origins after the human-chimpanzee split ($K_S = 0.011$) (Chen and Li 2001). The exclusion of 26 duplicate sets showing evidence of gene conversion did not alter the overall L-shaped frequency distribution of human gene duplicates, with a preponderance of evolutionarily recent gene duplicates since the human-chimpanzee split.

Genome Distance between Human Paralogs as a Function of Evolutionary Age

Where do newborn gene duplicates take up residence in the genome and does the pattern of distribution change with increasing evolutionary age? We used two measures to infer the genomic distribution of paralogs in the human genome, namely (i) the chromosomal location (*intra-* vs. *interchromosomal* locations for paralogs residing on the same and different chromosomes, respectively) and (ii) the genomic distance (unique sequence in bp) separating two *intrachromosomal* paralogs as a function of synonymous divergence, K_S. These two analyses were restricted to 172 gene duplicate pairs with known chromosomal locations for both paralogs.

With respect to chromosomal location, 83% (143/172) of the entire data set of 172 gene duplicate pairs comprise *intrachromosomal* duplications with both paralogs residing on the same chromosome; the remaining 17% (29/172) pairs display *interchromosomal* location of the two paralogs (Fig. 2). The exclusion of 26 duplicate pairs exhibiting gene conversion resulted in 82% (121/148) *intrachromosomal* and 18% (27/148) *interchromosomal* duplications, respectively. We further investigated whether the relative frequencies of *intrachromosomal vs. interchromosomal* duplicates was altered with increasing evolutionary age by classifying the human duplicate pairs into three evolutionary age-cohorts ($K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$). Although *intrachromosomal* duplicates dominate in frequency within each of the three age-cohorts, a clear decline in the frequency of *intrachromosomal* duplicates (and increase in the frequency of *intrachromosomal* duplicates) is apparent as a function of increasing

synonymous divergence: 100% (39/39), 88% (65/74), and 66% (39/59) from evolutionary younger to older age-cohorts (Fig. 2). A *G*-test of independence revealed chromosomal location to be significantly associated with synonymous divergence between paralogs ($G = 25.1$, $df = 2$, $p = 3.59e\text{-}06$). This significant trend of frequency decline of *intrachromosomal* duplicates with increasing evolutionary age remains unaltered even when the 26 duplicates pairs with signatures of gene conversion are excluded from the analyses ($G = 23.2$, $df = 2$, $p = 9.35e\text{-}06$). RNA-mediated gene duplicates appear to be older on average (higher K_S) and more likely to be found on different chromosomes. These biases in the features of RNA-mediated duplications may be responsible for the apparent relationship between chromosomal location (*intra-* vs. *interchromosomal*) and evolutionary age (K_S). However, when 21 putative RNA-mediated gene duplicate pairs were excluded from the analysis, we still found a significant increase in the proportion of *interchromosomal* duplicates with evolutionary age ($G = 10.2$, $df = 2$, $p = 0.006$).

When only *intrachromosomal* paralogs within our data set of duplicate pairs with $K_S \leq 0.1$ were analyzed (143 duplicate pairs), the correlation between K_S and log (distance) is not significant ($r = -0.08$, $df = 141$, $p = 0.84$) (Fig. 3), suggesting no increase in genomic distance between *intrachromosomal* paralogs over evolutionary time. The results were qualitatively the same when 22 *intrachromosomal* duplicate sets with a signature of gene conversion were omitted from the analysis ($r = -0.09$, $df = 119$, $p = 0.87$).

Chromosomal Distribution of Gene Duplicates

Are gene duplicates randomly distributed across all 24 chromosomes in the human genome or are they clustered on certain chromosomes? To correct for the variable number of protein-coding genes among chromosomes, we normalized the data by plotting the number of duplicate pairs/number of protein-coding genes per chromosome. Duplicated genes appear to be more frequent on the sex chromosomes than on the autosomes, but randomly distributed among autosomes. A *G*-test of differences in the frequency of *intrachromosomal* duplications among chromosomes was significant (G = 37.53, df = 23, p = 0.029), but not significant when only autosomes were considered (G = 24.52, df = 21, p = 0.27). When all duplicates (*intra-* and *interchromosomal*) in our study were considered, there was a significant difference in the frequency of duplications across chromosomes (G = 36.8, df = 23, p = 0.034) (Fig. 4), but no significant difference when only autosomes were considered (G = 21.9, df = 21, p = 0.405). Chromosomes X and Y have approximately three- and 17-fold more duplicates, respectively, than expected under an assumption of equal duplication frequencies across all chromosomes. The exclusion of 26 duplicate sets with evidence of gene conversion did not qualitatively change the above results (*intrachromosomal* duplications across all chromosomes: G = 43.99, df = 23, p = 0.0052; *intrachromosomal* duplications across all autosomes: G = 28.73, df = 21, p = 0.1206; *intra-* and *interchromosomal* duplications across all chromosomes: G = 42.07, df = 23, p = 0.0089; *intra-* and *interchromosomal* duplications across all autosomes: G = 25.3, df = 21, p = 0.234).

We further investigated if the distribution of human gene duplicates occurs in a random fashion along the length of a chromosome or exhibits a biased gradient of location, in proximity to the centromeres. The distribution of gene duplicates along the

length of chromosomes shows significant deviation from a random expectation based on gene density on chromosomes (G = 54.9, df = 14, p = 8.96e-07). Collectively, regions within 10 Mb distance from the centromeres appear to be particularly enriched for gene duplicates (Fig. 5). The exclusion of 26 duplicate sets with evidence of gene conversion did not qualitatively change the above results (G = 54.18, df = 14, p = 1.198e-06).

Equal Proportions of Intrachromosomal Paralogs with Direct and Inverse Transcriptional Orientation

Does the orientation of a duplicated gene relative to its ancestral gene influence its chances of survival? Of 143 young gene duplicates on the same chromosome, there are 46 % (66/143) and 54% (77/143) duplicates with *direct* and *inverse* transcriptional orientation, respectively. However, the proportion of inverted duplications is not significantly greater than those with the same (*direct*) transcriptional orientation (G = 0.844, df = 1, p = 0.36). The exclusion of 22 *intrachromosomal* duplicate sets with evidence of gene conversion did not qualitatively change the above results, finding no significant difference in the proportion of *direct* (54%; 54/121) versus *inverted* (55%; 67/121) duplicates (G = 1.39, df = 1, p = 0.24). A comparison of three age-cohorts of gene duplicates (K_S = 0, 0 < K_S ≤ 0.025, and 0.025 < K_S ≤ 0.1) detected no difference in the relative proportions of *direct vs. inverse* duplicates (G = 1.7949, df = 2, p = 0.41), suggesting no change in their frequencies with increasing evolutionary age. An identical trend was observed when 22 *intrachromosomal* duplicate sets with gene conversion were excluded from the analyses (G = 1.63, df = 2, p = 0.44).

Predominance of Young Gene Duplicates with Complete Structural Resemblance in the Human Genome

The structural resemblance between gene paralogs can influence their evolutionary dynamics. For DNA-mediated duplication events (N = 163 duplicate pairs), paralogs bearing *complete* structural resemblance dominate the sample of young human gene duplicates. The frequencies of *complete*, *partial*, and *chimeric* gene duplicates within our data set were 83%, 13%, and 4%, respectively. *Complete* duplicates represent the most common structural category even when gene duplicates of varying evolutionary age were analyzed (cohorts $K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$). However, the proportion of *complete* duplicates declines with evolutionary age (Fig. 6), comprising 93, 76, and 83% of the total duplicate pairs in the $K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$ age-cohorts, respectively. Furthermore, there was a significant difference in the relative proportions of the three structural categories of gene duplicates ($G = 11.9$, $df = 4$, $p = 0.018$) as a function of evolutionary age as represented by three different age-cohorts of gene duplicates ($K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$). This significant difference in the relative proportions of the three structural categories of gene duplicates as a function of K_S was also observed when 26 duplicate sets with gene conversion were excluded from the analyses ($G = 11.87$, $df = 4$, $p = 0.018$).

Duplication Span Exceeds the Average Gene Length in the Human Genome

The length of the duplication tract, which we refer to as the *duplication span*, is an important characteristic of gene duplicates that has bearing on the structural features

of newly duplicated genes as well as aspects relating to gene dosage. For example, short or abbreviated duplication spans are less likely to duplicate an ancestral ORF in its entirety. Very lengthy duplication spans are more likely to duplicate multiple ORFs and increase the probability of detrimental changes relating to gene dosage. What is the length distribution of duplication tracts involving protein-coding sequences in the human genome? The coding regions (from the initiation codon to the termination codon) of human protein-coding genes have a median and mean length of 25 and 65 kb, respectively. The duplication span within our data set of human gene duplicate pairs ranged from 136 bp - 1,055 kb, with a median and mean value of 36 and 86 kb, respectively. The duplication span of young human gene duplicates is significantly greater than the human gene length (Wilcoxon Rank Sum Test, $W = 2{,}102{,}894$, $p = 0.0015$) as well as the length of the coding region for protein-coding genes (Wilcoxon Rank Sum Test, $W = 2{,}367{,}542$, $p = 7.61\text{e-}11$) (Fig. 7). The span of DNA-mediated duplications shows a significant decrease with evolutionary age (Kendall's Tau = -0.258, $p = 2\times10^{-6}$) (Fig. 8). This significant reduction in the span of paralogs formed by DNA-mediated duplication events is observed even when 26 duplicate sets with gene conversion were excluded from the analyses (Kendall's Tau = -0.242, $p = 4.4\text{e-}05$). In contrast, there is no significant change in the span of putative retrotransposed duplicates as a function of K_S (entire data set, Kendall's Tau = 0, $p = 1$; exclusion of 26 duplicate sets with evidence of gene conversion, Kendall's Tau = -0.041, $p = 0.83$) (Fig. 8).

Smaller, but Persistent Presence of RNA-Mediated Duplications in Human Evolution

What is the frequency and fate of RNA-mediated duplication events relative to DNA-mediated ones in the human genome? Within our data set of 184 human duplicate pairs, 11.4% (21/184) were identified as putative retrotransposed gene duplicates. Interestingly, putative retroposed gene duplicates were completely absent in the youngest $K_S = 0$ age-cohort although their proportions appear to increase with age; 10% and 21% of all gene duplicates in the $0 < K_S \leq 0.025$ and $0.025 < K_S \leq 0.1$ age-cohorts, respectively. Furthermore, the genomic distribution of *retrotransposed* gene duplicates is significantly different from their DNA-mediated counterparts ($G = 76.04$, $df = 1$, $p = 2.2 \times 10^{-6}$). As expected, *retrotransposed* gene duplicates are predominantly *interchromosomal* whereas the majority of DNA-mediated duplication events yield *intrachromosomal* paralogs (Fig. 9). Of the 21 retrotransposed gene duplicates, seven and zero duplicate pairs had one paralog located on the X and Y chromosome, respectively. With respect to the seven retrotransposed duplicate pairs with one paralog residing on the X chromosome, four paralogs had intact introns and three paralogs were lacking introns, thereby suggesting approximately equal rates of traffic from and to the X chromosome.

Discussion

Structural and genomic features of recent gene duplicates can have important consequences for their evolutionary fate. For instance, gene duplications that contain the complete coding and regulatory sequences of the ancestral gene are more likely to have conserved the ancestral function compared to gene duplications that are incompletely duplicated. Similarly, gene duplicates that alter their genomic location or transcriptional orientation are more likely to be expressed differently from their ancestral paralogs. While human paralogs have been intensively studied in the last decade as a class of mutations within population-genomic studies investigating copy-number variants, a systematic and unbiased investigation delineating their basic structural and genomic features at, or close to inception, has been lacking.

We focused on 184 human gene duplicate pairs belonging to small gene families with <10% sequence divergence at synonymous sites (K_S), under the assumption that the degree of synonymous divergence is an appropriate proxy for evolutionary age for low estimates of K_S. Where possible, we compared the various genomic and structural features of different age-cohorts human paralogs to determine if any patterns are altered with increasing evolutionary age. We applied the same methodology to conduct our analyses of human gene duplicates as used previously for *C. elegans* and yeast paralogs (Katju and Lynch 2003; Katju et al. 2009) to facilitate direct comparison of the spectrum and properties of paralogs across these diverse eukaryotic genomes.

Ectopic gene conversion between homologous sequences, a form of concerted evolution, can homogenize the sequences of evolutionary older paralogs and lead to erroneous estimates of their evolutionary age as measured by the degree of synonymous divergence between paralogs (K_S). Although we currently lack any genome-wide direct empirical estimates of the spontaneous rate of ectopic gene conversion in humans or other species, it appears to be a ubiquitous process leading to sequence homogenization between paralogs in virtually all organisms that have been studied including humans (Deeb et al. 1994; Dumont and Eichler 2013; Fawcett and Innan 2013; Iatrou et al. 1984; Innan 2003; Katju and Bergthorsson 2010; Leigh Brown and Ish-Horowicz 1981; Liebhaber et al. 1981; Ollo and Rougeon 1983; Petes and Hill 1988; Rane et al. 2010; Santoyo and Romero 2005; Semple and Wolfe 1999). A growing list of human inherited diseases that result from ectopic gene conversion events between a pseudogene and its functional paralog would suggest an important and frequent role for gene conversion in the evolution of the human genome (Chen et al. 2007) although the dependency of the rate of gene conversion on various features of human paralogs (e.g. gene family size, age of paralogs, length of homologous sequence tract) is obscure. A high rate of ectopic gene conversion between members of duplicates pairs could contribute, in some part, to the higher frequencies of gene duplicates in the younger age-cohorts and thereby influence conclusions regards their evolutionary dynamics. While several studies have demonstrated evidence for frequent gene conversion among human paralogs (Dumont and Eichler 2013; Fawcett and Innan 2013), a study of four mammalian genomes including humans found a minimal contribution of ectopic gene conversion in the evolution of young gene duplicates (McGrath et al. 2009). Furthermore, Semple and

Wolfe (1999) demonstrated that the frequency of ectopic gene conversion events in *C. elegans* is positively correlated with gene-family size (Semple and Wolfe 1999). To guard against the confounding effects of gene conversion in our understanding of the early evolutionary dynamics of human paralogs, we restricted our data set to putatively young paralogs in small gene-families of five members or less. We additionally tested all duplicate pairs within our data set for a signature of gene conversion via GeneConv (Sawyer 1989) using chimpanzee orthologs as outgroup sequences, and determined 26 duplicate sets (14%; 26/184) displaying significant signatures of gene conversion. All analyses involving K_S as a variable was conducted on (i) the entire data set of 184 duplicate pairs, and (ii) 158 duplicate pairs with no signature of gene conversion. Inclusion or exclusion of the 26 converted duplicate pairs did not qualitatively alter our results pertaining to the evolutionary dynamics of human paralogs within our data set.

In concordance with genome-wide studies of extant gene duplicates in humans and other species (Lynch and Conery 2000), the distribution of human gene duplicates with low synonymous sequence divergence is strongly L-shaped, with 23% of the paralogs being identical at synonymous sites. The highest density of gene duplicates occurs in the youngest ($K_S = 0$) age-cohort followed by a strong decline in gene duplicate frequencies with increasing synonymous divergence. Although positive selection has been implicated in the spread and maintenance of some human gene duplicates, the most obvious explanation for this trend of continuing decline of duplicates with increasing synonymous divergence is a high rate of duplicate gene loss and suggests that a large fraction of the recent gene duplicates still lingering in our genomes are either evolving

neutrally under drift conditions, or being exposed to weak negative selection (Cotton and Page 2005). If recently duplicated genes are evolving neutrally, an association between their structural characteristics and K_S should either reflect (i) differences in the rate of loss of duplicate genes belong to different structural categories, or (ii) secondary changes to one or both of the paralogs subsequent to duplication. Alternatively, different structural categories of paralogs may be subject to different levels of purifying selection.

89% of genes duplicates within our data set bear signatures of origin from DNA-mediated events. This genomic proximity between paralogs suggests a major role for slippage and unequal exchange as major mutational mechanisms in the creation of human gene duplicates. Non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ) are two mechanisms of double-strand break repair that are implicated as common mutational mechanisms for the origin of gene duplicates. While we did not conduct sequence analysis of breakpoint junctions of paralogs within our data set to distinguish their relative contributions, both mechanisms likely contributed to the formation of gene duplicates from DNA-mediated events in our data set. The relative contributions of NAHR and NHEJ in generating structural variants in humans and other nonhuman primates is still under debate, with some studies favoring NAHR as the dominant mutational mechanism in the creation of copy-number variation (including duplications) (Gokcumen et al. 2013; Perry et al. 2008) and others implicating NHEJ in the creation of human structural variation across the genome (Korbel et al. 2007) and in the origin of segmental duplications in human subtelomeric regions (Linardopoulou et al. 2005). Furthermore, the role of *Alu SINE* elements in mediating human segmental

duplications remains to be resolved. Bailey et al. (2003) found that segmental duplications in the human genome with high sequence identity (<9% divergence) were significantly enriched for *Alu* elements in their breakpoint junctions, noting that *intrachromosomal* paralogs separated by 1 Mb of unique intervening sequence had the highest association with *Alu* elements (Bailey et al. 2003). Hence, they argue for a significant role of *Alu* elements in the origin of primate segmental duplications. In contrast, other studies of human structural variation including gene duplications have failed to find evidence for enrichment of *Alu* elements or other repeats in duplication breakpoint junctions (Korbel et al. 2007; Zhang et al. 2005).

The vast majority of gene duplicates in our data set (83%) tend to reside on the same chromosome (*intrachromosomal* duplicates), which may implicate NAHR in their formation. Zhang et al. (2005) also noted an excess of *intrachromosomal* gene duplicates for 15 human chromosomes. However, this pattern is not human-specific, having been observed for segmental duplications in orangutans and chimpanzees but not macaques (Gokcumen et al. 2013) as well as in *C. elegans* (Katju and Lynch 2003). Human *intrachromosomal* gene duplicates tend to be significantly larger in size and possess greater sequence similarity than their *interchromosomal* counterparts (Zhang et al. 2005). The latter could be explained if *intrachromosomal* duplicates represent (i) evolutionary recent duplicates, and/or (ii) experience higher rates of sequence homogenization. Indeed, the genomic proximity of paralogous genes is often thought to facilitate ectopic gene conversion (Petes and Hill 1988; Semple and Wolfe 1999) although Benovoy and Drouin (Benovoy and Drouin 2009) found no evidence for greater conversion frequencies

between humans paralogs in genomic proximity when the distribution of gene-family members was controlled for.

With respect to *intrachromosomal* duplicates, paralogs in *inverse* transcriptional orientation are equally frequent as paralogs in *direct* orientation. Inter-cohort comparisons found no significant difference in the proportions of *direct vs. inverted intrachromosomal* paralogs with increasing evolutionary age. This pattern of transcriptional orientation of putatively young human paralogs is in direct contrast to *C. elegans*. In *C. elegans*, a significant majority of *intrachromosomal* duplicates within the $K_S = 0$ age-cohort tend to occur as adjacent loci in inverted orientation but evolutionary older paralogs exhibit roughly equal proportions of *inverse vs. direct* orientation (Katju and Lynch 2003). Hence, humans appear to have a lower proportion of inverted duplications at birth than *C. elegans*. The results suggest that *direct* paralogs in the human genome are equally stable as *inverted* duplicates and local-scale inversion events do not play a major role in secondary movement or switching of transcriptional orientation with the progression of evolutionary time.

Studies of gene duplicates in eukaryotic genomes have detected an increase in distance between paralogs with increasing age (K_S), a trend frequently ascribed to secondary movement of genes (Achaz et al. 2001; Lercher et al. 2003). That is, the derived, duplicated locus originates in close proximity to the ancestral locus and at some later point in evolutionary time, secondary gene rearrangements lead one or both paralogs to new and more distant genomic locations. This 'secondary movement' hypothesis, if

true, would be manifest as a positive relationship between K_S and genomic distance. However, this positive correlation between duplicate age and genomic distance could also be explained by the differential survival of paralogs. The loss of duplicate genes may be facilitated by their proximity, for instance, by more frequent unequal crossing-over between closely-spaced paralogs. In the event that gene loss occurs by unequal crossing-over, there will be more intervening genetic material deleted the further apart the duplicates are, thereby increasing the magnitude of associated deleterious consequences of gene loss. There was a significant enrichment in the frequency of human *interchromosomal* paralogs with evolutionary time. This trend is still significant even when we exclude RNA-mediated duplications (characterized by high K_S values and occurrence on different chromosomes) from our analysis. All of the human duplicate pairs in the $K_S = 0$ cohort (39/39 pairs) have an *intrachromosomal* distribution, suggesting that new duplicates in the human genome overwhelmingly originate on the same chromosome as the parental copy, a pattern similar to that in *C. elegans* and *Drosophila melanogaster* (Katju and Lynch 2003; Zhou et al. 2008) but in contrast to small segmental duplications in *S. cerevisiae* (Katju et al. 2009). We did not find a significant correlation between K_S and the distance between extant *intrachromosomal* paralogs suggesting that (i) paralogs on the same chromosome do not migrate away from each other with evolutionary time, and (ii) nor do closer-spaced *intrachromosomal* paralogs suffer a higher loss rate. However, the decline in the frequency of *intrachromosomal* paralogs with evolutionary time can only be explained by (i) higher instability and loss rate of *intrachromosomal* duplicates, and/or (ii) secondary movement of paralogs to a new chromosome. We acknowledge that rearrangements do occur and

genes may get translocated further apart and onto different chromosomes. But the differential loss rate of gene duplicates, with higher rates of loss for paralogs in close proximity and a lower loss rate for duplicates further apart, may account for most of the observed relationship between distance and age (K_S). The findings that evolutionarily older gene duplicates possess higher proportions of *interchromosomal* duplicates and a lack of association between distance and K_S among *intrachromosomal* paralogs is similar to a previous result in *C. elegans* (Katju and Lynch 2003).

The chromosomal distribution of young gene duplicates can elucidate whether there exist certain mutational hotspots for their origin with respect to specific chromosomes as well as locations along the gradient of a chromosome. Regards chromosomal location, the distribution of gene duplicates on autosomes did not differ significantly from a random distribution, after normalizing for chromosome-specific gene density. Hence, the probability of a gene duplication or retention of gene duplicates does not appear to differ between the autosomes. However, there was an abundance of gene duplicates on the sex chromosomes (three- and 17-fold on the X and Y chromosomes, respectively), after accounting for the density of protein-coding genes. It is possible that the duplication rates are higher on the sex chromosomes than the autosomes, or the retention of sex-linked gene duplicates is higher (lower loss rate). The abundance of putative young gene duplicates on the Y chromosome is notable given that it is an especially gene depauperate environment. With respect to the location of gene paralogs along chromosomes, we found evidence for spatial clustering of duplicates with centromeric regions exhibiting a significant excess of gene duplicates. This nonrandom,

pericentromeric gradient of duplications in the human genome has been noted by preceding studies of rodent paralogs (Guryev et al. 2008), human gene duplicates on Chromosome 22 (Bailey et al. 2002a) as well as at a genome-wide scale (Bailey et al. 2001; Cheung et al. 2003; Fortna et al. 2004; Linardopoulou et al. 2005; Zhang et al. 2005). This pattern, moreover, is not restricted to humanoids. Emerson et al. (2008) observed an enrichment of duplications in pericentromeric regions in a population-genomic study of CNVs in 15 isofemale *D. melanogaster* lines (Emerson et al. 2008).

The degree of structural resemblance between paralogs has implications for the evolution of functionally novel genes following duplication. It has been argued that the evolution of novel functions in a new gene duplicate may be facilitated by radical changes in the exon-intron structure of the derived copy, typically manifest in structurally heterogeneous paralogs comprising *partial* and *chimeric* duplicates (Katju 2012). As such, *partial* and *chimeric* duplicates may be worthy candidate genes for investigations into the genetic basis of human-specific traits. Indeed, the novel human genes *PMCHL1* and *PMCHL2* arose from retrotransposition, partial duplication of an ancestral neuropeptide precursor in conjunction with recruitment of downstream noncoding DNA to yield novel ORFs (Courseaux and Nahon 2001). This complex sequence of partial duplication with recruitment and retrotransposition may have facilitated the novel function and expression patterns in the human testes and brain (Courseaux and Nahon 2001). The origin of a *partial* duplicate of an ancestral *SRGAP2* gene is implicated in enhanced cognitive abilities in humans since divergence from our primate ancestors (Charrier et al. 2012; Dennis et al. 2012). Our comparisons of the exon-intron structure

of paralogs revealed that *complete* duplicates are the dominant structural category of gene duplicates stemming from DNA-mediated duplication events within the human genome, comprising 83% of all gene duplicate pairs within our data set. The remaining 17% gene of duplicate pairs stemming from DNA-mediated duplication events comprise structurally heterogeneous duplicates (13% *partial* duplicates, and 4% *chimeric* duplicates). *Complete* duplicates represent the most frequent structural category of duplicates in all three age-cohorts although there is a noticeable decline in their frequency with increasing evolutionary age. This decline in frequency of *complete* duplicates in our set of human gene duplicates is in stark contrast to the pattern observed in macaques, orangutans and chimpanzees wherein the ratio of *complete*/*partial* gene duplications increased as a function of evolutionary age (Gokcumen et al. 2013).

The predominance of *complete* duplicates in the human genome is also notably different from the genomes of a handful of other multicellular eukaryotic species in which detailed structural characterization of paralogs has been conducted at a genome-wide scale. Structurally heterogeneous (*partial* and *chimeric*) duplicates exceed structurally homogenous (*complete*) duplicates in *C. elegans* (61%) (Katju and Lynch 2003), *D. melanogaster* (59%) (Zhou et al. 2008), and *Drosophila pseudoobscura* (56%) (Meisel 2009). The high frequency of *complete* duplicates in the human genome is especially intriguing given that the length of human protein-coding genes is quite substantial with a mean and median length of 65 and 25 kb, respectively. Because the duplication machinery is expected to be impervious to gene boundaries, the likelihood of capturing an entire ORF during duplication is more likely in compact genomes with a

shorter average gene length (Katju 2012). Given the larger genome size and average gene length in humans relative to worm and *Drosophila*, it is paradoxical that *complete* duplicates represent the most abundant structural class of gene duplicates within the human genome. However, our investigation into the distribution of duplication spans of human paralogs may provide some insight regards this paradox. The median duplication span for our data set of human gene duplicates was 36 kb, and is significantly greater than the median gene length of 25 kb for humans. Hence, the high prevalence of *complete* duplication events within our data set of young human gene duplicates may be explained by human duplicons having lengthier tracts, although the role of purifying selection against shorter duplication tracts yielding *partial* and *chimeric* duplicates cannot be ruled out. However, with increasing evolutionary age, we observed a significant increase in the frequency of both *partial* and *chimeric* duplicates as well as a concomitant attenuation of duplication spans. This pattern has two alternative explanations, namely (i) enhanced survivorship of *partial* and *chimeric* duplicates and/or stronger selection against *complete* duplicates, or (ii) gene rearrangements or deletion events that serve to erode the sequences of lengthier, *complete* duplicates and thereby reduce their detectable duplication spans. The large fraction of *complete* duplicates within our data set begs the question as to how the majority of newly minted human duplicate genes are able to rapidly assume unique species-specific functions. While the relationship between structural category of duplicates and signatures of accelerated evolution has not been conducted at a genome-wide scale in humans, there is some evidence to suggest that human paralogs can diverge rapidly. Zhang et al. (2003) found that for a large fraction of putatively young human paralogs ($Ks < 0.3$), one copy exhibited a signature of rapid

molecular evolution at the amino-acid level and less stringent selective constraints (high K_A/K_S ratios) (Zhang et al. 2003). Makova and Li (2003) demonstrated diverged spatial expression profiles for a large proportion of human paralogs, noting that the expression divergence increased approximately linearly with evolutionary time (K_S) (Makova and Li 2003). In a study of the expression of *complete* gene duplications in six tissues in humans and nonhuman primates, Gokcumen et al. (2013) found that the emergence of new *complete* duplicates often coincides with gene expression in new tissues (Gokcumen et al. 2013). In a similar vein, analysis of a human gene coexpression network revealed that even evolutionarily young gene duplicates rapidly gained new coexpressed partners (Chung et al. 2006). Studies of the patterns of sequence and functional divergence between human paralogs can be further elucidated by future investigations into whether, and the extent to which, structural resemblance between paralogs impinges on the evolution of novel function. Is the evolution of novel function primarily facilitated by changes to the intron-exon structure of the derived copy relative to its progenitor as manifest in *partial* and *chimeric* duplicates or do regulatory changes (rapid promoter evolution or the gain of novel promoters) play a significant role?

Although, DNA-mediated events are responsible for the origin of the vast majority of young gene duplicates in the human genome, we identified ~11% of duplicates (21 of 184 duplicate pairs) as putatively originating from RNA-mediated events. These putative retroduplicates pairs possessed several key diagnostic characteristics that implicated retrotransposition as the mutational mechanism of origin, namely (i) a single exon paralog lacking introns present in the other multiexonic paralog,

(ii) *interchromosomal* location of the two paralogs, and (iii) the lack of significant flanking region sequence homology between the two paralogs. The age distribution of putative retroposed human gene duplicates presented an interesting pattern, displaying increased frequencies with increasing evolutionary age (K_S), and a complete absence of retroposed duplicates in the $K_S = 0$ age-cohort. Although the small sample size of retroposed duplicates within our data set precludes a robust explanation for this trend, we speculate that this pattern may represent a burst of retrotranspositional activity yielding gene duplicates in our species' recent evolutionary past.

Our analyses of putative young gene duplicates in the human genome have revealed both similarities and differences with other species. As in *C. elegans*, there is a significant increase in the proportion of *interchromosomal* paralogs with increasing evolutionary age, but without a similar increase in distance with age within *intrachromosomal* paralogs. Two alternative hypotheses can account for the observed genomic distribution of human paralogs, namely (i) greater genomic stability of *interchromosomal* paralogs relative to *intrachromosomal* paralogs, or (ii) secondary movements of paralogs to nonhomologous chromosomes. Young human paralogs differ in some other aspects from their counterparts in *C. elegans* and *Drosophila*. For instance, inverted duplications are less common among the most recent paralogs in humans than in *C. elegans* (Katju and Lynch 2003), but their proportions are stable with age. This may indicate differences in prevailing duplication and duplication loss mechanisms in these species. In addition, human duplicates have, on average, much larger duplication spans which are more likely to capture entire ORFs leading to *complete* duplicates compared to

higher proportions of structurally heterogeneous duplicates (*partial* and *chimeric* duplications) in *Drosophila* and *C. elegans*. The change in the genomic and structural features of human paralogs with evolutionary time demonstrate that (i) genomic context and structural similarities have important consequences for the fate of duplicate genes, and (ii) the mutational spectrum of gene duplicates and their subsequent evolutionary dynamics can vary significantly among eukaryotic species. In conclusion, our study serves to bridge key characteristics of human gene duplicates upon origin in an evolutionary context with the plethora of data-rich population-genomic studies and also sets the stage for additional analyses of the gene duplication landscape in the genome of our closest relative, the chimpanzee.

Tables

Supplementary Table 1 – Evolutionary and genomic features of 184 gene duplicates with low synonymous divergence in the human genome.

Structural resemblance between paralogs within a duplicate was classified as (i) complete if sequence homology between the focal paralogs extended throughout their entire open reading frames (ORFs), from the start to the stop codon and possibly extending into one or both flanking regions; (ii) partial if one paralog possessed unique exon(s) and/or intron(s) in its ORF that are absent in the other paralog; (iii) chimeric if both paralogs contain unique exon(s) and/or intron(s) within their respective ORFs, to the exclusion of the other paralog ; 4) retroposed if the ORF of one paralog contained one or more introns which were absent in the other paralog's ORF. Accession numbers correspond to Ensembl ID version 72 released in June 2013.

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000232948	ENSG00000233050	0.0000	0.0000	0.5051	8/8	147895	-/-	Complete	98670	NO
ENSG00000230000	ENSG00000268181	0.0000	0.0000	0.4129	7/7	8473	+/-	Complete	147994	NO
ENSG00000182646	ENSG00000179304	0.0000	0.0000	0.5221	X/X	9934	-/+	Complete	44513	NO
ENSG00000215033	ENSG00000215020	0.0000	0.0033	99.0000	10/10	418855	+/+	Complete	88553	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000269358	ENSG00000269831	0.0001	0.0121	99.0000	1/1	223324882	-/-	Complete	152319	NO
ENSG00000197077	ENSG00000251180	0.0007	0.0000	0.0010	22/GL000242.1	NA	+/+	Partial	43479	NO
ENSG00000183474	ENSG00000145736	0.0036	0.0034	0.9352	5/5	1362518	+/-	Complete	95695	NO
ENSG00000205595	ENSG00000109321	0.0042	0.0000	0.0010	4/4	128793	+/+	Complete	40981	NO
ENSG00000168028	ENSG00000205246	0.0049	0.0045	0.9063	3/19	NA	+/+	Retroposed	2527	NO
ENSG00000204661	ENSG00000228259	0.0060	0.0656	10.9832	5/5	7724	-/+	Partial	6069	NO
ENSG00000197620	ENSG00000197021	0.0076	0.0101	1.3342	X/X	441777	+/-	Complete	29776	NO
ENSG00000182356	ENSG00000239511	0.0096	0.0084	0.8760	22/22	2614973	+/-	Complete	38208	NO
ENSG00000205076	ENSG00000178934	0.0152	0.0000	0.0010	19/19	8415	-/+	Complete	7597	NO
ENSG00000184945	ENSG00000185176	0.0157	0.0044	0.2810	2/2	3178	+/-	Complete	13058	NO
ENSG00000105835	ENSG00000229644	0.0215	0.0029	0.1368	7/10	NA	-/-	Retroposed	2701	NO
ENSG00000204936	ENSG00000204933	0.0231	0.0110	0.4752	19/19	3228	+/-	Partial	29644	NO
ENSG00000186825	ENSG00000197927	0.0291	0.0290	0.9971	2/2	19789	-/+	Chimeric	9690	NO
ENSG00000122852	ENSG00000185303	0.0294	0.0167	0.5684	10/10	34380	+/-	Complete	11159	NO
ENSG00000269337	ENSG00000268578	0.0300	0.0000	0.0010	9/9	25025042	+/-	Complete	151099	NO
ENSG00000143185	ENSG00000143184	0.0308	0.0084	0.2743	1/1	27416	-/+	Complete	6164	NO
ENSG00000148672	ENSG00000182890	0.0318	0.0220	0.6921	10/X	NA	-/+	Retroposed	3180	NO
ENSG00000115042	ENSG00000144199	0.0323	0.0098	0.3028	2/2	1668109	+/-	Complete	14879	NO
ENSG00000269099	ENSG00000130592	0.0326	0.0102	0.3130	13/11	NA	+/+	Partial	136	NO
ENSG00000175548	ENSG00000139133	0.0330	0.0250	0.7555	12/12	4364951	+/+	Complete	152707	NO
ENSG00000124172	ENSG00000180389	0.0421	0.0418	0.9932	20/13	NA	-/+	Retroposed	367	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000196459	ENSG00000256060	0.0451	0.0000	0.0010	X/19	NA	-/+	Retroposed	2680	NO
ENSG00000229924	ENSG00000171847	0.0464	0.0335	0.7224	4/12	NA	+/-	Complete	233500	NO
ENSG00000114547	ENSG00000065371	0.0472	0.0228	0.4840	3/3	1917158	+/-	Complete	61111	NO
ENSG00000033011	ENSG00000189366	0.0529	0.0384	0.7247	16/3	NA	+/-	Partial	24702	NO
ENSG00000128185	ENSG00000183628	0.0542	0.0224	0.4135	22/22	1376928	-/+	Complete	21921	NO
ENSG00000169469	ENSG00000169474	0.0554	0.0330	0.5954	1/1	44865	+/+	Complete	2205	NO
ENSG00000099721	ENSG00000125363	0.0564	0.0725	1.2863	Y/X	NA	-/+	Complete	10875	NO
ENSG00000188672	ENSG00000187010	0.0660	0.1021	1.5463	1/1	33395	-/+	Complete	61003	NO
ENSG00000110057	ENSG00000233094	0.0707	0.0205	0.2893	11/GL000222.1	NA	-/+	Partial	53411	NO
ENSG00000212643	ENSG00000169249	0.0724	0.0294	0.4059	5/X	NA	+/+	Retroposed	3083	NO
ENSG00000253797	ENSG00000156697	0.0780	0.0489	0.6265	13/X	NA	+/+	Retroposed	2481	NO
ENSG00000178700	ENSG00000228716	0.0785	0.0452	0.5760	3/5	NA	-/-	Retroposed	3499	NO
ENSG00000229314	ENSG00000228278	0.0793	0.0476	0.6000	9/9	1	+/+	Complete	7053	NO
ENSG00000166926	ENSG00000110077	0.0817	0.1230	1.5051	11/11	152178	+/-	Partial	7898	NO
ENSG00000243709	ENSG00000143768	0.0851	0.0210	0.2467	1/1	48103	-/-	Complete	2611	NO
ENSG00000243317	ENSG00000267889	0.0858	0.0199	0.2320	7/2	NA	+/+	Retroposed	848	NO
ENSG00000196659	ENSG00000197557	0.0876	0.0247	0.2815	2/2	59754	-/-	Complete	5829	NO
ENSG00000213714	ENSG00000124103	0.0896	0.0698	0.7784	20/20	3416	+/+	Complete	4416	NO
ENSG00000130741	ENSG00000180574	0.0932	0.0445	0.4777	X/12	NA	+/+	Retroposed	2925	NO
ENSG00000172115	ENSG00000269383	0.0982	0.1541	1.5690	7/2	NA	-/-	Complete	1210	NO
ENSG00000224389	ENSG00000244731	0.0025	0.0025	1.0186	6/6	1	+/+	Complete	32741	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000122543	ENSG00000135175	0.0259	0.0101	0.3897	7/7	91655248	+/-	Complete	51956	NO
ENSG00000185897	ENSG00000126251	0.0001	0.0076	99.0000	19/19	9518	+/+	Complete	2936	NO
ENSG00000253626	ENSG00000132507	0.0175	0.0088	0.5052	10/17	NA	+/+	Retroposed	1134	NO
ENSG00000187630	ENSG00000157326	0.0000	0.0000	0.4341	14/14	1	+/+	Complete	35350	NO
ENSG00000269011	ENSG00000268851	0.0000	0.0000	0.0010	21/GL000215.1	NA	-/-	Complete	172472	NO
ENSG00000170074	ENSG00000204677	0.0001	0.0089	99.0000	5/5	217864	-/+	Complete	51458	NO
ENSG00000072444	ENSG00000204147	0.0001	0.0057	99.0000	10/10	4485559	-/+	Partial	43204	NO
ENSG00000188611	ENSG00000204147	0.0000	0.0031	99.0000	10/10	540267	-/+	Partial	43204	NO
ENSG00000143556	ENSG00000184330	0.0747	0.0316	0.4227	1/1	16018	-/+	Complete	24384	NO
ENSG00000122696	ENSG00000141437	0.0323	0.0214	0.6618	9/18	NA	-/-	Retroposed	1272	NO
ENSG00000212899	ENSG00000212900	0.0688	0.0088	0.1272	17/17	5142	-/-	Complete	605	NO
ENSG00000253506	ENSG00000196531	0.0728	0.0676	0.9285	17/12	NA	-/-	Retroposed	836	NO
ENSG00000197110	ENSG00000183709	0.0207	0.0191	0.9224	19/19	13699	-/+	Complete	7838	NO
ENSG00000149531	ENSG00000220023	0.0226	0.0477	2.1115	20/GL000219.1	NA	+/-	Partial	73153	NO
ENSG00000188092	ENSG00000117262	0.0037	0.0000	0.0010	1/1	1561389	+/-	Complete	85051	NO
ENSG00000157322	ENSG00000157335	0.0104	0.0030	0.2846	16/16	157874	+/+	Complete	45147	NO
ENSG00000157335	ENSG00000140839	0.0034	0.0050	1.4431	16/16	4188011	+/-	Complete	39139	NO
ENSG00000184814	ENSG00000233701	0.0951	0.0770	0.8098	3/3	21031	-/-	Complete	2791	NO
ENSG00000184814	ENSG00000206260	0.0657	0.0492	0.7482	3/3	12807	-/-	Complete	1622	NO
ENSG00000204397	ENSG00000255221	0.0918	0.0886	0.9651	11/11	49331	-/-	Partial	6535	NO
ENSG00000171102	ENSG00000122136	0.0090	0.0439	4.8511	9/9	2290487	-/+	Complete	34319	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000184033	ENSG00000183678	0.0000	0.0000	53.4838	X/X	21622	-/+	Complete	35968	NO
ENSG00000197665	ENSG00000228157	0.0078	0.0040	0.5114	17/17	45996	-/+	Complete	45009	NO
ENSG00000228157	ENSG00000230493	0.0078	0.0040	0.5114	17/17	72764	+/+	Complete	53320	NO
ENSG00000182824	ENSG00000188280	0.0515	0.0264	0.5126	22/22	1945636	+/+	Chimeric	26142	NO
ENSG00000240247	ENSG00000239839	0.0000	0.0049	99.0000	8/8	1	-/-	Complete	19104	NO
ENSG00000240247	ENSG00000206047	0.0000	0.0000	0.0010	8/8	1	-/-	Complete	19104	NO
ENSG00000177710	ENSG00000164729	0.0237	0.0401	1.6873	8/17	NA	+/-	Complete	3057	NO
ENSG00000164729	ENSG00000259224	0.0242	0.0412	1.7006	17/17	26133000	-/+	Retroposed	1696	NO
ENSG00000254598	ENSG00000101266	0.0970	0.0548	0.5645	11/20	NA	-/-	Retroposed	1534	NO
ENSG00000166157	ENSG00000132958	0.0752	0.0881	1.1727	21/13	NA	-/-	Complete	94110	NO
ENSG00000170215	ENSG00000154537	0.0209	0.0000	0.0010	9/9	22291902	-/+	Complete	284036	NO
ENSG00000268942	ENSG00000173207	0.0503	0.0497	0.9889	5/1	NA	-/+	Retroposed	1746	NO
ENSG00000173432	ENSG00000134339	0.0591	0.0766	1.2958	11/11	15647	+/-	Partial	4464	NO
ENSG00000196312	ENSG00000148110	0.0834	0.0732	0.8785	9/9	2510563	-/+	Chimeric	78167	NO
ENSG00000099984	ENSG00000133433	0.0115	0.0018	0.1607	22/22	2128	+/-	Complete	30285	NO
ENSG00000186523	ENSG00000118894	0.0880	0.0388	0.4407	8/16	NA	-/-	Complete	49398	NO
ENSG00000237847	ENSG00000268991	0.0000	0.0000	0.4321	1/1	44330	+/-	Complete	35034	NO
ENSG00000203815	ENSG00000268991	0.0526	0.0549	1.0433	1/1	132612981	+/-	Complete	1684	NO
ENSG00000203815	ENSG00000268674	0.0635	0.0911	1.4349	1/1	132808823	+/+	Complete	1686	NO
ENSG00000099290	ENSG00000172661	0.0114	0.0062	0.5480	10/10	5503229	+/+	Complete	101659	NO
ENSG00000226784	ENSG00000171314	0.0093	0.0211	2.2660	X/10	NA	-/+	Retroposed	1689	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000212724	ENSG00000213417	0.0376	0.0000	0.0010	17/17	5218	-/-	Complete	576	NO
ENSG00000214518	ENSG00000212725	0.0000	0.0000	0.3027	17/17	7362	-/-	Complete	582	NO
ENSG00000212725	ENSG00000213417	0.0000	0.0029	99.0000	17/17	17932	-/-	Complete	661	NO
ENSG00000174428	ENSG00000196275	0.0011	0.0037	3.4274	7/7	203874	+/-	Complete	143572	NO
ENSG00000163283	ENSG00000163286	0.0632	0.0140	0.2211	2/2	18456	+/+	Complete	8712	NO
ENSG00000238083	ENSG00000176681	0.0040	0.0053	1.3368	17/17	19900	+/+	Complete	197813	NO
ENSG00000176681	ENSG00000176809	0.0478	0.0285	0.5950	17/17	18413859	+/-	Complete	66453	NO
ENSG00000139223	ENSG00000140350	0.0783	0.0629	0.8036	12/15	NA	+/-	Retroposed	646	NO
ENSG00000169763	ENSG00000169807	0.0000	0.0000	0.4611	Y/Y	1441567	-/-	Complete	168120	NO
ENSG00000269526	ENSG00000268964	0.0345	0.0239	0.6917	19/19	1	+/+	Complete	32741	NO
ENSG00000143954	ENSG00000172016	0.0973	0.0736	0.7568	2/2	127878	+/-	Complete	3641	NO
ENSG00000205456	ENSG00000205457	0.0071	0.0000	0.0010	16/16	634356	+/+	Complete	306492	NO
ENSG00000205456	ENSG00000261509	0.0071	0.0000	0.0010	16/16	865729	+/+	Complete	130935	NO
ENSG00000205457	ENSG00000183632	0.0000	0.0000	0.3835	16/16	214228	+/-	Complete	183089	NO
ENSG00000203817	ENSG00000215784	0.0001	0.0057	99.0000	1/1	5411200	-/-	Partial	35229	NO
ENSG00000188610	ENSG00000215784	0.0065	0.0142	2.1681	1/1	23007885	+/-	Complete	124962	NO
ENSG00000111775	ENSG00000226976	0.0219	0.0087	0.3964	12/6	NA	+/+	Retroposed	553	NO
ENSG00000227151	ENSG00000229665	0.0000	0.0000	0.3990	13/13	1	+/+	Complete	27725	NO
ENSG00000227151	ENSG00000234278	0.0000	0.0000	0.0010	13/13	1	+/+	Complete	27725	NO
ENSG00000204918	ENSG00000229665	0.0000	0.0000	0.0010	13/13	1	+/+	Complete	27725	NO
ENSG00000204918	ENSG00000204919	0.0000	0.0000	0.4396	13/13	1	+/+	Complete	27725	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000174876	ENSG00000187733	0.0000	0.0000	0.4701	1/1	19160	-/+	Complete	34804	NO
ENSG00000174876	ENSG00000240038	0.0724	0.0128	0.1769	1/1	107727	-/+	Complete	8790	NO
ENSG00000237763	ENSG00000187733	0.0000	0.0000	0.0592	1/1	45596	+/+	Complete	48555	NO
ENSG00000243480	ENSG00000240038	0.0550	0.0063	0.1137	1/1	37007	+/+	Complete	8829	NO
ENSG00000196507	ENSG00000204071	0.0708	0.0383	0.5408	X/X	1464257	+/-	Complete	3441	NO
ENSG00000183461	ENSG00000204375	0.0000	0.0000	0.0010	X/X	21643	+/-	Complete	28680	NO
ENSG00000244067	ENSG00000182793	0.0993	0.0667	0.6723	6/6	63506	-/-	Complete	17783	NO
ENSG00000244067	ENSG00000243955	0.0582	0.0257	0.4418	6/6	19782	-/-	Complete	20881	NO
ENSG00000228532	ENSG00000188612	0.0198	0.0057	0.2899	X/17	NA	-/-	Retroposed	2379	NO
ENSG00000177688	ENSG00000188612	0.0699	0.0785	1.1218	6/17	NA	+/-	Retroposed	1083	NO
ENSG00000220903	ENSG00000225899	0.0379	0.0267	0.7031	GL000222.1/10	NA	-/-	Complete	6409	NO
ENSG00000220903	ENSG00000172969	0.0000	0.0000	0.4768	GL000222.1/3	NA	-/+	Complete	186774	NO
ENSG00000205097	ENSG00000148828	0.0000	0.0000	0.4247	4/GL000228.1	NA	-/-	Complete	98341	NO
ENSG00000225899	ENSG00000148828	0.0037	0.0072	1.9219	10/GL000228.1	NA	-/-	Complete	74632	NO
ENSG00000169953	ENSG00000172468	0.0000	0.0000	0.0010	Y/Y	39643	-/+	Complete	190162	NO
ENSG00000185554	ENSG00000185945	0.0016	0.0000	0.0010	X/X	10765	+/-	Complete	140562	NO
ENSG00000237289	ENSG00000223572	0.0046	0.0000	0.0010	15/15	1	+/+	Complete	92931	NO
ENSG00000105889	ENSG00000164647	0.0062	0.0258	4.1599	7/7	67145303	-/+	Chimeric	48968	NO
ENSG00000196934	ENSG00000183246	0.0000	0.0005	99.0000	22/22	48673	+/-	Complete	70268	NO
ENSG00000196622	ENSG00000183246	0.0008	0.0033	3.9235	22/22	1344155	-/-	Complete	99520	NO
ENSG00000205810	ENSG00000205809	0.0813	0.1096	1.3484	12/12	4501	-/-	Partial	10755	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000206181	ENSG00000266996	0.0889	0.0865	0.9731	18/18	12842	-/-	Complete	3813	NO
ENSG00000266996	ENSG00000234298	0.0000	0.0017	99.0000	18/18	931	-/-	Complete	4992	NO
ENSG00000234298	ENSG00000183791	0.0000	0.0018	99.0000	18/18	923	-/-	Complete	4963	NO
ENSG00000223524	ENSG00000255940	0.0000	0.0000	0.0010	8/8	2970	+/-	Complete	14283	NO
ENSG00000223524	ENSG00000205176	0.0072	0.0006	0.0880	8/8	170367	+/-	Complete	3993	NO
ENSG00000119673	ENSG00000184227	0.0814	0.0374	0.4594	14/14	1452	+/+	Complete	26573	NO
ENSG00000158427	ENSG00000158164	0.0377	0.0000	0.0010	X/X	1426646	+/-	Complete	10733	NO
ENSG00000147059	ENSG00000186787	0.0038	0.0020	0.5334	X/X	10452	-/-	Complete	5259	NO
ENSG00000136488	ENSG00000213218	0.0571	0.0035	0.0605	17/17	7468	-/-	Complete	15106	NO
ENSG00000136488	ENSG00000259384	0.0953	0.0604	0.6332	17/17	16832	-/-	Complete	4504	NO
ENSG00000204807	ENSG00000204804	0.0646	0.1193	1.8476	9/9	1554	+/-	Complete	2935	NO
ENSG00000215356	ENSG00000215372	0.0206	0.0382	1.8576	8/8	63872	+/-	Complete	458748	Linked Set 1
ENSG00000171711	ENSG00000177257	0.0159	0.0000	0.0010	8/8	63872	+/-	Complete		
ENSG00000176797	ENSG00000177243	0.0000	0.0000	0.3882	8/8	63872	+/-	Complete		
ENSG00000178287	ENSG00000164871	0.0000	0.0000	0.4668	8/8	63872	+/-	Complete		
ENSG00000176782	ENSG00000177023	0.0000	0.0000	0.3248	8/8	63872	+/-	Complete		
ENSG00000186579	ENSG00000187082	0.0000	0.0000	0.5162	8/8	63872	+/-	Complete		
ENSG00000186562	ENSG00000186599	0.0000	0.0000	0.0010	8/8	63872	-/+	Complete		
ENSG00000186572	ENSG00000198129	0.0000	0.0000	0.0010	8/8	63872	-/+	Complete		
ENSG00000255378	ENSG00000255251	0.0000	0.0000	0.4910	8/8	63872	-/+	Complete		
ENSG00000103226	ENSG00000103512	0.0093	0.0024	0.2557	16/16	1257670	+/+	Complete	141089	

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000183889	ENSG00000183426	0.0095	0.0042	0.4623	16/16	1257670	+/+	Complete		Linked Set 2
ENSG00000146574	ENSG00000122674	0.0038	0.0000	0.0010	7/7	739684	-/+	Complete	99359	Linked Set 3
ENSG00000169402	ENSG00000155026	0.0017	0.0015	0.8940	7/7	739684	+/-	Complete		
ENSG00000187243	ENSG00000154545	0.0000	0.0000	0.4251	X/X	99855	-/+	Complete	36405	Linked Set 4
ENSG00000182776	ENSG00000179028	0.0000	0.0000	0.0010	X/X	99855	-/+	Complete		
ENSG00000182230	ENSG00000170074	0.0028	0.0105	3.6929	5/5	1318801	+/-	Complete	315967	Linked Set 5
ENSG00000248469	ENSG00000249109	0.0000	0.0000	0.4583	5/5	1318801	-/+	Complete		
ENSG00000214967	ENSG00000214940	0.0000	0.0000	99.0000	16/16	1569911	+/-	Complete	372537	Linked Set 6
ENSG00000183889	ENSG00000233024	0.0000	0.0000	0.4736	16/16	1569911	+/-	Complete		
ENSG00000103226	ENSG00000185164	0.0067	0.0011	0.1655	16/16	1569911	+/-	Complete		
ENSG00000174196	ENSG00000172661	0.0000	0.0000	0.0010	10/10	4760447	+/+	Partial	143021	Linked Set 7
ENSG00000174194	ENSG00000188234	0.0055	0.0039	0.7083	10/10	4760447	-/-	Complete		
ENSG00000233232	ENSG00000255524	0.0106	0.0318	0.8159	16/16	169291	-/+	Partial	134554	Linked Set 8
ENSG00000205609	ENSG00000184110	0.0000	0.0000	0.4863	16/16	169291	-/+	Complete		
ENSG00000198156	ENSG00000196993	0.0136	0.1021	0.8776	16/16	169291	-/+	Complete		
ENSG00000197859	ENSG00000215616	0.0000	0.0000	0.4657	9/GL000201.1	NA	+/+	Partial	36106	Linked Set 9
ENSG00000196990	ENSG00000215611	0.0181	0.0046	0.2548	9/GL000201.1	NA	-/-	Complete		
ENSG00000172058	ENSG00000205572	0.0000	0.0000	0.4789	5/5	381073	+/+	Complete	494845	Linked Set 10
ENSG00000172062	ENSG00000205571	0.0000	0.0145	6.7221	5/5	381073	+/+	Complete		
ENSG00000184040	ENSG00000148483	0.0000	0.0000	0.3925	10/10	50001	+/+	Complete	196904	Linked Set 11
ENSG00000120586	ENSG00000183748	0.0050	0.0009	0.1821	10/10	50001	+/+	Complete		

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000204807	ENSG00000232833	0.0003	0.0453	3.8050	9/9	21954191	+/-	Complete	145047	Linked Set 12
ENSG00000182368	ENSG00000170215	0.0000	0.0000	0.3932	9/9	21954191	+/-	Complete		
ENSG00000198444	ENSG00000185990	0.0000	0.0000	0.4466	X/X	66875	+/-	Complete	50563	Linked Set 13
ENSG00000198307	ENSG00000185978	0.0000	0.0000	1.3549	X/X	66875	+/-	Complete		
ENSG00000267985	ENSG00000268891	0.0000	0.0000	0.4675	7/7	2408306	+/-	Complete	199323	Linked Set 14
ENSG00000155428	ENSG00000178809	0.0259	0.0016	0.0625	7/7	2408306	-/+	Complete		
ENSG00000196313	ENSG00000135213	0.0192	0.0291	0.7299	7/7	2408306	+/-	Complete		
ENSG00000169627	ENSG00000183336	0.0000	0.0000	0.0010	16/16	592983	-/-	Complete	146356	Linked Set 15
ENSG00000132207	ENSG00000181625	0.0000	0.0000	0.0010	16/16	592983	+/+	Complete		
ENSG00000261052	ENSG00000213648	0.0000	0.0000	0.4986	16/16	592983	+/+	Complete		
ENSG00000198064	ENSG00000169203	0.0015	0.0000	0.0010	16/16	592983	-/-	Complete		
ENSG00000258150	ENSG00000258130	0.0000	0.0011	99.0000	16/16	592983	-/-	Complete		
ENSG00000189266	ENSG00000215700	0.0000	0.0000	0.3908	1/GL000191.1	NA	+/+	Complete	106432	Linked Set 16
ENSG00000188529	ENSG00000215699	0.0000	0.0030	99.0000	1/GL000191.1	NA	-/-	Complete		
ENSG00000204149	ENSG00000174194	0.0146	0.0066	0.4558	10/10	197048	+/-	Complete	181193	Linked Set 17
ENSG00000138297	ENSG00000204152	0.0001	0.0281	3.1339	10/10	197048	-/+	Partial		
ENSG00000235173	ENSG00000230567	0.0000	0.0011	99.0000	8/8	100001	+/+	Complete	69111	Linked Set 18
ENSG00000204775	ENSG00000170727	0.0000	0.0000	0.4028	8/8	100001	-/-	Partial		
ENSG00000268531	ENSG00000268343	0.0000	0.0046	99.0000	15/15	685431	+/+	Complete	321118	Linked Set 19
ENSG00000233917	ENSG00000230031	0.0000	0.0000	0.4397	15/15	685431	-/-	Complete		
ENSG00000152086	ENSG00000075886	0.0317	0.0060	0.1891	2/2	1080189	-/+	Complete	221851	

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000152076	ENSG00000163040	0.0175	0.0085	0.4843	2/2	1080189	-/+	Complete		Linked Set 20
ENSG00000168255	ENSG00000228049	0.0747	0.4109	0.3642	7/7	1	-/-	Complete	121324	Linked Set 21
ENSG00000005075	ENSG00000168255	0.0366	0.3899	0.4100	7/7	1	-/-	Chimeric		
ENSG00000205233	ENSG00000189093	0.0000	0.0000	0.0010	7/7	1	-/-	Complete		
ENSG00000170667	ENSG00000105808	0.0054	0.0011	0.2054	7/7	1	-/-	Complete		
ENSG00000205238	ENSG00000173678	0.0000	0.0011	99.0000	7/7	1	+/+	Complete		
ENSG00000237375	ENSG00000166351	0.0016	0.0027	1.6276	GL000213.1/21	NA	-/+	Complete	164238	Linked Set 22
ENSG00000269725	ENSG00000269011	0.0000	0.0000	0.0010	GL000213.1/21	NA	+/-	Complete		
ENSG00000152726	ENSG00000099290	0.0031	0.0015	0.4724	10/10	3855364	+/+	Partial	89925	Linked Set 23
ENSG00000072444	ENSG00000188611	0.0044	0.0096	0.6111	10/10	3855364	-/-	Partial		
ENSG00000222038	ENSG00000196834	0.0040	0.0029	0.7196	2/2	1	+/-	Complete	159213	Linked Set 24
ENSG00000136698	ENSG00000152093	0.0074	0.0019	0.2561	2/2	1	-/+	Complete		
ENSG00000183292	ENSG00000184761	0.0266	0.2235	0.5264	2/2	1	+/-	Complete		
ENSG00000188120	ENSG00000205916	0.0000	0.0025	99.0000	Y/Y	1621997	-/+	Complete	421202	Linked Set 25
ENSG00000183753	ENSG00000185894	0.0000	0.0000	0.3931	Y/Y	1621997	+/-	Complete		
ENSG00000183753	ENSG00000183795	0.0000	0.0000	99.0000	Y/Y	1238175	+/+	Complete	395549	Linked Set 26
ENSG00000188120	ENSG00000187191	0.0117	0.0091	0.7266	Y/Y	1238175	-/-	Complete		
ENSG00000185894	ENSG00000183795	0.0000	0.0000	0.4218	Y/Y	2139	-/+	Complete	408818	Linked Set 27
ENSG00000205916	ENSG00000187191	0.0078	0.0056	0.7106	Y/Y	2139	+/-	Complete		
ENSG00000196644	ENSG00000188092	0.0000	0.0026	99.0000	1/1	1260168	+/+	Partial	281531	Linked Set 28
ENSG00000152042	ENSG00000203836	0.0081	0.0102	0.4207	1/1	1260168	-/-	Complete		

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000172014	ENSG00000132498	0.0033	0.0054	1.6602	9/9	26193469	+/-	Complete	440115	Linked Set 29
ENSG00000204788	ENSG00000232866	0.0073	0.0037	0.5025	9/9	26193469	-/+	Complete		
ENSG00000204788	ENSG00000233434	0.0000	0.0029	99.0000	9/9	1340388	-/-	Complete	114813	Linked Set 30
ENSG00000172014	ENSG00000196774	0.0017	0.0048	2.7536	9/9	1340388	+/+	Complete		
ENSG00000232866	ENSG00000233434	0.0091	0.0066	0.7290	9/9	24738227	+/-	Complete	88713	Linked Set 31
ENSG00000132498	ENSG00000196774	0.0015	0.0028	1.8476	9/9	24738227	-/+	Complete		
ENSG00000198307	ENSG00000198082	0.0000	0.0037	99.0000	X/X	487550	+/+	Complete	9515	Linked Set 32
ENSG00000198444	ENSG00000197932	0.0000	0.0000	0.4034	X/X	487550	+/+	Complete		
ENSG00000185978	ENSG00000198082	0.0000	0.0037	99.0000	X/X	565704	-/+	Complete	9515	Linked Set 33
ENSG00000185990	ENSG00000197932	0.0000	0.0000	0.4045	X/X	565704	-/+	Complete		
ENSG00000172283	ENSG00000169763	0.0000	0.0000	0.0010	Y/Y	842578	+/-	Complete	1054777	Linked Set 34
ENSG00000269393	ENSG00000267935	0.0000	0.0000	0.0010	Y/Y	842578	-/+	Complete		
ENSG00000172288	ENSG00000172352	0.0000	0.0000	0.0010	Y/Y	842578	+/-	Complete		
ENSG00000169789	ENSG00000169807	0.0000	0.0000	0.3785	Y/Y	168948	+/-	Complete	283838	Linked Set 35
ENSG00000226941	ENSG00000169800	0.0031	0.0000	0.0010	Y/Y	168948	+/-	Complete		
ENSG00000171928	ENSG00000175106	0.0320	0.0400	0.6924	17/17	2872263	+/-	Partial	195489	Linked Set 36
ENSG00000171931	ENSG00000251537	0.0674	0.1966	0.5520	17/17	2872263	+/-	Complete		
ENSG00000108448	ENSG00000251537	0.0528	0.2444	0.4314	17/17	2872263	+/-	Complete		
ENSG00000249459	ENSG00000187607	0.0184	0.1020	0.6363	17/17	2872263	-/+	Complete		
ENSG00000189375	ENSG00000214946	0.0553	0.0778	0.3605	17/17	2872263	-/+	Complete		
ENSG00000187559	ENSG00000204793	0.0000	0.0022	99.0000	9/9	1562677	+/-	Complete	189814	

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	K_A	K_A/K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSG00000196873	ENSG00000204790	0.0068	0.0034	0.5039	9/9	1562677	+/-	Complete		Linked Set 37
ENSG00000228537	ENSG00000196400	0.0148	0.0000	0.0010	9/9	1562677	-/+	Complete		
ENSG00000196873	ENSG00000136682	0.0164	0.0105	0.6410	9/2	NA	+/+	Complete	149372	Linked Set 38
ENSG00000228537	ENSG00000238091	0.0000	0.0000	0.3029	9/2	NA	-/-	Complete		
ENSG00000187559	ENSG00000184492	0.0216	0.1924	0.5288	9/2	NA	+/+	Complete		
ENSG00000171129	ENSG00000171116	0.0000	0.0000	0.3977	X/X	164854	-/+	Complete	29139	Linked Set 39
ENSG00000123584	ENSG00000166008	0.0000	0.0000	0.4112	X/X	164854	-/+	Complete		
ENSG00000231997	ENSG00000204804	0.0984	0.1482	0.7801	9/9	486881	+/-	Complete	152361	Linked Set 40
ENSG00000237198	ENSG00000204807	0.0001	0.0115	99.0000	9/9	486881	-/+	Complete		
ENSG00000234295	ENSG00000157423	0.0051	0.0021	0.4157	GL000192.1/16	NA	-/-	Partial	357293	Linked Set 41
ENSG00000215642	ENSG00000157423	0.0050	0.0049	0.9702	GL000192.1/16	NA	-/-	Partial		
ENSG00000204382	ENSG00000204379	0.0000	0.0000	0.4144	X/X	8341	+/-	Complete	142156	Linked Set 42
ENSG00000155622	ENSG00000185751	0.0000	0.0000	0.5358	X/X	8341	+/-	Complete		
ENSG00000183461	ENSG00000204382	0.0000	0.0000	0.5037	X/X	179068	+/+	Complete	93872	Linked Set 43
ENSG00000204376	ENSG00000204379	0.0000	0.0000	0.4702	X/X	179068	-/-	Complete		

Figures

Figure 1. - Synonymous changes per synonymous site (K_S) based age distribution of 184 human gene duplicate pairs.

The average K_S value of 0.011 between coding regions of humans and chimpanzees (Chen and Li 2001) is shown for scale, and suggests that a large fraction of human gene duplicates within this data set may have originated since the human-chimpanzee split.

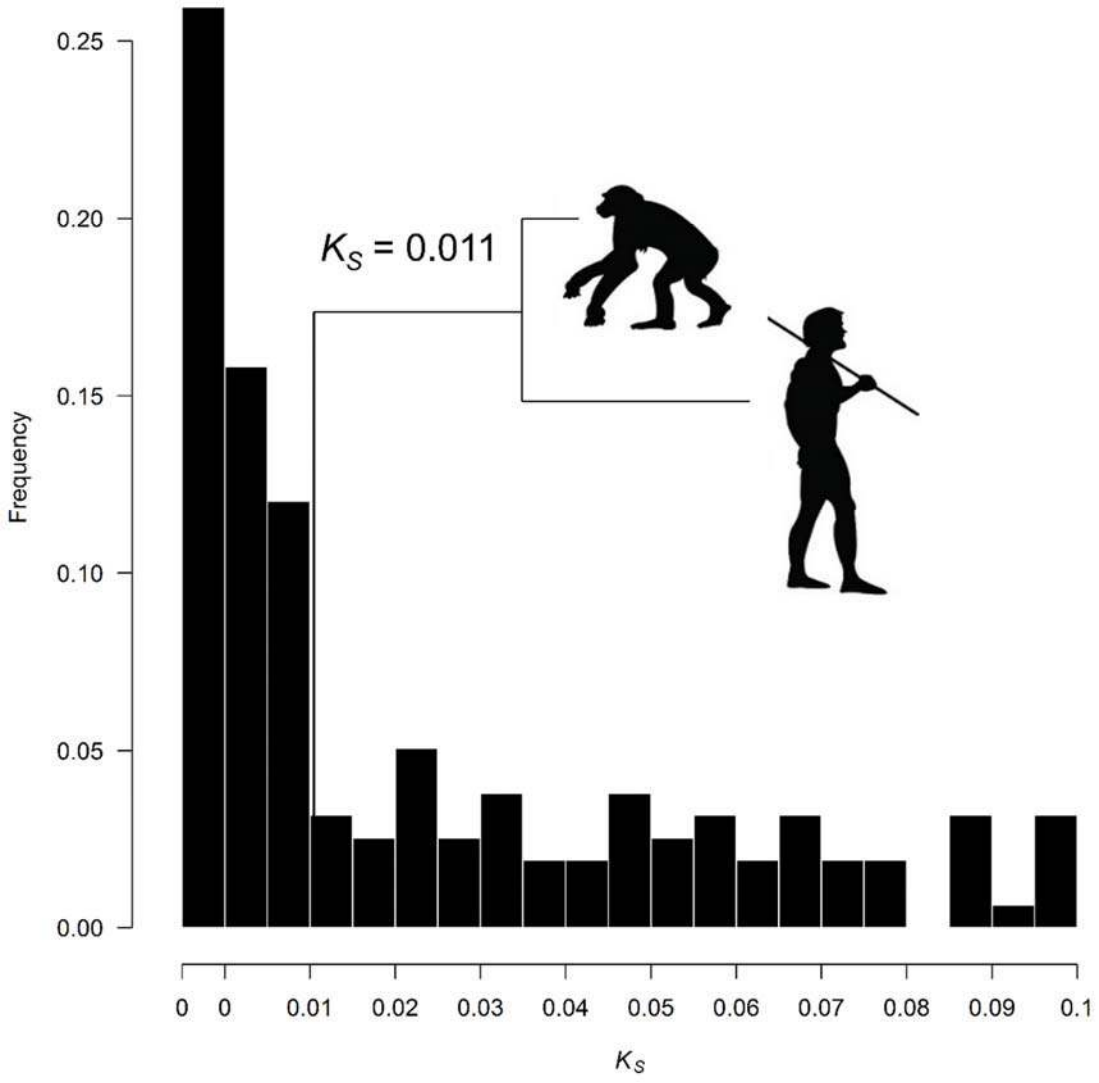


Figure 2. - Composition frequencies of *intra*- and *interchromosomal* duplication within three age-cohorts of human gene duplicate pairs.

The sample sizes of duplicate pairs within each age category ($K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$) are provided above the corresponding bars. The total sample size comprised 172 duplicate pairs with assigned chromosomal locations for both paralogs.

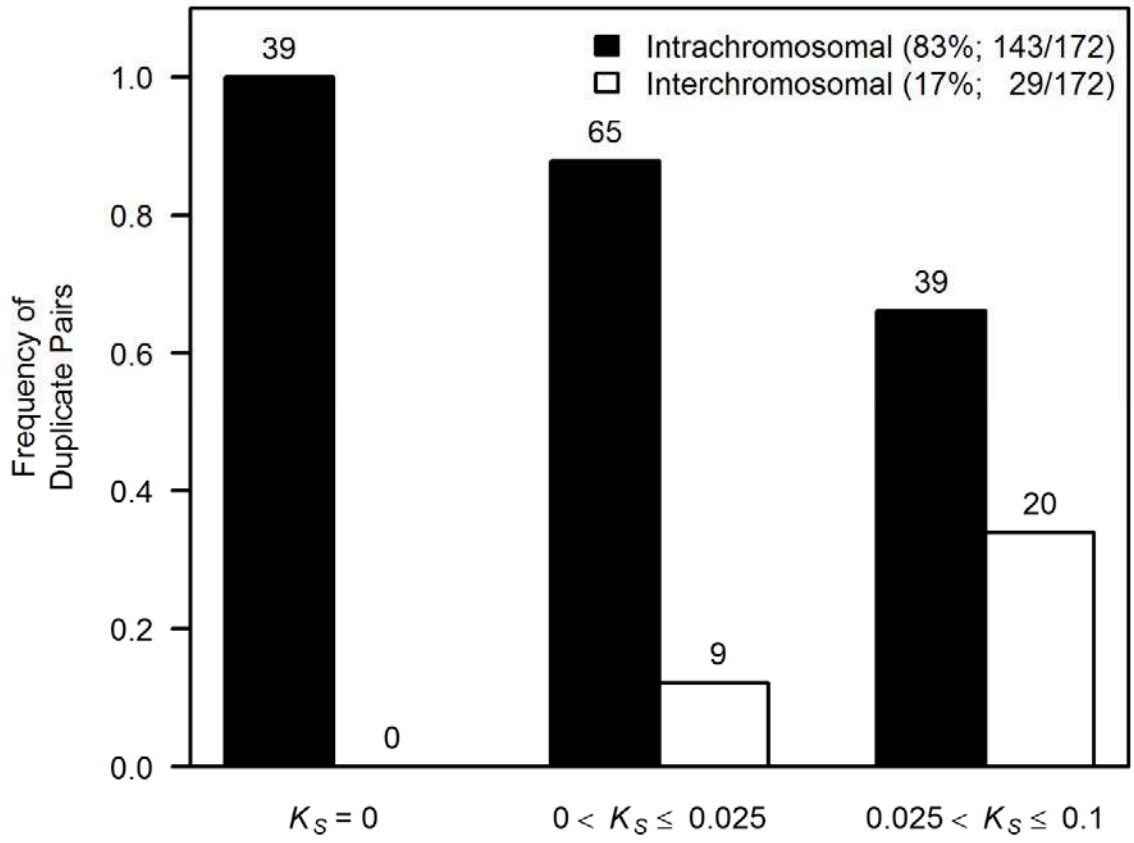


Figure 3. - The physical distance between *intrachromosomal* gene duplicates as a function of K_S.

The regression line represents the relationship between distance separating all *intrachromosomal* paralogs (143 pairs with $K_S \leq 0.1$) and K_S. The correlation between K_S and distance between paralogs is not significant ($r = -0.08$, $df = 141$, $p = 0.84$).

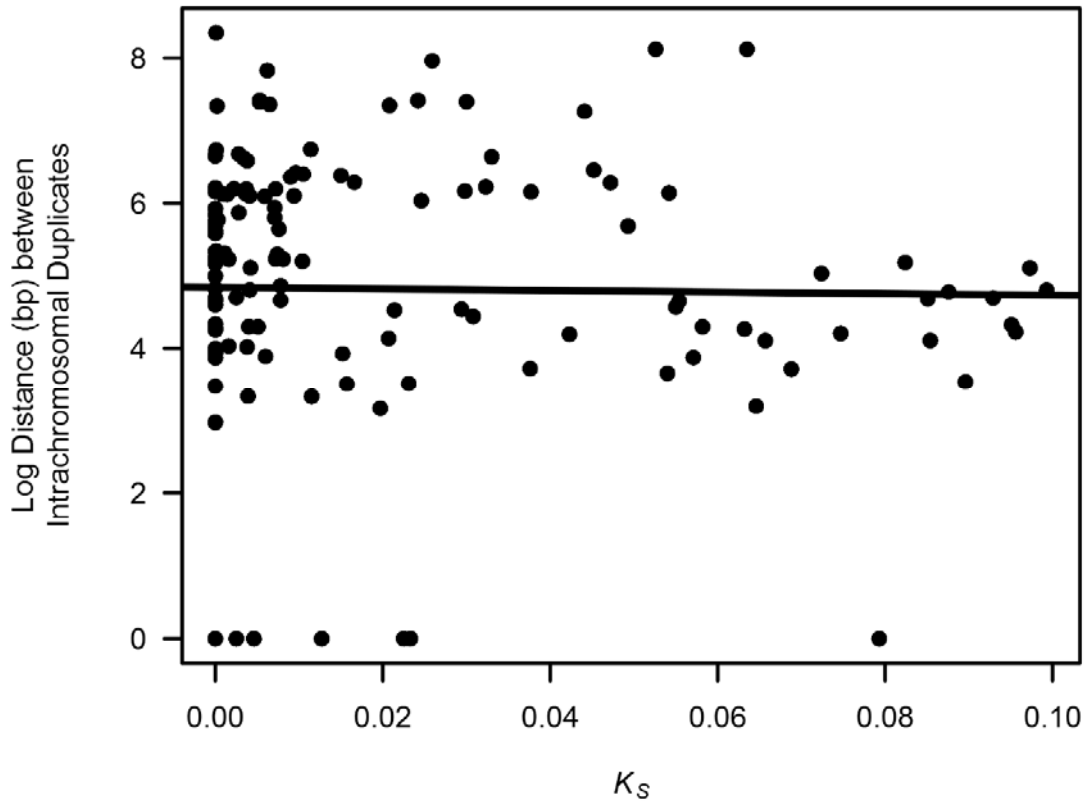


Figure 4. - Nonrandom chromosomal distribution of 172 pairs of young gene duplicates in the human genome.

The height of the blue bars indicates the relative duplication frequencies across 24 chromosomes, calculated as the ratio of the number of duplicate copies on a chromosome and the number of protein-coding genes on the same chromosome. The box plot displays the variation in these relative frequencies across 24 chromosomes, with the median represented by a solid line and the upper and lower quartiles in dotted lines.

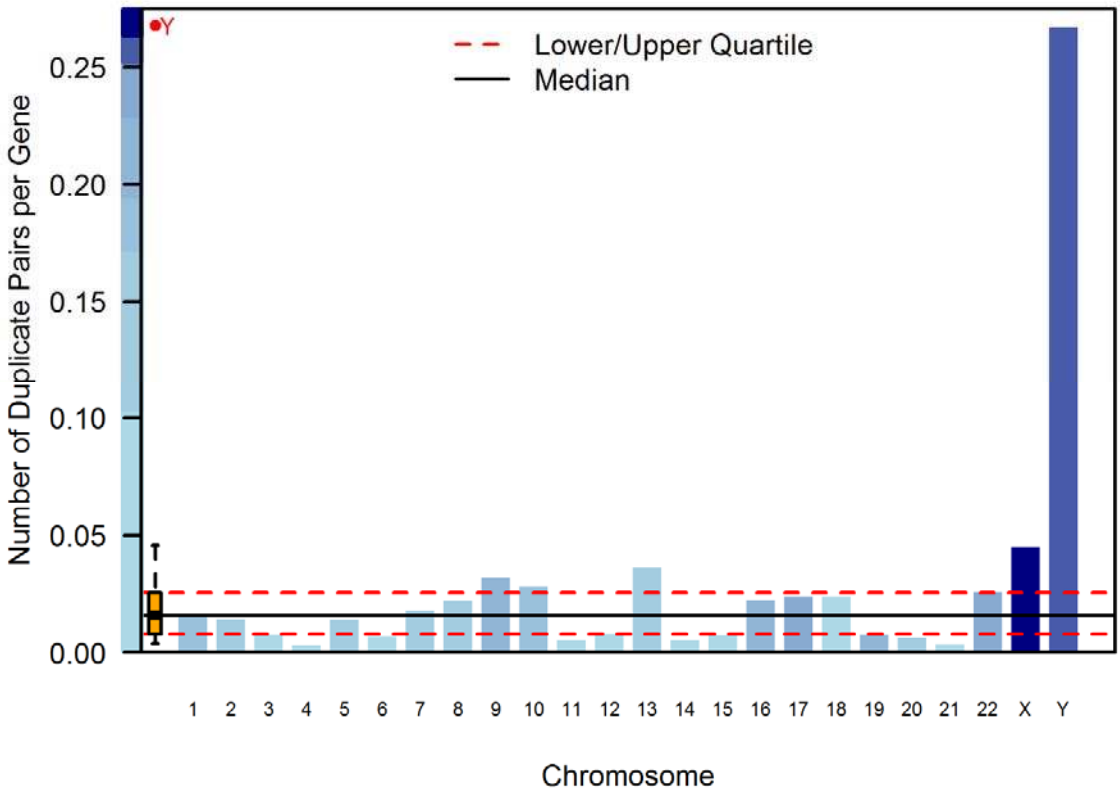


Figure 5. - Location of 172 human gene duplicates relative to the centromere.

The relative location of gene duplicates along chromosomal arms deviates significantly from an expected distribution based on protein-coding gene enrichment. Each chromosome was subdivided into 10 Mb bins representing increasing distance from the centromere. The proportions of gene duplicates and protein-coding genes (N = 20,172) within each bin are represented by black and white bars, respectively.

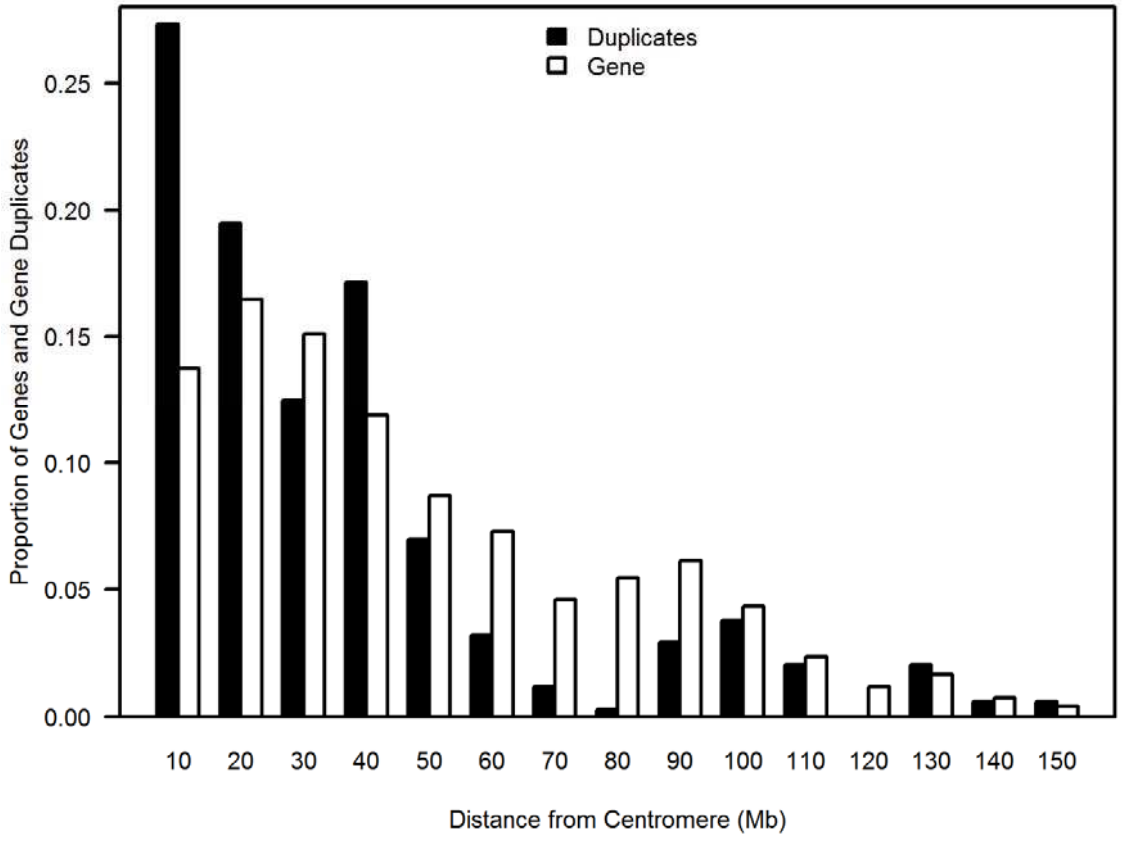


Figure 6. - Composition frequencies of three structural categories of DNA-mediated gene duplicates across three evolutionary age-cohorts.

The sample sizes of duplicate pairs within each of the three categories ($K_S = 0$, $0 < K_S \leq 0.025$, and $0.025 < K_S \leq 0.1$) are provided above the corresponding bars ($N = 163$ gene duplicate pairs).

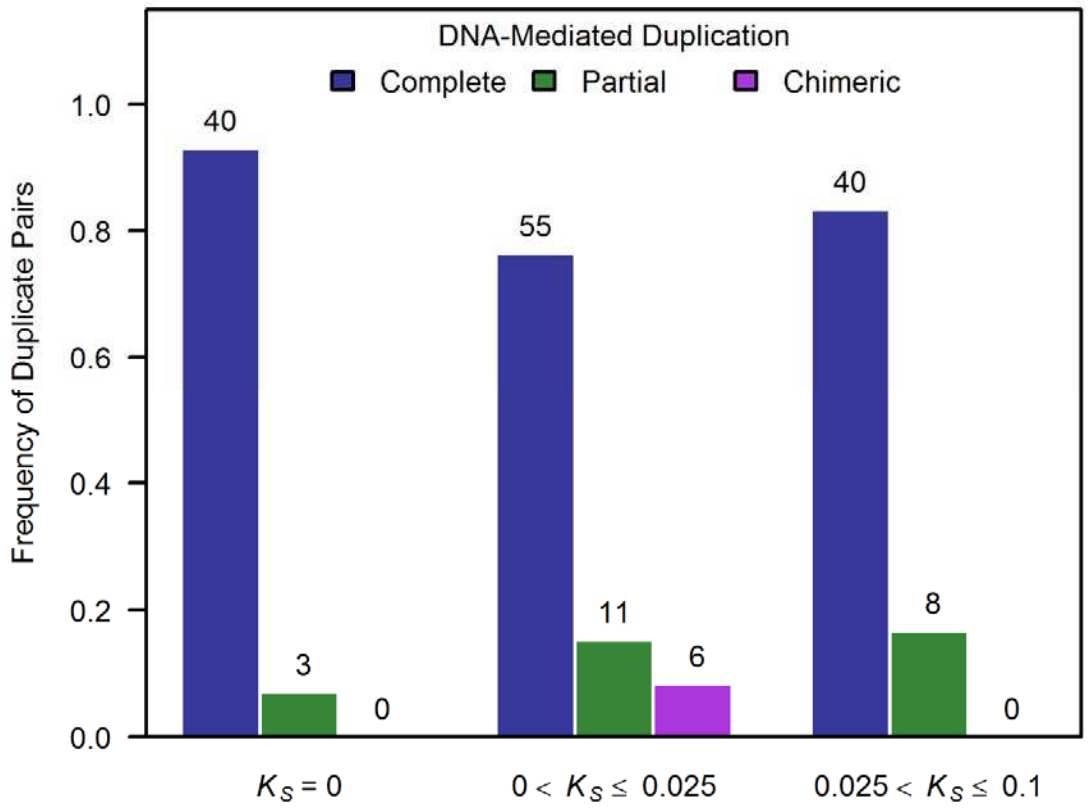


Figure 7. - Box plot displaying the distribution of minimum duplication span for 184 human young gene duplicates.

The range and median length of human protein-coding genes and their coding regions are displayed for comparison.

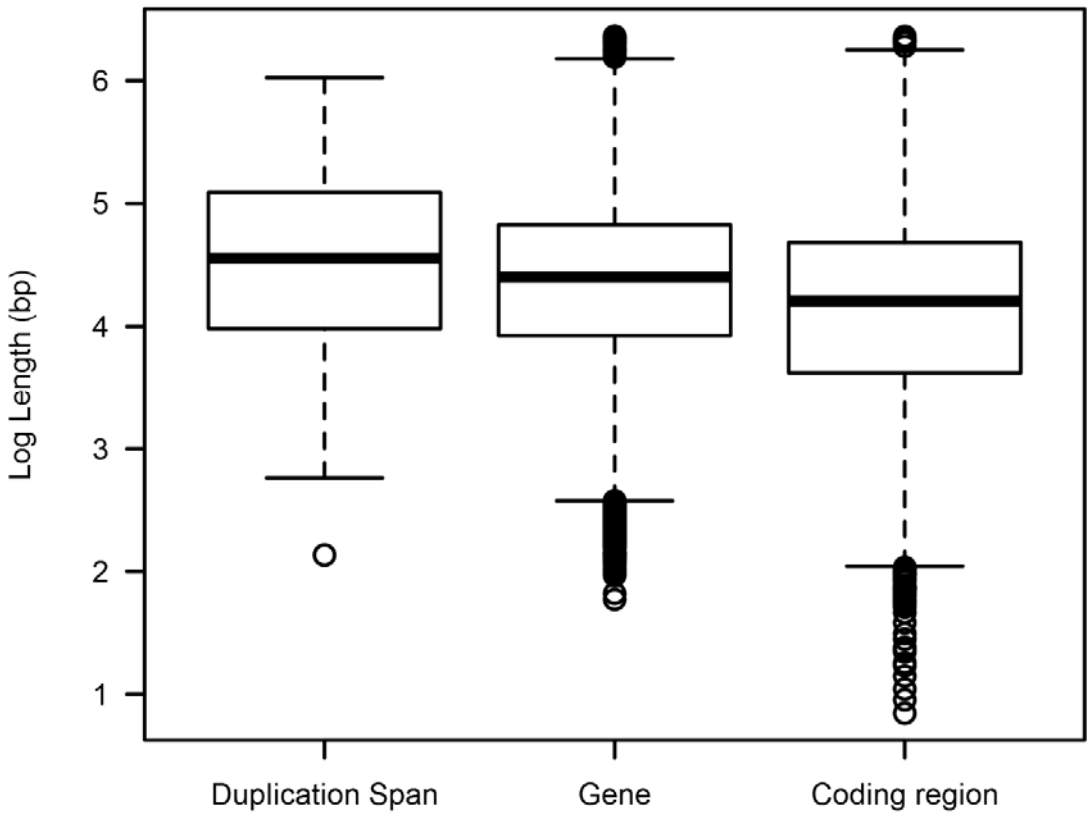


Figure 8. - Duplication span of DNA- and RNA- mediated duplicates as a function of evolutionary age (K_S).

The data set comprises 163 DNA-mediated duplicate pairs (blue) and 21 RNA-mediated duplicate pairs (orange).

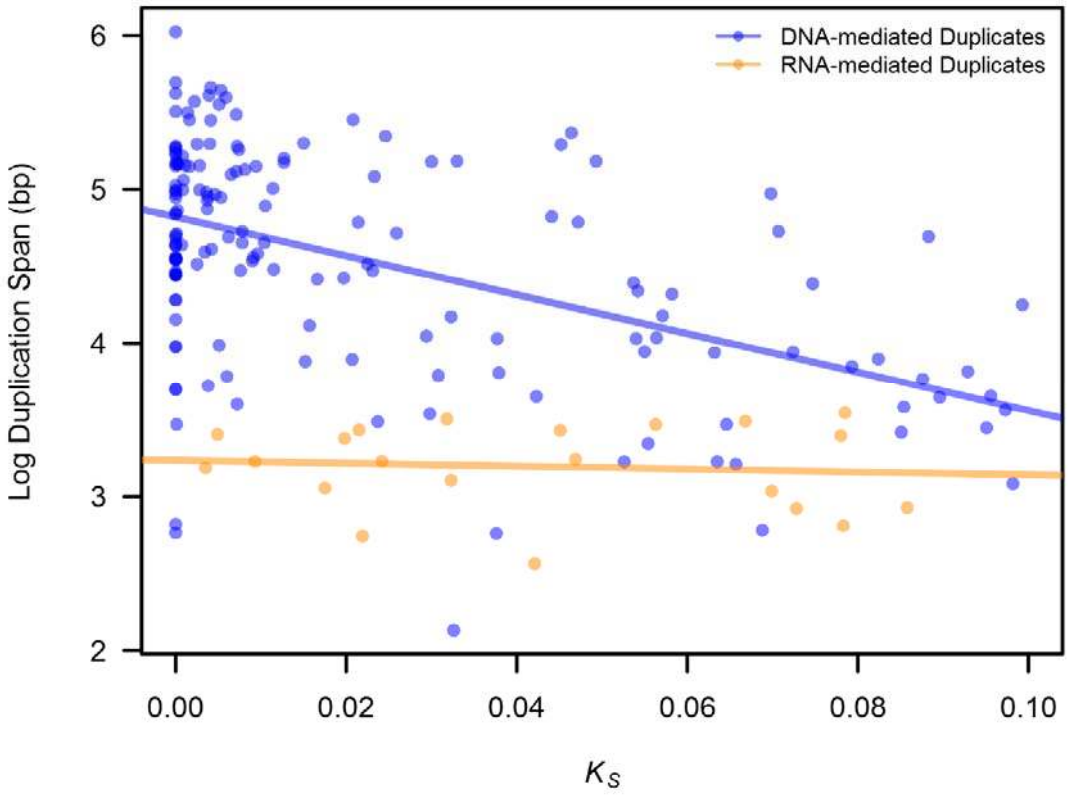
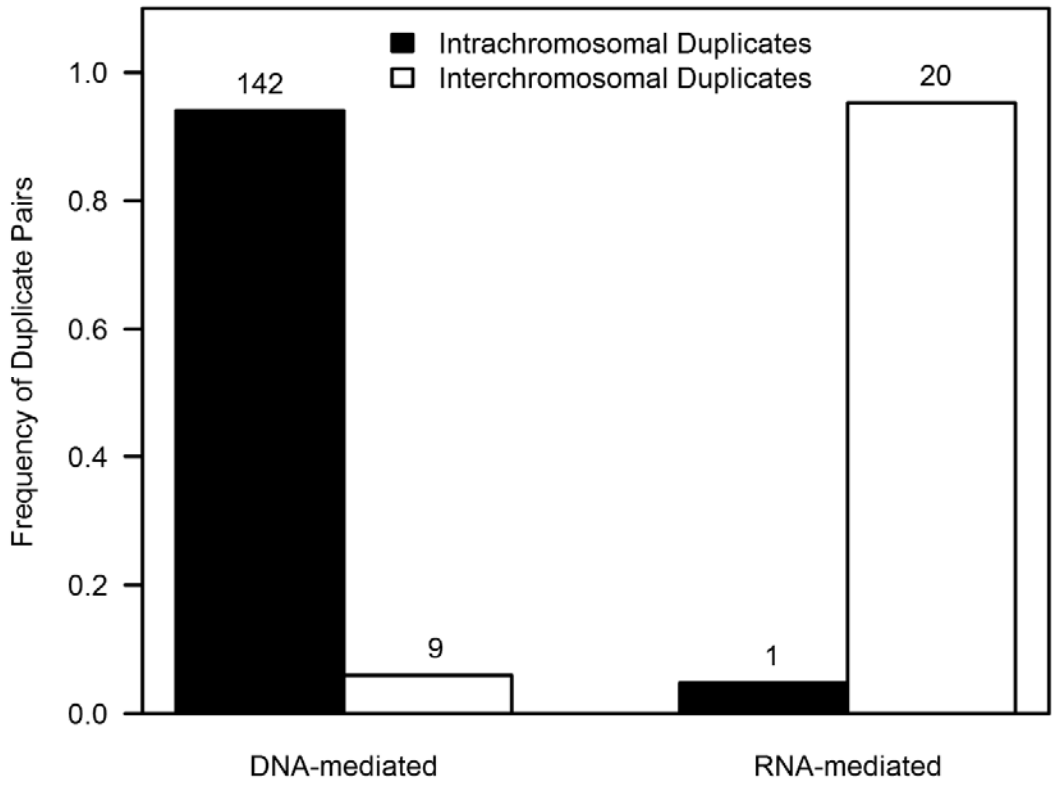


Figure 9. - Composition frequencies of *intra-* versus *interchromosomal* gene duplicates within DNA-mediated and RNA-mediated duplication events.



CHAPTER FOUR

Early Evolutionary Dynamics of Gene Paralogs in the Chimpanzee Genome Reveals a Divergent Duplication Landscape Relative to Humans

Abstract

Studies of segmental duplication and retrogenes have been performed separately to address their important role in primate genome evolution. A strict and systematic evolutionary framework for the population of young gene duplicates has been established and applied to few species, including *C. elegans*, *S. cerevisiae* and human. The unique pattern of duplication structural resemblance types observed in humans indicates a special composition of mechanisms for the origin of duplicates and drastically divergent evolutionary trajectories for the duplicates in humans. It is still unclear whether these features are uniquely human or are common across primates. We identified 181 gene duplicate pairs in small gene families with a synonymous sequence divergence of 10% or less within the chimpanzee draft genome. Active recent gene duplication events were detected in the chimpanzee genome while DNA-mediated and RNA-mediated gene duplicates each account for roughly 60% and 40% of the young gene duplicates in the chimpanzee genome. The abundance of RNA-mediated duplicates results in a large proportion of *interchromosomal* duplicates (97%; 65/67), while the majority of DNA-mediated duplicates (75%; 60/80) are located on the same chromosome. The sex chromosomes and chromosome seven have a significantly higher incidence of gene duplication per gene relative to all other chromosomes. DNA-mediated duplicates were found to have a preferentially enrichment within pericentromeric regions. Although the median DNA-mediated duplication span (11 kb) is not significantly larger than the median coding region (16 kb), the *complete* duplicates (80.9%; 89/110) out number *partial* (18.2%; 20/110) and *chimeric* (0.9%; 1/110) types in the chimpanzee genome.

The systematic analysis for young gene duplicates was applied to chimpanzees – the close relative to human. This analysis revealed informative shared and unique patterns of structural categories for young gene duplicates in the genome of chimpanzee in comparison to the genome of human and other species. And these patterns were connected to their special genomic architecture and trajectory of evolution. Also the results provide a list of candidate genes of functional novelties.

Introduction

Gene duplication has long been recognized as a major player in the functional and structural evolution of genomes (Bridges 1935; Bridges 1936; Ohno 1970). The rate of gene duplication has been empirically estimated to be relatively high, ranging from 10^{-7} to 10^{-3} per gene per generation (Katju and Bergthorsson 2013; Lipinski et al. 2011), which establishes that gene duplication is an important mutational force for introducing novel genetic material into the genome. These paralogous sequences can undergo a number of different fates (Innan and Kondrashov 2010). If there is selection for increased dosage of the ancestral gene product, this can lead to the retention of the redundant duplicate sequence, and prevent deterioration of the sequence via deleterious mutation (Bergthorsson et al. 2007; Sudmant et al. 2013). Alternatively, the gene duplicate may evolve new spatial expression patterns through regulatory subfunctionalization (Gokcumen et al. 2013; Makova and Li 2003). Complementary changes in the functional coding sequences of the two copies via subfunctionalization may lead to selection for the maintenance of both partially functional copies in order to maintain the ancestral gene function (Force et al. 1999). Neofunctionalizing mutations in one copy may facilitate the origin of novel function or shifts in functionality (Kaessmann 2010; Long et al. 2013). The abundance and evolutionary trajectories of duplicates originating from DNA-mediated mechanisms have been investigated in studies focusing on segmental duplication (Marques-Bonet et al. 2009; Samonte and Eichler 2002; She et al. 2006) and copy number variation (CNV) (Hastings et al. 2009; Perry et al. 2008; Sudmant et al. 2013). Although RNA-mediated or *retroposed* duplicates have also been

extensively explored (Kaessmann et al. 2009; Pan and Zhang 2009; Xing et al. 2006; Zhang 2013), few studies (Jun et al. 2008; Jun et al. 2009) have considered both DNA- and RNA-mediated mechanisms together to determine their differential contributions to genome architecture, if any.

In order to shed light on the origin and fate of gene duplicates originating through DNA- and RNA-mediated processes, a detailed examination of the exon-intron structure and structural resemblance between two paralogs is required. While RNA-mediated duplicates are commonly inserted into locations with new expression environments (Vinckenbosch et al. 2006), and have the potential to form new chimeric genes (Courseaux and Nahon 2001; Long et al. 2003; Wang et al. 2006; Zhou et al. 2008), most DNA-mediated duplicates have been assumed to maintain structural and functional resemblance to their ancestor upon duplication (Fisher 1935; Haldane 1933), though this is not necessarily the case. Quite a number of studies suggest that duplications may only cover part of the ancestral open reading frame and generate gene duplicates of structural *heterogeneity*, such as *partial* or *chimeric* duplicates (Katju and Lynch 2003, 2006; Katju et al. 2009; Meisel 2009; Zhou et al. 2008). *Heterogeneous* gene duplicates (*partial*, *chimeric*, and *retrotransposed*) experience higher death rates compared to *complete* duplicates, but also have the potential to be neo- or subfunctionalized upon duplication (Katju 2012), leading to selective pressures driving them towards fixation immediately in the post-duplication period.

Few studies have investigated young gene duplicates by systematically classifying structural resemblance between paralogs as a function of evolutionary age; however, the ones that have been completed have already revealed interesting results. C*omplete* duplicates make up the majority of young gene duplicates in the human genome (83%) (Bu and Katju *in review*), which is similar to *Saccharomyces cerevisiae* (89%) (Katju et al. 2009), but which contrasts with the abundance of *partial*/*chimeric* duplicates in *Caenorhabditis elegans* (61%) (Katju and Lynch 2003), *Drosophila melanogaster* (59%) (Zhou et al. 2008), and *Drosophila pseudoobscura* (56%) (Meisel 2009). It has been proposed that the abundance of certain types of duplicates is determined by whether the median duplication span is large enough to encompass the average coding region of genes (Katju et al. 2009). For example, the median length of young gene duplications in *C. elegans* (1.4 kb) is shorter than its average gene coding region length (2.5 kb) (Katju and Lynch 2003), which results in an abundance of *partial* or *chimeric* duplicates upon duplication. The excess frequency of *complete* duplicates in *S. cerevisiae* and *Homo sapiens* is due to a larger duplication span (2.5 kb and 36 kb) and a shorter gene coding length (1.1 kb and 25 kb), respectively (Katju et al. 2009; Bu and Katju *in review*). Studies have indicated a burst of retroposition in mammals (Pan and Zhang 2009), especially in primates (Marques et al. 2005). A higher proportion of functional RNA-mediated duplicates (11%; 21/184) were found in young gene duplicates of *H. sapiens* compared to *C. elegans* and *S. cerevisiae* (Bu and Katju *in review*).

Pan troglodytes (chimpanzee) and *H. sapiens* diverged roughly five to seven million years ago (Goodman et al. 1998; Langergraber and Prüfer 2012). Dramatic differences within their genomes in terms of repeat elements and the activity of

retrotransposition have been described (The Chimpanzee Sequencing and Analysis Consortium 2005). Currently, numerous segmental duplication (DNA-mediated) and retrogene (RNA-mediated) studies have been separately performed on genomes of the primate lineages. A comparison of human- and chimpanzee-specific patterns of duplicate structural classes and their contribution to the overall genomic architecture, as well as the evolutionary consequences thereof will help demonstrate the importance of structural features of gene duplicates to their future evolution.

Here we identified putative evolutionarily young gene duplicates in the chimpanzee genome by using their degree of synonymous divergence per synonymous site (K_S) as a proxy for evolutionary age (Zuckerkandl and Pauling 1962), and conducted a comparative analysis with their counterparts in the human genome (Bu and Katju *in review*). Our analysis focused on young gene duplicates in small gene families with five or fewer members and an estimated $K_S \leq 0.1$. Ectopic gene conversion has been demonstrated to reduce sequence divergence among paralogs in a large number of organisms (Chen et al. 2007; Deeb et al. 1994; Fawcett and Innan 2013; Iatrou et al. 1984; Innan 2003; Leigh Brown and Ish-Horowicz 1981; Liebhaber et al. 1981; Ollo and Rougeon 1983; Petes and Hill 1988; Rane et al. 2010). Prevalent gene conversion may act to reduce synonymous site divergence (K_S) between paralogs causing an underestimation of the actual evolutionary age of a gene duplicate pair. Therefore, where K_S has been used as an age indicator, duplicate genes that underwent gene conversion will likely appear to be younger than their real evolutionary age.

Both empirical (locus-specific) approaches and bioinformatic analyses of sequenced genomes have provided a wide range (e.g. ~10^{-10} to ~10^{-3} per cell division) of values for locus-specific gene conversion rates (Mansai et al. 2011). Although previous studies have indicated a high frequency (25.4%) of ectopic gene conversion in a subgroup of human SDs with four or more paralogous sequences (Dumont and Eichler 2013; Fawcett and Innan 2013), gene conversion was found to have minimal evolutionary effect on recent, lineage-specific gene duplicates in four mammals (McGrath et al. 2009). The contradictory observations that the gene conversion rate is negatively (Melamed and Kupiec 1992) or positively (Semple and Wolfe 1999) correlated with the gene family size is likely due to different cut-offs for sequence similarity searches and family grouping approaches used by different studies (Semple and Wolfe 1999). Duplicates with detectable gene conversion signal detected by the GENECONV program (Sawyer 1989) were excluded, in order to minimize possible bias caused by gene conversion. To our knowledge, this study is the first attempt to evaluate the relative contribution of *complete*, *partial*, *chimeric* and *retrotransposed* duplicates to the population of young gene duplicates in the *P. troglodytes* genome, as well as its comparison to *H. sapiens*.

Methods

Identification of Chimp Gene Duplicates and their Structural and Genomic Features

The detection of gene duplicate pairs, clustering of gene families, determination of duplicate boundaries, and the classification of structural resemblance of duplicates

followed the same workflow as previously described for the analysis of the human genome (Bu and Katju *in review*). Briefly, the workflow detects young gene duplicates ($K_S \leq 0.01$) within small gene families of five members or less based on a similarity search of protein sequences using the BLASTP algorithm (Altschul et al. 1990). This analysis utilized the chimpanzee genome assembly (CHIMP2.1.4) from Ensembl (version 74) (Flicek et al. 2013). Chimpanzee protein sequences of 18,759 canonical transcripts were included in the BLAST search with a cutoff E-value of $\leq 10^{-10}$ and an amino acid identity of at least 40%. Gene families were clustered using the single linkage principle.

Protein sequences were aligned for each duplicate pair using the CLUSTALW2 program (Larkin et al. 2007). The corresponding nucleotide sequences were aligned based on their protein sequence alignments using PAL2NAL (Suyama et al. 2006). The synonymous sequence divergence of paralogs was estimated using the *codeml* program (runmode = -2, pairwise model) of the PAML package (Yang 2007). Young gene duplicates with a K_S less than or equal to 0.1 were retained for further analyses. Linked sets of genes were treated as single genes while performing the sequence alignment and the K_S calculations, as these sets of genes may have been duplicated together in a single amplification event. Redundant pairs including *same-location* pairs and *shadow* pairs were removed before further analysis. *Same-location* pairs are paralogs with the same genomic location but are annotated as two different genes with different names/identification numbers. In gene families of more than two members, the number of redundant/shadow pairs to be removed increased exponentially with gene family size. For example, gene A duplicated to gene B, and gene B duplicated to gene C. The gene A

and C are *shadow* pairs as there are no true duplication events between gene A and C, although they share sequence similarity. Shadow pairs were removed based on a UPGMA tree generated from pairwise K_S values of the family.

Ectopic Gene Conversion Signal Detection between Paralogous Sequences

The closest human ortholog to each previously identified pair of chimpanzee duplicate genes was identified using the BLASTP algorithm. The coding sequences of the chimp paralogs and their closest human ortholog were aligned using the CLUSTALW2 program (Larkin et al. 2007). Potential gene conversion signals were detected using the statistical test implemented in the GENECONV program (version 1.81a) (Sawyer 1989) with default settings plus pairwise comparison (/lp; list pairwise). Significant gene conversion signals were listed with their *p* values from permutation tests corrected for multiple comparisons.

Detection of Duplication Boundaries and Structural Resemblance Types and Visual Verification

Potential duplication boundaries were identified by locating homologous regions within 200 kb flanking the previously identified gene duplicates (400 kb for few pairs) using the genomic alignment tool LASTZ (Harris 2007). The LASTZ alignments were imported into local GBrowse_syn, the Generic Synteny Browser (McKay et al. 2010) for visualization and manual verification. The duplication break points and the degree of structural resemblance between chimp paralogs (Katju and Lynch 2003) were assigned

based on visualizing the pattern of homology and the exon-intron structure of the two paralogs. Four structural resemblance types for duplicates were defined. *Complete* duplicates retain complete sequence homology for the entirety of the canonical coding region, encompassing at least everything from the start to the stop codon. *Partial* duplicates refer to the duplication of only part of the canonical coding region, with the derived paralog expected to be shorter. A *chimeric* duplicate refers to a duplication where part of the ancestral locus is shared with the derived copy, but the derived copy has includes new coding sequences. Finally, *retroposed* duplicates are defined as those pairs with one multi-exon paralog and a single exon paralog, with no homology in the flanking regions. Structurally, the derived paralog within retroposed duplicate pairs differs from the ancestral locus in the loss of introns and the gain of a poly-A tail.

Frequency Counting and Statistical Tests

To test if the chromosomal distribution of young duplicates matches the expected frequencies based on the number of genes on a particular chromosomes, the number of genes and duplicate events were counted for each chromosome. The duplication rate for each chromosome was calculated as half number of duplicates pairs found on the chromosome divided by the number of protein coding genes on that chromosome. In order to test the distance of the paralogs from the centromere, chromosomes were divided into 10 Mb bins starting from the centromere. The number of duplicates was compared to the number of genes within each bin. In cases where two paralogs were located in different bins, each paralog was counted as half. Statistical tests were performed using

the R program package (version 3.1.2) (R Core Team 2014). Instead of the default chi-square test, a *G*-test (the likelihood ratio test) was employed to test for goodness-of-fit.

Results

In a previous analysis of human gene duplicates (Bu and Katju *in review*), any conclusions regards the evolutionary dynamics and genomic features of young paralogs did not differ based on inclusion or exclusion of gene duplicate pairs with significant gene conversion signals. However, in order to prevent any bias in the final number of types of duplicates, chimpanzee duplicate pairs with signals of gene conversion (9%; 18/199) were excluded (**Table 1**). The same principle was applied previously in the case of human gene duplicates (14%; 26/184) (**Table 1**). All subsequent analyses and comparisons were performed on this filtered set of 181 and 158 chimpanzee and human duplicate pairs, respectively (**Table 1**).

In total, 181 gene duplication events (**Supplemental Table 1**) were identified with the same selection criteria imposed on human duplicates (Bu and Katju *in review*): 1) a synonymous sequence divergence of 10% or less, as longer evolutionary time results in multiple synonymous substitutions on the same site, causing an under-estimation of the evolutionary age of paralogs; 2) a restriction of gene families size to five members or less, as large multigene families may behave differently during evolution; 3) and the exclusion of gene duplicates with gene conversion signals in their coding sequences (18 pairs were filtered out by GENECONV program). Among these 181 pairs (Table 1 and Supplemental Table 1), 81.2% (147/181) which had both copies located on a defined chromosome (**Table 1**). For the remaining 34/181 pairs (18.8%; Table 1), at least one paralogs was located on a scaffold for which the exact chromosome was not known. The 181 chimpanzee gene duplicate pairss, as well as their synonymous divergence values,

chromosomal locations, structural classification, transcriptional orientation, duplication span (bp) and genomic distance are summarized in Supplementary Table 1.

Differences in Gene Duplicate Age Distributions between Humans and Chimpanzees

The synonymous sequence divergence (K_S values) between paralogs was used to represent the age of the duplication events under the assumption that synonymous mutations are neutral with respect to fitness and accumulate in a clock-like fashion. The distribution of putative evolutionarily young gene duplicates in chimpanzee is L-shaped (**Fig. 1**), i.e. the youngest age cohort contained the highest density of duplicates, with the relatively older classes rapidly dropping down to lower densities. Chimpanzee gene duplicates have a lower starting density within the $0 \leq K_S \leq 0.01$ age-cohort (35% 64/181 duplicate pairs) compared to their counterparts in the human genome (53%; 85/158, Table 1) (Bu and Katju *in review*). On average, human-chimp orthologs have a $K_S \approx 0.011$ and this is taken to roughly represent the evolutionary splitting of the human and chimpanzee lineages (Chen and Li 2001). For gene duplicate pairs with $K_S < 0.011$, the chimpanzee genome seems to have a lower birth rate and/or lower death rate of gene duplicates compared to human (age groups with $K_S > 0.01$). These differences lead to decreased concavity (or steeper average slope) in the age distribution of chimpanzee gene duplicates compared to that of human. Additionally, the frequency of chimpanzee gene duplicates decays at a lower rate relative to humans, which contributes to the difference between the human and chimpanzee duplicate age distribution (*G-test* of independence $G = 19.8$, $df = 9$, $p = 0.0193$).

Differences in the Genomic Location of Chimpanzee and Human Paralogs

In order to determine the genome location gene duplicates at birth and any possible changes of their distribution pattern with increased evolutionary age, we considered three features pertaining to the genomic distribution of paralogs: 1) *intra*- vs. *interchromosomal* duplications; 2) the physical distance (in bp) between two copies of *intrachromosomal* duplicates; 3) and the transcriptional orientation of *intrachromosomal* paralogs.

In terms of chromosomal location, 34% (62/181) and 66% (119/181) of chimpanzee paralogs are *intrachromosomal* and *interchromosomal* duplicates, respectively. The proportion of *interchromosomal* duplicates in chimpanzee is significantly higher than in human (23%; 37/158) ($G = 62.9$, $df = 1$, $p = 2.22e\text{-}15$). Duplicate pairs with one or both copies on a scaffold with an unknown chromosomal location were automatically classified as *interchromosomal* duplications. To remove any potential bias caused by duplications on scaffolds, we removed 34 duplication pairs with unassigned chromosomal location in the chimpanzee dataset (10 pairs in the human dataset) and repeated the test. Excluding the duplications located on scaffolds, the proportion of *interchromosomal* duplications remained significantly higher in the chimpanzee genome than in the human genome: 58% (85/147) in chimpanzee compared to 18% (27/148) in human ($G = 50.6$, $df = 1$, $p = 1.1e\text{-}12$) (**Table 1**). Gene duplicates originating from RNA-mediated duplication events are expected to randomly relocate to any of the 25 chromosomes regardless of the chromosomal location of the ancestral copy. Therefore, duplicates arising from retrotransposition are expected to have a 24:1 ratio of

inter- vs. *intrachromosomal* locations for chimpanzee (23:1 in the human data set). Therefore, it was necessary to further distinguish the genomic locations of duplicates arising from DNA-mediated vs. RNA-mediated events.

The chimpanzee data suggests that RNA-mediated duplicates account for 46% (67/147) of all gene duplicates within this genome, which is significantly higher than the 13% (19/148) observed in human ($G = 39.7$, $df = 1$, $p = 2.924e\text{-}10$). With respect to DNA-mediated duplicates only, the chimpanzee genome possesses a larger fraction of *interchromosomal* duplicates (25%; 20/80) compared to humans (7.0%; 9/129, **Table 1**) ($G = 12.8$, $df = 1$, $p = 0.0003$). Frequencies of *intra*- and *interchromosomal* duplicates within 10 age cohorts (**Fig. 2**) imply that the proportion of DNA-mediated duplicates and *intrachromosomal* duplicates did not change significantly as a function of evolutionary age in the chimpanzee genome. In chimpanzee-human comparisons, only the youngest age cohort ($0 \leq Ks \leq 0.01$) showed a significant difference in: (i) the proportion of *intrachromosomal* vs. *interchromosomal* gene duplicates ($G = 54.0$, $df = 1$, $p = 1.98e\text{-}13$), (ii) the proportion of RNA-mediated vs. DNA-mediated duplicates ($G = 45.1$, $df = 1$, $p = 1.84e\text{-}11$) (data not shown), and (iii) DNA-mediated *intrachromosomal* vs. *interchromosomal* gene duplicates ($G = 11.0$, $df = 1$, $p = 9e\text{-}4$).

We further compared the physical distance and orientation (*direct* vs. *inverse*) of 62 pairs of chimpanzee *intrachromosomal* duplicates, relative to 121 pairs in human. Log distances for *intrachromosomal* duplicates in the chimpanzee and human genome were plotted against their corresponding Ks values (**Fig. 3**). There was no significant

relationship between *intrachromosomal* paralog genomic distance and evolutionary age (chimpanzee: $r = -0.01$, $df = 60$, $p = 0.57$; human: $r = -0.007$, $df = 119$, $p = 0.64$).

The transcriptional orientation of *intrachromosomal* paralogs shows an equal proportion of *direct* (same orientation in paralogs) and *inverse* (opposing orientation of paralogs) for both genomes. The proportion of duplicates with *direct* transcriptional orientation is 50% (31/62) for chimpanzee compared to 44.6% (54/121) for human (**Table 1**). An inter-age cohort comparison (five equal cohorts for K_S from 0 to 0.1) showed no significant difference in the proportions of *direct* vs. *inverse* *intrachromosomal* duplicates with increasing evolutionary age ($G = 3.97$, $df = 4$, $p = 0.4095$).

In the human dataset, *directly* oriented duplicates are separated by shorter genomic distances (Wilcoxon rank sum test, $W = 1203$, $p = 0.0016$) and have shorter duplication spans ($W = 1348.5$, $p = 0.0165$) relative to duplicates in *inverse* orientation (**Supplementary Figure 1**). While the chimpanzee dataset shows similar trends with respect to genomic distance ($W = 324$, $p = 0.0281$), there was no significant difference in the duplication spans of *direct* and *inverse* duplicates in chimpanzees ($W = 401$, $p = 0.2681$).

Chromosomal Distribution of Gene Duplicates

To directly investigate the relationship between the number of genes per chromosome on gene duplication rate, we calculated the duplication rate for each

chromosome using "half" of the number of duplications (each duplication event counts as 1, so each copy stemming from the duplication event was counted as 0.5) divided by the number of protein coding gene on that chromosome (**Fig. 4**). Chromosomes Y, X and 7, in that order, were identified as outliers using Grubbs test ($G.Y = 4.2054$, $U = 0.2324$, $p =$ 1.2e-07; $G.X = 2.7387$, $U = 0.6597$, $p = 0.0332$; $G.7 = 3.0933$, $U = 0.5453$, $p = 0.0048$), as they have higher duplication rates relative to all the other chromosomes.

Using 147 pairs of chimpanzee gene duplicates of known chromosomal location, the search for biased duplication frequency toward the centromeres returned no significant result ($G = 8.5551$, $df = 10$, $p = 0.5748$). Upon excluding the products of RNA-mediated duplication events (**Fig. 5**), it was found that DNA-mediated duplicates are enriched within a 20 Mb region of the centromere ($G = 26.8855$, $df = 10$, $p = 0.0027$).

Structural Features of DNA-Mediated Duplicates in Chimpanzee Relative to Human

Differences in structural resemblance between duplicate pairs may dictate different evolutionary trajectories for gene paralogs. In the chimpanzee genome, the 110 gene duplicate pairs arising from DNA-mediated duplication events were dominated by *complete* duplications, as was the case in the human genome. The relative frequencies of *complete* (80.9%; 89/110), *partial* (18.2%; 20/110) and *chimeric* (0.9%; 1/110) gene duplicates within chimpanzee showed no significant deviation from the composition of similar gene duplicates in humans ($G = 4.3667$, $df = 2$, $p = 0.1127$) (**Table 1**). When we further classified the chimpanzee gene duplicates into 10 age cohorts, *complete* duplicates

are the most frequent structural category within each age cohort with no significant change in proportions across different age cohorts (**Fig. 6**).

Duplication Span in the Chimpanzee Genome

The duplication span is defined as the length of the homologous tract shared between a pair of duplicated genes. The range of duplication spans for 181 chimpanzee gene duplicates pairs was 212 bp - 454 kb (**Fig. 7**), with a median span of 3.4 kb which is significantly smaller than the median protein-coding gene length of 16 kb (Wilcoxon rank sum test two-tailed $W = 1078328$, $p < 2.2e\text{-}16$). However, upon excluding 71 retrotransposed gene duplicates, the median duplication span for DNA-mediated duplication events (range 389 - 454 kb; median duplication span of 11 kb) is not significantly different from the median protein-coding gene length in the chimpanzee genome ($W = 939436.5$, $p = 0.1051$).

The duplication span of both DNA- and RNA-mediated duplicates showed no significant correlation with their evolutionary ages K_S (**Fig. 8**. Kendall's rank correlation test two-sided: $tau = -0.0228$, $p = 0.7273$; $tau = 0.154$, $p = 0.0579$). The results remained nonsignificant when we excluded duplicate pairs having one copy on a scaffold with no chromosome assigned ($tau = -0.0972$, $p = 0.2050$; $tau = 0.147$, $p = 0.0795$). The duplication spans of the youngest age cohort ($0 \leq K_S \leq 0.01$) in the chimpanzee dataset are significantly shorter than their counterparts in human ($W = 4666$, $p = 8.6e\text{-}14$) (**Fig. 8**).

Higher Frequency of RNA-Mediated Duplications in the Chimpanzee Genome

A higher proportion of RNA-mediated duplicates are found in the chimpanzee genome (46%; 67/147), compared to the human genome (13%; 19/148) (**Fig. 9**) (**Table 1**). RNA-mediated duplication events have the largest contribution to the formation of *interchromosomal* duplicates in the chimpanzee genome. RNA-mediated duplicates appear to be evenly distributed across chromosomes (**Fig. 9**), except for the barren Y chromosome.

Discussion

Elucidating the early evolutionary features of gene duplicates can facilitate an understanding of the mutational mechanisms underlying their origin and the subsequent evolutionary forces that dictate their trajectory after birth. A comparative analysis of gene duplicates in the genomes of closely-related and diverse species can further determine if mutational mechanisms of duplicate origin and the evolutionary forces governing their spread/loss are shared across certain taxa/lineages or are species-specific. In this study, we analyzed putative evolutionarily young gene duplicates in the chimpanzee genome using approaches previously used for analysis of human gene duplicates. We restricted our analysis to chimpanzee gene duplicates belong to small gene families consisting of five or less members and a $Ks \leq 0.1$, and identified 181 relevant chimpanzee gene duplicate pairs. To eliminate any possible biases introduced by the inclusion of evolutionary older gene duplicate pairs which have been homogenized by gene conversion events, we excluded duplicate pairs with detectable gene conversion signals based on results generated by the GENECONV program (see Methods). We analyzed a number of genomic and structural features for the chimpanzee paralogs and compared these patterns to that of duplicates previously identified in the human genome, as well as previous analyses conducted on *C. elegans* and yeast paralogs (Katju and Lynch 2003; Katju et al. 2009).

Comparative genomic studies of gene duplicates in multiple model organisms, including humans, have revealed an L-shaped age distribution of gene duplicates which suggests a high birth rate and death rate for gene duplicates (Lynch and Conery 2000).

Although, young gene duplicates in chimpanzee have the same L-shaped distribution as humans (**Fig. 1**), there are two obvious differences: i) the chimpanzee genome has fewer gene duplicate pairs in the youngest age cohort compared to human, and ii) the rate of gene duplicate loss in the chimpanzee genome is less extreme compared to human. On average, the synonymous divergence between chimpanzee and human orthologs is estimated to be 0.011 (Chen and Li 2001). The smaller number of chimpanzee duplicates in the youngest age cohort may be correlated with fewer protein-coding genes in this species (18,759 compared to 22,691 genes in humans), but this could be a cause or an effect of the duplication process.

Few studies have compared the relative contributions of DNA-mediated versus RNA-mediated events in the formation of gene duplicates and suggest that unequal crossover events have a larger contribution than retrotransposition in the formation of evolutionarily young gene duplicates in the genomes of human and mouse (Pan and Zhang 2007; Bu and Katju *in review*). In chimpanzee, the frequency of DNA-mediated gene duplicates is higher than RNA-mediated ones (**Table 1**). However, both the absolute number of DNA-mediated duplicates and the ratio of DNA-mediated to RNA-mediated duplicates in the chimpanzee genome are lower compared to the human genome (**Table 1**). This indicates a larger contribution of retrotransposition in the formation of gene duplicates in the chimpanzee genome relative to human.

DNA-mediated duplications or segmental duplication (SD) commonly originate due to two molecular mechanisms of double-strand break repair: non-allelic homologous

recombination (NAHR) (Stankiewicz and Lupski 2002) and non-homologous end joining (NHEJ) (Gu et al. 2008; Lieber et al. 2003). In genome-wide studies of structural variation (SV) which includes SD, deletion and translocations, NHEJ is thought to be a major mechanism in the creation of structural variants (Korbel et al. 2007), or at least for SDs in subtelomeric regions (Linardopoulou et al. 2005). Although copy number variation (CNV, a major component of SV) and SDs are found to have associated genomic locations (Korbel et al. 2007), SVs do contain deletions and translocations in addition to SDs, which may rely on NHEJ/NAHR to different degrees. Studies focused on SDs found enrichment of *Alu* elements (the major type of short interspersed nucleotide elements SINE) on/near the duplication breakpoints (Babcock et al. 2003; Bailey et al. 2003). Also, the SDs formed by *Alu*-*Alu*-mediated recombination events together with other repetitive sequences can serve as hot-spots for further rounds of duplication by NAHR (Bailey et al. 2003; McVean 2010).

The divergent composition of DNA- vs. RNA-mediated young gene duplicates in humans and chimpanzees may reflect the divergent composition of the two genomes. Two non-long terminal repeat (LTR) families: *Alu* elements and L1 (long interspersed nucleotide elements LINE-1) represent about 30% of the human genome (Lander et al. 2001). Initial comparisons of the human and chimpanzee genomes suggested that the human genome has three times more lineage-specific insertion of *Alu* elements (7,082 to 2,340) and a slightly higher number of microsatellites (11,101 to 7,054) than does the chimpanzee genome (The Chimpanzee Sequencing and Analysis Consortium 2005). It is plausible that these extra homologous sequences provided additional recombination hot-

spots for NAHR in the human genome (Gu et al. 2008). The presence of a large number of human-specific SDs compared to chimpanzee-specific ones may corroborate this hypothesis (Cheng et al. 2005).

Additionally, the differential composition of *Alu* and L1 elements between humans and chimpanzees may also impact the origin of RNA-mediated duplicates. The movement of both *Alu* elements (Dewannieux et al. 2003) and retroposed genes (Esnault et al. 2000) relies mainly on the activity of L1 elements. In humans, an estimated 80-100 copies of activating L1 elements (Brouha et al. 2003) of a total 500,000 copies (Lander et al. 2001) have enabled the spread of *Alu* elements to up to 1,000,000 copies (Lander et al. 2001) in the past 65 million years (Deininger and Daniels 1986). The genome-wide LINE-1 amplification rate was found to be significantly greater in chimpanzees than in humans (Mathews et al. 2003). The larger amount of active L1s in chimpanzee may have provided more opportunities for mRNA to be transposed, and hence, to generate retroposed gene duplicates in this species. Additionally, the transposition of *Alu* elements, L1 itself and other normal gene coding mRNAs (candidate retroposed duplicates) all rely on the L1 reverse transcriptase (Dewannieux et al. 2003; Esnault et al. 2000). Based on this "substrate-enzyme" correlation between poly-A-tailed molecules and L1 reverse transcriptase, one can expect that there is competition among the transcript molecules of *Alu* elements and other normal genes. Given the large number of *Alu* elements in the human genome compared to the chimpanzee genome, there may fewer opportunities for the retrotransposition of common gene coding mRNAs via activating L1 elements. Thus, more *Alu* elements compete with coding mRNAs in the

human genome for L1-mediated translocation but may lead to more DNA-mediated gene duplications through NAHR. In contrast, the presence of fewer *Alu* elements in the chimpanzee genome enable retrotransposition by L1 elements. Although we lack the knowledge of the dynamics of L1-mediated transposition and the possibility of other retrotransposition mechanisms (Dewannieux and Heidmann 2005; Mandal et al. 2013), this *Alu* competition hypothesis is attractive. It fits the maximum parsimony principle in that it uses the minimum number of elements to explain both the dominance of DNA-mediated young gene duplicates in humans and the high proportion of RNA-mediated ones in chimpanzee. It would be interesting to test this hypothesis by performing a similar comparative study on the genome of the orang-utan, which has a different composition of *Alu* and L1 elements: the *Alu*/L1 ratios of human, chimpanzee and orang-utan genome are 5000/1800, 2300/2000 and 250/5000 (Locke et al. 2011). A prediction from this hypothesis is that orang-utan would contain the highest proportion of *retroposed* gene duplicates among these three closely related species.

The chimpanzee genome contains a larger amount of RNA-mediated duplicates with an even distribution across chromosomes, which indicates a high birth rate and survival rate of retroposed duplicates. As discussed in a preceding section, the hypothesis for the high birth rate of retroposed duplicates takes into account the (i) relative high activity of L1 elements and (ii) the presence of fewer copies of *Alu* elements for competition. Interestingly, the proportion of RNA-mediated duplicates in the chimpanzee genome does not change with increasing evolutionary age, thereby suggesting high rates of survivorship during their early evolution. In contrast to paralogs

originating from DNA-mediated duplication event, retroposed duplicates lack their ancestral regulatory element. The stringent requirement of inheriting a functional regulatory element in their new genomic location represents a challenge for the survivorship of retroposed gene duplicates. A study of retroposed gene duplicates in the human genome has previously suggested that transcribable retrocopies tend to be surrounded by higher active transcription environments than silent retrocopies. This in turn implies that retrocopies likely rely on the regulatory elements of neighboring genes or insertion into actively transcribed chromatin region for increasing their odds of survivorship (Vinckenbosch et al. 2006).

With respect to DNA-mediated gene duplicates residing on the same or different chromosomes in the chimpanzee genome, *intrachromosomal* duplicates outnumber *interchromosomal* duplicates, suggesting an important role of recombination or exchange of genetic material between homologous chromosomes in the origination of gene duplicates. However, the proportion of *intrachromosomal* duplicates in the chimpanzee genome remains significantly lower than that in the human genome. The pattern can be detected within each age cohort, with the highest abundance of *intrachromosomal* duplicates in the youngest age group (**Fig. 2**). This may suggest a relatively small contribution of NAHR in the formation of chimpanzee DNA-mediated gene duplicates, which could again be due to the comparatively smaller number of *Alu* elements in the chimpanzee genome compared to human (The Chimpanzee Sequencing and Analysis Consortium 2005).

The transcriptional orientation and inter-paralog distance of the *intrachromosomal* duplicates does not appear to have a large impact on the distribution of the duplicates in either the chimpanzee genome, or the human genome. The majority of *intrachromosomal* duplicates in *C. elegans* within the $K_S = 0$ age-cohort were found to occur in inverted orientation (Katju and Lynch 2003). In contrast, the proportions of paralogs in *direct* and *inverse* transcript orientation are roughly equal in both chimpanzees and humans, and no significant differences were found across different age cohorts of gene duplicates within these genomes. The data suggests that paralogs with *direct* and *inverse* orientation have equal probabilities of survivorship in the chimpanzee and human genomes.

Interestingly, human *direct intrachromosomal* duplicates are shorter and closer to each other than are the *inverse* ones, while in chimpanzees they are only observed to be closer but not shorter. A similar pattern has been noticed in the study focusing on *intrachromosomal* repeats in other eukaryotic genomes (Achaz et al. 2001). The pattern suggests unique divergence signatures, which may result from either the same mechanism acting on homologous and non-homologous chromosomes, or different duplication mechanisms producing *intra*- and *interchromosomal* duplicates.

No significant correlation was found between K_S and the distance for *intrachromosomal* duplicates in chimpanzees, a pattern similar to that in the human dataset (Bu and Katju *in review*). This pattern can be explained by two alternative hypotheses, namely (i) extremely limited occurrence of secondary rearrangements

leading to increase in genomic distance between *intrachromosomal* paralogs, or (ii) equal probabilities of survivorship of *intrachromosomal* paralogs irrespective of whether they are closely or distantly located on the same chromosome.

The chromosomal distribution of young gene duplicates may help determine the presence and locations of duplication hotspots within and between chromosomes. An abundance of young gene duplicates on the Y chromosome in both the chimpanzee and human genomes may be due to the presence of large palindromes (Skaletsky et al. 2003) and a relatively low gene density environment on this sex chromosome. After normalizing for the gene abundance on different chromosomes, the distribution of young gene duplicates on autosomes is not significantly different from random. Chromosomal 7, as well as the sex chromosomes seem to have an increased abundance of gene duplicates in chimpanzee (**Fig. 4**). These chromosomes (7, X and Y) may either have a higher birth rate of gene duplicates, or a higher retention rate. Pericentromeric regions may serve as duplication hotspots given an associated enrichment of DNA-mediated gene duplicates in these genomic locations. An abundance of segmental duplications and/or copy number variation within the pericentromeric regions had previously been observed in several eukaryotic genomes including human (Bailey et al. 2001, 2002b; Cheung et al. 2003; Fortna et al. 2004; Zhang et al. 2005), rat (Guryev et al. 2008), and *D. melanogaster* (Emerson et al. 2008).

The degree of structural resemblance between the ancestral and derived gene copy likely affects a duplicate's future evolutionary trajectory toward evolving functional

novelty or pseudogenization (Katju 2012). To evolve novel or shifts in function, *complete* duplicates are dependent on the accumulation of mutational events (single nucleotide or rare exon shuffling events) in the post-duplication period. *Partial*, *chimeric, and retroposed* duplicates, however, have higher probabilities of experiencing radical changes in their exon-intron structure relative to the ancestral copy due to the duplication process and extent of duplication span. Studies have shown that novel gene functions can be derived from structurally heterogeneous duplicates (Charrier et al. 2012; Courseaux and Nahon 2001; Dennis et al. 2012; Marques et al. 2005; Wang et al. 2006). Although the radical changes may bring a high death rate to these "*incomplete*" duplicates, it may take a shorter time for them to gain novel function if they can escape a fate of silencing. Although we found fewer DNA-mediated gene duplicates in the chimpanzee genome compared to the human genome, the proportion of *complete* duplicates is not significantly different between the two primates. This is interesting in the case of the chimpanzee genome, wherein the median duplication span of 11 kb is not significantly larger than the median length of a protein-coding region (16 kb) which in turn increases the probability of formation of *incomplete* gene duplicates. It is possible that *incomplete* gene duplicates (*partial*/*chimeric*) arise at high frequencies in the chimpanzee genome, but are rapidly eradicated from the genome via purifying selection if they bear a fitness cost to the carrier, eventually leading to a higher frequency of *complete* duplicates. However, *complete* duplicates represent the most abundant structural type in all age cohorts within both the chimpanzee and human genomes. Previous observation from macaques, orang-utans and chimpanzees have indicated that the ratio of fixed *complete*/*partial* gene duplicates (ones with at least one ortholog in each

of the primate genome) significantly increases with increasing evolutionary age (Gokcumen et al. 2013). However, the proportion of *complete* duplicates shows no significant change across age cohorts in our dataset of young gene duplicates, providing little evidence for the notion that *partial*/*chimeric* duplicates are being selected against.

Some signatures of functional novelty have been detected in *complete* duplicates within the human genome. For example, young paralogs experience rapid amino-acid substitution under relaxed selective constraints (Zhang et al. 2003), develop divergent special expression patterns (Gokcumen et al. 2013; Makova and Li 2003), and can quickly gain coexpressed partners (Chung et al. 2006). Doubts have been raised whether exon-intron structural changes (resulting in *partial*/*chimeric* duplicates) or the divergence of regulatory factors is a greater contributing factor more to the evolution of novel gene function (Bu and Katju *in review*). Among DNA-mediated duplicates, *complete* gene duplicates are the most abundant structural class in both chimpanzee and human genomes. However, DNA-mediated duplicates only account for 54% of the young gene duplicates in the chimpanzee genome, which contains a larger number of retained RNA-mediated duplicates. Therefore, the extent of functional novelty originating from DNA-mediated versus RNA-mediated duplications remains to be determined.

Tables

Table 1. Frequencies of gene duplicates included in different analyses. The total number of paralog pairs in each analysis is highlighted with bold font.

Duplicates Category	Human	Chimpanzee
Initially Identified Duplicate Pairs	**184**	**199**
Used in Current Study	86% (158/184)	91% (181/199)
Gene Conversion Detected	14% (26/184)	9% (18/199)
Duplicate Pairs Included	**158**	**181**
Located on Known Chromosome	93.7% (148/158)	81.2% (147/181)
Located on Scaffolds	6.3% (10/158)	18.8% (34/181)
Number of Duplicates (incl. Scaffolds)	**158**	**181**
Intrachromosomal	77% (141/158)	34% (62/181)
Interchromosomal	23% (37/158)	66% (119/181)
Number of Duplicates (excl. Scaffolds)	**148**	**147**
Intrachromosomal	82% (121/148)	42% (62/147)
Interchromosomal	18% (27/148)	58% (85/147)
Number of Duplicates (excl. Scaffolds)	**148**	**147**
DNA-Mediated	87% (129/148)	54% (80/147)
RNA-Mediated	13% (19/148)	46% (67/147)
Number of Duplicates (excl. Scaffolds)	**148**	**147**
DNA-Mediated *Intrachromosomal*	120	60
DNA-Mediated *Interchromosomal*	9	20
RNA-Mediated *Intrachromosomal*	1	2
RNA-Mediated *Interchromosomal*	18	65
Number of Duplicates o(excl. Scaffolds)	**148**	**147**
Complete	75.3% (119/158)	49.2% (89/181)
Partial	9.5% (15/158)	11% (20/181)
Chimeric	3.2% (5/158)	0.6% (1/181)
Retroposed	12% (19/158)	39.2% (71/181)
DNA-Mediated Duplicates Only		
Excluding Duplicates on Scaffolds	**139**	**110**
Complete	85.6% (119/139)	80.9% (89/110)
Partial	10.8% (15/139)	18.2% (20/110)
Chimeric	3.6% (5/139)	0.9% (1/110)

Supplementary Table 1 – Evolutionary and genomic features of 181 gene duplicates with low synonymous divergence in the chimpanzee genome.

Structural resemblance types of duplicate were defined as (i) *complete* if sequence homology between the focal paralogs extended throughout their entire open reading frames (ORF); (ii) partial if one paralog possessed unique exon(s) and/or intron(s) in its ORF that are absent in the other paralog; (iii) chimeric if both paralogs contain unique exon(s) and/or intron(s) within their respective ORFs, to the exclusion of the other paralog; 4) retroposed if the ORF of one paralog contained one or more introns which were absent in the other paralog's ORF. Accession numbers correspond to Ensembl ID version 74 released in December 2013.

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000005728	ENSPTRG00000031094	0.000006	13/7	NA	+/+	Retroposed	545	NO
ENSPTRG00000013708	ENSPTRG00000040730	0.000020	20/22	NA	+/+	Retroposed	380	NO
ENSPTRG00000015800	ENSPTRG00000041475	0.000002	3/1	NA	+/+	Retroposed	425	NO
ENSPTRG00000018039	ENSPTRG00000023212	0.000005	6/X	NA	+/+	Retroposed	539	NO
ENSPTRG00000028175	ENSPTRG00000022359	0.000005	X/X	88024019	-/+	Partial	727	NO
ENSPTRG00000031040	ENSPTRG00000019959	0.000006	8/AACZ03163003.1	NA	+/+	Partial	1786	NO
ENSPTRG00000040557	ENSPTRG00000031030	0.000020	AACZ03162641.1/8	NA	+/-	Partial	3284	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000039478	ENSPTRG00000000659	0.000006	16/1	NA	+/+	Retroposed	676	NO
ENSPTRG00000038837	ENSPTRG00000019106	0.000005	7/7	20267259	+/-	Partial	50635	NO
ENSPTRG00000041899	ENSPTRG00000040807	0.000006	Y/Y	8174546	+/+	Complete	33238	NO
ENSPTRG00000042556	ENSPTRG00000000752	0.000025	7/1	NA	+/-	Retroposed	1106	NO
ENSPTRG00000029670	ENSPTRG00000015176	0.000027	5/3	NA	+/-	Retroposed	509	NO
ENSPTRG00000040156	ENSPTRG00000005405	0.000031	12/1	NA	+/-	Retroposed	1081	NO
ENSPTRG00000011298	ENSPTRG00000023483	0.000043	19/16	NA	+/-	Retroposed	659	NO
ENSPTRG00000033835	ENSPTRG00000040582	0.000042	2A/12	NA	-/-	Retroposed	1981	NO
ENSPTRG00000009065	ENSPTRG00000039971	0.000099	17/19	NA	-/+	Retroposed	1318	NO
ENSPTRG00000001650	ENSPTRG00000001649	0.000171	1/1	30885336	+/-	Complete	6219	NO
ENSPTRG00000018547	ENSPTRG00000041140	0.003209	6/GL393552.1	NA	+/+	Retroposed	2702	NO
ENSPTRG00000038711	ENSPTRG00000040238	0.004875	3/GL390583.1	NA	+/-	Complete	16546	NO
ENSPTRG00000011423	ENSPTRG00000009856	0.005244	19/18	NA	-/-	Retroposed	2467	NO
ENSPTRG00000042329	ENSPTRG00000005860	0.005590	13/10	NA	+/+	Retroposed	1869	NO
ENSPTRG00000014555	ENSPTRG00000012386	0.005662	2B/22	NA	+/-	Complete	31154	NO
ENSPTRG00000022808	ENSPTRG00000034348	0.006115	9/3	NA	-/-	Retroposed	655	NO
ENSPTRG00000039432	ENSPTRG00000003768	0.006529	7/11	NA	-/-	Retroposed	1221	NO
ENSPTRG00000022841	ENSPTRG00000028538	0.006666	3/5	NA	+/-	Retroposed	999	NO
ENSPTRG00000023861	ENSPTRG00000017556	0.007789	5/5	110478390	+/+	Complete	31711	NO
ENSPTRG00000007228	ENSPTRG00000023851	0.007926	15/21	NA	+/+	Retroposed	516	NO
ENSPTRG00000041299	ENSPTRG00000041900	0.008234	2B/GL389464.1	NA	-/+	Complete	11886	NO
ENSPTRG00000001693	ENSPTRG00000042455	0.008780	8/1	NA	+/+	Retroposed	1682	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000005075	ENSPTRG00000024159	0.008966	12/15	NA	-/-	Retroposed	414	NO
ENSPTRG00000041287	ENSPTRG00000031237	0.009343	13/13	5994076	-/+	Complete	14675	NO
ENSPTRG00000011299	ENSPTRG00000041745	0.010327	19/5	NA	+/-	Retroposed	566	NO
ENSPTRG00000040245	ENSPTRG00000007663	0.010516	2B/16	NA	+/-	Retroposed	1184	NO
ENSPTRG00000022977	ENSPTRG00000039212	0.011432	19/17	NA	-/+	Retroposed	2251	NO
ENSPTRG00000041906	ENSPTRG00000042297	0.011732	19/19	19651340	-/+	Complete	4987	NO
ENSPTRG00000010777	ENSPTRG00000014403	0.012663	19/22	NA	-/+	Retroposed	974	NO
ENSPTRG00000029890	ENSPTRG00000002501	0.014392	10/10	7293064	+/-	Partial	42684	NO
ENSPTRG00000017060	ENSPTRG00000026567	0.014636	3/5	NA	-/-	Retroposed	321	NO
ENSPTRG00000039330	ENSPTRG00000022785	0.015219	10/8	NA	+/-	Complete	43780	NO
ENSPTRG00000007823	ENSPTRG00000038867	0.015299	16/4	NA	-/-	Retroposed	1826	NO
ENSPTRG00000002673	ENSPTRG00000002671	0.016734	10/10	37635165	-/+	Complete	11144	NO
ENSPTRG00000040648	ENSPTRG00000031173	0.016928	GL392082.1/13	NA	+/+	Partial	1316	NO
ENSPTRG00000013914	ENSPTRG00000031426	0.016954	15/21	NA	+/-	Retroposed	2258	NO
ENSPTRG00000018896	ENSPTRG00000019432	0.017072	7/7	37085604	-/+	Complete	55178	NO
ENSPTRG00000040617	ENSPTRG00000021492	0.018602	5/9	NA	+/-	Retroposed	1360	NO
ENSPTRG00000028438	ENSPTRG00000028442	0.019737	7/7	20246708	-/-	Complete	13092	NO
ENSPTRG00000005737	ENSPTRG00000013681	0.020541	13/20	NA	+/-	Retroposed	212	NO
ENSPTRG00000034202	ENSPTRG00000003333	0.020768	6/11	NA	-/+	Retroposed	489	NO
ENSPTRG00000019571	ENSPTRG00000041649	0.021280	7/10	NA	-/-	Retroposed	2675	NO
ENSPTRG00000039470	ENSPTRG00000017053	0.021315	19/5	NA	+/+	Retroposed	699	NO
ENSPTRG00000001967	ENSPTRG00000041487	0.021609	1/19	NA	-/+	Retroposed	1844	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000017536	ENSPTRG00000008742	0.021693	14/17	NA	+/+	Retroposed	493	NO
ENSPTRG00000004822	ENSPTRG00000029678	0.026714	12/12	724055	-/-	Complete	133709	NO
ENSPTRG00000029776	ENSPTRG00000041526	0.022484	12/12	41049920	-/+	Complete	6784	NO
ENSPTRG00000040744	ENSPTRG00000010200	0.027382	2B/19	NA	-/+	Retroposed	467	NO
ENSPTRG00000004634	ENSPTRG00000019957	0.031053	8/12	NA	+/-	Complete	255122	NO
ENSPTRG00000040903	ENSPTRG00000019672	0.033308	15/7	NA	-/+	Retroposed	653	NO
ENSPTRG00000039319	ENSPTRG00000001213	0.035048	1/1	2512048	-/+	Partial	23207	NO
ENSPTRG00000008384	ENSPTRG00000039664	0.035731	16/9	NA	+/+	Retroposed	3084	NO
ENSPTRG00000015558	ENSPTRG00000007737	0.037685	3/16	NA	-/+	Partial	9868	NO
ENSPTRG00000019987	ENSPTRG00000042155	0.037921	GL390916.1/17	NA	+/+	Retroposed	1261	NO
ENSPTRG00000012250	ENSPTRG00000012214	0.041871	2A/2A	2544102	-/+	Complete	14894	NO
ENSPTRG00000041285	ENSPTRG00000021674	0.045122	19/X	NA	+/-	Retroposed	2691	NO
ENSPTRG00000041063	ENSPTRG00000000426	0.046539	14/1	NA	+/-	Retroposed	1946	NO
ENSPTRG00000041405	ENSPTRG00000018884	0.051001	7/7	2863656	-/+	Partial	28394	NO
ENSPTRG00000011615	ENSPTRG00000034468	0.052116	2A/13	NA	+/-	Retroposed	637	NO
ENSPTRG00000028666	ENSPTRG00000029267	0.052401	7/9	NA	+/+	Retroposed	1818	NO
ENSPTRG00000022548	ENSPTRG00000040111	0.053040	22/X	NA	+/-	Retroposed	2253	NO
ENSPTRG00000022224	ENSPTRG00000039627	0.057938	X/8	NA	-/-	Retroposed	824	NO
ENSPTRG00000030418	ENSPTRG00000029608	0.057921	1/GL389124.1	NA	+/+	Complete	1684	NO
ENSPTRG00000022474	ENSPTRG00000021659	0.060684	Y/X	NA	+/+	Complete	19030	NO
ENSPTRG00000011449	ENSPTRG00000041697	0.061595	19/22	NA	+/+	Retroposed	2002	NO
ENSPTRG00000016122	ENSPTRG00000031183	0.063777	4/4	18555797	-/-	Complete	11504	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000030380	ENSPTRG00000030381	0.065014	1/1	15185046	+/+	Complete	2241	NO
ENSPTRG00000009149	ENSPTRG00000030932	0.065386	17/17	17618580	+/+	Complete	614	NO
ENSPTRG00000007956	ENSPTRG00000031090	0.066736	16/GL392457.1	NA	+/+	Partial	26167	NO
ENSPTRG00000039245	ENSPTRG00000038958	0.069699	19/3	NA	+/+	Retroposed	307	NO
ENSPTRG00000038932	ENSPTRG00000011128	0.071864	19/19	20414530	+/+	Complete	5504	NO
ENSPTRG00000021693	ENSPTRG00000017138	0.076796	X/5	NA	+/+	Retroposed	2023	NO
ENSPTRG00000039638	ENSPTRG00000010157	0.081317	17/19	NA	-/-	Retroposed	3392	NO
ENSPTRG00000017042	ENSPTRG00000015131	0.082508	5/3	NA	+/-	Retroposed	3479	NO
ENSPTRG00000040271	ENSPTRG00000042161	0.085793	2B/7	NA	+/+	Retroposed	851	NO
ENSPTRG00000004091	ENSPTRG00000041165	0.089870	11/3	NA	-/+	Retroposed	3410	NO
ENSPTRG00000023480	ENSPTRG00000010686	0.097544	2A/19	NA	-/+	Retroposed	618	NO
ENSPTRG00000003711	ENSPTRG00000003709	0.097474	11/11	4471425	+/-	Complete	11464	NO
ENSPTRG00000003834	ENSPTRG00000040383	0.098626	11/2A	NA	-/-	Retroposed	590	NO
ENSPTRG00000039915	ENSPTRG00000006638	0.004883	4/14	NA	+/-	Retroposed	4527	NO
ENSPTRG00000040120	ENSPTRG00000001371	0.000006	10/1	NA	-/+	Retroposed	346	NO
ENSPTRG00000013348	ENSPTRG00000038678	0.037251	GL393533.1/11	NA	+/+	Retroposed	736	NO
ENSPTRG00000007960	ENSPTRG00000007825	0.086061	16/16	12436584	-/-	Partial	74604	NO
ENSPTRG00000013615	ENSPTRG00000034248	0.017867	20/3	NA	-/-	Retroposed	648	NO
ENSPTRG00000030689	ENSPTRG00000005536	0.098358	14/12	NA	-/+	Retroposed	661	NO
ENSPTRG00000028209	ENSPTRG00000028208	0.072454	X/X	59165878	-/-	Complete	5199	NO
ENSPTRG00000001350	ENSPTRG00000023699	0.000006	1/1	15629624	-/-	Complete	3539	NO
ENSPTRG00000001350	ENSPTRG00000001349	0.000006	1/1	15616893	-/+	Complete	24127	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000041809	ENSPTRG00000029302	0.026812	18/18	41478698	-/-	Complete	6322	NO
ENSPTRG00000040649	ENSPTRG00000042272	0.009772	7/7	97715849	-/+	Complete	34110	NO
ENSPTRG00000002020	ENSPTRG00000042272	0.043510	GL389157.1/7	NA	+/+	Complete	43449	NO
ENSPTRG00000034210	ENSPTRG00000034284	0.018070	5/1	NA	-/-	Complete	1145	NO
ENSPTRG00000010957	ENSPTRG00000023285	0.028677	19/19	14763470	-/+	Complete	6942	NO
ENSPTRG00000007305	ENSPTRG00000007409	0.028730	15/GL392289.1	NA	-/+	Partial	10460	NO
ENSPTRG00000041839	ENSPTRG00000040123	0.053037	GL392644.1/19	NA	+/-	Complete	2790	NO
ENSPTRG00000028859	ENSPTRG00000040101	0.000008	19/16	NA	-/+	Retroposed	1007	NO
ENSPTRG00000024222	ENSPTRG00000041434	0.061253	4/1	NA	+/+	Retroposed	1038	NO
ENSPTRG00000042403	ENSPTRG00000028201	0.044868	X/X	21863542	-/-	Complete	1860	NO
ENSPTRG00000040429	ENSPTRG00000022445	0.070807	X/X	92967691	-/+	Complete	35855	NO
ENSPTRG00000007377	ENSPTRG00000023171	0.088174	6/22	NA	-/+	Complete	510	NO
ENSPTRG00000023171	ENSPTRG00000007379	0.020760	22/15	NA	+/-	Retroposed	463	NO
ENSPTRG00000013328	ENSPTRG00000013329	0.030297	20/20	2164758	-/-	Complete	56278	NO
ENSPTRG00000041835	ENSPTRG00000006073	0.022955	17/GL393546.1	NA	+/+	Complete	3072	NO
ENSPTRG00000029333	ENSPTRG00000020954	0.036917	18/9	NA	-/-	Complete	1007	NO
ENSPTRG00000041759	ENSPTRG00000039681	0.003600	GL393474.1/1	NA	-/+	Complete	2791	NO
ENSPTRG00000039681	ENSPTRG00000041740	0.010940	1/7	NA	+/+	Complete	2764	NO
ENSPTRG00000003416	ENSPTRG00000022627	0.022965	11/11	32613415	-/+	Complete	3472	NO
ENSPTRG00000020996	ENSPTRG00000020994	0.084168	9/9	44909904	-/+	Partial	4696	NO
ENSPTRG00000010842	ENSPTRG00000029043	0.073406	19/19	10859937	+/+	Complete	3596	NO
ENSPTRG00000030111	ENSPTRG00000015455	0.015916	3/3	47508654	-/-	Complete	1643	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000040873	ENSPTRG00000015455	0.017286	3/3	47470054	-/-	Complete	1208	NO
ENSPTRG00000009205	ENSPTRG00000009204	0.035712	17/17	19092770	-/-	Complete	6197	NO
ENSPTRG00000039807	ENSPTRG00000009506	0.005084	17/17	10980488	+/+	Retroposed	3201	NO
ENSPTRG00000031148	ENSPTRG00000040336	0.023709	16/GL392546.1	NA	-/+	Partial	13250	NO
ENSPTRG00000021134	ENSPTRG00000021159	0.070620	9/9	45202277	+/-	Partial	13994	NO
ENSPTRG00000029643	ENSPTRG00000031401	0.012154	4/10	NA	-/-	Complete	6501	NO
ENSPTRG00000041100	ENSPTRG00000031401	0.090139	3/10	NA	+/-	Complete	6501	NO
ENSPTRG00000009139	ENSPTRG00000009140	0.000007	17/17	17558624	+/+	Complete	979	NO
ENSPTRG00000039092	ENSPTRG00000033900	0.053105	GL392675.1/17	NA	+/+	Complete	7917	NO
ENSPTRG00000030036	ENSPTRG00000033900	0.000007	GL392675.1/17	NA	+/+	Complete	5047	NO
ENSPTRG00000015309	ENSPTRG00000015326	0.086157	3/3	34269102	+/-	Complete	55033	NO
ENSPTRG00000041166	ENSPTRG00000041024	0.009470	GL393537.1/GL394961.1	NA	+/+	Complete	10863	NO
ENSPTRG00000039628	ENSPTRG00000007412	0.082819	GL393475.1/AACZ03172463.1	NA	+/-	Complete	1290	NO
ENSPTRG00000039628	ENSPTRG00000042298	0.031573	GL393475.1/Y	NA	+/+	Complete	69132	NO
ENSPTRG00000020751	ENSPTRG00000032572	0.000007	9/9	39649058	-/-	Complete	5884	NO
ENSPTRG00000039453	ENSPTRG00000039098	0.006504	GL391127.1/15	NA	+/-	Complete	18460	NO
ENSPTRG00000015723	ENSPTRG00000034098	0.000005	3/6	NA	-/-	Complete	389	NO
ENSPTRG00000023887	ENSPTRG00000034098	0.000006	X/6	NA	-/-	Partial	952	NO
ENSPTRG00000040434	ENSPTRG00000005532	0.015539	6/12	NA	+/+	Retroposed	553	NO
ENSPTRG00000022966	ENSPTRG00000019132	0.005614	7/7	41756108	-/-	Chimeric	15012	NO
ENSPTRG00000042540	ENSPTRG00000042282	0.010878	7/7	3156690	+/+	Complete	229214	NO
ENSPTRG00000042540	ENSPTRG00000042585	0.058293	7/GL390634.1	NA	+/+	Complete	105804	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000029283	ENSPTRG00000034374	0.000014	9/9	9887627	-/-	Complete	18554	NO
ENSPTRG00000039103	ENSPTRG00000033768	0.000035	22/14	NA	+/+	Retroposed	339	NO
ENSPTRG00000042141	ENSPTRG00000038938	0.008143	5/6	NA	+/-	Complete	25641	NO
ENSPTRG00000042141	ENSPTRG00000019244	0.008897	5/7	NA	+/-	Partial	13496	NO
ENSPTRG00000028845	ENSPTRG00000028810	0.007758	Y/Y	2780851	-/+	Complete	454929	NO
ENSPTRG00000012114	ENSPTRG00000012113	0.007677	2A/2A	11609943	+/-	Complete	6121	NO
ENSPTRG00000030794	ENSPTRG00000040078	0.000005	1/GL389125.1	NA	+/-	Complete	5323	NO
ENSPTRG00000042347	ENSPTRG00000040078	0.043759	AACZ03149932.1/GL389125.1	NA	+/-	Complete	9769	NO
ENSPTRG00000042347	ENSPTRG00000040171	0.036633	AACZ03149932.1/GL389118.1	NA	+/+	Complete	1303	NO
ENSPTRG00000029254	ENSPTRG00000023039	0.041487	GL391203.1/9	NA	+/+	Complete	75434	NO
ENSPTRG00000013385	ENSPTRG00000042207	0.016846	20/8	NA	+/-	Retroposed	812	NO
ENSPTRG00000014802	ENSPTRG00000041780	0.000013	3/5	NA	-/+	Retroposed	1318	NO
ENSPTRG00000041964	ENSPTRG00000007219	0.036441	15/15	13799808	+/-	Retroposed	1107	NO
ENSPTRG00000042315	ENSPTRG00000012817	0.000100	1/2B	NA	+/-	Retroposed	2003	NO
ENSPTRG00000039108	ENSPTRG00000002992	0.003825	X/10	NA	+/-	Retroposed	1556	NO
ENSPTRG00000000408	ENSPTRG00000000409	0.035695	X/1	NA	+/-	Retroposed	1787	NO
ENSPTRG00000013035	ENSPTRG00000013034	0.063626	2B/2B	101713596	+/+	Complete	14419	NO
ENSPTRG00000013331	ENSPTRG00000014177	0.081973	20/22	NA	-/+	Partial	5227	NO
ENSPTRG00000014177	ENSPTRG00000031326	0.058931	22/GL393073.1	NA	+/-	Partial	16107	NO
ENSPTRG00000033753	ENSPTRG00000034095	0.023834	15/15	57236454	-/+	Complete	59125	NO
ENSPTRG00000042401	ENSPTRG00000038794	0.000007	7/AACZ03179779.1	NA	+/+	Complete	27260	NO
ENSPTRG00000042401	ENSPTRG00000040172	0.000006	7/7	99903263	+/+	Complete	264827	NO

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000041540	ENSPTRG00000040172	0.000005	7/7	98586984	+/+	Complete	347562	NO
ENSPTRG00000034212	ENSPTRG00000034560	0.035168	17/17	28624643	-/-	Complete	5316	NO
ENSPTRG00000038904	ENSPTRG00000039379	0.000005	7/7	1871044	-/+	Complete	92822	NO
ENSPTRG00000033706	ENSPTRG00000020961	0.026236	GL391077.1/9	NA	-/-	Complete	291776	NO
ENSPTRG00000028235	ENSPTRG00000028228	0.025396	X/X	41866236	+/-	Complete	3406	NO
ENSPTRG00000019967	ENSPTRG00000042326	0.091919	GL389982.1/8	NA	+/+	Complete	200515	NO
ENSPTRG00000021644	ENSPTRG00000028322	0.000017	Y/Y	13618447	+/-	Complete	17374	NO
ENSPTRG00000021644	ENSPTRG00000028321	0.047726	Y/X	NA	+/-	Complete	7354	NO
ENSPTRG00000021637	ENSPTRG00000028321	0.000065	X/X	50989580	+/-	Complete	9628	NO
ENSPTRG00000028321	ENSPTRG00000028792	0.058930	X/X	52945949	-/-	Complete	8353	NO
ENSPTRG00000009573	ENSPTRG00000041217	0.095594	GL392695.1/10	NA	+/-	Retroposed	6353	NO
ENSPTRG00000041325	ENSPTRG00000021132	0.041067	9/9	47659263	+/-	Complete	30980	NO
ENSPTRG00000021132	ENSPTRG00000033832	0.097069	9/GL391687.1	NA	-/+	Complete	14912	NO
ENSPTRG00000016203	ENSPTRG00000028248	0.078359	4/X	NA	-/+	Partial	4011	NO
ENSPTRG00000039710	ENSPTRG00000040425	0.026170	GL388884.1/1	NA	-/-	Complete	1312	NO
ENSPTRG00000039479	ENSPTRG00000029238	0.014891	9/9	32658818	-/+	Complete	6941	NO
ENSPTRG00000029238	ENSPTRG00000042546	0.016594	9/9	38688814	+/+	Complete	6314	NO
ENSPTRG00000023647	ENSPTRG00000039445	0.000002	7/1	NA	-/-	Complete	291025	Linked set 1
ENSPTRG00000023135	ENSPTRG00000039924	0.000000	7/1	NA	+/+	Complete		
ENSPTRG00000040891	ENSPTRG00000040633	0.000000	7/7	60469903	+/+	Complete	43445	Linked set 2
ENSPTRG00000028524	ENSPTRG00000028440	0.039407	7/7	60469903	+/+	Complete		
ENSPTRG00000031265	ENSPTRG00000034221	0.030500	4/8	NA	+/+	Complete	252549	

Paralog A Ensembl Gene ID	Paralog B Ensembl Gene ID	K_S	Chr. Location	Distance, if on same chromosome (bp)	Transcription Orientation	Structure Resemblance	Duplication Span (bp)	Linked?
ENSPTRG00000034337	ENSPTRG00000034308	0.086580	8/4	NA	-/-	Complete		Linked set 3
ENSPTRG00000039292	ENSPTRG00000015911	0.032400	8/4	NA	+/+	Complete		

Figures

Figure 1. - Synonymous changes per synonymous site (K_S) based on age distribution of chimpanzee and human gene duplicate pairs.

The large number of duplicates in the youngest age cohort (left most) suggests that a large fraction of gene duplicates in both species may have originated since the human-chimpanzee split, which occurred at $K_S \sim 0.011$ (Chen and Li 2001). The initial death rate is higher for chimpanzee gene duplicates (black), while human gene duplicates (white) decline more gradually over time.

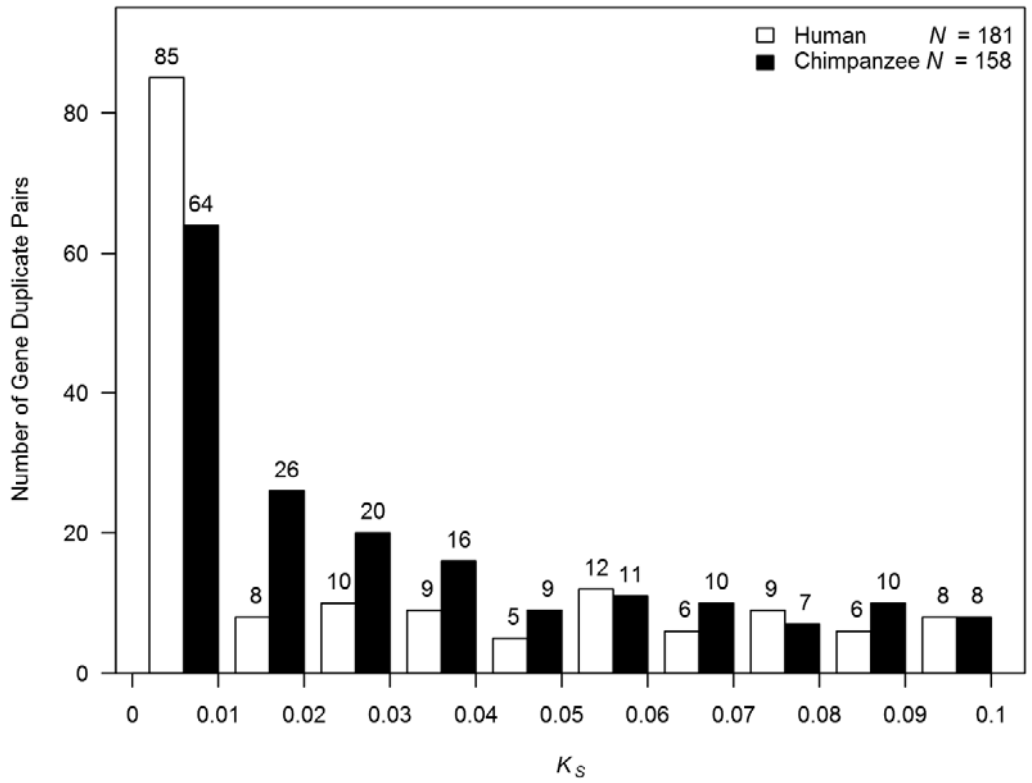


Figure 2. - Composition frequencies of *intra*- and *interchromosomal* duplication within 10 age-cohorts of DNA-mediated duplicates in the human and chimpanzee genomes.

The sample sizes of duplicate pairs within each age cohort are provided above the corresponding bars. The total sample size comprised 129 human and 80 chimpanzee duplicate pairs with assigned chromosomal locations for both paralogs. Only the youngest cohort ($0 \leq Ks \leq 0.01$) shows a significantly different proportion between the human duplicate pairs and those of chimpanzee ($G = 11.0$, $df = 1$, $p = 9e\text{-}4$).

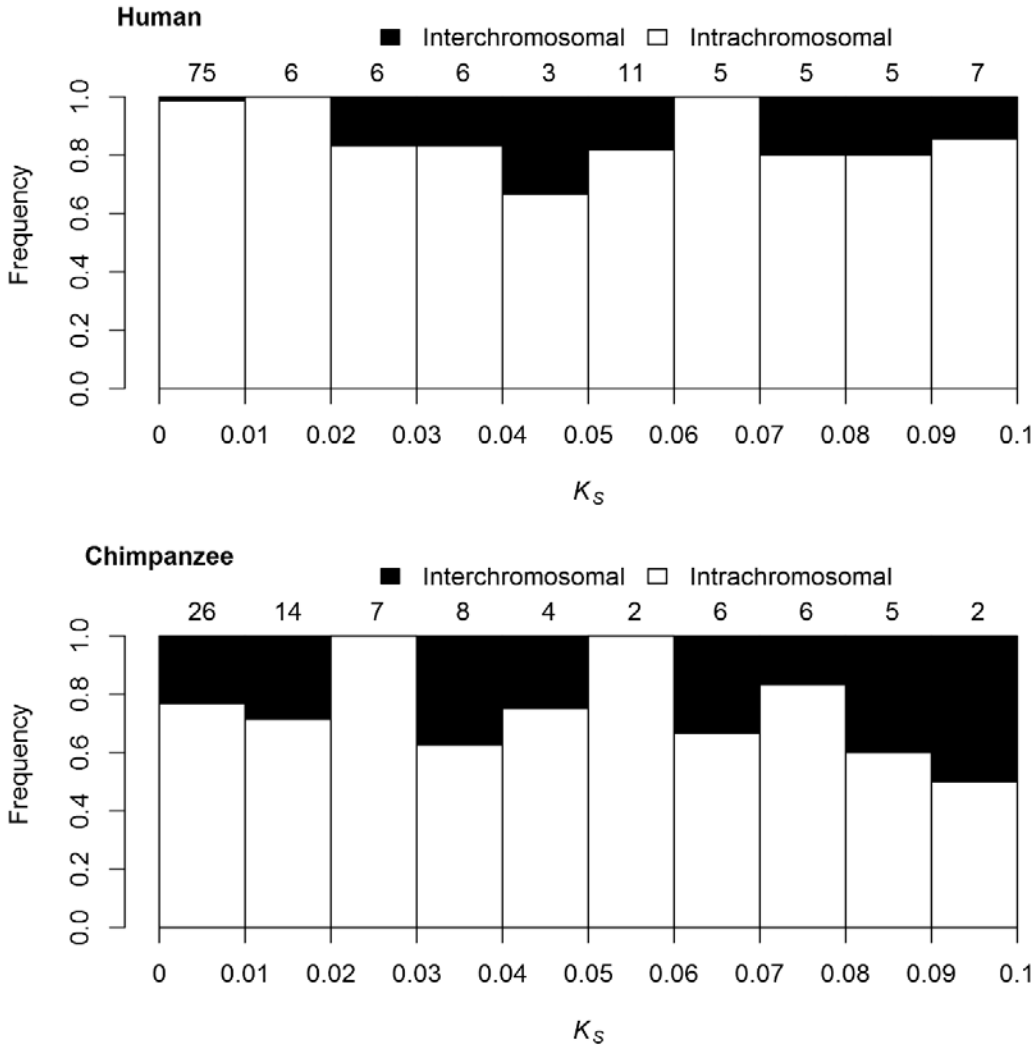


Figure 3. - The physical distance between *intrachromosomal* gene duplicates as a function of K_S in the human and chimpanzee genomes.

The regression line represents the relationship between distance between *intrachromosomal* paralogs (121 and 62 pairs in human and chimpanzee, respectively, with $K_S \leq 0.1$) and K_S. No significant correlation between K_S and paralog distance was found in either species.

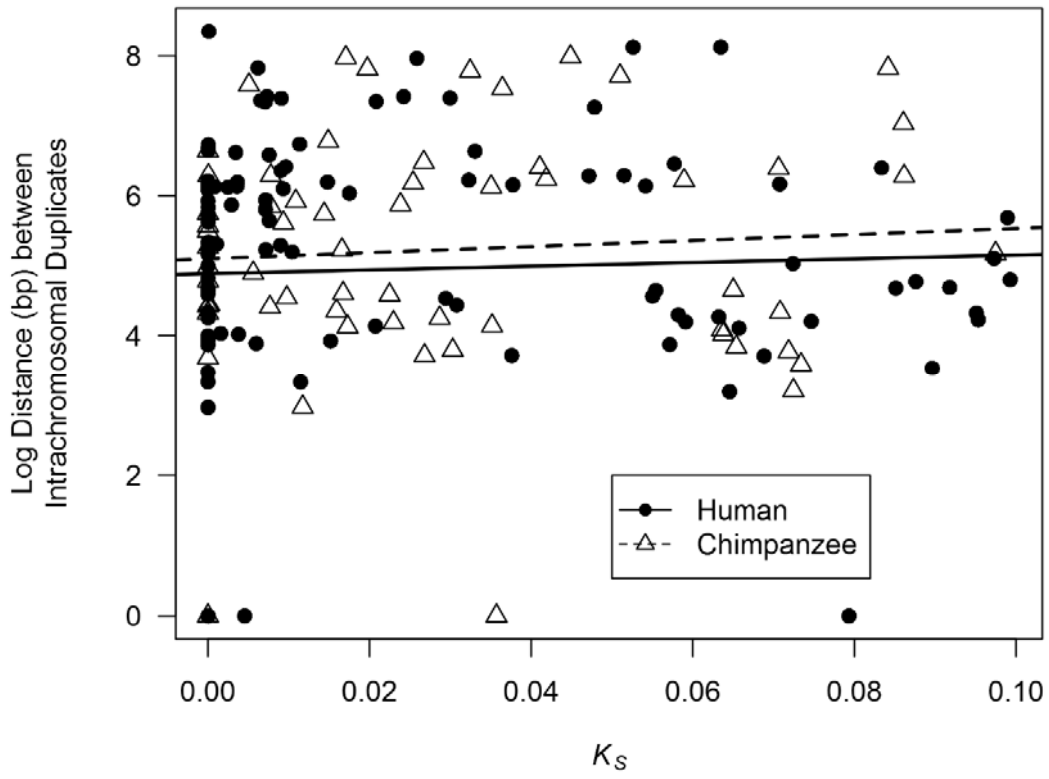


Figure 4. – Nonrandom chromosomal distribution of 147 pairs of young gene duplicates in the chimpanzee genome.

The height of the blue bars indicates the relative duplication frequencies across the 25 chimpanzee chromosomes, calculated as the ratio of the number of duplicate copies on a chromosome and the number of protein-coding genes on the same chromosome. The box plot displays the variation in these relative frequencies across 25 chromosomes, with the median represented by a solid line and the upper and lower quartiles in dotted lines.

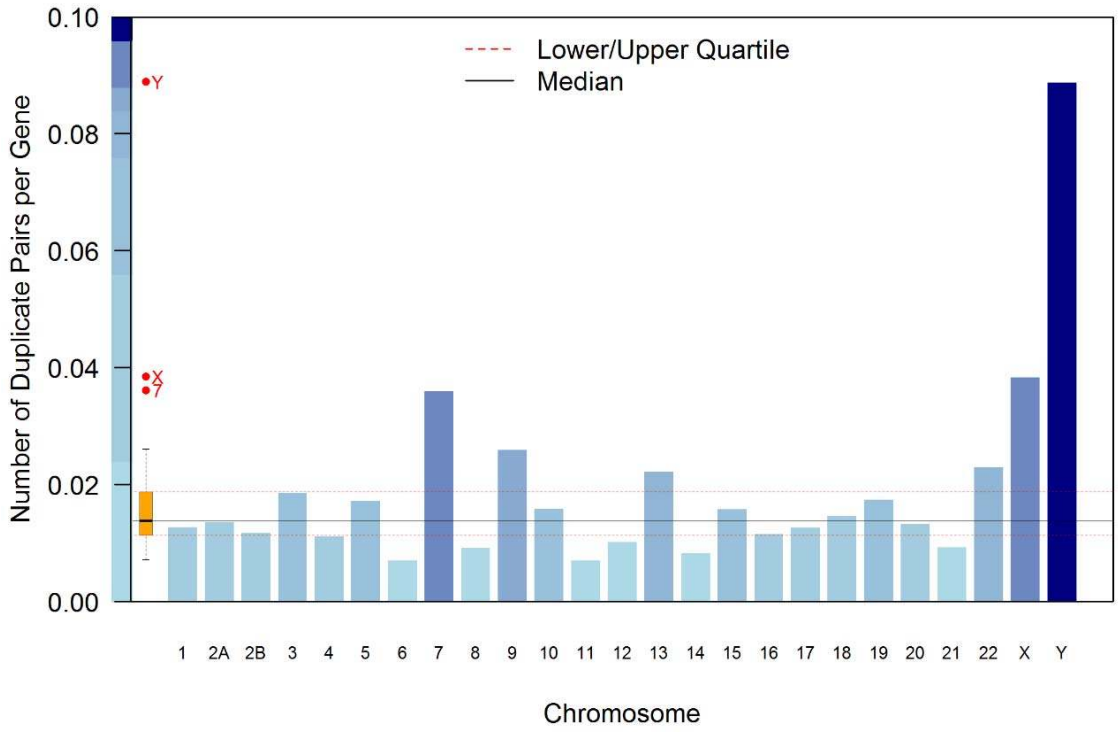


Figure 5. – Proximity of 147 chimpanzee gene duplicates to the centromere

The relative location of DNA-mediated gene duplicates (squares) along chromosomal arms deviates significantly from an expected distribution based on protein-coding gene enrichment (diamonds). No significant deviation was detected for RNA-mediated duplicates (closed circles), or the population of all duplicates (triangles), regardless of duplication mechanism. Each chromosome was subdivided into 10 Mb bins representing increasing distance from the centromere. The proportions of DNA-mediated ($N = 80$), RNA-mediated ($N = 67$), all duplicates (triangles, $N = 147$), and protein coding genes (diamonds, $N = 18{,}759$) were compared. The proportions of gene duplicates and protein-coding genes ($N = 20{,}172$) within each bin are represented by black and white bars, respectively.

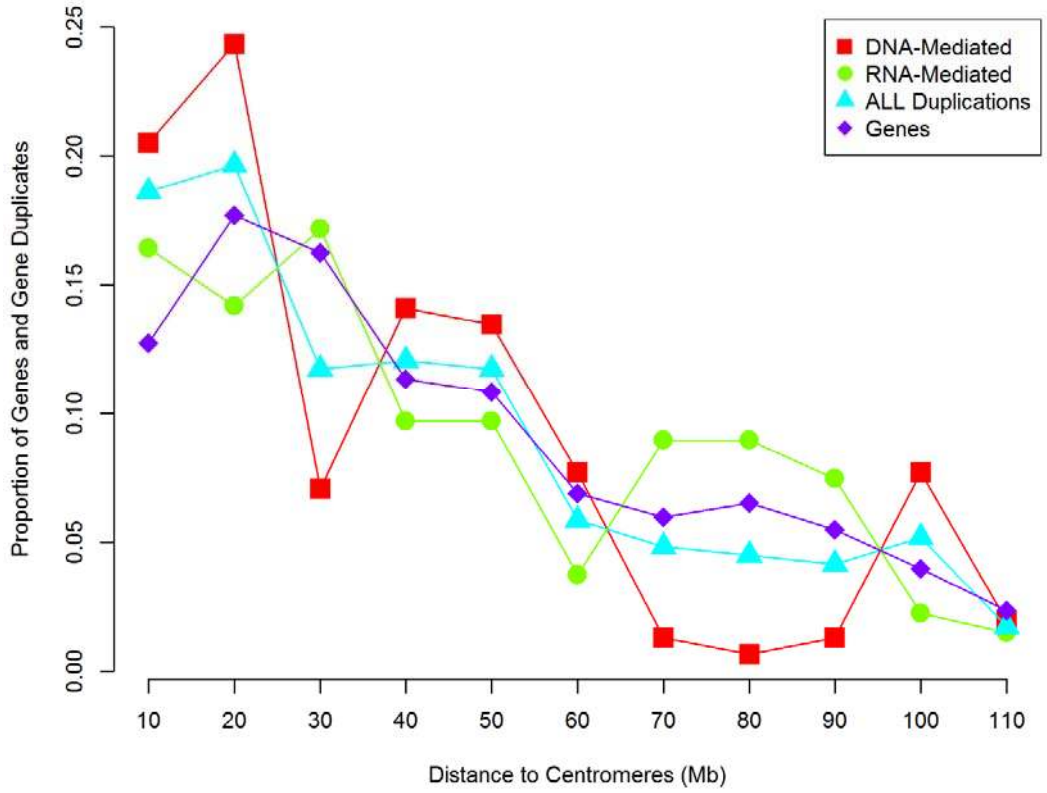


Figure 6. - Composition frequencies of three structural categories of DNA-mediated gene duplicates across 10 evolutionary age-cohorts in the human and chimpanzee genomes.

The total sample size is 139 duplicate pairs for human and 110 pairs for chimpanzee, including duplicates located on scaffolds with unknown chromosome coordinates. Sample sizes for each age cohort are indicated by the numbers in each bar.

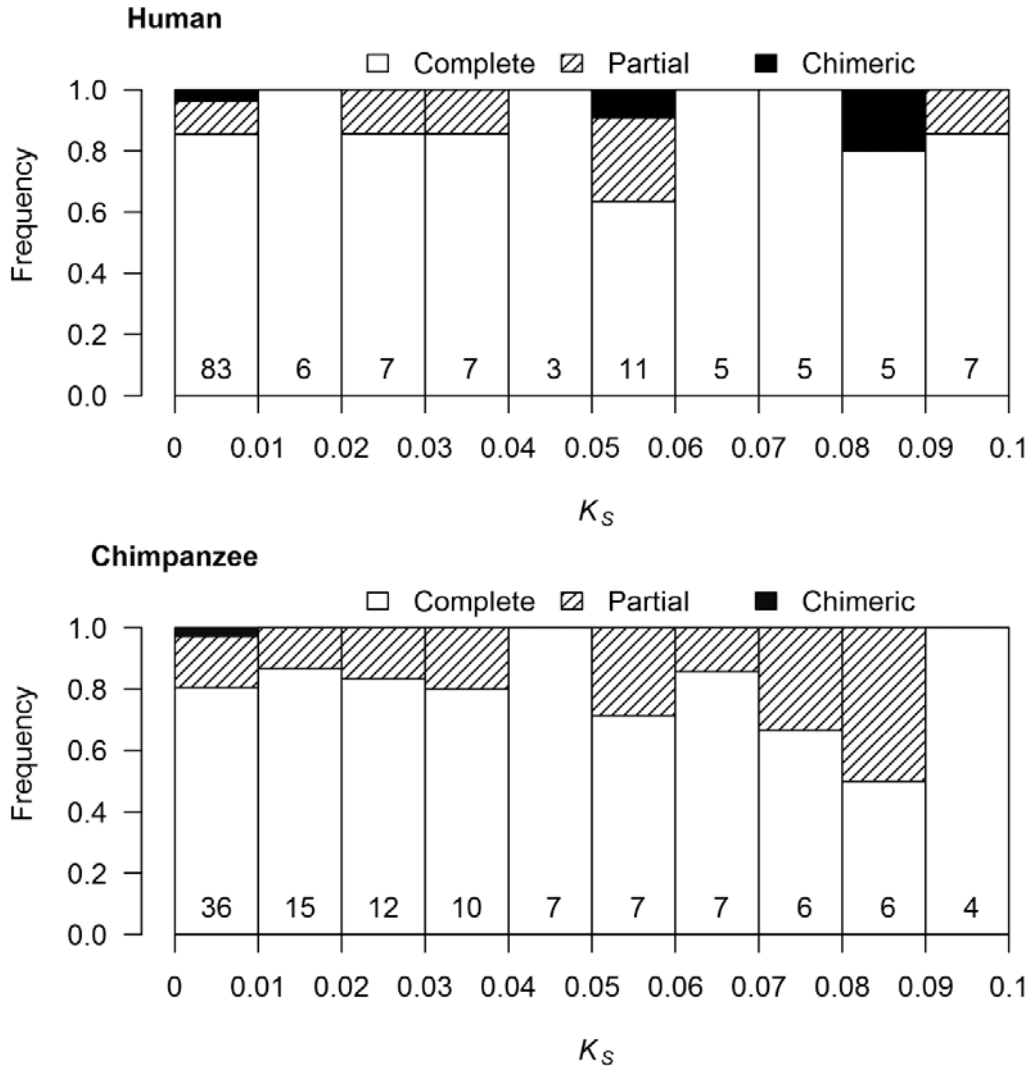


Figure 7. – Violin plots displaying the minimum duplication span of young DNA- and RNA-mediated gene duplicates, as well as the gene and coding region length within the chimpanzee genome.

The range, median, and density for all young gene duplicates (N = 181), DNA-mediated duplicates (N = 110), RNA-mediated duplicates (N = 71) are displayed and compared to the length of genes and coding regions of all protein-coding genes (N = 18,759) within the chimpanzee genome.

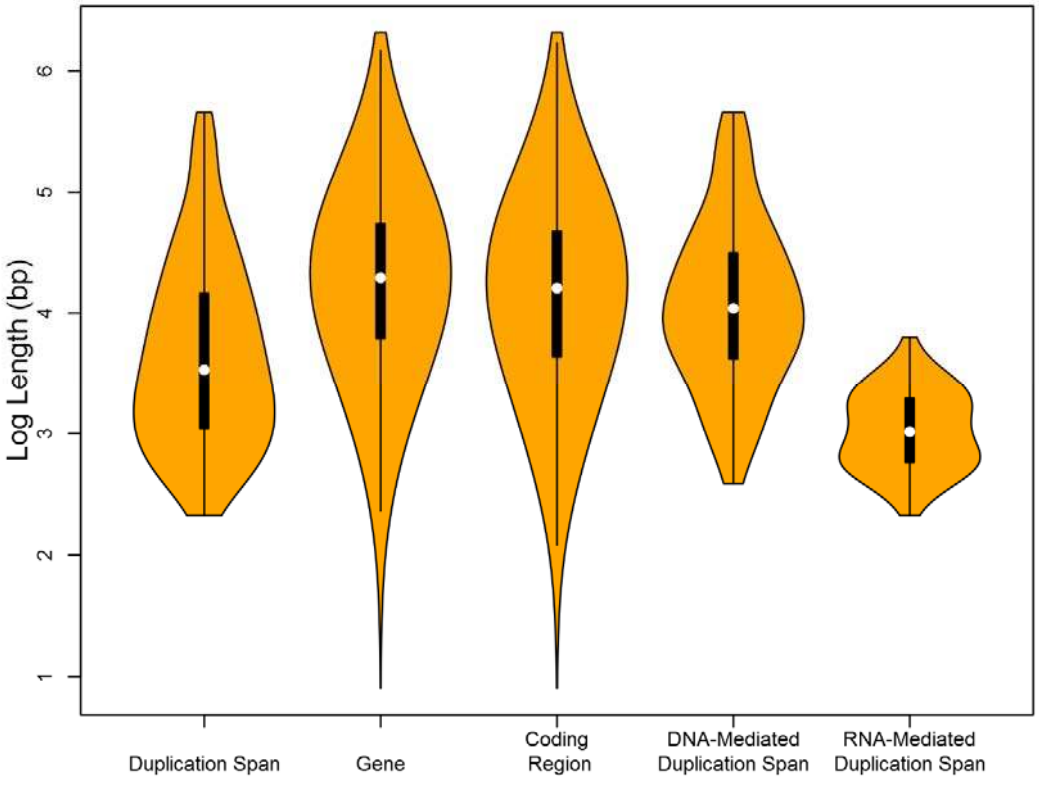


Figure 8. - Duplication span of DNA- and RNA- mediated duplicates as a function of evolutionary age (K_S) in the chimpanzee genome.

The data set contains 80 DNA-mediated duplicate pairs (closed circles) and 67 RNA-mediated duplicate pairs (open diamonds). No significant change of duplication span over evolutionary time is detected.

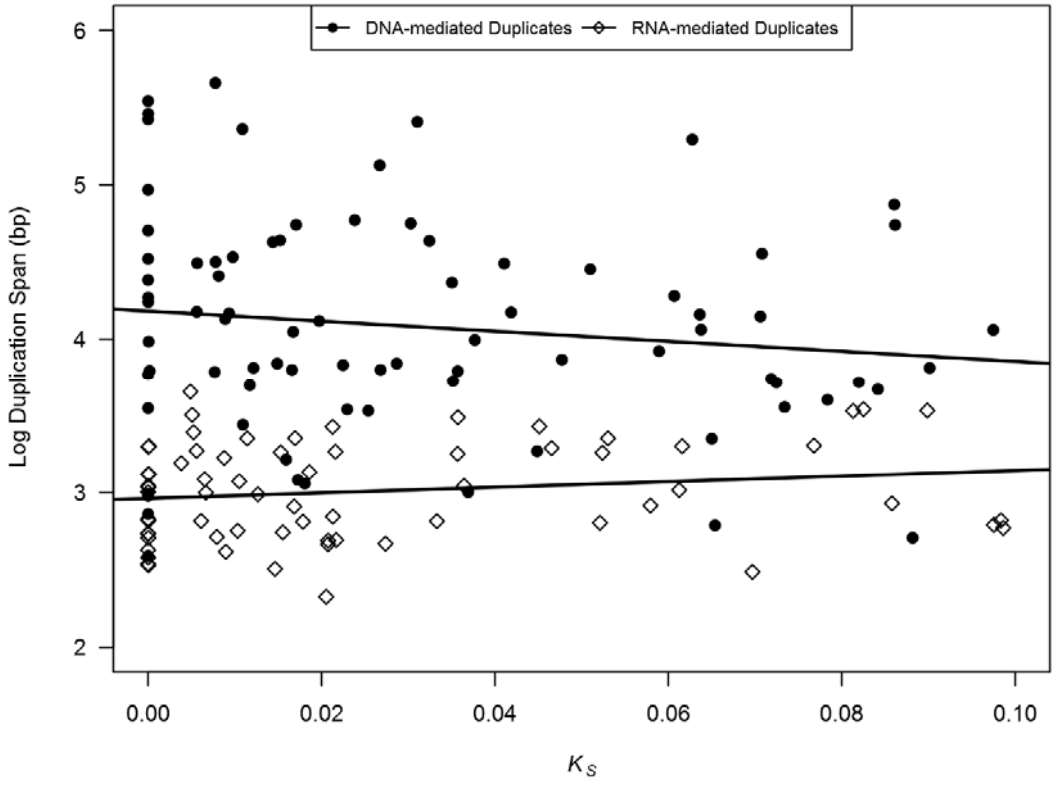
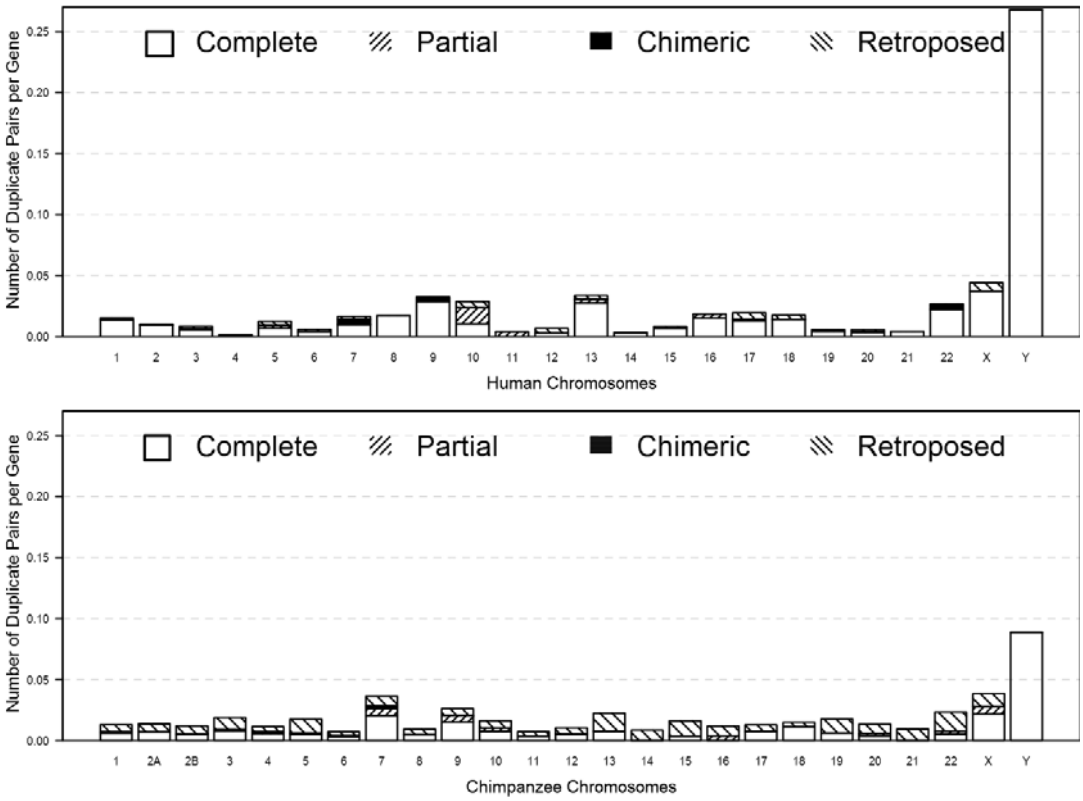


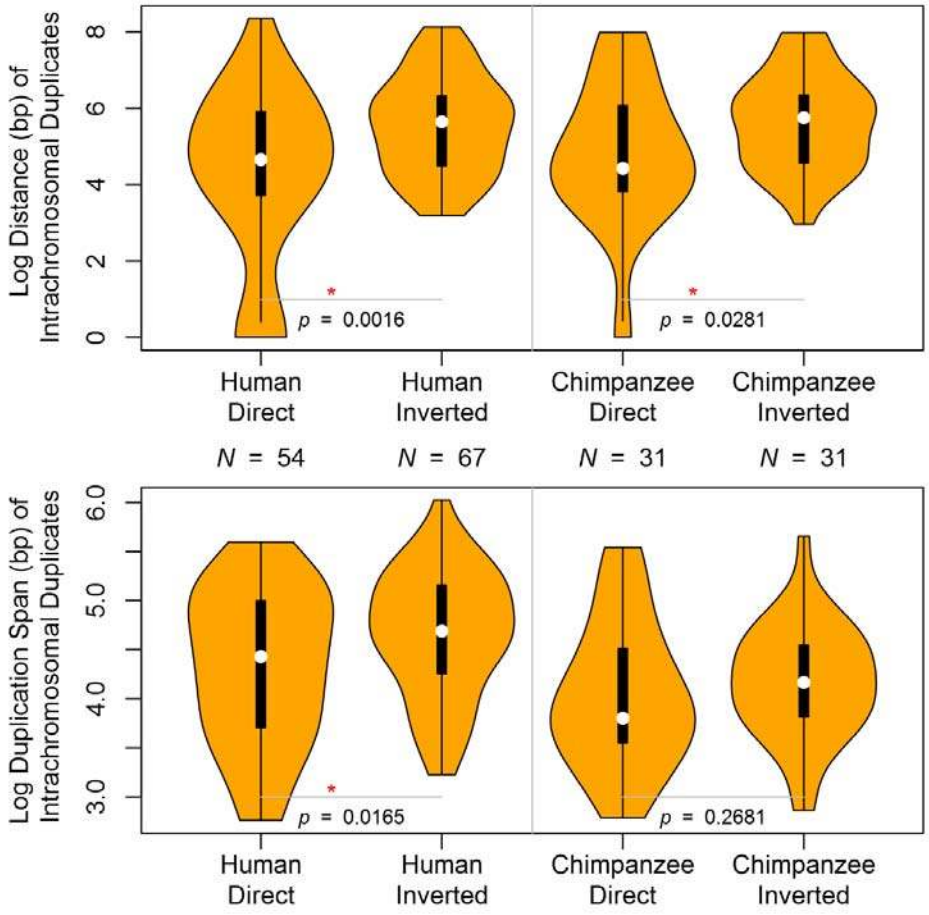
Figure 9. – Chromosome specific duplication frequency for the four structural resemblance types (*complete*, *partial*, *chimeric*, and *retroposed*) within the human and chimpanzee genomes.

Gene duplication frequency for each chromosome (the number of duplicates per gene) was calculated for each structural category.



Supplementary Figure 1 – Log distance between paralogs and duplication span (bp) of *direct* and *inverted intrachromosomal* duplicates in the human and chimpanzee genomes.

Within *intrachromosomal* paralogs, the median distance between two copies of direct orientation is shorter than for inverted ones in both species. In the human genome, the median duplication span for duplicates with *direct* orientation is also significantly shorter than the median duplication span of *inverted* duplicates, which is not the case in the chimpanzee genome. The number of duplicate pairs in each group is given below each plot.



Literature cited

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. (2002). Evidence of *en bloc* duplication in vertebrate genomes. Nat. Genet. *31*, 100–105.

Achaz, G., Netter, P., and Coissac, E. (2001). Study of intrachromosomal duplications among the eukaryote genomes. Mol. Biol. Evol. *18*, 2280–2288.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Babcock, M., Pavlicek, A., Spiteri, E., Kashork, C.D., Ioshikhes, I., Shaffer, L.G., Jurka, J., and Morrow, B.E. (2003). Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. Genome Res. *13*, 2519–2532.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. *11*, 1005–1017.

Bailey, J.A., Yavor, A.M., Viggiano, L., Misceo, D., Horvath, J.E., Archidiacono, N., Schwartz, S., Rocchi, M., and Eichler, E.E. (2002a). Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am. J. Hum. Genet. *70*, 83–100.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R. V, Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002b). Recent segmental duplications in the human genome. Science (80-.). *297*, 1003–1007.

Bailey, J.A., Liu, G., and Eichler, E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. *73*, 823–834.

Benovoy, D., and Drouin, G. (2009). Ectopic gene conversions in the human genome. Genomics *93*, 27–32.

Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. U. S. A. *104*, 17004–17009.

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. *33 Suppl*, 228–237.

Bridges, C.B. (1935). Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. J. Hered. *26*, 60–64.

Bridges, C.B. (1936). THE BAR "GENE" A DUPLICATION. Science *83*, 210–211.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J. V, and Kazazian, H.H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. *100*, 5280–5285.

Bu, L., and Katju, V. (In review) Early evolutionary history and genomic features of gene duplicates in the human genome.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. *15*, 1456–1461.

Carbone, a., Zinovyev, a., and Kepes, F. (2003). Codon adaptation index as a measure of dominating codon bias. Bioinformatics *19*, 2005–2015.

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.-E., Lambert, N., de Marchena, J., Jin, W.-L., Vanderhaeghen, P., Ghosh, A., Sassa, T., et al. (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell *149*, 923–935.

Chen, F.C., and Li, W.H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. *68*, 444–456.

Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C., and Patrinos, G.P. (2007). Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. *8*, 762–775.

Chen, L., DeVries, a L., and Cheng, C.H. (1997). Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. U. S. A. *94*, 3811–3816.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R.K., Pääbo, S., et al. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature *437*, 88–93.

Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.C., and Scherer, S.W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. *4*, R25.

Chung, W.-Y., Albert, R., Albert, I., Nekrutenko, A., and Makova, K.D. (2006). Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. BMC Bioinformatics *7*, 46.

Clark, A.G. (1994). Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. U. S. A. *91*, 2950–2954.

Cliften, P.F., Fulton, R.S., Wilson, R.K., and Johnston, M. (2006). After the duplication: gene loss and adaptation in *Saccharomyces* genomes. Genetics *172*, 863–872.

Conant, G.C., and Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. Genome Res. *13*, 2052–2058.

Connant, G.C., and Wagner, A. (2002). GenomeHistory: a software tool and its application to fully sequenced genomes. Nucleic Acids Res *30*, 3378–3386.

Cotton, J.A., and Page, R.D.M. (2005). Rates and patterns of gene duplication and loss in the human genome. Proc. R. Soc. B. *272*, 277–283.

Courseaux, A., and Nahon, J.-L. (2001). Birth of two chimeric genes in the *Hominidae* lineage. Science *291*, 1293–1297.

Cronn, R.C., Small, R.L., and Wendel, J.F. (1999). Duplicated genes evolve independently after polyploid formation in cotton. Proc Natl Acad Sci U S A *96*, 14406–14411.

Cusack, B.P., and Wolfe, K.H. (2007). Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol. Biol. Evol. *24*, 679–686.

Deeb, S.S., Jorgensen, A.L., Battisti, L., Iwasaki, L., and Motulsky, A.G. (1994). Sequence divergence of the red and green visual pigments in great apes and humans. Proc. Natl. Acad. Sci. U. S. A. *91*, 7262–7266.

Deininger, P.L., and Daniels, G.R. (1986). The recent evolution of mammalian repetitive DNA elements. Trends Genet. *2*, 76–80.

Deng, C., Cheng, C.-H.C., Ye, H., He, X., and Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc. Natl. Acad. Sci. U. S. A. *107*, 21593–21598.

Dennis, M.Y., Nuttle, X., Sudmant, P.H., Antonacci, F., Graves, T. a, Nefedov, M., Rosenfeld, J. a, Sajjadian, S., Malig, M., Kotkiewicz, H., et al. (2012). Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. Cell *149*, 912–922.

Dewannieux, M., and Heidmann, T. (2005). LINEs, SINEs and processed pseudogenes: Parasitic strategies for genome modeling. Cytogenet. Genome Res. *110*, 35–48.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., et al. (2004). The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science *304*, 304–307.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A. *102*, 14338–14343.

Drummond, D.A., Raval, A., and Wilke, C.O. (2006). A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. *23*, 327–337.

Dumont, B.L., and Eichler, E.E. (2013). Signals of historical interlocus gene conversion in human segmental duplications. PLoS One *8*, e75949.

Emanuel, B.S., and Shaikh, T.H. (2001). Segmental duplications: an "expanding" role in genomic instability and disease. Nat. Rev. Genet. *2*, 791–800.

Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O., and Long, M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science *320*, 1629–1631.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. *24*, 363–367.

Fawcett, J.A., and Innan, H. (2013). The role of gene conversion in preserving rearrangement hotspots in the human genome. Trends Genet. *29*, 561–568.

Fisher, R.A. (1935). The Sheltering of Lethals. Am. Nat. *69*, 446–455.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucleic Acids Res. *41*, D48–D55.

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

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. *2*, E207.

Gilbert, W. (1978). Why genes in pieces? Nature *271*, 501.

Gokcumen, O., Tischler, V., Tica, J., Zhu, Q., Iskow, R.C., Lee, E., Fritz, M.H.-Y., Langdon, A., Stütz, A.M., Pavlidis, P., et al. (2013). Primate genome architecture

influences structural variation mechanisms and functional consequences. Proc. Natl. Acad. Sci. U. S. A. *110*, 15764–15769.

Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C.P. (1998). Toward a Phylogenetic Classification of Primates Based on DNA Evidence Complemented by Fossil Evidence. Mol. Phylogenet. Evol. *9*, 585–598.

Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. PLoS Genet. *5*, e1000485.

Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D.C., and Jahn, D. (2005). JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res *33*, W526–W531.

Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. Pathogenetics *1*, 4.

Guo, X., Zhang, Z., Gerstein, M.B., and Zheng, D. (2009). Small RNAs originated from pseudogenes: cis- or trans-acting? PLoS Comput. Biol. *5*, e1000449.

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S.A.A.C., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J.D., et al. (2008). Distribution and functional impact of DNA copy number variation in the rat. Nat. Genet. *40*, 538–545.

Haldane, J.B.S. (1933). The Part Played by Recurrent Mutation in Evolution. Am. Nat. *67*, 5–19.

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

Harris, R.S. (2007). Improved pairwise alignment of genomic dna. Ph.D. Thesis, Pennsylvania State Univ.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat. Rev. Genet. *10*, 551–564.

Hughes, a L. (1994). The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. *256*, 119–124.

Hughes, M.K., and Hughes, A.L. (1993). Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Mol. Biol. Evol. *10*, 1360–1369.

Iatrou, K., Tsitilou, S.G., and Kafatos, F.C. (1984). DNA sequence transfer between two high-cysteine chorion gene families in the silkmoth *Bombyx mori*. Proc. Natl. Acad. Sci. U. S. A. *81*, 4452–4456.

Innan, H. (2003). A two-locus gene conversion model with selection and its application to the human *RHCE* and *RHD* genes. Proc. Natl. Acad. Sci. U. S. A. *100*, 8793–8798.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Inoue, K., and Lupski, J.R. (2002). Molecular mechanisms for genomic disorders. Annu. Rev. Genomics Hum. Genet. *3*, 199–242.

Jeffreys, A.J. (1979). DNA sequence variants in the Gγ-, Aγ-, δ- and β-globin genes of man. Cell *18*, 1–10.

Jun, J., Ryvkin, P., Hemphill, E., Ion, M., Nelson, C., and Măndoiu, I. (2008). Estimating the Relative Contributions of New Genes from Retrotransposition and Segmental Duplication Events During Mammalian Evolution. In Comparative Genomics, C. Nelson, and S. Vialette, eds. (Springer Berlin Heidelberg), pp. 40–54.

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

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. *10*, 19–31.

Katju, V. (2012). In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. Int. J. Evol. Biol. *2012*, 341932.

Katju, V., and Bergthorsson, U. (2010). Genomic and Population-Level Effects of Gene Conversion in *Caenorhabditis* Paralogs. Genes (Basel). *1*, 452–468.

Katju, V., and Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front. Genet. *4*, 273.

Katju, V., and Lynch, M. (2003). The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics *165*, 1793–1803.

Katju, V., and Lynch, M. (2006). On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. Mol. Biol. Evol. *23*, 1056–1067.

Katju, V., LaBeau, E.M., Lipinski, K.J., and Bergthorsson, U. (2008). Sex change by gene conversion in a *Caenorhabditis elegans* fog-2 mutant. Genetics *180*, 669–672.

Katju, V., Farslow, J.C., and Bergthorsson, U. (2009). Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*. Genome Biol. *10*, R75.

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

Kim, S.H., and Yi, S. V (2006). Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. Mol. Biol. Evol. *23*, 1068–1075.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E. V (2002). Selection in the evolution of gene duplications. Genome Biol. *3*, RESEARCH0008.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science (80-.). *318*, 420–426.

Koshi, J.M., and Goldstein, R. a (1996). Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. *42*, 313–320.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langergraber, K., and Prüfer, K. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc. … *109*, 15716–15721.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Leigh Brown, A.J., and Ish-Horowicz, D. (1981). Evolution of the 87A and 87C heat-shock loci in *Drosophila*. Nature *290*, 677–682.

Lercher, M.J., Blumenthal, T., and Hurst, L.D. (2003). Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. *13*, 238–243.

Lieber, M.R., Ma, Y., Pannicke, U., and Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. Nat. Rev. Mol. Cell Biol. *4*, 712–720.

Liebhaber, S.A., Goossens, M., and Kan, Y.W. (1981). Homology and concerted evolution at the alpha 1 and alpha 2 loci of human alpha-globin. Nature *290*, 26–29.

Linardopoulou, E. V, Williams, E.M., Fan, Y., Friedman, C., Young, J.M., and Trask, B.J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature *437*, 94–100.

Lipinski, K.J., Farslow, J.C., Fitzpatrick, K.A., Lynch, M., Katju, V., and Bergthorsson, U. (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. Curr Biol *21*, 306–310.

Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P., and Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. Nat. Biotechnol. *27*, 652–658.

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L. V, Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P., et al. (2011). Comparative and demographic analysis of orang-utan genomes. Nature *469*, 529–533.

Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. *4*, 865–875.

Long, M., VanKuren, N.W., Chen, S., and Vibranovski, M.D. (2013). New gene evolution: little did we know. Annu. Rev. Genet. *47*, 307–333.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science (80-.). *290*, 1151–1155.

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

Lynch, M., and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. Trends Genet *20*, 544–549.

Makova, K.D., and Li, W.H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. *13*, 1638–1645.

Mandal, P.K., Ewing, A.D., Hancks, D.C., and Kazazian, H.H. (2013). Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. Hum. Mol. Genet. *22*, 3730–3748.

Mansai, S.P., Kado, T., and Innan, H. (2011). The rate and tract length of gene conversion between duplicated genes. Genes (Basel). *2*, 313–331.

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

Marques-Bonet, T., Girirajan, S., and Eichler, E.E. (2009). The origins and impact of primate segmental duplications. Trends Genet. *25*, 443–454.

Mathews, L.M., Chi, S.Y., Greenberg, N., Ovchinnikov, I., and Swergold, G.D. (2003). Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. Am. J. Hum. Genet. *72*, 739–748.

Maydan, J.S., Lorch, A., Edgley, M.L., Flibotte, S., and Moerman, D.G. (2010). Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*. BMC Genomics *11*, 62.

McGrath, C.L., Casola, C., and Hahn, M.W. (2009). Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. Genetics *182*, 615–622.

McKay, S.J., Vergara, I.A., and Stajich, J.E. (2010). Using the Generic Synteny Browser (GBrowse_syn). Curr. Protoc. Bioinformatics *Chapter 9*, Unit 9.12.

McVean, G. (2010). What drives recombination hotspots to repeat DNA in humans? Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1213–1218.

Meisel, R.P. (2009). Evolutionary dynamics of recently duplicated genes: Selective constraints on diverging paralogs in the *Drosophila pseudoobscura* genome. J. Mol. Evol. *69*, 81–93.

Melamed, C., and Kupiec, M. (1992). Effect of donor copy number on the rate of gene conversion in the yeast *Saccharomyces cerevisiae*. Mol. Gen. Genet. *235*, 97–103.

Muller, H.J. (1936). BAR DUPLICATION. Science *83*, 528–530.

Nair, S., Miller, B., Barends, M., Jaidee, A., Patel, J., Mayxay, M., Newton, P., Nosten, F., Ferdig, M.T., and Anderson, T.J.C. (2008). Adaptive copy number evolution in malaria parasites. PLoS Genet. *4*, e1000243.

Nembaware, V., Crum, K., Kelso, J., and Seoighe, C. (2002). Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res. *12*, 1370–1376.

Notebaart, R. a, Huynen, M. a, Teusink, B., Siezen, R.J., and Snel, B. (2005). Correlation between sequence conservation and the genomic context after gene duplication. Nucleic Acids Res. *33*, 6164–6171.

Ohno, S. (1970). Evolution by gene duplication (Berlin: Springer-Verlag).

Ollo, R., and Rougeon, F. (1983). Gene conversion and polymorphism: generation of mouse immunoglobulin gamma 2a chain alleles by differential gene conversion by gamma 2b chain gene. Cell *32*, 515–523.

Pan, D., and Zhang, L. (2007). Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. Genome Biol. *8*, R158.

Pan, D., and Zhang, L. (2009). Burst of young retrogenes and independent retrogene formation in mammals. PLoS One *4*, e5040.

Panchin, A.Y., Gelfand, M.S., Ramensky, V.E., and Artamonova, I.I. (2010). Asymmetric and non-uniform evolution of recently duplicated human genes. Biol. Direct *5*, 54.

Van de Peer, Y., Taylor, J.S., Braasch, I., and Meyer, A. (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J. Mol. Evol. *53*, 436–446.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F. a, Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., et al. (2008). Copy number variation and evolution in humans and chimpanzees. Genome Res. *18*, 1698–1710.

Petes, T.D., and Hill, C.W. (1988). Recombination between repeated genes in microorganisms. Annu. Rev. Genet. *22*, 147–168.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., and Carter, D.R.F. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? RNA *17*, 792–798.

R Core Team (2014). R: A Language and Environment for Statistical Computing.

Rane, H.S., Smith, J.M., Bergthorsson, U., and Katju, V. (2010). Gene Conversion and DNA Sequence Polymorphism in the Sex-Determination Gene *fog-2* and Its Paralog *ftr-1* in *Caenorhabditis elegans*. Mol. Biol. Evol. *27*, 1561–1569.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Robinson-Rechavi, M., and Laudet, V. (2001). Evolutionary rates of duplicate genes in fish and mammals. Mol. Biol. Evol. *18*, 681–683.

Samonte, R.V., and Eichler, E.E. (2002). Segmental duplications and the evolution of the primate genome. Nat. Rev. Genet. *3*, 65–72.

Santoyo, G., and Romero, D. (2005). Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol. Rev. *29*, 169–183.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. Mol. Biol. Evol. *6*, 526–538.

Scannell, D.R., and Wolfe, K.H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res. *18*, 137–147.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., and Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature *440*, 341–345.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.

Semple, C., and Wolfe, K.H. (1999). Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. *48*, 555–564.

Sharp, P.M., and Li, W.H. (1987). The Codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res *15*, 1281–1295.

She, X., Liu, G., Ventura, M., Zhao, S., Misceo, D., Roberto, R., Cardone, M.F., Rocchi, M., Green, E.D., Archidiacano, N., et al. (2006). A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. Genome Res. *16*, 576–583.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature *423*, 825–837.

Stankiewicz, P., and Lupski, J.R. (2002). Genome architecture, rearrangements and genomic disorders. Trends Genet. *18*, 74–82.

Sudmant, P.H., Huddleston, J., Catacchio, C.R., Malig, M., Hillier, L.W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R.E., Persengiev, S., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. Genome Res. *23*, 1373–1382.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609–W612.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. Genetics *135*, 599–607.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. *24*, 1596–1599.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A *103*, 3220–3225.

Wagner, A. (2002). Asymmetric functional divergence of duplicate genes in yeast. Mol. Biol. Evol. *19*, 1760–1768.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S., et al. (2006). High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell *18*, 1791–1802.

Wolfe, K. (2000). Robustness--it's not where you think it is. Nat Genet *25*, 3–4.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708–713.

Wong, S., Butler, G., and Wolfe, K.H. (2002). Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc. Natl. Acad. Sci. U. S. A. *99*, 9272–9277.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., and Batzer, M.A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci U S A *103*, 17608–17613.

Yang, Z. (2006). Computational Molecular Evolution.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, Z., Kumar, S., and Nei, M. (1995). A New Method of Inference of Ancestral Nucleotide and.

Zhang, Q. (2013). The role of mRNA-based duplication in the evolution of the primate genome. FEBS Lett. *587*, 3500–3507.

Zhang, J., Rosenberg, H.F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. U. S. A. *95*, 3708–3713.

Zhang, L., Vision, T.J., and Gaut, B.S. (2002). Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol. Biol. Evol. *19*, 1464–1473.

Zhang, L., Lu, H.H., Chung, W.Y., Yang, J., and Li, W.H. (2005). Patterns of segmental duplication in the human genome. Mol. Biol. Evol. *22*, 135–141.

Zhang, P., Gu, Z., and Li, W.-H.H. (2003). Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol. *4*, R56.

Zheng, D., and Gerstein, M.B. (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet. *23*, 219–224.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in *Drosophila*. Genome Res. *18*, 1446–1455.

Zuckerkandl, E., and Pauling, L.B. (1962). Molecular disease, evolution, and genetic heterogeneity. In Horizons in Biochemistry, M. Kasha, and B. Pullman, eds. (New York: Academic Press), pp. 189–225.