7-1-2015

Evolution of Young Gene Duplicates in the Caenorhabditis briggsae Genome

Lucille Packard

Lucille Boreal Packard

Candidate

Biology

Department

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

Approved by the Thesis Committee:

Vaishali Katju, Ph. D., Chairperson

Ulfar Bergthorsson, Ph. D.

Don Natvig, Ph. D.

Evolution of Young Gene Duplicates in the *Caenorhabditis briggsae* Genome

By

Lucille Boreal Packard
B.S., Biology, University of New Mexico, 2012

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

Biology

The University of New Mexico
Albuquerque, New Mexico

July, 2015

Evolution of Young Gene Duplicates in the *Caenorhabditis briggsae* Genome

By

Lucille Boreal Packard

B.S., Biology, University of New Mexico, 2012
M.S., Biology, University of New Mexico, 2015

Abstract

Gene duplication plays a significant role in the evolution of novel function. Investigations concerning the genomic features of young gene duplicates can enhance our understanding of the nature of gene duplicates at inception and their fates over evolutionary time. Previous analysis of young gene duplicates in *Caenorhabditis elegans* has revealed genomic features of evolutionarily young gene duplicates with respect to structure, span, location and orientation of paralogous genes. The analysis of young gene duplicate pairs in *Caenorhabditis briggsae* provides insight into the characteristics of gene duplication in another member of the genus *Caenorhabditis.* The identification of 376 evolutionarily young gene duplicate pairs with less than 10% sequence divergence at synonymous sites and subsequent exploration of certain genomic, structural, and evolutionary features enhances our understanding of gene duplication in eukaryotic species. The mean duplication span of 3.6 kb in young *C. briggsae* gene duplicates is large compared to the mean transcript length of 2.7 kb. The relatively large size of a gene duplicate likely contributes to the abundance of structurally complete *C. briggsae*

gene duplicates. The majority of very young duplicates ($K_S = 0$) are in *direct* orientation on the same chromosome. Though evolutionarily older gene duplicates ($0 < K_S \leq 0.10$) tend to remain on the same chromosome, they are more likely to be in *inverse* orientation. Analysis of selective constraints acting on *C. briggsae* paralogs reveals that the majority of young duplicates are under weak purifying selection regardless of duplication structure. Here I show that, despite its relatively close evolutionary proximity to *C. elegans*, gene duplicates in *C. briggsae* exhibit some strikingly unique genomic characteristics.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Existing analyses of gene duplicates in *Caenorhabditis elegans* have elucidated their characteristics with respect to gene structure, age, and relative location in the genome (Katju and Lynch 2003). The availability of the complete *Caenorhabditis briggsae* genome enables comparative genomic analyses of paralogous genes in another member of the genus *Caenorhabditis* (Stein *et al.* 2003). Although the two species are almost morphologically indistinguishable, they are estimated to have diverged as little as ~30 mya (Cutter 2008) and up to 120 mya (Coghlan and Wolfe 2002). Despite the long evolutionary time separating them, both chromosome organization and synteny are highly conserved between *C. elegans* and *C. briggsae* (Stein *et al.* 2003; Vergara and Chen 2010). Single nucleotide polymorphisms between strains of *C. briggsae* suggest that the effective population size in this species may be several fold greater than in *C. elegans* (Hillier *et al.* 2007). Furthermore, the genome of *C. briggsae* is slightly larger than that of *C. elegans* at 104 Mbp and 100.3 Mbp, respectively. This slight discrepancy in genome size is due, almost entirely, to excess repetitive sequence in the *C. briggsae* genome which is comprised of 22.4% repetitive sequence compared to 16.5% in *C. elegans.* The excess repetitive sequences in *C. briggsae* is suspected to be largely due to transposable elements (Stein *et al.* 2003). In a mutation accumulation study, the genomic mutation rate of *C. briggsae* was estimated to be twice as high as that of *C. elegans* (Baer *et al.* 2005). It is possible that the higher mutation rate of *C. briggsae*

contributes to the increased repetitive sequence in the *C. briggsae* genome, perhaps by increased rate of duplication.

The rate, process, and products of gene duplication have the potential to shape genomes dramatically by providing novel genomic substrate on which evolution may act (Ohno 1970; Lynch and Conery 2000). Tandem duplications in zebra fish, for example, are suspected to have led to significant genome expansion (Lu *et al.* 2012). Understanding the characteristics of gene duplication across species may shed light on the contribution of gene duplication to genome architecture. Previous studies have explored the structural attributes of gene duplication in distantly related eukaryotic species, *C. elegans* (Katju & Lynch 2003) and *S. cerevisiae* (Katju *et al.* 2009). Gene duplicates in *C. elegans* exhibit structural variation from birth, median duplication span does not exceed average gene length resulting in mainly *partial* duplications, and duplicates generally reside close together in the genome, often in *inverse* orientation (Katju and Lynch 2003). In *S. cerevisiae*, different patterns were observed regardless of whether the duplicates considered were ohnologs (the products of whole genome duplication) or paralogs. The majority of gene duplicates in the *S. cerevisiae* genome are located on different chromosomes and are structurally *complete*. The availability of the complete *C. briggsae* genome facilitates a similar analysis in a different species in the genus *Caenorhabditis*. Such analyses can elucidate whether there are structural and mechanistic differences in gene duplicates and gene duplication events between closely related species.

Here I explore the distribution of *complete*, *partial*, and *chimeric* gene duplicates in the genome of *C. briggsae*. *Complete* duplicates possess sequence homology for the entire open reading frame of the duplicated gene, *partial* duplicates are defined by the presence of unique exonic sequence in one copy that is absent in the other paralog, and *chimeric* duplicates each possess unique exonic sequence not present in the other paralog (Katju and Lynch 2003). In a variety of settings, gene duplicates may be transcriptionally active regardless of gene structure (Katju and Lynch 2006). For example, the gene that confers spermatogenesis in *C. elegans*, *fog-2*, is the result of a partial duplication of the gene *ftr-1* (Nayak *et al.* 2004; Katju *et al.* 2008). The SETMAR protein in primates, which is highly transcribed in humans, resulted from a *chimeric* gene duplication that recruited a sequence from a transposase (Cordaux *et al.* 2006). As several studies have demonstrated, duplicates may be retained in plant (Hanada *et al.* 2009; Moody *et al.* 2012), yeast (Gu *et al.* 2003), and nematode (Conant and Wagner 2004) genomes for functional redundancy, and may even contribute to genetic robustness. However, the rate of gene duplicate loss is likely high and genes are commonly lost following duplication (Lynch and Conery 2000; Brunet *et al.* 2006). Regardless of the selective pressures on gene duplicates, they arise in eukaryotic genomes at a high rate. Lipinski *et al*. (2011) found the spontaneous duplication rate in *C. elegans* to be 10^{-7} duplications/gene/generation. This high duplication rate suggests that gene duplication has the potential to contribute significantly to genome composition.

This analysis of 376 evolutionarily young *C. briggsae* gene duplicate pairs with up to 10% synonymous divergence describes the structure, relative genomic location,

and size of young duplicates within the *C. briggsae* genome. *C. briggsae* is a member of the sister clade to *C. elegans.* Yet, despite this evolutionary proximity between these two congeneric species, I find differences in the nature of duplications in *C. briggsae* compared to *C. elegans*. Differences in average duplication span, structure, orientation, and abundance in the *C. briggsae* genome compared to *C. elegans* suggest potential differences in the size and dominant process of duplication between species.

Chapter 2

Methods

Identification of young C. briggsae gene duplicates

Sequences for each duplicate pair were downloaded from Wormbase [www://www.wormbase.org/], release WS235. The longest transcript for each gene was selected using in-house Perl scripts. An all-against-all BLASTP program with a cutoff E value $\leq 10^{-10}$ and identity of $\leq 40\%$ was employed for the similarity search. The low identity cutoff was used with the intention of retaining *partial* and *chimeric* gene duplicates which may be lost in similar searches with higher cutoff identities.

Only duplicate pairs from two member gene families were included in this analysis. The initial list of gene duplicates was composed of 467 young gene duplicate pairs with reported nucleotide substitutions per synonymous site (K_S) values of ≤ 0.10. K_S and nucleotide substitutions per non-synonymous site (K_A) were recalculated using nucleotide and amino acid alignments manually aligned in Se-Al v2.0a11, to ensure an accurate estimate of K_S and K_A. Both nucleotide and amino acid sequences were downloaded from Wormbase (WS235). K_S, K_A, and K_A/K_S calculations were derived by manually entering amino acid and nucleotide alignments into PAL2NAL v14 [http://www.bork.embl.de/pal2nal/] for calculation. Of the 467 initial gene duplicate pairs, seven were removed because K_S values were shown to be higher than the 0.10 cutoff. An additional 13 gene duplicate pairs were excluded from this analysis after manual alignment and analysis of genomic location proved them to be either (i) splice

variants or (ii) structurally ambiguous due to gaps in genomic sequence. The exclusion of 20 pairs from the initial list resulted in a final list of 447 gene duplicate pairs.

Information concerning additional genomic features was obtained, along with both nucleotide and amino acid sequence, from Wormbase (WS235). Chromosomal location, including chromosome number, gene coordinates, and transcript orientation of all paralogous genes was harvested from Wormbase.

Analysis of duplication span

Duplication span for each set of paralogs was obtained from the manual alignment of paired sequences. A break in homology between duplicate pairs was taken to be the duplication breakpoint. As duplicates age, large indels are commonly observed between sequences. For this reason, a break in homology of $\geq$ 1,000 bp was taken to represent a duplication breakpoint. 100 bp N-tracks are prevalent in the *C. briggsae* genome. Correspondence with Wormbase revealed that these 100 bp N-tracks are used as spacers in the genome in place of sequence gaps of unknown length. N-tracks of $\geq$ 100 bps commonly flanked homologous sequence on the 3′ and/or 5′ ends. To maintain a consistent estimate of the minimum conservative duplication span, $\geq$ 100 bp N-tracks were interpreted as the end of sequence homology. Such gaps were common and observed in 262 of the 447 total gene duplicate pairs on the 3′, 5′, or both 3′ and 5′ ends. Along with this systematic underrepresentation of duplication span comes an inherent

overrepresentation of the distance between duplicate pairs residing on the same chromosome.

Due to the large relative size of a duplication event compared to the average gene size in *C. briggsae*, duplication of multiple genes in one event was not uncommon. Multiple genes captured in a single duplication event are referred to here as linked sets. Linked sets were identified using an in-house genome browser [129.24.144.246/test/cbgiggsae/summary.html] and the Wormbase (WS245) genome browser for confirmation. Linked sets were manually aligned and genomic distance between duplicates was calculated using information on chromosomal location obtained from Wormbase (where available). K_S, K_A, and K_A/K_S were calculated by contiguously aligning nucleotide and amino acid sequences of all genes in a linked set, excluding stop codons. The averaged K_S, K_A, and K_A/K_S values were used for the representative pair of each set in later analyses of gene duplicate characteristics. All redundant pairs from a single linked set were excluded from all final analyses, except for the analysis of structural variants, to avoid overrepresentation of genomic features.

Structural categorization of gene duplicate pairs

In order to tease apart differential levels of selection and altered evolutionary trajectories of duplicates of different structural types, I categorized all duplicate pairs in the analysis as belonging to one of three structural categories (i) *complete*, (ii) *partial*, or (iii) *chimeric*. Structural classification depends on the relative exon-intron structure

between the two paralogs in a duplicate pair. The product of capture of an entire open reading frame by a single duplication event is here considered a complete duplication event and the involved pairs are considered structurally *complete.* Sequence homology between two *complete* duplicate pairs often extends from the open reading frame into the 3′ and 5′ flanking sequence. Duplicates were classified as *partial* if unique exonic sequence was observed in one copy and absent in the other. *Partial* gene duplicates are a result of a duplication event capturing only part of a gene's open reading frame. A gene duplicate pair was classified as *chimeric* when each paralog possessed unique exonic sequence not present in the other paralog. In a few cases, large N-tracks, in either flanking exonic sequence or internal exonic and intronic sequence, obscured the ability to confidently identify and structurally categorize gene duplicates. These pairs were excluded from the analyses.

Analysis of relative duplicate location

Due to the incomplete annotation of the *C. briggsae* genome, many genes have yet to be mapped to chromosomes and remain on contigs. According to Wormbase (WS235) these unmapped genes reside on UN, unknown chromosomes. Of the 376 pairs considered in the analysis of relative chromosomal location, 92 were excluded from the analysis because the genomic location of one or both paralog(s) could not be determined. The remaining duplicates were categorized as being on either the same chromosome or on different chromosomes. The distance between the 240/284 gene duplicate pairs with both copies residing on the same chromosome was calculated using

the genomic coordinates obtained from Wormbase and subtracting 3′ or 5′ homologous flanking sequence from both pairs, when present.

Statistical analysis

Statistical analyses were carried out in both R and Microsoft Excel. All *G* - tests for goodness of fit and *G* - tests of independence were performed manually in Microsoft Excel. Initially gene duplicate pairs were divided into six age-cohorts, $K_S = 0$, $0 < K_S \leq 0.01$, $0.01 < K_S \leq 0.03$, $0.03 < K_S \leq 0.05$, $0.05 < K_S \leq 0.07$, and $0.07 < K_S \leq 0.10$. Due to the small size of the two oldest age-cohorts, duplicates were condensed into three age-cohorts, $K_S = 0$, $0 < K_S \leq 0.03$, $0.03 < K_S \leq 0.10$. Important statistical differences were well defined between the youngest age-cohort and the two older cohorts, with the two older cohorts grouping together. So, for all relevant analyses, two age-cohorts were considered, $K_S = 0$, and $0 < K_S \leq 0.10$. All regression analyses, correlation analyses, and Wilcoxon tests were performed in R. The Kolmogorov-Smirnov test was performed using an online generator, [www.physics.csbsju.edu/stats/KS-test.n.plot_form.html].

Chapter 3

Results

Duplication structure in the C. briggsae genome

Gene duplicates were divided into three structural categories, *complete*, *partial*, and *chimeric* based on relative structural resemblance between paralogs. Genes belonging to linked sets, sets of genes copied in a single duplication event, were included in the analysis of duplicate structure. A single duplication event may capture multiple genes depending on the span of the event and the area covered by the duplication. Duplication structure may vary between genes captured in a single linked duplication event. For this reason, the structure of each pair in each linked set was considered a unique entity. Of the 447 gene duplicate pairs considered, 257/447 were structurally *complete*, 158/447 pairs were *partial* gene duplicates, and 32/447 were considered *chimeric*. Duplicates were divided into two cohorts, $K_S = 0$ and $0 < K_S \leq 0.10$ and a G – test of independence was performed comparing the two cohorts. No significant difference was found between the frequencies of the three structural categories within the two cohorts ($G = 0.58$, $d.f. = 2$, $P > 0.5$) (**Fig. 1**).

The majority of young duplicate pairs in the *C. briggsae* genome were structurally *complete*. *Complete* gene duplicates comprised roughly 57% of the pairs in this analysis with the remaining 35% and 7% being of *partial* and *chimeric* structure, respectively. It is not surprising that a significant proportion of gene duplicates in this

analysis are structurally *complete* given that the mean duplication span of 3,637 bp greatly exceeds the average gene transcript length of 2,731 bp.

Relative genomic location of C. briggsae paralogs

Due to the incomplete annotation of the *C. briggsae* genome, a large fraction of the duplicate pairs in this analysis have not been assigned unique chromosomal locations. Of the 376 pairs remaining after excess pairs belonging to the same linked set were removed to avoid over representation of certain genomic features, 92/376 pairs had either one or both paralogs located on unknown chromosomes according to Wormbase (WS235). Gene duplicates on unknown chromosomes could not be included in the analysis of chromosomal location because location relative to the other gene in a pair could not be known. Of the 284 gene duplicate pairs for which location was known, 240/284 were located on the same chromosome while the remaining 44/284 pairs were located on different chromosomes. The predominance of *intrachromosomal* duplications is to be expected due the nature of formation of gene duplicates by unequal crossing-over. However, proximal duplicates formed by unequal crossing-over have the potential to be lost through the same mechanism by which they are formed suggesting that gene duplicates residing on different chromosomes may have a higher rate of survival. To test this hypothesis duplicate pairs were again divided into two cohorts, $K_S = 0$ and $0 < K_S \leq 0.10$ **(Fig. 2**). A significant difference in relative location of gene duplicates was observed between the two cohorts ($G = 34.12$, $d.f. = 1$, $P << 0.01$). Roughly 96% of duplicates within the $K_S = 0$ age-cohort reside on the same chromosome

while only 72% duplicates of $0 < K_S \leq 0.10$ reside on the same chromosome with the remaining 28% located on different chromosomes compared to only 4% of duplicates in the $K_S = 0$ cohort. It is likely that duplicates born on different chromosomes have an increased likelihood of survival compared to those born on the same chromosome.

Distance between intrachromosomal gene duplicates

Because gene duplicates are more likely to be lost when they are in close proximity in the genome, one may expect to find differential rates of survival between genes in close chromosomal proximity compared to genes that are farther apart on the same chromosome. A regression analysis of the 240 *intrachromosomal* duplicate pairs found no evidence suggesting that older gene duplicates are farther apart than younger pairs (*adj* $R^2 = -0.003459$, $P > 0.5$). There appears to be a slight decrease in genomic distance with evolutionary age but the regression does not explain an adequate amount of the variation to draw any conclusions (**Fig. 4**). Results did not change significantly when the youngest cohort ($K_s = 0$) was removed from the analysis.

Density of gene duplicates per chromosome

Additional statistical analyses were performed to investigate whether the density of gene duplicates differs among the six *C. briggsae* chromosomes. No evidence for a duplication bias on any of the six *briggsae* chromosomes was observed **(Fig. 3**). The number of duplicates per chromosome was calculated by dividing the number of duplicated genes on a single chromosome by the total number of genes on that

chromosome. A G – test for goodness-of-fit found no evidence for increased duplication on any given chromosome(s) ($G = 3.56$, $d.f. = 5$, $P > 0.25$). The number of duplications per chromosome seems to correlate with the gene density of each chromosome. Because *Caenorhabditis* chromosomes are holocentric, no analysis on distance from centromere was performed.

Transcriptional orientation of duplicate pairs residing on the same chromosome

To investigate whether the relative transcriptional orientation of gene duplicates affects their rate of survival, the 240 gene duplicate pairs known to be located on the same chromosome were examined. Duplicate pairs with paralogs in *direct* and *inverse* orientation comprise copies transcribed in the same and opposite direction, respectively. Of the 240 duplicate pairs known to reside on the same chromosome, 131 and 109 were categorized as having *direct* and *inverse* orientation, respectively. The data set was divided into two K_S cohorts ($K_S = 0$ and $0 < K_S \leq 0.10$) to test for a difference in transcriptional orientation between very young gene duplicates ($K_S = 0$) and older gene duplicates with increased divergence at synonymous sites (**Fig. 5)**. A G – test of independence revealed a significant difference in the proportion of *direct* and *inverse* duplicates between the two cohorts ($G = 15.69$, $d.f. = 1$, $P < 0.0005$) suggesting that orientation is not independent of evolutionary age. The $K_S = 0$ cohort comprised predominantly *direct* paralogs while the $0 < K_S \leq 0.10$ was comprised of more *inverted* than *direct* duplicates. The predominant mechanism by which gene duplications in *Caenorhabditis briggsae* are formed may favor the creation of *direct* duplicates but such

duplicates may, over time, be preferentially lost or inverted as a result of chromosomal rearrangements.

Analysis of gene duplicate span

The analysis of duplication span excluded redundant pairs from linked sets to avoid over representation of the span of a single duplication event. *C. briggsae* gene duplicates were aligned manually in Se-Al. Either a break in sequence homology by an N-track of ≥ 100 bp or a break in sequence homology for ≥ 1,000 bp was taken as the duplication breakpoint. N-tracks of ≥ 100 bp were common in the *C. briggsae* genome due to its incomplete annotation and low sequencing coverage. Flanking N-tracks determined the 3′, 5′ or both breakpoints in 262/447 duplicate pairs. Duplicates ranged in span from 161 – 13,397 bp with a mean and median duplication span of 3,637 bp and 3,274 bp, respectively. Despite potentially underestimating the duplication span, the average duplication span is significantly larger than the average gene transcript length of 2,731 bp in the *C. briggsae* genome (Wilcoxon rank sum test two-tailed $W = 3032555$, $P < 2.2e\text{-}16$). This relatively large duplication span may explain the predominance of *complete* duplicates in *C. briggsae.* Duplication span appears to be negatively correlated with divergence at synonymous sites (Kendall's tau = -0.2339662, $P = 7.698e\text{-}10$) indicating that duplication span is likely to decrease with age in *C. briggsae* (**Fig. 6**). This strong negative correlation between duplication span and divergence at synonymous sites could indicate either an increased rate of survival of smaller gene duplicates or a loss/shortening of sequence with evolutionary time possibly due to

deletions in sequence. The frequency distribution of duplication span shows that the majority of gene duplicates are small (**Fig. 7**). 60% of duplicates in this analysis were under 4,000 bp in length whereas only ~11% are between 8 – 13 kb in length.

Frequency distribution of K_S

The distribution of K_S values of gene duplications in this analysis is strongly L – shaped (**Fig. 8**) with 218 gene duplicate pairs belonging to the $K_S = 0$ cohort and the remaining 158 pairs in the $0 < K_S \leq 0.10$. This strongly left skewed distribution may be due to selection against the accumulation of extraneous sequence in the *C. briggsae* genome suggesting that gene duplicates are lost at a high rate after birth.

Selective constraints on gene duplicate pairs

K_A/K_S ratios provide information regarding selective constrains acting on genes in a duplicate pair. A K_A/K_S ratio of one suggests duplicate pairs are evolving under neutral selection and neither synonymous nor nonsynonymous changes are favored, both being tolerated equally by selection. When the proportion of synonymous substitutions between two paralogs increases relative to the proportion of nonsynonymous substitutions, purifying selections is implied. Purifying selection acts to prevent amino acid changes between sequences and may be interpreted as selective pressure for conserved function of both duplicate pairs. An increase in nonsynonymous substitutions relative to synonymous substitutions between two paralogous genes suggests positive

selection. Positive selection relaxes selection against amino acid changes between two paralogs thus increasing the possibility of functional divergence. The majority of young gene duplicates considered in this analysis appear to be under purifying selective constraints. 107/158 gene duplicates had K_A/K_S of < 1, suggesting varied levels of purifying selection on these duplicate pairs. The remaining 51 duplicate pairs had K_A/K_S values of >1, indicating positive selection. Gene duplicates were divided into two structural categories, *complete* and *partial/chimeric* to explore whether duplicate structure influences selective constraints on paralogs. For each of the two structural categories, K_A versus K_S was plotted and the slopes of the lines were interpreted as the average level of selection acting in each category (**Fig. 9a and 9b**). The slopes of the lines for *complete* duplicates and *partial/chimeric* gene duplicates were 0.5323 and 0.4306, respectively. The observed slopes of less than one indicate weak purifying selection acting on the majority of duplicate pairs in both structural categories. A Kolmogorov-Smirnov test was performed to test for a difference in distribution between K_A/K_S ratios for each of the two structural categories. No significant difference was observed indicating similar levels of selection regardless of structural type ($D = 0.1214$, $P > 0.50$). Duplicate pairs of $K_S = 0$ were not included in this analysis, only the 158 duplicate pairs of $0 < K_S \leq 0.10$ were considered to ensure some divergence time between paralogs.

Chapter 4

Discussion

The analysis of young gene duplicates in the *Caenorhabditis briggsae* genome and a comparison with the features of gene duplicates in the *C. elegans* genome provide a unique opportunity to explore the products of gene duplication and their evolutionary dynamics in two closely-related species. Though the estimate of divergence time between *C. elegans* and *C. briggsae* ranges from ~30mya (Cutter 2008) to ~120mya (Coghlan and Wolfe 2002), both *C. elegans* and *C. briggsae*, a member of the *C. elegans* sister clade, are self-fertilizing and are essentially morphologically indistinguishable. The completion of the *C. briggsae* genome in 2003 created significant potential for comparative genomics analysis between the extensively studied *C. elegans* and *C. briggsae,* which could be used as a reference for comparative genomics across eukaryotic taxa. Despite the *C. briggsae* genome being described as essentially complete by Harris *et al.* in 2003, significant gaps in the most current version of the genome were a hindrance to this analysis. The cursory nature of the most recent release of the *C. brigssae* genome and its rough annotation, despite its release over ten years ago, created a formidable challenge to the identification and analysis of certain genomic features of gene duplicates. In 92 of the 376 identified gene duplicate pairs, either one or both paralogs could not be mapped to a specific chromosome/location and were instead placed on contigs and categorized as residing on an "unknown" chromosome by Wormbase (WS235). The incomplete annotation posed a significant barrier to the thorough analysis of relative duplicate location in the *C. briggsae* genome. The 92 duplicate pairs of unknown chromosomal location were excluded from the analysis of

relative location. However, this should not pose a significant hindrance to my findings as the sample size of duplicates of known location is still large (284 pairs).

The prevalence of large N-tracks in the published *C. briggsae* genome presented another significant barrier to this analysis. Flanking N-tracks ≥ 100 bps were observed in one or both paralogs in roughly 59% of gene duplicate pairs. N-tracks at duplication breakpoints were commonly observed in the analysis of duplication span. Correspondence with Wormbase revealed that often 100 bp N-tracks are substituted in place of gaps of unknown length. This is necessary due to the lack of sequencing depth and quality compared to more complete genomes. In this analysis, N-tracks ≥ 100 bps flanking homologous sequence were taken as the end of homology between duplicate pairs. Often highly repetitive DNA is difficult to sequence using current technology (Treangen and Salzberg 2011; Sipos *et al.* 2012). The observed prevalence of gaps flanking duplicate pairs could be the result of extant highly repetitive sequence in the genome. DNA-mediated gene duplication events are commonly facilitated by highly repetitive genomic sequence (Meisel 2009; Behura and Severson 2013). In their study, Behura and Severson (2013) found that genomic regions rich in repetitive microsatellites facilitate the formation of segmental gene duplicates thus leading to an excess of repetitive sequence in certain genomic regions. Not surprisingly then, these 100 bp spacers were often observed to accurately mark the end of sequence homology for ≥ 1,000 bp (beyond this number of bp was not commonly checked).

Although the above described issues arose frequently in the analysis of the genomic features of young gene duplicates, there is no reason to suspect that the incomplete nature of the *C. briggsae* genome should bias the results in any way. The sample size in the analyses of chromosomal location (284 pairs) and *intrachromosomal* distance between duplicates (240 pairs) was still relatively large despite the number of pairs excluded. Here I present a conservative analyses of duplication span in *C. briggsae*. It is possible that the duplication span in *C. briggsae* is downwardly biased due to the prevalence of gaps and N-tracks in the genome. Despite this possible systematic underestimation of duplication span, I found that the average duplicate in *C. briggsae* is, (i) larger than the average transcript length of a protein coding gene in *C. briggsae* and, (ii) larger than the average gene duplication span in *C. elegans*. An effort to increase the quality and coverage of extant genomes could contribute significantly to future comparative genomics efforts between *Caenorhabditis* species.

The majority of young gene duplicates identified in this study were located in *direct* orientation (**Fig. 5**) on the same (**Fig. 2**) chromosome, suggesting unequal crossing-over and slippage as the primary mechanisms of gene duplication. Gene duplicates born in close proximity are expected to be transcribed in the same orientation (Semple and Wolf 1999; Katju and Lynch 2003). These two common mechanisms of gene duplication, slippage and unequal crossing-over, are facilitated by shared repetitive sequence. The slight increase in size of the *C. briggsae* genome compared to *C.elegans* is largely due to an increase in repetitive sequence, including transposable element sequence, in *C. briggsae* (Stein *et al.* 2003). This observed increase in repetitive

elements could lead to an increased occurrence of slippage, unequal crossing-over events and possibly non-allelic homologous recombination (Bailey *et al.* 2003; Yang *et al.* 2008) leading to an increased occurrence of small-scale duplication events in *C. briggsae*.

We observed a striking deviation from the organization of gene duplicates in the *C. elegans* genome with respect to gene duplicate orientation in *C. briggsae.* Gene duplicates in the newborn cohort ($K_S = 0$) in *C. briggsae* were most commonly found in *direct* orientation (**Fig. 5**), with almost twice as many *direct* gene duplications as *inverted* in the youngest cohort. Abundant young inverted duplications in both *C. elegans* and *Saccharomyces* are thought to be the result of inverted duplications, duplicates that are inverted at inception (Katju and Lynch 2003, Fischer *et al.* 2001). It is possible that the mechanism producing such duplicates in *C. elegans* and *Saccharomyces* is far less common in *C. briggsae*. Here I find that as *C. briggsae* duplicate pairs age ($0 < K_S \leq 0.10$ cohort) they are more likely to be *inverted* (**Fig. 5**). This finding suggests either a differential loss of paralogs in *direct* orientation over time or an increase in secondary chromosomal rearrangement, mainly inversions, in *C. briggsae.* The fact that young gene duplicates residing in close, often tandem, orientation are unstable and therefore likely to be lost by unequal crossing-over (Koszul *et al.* 2006) may account for the decrease in frequency of *direct* duplicates over evolutionary time. Chromosomal rearrangements including large indels exceeding 100 bp were also common and were observed to increase in frequency with evolutionary age and are common in the *C. elegans* genome, particularly insertions (Denver *et al.* 2004).

The patterns of duplicate orientation observed in this analysis agree with the proposed model of *intrachromosomal* duplication suggested in the Achaz *et al.* (2001) study of duplication across several eukaryotic genomes. Their findings suggest that gene duplicates often arise in tandem *direct* orientation and are later separated and inverted via secondary chromosomal rearrangements. This model may also explain the increase in inverted duplications I observe over increased evolutionary time in *C. briggsae*.

In contrast to previous analyses of the characteristics, frequency and mechanism of gene duplicates in humans (Jun *et al.* 2009; Bu and Katju *in review*), chimpanzees (Bu and Katju *in prep*) and, to a far lesser extent, in *C. elegans* (Katju and Lynch 2003), here I find no evidence for RNA-mediated gene duplication in *C. briggsae*. Gene duplication by reverse transcription is identified by a lack of introns in one copy, a relatively large genomic distance separating duplicates, and a lack of homologous flanking sequence between gene duplicates (Zhang 2003). I found no processed duplicates, duplicates with one copy missing introns and homologous genomic flanking sequence, in this analysis. This result suggests that gene duplication by reverse transcription is uncommon in *C. briggsae* or that, if retrotransposition does occur; genes are immediately pseudogenised (Zhang 2003). This result is similar to observations made in *C. elegans* genome where less than 0.001% of observed duplicates were suspected to have arisen via retrotransposition (Katju and Lynch 2003). Based on the results in *C. elegans* and *C. briggsae*, RNA-mediated duplicates are uncommon across *Caenorhabditis* in comparison to other lineages. This result is similar to findings in *Saccharomyces cerevisiae* (Katju *et al.* 2009).

Some genomic regions may be considered duplication hot spots. Likewise, in some cases, entire chromosomes may experience an increased incidence of gene duplication compared to the rest of the genome. Data concerning the relative chromosomal frequency of gene duplicates in both humans and chimpanzees shows a massive increase in duplication frequency on the Y chromosome in both species (Bu and Katju *in prep*). It has been proposed that the abundance of repetitive sequence and the low number of coding genes on the Y chromosome in humans and chimpanzees may lead to an increased incidence of duplication (Bu and Katju *in prep*). In *C. briggsae*, however, duplication frequency per chromosome seems to correlate well with gene density per chromosome (**Fig. 3**) and no apparent increase in duplication activity per any given chromosome is observed, including the sex chromosome, X. In contrast to Achaz *et al.* 2001 findings in *C. elegans*, I find no evidence to suggest a decrease in the abundance of gene duplicates on the X chromosome compared to the five *C. briggsae* autosomes.

K_A/K_S ratios calculated for each duplicate pair were examined to investigate selective constraints acting on young gene duplicates. The majority of gene duplicate pairs in both prokaryotes and eukaryotes are constrained by purifying selection (Lynch and Conery 2000; Kondrashov *et al.* 2002; Bavishi *et al.* 2010; Li *et al.* 2015). The same is true in *C. briggsae* with the majority of paralogs under varying intensities of purifying selection ($K_A/K_S < 1$ - $K_A/K_S << 1$). Only gene duplicate pairs of $0 < K_S \leq 0.10$ were considered in this analysis to allow evolutionary time for sequence divergence

between paralogs. To explore whether there are differential selective constraints between the major structural variants of gene duplicates the dataset was divided into two categories, *complete* gene duplicates and *partial/chimeric* gene duplicates. No evidence was found to suggest varied levels of selection between these structural categories (**Fig. 9a 9b**) suggesting that the majority of young gene duplicates are constrained by purifying selection, regardless of duplicate structure.

Due to the early release of the *C. briggsae* genome and the relatively close evolutionary relationship between *C. elegans* and *C. briggsae,* significant comparative genomics analysis between these two species has been conducted thus far. Markov *et al*. (2015) found an absence of one-to-one orthology between 11 nematode species, including *C. elegans* and *C. briggsae.* Their findings suggest that lineage-specific patterns of amplification and loss are common in nematodes. Similar findings in the Stein *et al.* (2003) comparative analysis found evidence of significant species-specific gene family expansions in chemosensory genes. Conversely, Aevermann and Waters (2008) found evidence of strong selection for conserved function in small HSPs between *C. elegans* and *C. briggsae*. The identification and classification of gene duplicates in *C. briggsae* paves the way for future studies to compare young gene duplicates between *C. elegans* and *C. briggsae* and investigate differential evolutionary constraints on shared gene duplicates between these two species.

Despite potentially underestimating duplication span, the average span of a gene duplicate is large compared to the average transcript length of a protein coding gene in

C. briggsae. I calculated the the mean and median coding transcript in *C. briggsae* to be 2,731 bp and 1,748 bp, respectively. The analysis of young gene duplication span yields a mean and median span of 3,637 bp and 3,274 bp, respectively. Both estimates of span are larger than the mean and median coding transcript length in the genome. The large average duplication span observed in *C. briggsae* likely contributes to the abundance of gene duplicates of *complete* structural resemblance. Gene duplication events likely capture entire open reading frames by chance due to the large relative size of a gene duplication event. These findings stand in contrast to the known composition of structural categories of gene duplicates in both the *C. elegans* (Katju and Lynch 2003) and *D. melanogaster* (Zhou *et al.* 2008) genomes where *partial*/*chimeric* gene duplicates were found to outnumber gene duplicates of *complete* structural resemblance. I also find that the duplication span in *C. briggsae* exceeds the median duplication span of 1.4 kb in *C. elegans* (Katju and Lynch 2003). This relatively large duplication span may, in part, contribute to the slight increase in genome size between *C. briggsae* and *C. elegans* which is thought to be a result of increased repetitive sequences in *C. briggsae.* Likewise, this observed abundance of repetitive sequence in *C. briggsae* compared to *C. elegans* may contribute to an increase in gene duplication events in the *C. briggsae* genome. These subtle genome specific trends in composition may, overtime, contribute to the overall architectural differences observed between extant genomes.

Tables

Supplementary Table 1. – Gene identifications and relevant evolutionary and physical characteristics of 376 gene duplicate pairs used in all analyses excluding the analysis of structural variants.

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG21453	II	14,858,839..14,859,078	+	14,809	0.0000	0.0000	Complete	5,837
CBG21448	II	14,843,058..14,843,297	-					
CBG14454	X	9,648,196..9,648,781	+	1,873,291	0.0000	0.0000	Complete	8,469
CBG14040	X	7,765,809..7,767,252	+					
CBG23640	UN	169,163..170,732	-	N/A	0.0000	0.0000	Complete	5,098
CBG24780	UN	1,965,449..1,967,018	-					
CBG06604	V	7,014,446..7,015,600	-	907,415	0.0000	0.0000	Complete	4,213
CBG06339	V	6,102,800..6,103,954	-					
CBG05675	V	19,360,815..19,362,159	-	1,002	0.0000	0.0000	Complete	7,002
CBG05674	V	19,352,813..19,354,158	-					
CBG22594	IV	3,571,197..3,573,667	+	118,776	0.0000	0.0000	Complete	4,274
CBG26272	IV	3,694,245..3,694,364	+					
CBG24939	UN	2,437,899..2,438,392	+	N/A	0.0000	0.0000	Partial	5,363
CBG20428	I	12,493,529..12,495,076	-					
CBG12943	II	6,585,338..6,585,598	+	6115	0.0000	0.0000	Complete	4,130
CBG12946	II	6,575,091..6,575,351	+					
CBG07468	X	19,080,926..19,081,364	-	133,681	0.0000	0.0000	Complete	3,613
CBG07494	X	18,946,807..18,947,245	-					
CBG14798	X	11,353,104..11,353,327	-	484,517	0.0000	0.0000	Complete	8,123
CBG14697	X	10,857,406..10,857,629	-					
CBG08109	X	290,212..291,866	+	N/A	0.0000	0.0000	Partial	11,827
CBG26680	UN	812,843..813,127	+					
CBG00034	III	12,301,886..12,302,720	-	N/A	0.0000	0.0000	Complete	1,059
CBG23653	UN	222,575..223,409	-					
CBG07058	II	16,288,912..16,292,661	+	149,952	0.0000	0.0000	Chimeric	8,054
CBG24273	II	16,130,807..16,134,643	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG17252	X	6,678,686..6,678,970	+	1,034	0.0000	0.0000	Partial	3,774
CBG17251	X	6,681,246..6,691,756	-					
CBG23712	UN	329,176..329,384	-	N/A	0.0000	0.0000	Partial	894
CBG00255	X	15,098,812..15,100,185	-					
CBG05346	X	17,417,032..17,417,282	+	4,473	0.0000	0.0000	Partial	608
CBG05345	X	17,410,245..17,412,288	-					
CBG24539	UN	1,570,541..1,572,088	+	N/A	0.0000	0.0000	Partial	2,685
CBG17687	IV	11,671,679..11,673,483	+					
CBG02263	II	10,804,607..10,804,870	+	N/A	0.0000	0.0000	Complete	5,690
CBG24807	UN	2,049,847..2,050,110	-					
CBG27787	X	12,757,978..12,758,190	-	1,906	0.0000	0.0000	Complete	4,716
CBG27788	X	12,762,911..12,763,123	-					
CBG16492	III	7,405,251..7,407,505	+	N/A	0.0000	0.0000	Complete	7,599
CBG24567	UN	1,647,182..1,649,436	+					
CBG00322	I	652,393..653,432	+	N/A	0.0000	0.0000	Complete	1,911
CBG24343	UN	1,278,652..1,279,275	-					
CBG10123	III	9,767,611..9,768,802	-	646	0.0000	0.0000	Complete	5,891
CBG10122	III	9,761,075..9,762,266	-					
CBG25333	I	6,305,076..6,307,273	+	N/A	0.0000	0.0000	Partial	4,278
CBG24859	UN	2,138,661..2,147,140	-					
CBG00245	X	15,038,515..15,038,763	+	1,001	0.0000	0.0000	Complete	5,638
CBG00244	X	15,031,877..15,032,125	+					
CBG08173	X	597,001..600,474	-	55,247	0.007358	0.005353	Partial	8,709
CBG08190	X	660,954..666,947	-					
CBG06958	II	16,600,956..16,602,440	+	302	0.0000	0.0000	Complete	7,190
CBG06955	II	16,608,447..16,609,931	+					
CBG24960	UN	2,504,739..2,505,232	+	N/A	0.0000	0.0000	Partial	4,844
CBG16021	II	15,255,746..15,257,440	-					
CBG24066	IV	15,062,570..15,063,801	+	N/A	0.0000	0.0000	Complete	6,312
CBG24928	UN	2,408,504..2,409,735	+					
CBG22485	III	2,877,276..2,886,428	-	N/A	0.0000	0.0000	Partial	3,250

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG25011	UN	2,687,346..2,687,594	+					
CBG04525	X	6,294,097..6,294,454	+	1,137	0.0000	0.0000	Complete	2,972
CBG04526	X	6,289,991..6,290,347	+					
CBG24831	UN	2,077,779..2,078,710	-	N/A	0.0000	0.0000	Partial	161
CBG24872	UN	2,210,085..2,210,237	+					
CBG26669	V	16,296,068..16,297,874	-	N/A	0.0000	0.0000	Complete	3,294
CBG24367	UN	1,340,196..1,342,002	-					
CBG00932	II	610,631..611,426	+	219595	0.0000	0.0000	Complete	2,656
CBG00900	II	832,880..833,675	+					
CBG21797	IV	4,253,769..4,258,626	+	353	0.0000	0.0000	Partial	916
CBG26282	IV	4,259,652..4,260,131	-					
CBG15349	III	14,503,654..14,514,467	+	249,577	0.0000	0.0000	Partial	7,494
CBG00609	III	14,253,133..14,253,968	-					
CBG04255	II	4,844,300..4,845,966	+	8,672	0.0000	0.0000	Partial	3,669
CBG04251	II	4,832,153..4,833,590	+					
CBG20144	III	11,687,131..11,690,509	-	0	0.0000	0.0000	Chimeric	4,183
CBG09120	III	11,691,415..11,694,724	-					
CBG24506	UN	1,496,488..1,497,217	-	N/A	0.0000	0.0000	Partial	4,047
CBG22943	III	14,071,790..14,075,341	-					
CBG25087	UN	2,900,875..2,901,848	+	N/A	0.0000	0.0000	Complete	2,011
CBG16452	IV	14,609,459..14,610,428	-					
CBG24893	UN	2,279,276..2,280,208	-	N/A	0.0000	0.0000	Partial	5,441
CBG15211	III	328,447..332,178	+					
CBG06506	V	6,694,877..6,698,501	-	494	0.0000	0.0000	Complete	7,157
CBG06509	V	6,702,526..6,706,150	-					
CBG12018	I	7,523,335..7,524,906	-	962	0.0000	0.0000	Complete	3,715
CBG12019	I	7,528,010..7,529,581	-					
CBG24855	UN	2,113,047..2,114,538	+	N/A	0.0000	0.0000	Partial	3,709
CBG01856	X	3,052,311..3,060,790	+					
CBG09731	V	10,159,324..10,160,198	+	33,714	0.0000	0.0000	Partial	3,649
CBG09743	V	10,122,224..10,122,827	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG24318	II	14,626,651..14,646,422	+	N/A	0.016327	0.014513	Partial	5,888
CBG24958	UN	2,493,433..2,495,626	+					
CBG04373	I	13,812,952..13,814,496	-	859	0.0000	0.0000	Complete	2,435
CBG04372	I	13,810,072..13,811,615	+					
CBG25484	II	2,867,504..2,868,275	+	31,285	0.0000	0.0000	Complete	4,959
CBG13419	II	2,903,874..2,904,516	+					
CBG25012	UN	2,691,567..2,692,907	+	NA	0.0000	0.0000	Partial	3,538
CBG19720	I	14,700,620..14,704,468	+					
CBG23856	I	12,008,337..12,008,807	+	83,985	0.0000	0.0000	Partial	688
CBG18679	I	11,923,425..11,924,300	+					
CBG17941	V	5,879,526..5,881,478	-	159,734	0.0000	0.0000	Complete	4,852
CBG17888	V	5,714,934..5,715,816	-					
CBG12271	III	12,530,078..12,537,305	-	1,158	0.0000	0.0000	Compete	11,106
CBG12272	III	12,542,341..12,549,568	-					
CBG08624	V	7,967,532..7,969,237	-	1,134	0.0000	0.0000	Partial	3,338
CBG08626	V	7,963,062..7,963,375	-					
CBG13717	IV	1,089,630..1,090,885	-	N/A	0.0000	0.0000	Complete	3,244
CBG25025	UN	2,727,595..2,728,850	-					
CBG23654	UN	224,659..224,895	+	N/A	0.0000	0.0000	Partial	1,912
CBG00033	III	12,298,675..12,301,375	+					
CBG22217	I	2,837,204..2,845,394	+	N/A	0.0000	0.0000	Partial	3,185
CBG25027	UN	2,732,830..2,733,174	-					
CBG10567	IV	16,630,881..16,634,737	+	29,831	0.0000	0.0000	Partial	5,437
CBG10577	IV	16,595,640..16,597,137	+					
CBG10599	X	21,443,798..21,444,174	+	737	0.0000	0.0000	Complete	3,513
CBG10600	X	21,441,336..21,442,177	-					
CBG06907	III	10,157,624..10,158,444	+	402	0.0000	0.0000	Complete	5,314
CBG06906	III	10,163,338..10,164,158	+					
CBG15019	IV	3,881,437..3,881,899	+	2,265	0.0000	0.0000	Complete	1,813
CBG15020	IV	3,885,583..3,886,045	-					
CBG24545	UN	1,591,827..1,592,168	+	N/A	0.0000	0.0000	Complete	3,954

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG24974	UN	2,549,560..2,549,901	-					
CBG13089	II	5,858,808..5,859,842	+	N/A	0.0000	0.0000	Partial	4,327
CBG26633	UN	66,943..67,050	-					
CBG24121	UN	920,932..924,384	+	N/A	0.0000	0.0000	Complete	6,716
CBG22786	X	16,354,546..16,357,998	-					
CBG08811	V	1,795,120..1,796,694	+	1,754	0.0000	0.0000	Complete	4,619
CBG08813	V	1,801,491..1,803,065	+					
CBG24233	UN	1,061,913..1,064,191	-	N/A	0.0000	0.0000	Complete	6,093
CBG21877	II	14,505,474..14,506,615	+					
CBG23792	UN	489,949..490,803	-	N/A	0.0000	0.0000	Complete	2,454
CBG27290	V	17,119,976..17,120,095	+					
CBG06795	III	10,636,170..10,636,674	+	1,002	0.0000	0.0000	Complete	5,805
CBG06793	III	10,642,975..10,643,479	+					
CBG08942	X	20,882,336..20,883,886	-	N/A	0.0000	0.0000	Complete	3,306
CBG24758	UN	1,923,259..1,925,236	+					
CBG19699	III	3,681,641..3,683,417	-	N/A	0.0000	0.0000	Partial	3,616
CBG24955	UN	2,486,928..2,487,352	+					
CBG15797	III	1,652,123..1,655,650	+	499,057	0.0000	0.0000	Partial	6,657
CBG15701	III	1,133,862..1,149,938	+					
CBG22318	IV	10,259,010..10,260,040	-	132,592	0.0000	0.0000	Complete	6,311
CBG24368	IV	10,120,109..10,121,139	-					
CBG24756	I	8,031,339..8,032,334	-	11,624	0.0000	0.0000	Complete	8,842
CBG02255	I	8,052,924..8,053,919	+					
CBG26109	III	14,401,969..14,402,053	+	1,003	0.0000	0.0000	Complete	7,701
CBG26110	III	14,410,674..14,410,758	+					
CBG23817	V	18,076,168..18,078,310	+	1,709	0.0000	0.0000	Complete	6,433
CBG27335	V	18,084,308..18,086,452	+					
CBG17262	X	6,618,110..6,618,397	+	78,053	0.0000	0.0000	Partial	2,602
CBG17249	X	6,700,976..6,710,486	-					
CBG24702	IV	12,296,015..12,299,717	+	N/A	0.0000	0.0000	Partial	6,636
CBG24882	UN	2,238,419..2,240,012	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG24168	UN	947,774..948,291	+	N/A	0.0000	0.0000	Partial	4,241
CBG21677	IV	5,072,997..5,076,000	+					
CBG17455	III	3,644,372..3,645,086	-	N/A	0.0000	0.0000	Complete	2,343
CBG25066	UN	2,844,759..2,845,638	-					
CBG20401	I	12,365,653..12,366,130	-	137,918	0.0000	0.0000	Complete	4,889
CBG20432	I	12,508,458..12,508,935	-					
CBG05335	X	17,379,583..17,379,816	+	7,640	0.0000	0.0000	Partial	366
CBG05334	X	17,372,091..17,374,343	-					
CBG24235	UN	1,073,960..1,074,232	-	N/A	0.0000	0.0000	Partial	1,214
CBG07191	II	15,703,946..15,707,124	-					
CBG15091	IV	4,211,261..4,214,638	-	N/A	0.0000	0.0000	Complete	6,102
CBG15092	I	3,230,380..3,232,395	-					
CBG12295	III	12,689,630..12,692,884	-	451	0.0000	0.0000	Complete	10,340
CBG12292	III	12,678,841..12,682,094	-					
CBG21908	I	6,867,663..6,868,978	+	7,044	0.0000	0.0000	Complete	1,765
CBG21906	I	6,859,213..6,859,770	-					
CBG25374	I	10,601,743..10,601,924	+	1,002	0.0000	0.0000	Complete	8,740
CBG25375	I	10,611,483..10,611,664	+					
CBG18444	II	1,780,070..1,780,441	-	70,856	0.0000	0.0000	Complete	5,596
CBG18452	II	1,856,616..1,856,987	-					
CBG05347	X	17,418,896..17,419,735	-	8,541	0.0000	0.0000	Partial	1,798
CBG05344	X	17,406,471..17,408,590	+					
CBG04923	V	15,696,054..15,696,502	+	1,002	0.0000	0.0000	Complete	8,003
CBG04926	V	15,705,057..15,705,505	+					
CBG13321	III	6,368,597..6,369,596	+	4,443	0.0000	0.0000	Complete	8,338
CBG13316	III	6,355,818..6,356,817	+					
CBG20004	IV	5,560,459..5,561,112	-	N/A	0.0000	0.0000	Partial	1,021
CBG24389	UN	1,351,207..1,351,551	+					
CBG12949	II	6,565,080..6,566,633	+	5,496	0.0000	0.0000	Complete	6,094
CBG12945	II	6,576,669..6,578,222	+					
CBG23931	I	8,002,238..8,003,053	+	419,214	0.0000	0.0000	Complete	6,107

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG12192	I	8,427,557..8,428,372	+					
CBG06598	V	7,000,908..7,001,234	+	15,221	0.0000	0.0000	Complete	8,297
CBG06608	V	7,024,424..7,024,645	+					
CBG03094	II	7,473,659..7,475,256	-	8,597	0.0000	0.0000	Complete	5,509
CBG03092	II	7,487,763..7,489,360	-					
CBG04881	V	15,598,117..15,598,555	-	70,656	0.0000	0.0000	Complete	5,962
CBG04856	V	15,521,217..15,521,655	+					
CBG19646	II	12,449,942..12,450,757	-	463,642	0.0000	0.0000	Complete	6,245
CBG19532	II	12,919,828..12,920,643	-					
CBG25649	II	10,546,189..10,546,769	+	432	0.0000	0.0000	Partial	864
CBG02334	II	10,544,722..10,545,753	-					
CBG20995	II	12,265,425..12,266,109	+	N/A	0.0000	0.0000	Complete	3,463
CBG24919	UN	2,366,844..2,367,528	+					
CBG01933	X	3,446,812..3,448,233	-	501	0.000041	0.004095	Partial	2,278
CBG01930	X	3,444,055..3,444,318	+					
CBG23652	UN	217,353..217,711	+	N/A	0.0000	0.0000	Complete	6,019
CBG11626	IV	3,147,647..3,148,005	+					
CBG13233	III	6,047,409..6,049,400	-	468,792	0.0000	0.0000	Complete	5,387
CBG18240	III	5,573,232..5,575,223	-					
CBG17261	X	6,620,002..6,621,196	-	N/A	0.0000	0.0000	Partial	1,735
CBG24079	UN	849,375..849,647	-					
CBG17140	X	7,267,003..7,267,245	-	114	0.0000	0.0000	Complete	5,958
CBG17143	X	7,260,921..7,261,163	-					
CBG04350	II	13,201,572..13,202,539	+	7,475	0.0000	0.0000	Partial	1,762
CBG04353	II	13,210,807..13,214,362	-					
CBG02425	II	10,198,494..10,198,978	-	6,540	0.0000	0.0000	Complete	2,685
CBG02427	II	10,191,081..10,191,565	+					
CBG27169	V	12,792,759..12,794,782	-	4,978	0.0000	0.0000	Complete	6,366
CBG19115	V	12,804,100..12,806,123	-					
CBG24396	UN	1,368,608..1,369,861	-	N/A	0.0000	0.0000	Complete	4,278
CBG08010	I	14,407,946..14,409,289	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG25476	II	2,614,119..2,615,423	-	N/A	0.0000	0.0000	Chimeric	480
CBG10136	III	9,815,878..9,822,035	-					
CBG23558	UN	70,524..71,182	+	N/A	0.0000	0.0000	Complete	1,418
CBG23560	UN	73,570..74,228	-					
CBG25177	I	1,579,111..1,579,371	-	3,901	0.0000	0.0000	Partial	626
CBG05109	I	1,570,852..1,574,821	+					
CBG23870	UN	619,882..620,902	+	N/A	0.0000	0.0000	Complete	3,992
CBG04267	II	4,929,238..4,930,219	+					
CBG15234	III	252,207..253,497	+	2,901	0.0000	0.0000	Partial	3,160
CBG15231	III	260,362..263,645	-					
CBG19212	V	13,145,071..13,146,091	-	519,796	0.017385	0.011939	Complete	13,397
CBG19053	V	12,612,240..12,613,260	-					
CBG08823	V	1,825,786..1,826,067	-	N/A	0.0000	0.0000	Partial	503
CBG24068	UN	831,017..832,519	+					
CBG27054	V	9,410,544..9,410,687	+	7,436	0.0000	0.0000	Complete	2,420
CBG27053	V	9,400,690..9,400,833	+					
CBG00634	II	3,875,762..3,876,411	+	N/A	0.0000	0.0000	Complete	5,145
CBG24969	UN	2,530,393..2,531,042	-					
CBG25022	UN	2,714,799..2,715,304	-	N/A	0.0000	0.0000	Partial	3,247
CBG16984	V	3,073,564..3,075,583	-					
CBG24453	III	13,434,185..13,436,261	+	N/A	0.0000	0.0000	Complete	5,251
CBG24900	UN	2,297,350..2,301,659	-					
CBG17823	V	17,743,012..17,745,062	-	N/A	0.0000	0.0000	Partial	10,881
CBG24865	UN	2,158,401..2,159,129	+					
CBG12673	I	9,838,324..9,840,719	-	1,343	0.0000	0.0000	Complete	5,640
CBG12671	I	9,831,343..9,833,870	-					
CBG18379	III	4,990,431..4,991,147	-	204,457	0.0000	0.0000	Partial	2,511
CBG18324	III	5,197,397..5,199,447	-					
CBG08031	I	14,513,703..14,513966	-	1,238	0.0000	0.0000	Complete	1,422
CBG08030	I	14,511,917..14,512,180	+					
CBG24798	UN	2,012,296..2,013,248	-	N/A	0.0000	0.0000	Complete	7,588

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG24300	UN	1,187,102..1,188,052	-					
CBG00538	II	4,142,761..4,142,959	-	12	0.0000	0.0000	Complete	4,272
CBG00540	II	4,147,043..4,147,241	-					
CBG22202	I	2,797,728..2,798,218	+	2,998	0.0000	0.0000	Complete	1,490
CBG22203	I	2,802,801..2,803,291	-					
CBG04885	V	15,605,645..15,606,508	+	83,427	0.0000	0.0000	Complete	6,785
CBG04852	V	15,512,699..15,513,562	-					
CBG02468	II	10,058,164..10,061,365	-	1,021	0.0000	0.0000	Complete	4,908
CBG02470	II	10,052,236..10,055,437	-					
CBG26557	IV	3,146,069..3,146,548	+	387,236	0.0000	0.0000	Partial	632
CBG22449	IV	2,754,781..2,758,682	+					
CBG05799	IV	7,981,769..7,983,078	+	N/A	0.0135	0.0169	Complete	2,967
CBG21419	II	13,650,472..13,651,795	-					
CBG20274	I	10,978,467..10,978,733	+	7,280	0.0222	0.0000	Complete	777
CBG20278	I	10,970,030..10,970,296	-					
CBG08583	V	8,109,682..8,111,738	+	858	0.0317	0.0043	Chimeric	481
CBG27014	V	8,108,439..8,108,824	+					
CBG24002	IV	10,107,871..10,108,140	-	9,213	0.0000	0.0000	Complete	1,627
CBG24003	IV	10,118,709..10,118,978	-					
CBG01844	IV	10,287,521..10,288,240	+	21,627	0.0000	0.0000	Complete	4,122
CBG01837	IV	10,313,268..10,313,987	+					
CBG07912	II	14,150,635..14,151,306	+	5,481	0.0000	0.0000	Complete	1,312
CBG07911	II	14,143,605..14,144,378	-					
CBG13343	II	3,264,893..3,267,532	+	370,423	0.0000	0.0000	Complete	6,829
CBG13423	II	2,886,990..2,889,629	-					
CBG20860	II	11,670,932..11,673,465	+	6,417	0.0000	0.0000	Partial	6,790
CBG20861	II	11,684,137..11,687,242	+					
CBG24392	UN	1,356,484..1,357,746	-	N/A	0.0000	0.0000	Partial	6,685
CBG20721	V	18,590,295..18,592,712	-					
CBG25390	I	11,017,988..11,018,216	-	8,249	0.0000	0.0000	Complete	749
CBG25394	I	11,026,985..11,027,212	-					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG04868	V	15,561,689..15,562,769	+	1,002	0.0000	0.0000	Complete	6,527
CBG04871	V	15,569,216..15,570,296	+					
CBG18592	V	17,436,515..17,439,160	-	70,295	0.0145	0.0088	Complete	4,462
CBG18605	V	17,363,925..17,366,161	+					
CBG24138	II	14,568,638..14,569,098	+	4,817	0.0000	0.0000	Partial	4,028
CBG24136	II	14,553,643..14,559,718	-					
CBG00584	III	14,173,045..14,176,109	+	399,958	0.0000	0.0000	Partial	2,713
CBG26111	III	14,577,612..14,578,851	+					
CBG17573	X	6,028,500..6,031,737	-	1,077	0.0001	0.0142	Chimeric	2,306
CBG17572	X	6,020,829..6,023,117	+					
CBG08315	IV	12,765,669..12,766,553	+	N/A	0.0932	0.0432	Partial	1,189
CBG19495	II	13,066,171..13,066,947	-					
CBG25513	II	4,797,062..4,799,041	+	408,358	0.0058	0.0133	Chimeric	1,108
CBG11242	II	5,208,112..5,217,952	-					
CBG23637	IV	3,479,967..3,489,571	+	44,641	0.0000	0.0000	Partial	6,051
CBG22665	IV	3,538,601..3,540,730	-					
CBG05204	III	4,566,278..4,567,895	-	35,639	0.0000	0.0000	Partial	3,581
CBG05213	III	4,527,060..4,527,590	-					
CBG14494	X	9,871,976..9,874,732	+	4,792	0.0000	0.0000	Complete	6,866
CBG14492	X	9,860,320..9,863,076	+					
CBG00857	II	1,136,862..1,137,819	-	N/A	0.0000	0.0000	Partial	2,550
CBG24477	UN	1,469,726..1,469,965	+					
CBG27304	V	16,763,528..16,764,048	-	1,360	0.0000	0.0000	Partial	2,510
CBG22911	V	16,765,548..16,766,260	+					
CBG22569	UN	29,625..30,029	+	N/A	0.0000	0.0000	Complete	5,222
CBG22568	UN	28,987..29,391	-					
CBG22961	III	13,965,253..13,974,293	-	N/A	0.0000	0.0000	Partial	4,725
CBG23773	UN	430,710..433,553	+					
CBG13798	IV	1,486,997..1,491,461	+	97,798	0.0000	0.0010	Chimeric	4,147
CBG13831	IV	1,589,967..1,593,563	+					
CBG11945	I	7,245,775..7,248,509	+	4,647	0.0000	0.0014	Partial	7,239

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG11942	I	7,196,249..7,236,624	+					
CBG24515	UN	1,501,500..1,506,652	-	N/A	0.0000	0.0000	Partial	6,113
CBG24468	V	17,864,590..17,870,799	+					
CBG20841	II	11,601,067..11,601,798	+	1,329	0.0000	0.0017	Complete	972
CBG20840	II	11,598,704..11,599,435	-					
CBG06889	III	10,234,807..10,235,580	+	1,014	0.0000	0.0021	Complete	1,027
CBG06888	III	10,236,967..10,237,740	-					
CBG13598	I	11,535,011..11,540,336	+	N/A	0.0000	0.0000	Partial	4,230
CBG24986	UN	2,608,357..2,612,274	+					
CBG11821	X	7,497,580..7,498,516	-	12,573	0.0000	0.0029	Complete	3,831
CBG11825	X	7,514,142..7,515,077	-					
CBG22905	V	16,790,128..16,791,129	+	388	0.0000	0.0024	Partial	4,104
CBG22904	V	16,794,619..16,797,129	+					
CBG05740	IV	8,177,933..8,178,355	-	1,667	0.0000	0.0028	Complete	1,721
CBG05739	IV	8,181,578..8,182,483	+					
CBG20472	I	6,758,167..6,760,183	+	790	0.0000	0.0030	Chimeric	2,226
CBG20473	I	6,754,939..6,756,898	+					
CBG25417	I	13,335,803..13,336,216	+	N/A	0.0000	0.0030	Complete	550
CBG24642	UN	1,722,009..1,722,422	-					
CBG22357	IV	2,881,802..2,882,659	-	1,017	0.0000	0.0032	Partial	7,347
CBG22359	IV	2,890,164..2,891,183	-					
CBG20948	II	12,099,536..12,100,626	+	720	0.0000	0.0033	Complete	1,849
CBG20950	II	12,102,691..12,103,781	-					
CBG14342	X	9,211,437..9,211,940	-	285	0.0000	0.0033	Complete	1,771
CBG14341	X	9,208,291..9,208,794	+					
CBG13957	IV	15,827,314..15,827,691	+	13,307,524	0.0000	0.0045	Complete	1,711
CBG18753	IV	2,516,907..2,518,931	-					
CBG08338	IV	12,675,462..12,675,824	+	N/A	0.0001	0.0051	Complete	1,284
CBG10673	X	21,174,240..21,174,602	-					
CBG24371	IV	10,126,197..10,265,502	-	N/A	0.0000	0.0046	Complete	3,504
CBG24825	X	681,322..681,579	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG04349	II	13,197,769..13,198,460	-	26,294	0.0000	0.0048	Partial	3,858
CBG04342	II	13,163,833..13,167,576	+					
CBG05375	IV	7,194,744..7,195,357	+	1,978	0.0001	0.0050	Complete	888
CBG05376	IV	7,191,704..7,192,326	-					
CBG24663	V	16,550,364..16,551,161	-	N/A	0.0001	0.0057	Complete	2,968
CBG25903	III	9,286,198..9,286,404	+					
CBG03050	II	7,627,771..7,628,701	+	595	0.0000	0.0000	Partial	810
CBG03052	II	7,625,867..7,626,183	-					
CBG18879	I	13,016,110..13,016,636	+	N/A	0.0001	0.0066	Complete	1,235
CBG23858	UN	570,392..570,922	+					
CBG08653	V	7,877,479..7,878,129	+	4,959,885	0.0001	0.0079	Complete	653
CBG22434	V	2,917,275..2,917,469	-					
CBG22935	III	14,108,814..14,110,168	+	5,784	0.0000	0.0045	Partial	574
CBG22932	III	14,116,003..14,116,356	+					
CBG01988	X	3,717,635..3,717,853	-	899	0.0001	0.0069	Partial	1,494
CBG01989	X	3,718,787..3,719,059	+					
CBG24994	UN	2,638,489..2,640,055	+	N/A	0.005343	0.004739	Partial	3,667
CBG00638	II	3,861,699..3,866,969	-					
CBG05428	IV	6,999,304..7,003,079	+	3,450	0.0000	0.0000	Partial	857
CBG05426	IV	7,007,007..7,007,615	-					
CBG15122	III	747,863..757,248	-	151	0.0001	0.0103	Partial	4,242
CBG26093	III	761,498..761,641	-					
CBG24817	UN	2,056,171..2,057,356	-	N/A	0.0000	0.0000	Partial	3,646
CBG10316	II	14,320,039..14,323,273	+					
CBG20945	II	12,090,439..12,093,514	-	341,365	0.0001	0.0138	Partial	878
CBG24975	II	12,435,071..12,435,579	+					
CBG02654	II	9,311,733..9,312,396	+	N/A	0.0002	0.0190	Partial	472
CBG05339	X	17,390,481..17,390,633	+					
CBG15903	V	16,697,434..16,697,637	-	422	0.0002	0.0222	Partial	307
CBG15904	V	16,698,137..16,698,585	+					
CBG16882	X	13,810,528..13,811,325	+	N/A	0.0000	0.0000	Partial	486

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG21572	IV	2,061,393..2,061,832	+					
CBG10405	V	3,637,310..3,637,787	-	N/A	0.0003	0.0268	Complete	1,609
CBG12242	III	12,400,402..12,400,894	-					
CBG19067	V	12,649,661..12,650,272	+	5,827,112	0.0003	0.0315	Complete	1,503
CBG06548	V	6,819,056..6,819,670	-					
CBG18332	III	5,181,841..5,182,164	+	N/A	0.0003	0.0316	Complete	745
CBG04793	V	15,312,236..15,312,559	-					
CBG15269	III	111,716..114,036	-	746,174	0.0000	0.0000	Partial	2,085
CBG15105	III	861,733..862,807	+					
CBG24770	UN	1,943,937..1,951,048	-	N/A	0.0000	0.0006	Partial	7,427
CBG24038	III [random]	206..2,651	+					
CBG18080	III	6,814,599..6,816,030	+	5,047,623	0.0037	0.0073	Complete	7,521
CBG00480	III	1,759,417..1,760,849	+					
CBG28015	X	17,112,489..17,113,406	+	N/A	0.0032	0.0000	Complete	10,324
CBG24109	UN	892,473..893,390	-					
CBG23482	V	8,567,749..8,568,773	+	N/A	0.0048	0.0049	Complete	3,619
CBG15876	I	13,076,312..13,077,334	+					
CBG04119	I	3,922,657..3,923,685	-	N/A	0.0046	0.0015	Complete	1,302
CBG24224	UN	1,025,602..1,026,637	-					
CBG07276	X	20,056,650..20,063,316	+	37,494	0.0047	0.0092	Partial	4,069
CBG07268	X	20,105,249..20,106,927	-					
CBG26257	IV	2,342,331..2,345,840	-	164	0.0083	0.0048	Complete	7,153
CBG26256	IV	2,335,298..2,338,357	-					
CBG26156	II	1,593,837..1,595,015	+	N/A	0.0059	0.0110	Complete	3,035
CBG27949	X	15,498,640..15,500,028	-					
CBG13610	I	11,571,514..11,572,845	-	167,345	0.0037	0.0038	Complete	5,278
CBG08458	I	11,402,122..11,403,450	+					
CBG25037	UN	2,783,462..2,784,057	-	N/A	0.0057	0.0000	Partial	2,396
CBG24795	UN	2,003,605..2,006,080	-					
CBG17686	IV	11,674,968..11,675,641	+	10,368,203	0.0066	0.0144	Complete	3,538
CBG13753	IV	1,303,226..1,303,899	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG17494	III	3,472,458..3,474,628	-	7,286,973	0.0061	0.0075	Partial	1,286
CBG24037	III	10,762,115..10,762,877	-					
CBG20790	II	11,386,481..11,388,535	-	3,506	0.0041	0.0041	Complete	6,431
CBG20787	II	11,377,079..11,378,598	-					
CBG01118	V	4,965,216..4,970,030	-	206,153	0.0086	0.0074	Complete	8,785
CBG01182	V	4,750,235..4,755,049	-					
CBG03690	II	2,485,102..2,485,653	-	27,879	0.0077	0.0024	Complete	2,850
CBG03697	II	2,514,819..2,515,784	+					
CBG21457	II	14,871,580..14,871,993	+	59,420	0.0164	0.0183	Complete	2,294
CBG21442	II	14,808,262..14,808,711	-					
CBG12887	I	10,804,493..10,804,891	+	N/A	0.0185	0.0035	Partial	826
CBG10260	X	20,751,723..20,760,652	+					
CBG08677	V	7,796,666..7,798,877	+	8,566	0.0000	0.0000	Complete	3,746
CBG08673	V	7,810,780..7,812,599	-					
CBG13235	III	6,052,409..6,053,641	+	4,279	0.0041	0.0000	Complete	3,264
CBG13231	III	6,040,306..6,043,109	-					
CBG16630	III	7,976,849..7,977,558	-	N/A	0.006916	0.005123	Complete	5,687
CBG24093	V	16,357,378..16,358,087	-					
CBG07328	X	19,838,836..19,842,035	-	N/A	0.0100	0.0024	Complete	5,136
CBG15607	IV	14,824,738..14,827,793	-					
CBG16819	IV	16,530,321..16,530,599	+	566	0.0164	0.0000	Partial	564
CBG16817	IV	16,527,752..16,529,496	-					
CBG03373	IV	9,562,446..9,564,615	-	N/A	0.0091	0.0012	Partial	5,613
CBG08954	X	20,842,869..20,844,369	-					
CBG27678	X	9,488,467..9,489,349	+	1,352	0.0099	0.0162	Partial	3,000
CBG14412	X	9,492,018..9,495,684	-					
CBG19483	II	13,106,114..13,107,904	+	788	0.0112	0.0147	Complete	3,819
CBG19485	II	13,103,357..13,105,104	-					
CBG22861	V	16,933,158..16,933,717	-	1,003	0.0080	0.0000	Complete	4,560
CBG22864	V	16,927,598..16,928,156	-					
CBG18155	III	6,504,068..6,504,679	-	601	0.0113	0.0143	Complete	1,102

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG18154	III	6,506,016..6,506,627	+					
CBG18260	III	5,485,285..5,491,241	-	488,105	0.0000	0.0000	Partial	4,608
CBG18378	III	4,992,895..4,997,094	+					
CBG10423	V	3,579,277..3,580,111	-	952	0.0117	0.0055	Complete	1,228
CBG10424	V	3,577,237..3,578,071	+					
CBG19895	IV	6,150,436..6,151,335	-	3,277	0.0128	0.0055	Complete	1,075
CBG19894	IV	6,154,757..6,155,656	+					
CBG09454	V	11,064,199..11,077,607	+	102	0.0050	0.0015	Partial	8,082
CBG29115	V	11,056,017..11,060,596	+					
CBG08153	X	492,771..493,106	+	6,549,134	0.0130	0.0357	Complete	1,032
CBG17189	X	7,042,699..7,043,034	-					
CBG19881	IV	6,195,649..6,197,488	-	2,690	0.0091	0.0000	Partial	1,457
CBG19883	IV	6,190,550..6,190,891	+					
CBG24983	UN	2,601,736..2,603,152	-	N/A	0.0000	0.0000	Partial	3,169
CBG24167	III	3,125,010..3,130,039	+					
CBG18660	V	17,184,279..17,185,443	+	N/A	0.0121	0.0448	Complete	2,924
CBG26701	UN	1,187,860..1,188,273	+					
CBG06057	IV	8,452,405..8,456,135	+	7,420	0.0168	0.0142	Partial	5,133
CBG06054	IV	8,437,813..8,443,584	+					
CBG21155	III	13,294,044..13,294,499	+	7,188,345	0.0144	0.0033	Partial	839
CBG13251	III	6,104,692..6,105,317	+					
CBG25008	UN	2,678,796..2,679,586	+	N/A	0.0175	0.0150	Partial	3,457
CBG22833	V	17,014,660..17,015,237	+					
CBG24916	UN	2,360,183..2,362,261	+	N/A	0.0000	0.0000	Partial	6,904
CBG12925	V	18,409,560..18,422,952	+					
CBG21997	V	5,548,689..5,549,998	+	3,457	0.0115	0.0000	Complete	1,664
CBG21999	V	5,553,586..5,554,895	-					
CBG12702	I	9,980,020..9,983,765	-	1,021	0.0154	0.0163	Partial	841
CBG12701	I	9,978,188..9,978,948	+					
CBG01767	IV	10,522,466..10,522,693	-	911	0.0160	0.0126	Complete	1,164
CBG01768	IV	10,520,392..10,520,619	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG24555	I	12,017,205..12,021,743	-	N/A	0.0162	0.0033	Patrial	3,243
CBG06341	V	6,106,642..6,108,406	-					
CBG05349	IV	15,379,631..15,386,475	-	N/A	0.0000	0.0000	Partial	6,269
CBG24006	UN	741,138..742,498	-					
CBG06364	V	6,176,875..6,177,054	-	909	0.0175	0.0000	Partial	342
CBG06366	V	6,178,124..6,179,523	-					
CBG18817	I	12,723,062..12,723,916	-	111,557	0.0152	0.0000	Complete	3,832
CBG18839	I	12,838,889..12,839,335	-					
CBG19549	II	12,841,658..12,841,897	-	N/A	0.0179	0.0056	Complete	484
CBG08537	V	8,266,969..8,267,208	+					
CBG22574	IV	3,684,961..3,685,478	-	1,856,982	0.0173	0.0360	Partial	728
CBG20010	IV	5,542,798..5,543,159	-					
CBG22842	V	16,973,041..16,973,998	+	231	0.0179	0.0084	Complete	1,571
CBG22843	V	16,971,664..16,972,623	-					
CBG24892	UN	2,273,319..2,275,382	-	N/A	0.0033	0.0011	Partial	6,808
CBG15209	III	340,236..343,550	+					
CBG06485	V	6,635,986..6,636,996	+	112	0.0000	0.0000	Partial	1,563
CBG06486	V	6,637,239..6,638,107	-					
CBG09979	III	9,249,024..9,250,121	+	54,786	0.0089	0.01595	Partial	5,818
CBG10003	III	9,309,626..9,312,196	+					
CBG23750	IV	8,365,205..8,368,291	+	1,136,174	0.0000	0.0008	Complete	3,583
CBG06036	IV	7,225,450..7,228,472	+					
CBG19394	V	13,922,715..13,927,407	+	N/A	0.0000	0.0000	Partial	3,038
CBG25032	UN	2,758,625..2,760,600	-					
CBG02825	II	8,550,980..8,551,416	-	N/A	0.0192	0.0044	Complete	1,479
CBG01434	V	3,900,067..3,904,699	+					
CBG29116	III	4,797,081..4,797,513	+	103	0.019582	0.01832	Partial	4,955
CBG24619	III	4,802,121..4,805,499	+					
CBG21261	IV	13,543,031..13,544,369	-	N/A	0.0225	0.0157	Partial	840
CBG19576	II	12,727620..12,727,853	-					
CBG20991	II	12,253,968..12,255,151	-	4,443	0.0000	0.0000	Chimeric	479

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG20994	II	12,260,077..12,264,773	+					
CBG10582	X	21,517,870..21,522,461	+	14,902,985	0.0179	0.0236	Partial	3,656
CBG04471	X	6,611,608..6,612,681	-					
CBG08521	I	11,118,669..11,119,264	-	738	0.0165	0.0000	Partial	293
CBG08520	I	11,120,013..11,120,238	+					
CBG02428	II	10,188,376..10,189,031	-	3,364	0.0000	0.0029	Partial	3,346
CBG02426	II	10,195,084..10,196,484	-					
CBG09441	V	11,128,812..11,133,469	-	370	0.0000	0.0000	Partial	5,661
CBG09439	V	11,134,840..11,144,427	-					
CBG23485	V	8,563,132..8,563,658	-	N/A	0.0264	0.0000	Partial	566
CBG24705	UN	1,805,785..1,806,136	+					
CBG27680	X	9,525,927..9,526,694	-	1,763,393	0.0000	0.0000	Partial	1,108
CBG14039	X	7,756,924..7,762,394	+					
CBG24867	UN	2,168,162..2,169,124	+	N/A	0.0000	0.0000	Partial	4,496
CBG08174	X	603,296..605,585	+					
CBG23480	V	8,571,690..8,572,717	+	N/A	0.02604	0.010511	Complete	5,599
CBG15874	I	13,080,452..13,081,408	+					
CBG05734	IV	8,210,217..8,211,250	-	N/A	0.0226	0.0281	Complete	1,302
CBG06875	III	10,274,733..10,275,766	+					
CBG08302	IV	12,813,449..12,813,745	-	N/A	0.0207	0.0258	Complete	1,053
CBG06619	V	7,059,633..7,059,890	+					
CBG18712	II	1,446,228..1,447,181	+	9,566	0.0321	0.0102	Complete	2,149
CBG18708	II	1,433,430..1,434,251	-					
CBG19931	IV	5,952,521..5,953,513	+	3,777	0.0233	0.0191	Complete	1,316
CBG19932	IV	5,947,676..5,948,131	-					
CBG17828	V	17,755,291..17,755,615	-	10,955,241	0.0210	0.0000	Complete	6,517
CBG06540	V	6,796,051..6,796,687	+					
CBG24257	UN	1,117,706..1,118,790	-	N/A	0.0000	0.0000	Chimeric	474
CBG24256	UN	1,116,268..1,117,419	+					
CBG25014	UN	2,698,097..2,698,786	+	N/A	0.0241	0.0176	Patrial	2,239
CBG14550	X	10,151,669..10,154,842	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG12379	I	8,863,815..8,866,875	-	N/A	0.0246	0.0306	Complete	4,609
CBG18622	V	17,305,566..17,308,598	+					
CBG23971	UN	721,715..723,792	-	N/A	0.0238	0.0065	Partial	1,909
CBG23970	UN	713,117..714,570	-					
CBG06899	III	10,189,284..10,189,778	-	1,826	0.0334	0.0517	Complete	1,168
CBG06898	III	10,192,077..10,192,595	-					
CBG21014	II	12,336,925..12,337,497	+	1,250	0.0262	0.0174	Complete	807
CBG21012	II	12,334,624..12,335,196	-					
CBG13850	IV	1,666,878..1,667,705	-	2,491,911	0.0162	0.0037	Chimeric	873
CBG15079	IV	4,159,685..4,161,218	-					
CBG24810	II	14,527,263..14,528,387	-	1,471	0.0277	0.0394	Complete	2,457
CBG21869	II	14,523,335..14,524,459	-					
CBG20274	I	10,978,467..10,978,733	+	7,280	0.0222	0.0000	Complete	777
CBG20278	I	10,970,030..10,970,296	-					
CBG24763	V	8,323,412..8,323,825	+	284,203	0.0000	0.0000	Partial	3,399
CBG08605	V	8,034,475..8,036,225	+					
CBG17449	X	5,665,978..5,670,096	-	26,901	0.0325	0.0272	Complete	6,293
CBG17504	X	5,699,170..5,703,289	-					
CBG09186	III	12,028,459..12,028,611	-	1,197	0.0247	0.0096	Complete	224
CBG09187	III	12,029,878..12,030,030	-					
CBG25073	UN	2,860,624..2,861,642	-	N/A	0.0000	0.0000	Partial	2,039
CBG21498	IV	1,769,811..1,774,433	+					
CBG16454	IV	14,613,019..14,616,734	+	338	0.0000	0.0000	Chimeric	311
CBG16453	IV	14,612,358..14,612,672	-					
CBG07487	X	18,968,622..18,969,161	+	2,578	0.0000	0.0000	Complete	2,765
CBG07490	X	18,963,281..18,963,547	+					
CBG14016	X	7,669,223..7,672,442	+	N/A	0.0329	0.0245	Partial	1,184
CBG07839	II	13,872,910..13,873,466	-					
CBG10434	V	3,529,806..3,532,426	-	N/A	0.0361	0.0291	Complete	3,355
CBG25013	UN	2,694,995..2,697,798	-					
CBG11564	V	14,129,371..14,129,585	-	351	0.0380	0.0303	Partial	665

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG11563	V	14,127,708..14,129,016	-					
CBG19974	IV	5,736,333..5,736,854	-	N/A	0.0367	0.0419	Complete	1,103
CBG23852	I	12,003,477..12,003,997	+					
CBG09125	III	11,712,582..11,713,175	-	5,193	0.0380	0.0042	Complete	1,141
CBG09126	III	11,718,944..11,719,555	-					
CBG15703	III	1,150,128..1,161,425	-	N/A	0.0000	0.0000	Partial	6,737
CBG24914	UN	2,352,681..2,355,192	-					
CBG10446	V	3,485,206..3,486,400	-	496	0.0000	0.0000	Partial	1,970
CBG10445	V	3,487,422..3,491,951	-					
CBG00949	II	522,093..523,034	-	N/A	0.0391	0.0206	Complete	3,284
CBG08187	X	654,271..655,065	-					
CBG21590	IV	2,141,858..2,143,947	+	1,863,723	0.0347	0.0088	Complete	12,211
CBG15048	IV	4,019,641..4,024,466	-					
CBG25344	I	9,285,017..9,285,351	-	1,532	0.0000	0.0000	Chimeric	382
CBG12516	I	9,286,941..9,291,504	+					
CBG20446	II	14,711,152..14,712,282	+	N/A	0.0394	0.0496	Complete	3,680
CBG23830	UN	549,096..550,227	+					
CBG04301	II	5,075,786..5,076,223	-	1,476	0.0399	0.0188	Complete	744
CBG04302	II	5,078,133..5,078,570	+					
CBG26920	V	1,587,566..1,588,668	-	498	0.0424	0.0234	Partial	1,169
CBG00831	V	1,585,488..1,587,069	-					
CBG13639	I	11,725,993..11,734,337	-	N/A	0.0000	0.0000	Partial	464
CBG24048	UN	825,142..825,450	-					
CBG24951	UN	2,469,090..2,470,400	+	N/A	0.0198	0.0075	Partial	2,565
CBG24657	V	16,509,571..16,512,224	+					
CBG19217	V	13,159,747..13,163,186	-	1,698	0.0283	0.0032	Partial	1,039
CBG19216	V	13,157,069..13,158,002	-					
CBG22115	I	2,416,303..2,426,644	-	785,644	0.0025	0.0008	Partial	8,453
CBG22304	I	3,219,616..3,222,299	-					
CBG06050	IV	8,418,418..8,419,129	-	5,223	0.067185	0.040629	Complete	1,737
CBG06047	IV	8,412,586..8,412,944	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG10517	V	3,252,675..3,253,379	+	2,893	0.0973	0.0709	Complete	900
CBG10518	V	3,248,786..3,249,499	-					
CBG17970	III	7,245,127..7,245,342	-	N/A	0.0474	0.0215	Complete	352
CBG05377	IV	7,190,570..7,190,869	+					
CBG01859	X	3,067,421..3,068,354	+	2,690	0.0411	0.0000	Partial	676
CBG01858	X	3,064,057..3,064,732	-					
CBG19434	V	14,045,374..14,047,157	+	602,027	0.014844	0.006881	Complete	6,948
CBG19274	V	13,436,400..13,438,238	+					
CBG20237	II	1,428,654..1,428,827	-	11,690,821	0.0375	0.0336	Partial	283
CBG04334	II	13,119,666..13,129,981	+					
CBG05481	IV	6,792,540..6,794,285	-	41,826	0.0000	0.0000	Chimeric	6,165
CBG05467	IV	6,842,274..6,847,551	-					
CBG01045	II	66,162..69,975	+	0	0.0000	0.0000	Chimeric	5,071
CBG01044	II	71,581..76,046	+					
CBG12676	I	9,853,156..9,854,117	+	18,320	0.0000	0.0036	Complete	2,081
CBG25360	I	9,873,257..9,873,706	+					
CBG06365	V	6,177,398..6,177,988	+	2,002	0.0528	0.0001	Complete	906
CBG06367	V	6,180,304..6,182,141	+					
CBG05799	IV	7,981,769..7,983,078	+	N/A	0.0746	0.0677	Complete	3,833
CBG21419	II	13,650,472..13,651,795	-					
CBG21420	II	13,654,706..13,655,869	+	N/A	0.0552	0.0703	Complete	2,514
CBG05796	IV	7,987,605..7,988,769	-					
CBG26395	IV	9,840,061..9,840,306	+	2,786	0.0518	0.0348	Partial	372
CBG03446	IV	9,835,199..9,837,164	-					
CBG15084	IV	4,184,470..4,184,817	+	914,610	0.0526	0.0329	Complete	3,562
CBG21685	IV	5,102,640..5,104,091	+					
CBG05276	X	17,115,446..17,119,591	-	N/A	0.0469	0.0088	Complete	8,069
CBG16655	III	8,057,822..8,061,895	-					
CBG05863	IV	7,742,375..7,748,310	-	N/A	0.0002	0.0174	Partial	1,652
CBG16417	X	772,174..772,344	+					
CBG26522	IV	16,438,646..16,438,881	+	N/A	0.0000	0.0000	Partial	319

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG18911	V	12,052,615..12,055,340	+					
CBG07521	X	18,841,021..18,841,548	-	N/A	0.0563	0.0557	Partial	998
CBG07805	II	13,764,060..13,766,255	+					
CBG02709	II	9,064,327..9,064,957	+	N/A	0.0485	0.0604	Partial	2,338
CBG20107	IV	5,145,701..5,147,834	+					
CBG10791	I	6,443,125..6,443,971	-	114,529	0.0498	0.0065	Partial	800
CBG10768	I	6,327,791..6,328,597	+					
CBG04656	V	14,803,833..14,804,024	-	1,416	0.0406	0.0380	Partial	300
CBG04658	V	14,805,547..14,805,917	-					
CBG18591	V	17,440,707..17,445,741	+	76,643	0.0561	0.0152	Complete	5,483
CBG18606	V	17,357,827..17,362,356	-					
CBG24124	UN	930,380..930,890	+	N/A	0.0562	0.0258	Complete	1,055
CBG19817	I	15,222,649..15,223,158	-					
CBG12855	I	10,671,324..10,672,018	-	N/A	0.0387	0.0165	Chimeric	3,179
CBG00165	X	14,663,306..14,665,954	-					
CBG26226	IV	803,147..803,533	+	12,755	0.0581	0.0430	Partial	552
CBG26225	IV	789,293..790,235	-					
CBG15593	X	15,249,517..15,250,011	-	399,417	0.0000	0.0000	Partial	546
CBG15502	X	15,649,475..15,657,637	-					
CBG18604	V	17,366,445..17,367,495	-	196,981	0.0099	0.0167	Chimeric	883
CBG27311	V	17,167,835..17,168,710	+					
CBG04252	II	4,828,494..4,838,406	-	0	0.0000	0.0000	Chimeric	5,596
CBG04249	II	4,823,258..4,827,933	-					
CBG14812	X	11,420,486..11,420,581	-	70,462	0.00437	0.002693	Complete	6,426
CBG14795	X	11,342,264..11,343,695	-					
CBG10074	III	9,554,810..9,557,116	+	543	0.0000	0.0000	Chimeric	1,778
CBG10073	III	9,550,645..9,552,568	-					
CBG24720	UN	1,831,727..1,837,191	-	N/A	0.0000	0.0015	Partial	5,192
CBG24768	V	8,334,964..8,339,755	-					
CBG10216	X	20,565,484..20,567,283	-	1,254	0.0000	0.0000	Complete	4,438
CBG10214	X	20,560,272..20,561,967	+					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG04851	V	15,502,930..15,510,068	-	45,290	0.0000	0.0000	Partial	2,623
CBG04838	V	15,455,880..15,456,826	-					
CBG21711	IV	4,791,138..4,791,509	-	343	0.0636	0.0054	Complete	438
CBG21710	IV	4,791,955..4,792,321	+					
CBG16031	II	15,289,661..15,290,999	+	372	0.0628	0.0800	Complete	614
CBG16032	II	15,291,420..15,291,998	-					
CBG17030	V	3,216,158..3,217,239	-	N/A	0.0578	0.0384	Complete	2,795
CBG24205	UN	975,630..976,893	-					
CBG11901	I	416,982..417,224	-	N/A	0.0726	0.0160	Complete	570
CBG25067	UN	2,846,005..2,846,178	-					
CBG08516	I	11,131,416..11,133,442	+	685	0.0721	0.0066	Chimeric	660
CBG08517	I	11,129,662..11,129,925	-					
CBG04592	V	14,583,006..14,583,494	+	52,153	0.071003	0.02442	Partial	1,823
CBG04607	V	14,635,646..14,636,164	-					
CBG04636	V	14,764,022..14,765,147	-	N/A	0.0621	0.0247	Complete	6,336
CBG19949	IV	5,880,019..5,881,129	-					
CBG24364	UN	1,336,418..1,336,771	-	N/A	0.0303	0.0188	Complete	1,911
CBG13327	III	6,387,354..6,388,646	-					
CBG00870	II	1,019,210..1,054,689	+	N/A	0.0000	0.0000	Partial	6,016
CBG24954	UN	2,483,029..2,486,348	+					
CBG04698	V	14,907,246..14,908,067	+	604	0.0676	0.0451	Complete	2,399
CBG04695	V	14,902,728..14,903,732	-					
CBG18448	II	1,817,302..1,818,680	-	N/A	0.0002	0.0149	Chimeric	819
CBG22853	V	16,951,703..16,952,344	-					
CBG16381	X	933,858..938,788	-	172,894	0.0000	0.0000	Partial	7,869
CBG16347	X	1,117,596..1,119,398	-					
CBG15892	V	16,662,460..16,664,819	-	76,987	0.0000	0.0000	Partial	6,759
CBG24530	V	16,746,752..16,748,563	-					
CBG18376	III	5,009,481..5,010,434	-	118,719	0.0500	0.0570	Complete	2,377
CBG18345	III	5,131,110..5,132,090	+					
CBG07220	II	15,555,390..15,558,638	+	N/A	0.0682	0.0223	Partial	1,205

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG20314	III	2,294,372..2,924,741	-					
CBG20418	I	12,453,487..12,455,648	-	1,268	0.0734	0.0360	Complete	2,180
CBG20419	I	12,457,599..12,459,432	+					
CBG18729	IV	2,360,056..2,361,768	+	N/A	0.0782	0.0682	Chimeric	1,567
CBG00898	II	836,950..841,570	-					
CBG05787	IV	8,009,652..8,014,265	+	1,013	0.0343	0.0000	Chimeric	644
CBG26361	IV	8,015,784..8,016,904	-					
CBG22303	I	3,212,459..3,219,426	-	55,730	0.0000	0.0000	Chimeric	7,257
CBG15102	I	3,275,155..3,278,497	+					
CBG04801	V	15,331,323..15,331,719	+	31,887	0.0731	0.0179	Complete	661
CBG04817	V	15,363,872..15,364,268	+					
CBG26278	IV	4,026,721..4,030,156	+	N/A	0.0758	0.0661	Complete	4,114
CBG27951	X	15,502,685..15,502,810	-					
CBG01243	V	4,545,898..4,547,926	-	1,999	0.0399	0.0429	Chimeric	2,352
CBG01245	V	4,540,181..4,543,168	+					
CBG25007	UN	2,665,308..2,667,959	+	N/A	0.0593	0.0510	Partial	3,759
CBG21359	II	13,384,222..13,388,555	-					
CBG18649	V	17,220,127..17,220,921	+	16,142	0.073832	0.066629	Complete	6,077
CBG18655	V	17,197,859..17,198,654	+					
CBG02900	II	8,284,207..8,286,129	+	7,772	0.0000	0.0009	Partial	2,038
CBG02898	II	8,294,085..8,301,241	-					
CBG24931	UN	2,419,963..2,424,032	+	N/A	0.0590	0.0208	Partial	4,648
CBG05725	IV	8,243,796..8,245,589	+					
CBG22684	X	16,808,604..16,810,793	-	5,160	0.0162	0.0024	Complete	1,763
CBG22683	X	16,816,396..16,816,937	+					
CBG20983	II	12,230,782..12,232,616	+	668	0.0426	0.0192	Patrial	655
CBG20984	II	12,233,284..12,233,715	-					
CBG07269	X	20,103,081..20,103,797	+	1,697	0.0000	0.0000	Chimeric	1,490
CBG07271	X	20,096,337..20,100,589	+					
CBG09964	III	9,204,050..9,205,083	-	1,328	0.0000	0.0000	Partial	3,543
CBG09966	III	9,208,609..9,211,310	-					

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG03701	I	5,613,461..5,615,829	-	5,187	0.0000	0.0000	Partial	7,133
CBG03760	I	5,601,142..5,601,396	-					
CBG18674	V	17,143,753..17,149,390	-	N/A	0.0000	0.0000	Partial	7,510
CBG24902	UN	2,313,449..2,314,856	+					
CBG24879	UN	2,227,431..2,228,087	+	N/A	0.0000	0.0000	Complete	9,339
CBG04971	V	15,823,157..15,823,813	+					
CBG00856	II	1,139,026..1,139,485	-	76,817	0.0000	0.0000	Partial	5,672
CBG00845	II	1,221,513..1,223,697	-					
CBG21481	II	14,979,856..14,986,860	+	N/A	0.0000	0.0000	Chimeric	2,830
CBG24535	IV	15,509,243..15,512,565	-					
CBG16742	X	13,179,691..13,184,250	-	2,414	0.0350	0.0096	Partial	1,187
CBG16741	X	13,176,234..13,176,800	-					
CBG11861	I	176,406..177,846	+	1,002	0.0000	0.0000	Partial	4,963
CBG11853	I	118,887..171,882	+					
CBG27527	X	5,125,939..5,130,331	+	0	0.0000	0.0000	Chimeric	3,681
CBG17336	X	5,109,342..5,125,650	+					
CBG03230	II	6,919,565..6,920,135	+	1,201,114	0.0000	0.0034	Complete	5,703
CBG11214	II	5,712,515..5,713,321	+					
CBG25153	I	921,231..924,230	-	44,934	0.0000	0.0000	Partial	1,925
CBG19473	I	874,503..874,944	-					
CBG24788	UN	1,995,533..2,000,484	-	N/A	0.00507	0.003913	Partial	12,331
CBG11320	V	750,843..763,241	-					
CBG24833	UN	2,083,161..2,090,222	+	N/A	0.0026	0.0000	Partial	1,518
CBG23992	UN	732,654..733,973	+					
CBG06735	V	7,511,466..7,513,069	+	422,250	0.0000	0.0000	Partial	10,564
CBG06625	V	7,078,026..7,080,876	+					
CBG21090	III	3,178,220..3,179,616	+	12	0.0000	0.0000	Complete	7,476
CBG21094	III	3,186,070..3,187,104	+					
CBG23634	IV	3,467,735..3,473,364	+	13,188	0.0053	0.0000	Partial	7,743
CBG24733	IV	3,440,297..3,441,160	-					
CBG17838	V	17,780,504..17,781,290	-	0	0.0000	0.0000	Chimeric	3,512

Paralogs	Chr	Coordinates (approx)	Orientation	Genomic Distance	K_S	K_A	Structure	Min Dup Span
CBG17837	V	17,776,367..17,778,726	-					
CBG10679	X	21,154,415..21,162,695	-	27,129	0.01485	0.01115	Partial	7,858
CBG10664	X	21,195,156..21,197,274	-					
CBG15116	III	795,602..796,042	-	N/A	0.0245	0.0167	Complete	5,703
CBG24824	X	679,715..680,179	+					
CBG13823	IV	569,032..570,317	+	1,194	0.0000	0.0034	Chimeric	2,248
CBG13825	IV	575,304..577,762	-					
CBG11903	I	7,028,439..7,030,341	-	848,857	0.0000	0.0000	Complete	7,052
CBG12104	I	7,884,969..7,886,248	-					
CBG19694	III	3,710,562..3,711,379	+	863	0.0000	0.0031	Chimeric	1,324
CBG19693	III	3,713,322..3,714,835	-					
CBG20696	II	11,168,867..11,170,032	-	282	0.0001	0.0108	Chimeric	593
CBG20697	II	11,170,313..11,171,916	+					

Table 2. – Linked sets containing two, three and four duplicate pairs. The gene duplicate pair in bold at the beginning of each block was the representative pair used in the above list. K_A, K_S, and K_A/K_S values were calculated using a contiguous alignment of all genes in each linked set. Structural heterogeneity of genes in linked sets was incorporated into analysis of the frequencies of different structural resemblance categories of gene duplicates within this genome.

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
1	**CBG21453**	II	14,858,839..14,859,078	+	Complete
	CBG21448	II	14,843,058..14,843,297	-	
2	CBG21447	II	14,838,649..14,841,630	+	Complete
	CBG21454	II	14,860,506..14,863,487	-	
1	**CBG06604**	V	7,014,446..7,015,600	-	Complete
	CBG06339	V	6,102,800..6,103,954	-	
2	CBG06340	V	6,104,972..6,105,681	+	Complete
	CBG06605	V	7,016,618..7,017,327	+	
1	**CBG25022**	UN	2,714,799..2,715,304	-	Partial
	CBG16984	V	3,073,564..3,075,583	-	
2	CBG16983	V	3,071,191..3,073,175	-	Complete
	CBG25021	UN	2,712,426..2,714,410	-	
1	**CBG14798**	X	11,353,104..11,353,327	-	Complete
	CBG14697	X	10,857,406..10,857,629	-	
2	CBG14797	X	11,349,866..11,350,108	-	Complete
	CBG14696	X	10,854,169..10,854,411	-	
1	**CBG08109**	X	290,212..291,866	+	Partial
	CBG26680	UN	812,843..813,127	+	
2	CBG24030	UN	813,567..815,156	+	Complete
	CBG08110	X	292,306..293,895	+	
1	**CBG24994**	UN	2,638,489..2,640,055	+	Partial
	CBG00638	II	3,861,699..3,866,969	-	
2	CBG00639	II	3,860,811..3,861,091	+	Complete
	CBG24995	UN	2,640,663..2,640,943	-	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
1	**CBG20790**	II	11,386,481..11,388,535	-	Complete
	CBG20787	II	11,377,079..11,378,598	-	
2	CBG20788	II	11,379,131..11,380,865	-	Complete
	CBG20791	II	11,389,068..11,390,802	-	
1	**CBG10599**	X	21,443,798..21,444,174	+	Complete
	CBG10600	X	21,441,336..21,442,177	-	
2	CBG10602	X	21,439,372..21,439,781	-	Complete
	CBG10597	X	21,445,661..21,446,034	+	
1	**CBG06506**	V	6,694,877..6,698,501	-	Complete
	CBG06509	V	6,702,526..6,706,150	-	
2	CBG06508	V	6,699,670..6,702,450	+	Partial
	CBG06505	V	6,691,446..6,694,801	+	
1	**CBG23870**	UN	619,882..620,902	+	Complete
	CBG04267	II	4,929,238..4,930,219	+	
2	CBG23869	UN	618,352..619,267	-	Complete
	CBG04266	II	4,927,670..4,928,584	-	
1	**CBG24702**	IV	12,296,015..12,299,717	+	Partial
	CBG24882	UN	2,238,419..2,240,012	+	
2	CBG24883	UN	2,242,488..2,244,267	+	Complete
	CBG24704	IV	12,302,193..12,303,140	+	
1	**CBG09731**	V	10,159,324..10,160,198	+	Partial
	CBG09743	V	10,122,224..10,122,827	+	
2	CBG09730	V	10,161,270..10,161,529	-	Complete
	CBG27068	V	10,123,899..10,124,158	-	
1	**CBG02468**	II	10,058,164..10,061,365	-	Complete
	CBG02470	II	10,052,236..10,055,437	-	
2	CBG02469	II	10,055,470..10,056,692	+	Complete
	CBG02467	II	10,061,398..10,062,620	+	
1	**CBG04923**	V	15,696,054..15,696,502	+	Complete
	CBG04926	V	15,705,057..15,705,505	+	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
2	CBG04924	V	15,698,402..15,699,658	+	Complete
	CBG04927	V	15,707,405..15,708,661	+	
1	**CBG22318**	IV	10,259,010..10,260,040	-	Complete
	CBG24368	IV	10,120,109..10,121,139	-	
2	CBG22320	IV	10,262,959..10,263,302	-	Complete
	CBG24370	IV	10,124,058..10,124,401	-	
1	**CBG13089**	II	5,858,808..5,859,842	+	Partial
	CBG26633	UN	66,943..67,050	-	
2	CBG23555	UN	64,536..66,816	+	Complete
	CBG13090	II	5,859,969..5,862,280	-	
1	**CBG26109**	III	14,401,969..14,402,053	+	Complete
	CBG26110	III	14,410,674..14,410,758	+	
2	CBG15326	III	14,398,675..14,401,510	-	Complete
	CBG15328	III	14,407,380..14,410,215	-	
1	**CBG24168**	UN	947,774..948,291	+	Partial
	CBG21677	IV	5,072,997..5,076,000	+	
2	CBG21678	IV	5,078,231..5,085,180	+	Partial
	CBG24170	UN	950,523..951,109	+	
1	**CBG20401**	I	12,365,653..12,366,130	-	Complete
	CBG20432	I	12,508,458..12,508,935	-	
2	CBG20402	I	12,366,863..12,367,904	+	Complete
	CBG20433	I	12,509,669..12,510,710	+	
1	**CBG19434**	V	14,045,374..14,047,157	+	Complete
	CBG19274	V	13,436,400..13,438,238	+	
2	CBG19273	V	13,435,029..13,436,171	-	Complete
	CBG19433	V	14,044,003..14,045,145	-	
1	**CBG04881**	V	15,598,117..15,598,555	-	Complete
	CBG04856	V	15,521,217..15,521,655	+	
2	CBG04857	V	15,521,777..15,522,959	-	Complete
	CBG04880	V	15,596,813..15,597,995	+	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
1	**CBG14812**	X	11,420,486..11,420,581	-	Complete
	CBG14795	X	11,342,264..11,343,695	-	
2	CBG14811	X	11,416,953..11,418,666	+	Complete
	CBG14794	X	11,340,067..11,341,780	+	
1	**CBG27054**	V	9,410,544..9,410,687	+	Complete
	CBG27053	V	9,400,690..9,400,833	+	
2	CBG23257	V	9,409,721..9,410,381	-	Complete
	CBG23259	V	9,399,867..9,400,527	-	
1	**CBG00634**	II	3,875,762..3,876,411	+	Complete
	CBG24969	UN	2,530,393..2,531,042	-	
2	CBG24970	UN	2,531,274..2,532,259	-	Complete
	CBG00635	II	3,874,545..3,875,530	+	
1	**CBG17823**	V	17,743,012..17,745,062	-	Partial
	CBG24865	UN	2,158,401..2,159,129	+	
2	CBG24861	UN	2,150,778..2,151,796	+	Complete
	CBG17826	V	17,751,795..17,752,813	-	
1	**CBG18379**	III	4,990,431..4,991,147	-	Partial
	CBG18324	III	5,197,397..5,199,447	-	
2	CBG18380	III	4,989,537..4,990,170	+	Complete
	CBG18325	III	5,196,503..5,197,136	+	
1	**CBG24763**	V	8,323,412..8,323,825	+	Partial
	CBG08605	V	8,034,475..8,036,225	+	
2	CBG24764	V	8,323,978..8,325,049	-	Complete
	CBG08604	V	8,036,378..8,037,449	-	
1	**CBG24392**	UN	1,356,484..1,357,746	-	Partial
	CBG20721	V	18,590,295..18,592,712	-	
2	CBG25671	V	18,589,572..18,589,805	-	Complete
	CBG24391	UN	1,355,049..1,356,024	-	
1	**CBG05481**	IV	6,792,540..6,794,285	-	Chimeric
	CBG05467	IV	6,842,274..6,847,551	-	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
2	CBG05466	IV	6,849,687..6,852,169	+	Complete
	CBG05480	IV	6,796,428..6,798,910	+	
1	**CBG24318**	II	14,626,651..14,646,422	+	Partial
	CBG24958	UN	2,493,433..2,495,626	+	
2	CBG24957	UN	2,495,956..2,497,253	-	Complete
	CBG24321	II	14,646,752..14,648,048	-	
1	**CBG23480**	V	8,571,690..8,572,717	+	Complete
	CBG15874	I	13,080,452..13,081,408	+	
2	CBG23478	V	8,574,868..8,575,229	-	Complete
	CBG15872	I	13,083,525..13,083,886	-	
1	**CBG29116**	III	4,797,081..4,797,513	+	Partial
	CBG24619	III	4,802,121..4,805,499	+	
2	CBG24620	III	4,798,858..4,801,667	-	Complete
	CBG23800	III	4,793,836..4,796,612	-	
1	**CBG06050**	IV	8,418,418..8,419,129	-	Complete
	CBG06047	IV	8,412,586..8,412,944	+	
2	CBG06051	IV	8,419,403..8,420,331	-	Complete
	CBG06046	IV	8,411,421..8,412,114	+	
1	**CBG04592**	V	14,583,006..14,583,494	+	Partial
	CBG04607	V	14,635,646..14,636,164	-	
2	CBG04591	V	14,582,163..14,582,654	-	Complete
	CBG04608	V	14,636,509..14,636,997	+	
1	**CBG18649**	V	17,220,127..17,220,921	+	Complete
	CBG18655	V	17,197,859..17,198,654	+	
2	CBG18656	V	17,195,219..17,196,345	-	Complete
	CBG18650	V	17,217,486..17,218,613	-	
1	**CBG11945**	I	7,245,775..7,248,509	+	Partial
	CBG11942	I	7,196,249..7,236,624	+	
2	CBG11944	I	7,237,071..7,240,024	+	Complete
	CBG11946	I	7,248,956..7,251,909	+	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
1	**CBG01933**	X	3,446,812..3,448,233	-	Partial
	CBG01930	X	3,444,055..3,444,318	+	
2	CBG01932	X	3,445,823..3,446,639	-	Complete
	CBG01931	X	3,444,491..3,445,290	+	
1	**CBG00584**	III	14,173,045..14,176,109	+	Partial
	CBG26111	III	14,577,612..14,578,851	+	
2	CBG00583	III	14,171,833..14,172,440	+	Partial
	CBG24156	III	14,576,066..14,577,007	+	
1	**CBG08173**	X	597,001..600,474	-	Partial
	CBG08190	X	660,954..666,947	-	
2	CBG08172	X	594,341..594,780	-	Complete
	CBG08189	X	658,295..658,734	-	
3	CBG08188	X	656,691..657,729	+	Complete
	CBG08171	X	592,905..593,775	+	
1	**CBG02263**	II	10,804,607..10,804,870	+	Complete
	CBG24807	UN	2,049,847..2,050,110	-	
2	CBG02264	II	10,802,677..10,803,474	-	Complete
	CBG24808	UN	2,051,242..2,052,039	+	
3	CBG24809	UN	2,053,709..2,054,604	+	Complete
	CBG02265	II	10,800,112..10,801,007	-	
1	**CBG09979**	III	9,249,024..9,250,121	+	Partial
	CBG10003	III	9,309,626..9,312,196	+	
2	CBG09980	III	9,250,225..9,251,091	-	Complete
	CBG10004	III	9,310,584..9,311,693	-	
3	CBG10002	III	9,306,565..9.309,187	+	Complete
	CBG09978	III	9,245,962..9,248,585	+	
1	**CBG22303**	I	3,212,459..3,219,426	-	Chimeric
	CBG15102	I	3,275,155..3,278,497	+	
2	CBG15104	I	3,281,822..3,282,205	-	Complete
	CBG22301	I	3,208,751..3,209,134	+	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
3	CBG15103	I	3,279,452..3,281,417	+	Complete
	CBG22302	I	3,209,539..3,211,324	-	
1	**CBG03701**	I	5,613,461..5,615,829	-	Partial
	CBG03760	I	5,601,142..5,601,396	-	
2	CBG03699	I	5,607,948..5,608,249	+	Complete
	CBG03762	I	5,595,630..5,595,931	+	
3	CBG03700	I	5,610,194..5,613,207	+	Complete
	CBG03761	I	5,597,875..5,600,888	+	
1	**CBG10679**	X	21,154,415..21,162,695	-	Partial
	CBG10664	X	21,195,156..21,197,274	-	
2	CBG10681	X	21,149,279..21,149,530	+	Complete
	CBG10666	X	21,190,020..21,190,271	+	
3	CBG10665	X	21,194,229..21,194,663	+	Complete
	CBG10680	X	21,153,488..21,153,922	+	
1	**CBG24756**	I	8,031,339..8,032,334	-	Complete
	CBG02255	I	8,052,924..8,053,919	+	
2	CBG02254	I	8,054,182..8,054,560	-	Complete
	CBG24755	I	8,030,698..8,031,076	+	
3	CBG02253	I	8,055,578..8,056,140	+	Complete
	CBG24754	I	8,029,118..8,029,680	-	
1	**CBG19212**	V	13,145,071..13,146,091	-	Complete
	CBG19053	V	12,612,240..12,613,260	-	
2	CBG19211	V	13,141,692..13,143,046	+	Complete
	CBG19052	V	12,608,860..12,610,231	+	
3	CBG19050	V	12,604,347..12,605,969	+	Complete
	CBG19209	V	13,137,180..13,138,802	+	
1	**CBG16630**	III	7,976,849..7,977,558	-	Complete
	CBG24093	V	16,357,378..16,358,087	-	
2	CBG16631	III	7,978,406..7,979,998	+	Complete
	CBG24094	V	16,358,935..16,360,527	+	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
3	CBG16629	III	7,975,848..7,976,538	-	Complete
	CBG24092	V	16,356,383..16,357,067	-	
1	**CBG17140**	X	7,267,003..7,267,245	-	Complete
	CBG17143	X	7,260,921..7,261,163	-	
2	CBG17139	X	7,267,385..7,267,723	-	Complete
	CBG17142	X	7,261,303..7,261,641	-	
3	CBG17144	X	7,258,902..7,259,196	+	Complete
	CBG17141	X	7,264,741..7,265,279	+	
1	**CBG24788**	UN	1,995,533..2,000,484	-	Partial
	CBG11320	V	750,843..763,241	-	
2	CBG11321	V	746,629..748,380	+	Complete
	CBG24785	UN	1,988,820..1,990,571	+	
3	CBG11322	V	744,140..746,295	-	Complete
	CBG24784	UN	1,986,330..1,988,486	-	
1	**CBG15349**	III	14,503,654..14,514,467	+	Partial
	CBG00609	III	14,253,133..14,253,968	-	
2	CBG15351	III	14,514,594..14,515,130	-	Complete
	CBG00608	III	14,252,470..14,253,006	+	
3	CBG15353	III	14,517,818..14,519,857	+	Complete
	CBG00606	III	14,247,747..14,249,784	-	
1	**CBG06598**	V	7,000,908..7,001,234	+	Complete
	CBG06608	V	7,024,424..7,024,645	+	
2	CBG06599	V	7,001,940..7,003,831	-	Complete
	CBG06609	V	7,025,509..7,027,366	-	
3	CBG06610	V	7,028,756..7,030,306	+	Complete
	CBG06600	V	7,005,221..7,006,771	+	
4	CBG06611	V	7,030,350..7,032,025	-	Complete
	CBG06601	V	7,006,950..7,008,493	-	
1	**CBG13321**	III	6,368,597..6,369,596	+	Complete
	CBG13316	III	6,355,818..6,356,817	+	

# Pairs in Linked Set	Paralogs	Chr	Coordinates (approx)	Strand Orientation	Structure
2	CBG13322	III	6,369,878..6,370,818	+	Complete
	CBG13317	III	6,357,098..6,358,038	+	
3	CBG13315	III	6,354,264..6,355,303	-	Complete
	CBG13320	III	6,367,044..6,368,083	-	
4	CBG13314	III	6,353,076..6,353,507	-	Complete
	CBG13319	III	6,365,856..6,366,287	-	
1	**CBG27169**	V	12,792,759..12,794,782	-	Complete
	CBG19115	V	12,804,100..12,806,123	-	
2	CBG19116	V	12,806,316..12,808,212	-	Complete
	CBG19111	V	12,794,975..12,796,871	-	
3	CBG19117	V	12,808,597..12,809,844	+	Complete
	CBG19112	V	12,797,256..12,798,503	+	
4	CBG19114	V	12,800,895..12,803,980	+	Complete
	CBG19110	V	12,791,149..12,792,639	+	

Figures

Figure 1. – Frequency distribution of three structural categories of gene duplicates between the $K_S = 0$ cohort and the $0 < K_S \leq 0.10$ cohort. A G – test of independence found no difference in frequency of structural categories between age-cohorts ($G = 0.58$, $d.f. = 2$, $P > 0.5$).

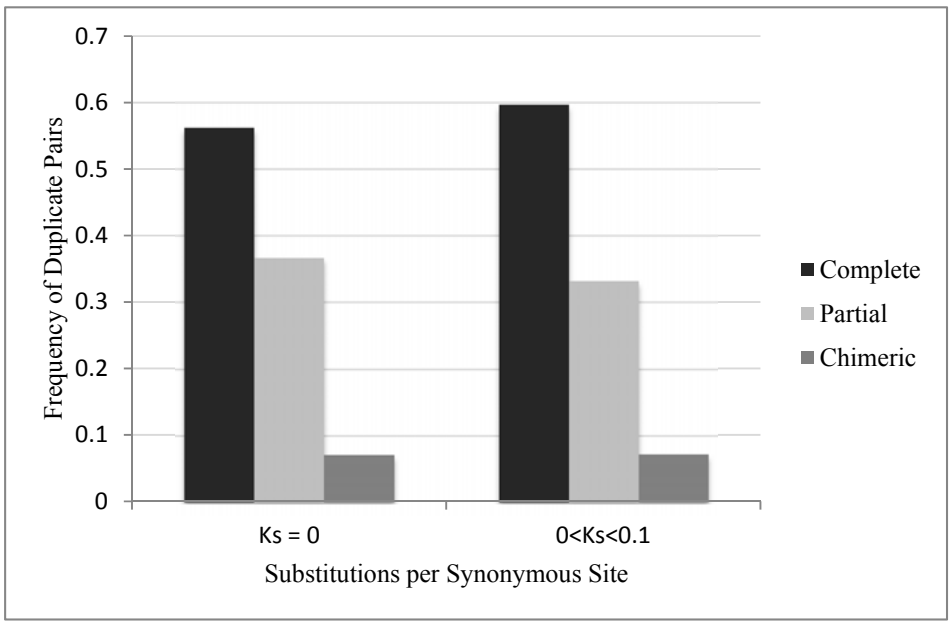


Figure 2. – Relative frequencies of gene duplicate pairs residing on the same chromosome versus different chromosomes between the $K_S = 0$ cohort and the $0 < K_S \leq 0.10$ cohort. A G – test of independence showed a significant difference in relative location of gene duplicates between the two age-cohorts ($G = 34.12$, $d.f. = 1$, $P << 0.01$).

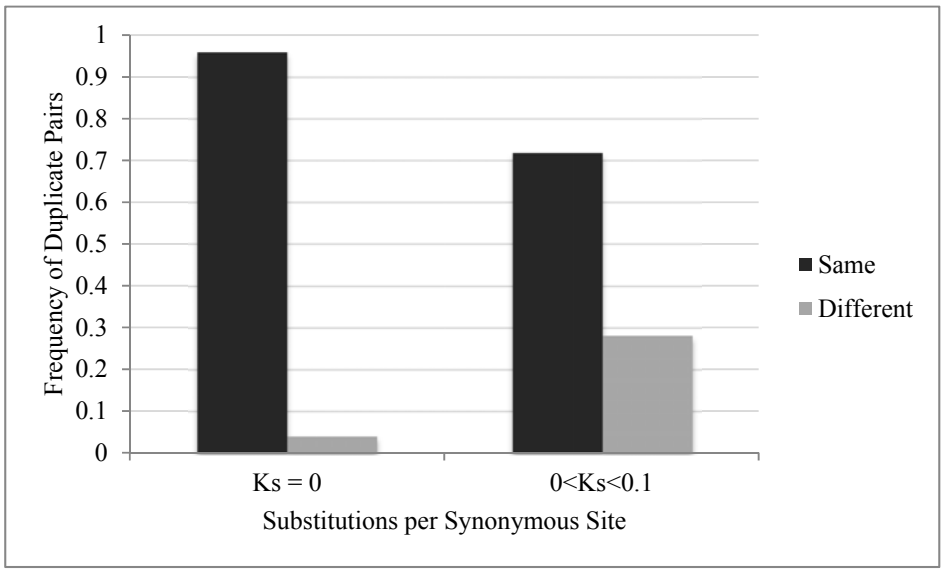


Figure 3. – Density of young duplicate pairs per chromosome. The number of gene duplicates per chromosome was corrected for by dividing by the number of genes per chromosome. A *G* – test for goodness of fit found no evidence for increased duplication on any given chromosome(s) ($G = 3.56$, $d.f. = 5$, $P > 0.25$).

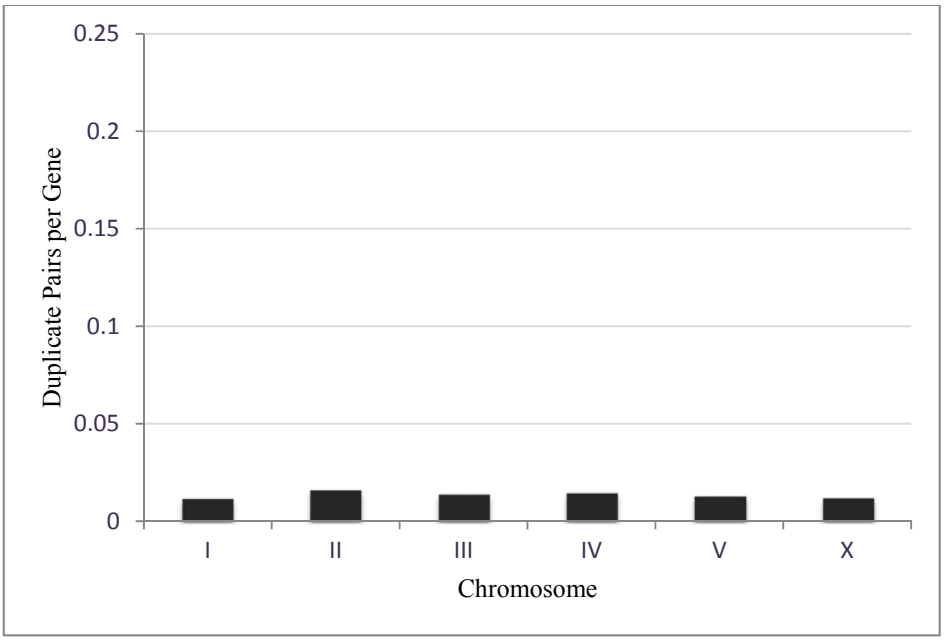


Figure 4. – Correlation between Log *intrachromosomal* distances for 240 gene duplicate pairs residing on the same chromosome and evolutionary age of duplicate pairs represented by sequence divergence at synonymous sites (K_S). No evidence was found for a significant correlation between genomic distance and K_S (*adj* R^2 = -0.003459, $P >$ 0.5).

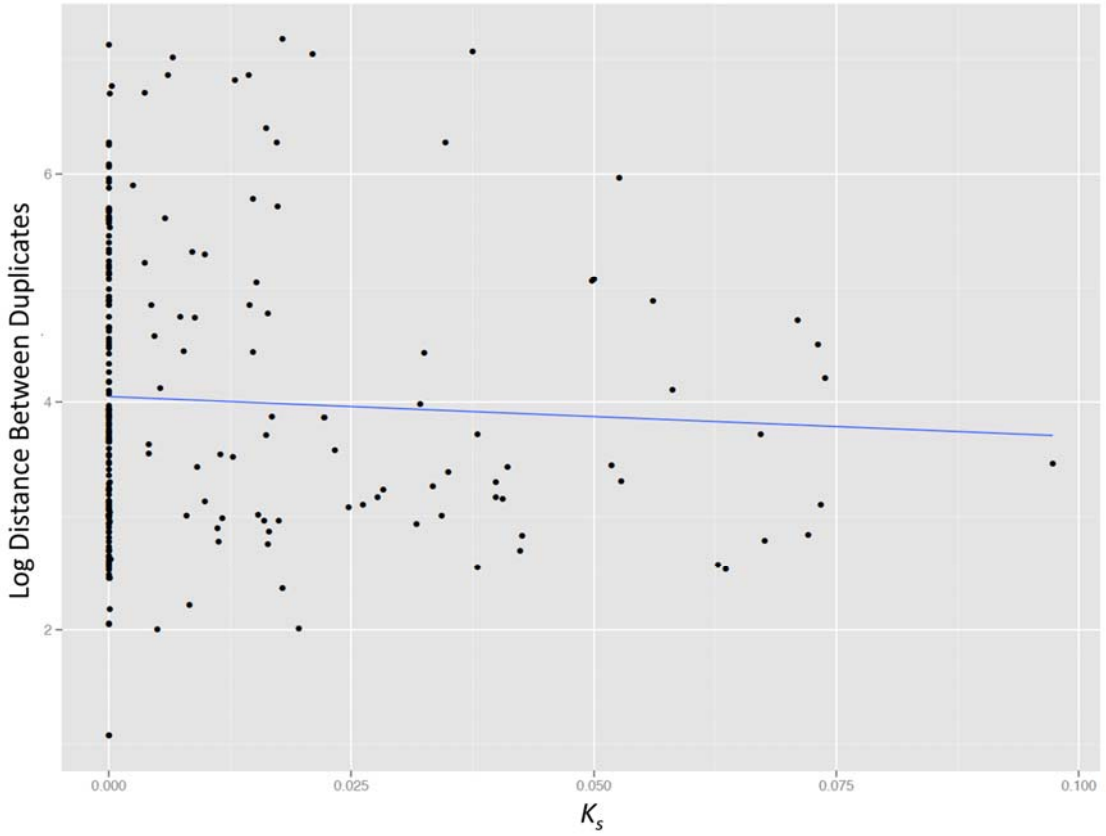


Figure 5. – Frequency distribution of *direct* and *inverse* gene duplicate pairs in the $K_S = 0$ cohort compared to the $0 < K_S \leq 0.10$ cohort. A G – test of independence revealed a significant difference in the proportion of *direct* and *inverse* duplicates between the two age-cohorts ($G = 15.69$, $d.f. = 1$, $P < 0.0005$).

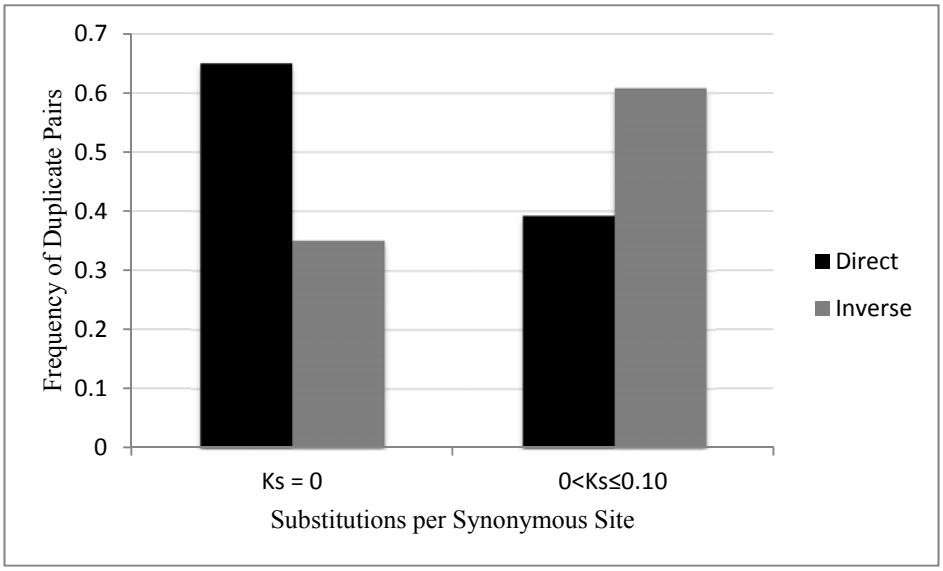


Figure 6. – Relationship between duplication span (bp) and K_S of 376 gene duplicate pairs. The regression indicates a strong negative relationship between duplication span and K_S (Kendall's tau = -0.2339662, P = 7.698e-10).

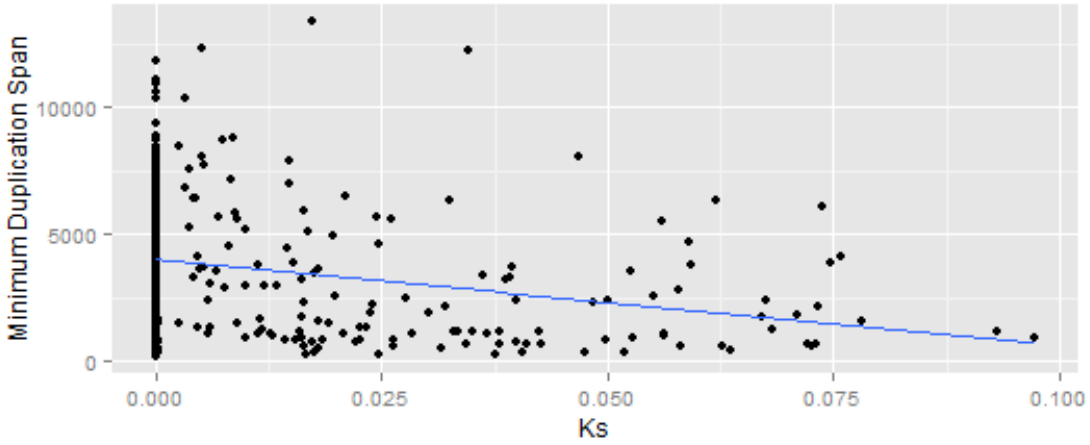


Figure 7. – Frequency distribution of duplication span (kb).

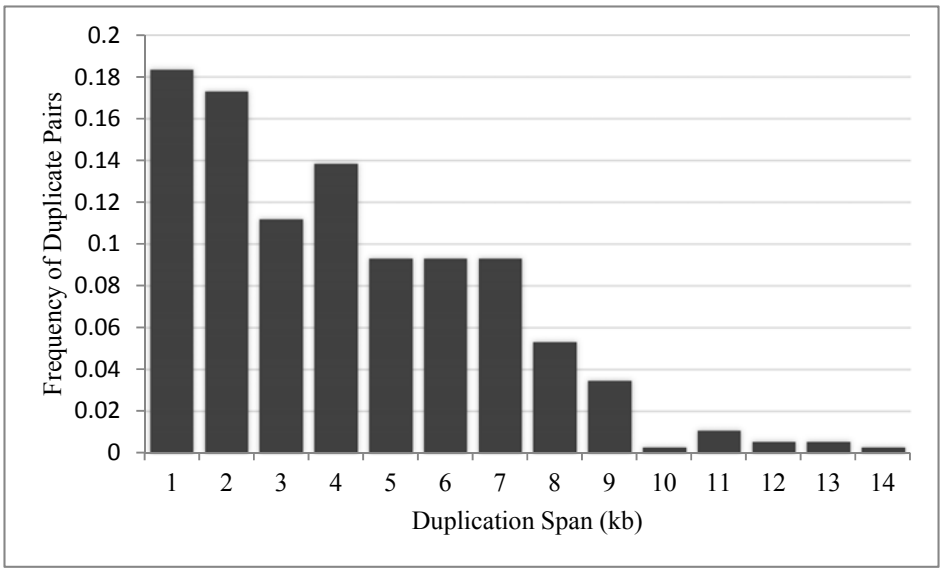


Figure 8. – Frequency distribution of synonymous nucleotide substitutions per synonymous site. L-shaped distribution suggests low retention rate of gene duplicates over evolutionary time.

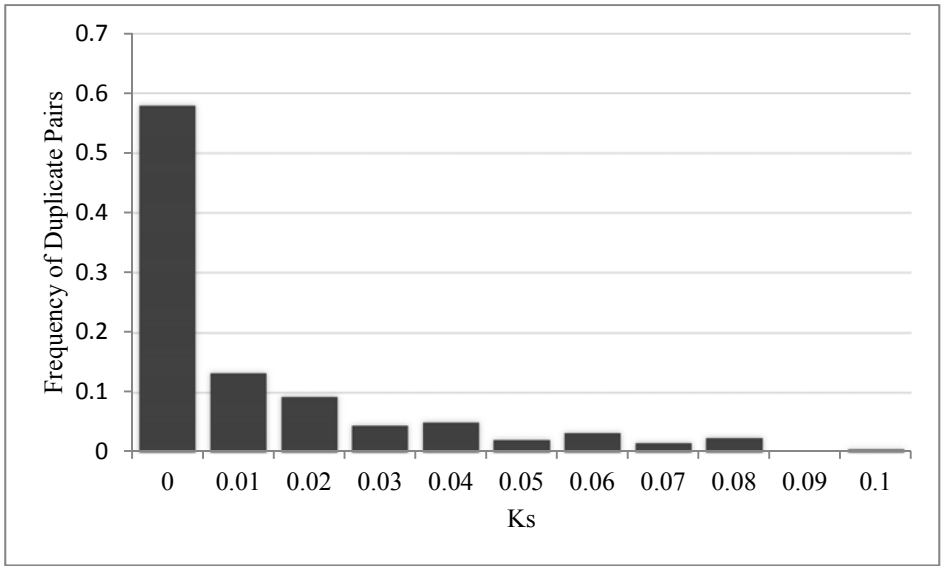
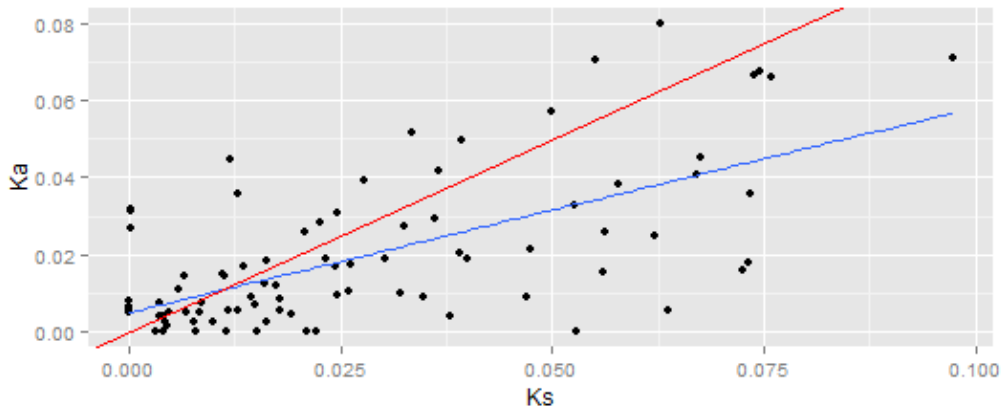
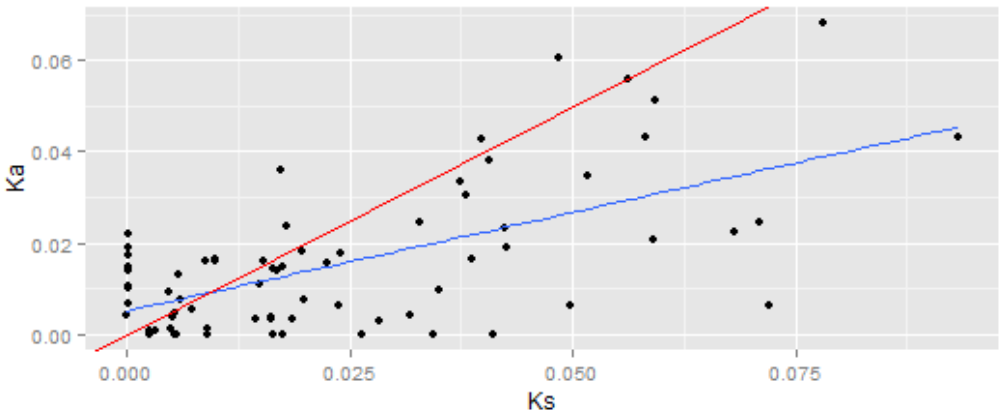


Figure 9. – K_A/K_S distributions. (a) Distribution of K_A and K_S values of structurally *complete* gene duplicates of $0 < K_S \leq 0.1$ (b) Distribution of K_A and K_S values of structurally *partial/chimeric* gene duplicates of $0 < K_S \leq 0.1$. All data points below the red line with a slope of one and an intercept of zero are under varying levels of purifying selection ($K_A/K_S < 1$). The blue regression line for *complete* gene duplicates has a slope of 0.5323 while the blue regression line for *partial/chimeric* duplicates has a slope of 0.4306. A Kolmogorov-Smirnov test showed no difference between K_A/K_S ratios between two different categories of structural resemblance ($D = 0.1214$, $P > 0.50$).

(a)



(b)



Work Cited

Achaz, G., Netter, P., Coissac E. (2001). Study of intrachromosomal duplications among the eukaryote genomes. *Molecular Biology and Evolution*. *18*(12), 2280-8.

Aevermann, D. B., Waters, R. E. (2008). A comparative genomic analysis of the small heat shock proteins in *Caenorhabditis elegans* and *C. briggsae. Genetica. 133*(13), 307 – 19.

Baer, C. F., Shaw, F., Steding, C., Baumgertner, M., Hawkins, A., Houppert, A., . . . Lynch, M. (2005). Comparative evolutionary genetics of spontaneous mutations affecting fitness in rhabditid nematodes. *Proceedings of the National Academy of Sciences of the United States of America*. *102*(16), 5785-90.

Bailey, J. A., Liu, G., Eichler, E. E. (2003). An *Alu* transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics 73*(4), 823-34.

Bavishi, A., Lin, L., Schroeder, K., Peters, A., Cho, H., Choudhary, M. (2010). The prevalence of gene duplications and their ancient origin in *Rhodobacter sphaeroides* 2.4.1. *BMC Microbiology. 10*, 331

Behura, S. K., Severson, D. W. (2013). Association of microsatellite pairs with segmental duplications in insect genomes. *BMC Genomics*. *14*, 907.

Brunet, G. F., Crollius, R. H., Paris, M., Aury, J., Gibert, P., Jaillon, O., … Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution. 23*(9), 1808-16.

Bu, L., Katju, V. (2015). Comparative study of the early evolutionary history of gene duplicates between the human and chimpanzee genome. *In prep*.

Bu, L., Katju, V. (2015). Early Evolutionary history and genomic features of gene duplicates in the human genome. *In review.*

Coghlan, A., Wolfe, K. H. (2002). Fourfold faster rate of genome rearrangement in nematodes that in *Drosophila. Genome Research.* 16, 857-67.

Conant, G. C., Wagner, A. (2004). Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans. Proceedings of the Royal Society.* 271(1534), 89-96.

Cordaux, R., Udit, S., Batzer, M. A., Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proceedings of the National Academy of Sciences of the United States of America*. *103*(21), 8101-6.

Cutter, A. D., (2008). Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Molecular Biology and Evolution*. *25*(4), 778-86.

Denver, D. R., Morris, K., Lynch, M., Thomas W. K. (2004). High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature*. *430*, 679-82.

Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C., Dujon, B. (2001). Evolution of gene order in the genomes of two related yeast species. *Genome Research*. *11*, 2009-2019

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., Li, W. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*. *421*, 63-66.

Hanada, K., Kuromori, T., Myouga, F., Toyoda, T., Li, W., Shinozaki, K. (2009). Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. *Genome Biology and Evolution*. *1*, 409-14.

Harris, T. W., Lee, R., Schrwarz, E., Bradnam, K., Lawson, D., Chen, W., … Stein, D. (2003). Wormbase: a cross-species database for comparative genomics. *Nucleic Acids Research*. *31(1)*, 133 – 7.

Hillier, L. W., Miller, R. D., Baird, S. E., Chinwalla, A., Fulton, L. A., Koboldt, D. C., Waterston, R. H. (2007). Comparison of *C. elegans* and *C. briggsae* genome sequence reveals extensive conservation of chromosome organization and synteny. *PLOS Biology*. *5*(7), e167.

Jun, J., Ryvkin, P., Hemphill, E., Mandoiu, I., Nelson, C. (2009). The birth of new genes by RNA- and DNA-mediated duplication during mammalian evolution. *Journal of Computational Biology*. *16*(10), 1429-44.

Katju, V., Farslow, J. C., Bergthorsson, U. (2009). Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*. *Genome Biology*. *10*(7), R75.

Katju, V., LaBeau, E. M., Lipinski, K. J., Bergthorsson, U. (2008). Sex change by gene conversion in a *Caenorhabditis elegans fog-2* mutant. *Genetics*. *180*, 669–72.

Katju, V., Lynch, M. (2003). The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics*. *165*(4),1793–1803.

Katju, V., Lynch, M., (2006). On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Molecular Biology and Evolution*. *23*(5), 1056–67.

Kondrashov, A. F., Rogozin, B. I., Wolf, I. Y., Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*. *3*(2).

Koszul, R., Dujon, B., Fischer, G. (2006). Stability of large segmental duplications in the yeast genome. *Genetics*. *172*, 2211-22.

Li, J., Hou, Y., Kong, F., Li, Y., Zeng, M., Li, D., … Sun, W. (2015). The fate of recent duplicated genes following a fourth-round whole genome duplication in tetraploid fish, common carp (*Cyprinus carpio*). *Scientific Reports*. *5*, 8199.

Lipinski, K. J., Farslow, J. C., Fitzpatrick, K, A., Lynch, M., Katju, V., Bergthorsson, U. (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology*. *21*, 306-10.

Lu, J., Peatman, E., Tang, H., Lewis, J., Liu, Zhanjiang. (2012). Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. *BMC Genomics*. *13*, 246.

Lynch, M., Blanchard, J., Houle, D., Kibota, T., Schultz, S., Vassilieva, L., Willis, J. (1999). Perspective: spontaneous deleterious mutation. *Evolution*. *53*(3), 645-63.

Lynch, M., Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science. 290*, 1151-5.

Markov, V. G., Baskaran, P. (2015). The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in Nematodes. *Journal of Molecular Evolution 80*(1), 18-36.

Meisel, R. P. (2009). Repeat mediated gene duplication in the *Drosophila pseudoobscura* genome. *Gene*, *438*, 1-7.

Moody, L. A., Saidi, Y., Smiles, E. J., Bradshaw, S. J., Meddings, M., Winn, P. J., Coates, J. C. (2012). *ARABIDILLO* gene homologues in basal land plants: species-specific gene duplication and likely functional redundancy. *Planta*. *236*, 1927-41.

Nayak, S., Goree, J., Schedl, T. (2005). *fog-2* and the evolution of self-fertile hermaphroditism in caenorhabditis. *PLoS Biology*. *3*(1), e6.

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

Sipos, B., Massingham, T., Stutz, A. M., Goldman, N. (2012). An improved protocol for sequencing of repetitive genomic regions and structural variations using mutagenesis and next generation sequencing. *PLoS One*. *7*(8), e43359.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., … Waterston, R. H. (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biology. 1*(2), 166-92.

Treangen, T. J., Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Review Genetics*. *13*, 36-46.

Vergara, A. I., Chen, N. (2010). Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics. 11,* 516.

Yang, S., Arguella, R. J., Li, X., Ding, Y., Zhou, Q., Chen, Y., Zhang, Y., …Wang, W. (2008). Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila. PLoS Genetics. 4*(1), e3.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*. *18*(6) 292 – 98.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., … Wang, W. (2008). On the origin of new genes in *Drosophila. Genome Research 18(9),* 1446 – 55.