EXTINCTION OF CHROMOSOMES DUE TO SPECILIZATION IS A UNIVERSAL OCCURRENCE

A THESIS IN
Bioinformatics

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by
JASON RYON WILSON

B.S., Kansas State University, 2016

Kansas City, Missouri
2019

EXTINCTION OF CHROMOSOMES DUE TO SPECILIZATION IS A UNIVERSAL OCCURENCE

Jason Ryon Wilson, Candidate for the Master of Science Degree

University of Missouri-Kansas City, 2019

ABSTRACT

The human X and Y chromosomes evolved from a pair of autosomes approximately 180 million years ago. Despite their shared evolutionary origin, extensive genetic decay has resulted in the human Y chromosome losing 97% of its ancestral genes while gene content and order remain highly conserved on the X chromosome. Five ‘stratification' events, most likely inversions, reduced the Y chromosome's ability to recombine with the X chromosome across the majority of its length and subjected its genes to the erosive forces associated with reduced recombination. The remaining functional genes are ubiquitously expressed, functionally coherent, dosage-sensitive genes, or have evolved male-specific functionality. It is unknown, however, whether functional specialization is a degenerative phenomenon unique to sex chromosomes, or if it conveys a potential selective advantage aside from sexual antagonism. We examined the evolution of mammalian orthologs to determine if the selective forces that led to the degeneration of the Y chromosome are unique in the genome. The results of our study suggest these forces are not exclusive to the Y chromosome, and

chromosomal degeneration may have occurred throughout our evolutionary history. The reduction of recombination could additionally result in rapid fixation through isolation of specialized functions resulting in a cost-benefit relationship during times of intense selective pressure.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Medicine have examined a thesis titled "Extinction of Chromosomes due to Specialization is a Universal Occurrence," presented by Jason R. Wilson, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance

Supervisory Committee

Gerald J. Wyckoff, Ph.D., Committee Chair
Department of Biomedical and Health Informatics
Department of Molecular Biology and Biochemistry

Monica Gaddis, Ph.D.
Department of Biomedical and Health Informatics

Mark A. Hoffman, Ph.D.
Department of Biomedical and Health Informatics

CONTENTS

TABLES

Table Page

ILLUSTRATIONS

Figure Page

ACKNOWLEDGMENTS

I would like to acknowledge my co-authors Joshua Staley and Gerald J. Wyckoff for their contributions to the resulting manuscript for this work. I would also like to thank Christine Malcom and Neil Miller for help in preparing and reviewing this manuscript. This work was supported in part by Bionexus KC.

CHAPTER 1

INTRODUCTION

The human Y chromosome has lost its ability to recombine with its once homologous partner, the X chromosome, except in its pseudoautosomal regions (PARs) at the termini of the X and Y chromosomes[1-3]. This has resulted in the majority of the Y chromosome's gene content being inherited as a unit, known as the human MSY (male-specific region of the Y chromosome)[4]. Suppression of recombination occurred at five discrete time points, probably caused by inversions, that integrated each segment into the MSY and initiated the degradative processes[5] that resulted in wide-spread gene deletion and loss[1-2,6-7]. These evolutionary strata show a continuum of degeneration that is highly correlated with the age of X-Y gene pairs within each stratum[1-2,4]. The oldest of which contains only four remaining genes, including the sex-determining factor SRY[8]. The degenerative nature of the Y chromosome has led some researchers to suggest it may lose all functional genes and become extinct in as little as 5 million years[8-10], an evolutionary phenomenon that has been observed in other species[11-13]. Recent research, however, suggests that the Y chromosome has maintained a stable assortment of genes for the last 25 million years[3,14-15] through effective purifying selection on single-copy genes[16], and intrachromosomal gene conversion of ampliconic sequences[17-21]. Despite conflicting views on the terminal fate of the Y chromosome, functional specialization and biased gene retention[22-24] on the Y chromosome is believed to be unique in the genome[25] and may have played an essential role in Y chromosome degeneration.

The remaining functional genes in the human MSY fall into three classes: X-degenerate, ampliconic, and X-transposed[3-4,26]. The X-transposed sequences are a result of an X-to-Y transposition that occurred after the divergence of the human and chimpanzee lineages, approximately 3-4 million years ago[4,26]. These sequences remain 99% identical to their X counterparts[4]. In contrast, the X-degenerate sequences are single-copy MSY genes that are surviving relics of the ancestral autosomes from which the sex chromosomes evolved[4]. With the notable exception of SRY, these genes are functionally coherent[25], and ubiquitously expressed[1,14,17]. Their homologous X counterparts also disproportionately escape X-inactivation and are subject to stronger purifying selection than other X-linked genes[17]. Thus, researchers have suggested that this class of sequences is dosage-sensitive and potentially haplolethal[17]. The last class of functional genes in the MSY consists of nine protein-coding gene families that have undergone various levels of amplification[4]. Unlike the ubiquitously expressed X-degenerate genes, the ampliconic gene families are expressed primarily or exclusively in the testes[4,18] and rely on intrachromosomal gene conversion to offset the degenerative nature of the MSY[18-21]. Surviving Y-linked genes were therefore retained through two evolutionary mechanisms: effective purifying selection on single copy dosage-sensitive genes[16] and intrachromosomal gene conversion of ampliconic sequences[17-21].

Wide-spread gene loss accompanied by preferential retention appears to be a unique phenomenon. A review of genomic evolution, however, suggests that these trends are not unique to the Y chromosome, with the relevant literature rarely being cross-cited. Recent research suggests that the ancestral vertebrate karyotype was much larger than previously estimated, consisting of an estimated 54 chromosomes[27] resulting from two ancestral whole-genome duplication (WGD) events[27-30]. The majority of genes following a WGD event are rapidly lost or pseudogenized due to loss of function mutations[7,31-33]. This loss has also been shown to continue on a power scale[32,34]. Consequently, a large portion of the ancestral vertebrate chromosomes has been subsequently lost through fusion in the descent of the human lineage[27,30], explaining the apparent haphazard gene content of most autosomes[1]. Highly expressed genes[35], dosage-sensitive protein complexes[33,36], and transcriptional and developmental regulators and signal transducers, however, are preferentially retained[29,32,37-39]. Furthermore, these genes have been maintained through purifying selection[36], a trend that has been observed in ubiquitously expressed genes throughout the genome[40-44]. The factors that led to the biased retention of ubiquitously expressed single-copy genes, therefore, appear not to be restricted to the evolutionary history of the Y chromosome and have been observed in the events following large scale duplications. The biased acquisition of male-advantage traits on the Y chromosome is a subject of more considerable ambiguity in the context of genomic duplications.

Subfunctionalization has been shown to increase the likelihood a gene will be preserved in duplicate due to partial loss of function mutations in both copies[45]. This targeted divergence of the duplicates may lead to differential tissue expression of the paralogs[33-34,44,46-48] and has been proposed to occur frequently following WGD events[49]. If the remaining functions are under selective constraint, the duplicates will likely remain in the population[46]. A lack of genome-wide representation of subfunctionalized gene pairs, however, suggests that this may be a transition phase to neofunctionalization due to an absence of purifying selection on the redundant portions of the gene[50], an evolutionary phenomenon known as the subneofunctionalization model[51]. In 2009, Wilson and Makova suggested that suppression of recombination could be thought of as a duplication event and showed X-Y genes followed similar patterns of evolution following recombination suppression as duplicated paralogs[52-53]. Following a review of experimental data, they also concluded that the acquisition of unique expression patterns and functions might have contributed to the retention of Y-linked genes. Strong expression reduction has also been implicated in the evolution of Y genes towards testis specificity[54]. The biased content of male reproductive genes on both sex chromosomes[4,55-57], therefore, suggests that subfunctionalization of Y-linked genes could explain the initial retention and accelerated divergence of male-advantage genes, as new evolutionary features typically bear marks of their ancestry[49].

The WGD events at the origin of the vertebrate lineage may have had significant impacts on biological complexity and evolutionary novelties of the time due to the large-scale increase in genetic redundancy[37]. The mechanisms by which this was achieved and the selective pressures resulting in differential chromosome survival remain unknown. If the evolutionary history of the Y chromosome provides a model of genomic evolution, it would suggest that large scale duplication events allowed genes to subfunctionalize and experience periods of relaxed purifying selection through relief from pleiotropic constraints that were operating on single-gene loci[45]. It has also been hypothesized that the Y chromosome's long-term fragility may be driven by short-term selective pressures[58], the most obvious of which is the accumulation of sexually antagonistic alleles in a non-recombining portion of the genome[58-64], a phenomenon that is supported by the transposition of male-advantage genes into the MSY from autosomes[1,3-4,14,65-67]. The rapid evolution of male reproductive genes[44,68-70] and the implication of inversions in local adaption[7,71], however, suggest that functional isolation may become selectively favored even in the absence of sexually antagonistic traits under certain circumstances, despite the deleterious effects of reduced recombination. In order to test these hypotheses, and in lieu of the large amount of literature pertaining to expression, we analyzed the nonsynonymous to synonymous mutation rate (K_a/K_s) of 6734 human genes with surviving mammalian orthologs in the context of their Gene Ontology (GO) annotations and chromosomal locations to determine if functional specialization and genomic isolation convey a selective advantage, respectively.

CHAPTER 2

METHODOLOGY

In the present analysis, we examined the divergence of human genes from their mammalian orthologs with respect to their GO annotations and cytological positions. For mammalian genes, the probability a newly arisen nonsynonymous mutation is fixed, relative to what is expected under neutrality, is resolved by the strength of selection[72]. The K_a/K_s ratio is commonly utilized as a measurement of this selective strength, with low values suggesting strong purifying selection and high values indicating relaxed purifying selection and/or positive selection[72]. By definition, human orthologs are identical by descent[73] with at least one other species in our analysis. The use of mammalian ortholog comparisons in conjunction with GO annotations, therefore, allowed us to analyze how gene protein functions influence the patterns of selection that led to the differential divergence of ancestral mammalian genes.

Data Collection

Divergence data was collected from The Searchable Prototype Experimental Evolutionary Database (SPEED)[74]. SPEED contains orthologous sequence comparisons of nine species including human (*Homo sapiens*), chimp *(Pan troglodytes*), rhesus macaque *(Macaca mulatta*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*), cow (*Bos Taurus*), opossum (*Momodelphis domestica*), and chicken (*Gallus gallus*) as a true outgroup. Methodology on the identification of orthologous groups and calculation of divergence data can be found elsewhere[74].

Orthologous sequence pairs were queried from SPEED for genetic summary information and their related K_a/K_s values. Data cleaning was performed using PySpark (Spark version 2.3.1, Python version 2.7.10). Sequence comparisons containing a K_s value of zero or less due to computational error were removed from the analysis. Two sequence comparisons containing K_a/K_s values greater than 50,000 due to unusually small K_s values were also removed to prevent these outliers from biasing results. Where multiple comparisons existed, divergence data inconsistencies were resolved by computing a zero-corrected harmonic mean; therefore, more significant weight was given to conservative estimates[75], and comparisons containing at least one zero K_a/K_s value were assigned a K_a/K_s value of zero. Lastly, sequence comparisons that did not include a human comparison with an associated gene name and chromosomal location were excluded. Our resulting dataset included a total of 68,006 comparisons across 10,849 genes.

Gene ontology information was collected from the European Bioinformatics Institute[76]. The most recent version of human gene ontology annotations (9/19/19) was downloaded and joined to their respective genes. The dataset included 19,395 genes and 18,211 GO terms. The validity of GO terms with IEA evidence codes has been questioned due to their inferential nature[77]. The quality of IEA terms, however, has significantly improved and rival those inferred by curators[78]. To alleviate potentially biased numbers of GO annotations on well-studied genes, IEA terms were also retained. After joining with the ortholog dataset and removing genes that lacked annotation, our final dataset included 6,734 annotated genes across 14,121 GO terms. IEA (inferred from electron annotation), IDA (inferred from direct assay), ISS (Inferred from Sequence or structural Similarity), IBA (Inferred from Biological aspect of Ancestor), IMP (Inferred from Mutant Phenotype) and TAS (Traceable Author Statement) evidence codes were the primary methods of annotation in our dataset at 30.2%, 20.96%, 13.2%, 12.5%, 10.3%, and 7.3%, respectively.

Data Preparation

Single value human gene K_a/K_s rates were derived by averaging their respective K_a/K_s values across all species comparisons present in the dataset. GO annotation K_a/K_s values were obtained by averaging the K_a/K_s values of all related genes across all species comparisons. Chromosome arm K_a/K_s values were calculated by averaging the K_a/K_s values of all genes present on the respective chromosome arm across all species comparisons. Ortholog density was calculated by dividing the number of orthologs present on a given chromosome arm by arm size in Mb. Lastly, the chromosome arms and their related GO annotations were cross-tabulated to obtain the number of times a given function occurs on each arm.

Due to their hierarchical nature, GO terms can be broad[77]. This issue was addressed in a context dependent manner for each analysis. Prior to clustering the chromosome arms based on functional relatedness, the number of times each GO annotation occurred on each chromosome arm was weighted using an algorithm adapted from Martinez and Reyes-Valdés[79]. We considered the average frequency of the ith GO term among j chromosome arms as,

$$1)\quad p_i = \frac{1}{t}\sum_{j=1}^{t} p_{ij}$$

and defined GO term specificity as the information that its expression provides about the identity of the chromosome arm as

$$2)\quad S_i = \frac{1}{t}\left(\sum_{j=1}^{t} \frac{p_{ij}}{p_i}\log_2 \frac{p_{ij}}{p_i}\right)$$

S_i will give zero if the GO term is expressed on all chromosome arms and max $\log_2(t)$ if the function is exclusively expressed on a single chromosome arm. We then assigned a weighted frequency for each GO term on each chromosome arm as the product of the GO term specificity and its frequency on a given chromosome arm.

$$3)\quad \delta_{ij} = p_{ij}s_i$$

Thus, a higher degree of functional similarity would be found between chromosome arms if their shared functions were absent elsewhere in the genome. This method was also applied to the relationship between genes and their related GO terms. Weighted GO term counts were derived for each gene by summing the specificities of their related GO terms in Eq. 3. However, the weighted GO counts did not alter the distributions or significances of our ortholog regression analyses. Therefore, raw counts were used for ease of interpretability.

A primary goal of our GO annotation regression analyses was to determine if genomic representation influences the selective pressures exerted on a function. Therefore, all ontology terms were retained in this analysis in order to examine our hypothesis that large scale duplication events may relieve pleiotropic constraints in a subset of genes through increased dosage of essential functions. However, where multiple GO terms contained the same set of related genes, only one term was retained to remove redundant data points. 11,016 terms of the original 14,121 were found to have unique sets of related genes.

Statistical Analysis

All regression and distribution analyses were performed in Python (see version above) using the statsmodels API. Due to the strong positive skew of several variables in our dataset, generalized linear models (GLM) were used where appropriate. Fitting lognormally distributed continuous data with a gamma distribution has been shown to perform comparably or outperform lognormal transformations without the need for manual manipulation of the variables[80-81]. A log link was used to maintain a non-linear fit while respecting the domain of the gamma function. Where statistically meaningful zero values were present, a hurdle method was employed to counteract the calculation error introduced. This entails fitting a gamma distribution to all non-zero data, as well as a binomial distribution to the full dataset to determine the influence of the predictor variable on the probability that the dependent variable is zero[82-83]. The linear relationship of chromosome arm's number of related genes and GO annotations was fit with ordinary least squares regression without an intercept, as it was nonsensical in the given context. Normality of distributions was determined using the Shapiro-Wilk test which tests the null hypothesis that the data are normally distributed.

Hierarchical clustering of the chromosome arms based on GO annotation content was performed using the cluster package in R (Version 3.5.3). GO term counts were not scaled before distance calculation due to the homogenous nature of the variables. The distance was calculated using Euclidean distance. The linkage measure was determined by obtaining the agglomerative coefficient (amount of clustering structure found) for single, complete, average linkage and Ward's method using the agnes() function. For our dataset, Ward's method resulted in the highest agglomerative coefficient (data not shown) and was subsequently used in our clustering analysis. Therefore, multi-node clusters were joined based on the minimum increase in within-group variance.

CHAPTER 3

RESULTS

Functional Diversity of Orthologs

Two primary paths to survival have occurred on the Y chromosome. Broadly expressed dosage-sensitive genes have been maintained through purifying selection[16], while amplification and gene conversion have supported testis-specific genes[17-21]. Selection for conservation through amplification and gene conversion of testis-specific sequences and the rapid evolution of male reproductive genes[44,68-70] suggest that adaptability may be a selected phenomenon. Evolution of testis-specific functions is also believed to have preceded amplification on the Y chromosome[17], suggesting subfunctionalization may have facilitated their initial retention[44] and subsequent divergence by relieving redundant portions of the genes from adaptive constraint.

To determine if this is a universal trend, we analyzed the divergence of human genes across their mammalian orthologs to provide a conservative estimate of the degree to which newly duplicated genes may diverge[84] following subfunctionalization. The results of our analysis suggest that a human gene's average K_a/K_s across its related orthologs and number of GO annotations are positively skewed (skew = 3.92 & 3.06, respectively) (Supplementary Figure 1). Additionally, average K_a/K_s is zero-inflated. This suggests that the majority of orthologous genes are under purifying selection and related to a small set of functions. As a gene's average K_a/K_s value appears to be negatively associated with its number of GO annotations (Fig. 1), a gamma-hurdle model was employed (see methods) to determine the statistical significance of this relationship. Our results suggest that a gene's average K_a/K_s decreases with increasing numbers of functional annotations ($p = 8.25 \times 10^{-20}$) (Supplementary Figure 2), and the probability that it is entirely conserved increases ($p = 5.25 \times 10^{-9}$, odds-ratio = 1.011) (Supplementary Figure 3).

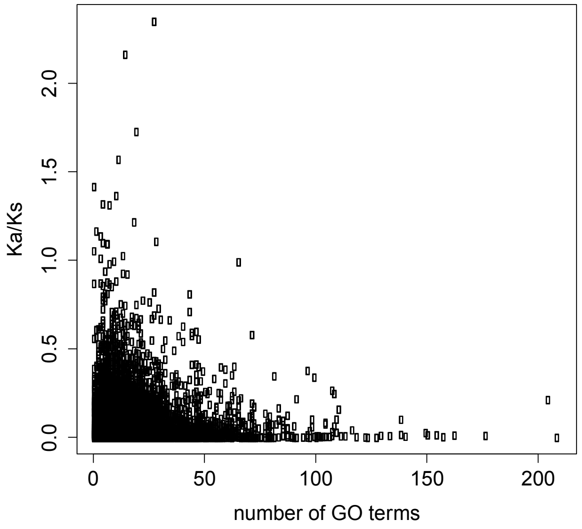


Figure 1 Ortholog K_a/K_s-GO Relationship: Average K_a/K_s values of 6734 human genes across their related orthologs plotted against their number of GO terms

The results of this analysis suggest that genes with high functional diversity are under more intensive purifying selection than their more functionally specific counterparts. These findings parallel those showing higher levels of purifying selection on broadly expressed essential genes throughout the genome[40-44] as well as on the Y chromosome[17] and suggest an association between the two. We conclude that functional specificity and reduced expression are associated with relaxed purifying selection, suggesting that subfunctionalization of duplicated paralogs could result in differential tissue expression[33-34,44,46-48] and accelerated protein divergence.

Genomic Isolation of functional annotations

Next, we were interested in determining if functional isolation can provide a selective advantage in the absence of sexual antagonism. The rapid evolution of male reproductive genes[44,68-70] and the implication of inversions in local adaption[7,71] suggest that the localization of functionally related genes may accelerate protein divergence and facilitate adaptability. To determine if localization of genes related to a given function is associated with reduced purifying selection, we analyzed the average K_a/K_s of GO annotation's related sequence comparisons with respect to the genomic distribution of their related genes. GO annotations show positively skewed distributions for their number of associated genes (skew = 34.68), number of chromosome arms they are expressed on (skew = 2.58), and average K_a/K_s (skew = 2.85) (Supplementary Figure 4). This suggests that the majority of functional annotations we analyzed are carried out by a limited number of genes, are expressed in specific locations and under purifying selection. Their relationships with one another, however, suggest all three trends are not typically present at the same time.

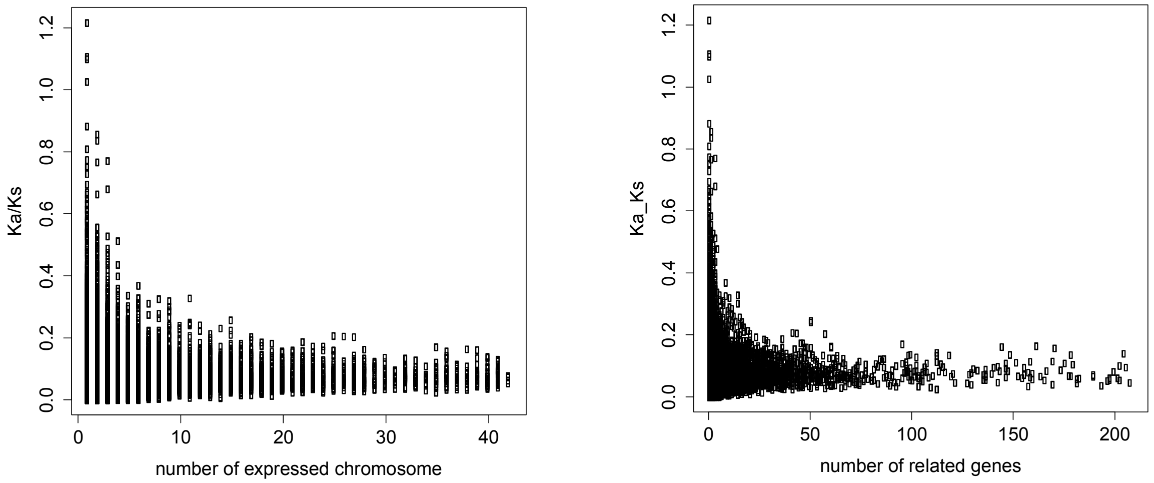


Figure 2 GO K_a/K_s Relationships: Left) Average K_a/K_s values of the 11016 GO terms with unique gene sets in our dataset across each term's related genes and human sequence comparisons plotted against the number of chromosome arms a given GO term was found. Right) Average K_a/K_s values of the same GO terms plotted against their number of related genes (trimmed within 3 SD of the mean number of genes for clarity)

A GO term's average K_a/K_s value appears to be negatively associated with the number of chromosome arms it is expressed on and its number of related genes (Fig. 2). Due to the positive skewed nature of these distributions (Supplementary Figure 5), a gamma model with a log link was used to determine if a function's number of related genes or expressed chromosome arms is significantly associated with its average K_a/K_s. Increasing the number of genes or expressed chromosome arms related to a given function, however, increases the probability one K_a/K_s value is non-zero. Zero average K_a/K_s values in the context of this analysis, therefore, are not informative and were removed from the analysis, negating the need for a hurdle method. The two predictive variables were fit separately to determine their individual effects. The results of our analyses suggest a function's average K_a/K_s decreases with the number of chromosome arms it is expressed on ($p = 7.01 \times 10^{-19}$, Supplementary Figure 6), however, a function's number of related genes was non-significant ($p = 0.05$, Supplementary Figure 7). This suggests that genomic isolation is more strongly associated with relaxed purifying selection than a function's number of related genes. The non-significance of a function's number of related genes additionally suggests that low K_a/K_s values for annotations expressed on a large number of chromosome arms cannot be attributed to convergence to the genome-wide average alone.

We expect that the majority of functions related to a large number of genes and expressed throughout the genome are higher-level ontology functions. Genes that are beneficial in increased dosage, however, are preferentially retained following duplication events[84-85]. Thus, large scale duplications may have resulted in the stability of higher-level functions, while relieving more redundant duplicates from adaptive conflict. We conclude that functional isolation is associated with relaxed purifying selection on the genes related to that function, potentially through relief from background selection acting on more highly conserved linked sites[86]. This finding parallels the accelerated evolution of Y-linked genes following recombination suppression and suggests isolation of functions may accelerate sequence divergence of their related genes through relaxation of purifying selection. These findings provide only a modest estimate of the extent to which protein functions may diverge in isolation when recombination is suppressed or following a WGD event when genetic redundancy is at its peak.

Potential Retention of Functionally Related Haplogroups

A GO annotation's number of associated genes also appears to increase exponentially with the number of chromosome arms it is expressed on (Fig. 3). This relationship was fit using gamma regression and a log link, the results of which were highly significant ($p < 0.0005$, Supplementary Figure 8). For a GO annotation's number of related genes to increase in this manner, the genes on a given chromosome arm must be moderately functionally related. This suggests that the retention of genes following large-scale duplication events may operate at the haplogroup level, a trend that is predicted due to the dosage-sensitive nature of protein complexes. Chromosomes enriched with blocks of functionally related genes that are beneficial in increased dosage would show the highest levels of gene retention.

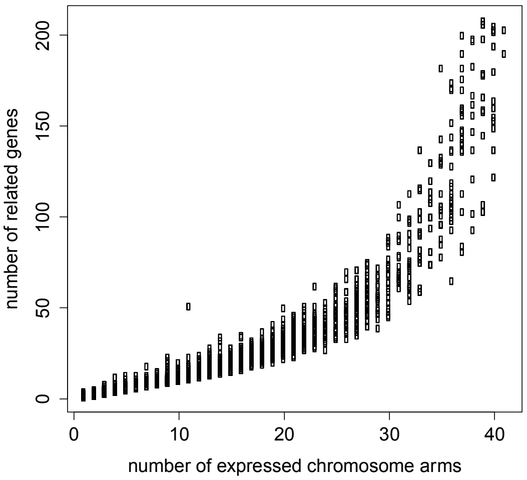


Figure 3 GO Chromosome Arm and Gene Relationship: Number of genes related to each GO term plotted against the number of chromosome arms it was found (trimmed to within 3 SD of the mean number of genes for clarity)

Thus, the functional coherence of the Y chromosome could be attributed to a lower content of functionally related haplogroups that were beneficial in increased dosage on the ancestral autosomes.

Biased Retention of Orthologs on Existing Chromosomes

The ancestral vertebrate chromosomes displayed substantial differences in gene number, potentially as a result of more significant gene deletion and loss on chromosomes with a smaller number of resulting genes[27]. This has led to speculation of systematic biases in the deletion of duplicates on a subset of chromosomes following rediploidization, which may have resulted in the chromosome's eventual loss[27]. We were interested in determining if this systematic bias could be attributed to the gene content of the pre-duplicated chromosomes from which they were derived. The results of our chromosome analysis show human chromosome arms have normally distributed numbers of orthologous genes (Shapiro-Wilk 0.97, $p = 0.34$) and average K_a/K_s values (Shapiro-Wilk 0.98, $p = 0.72$) (Supplementary Figure 8), and that a chromosome arm's number of genes and GO annotations are linearly related ($p = 5.68x10^{-40}$, adjusted $R^2 = 0.985$, Supplementary Figure 9) (Fig. 4). Density of orthologs on existing chromosome arms, however, was found to be non-normally distributed (Shapiro-Wilk 0.912, $p = 0.002$). This is due to biased ancestral gene conservation on a subset of chromosomes. Despite differential average K_a/K_s rates, selection at the chromosome arm level since the divergence of mammals does not appear to influence the number or density of orthologs on a given chromosome (Supplementary Figure 10).

In contrast, we found that arms of chromosomes that have retained large clusters of genes resulting from the ancestral WGD events contain a disproportionate number of

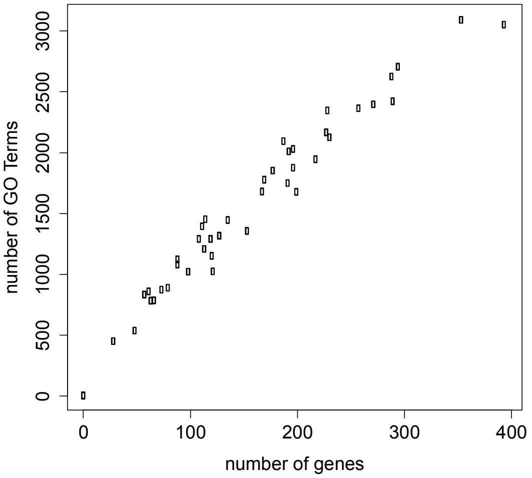
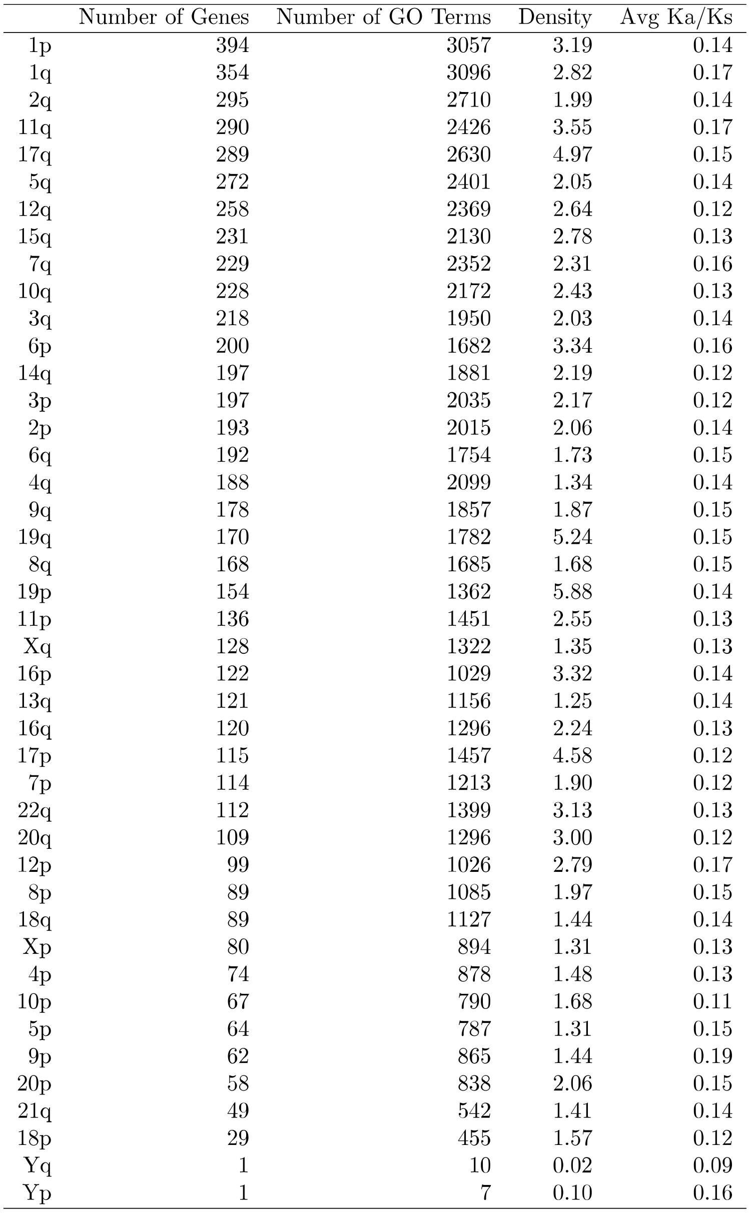


Figure 4 Chromosome Arm Gene-GO Relationship: A chromosome arm's number of GO terms plotted against its number of orthologous genes within our dataset

mammalian orthologs in our analysis. Approximately 35% of genes still exist in duplicate copies[38], and several ohnologs (paralogs resulting from WGD events[30-31]) were retained in quartets[27]. These include clusters containing the four HOX regions on chromosomes 2, 7, 12, and 17, as well as the MHC region on chromosome 6 containing ohnologs on chromosomes 1, 9, and 19 that are a result of single pre-duplicated regions[27]. The gene content of chromosomes 14 and 15 have also been shown to be almost entirely derived from individual pre-duplicated chromosomes[27]. The arms of these chromosomes show some of the higher



	Number of Genes	Number of GO Terms	Density	Avg Ka/Ks
1p	394	3057	3.19	0.14
1q	354	3096	2.82	0.17
2q	295	2710	1.99	0.14
11q	290	2426	3.55	0.17
17q	289	2630	4.97	0.15
5q	272	2401	2.05	0.14
12q	258	2369	2.64	0.12
15q	231	2130	2.78	0.13
7q	229	2352	2.31	0.16
10q	228	2172	2.43	0.13
3q	218	1950	2.03	0.14
6p	200	1682	3.34	0.16
14q	197	1881	2.19	0.12
3p	197	2035	2.17	0.12
2p	193	2015	2.06	0.14
6q	192	1754	1.73	0.15
4q	188	2099	1.34	0.14
9q	178	1857	1.87	0.15
19q	170	1782	5.24	0.15
8q	168	1685	1.68	0.15
19p	154	1362	5.88	0.14
11p	136	1451	2.55	0.13
Xq	128	1322	1.35	0.13
16p	122	1029	3.32	0.14
13q	121	1156	1.25	0.14
16q	120	1296	2.24	0.13
17p	115	1457	4.58	0.12
7p	114	1213	1.90	0.12
22q	112	1399	3.13	0.13
20q	109	1296	3.00	0.12
12p	99	1026	2.79	0.17
8p	89	1085	1.97	0.15
18q	89	1127	1.44	0.14
Xp	80	894	1.31	0.13
4p	74	878	1.48	0.13
10p	67	790	1.68	0.11
5p	64	787	1.31	0.15
9p	62	865	1.44	0.19
20p	58	838	2.06	0.15
21q	49	542	1.41	0.14
18p	29	455	1.57	0.12
Yq	1	10	0.02	0.09
Yp	1	7	0.10	0.16

Figure 5 Chromosome Arm Summary Statistics: Includes number of genes, number of GO terms, orthologs/Mb, and average K_a/K_s of all sequences on a given chromosome arm

levels of mammalian ortholog retention in our analysis, and chromosomes 17 and 19 have the highest ortholog densities in the genome (Fig. 5).

To determine the extent to which these chromosomes remain functionally related, aside from their conserved gene families, we performed hierarchical clustering of the

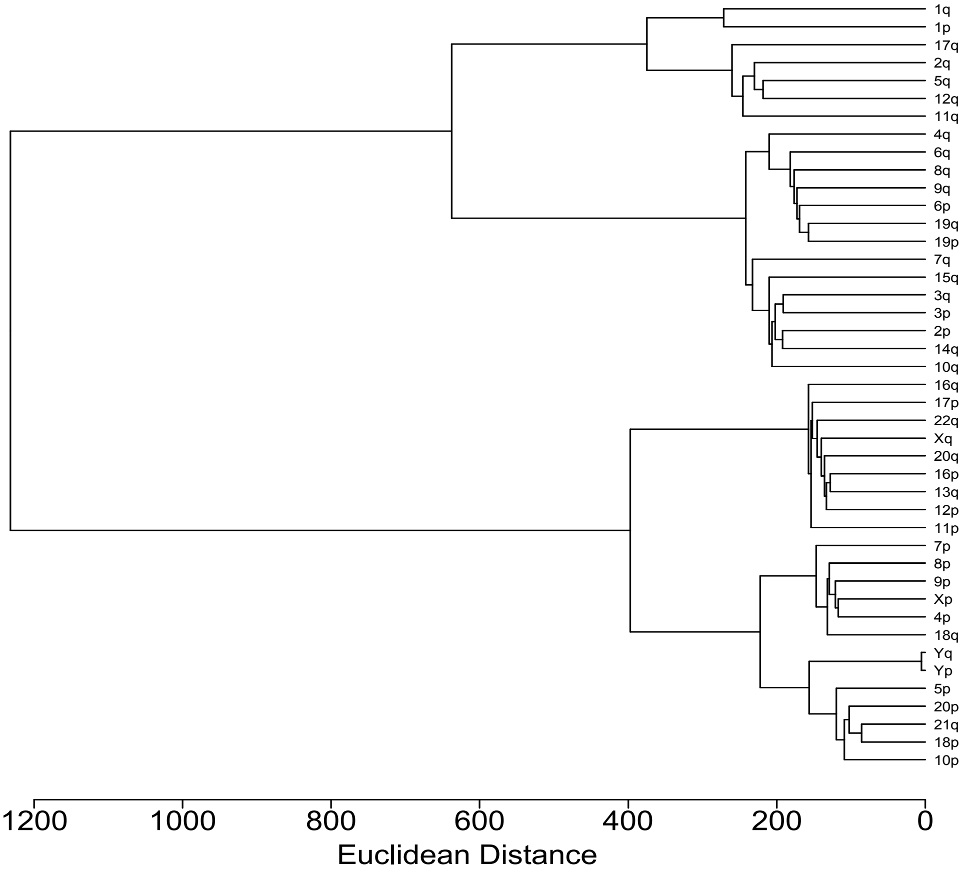


Figure 6 Chromosome Arm Dendrogram: Hierarchical clustering of the chromosome arms based on functional relatedness. Each chromosome arm was assigned a weighted term count for the 14121 GO terms present in our dataset. The arms were subsequently clustered using Euclidean distance and Ward's method. Multi-node clusters are therefore clustered based on minimum increase in within group variance.

chromosome arms based on a weighted frequency (see methods) for each GO annotation on a given arm in our dataset. The results of our dendrogram (Fig. 6) indicate several trends in the functional relationships between chromosome arms. The two top-level clusters appear to be differentiated based on the number of related functions retained on the chromosome arms (Fig. 5), a result that was expected given clustering with Euclidean distance. We additionally found lower level clustering of chromosome arms that include both the ancestral HOX and MHC regions. These include the clustering of chromosome 2q, 12q, and 17q, as well as 6pq, 19pq, and 9q. This suggests that the ancestral WGD events have had a profound impact on the retention and organization of mammalian orthologs throughout the human genome.

As stated earlier, specific classes of genes are preferentially retained following WGD events. Chromosomes that have maintained a large portion of their ancestral genes are therefore a result of the gene content and functional annotations of the pre-duplicated chromosomes from which they were derived. It has been hypothesized that the specialization of the Y chromosome is a result of the number of functional genes initially present on the ancestral autosomes[56], a hypothesis supported by the low functional gene density of the X chromosome[2,57]. In our present analysis, we also found low ortholog density on both the X chromosome, and chromosomes that are orthologous to the chicken Z chromosome (chromosomes 5, 9, and 18[57]). However, we did not find as significant of a disparity in ortholog density between the X chromosome and the human genome-wide average (2.33) relative to overall gene density comparisons[57], suggesting that ancestral gene density may be less heavily influenced by the invasion of interspersed repeats. Biased gene deletion following the ancestral WGD events resulting in low gene density on a subset of chromosomes suggests that several chromosomes were pre-adapted to specialize similar to the sex chromosomes. Furthermore, chromosomal rearrangements would be under less negative selection in these regions.

CHAPTER 4

DISCUSSION

Since its discovery, the perceived functional importance of the Y chromosome has grown exponentially within the scientific community and now may provide further insight into chromosomal evolution following the ancestral WGD events at the origin of the vertebrate lineage. Our present analysis, in conjunction with existing literature, has shown that evolutionary trends believed to be unique to the Y chromosome are observed in the events following large-scale duplications and are still present in mammalian ortholog comparisons. These include higher levels of purifying selection on functionally diverse, ubiquitously expressed genes[40-44], as well as reduced purifying selection on genomically isolated protein functions. The biased distribution of ancestral mammalian genes on chromosomes primarily derived from single pre-duplication chromosomes additionally suggests that gene retention was dependent on the gene content of the ancestral chromosome from which they were derived, and this retention may persist over long periods of evolutionary time. The conservation of the functionally coherent, potentially haplolethal X-degenerate sequences through purifying selection[16], the rapid evolution of ampliconic sequences expressed primarily in the testes[14], and the pseudogenization and loss of redundant sequences are consistent with a large-scale duplication event. This suggests that the Y chromosome may serve as a model for chromosome evolution following a large-scale duplication event.

Examining suppression of recombination on the Y chromosome in the light of large-scale duplication events has essential implications for karyotypic evolution at the onset of the vertebrate lineage. It has been proposed that the ancestral WGD events contributed to the proliferation of vertebrates during the Cambrian period due to the increase in genetic variation and tolerance to environmental conditions[37,87]. Recent research suggests that all extant vertebrate karyotypes are descendants of an ancestral marine chordate consisting of 17 chromosomes that underwent two successive WGDs[27]. Rapid loss through the fusion of seven chromosomes between duplications and the loss of an additional five chromosomes following the second duplication resulted in an ancestral *Amniota* karyotype of 49 chromosomes with highly differential gene content[27]. The smaller size of extant genomes, therefore, suggests a consistent pattern of karyotype reduction following the ancestral WGD events[27], and speciation rates have been shown to be strongly correlated with chromosomal evolution rates[88].

Duplication events should occur at a fitness cost, and an optimal gene copy number should exist[84]. Duplicate genes, therefore, would be subjected to three potential evolutionary fates: retention of genes that are beneficial in increased dosage, inactivation of genes that are harmful in increased dosage, and a period of neutral evolution of redundant sequences. Consequently, the observed gene retention on a given chromosome could be attributed to its density of genes that were beneficial in an increased dosage and subsequently retained through purifying selection. Loss of chromosomes due to widespread gene inactivation of detrimental duplicates, however, should have occurred early and ubiquitously, contributing little to the evolutionary novelties and speciation observed at the time. If the divergence of the Y chromosome serves as an evolutionary model, it would suggest an alternative hypothesis: chromosomal rearrangements resulting in large regions of the genome being protected from gene flow allow isolated genes to diverge until a complete reproductive barrier exists[89].

The specialization of SRY as the sex-determining factor appears to have played a significant role in X-Y divergence, as its emergence is correlated with the first stratification event that reduced recombination between the neo-sex chromosomes[90]. Single gene sex determination alone should not select for recombination suppression[91]. However, the presence of gonadal dysgenesis in XY individuals with an SRY deletion[92] and sterility of XX individuals containing an inactivated copy of SRY[93-98] suggests that multiple genes are required to produce fertile offspring. The accumulation of sexually antagonistic alleles in a non-recombining portion of the genome could have provided a sufficient selective advantage that outweighed the deleterious effects of reduced recombination due to their synergistic effects on fertility. Despite Mueller's ratchet being implicated in the early stages of Y chromosome degeneration[99], genetic decay due to strong positive selection resulting in hitchhiking events is believed to be responsible for its extensive divergence and continued degeneration[99-102]. This is supported by the stepwise repression of recombination[1] and the correlation of Y-degeneration with levels of female promiscuity in related species comparisons[3,14-15,103]. This suggests that strong positive selection on mutually beneficial alleles at linked sites may drive recombination suppression to become selectively favored.

Ohno's original model of genetic evolution suggesting that newly duplicated genes would be functionally redundant and able to escape purifying selection[104] has mixed empirical evidence[53,84]. The events of large-scale duplications, however, create an environment in which newly duplicated genes or complexes that are beneficial in increased dosage may be retained through purifying selection, while the remainder of duplicates would show a continuum of redundancy. The scale of such duplications would allow a small subset of genes to achieve a beneficial mutation. If one mutation resulted in a novel function or further specialized a gene towards one of its respective functions, the likelihood of the gene being retained would increase[45]. Our analysis has also shown that increasing its functional specificity may relax purifying selection, resulting in further divergence. In the event that this new, beneficial mutation occurred on a highly redundant chromosome, the additional reduction of purifying selection due to isolation of a function in an environment with little to no background selection may selectively specialize the chromosome. The survival of the remaining neutrally evolving sequences on that chromosome would depend on their acquisition of functionally related beneficial mutations. If the chromosome bearing this specialized function captured a pair of alleles that together significantly increased the organism's fitness, selection for recombination suppression may result in an inversion becoming prevalent in the population. As observed on the Y chromosome, the resulting chromosome would now contain a complex of genes maintained through purifying selection, as well as a subset of specialized genes that are rapidly evolving resulting in a period of extensive divergence from its homologous counterpart.

The probability that a new inversion captures an advantageous haplotype can be high[71]; however, for an inversion to become fixed when sexually antagonistic alleles are not present the selective advantage would have to strongly outweigh the negative fitness consequences of reduced recombination[105]. In 1973, Leigh Van Valen showed that the probability of extinction of a population was constant over time and suggested an evolutionary arms race where survival is dependent on a population's ability to adapt to changing selective pressures[106]. During times of intense selective pressure, selection for rapid fixation of a highly advantageous haplotype may have driven recombination suppression to become selectively favored due to the reduced effective population size and increased fixation rate. Recombination suppression events, such as inversions, in the absence of sexual antagonism, would have markedly different evolutionary consequences. This is due in part to inversions only reducing recombination in heterozygotes[7]. If the inversion is driven to fixation, recombination would resume between the new homologous chromosomes. In isolated populations, this divergence from the ancestral chromosome may have been sufficient to create a reproductive barrier, such as in the divergence of ancestral *Equus* populations[107]. As evidenced by the Y chromosome, recombination suppression can also occur progressively[91] and may be related to continued selection for newly introduced, functionally related, beneficial alleles. Subsequent inversions resulting from extended periods of intense selective pressure on the associated functions would continue to drive the degeneration of the chromosome through successive hitchhiking events, increasing its long-term fragility.

For the Y chromosome, or any significantly degraded chromosome to go extinct, its functions would need to be replaced elsewhere or no longer under selective constraint to prevent fitness consequences[24]. Relaxation of selection at the locally adapted sites (e.g., predator/prey adaption), however, would render the genes functionally inert, and selection for recombination resumption between ubiquitously expressed genes would result in fusion events becoming selectively favored. Thus, prolonged strong selection for specialization would have driven a subset of ancestral chromosomes to extinction. As this pertains to the terminal fate of the human Y chromosome, continued selection for localization of sexually antagonistic traits, reduction of female promiscuity resulting in less intensive sexual selection, and its recent stability may suggest it is here to stay. Its continued survival in species still experiencing strong sexual selection, however, may be suspect.

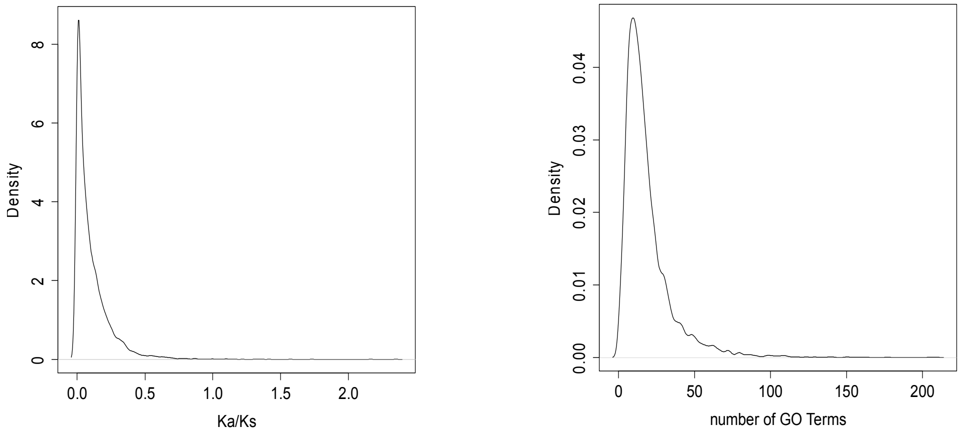
It is worth noting that an apparent contradiction in this logic is the ZW sex-determining chromosomes in avian lineages in which females are the heterogametic sex. Similar to the evolution of the Y chromosome, suppression of recombination in the W chromosome has resulted in significant degeneration of its ancestral gene content[108]. Those that remain functionally active have been shown to be ubiquitously expressed and are believed to be essential to both sexes[108]. In contrast, the W chromosome lacks genes coding for female-advantage traits[108], suggesting that selection for specialization has not resulted in its degeneration. To reduce the recombination between mutually beneficial, sexually antagonistic alleles, however, an inversion can occur in either chromosome[91]. DMRT1 has also been implicated as the sex-determining locus in the ZW system and is present only on the Z chromosome[109-110], suggesting testes development may function through a dosage-dependent mechanism. In conjunction with a lack of dosage-compensation observed in the ZW system[111], the degeneration of the W chromosome is still a result of male-driven positive selection, only on the opposite chromosome.

The chicken Z has been found to be orthologous to portions of human chromosomes 5, 9, and 18 while the human X is orthologous to chicken chromosomes 1 and 4[57,108]. Due to a lack of structural similarity with their respective orthologous regions, researchers have suggested the chicken Z and human X chromosome were not predisposed to become sex chromosomes and their low gene density is a result of convergent evolution[57]. Chromosomes 5, 9, and 18, however, show similar levels of ortholog density as the X chromosome in our analysis, as well as, some of the most substantial differences in mammalian ortholog content between the arms of individual chromosomes. This phenomenon is most likely a result of the arms being derived from different ancestral chromosomes that underwent fusion events[27]. Chromosome 9p also contains the ortholog of the believed avian sex-determining locus DMRT1 and is functionally related to the short arm of the X chromosome in our analysis. This may indicate that the convergent evolution of the sex chromosomes is a result of the differential fusion of ancestral chromosomes containing sex-related genes, and the process of sex chromosome evolution further lowered overall gene density to their current state. However, further research needs to be conducted to determine the significance of this relationship.

The results of our analysis in the context of existing literature present a model by which chromosomes, and therefore populations, rapidly evolved at the onset of the vertebrate lineage. The large-scale duplication events allowed a subset of genes to subfunctionalize, thereby reducing pleiotropic constraints and accelerating evolutionary rates. The isolation of these genes on redundant chromosomes further relieved purifying selection, resulting in a period of rapid chromosomal evolution and divergence due to specialization. If this divergence alone did not create a reproductive barrier, the chromosome's eventual loss due to a change in adaptive pressures would have resulted in differential karyotypes of isolated populations. Thus, the extinction of chromosomes due to specialization is not unique to the Y chromosome, or sex chromosomes in general.

APPENDIX

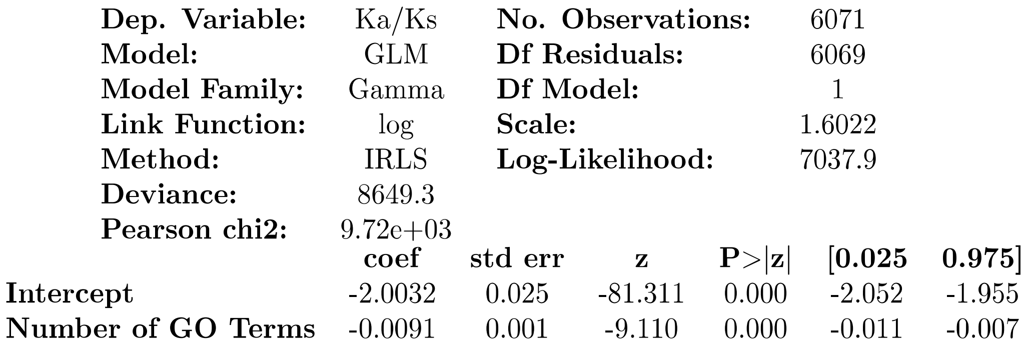
SUPPLEMENTARY MATERIAL



Supplementary Figure 1 Ortholog Density Plots: Distribution of average K_a/K_s values (left) and number of related GO terms (right) for the 6734 human genes with orthologous sequence comparisons in our dataset. For summary statistics see table below. The distribution of average K_a/K_s and number of GO terms both show strong positive skew (skew statistic = 3.92 and 3.06, respectively). The distribution of average K_a/K_s values is also zero-inflated, containing 663 genes that were found to have entirely conserved protein sequences.

Supplementary Table 1 Gene Summary Statistics

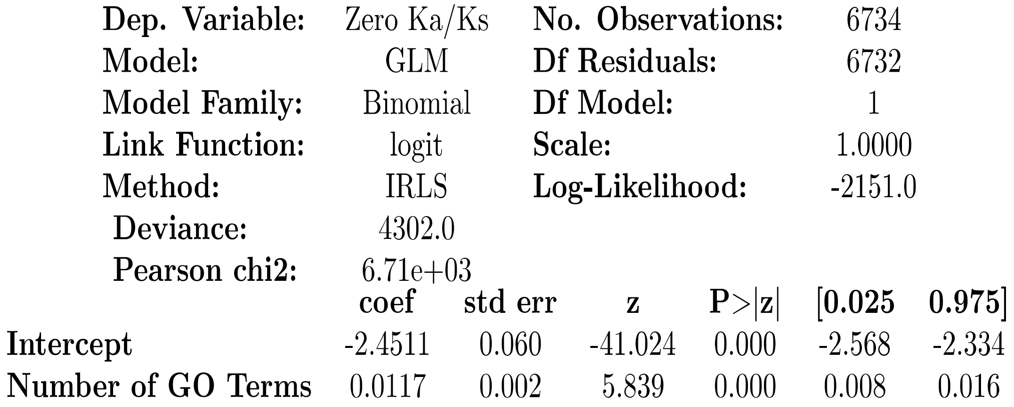
Statistic	Number of GO Terms	K a /K s
mean	18.87	0.1039
stdev	16.61	0.1423
min	1	0
max	209	2.35



Dep. Variable:	Ka/Ks	**No. Observations:**	6071
Model:	GLM	**Df Residuals:**	6069
Model Family:	Gamma	**Df Model:**	1
Link Function:	log	**Scale:**	1.6022
Method:	IRLS	**Log-Likelihood:**	7037.9
Deviance:	8649.3		
Pearson chi2:	9.72e+03		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-2.0032	0.025	-81.311	0.000	-2.052	-1.955
Number of GO Terms	-0.0091	0.001	-9.110	0.000	-0.011	-0.007

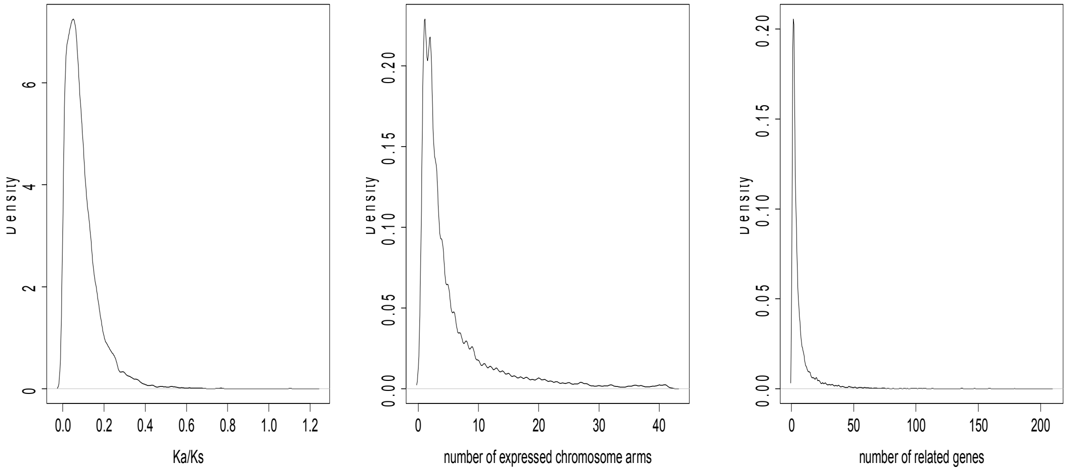
Supplementary Figure 2 GLM Gamma Regression Results 1: Using gamma regression and a log link, non-zero average K_a/K_s values of human genes across their surviving orthologs were fit using an intercept and their number of associated GO terms as predictor variables. The exponential of the coefficient for the intercept and number of GO terms, therefore, represent the initial predicted K_a/K_s value and rate of change for a one-unit increase in number of GO terms, respectively. The intercept, as well as a gene's number of associated GO terms were found to be significant.



Dep. Variable:	Zero Ka/Ks	**No. Observations:**	6734
Model:	GLM	**Df Residuals:**	6732
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2151.0
Deviance:	4302.0		
Pearson chi2:	6.71e+03		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-2.4511	0.060	-41.024	0.000	-2.568	-2.334
Number of GO Terms	0.0117	0.002	5.839	0.000	0.008	0.016

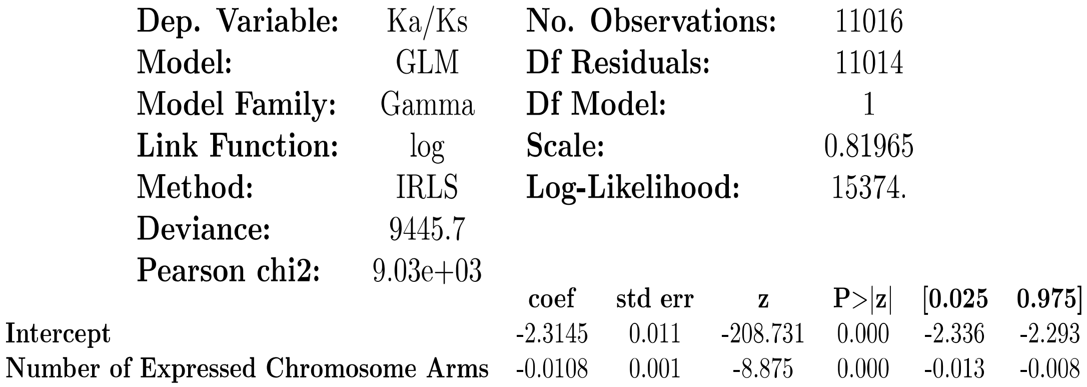
Supplementary Figure 3 GLM Binomial Regression Results: Using binomial regression, the probability a human gene's average K_a/K_s value was zero was fit using an intercept and number of associated GO terms as the predictor variables. The intercept and a gene's number of related GO terms were both found to be significant. The odds-ratio of a gene being entirely conserved to not can, therefore, be determined by the exponential of the linear equation of predictors.



Supplementary Figure 2 GO Annotation Density Plots: Distribution of average K_a/K_s values (left), number of expressed chromosome arms (middle), and number of related genes (right) for the 11016 GO terms with unique gene sets in our dataset. For summary statistics, see table below.

Table 1 Ontology Summary Statistics

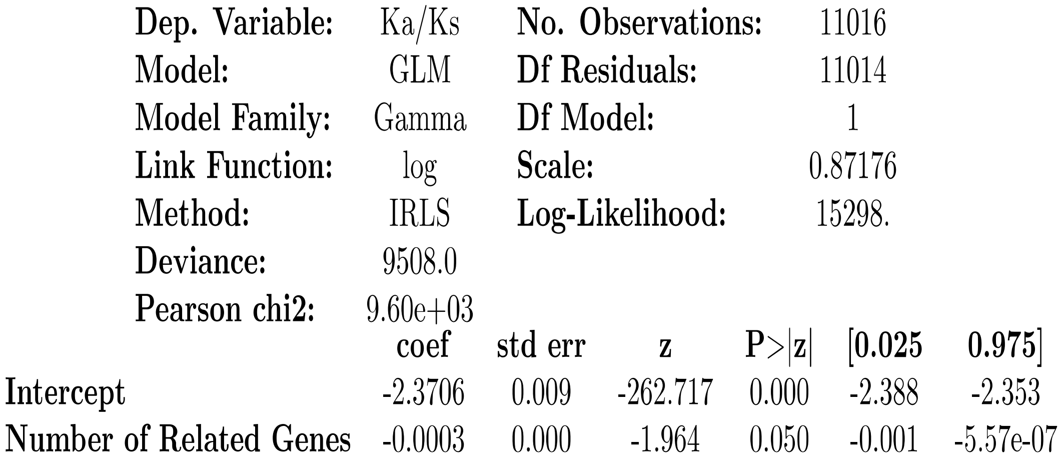
Statistic	Number of Related Genes	Number of Expressed Chromosome arms	Ka/Ks
mean	11.216	5.71	0.093
stdev	66.019	7.07	0.087
min	1	1	0
max	4105	42	1.217



Dep. Variable:	Ka/Ks	**No. Observations:**	11016
Model:	GLM	**Df Residuals:**	11014
Model Family:	Gamma	**Df Model:**	1
Link Function:	log	**Scale:**	0.81965
Method:	IRLS	**Log-Likelihood:**	15374.
Deviance:	9445.7		
Pearson chi2:	9.03e+03		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-2.3145	0.011	-208.731	0.000	-2.336	-2.293
Number of Expressed Chromosome Arms	-0.0108	0.001	-8.875	0.000	-0.013	-0.008

Supplementary Figure 3 GLM Gamma Regression Results 2: Using gamma regression and a log link, non-zero K_a/K_s values of GO annotations averaged across all related human genes and their surviving orthologs were fit using an intercept and the number of chromosome arms they are expressed on. The exponential of the coefficient for the intercept and number of expressed chromosome arms, therefore, represent the initial predicted K_a/K_s value and rate of change for a one-unit increase in chromosome arms expressed, respectively. The intercept, as well as the number of chromosome arms a GO term is expressed on were found to be significant.



Dep. Variable:	Ka/Ks	**No. Observations:**	11016
Model:	GLM	**Df Residuals:**	11014
Model Family:	Gamma	**Df Model:**	1
Link Function:	log	**Scale:**	0.87176
Method:	IRLS	**Log-Likelihood:**	15298.
Deviance:	9508.0		
Pearson chi2:	9.60e+03		

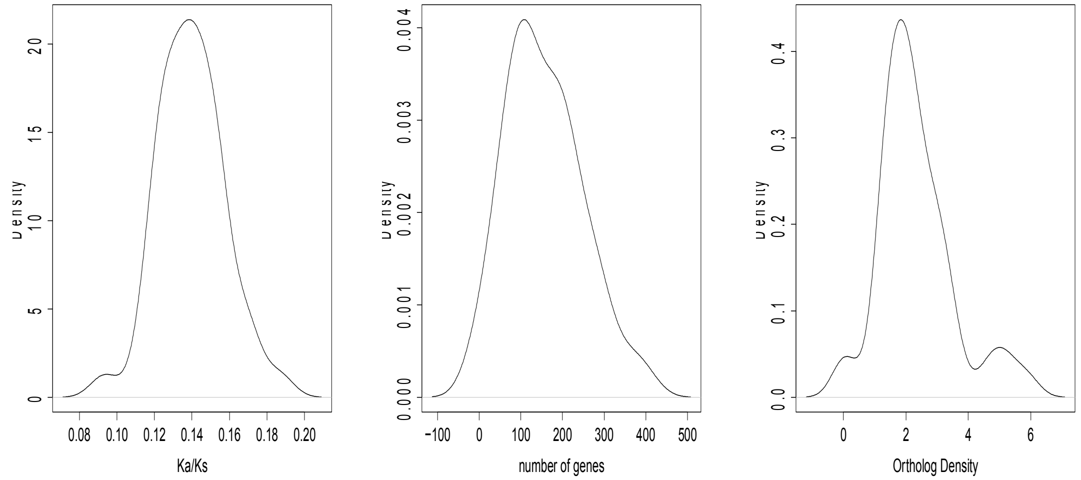
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-2.3706	0.009	-262.717	0.000	-2.388	-2.353
Number of Related Genes	-0.0003	0.000	-1.964	0.050	-0.001	-5.57e-07

Supplementary Figure 4 GLM Gamma Regression Results 3: Using gamma regression and a log link, non-zero K_a/K_s values of GO annotations averaged across all related human genes and their surviving orthologs were fit using an intercept and the number of genes related to a given GO term. The exponential of the coefficient for the intercept and number of related genes, therefore, represent the initial predicted K_a/K_s value and rate of change for a one-unit increase in number or related genes, respectively. The intercept was found to be significant. However, a GO term's number of related genes did not significantly influence its average K_a/K_s value.

Dep. Variable:	Number of Genes	**No. Observations:**	11020
Model:	GLM	**Df Residuals:**	11018
Model Family:	Gamma	**Df Model:**	1
Link Function:	log	**Scale:**	0.17410
Method:	IRLS	**Log-Likelihood:**	-21580.
Deviance:	2125.0		
Pearson chi2:	1.92e+03		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	0.5542	0.005	108.464	0.000	0.544	0.564
Number of Expressed Chromosome Arms	0.1554	0.001	276.393	0.000	0.154	0.157

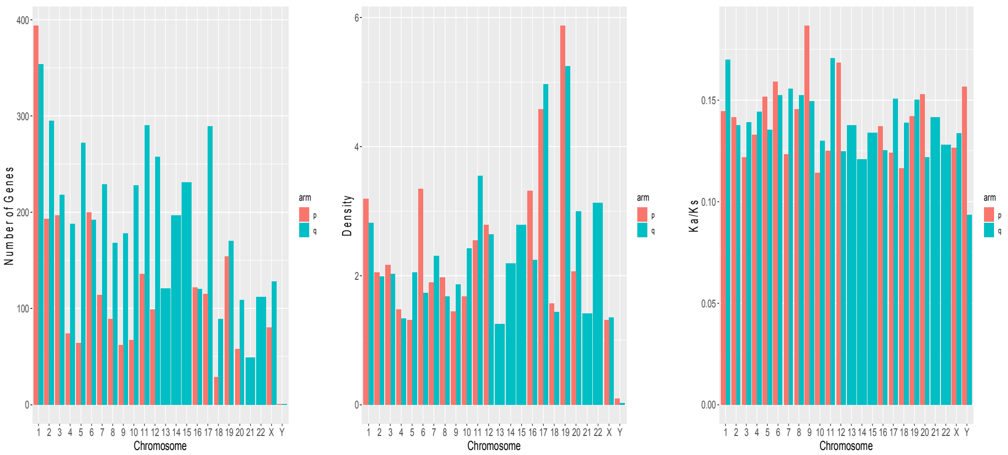
Supplementary Figure 5 GLM Gamma Regression Results 4: Using gamma regression and a log link, GO term's number of related genes were fit using an intercept and the number of chromosome arms they were expressed on. The exponential of the coefficient for the intercept and number of related genes, therefore, represent the initial predicted number of genes and rate of change for a one-unit increase in number of expressed chromosome arms, respectively. The intercept, as well as the number of chromosome arms a GO term is expressed on were found to be significant.



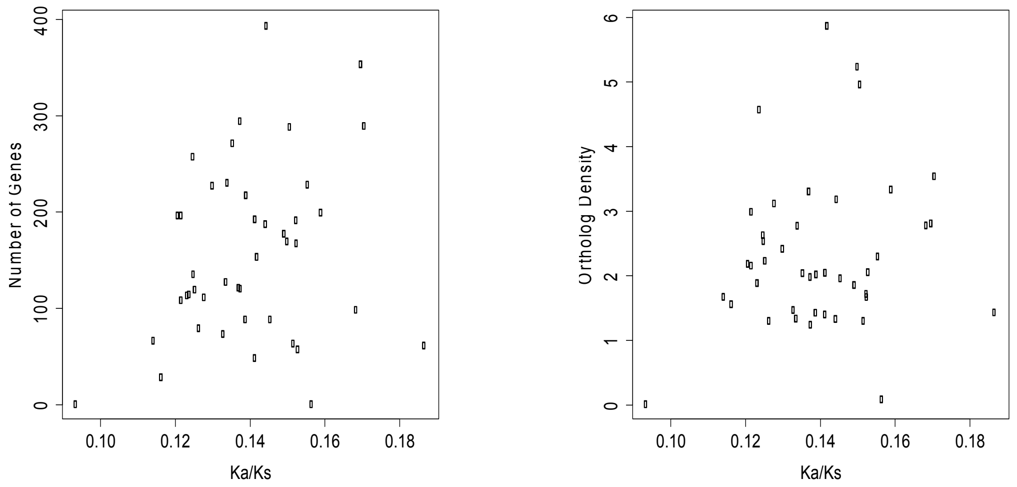
Supplementary Figure 6 Chromosome Arm Density Plots: Distribution of average K_a/K_s values (left), number of orthologous genes (middle), and orthologs/Mb (right) on the 43 chromosome arms. For summary statistics, see table below. The human chromosome arms have normally distributed numbers of orthologous genes (Shapiro-Wilk 0.97, p = 0.34) and average K_a/K_s values (Shapiro-Wilk 0.98, p = 0.72). Density of orthologs, however, is not normally distributed (Shapiro-Wilk 0.91, p = 0.002)

Table 2 Chromosome Arm Summary Statistics

Statistic	Number of Genes	Number of GO Terms	Density	Ka/Ks
mean	156.6	1543.44	2.33	0.1397
stdev	91.56	746.434	1.20	0.0175
min	1	7	0.02	0.0935
max	394	3096	5.88	0.1867



Supplementary Figure 7 Chromosome Arm Bar Charts: Number of orthologous genes (left), orthologs/Mb (middle), and average K_a/K_s (right). The bar graphs provide visual representation of the disparity between the arms of different chromosomes, as well as the separate arms of individual chromosomes.



Supplementary Figure 10 Chromosome Arm Scatter Plots: Number of orthologous genes (left) and orthologs/Mb (right) versus average K_a/K_s values. As stated in the main text, we were unable to find a significant relationship between a chromosome arm's number of related genes or orthologs/Mbp with its average K_a/K_s value. The random distribution of these variables among average K_a/K_s values, therefore, suggests selection at the chromosome arm level has not significantly impacted gene retention.

Dep. Variable:	Number of GO Terms	**R-squared:**	0.985
Model:	OLS	**Adj. R-squared:**	0.985
Method:	Least Squares	**F-statistic:**	2769.
No. Observations:	43	**Prob (F-statistic):**	5.68e-40
Df Residuals:	42	**Log-Likelihood:**	-290.76
Df Model:	1		
AIC:	583.5		
BIC:	585.3		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Number of Genes	9.3872	0.178	52.617	0.000	9.027	9.747

Supplementary Figure 11 Chromosome Arm Ordinary Least Squares (OLS) Regression Results. Using linear regression, a chromosome arm's number of related GO terms was predicted based on the number of orthologous genes found on the chromosome arm, the results of which were found to be highly significant. A chromosome arm's number of functional annotations, therefore, is linearly related to the number of genes on a given chromosome arm by a factor of 9.3872. Our adjusted R-squared value also indicates that this trend should hold across multiple datasets.

REFERENCES

1. Lahn, B. T. & Page, D. C. Four Evolutionary Strata on the Human X Chromosome. *Science* **286**, 964–967 (1999).
2. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
3. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012).
4. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
5. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**, 1563–1572 (2000).
6. Lemaitre, C. *et al.* Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. *Genome Biol Evol* **1**, 56–66 (2009).
7. Kirkpatrick, M. How and Why Chromosome Inversions Evolve. *PLoS Biol* **8**, (2010).
8. Graves, J. A. M. Sex Chromosome Specialization and Degeneration in Mammals. *Cell* **124**, 901–914 (2006).
9. Aitken, R. J. & Graves, J. A. M. Human spermatozoa: The future of sex. *Nature* **415**, 963 (2002).
10. Graves, J. A. M. The rise and fall of SRY. *Trends in Genetics* **18**, 259–264 (2002).

11. Arakawa, Y., Nishida-Umehara, C., Matsuda, Y., Sutou, S. & Suzuki, H. X-chromosomal localization of mammalian Y-linked genes in two XO species of the Ryukyu spiny rat. *Cytogenet. Genome Res.* **99**, 303–309 (2002).

12. Just, W. *et al.* Absence of Sry in species of the vole Ellobius. *Nature Genetics* **11**, 117 (1995).

13. Kuroiwa, A., Ishiguchi, Y., Yamada, F., Shintaro, A. & Matsuda, Y. The process of a Y-loss event in an XO/XO mammal, the Ryukyu spiny rat. *Chromosoma* **119**, 519–526 (2010).

14. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).

15. Hughes, J. F. *et al.* Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437**, 100 (2005).

16. Rozen, S., Marszalek, J. D., Alagappan, R. K., Skaletsky, H. & Page, D. C. Remarkably Little Variation in Proteins Encoded by the Y Chromosome's Single-Copy Genes, Implying Effective Purifying Selection. *Am J Hum Genet* **85**, 923–928 (2009).

17. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).

18. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).

19. Teitz, L. S., Pyntikova, T., Skaletsky, H. & Page, D. C. Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages. *Am J Hum Genet* **103**, 261–275 (2018).

20. Marais, G. A. B., Campos, P. R. A. & Gordo, I. Can Intra-Y Gene Conversion Oppose the Degeneration of the Human Y Chromosome? A Simulation Study. *Genome Biol Evol* **2**, 347–357 (2010).

21. Connallon, T. & Clark, A. G. Gene Duplication, Gene Conversion and the Evolution of the Y Chromosome. *Genetics* **186**, 277–286 (2010).

22. Bachtrog, D. A dynamic view of sex chromosome evolution. *Current Opinion in Genetics & Development* **16**, 578–585 (2006).

23. Kaiser, V. B., Zhou, Q. & Bachtrog, D. Nonrandom Gene Loss from the Drosophila miranda Neo-Y Chromosome. *Genome Biol Evol* **3**, 1329–1337 (2011).

24. Bachtrog, D. Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nat Rev Genet* **14**, 113–124 (2013).

25. Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278**, 675–680 (1997).

26. Page, D. C., Harper, M. E., Love, J. & Botstein, D. Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* **311**, 119–123 (1984).

27. Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biology* **19**, 166 (2018).

28. Dehal, P. & Boore, J. L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol* **3**, (2005).

29. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).

30. Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254–1265 (2007).

31. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* **2**, 333 (2001).

32. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7**, R43 (2006).

33. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**, 97–108 (2010).

34. Lynch, M. & Conery, J. S. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290**, 1151–1155 (2000).

35. Gout, J.-F., Kahn, D., Duret, L. & Consortium, P. P.-G. The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. *PLOS Genetics* **6**, e1000944 (2010).

36. Birchler, J. A. & Veitia, R. A. The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. *Plant Cell* **19**, 395–402 (2007).

37. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10**, 725–732 (2009).

38. Singh, P. P., Arora, J. & Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLOS Computational Biology* **11**, e1004394 (2015).

39. Makino, T. & McLysaght, A. Positionally biased gene loss after whole genome duplication: Evidence from human, yeast, and plant. *Genome Res* **22**, 2427–2435 (2012).

40. Tu, Z. *et al.* Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* **7**, 31 (2006).

41. Zhang, L. & Li, W.-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21**, 236–239 (2004).

42. Yang, J., Gu, Z. & Li, W.-H. Rate of Protein Evolution Versus Fitness Effect of Gene Deletion. *Mol Biol Evol* **20**, 772–774 (2003).

43. Duret, L. & Mouchiroud, D. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Mol Biol Evol* **17**, 68–070 (2000).

44. Torgerson, D. G. & Singh, R. S. Rapid Evolution Through Gene Duplication and Subfunctionalization of the Testes-Specific α4 Proteasome Subunits in Drosophila. *Genetics* **168**, 1421–1432 (2004).

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

46. Wagner, A. The fate of duplicated genes: loss or new function? *BioEssays* **20**, 785–788 (1998).

47. Dermitzakis, E. T. & Clark, A. G. Differential Selection After Duplication in Mammalian Developmental Genes. *Mol Biol Evol* **18**, 557–562 (2001).

48. Li, W.-H., Yang, J. & Gu, X. Expression divergence between duplicate genes. *Trends in Genetics* **21**, 602–607 (2005).

49. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics* **9**, 938–950 (2008).

50. Rastogi, S. & Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5**, 28 (2005).

51. He, X. & Zhang, J. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics* **169**, 1157–1164 (2005).

52. Wilson, M. A. & Makova, K. D. Evolution and Survival on Eutherian Sex Chromosomes. *PLoS Genet* **5**, (2009).

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

54. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014).

55. Mueller, J. L. *et al.* Independent specialization of the human and mouse X chromosomes for the male germline. *Nat Genet* **45**, 1083–1087 (2013).

56. Saifi, G. M. & Chandra, H. S. An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proc Biol Sci* **266**, 203–209 (1999).

57. Bellott, D. W. *et al.* Convergent Evolution of Chicken Z and Human X Chromosomes by Expansion and Gene Acquisition. *Nature* **466**, 612–616 (2010).

58. Blackmon, H. & Brandvain, Y. Long-Term Fragility of Y Chromosomes Is Dominated by Short-Term Resolution of Sexual Antagonism. *Genetics* **207**, 1621–1629 (2017).

59. Rice, W. R. The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination Between Primitive Sex Chromosomes. *Evolution* **41**, 911–914 (1987).

60. Rice, W. R. Evolution of the Y Sex Chromosome in AnimalsY chromosomes evolve through the degeneration of autosomes. *BioScience* **46**, 331–343 (1996).

61. Charlesworth, D. & Charlesworth, B. Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genet. Res.* **35**, 205–214 (1980).

62. van Doorn, G. S. & Kirkpatrick, M. Turnover of sex chromosomes induced by sexual conflict. *Nature* **449**, 909–912 (2007).

63. Matsumoto, T. & Kitano, J. The intricate relationship between sexually antagonistic selection and the evolution of sex chromosome fusions. *J. Theor. Biol.* **404**, 97–108 (2016).

64. Charlesworth, B. The evolution of chromosomal sex determination and dosage compensation. *Current Biology* **6**, 149–162 (1996).

65. Saxena, R. *et al.* The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nature Genetics* **14**, 292 (1996).

66. Lahn, B. T. & Page, D. C. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat. Genet.* **21**, 429–433 (1999).

67. Bhowmick, B. K., Satta, Y. & Takahata, N. The origin and evolution of human ampliconic gene families and ampliconic structure. *Genome Res* **17**, 441–450 (2007).

68. Zhou, Q. & Bachtrog, D. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* **337**, 341–345 (2012).

69. Wyckoff, G. J., Li, J. & Wu, C.-I. Molecular Evolution of Functional Genes on the Mammalian Y Chromosome. *Mol Biol Evol* **19**, 1633–1636 (2002).

70. Wyckoff, G. J., Wang, W. & Wu, C.-I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304 (2000).

71. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).

72. Wyckoff, G. J., Malcom, C. M., Vallender, E. J. & Lahn, B. T. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends in Genetics* **21**, 381–385 (2005).

73. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* **39**, 309–338 (2005).

74. Vallender, E. J., Paschall, J. E., Malcom, C. M., Lahn, B. T. & Wyckoff, G. J. SPEED: a molecular-evolution-based database of mammalian orthologous groups. *Bioinformatics* **22**, 2835–2837 (2006).

75. Ferger, W. F. The Nature and Use of the Harmonic Mean. *Journal of the American Statistical Association* **26**, 36–40 (1931).

76. Huntley, R. P. *et al.* The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.* **43**, D1057-1063 (2015).

77. du Plessis, L., Škunca, N. & Dessimoz, C. The what, where, how and why of gene ontology—a primer for bioinformaticians. *Brief Bioinform* **12**, 723–735 (2011).

78. Škunca, N., Altenhoff, A. & Dessimoz, C. Quality of Computationally Inferred Gene Ontology Annotations. *PLOS Computational Biology* **8**, e1002533 (2012).

79. Martínez, O. & Reyes-Valdés, M. H. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *PNAS* **105**, 9709–9714 (2008).

80. Gustavsson, S., Fagerberg, B., Sallsten, G. & Andersson, E. M. Regression Models for Log-Normal Data: Comparing Different Methods for Quantifying the Association between Abdominal Adiposity and Biomarkers of Inflammation and Insulin Resistance. *Int J Environ Res Public Health* **11**, 3521–3539 (2014).

81. Gustavsson, S. Evaluation of Regression Methods for Log-Normal Data. 63.

82. Tong, E. N. C., Mues, C. & Thomas, L. A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting* **29**, 548–562 (2013).

83. Nobre, A. A. *et al.* Multinomial model and zero-inflated gamma model to study time spent on leisure time physical activity: an example of ELSA-Brasil. *Rev Saude Publica* **51**, (2017).

84. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biology* **3**, research0008.1 (2002).

85. Kondrashov, F. A. & Koonin, E. V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics* **20**, 287–290 (2004).

86. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLOS Genetics* **5**, e1000471 (2009).

87. Crow, K. D. & Wagner, G. P. What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Mol Biol Evol* **23**, 887–892 (2006).

88. Bush, G. L., Case, S. M., Wilson, A. C. & Patton, J. L. Rapid speciation and chromosomal evolution in mammals. *Proc Natl Acad Sci U S A* **74**, 3942–3946 (1977).

89. Livingstone, K. & Rieseberg, L. Chromosomal evolution and speciation: a recombination-based approach: Research review. *New Phytologist* **161**, 107–112 (2003).

90. Foster, J. W. & Graves, J. A. An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 1927–1931 (1994).

91. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118 (2005).

92. Hughes, J. F. & Rozen, S. Genomics and Genetics of Human and Primate Y Chromosomes. *Annu. Rev. Genom. Hum. Genet.* **13**, 83–108 (2012).

93. Chapelle, A. de la, Tippett, P. A., Wetterstrand, G. & Page, D. Genetic evidence of X–Y interchange in a human XX male. *Nature* **307**, 170–171 (1984).

94. Page, D. C., Chapelle, A. de la & Weissenbach, J. Chromosome Y-specific DNA in related human XX males. *Nature* **315**, 224–226 (1985).

95. Pepene, C. E., Coman, I., Mihu, D., Militaru, M. & Duncea, I. Infertility in a new 46, XX male with positive SRY confirmed by fluorescence in situ hybridization: a case report. *Clin Exp Obstet Gynecol* **35**, 299–300 (2008).

96. Anık, A., Çatlı, G., Abacı, A. & Böber, E. 46,XX Male Disorder of Sexual Development: A Case Report. *J Clin Res Pediatr Endocrinol* **5**, 258–260 (2013).

97. Wang, T., Liu, J. H., Yang, J., Chen, J. & Ye, Z. Q. 46, XX male sex reversal syndrome: a case report and review of the genetic basis. *Andrologia* **41**, 59–62 (2009).

98. Wu, Q.-Y. *et al.* Clinical, molecular and cytogenetic analysis of 46, XX testicular disorder of sex development with SRY-positive. *BMC Urology* **14**, 70 (2014).

99. Bachtrog, D. The Temporal Dynamics of Processes Underlying Y Chromosome Degeneration. *Genetics* **179**, 1513–1525 (2008).

100. Kuroki, Y. *et al.* Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nature Genetics* **38**, 158 (2006).

101. Bachtrog, D. Evidence that positive selection drives Y-chromosome degeneration in Drosophila miranda. *Nature Genetics* **36**, 518 (2004).

102. Rice, W. R. Genetic Hitchhiking and the Evolution of Reduced Genetic Activity of the Y Sex Chromosome. *Genetics* **116**, 161–167 (1987).

103. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nature Genetics* **36**, 1326–1329 (2004).

104. Ohno, S. *Evolution by Gene Duplication*. (Springer-Verlag, New York, 1970).

105. Charlesworth, B., Betancourt, A. J., Kaiser, V. B. & Gordo, I. Genetic Recombination and Molecular Evolution. *Cold Spring Harb Symp Quant Biol* **74**, 177–186 (2009).

106. Van Valen, L. A New Evolutionary Law. *Evolutionary Theory*, **1**, 1-30 (1973).

107. Renaud, G. *et al.* Improved de novo genomic assembly for the domestic donkey. *Science Advances* **4**, eaaq0392 (2018).

108. Bellott, D. W. *et al.* Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nature Genetics* **49**, 387–394 (2017).

109. Smith, C. A. *et al.* The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. *Nature* **461**, 267–271 (2009).

110. Marshall Graves, J. A. Sex determination: Birds do it with a Z gene. *Nature* **461**, 177–178 (2009).

111. Vicoso, B. & Bachtrog, D. Progress and prospects toward our understanding of the evolution of dosage compensation. *Chromosome Res* **17**, (2009)

VITA

Jason Ryon Wilson was born on October 7, 1994, in Overland Park, Kansas. He was educated in the Blue Valley School District and graduated from Blue Valley Southwest in 2013. He graduated from Kansas State University in 2016 with a Bachelor of Science in Biology degree.

After graduation, he worked as a laboratory assistant for Dr. Mark Ungerer at Kansas State. His work focused on plant genomics using bioinformatic and molecular biology approaches. This work generated a manuscript on intraspecific nuclear genome size variation and is nearing submission. This research experience ultimately led him to pursue a degree in bioinformatics.

In 2018, Jason began working toward a M.S. in bioinformatics at the University of Missouri-Kansas City School of Medicine. Upon completion of his degree requirements, Jason intends to pursue a Ph.D. in genetics.

Jason is a member of the National Society of Collegiate Scholars, UMKC Golden Key, and Phi Beta Kappa Honor Societies.