THE UNIVERSITY OF CHICAGO

READING-FRAME SHIFT MECHANISMS OF INTRODUCING GENETIC NOVELTY TO THE HUMAN GENOME

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY

ALEXANDER ADVANI

CHICAGO, ILLINOIS

JUNE 2020

To my family and friends.

You have encouraged me to aim higher.

You have supported me in every endeavor.

You have pushed me to be better.

Thank you all.

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are many people I have to thank who have supported me, encouraged and made it possible for me to survive and succeed at this journey. First and foremost, I would like to thank my advisor, Manyuan Long, for his unwavering support and constant consideration throughout my time at the University of Chicago. I am very grateful for the opportunities he gave me to develop a research project I was passionate about and for making my well-being and happiness his primary concern. I was always inspired by his passion for science and impressed by his diverse interests and encyclopedic knowledge.

I would also like to thank the many faculty who have mentored and taught me in my career thus far. My committee members and the chairs of my program, Dick Hudson, Chip Ferguson, Urs Schmidt-Ott, Yoav Gilad and Marcelo Nobrega, who always made sure I was progressing throughout graduate school. Greg Wray who took me in as an undergraduate and started me down this path. The many high school teachers who taught me to love science and to strive to make an impact in the world.

I would also like to thank the Long Lab members, past and present, of whom there are too many to name but who have made countless contributions to my work over the years and given me the encouragement and assistance I needed to succeed. In particular I would very much like to thank Claus Kemkemer who trained me, Nick VanKuren who gave me the pointers I needed to get my project off the ground and UnJin Lee who gave me the creative discussions and support I needed to get over the finish line.

I am especially grateful for Sue Levison without whom none of this would have been possible. She has chased me relentlessly over the years with my best interests at heart and I am very much the better for it. I would not have been able to navigate through graduate school without her.

Perhaps my most significant acknowledgement is to Phil Ross. My friend, roommate and collaborator, Phil has been an invaluable addition to my life and through our collaboration an immeasurable benefit to my career. I consider myself very lucky to have been paired up with Phil during recruitment and to have developed this friendship.

My friends and colleagues at the University of Chicago, you have made my time here a joy and probably kept me sane over the years. Andrei Anghel you have been an incredible friend and I am very fortunate to have met you. When we first met at recruitment seven years ago I had no idea we would bond over rewards programs and usually being the weedsiest person in the room but I'm very glad I got the opportunity to find out. Katie Mika you have been the most patient, giving and kind friend anyone could ask for and I would like to thank you very much for putting up with me. Bill Richter, Diedre Reitz, Alex Gileta, Bryan Pavlovic and Aarti Venkat you have taught me so many things and we have had so many adventures together. Thank you for the game nights and poker nights and road trips to retreat and the countless lunches that made graduate school so memorable.

My friends from Cyprus your friendship for just about two decades means more to me than I can write here. Bambo, Polis, Andri, Annabel, Demetris, Charis, Eugenie, George, Jovanna, Stavria, Angelina, Markella, Evgenia, Marinos and everyone there, I wouldn't be who I am today without you.

To my family I love you very much and I will always be grateful for your constant love, support and guidance. My parents, Sudi and Ioli, I am so privileged to have been born your son. You have sacrificed everything to give me every opportunity that you could and I am incredibly grateful for that. For every night you stayed up late for me, for every New Year's Party we almost missed because I was submitting an application, for every trip and experience I have had thank you so much. My amazing sisters, Nolo and Dani, thank you for always being there for me and your infinite patience. I'm so proud of both of you. My aunt Nalini, you have always been there for me and given me everything I could have ever asked for. The books you bought me as a child are probably the reason I ended up in graduate school and on this career path. To Yiannis, Beverly, Thalia and Vania, thank you for your love and support. You have helped make me who I am and I am extremely fortunate to have you in my life. To my grandparents, Vassos and Popi, thank you for your love and care and patience. I know you have been waiting endlessly for this to be written.

Last but definitely not least, my greatest supporter, Laura Cioffi. I definitely would not be here without you. You have been my rock and have helped me climb this mountain. Thank you for your love and unwavering encouragement. Thank you for introducing me to so many new things and helping me discover aspects of my life I didn't even know I was missing. The day I met you was one of the luckiest days of my life and I will always be grateful for you.

Abstract

How genetic novelty arises is one of the most elusive questions within genetics and has implications across numerous and diverse fields. Although there are many possible answers, reading frame-shifts in protein-coding DNA are known to create dramatically different peptides and have the potential to enable large evolutionary steps. The radical nature of these mutations have led to the assumption that they do not often survive and are strongly selected against. Nevertheless, when they occur in a duplicated gene, many of the negative selective pressures are alleviated. In this thesis I use a conservative method which proves that this is a mechanism by which genetic material has been commonly introduced to the human genome. In addition, I determine the characteristics which human genes formed by this mechanism most commonly share and the roles they have played in human evolution. Finally, I discuss the effects of frameshifting and the cooption of frameshifted genes on human evolution and more broadly genomic adaptation across the tree of life.

Chapter 1: Introduction

One of the central questions of evolutionary genetics is where genes come from and how they adapt over time. Understanding the origination of genetic novelties has been a challenge since the advent of the study of genetics [1] [2] [3] and yet there are still many unanswered questions about mechanisms of creating and maintaining genetic novelties underlying the evolutionary leaps in phenotypes that lead to the extant biodiversity we can observe today. The interpretation of the data available to us has fluctuated dramatically over the decades between gradual change [1] [2] [4] [5] and punctuated equilibrium [6] [7] [8] [9] and many intermediate theories [10] [11] [12] [13]. Current consensus on this issue is that evolution is heterogeneous and opportunistic, taking large adaptive leaps when possible and small gradual steps when not. Recent observations and analyses have resulted in a hybrid model which is often applied to gain insights into cancer genome or pathogen evolution [14] [15] [16] [17]. Consequently, understanding how and when genetic novelty arises, permitting a large adaptive step, is critical to understanding how genomes evolve and to advancing many diverse and important fields.

Mechanisms of new gene origination

There are eleven currently known mechanisms of gene transformation that lead to novel genetic material [18]. They can be divided into two categories: mechanisms of new gene formation and mechanisms of introducing genetic novelty. Under the first category there are five mechanisms which create a functional DNA sequence in the genome where there was none before. These are DNA-based gene duplication, retrotransposition, Transposable Element domestication, lateral gene transfer and *de novo* origination. The second category comprises of six mechanisms which drastically change functional genes and create the potential for new functions. These are exon or domain shuffling, gene fusion or fission, reading-frame shift, novel alternative splicing, coding adoption of non-coding RNA and pseudogene as an RNA regulator. Each of these mechanisms has been studied but to varying degrees, usually corresponding to the perceived frequency with which they occur. However, these mechanisms are almost always studied independently and are rarely considered in conjunction with one another, despite the fact that most of them can work together to shape a new gene simultaneously. This dissertation will focus on mechanisms involving a duplication event within the same genome, which are DNA-based gene duplication, retrotransposition or Transposable Element domestication, followed by a reading-frame shift, also called a frameshift.

The goals of this dissertation are to 1) identify genes involved in this combination of mechanisms, 2) identify their most common characteristics and 3) further our understanding of how this combination can lead to an increase in genetic novelty and diversity.

Duplication mechanisms

There are three common mechanisms which duplicate genetic material. Firstly and by far the most common source of new genes is DNA-based duplication mechanisms. These duplications are the result of replication errors caused by the dissociation and incorrect re-association of polymerases during DNA replication [19] [20] [21]. The majority of DNA-based duplications are also tandem duplications [20] [21]. These duplications usually occur during cell division when polymerases slip on their template strand and create an additional segment of DNA which is a duplicate of the segment immediately adjacent to it [20] [21] [22]. The close proximity of the offspring gene to the parent gene means they often inherit the parental regulatory elements and characteristics as well. Less common events involve inverted duplications caused by palindromic sequences or non-tandem duplications which typically rely on sequences with a significant degree of sequence similarity if not homology [22].

The second mechanism is retrotransposition. This involves a DNA sequence being transcribed into RNA and then being reverse transcribed into cDNA which is then incorporated back into the genome [23] [24]. Retrotransposed genes are rarely found near their parent genes and initially lack all features of a mature gene such as introns or gene specific regulatory elements which makes them easily identifiable [23] [24]. However, these features are acquired over time and it is rarely possible to tell if an ancient gene was formed by DNA-based duplication or retrotransposition without identifying the parent gene [23] [24].

The third mechanism this dissertation will touch on is Transposable Element (TE) domestication. Transposable Elements (TEs) are genetic units which exploit cellular functions to proliferate and relocate in a genome [25] [26] [27]. They are usually self-contained entities which include genes to encode the proteins they require to exploit their host cells but a genome will often contain many copies of the same TE [25] [26] [27]. TE proteins can sometimes be domesticated to perform host functions by their host cells and incorporated into new host genes [25] [26] [27]. They are often co-opted to defend the host cell against pathogens or other invaders or to relieve an evolutionary pressure on a pre-existing gene with multiple functions and conflicting selective pressures [25] [27].

Reading-Frame shifts

A reading-frame shift (RFS) is the most common consequence of an insertion or deletion occurring in a coding sequence. Due to the significant impact this can have on a gene there is an expectation that the results of this would be highly deleterious and there is evidence to support this. A RFS can often produce a protein that bears very little resemblance to the unframeshifted version after the point of mutation [28] [29]. Early STOP codons are often observed and the frame-shifted protein is usually predicted to form a non-functional peptide that folds non-specifically [28] [29]. Misfolded proteins are expected to impose a negative fitness cost due to interference with cellular activities and depletion of cellular resources [30] [31]. In addition, frame-shifted proteins have been repeatedly shown to be potentially causal for human disease [32]. However, I would argue that this evidence is a one sided view of this phenomenon and does not encapsulate the full picture.

The majority of studies on frameshift mutations are done in clinical settings or within a medical context. As a result of this focus there is a significant amount of evidence available for the negative consequences of RFS mutations. In humans frameshifts have often been associated with diseases [33] [34] [35] or infertility [36] [37] and the occurrence of a RFS in a pathogen is often reported to have significant adverse effects for human health, such as antibiotic resistance [38] or increased virulence [39]. Far fewer studies have been done on a genomic or evolutionary scale to account for the advantages of a frameshift mutation or on the potential of co-opting frameshifting for adaptation. As a result, while we have a clear picture of the downside of RFS mutations we must be wary of assuming that it is representative of the overall effects of frameshifting. Unfortunately, due to the difficulty of performing genomic analyses until approximately 20 years ago and the consequent lack of evidence for the upside of RFS mutations, that assumption has become very pervasive.

When contextualized properly, reexamining the literature for evidence of how frameshift mutations function and approaching them with a neutral perspective reveals a more complete picture. While there is undoubtable evidence for the drawbacks that are caused by individual RFS mutations, there is some support for the idea that these mutations have had a lasting impact on the evolutionary trajectory of several species, including our own. RFS mutations have been shown to create novel coding sequences [18] but, given the dramatic impact of a frameshift on a pre-existing protein, many sources claim frameshifts are rapidly eliminated from the genome except in rare cases [40] [41] [29] [28]. Though rapid degradation and purging of frameshifted genes has been observed [40] [28], only recently have studies been conducted on the genome wide survival rate of frameshifts [42] [43] [44]. They suggest that frameshift mutations can survive and may indeed play a larger role in our evolutionary history than previously suspected. There is also evidence that the genetic code is far more optimized for frameshift tolerance than previously thought [45]. Additionally, there is recent evidence to suggest that frameshift survival may be a strategy that evolved early on in our history and may not just be a human specific phenomenon [46].

In addition, there are several mechanisms of post-transcriptional frameshifting which add a layer of novelties to protein diversity. This mechanism allows for increased genomic efficiency as multiple peptides can be encoded by the same gene. As these frameshifted peptides are a small percentage of the total output from these genes, this also allows organisms to encode uncommonly used peptides this way and conserve cellular resources. Ribosomal frameshifting is common in prokaryotes [47] [48] [49]and has been observed in eukaryotes as well [50] [51], including in humans [52] [53]. These post-transcriptional frameshifts usually involve one of three mechanisms which can work independently or in conjunction with each other: slippery sequences [47] [52], pseudoknot structures [49] [50] [52] and hypomodified tRNAs [48] [51] [52] [53]. Slippery sequences are short repetitive sequences that correspond to rare tRNAs depending on the species codon bias [47] [52]. They result in a pause in translation until the rare tRNA arrives at the ribosome [47] [52]. During this time due to molecular dissociation/association processes the ribosome can slip on the mRNA and cause a frameshift [47] [51] [52]. Pseudoknot structures involve 3D folds in the mRNA that form an RNA knot via complementary base pairing [49] [50] [52]. When the mRNA is being translated the knot bumps against the ribosome and can cause the mRNA to move back resulting in a frameshift mutation [49] [50]. Finally, hypomodified tRNAs have a modified base which allows them to recognize more than one codon [48] [51] [53]. This can result in the tRNA slipping on the mRNA and causing a frameshift during translation [48] [51] [53]. This can be compounded by a slippery sequence [51] [52]. Hypomodified tRNAs have been identified across eukaryotes and have been shown to be highly conserved [52] [53]. In addition, there is evidence that a hypomodified tRNA can sometimes counteract an encoded frameshift mutation [48] [52].

Models of gene evolution

Starting with Muller's initial description of duplicate genes acquiring new functions, there have been many models used to describe the evolutionary pressures on duplicated genes [54]. To varying extents they all rely on the premise that a duplicated gene will be maintained if the organism benefits from increased abundance or activity of the gene product, something that has been observed multiple times [55] [56] [57]. Some models such as neofunctionalization and subfunctionalization, are older and broader but can still be useful when reconstructing the ancestry of a duplicated gene, particularly in a multi-species comparative analysis [3] [58] [59] [60]. Ohno first described neofunctionalization as a method of increasing genetic variation and novelty in a species but presupposed that duplicate genes are maintained due to the benefits associated with their increased function [3]. The contradictory forces of selecting for both a conserved function and a new function became known as "Ohno's dilemma" [61]. To resolve this, more recent detailed models which address that paradox and apply to specific contexts have been published and testing their predictions against datasets of gene pairs can greatly enhance our understanding of the gene pairs' lineage [62] [63]. Currently, the most commonly used models are Innovation-Amplification-Divergence (IAD), Escape from Adaptive Conflict (EAC) and Adaptive Radiation (AR) [63] [64] [65].

This dissertation examines the prevalence of genes with a duplication event followed by a reading-frame shift mutation in their evolutionary history and explores their potential as a source of genetic novelty. I found that genes formed by this mechanism, called Reading-Frame Shift aided by Duplication (RFSD) genes in this dissertation, are prevalent in the human genome. I investigated their characteristics including their ages, expression patterns and functions and whether they share these with their parent genes. I have determined that they commonly share many of the characteristics of their parent genes but not their molecular functions. This suggests a model whereby the genetic novelty introduced by a novel peptide attached to one or more functional protein domains provides the opportunity to expand the original function of a gene and build more complex biological networks. In addition the survivability of a RFSD gene is potentially increased due to the duplicate unframeshifted copy, which allows for a RFS of a functional protein without severely compromising the original role it held.

Chapter 2: Reading frame-shifts provided an important source of genetic novelty for sex-dependent and signaling functions in *Homo sapiens*

Abstract

Understanding sources of genetic novelty and how our DNA manages mutations is a key concern to many disciplines. One of the most radical and significant forms of genetic change is a reading frame-shift in protein-coding genes. However, this mechanism is conventionally viewed as an extremely rare opportunity for a cell or an organism to take a major adaptive step by co-opting a novel protein sequence attached to one or more functional domains. In this study we identified a large number of human genes formed by this mechanism using conservative criteria in identification pipelines to compare human genomes with those of other vertebrates. Further examination of these frameshift-derived genes found excess novel genes that are located on Y and X chromosomes, indicating extensive generation of novel proteins during the evolution of human sex chromosomes. Frameshift-derived genes also show a high excess of functions related to signaling suggesting this mechanism may have played a significant role in the evolution of the extant diversity of signaling pathways. Finally these genes are significantly more likely to be expressed in mitochondrial proteomes, suggesting that frameshifting may also have played a role in the evolution of energy/metabolism networks. These findings reveal frequent evolution of human genomes by acquiring new sex- and signaling- related gene functions from previously little used reading frames.

Introduction

One of the most central questions to our understanding of modern genetics and biological diversity is how does genetic novelty arise? Studying how genes arise and the way they integrate into preexisting genetic networks is key to many fields, including evolutionary genetics, molecular biology and systems biology. While many mechanisms have been researched extensively [18] [66] [67], genes which undergo a frameshift mutation have been an understudied area, especially when these mutations occur in conjunction with other methods of novel gene origination.

New genes are known to commonly arise via duplication events, either DNA-based or retrotransposed [18] [40] [68] [69]. New genes formed via a frameshift mechanism have also been observed [18] but most sources claim these are rare or short-lived events and play a minor role in evolution [28] [29] [40] [41]. Studies have also suggested that frameshifted genes often produce proteins that misfold or fold non-specifically [70] and misfolded proteins are known to impose a fitness cost on the host organism [30] [31]. However, a few studies have suggested that frameshifted genes survive more frequently than previously thought [42] [43] [71] and that frameshifted protein folding is a lot more plastic [44].

In particular, Okamura et al. published findings in 2006 that strongly suggested frameshifting played a significant role in mammalian evolution [43]. They identified 470 possible frameshift-derived genes in the human genome and 108 in mice, a much higher number than previously suspected [43]. This made the possibility of identifying genes previously unknown to be frameshifted far more likely and raised the importance of several unanswered questions that could now have far broader implications. This study will build on their method of identification and determine with a high level of confidence which genes in the human genome have a frameshift in their evolutionary history and what characteristics this allowed them to acquire. The Okamura et al. study, along with a few others that have raised similar questions, have also resulted in a reevaluation of the role frameshifts play and how they fit in to the broader context of human evolution [42] [43] [71].

Frameshift mutations have been suggested as a source of evolutionary novelty and identified as having a crucial role in the appearance of several new gene families [42] [72]. These families are usually formed by sequential duplication events subsequently followed by frameshift mutations [42] [72]. This suggests a model of evolution in which reading frame-shifts (RFS), often combined with gene duplication so the RFS would not destroy ancestral functions a paralogue encodes, are critical to taking the large adaptive steps required to rapidly escape adaptive constraint and enable neo-functionalization [42]. Furthermore, it is likely that the preservation of a portion of the original protein product is a determining factor in ensuring that these genes can perform a specialized but related function [42] [72]. In addition, newly duplicated genes often have duplicated regulatory elements which can be co-opted to express the novel peptides generated via RFS mutations, bypassing another hurdle towards fully functional and useful additions to the genome [42]. Finally, there is some evidence suggesting that the standard genetic code is optimized for frameshifting, further supporting the theory that RFS mutations have had a meaningful impact on our evolutionary history [46].

Given the potential for frameshift mutations to rapidly diversify duplicated genes and circumvent the initial difficulties with generating new useful peptides, we believe the impact of RFS mutations on evolution is undercounted and undervalued. Large scale genomic analysis is required to thoroughly test the prevalence and significance of RFS mutations in the human genome. Our study is a step towards addressing that gap in the literature by taking a broad view of frameshift mutations in the human genome and determining what genes with a RFS in their evolutionary history are most likely to have in common.

In this study we identify human genes which have been generated by a reading frame-shift aided by gene duplication (RFSD), hereafter called a RFSD event. We then determine the characteristics of genes involved in these events and search for patterns that might provide insights into which genes are more susceptible to being involved in a RFSD event. Finally, we examine the RFSD genes detected, in conjunction with available data, to determine what properties and functions are enriched in these genes and what their impact is on human life. Remarkably we found that these events have occurred with an unexpectedly high frequency throughout the human lineage, revealing peculiar patterns shaped by functional adaptation due to their increase in activity coupled with the genetic novelty that a frameshift can provide.

Methods

Identifying RFSD events

A dataset of all expressed human cDNAs that correspond to known proteins was obtained from Ensembl via Biomart [73]. All analysis was done with data from Ensembl version 91 last accessed in January 2018. This dataset was compared to itself using a custom tblastx function. The tblastx function is available as part of the blast+ package from NCBI. This translated the cDNA of each gene into 6 different possible frames and compared them as a query to the translations of the other genes in the dataset in 6 possible frames (Figure 1). The custom output included the identities and frames of the gene being matched and the gene identified as a match for it, as well as the alignment length, the e-value of each match, the percentage identity, the number of identical positions and number of mismatches in the alignment.

All results were then filtered using the following criteria. The term query gene refers to the gene being matched against the dataset of frameshifted genes. The term subject gene refers to the gene from the dataset being tested for a match to the query gene.

1. Query gene is not the same as the subject gene;
2. Query frame does not match the subject frame;
3. The percentage identity of the match is at least 80%;
4. The e-value of the match is at most 1^e-10;
5. In the event of multiple matches between a query gene and a subject gene, only the match with the longest alignment length is retained.

The output data were then sorted into two datasets, the Standard dataset and the Conservative dataset. The Standard dataset comprises of all data that were filtered as described above and matches the following two criteria:

1. The minimum alignment length of the match is at least 50 amino acids;
2. The query gene and the subject gene must have a reciprocal match.

The Conservative dataset, as a subset of the Standard dataset, comprises of all data that was filtered as described above and matches the following two criteria:

1. The minimum alignment length of the match is at least 100 amino acids;
2. The query gene and the subject gene must have a longer reciprocal match than they have with any other gene, called a reciprocal best match.

Each event identified by the Conservative criteria was then individually examined by running a pairwise blast comparison to confirm that the RFSD model of origination as a true event. This was done to ensure our method is not identifying any artefacts, allowing us to have extremely high confidence in our conclusions.

The choices made in our selection criteria were made to ensure we were able to manage the data output of our search while still having confidence in the datasets we compiled. Firstly, we mandated that all matches were between distinct genes. As a result we automatically excluded any frameshifting that occurs within a gene and any alternative splicing that occurs in a different frame, both of which are known to happen [43] [74]. However, permitting those categories would have meant we had to verify most of those matches experimentally ourselves to ensure that the genes are actually expressed in those frames. If we did not verify them, we would not be able to confidently state that these matches correspond to true events.

Secondly, the minimum percentage identity match required and maximum e-value were chosen to be conservative while not dramatically increasing the false negative rate. The primary concern when we chose these criteria was to ensure that the data identified were undoubtedly real events. However, it is important to note that we did not permit gaps when we ran the blast comparison as a gap could negate a frameshift and give us spurious results.

Thirdly, we chose to only retain the longest match if there were multiple matches between two genes. Multiple matches can occur if the original frame is restored for part of the gene or a secondary frameshift occurs. We elected not to maintain multiple results per gene pair to simplify our data. This choice meant we did not lose any pairs while also ensuring that duplicate matches would not skew our analysis downstream by giving more weight to gene pairs with multiple matches.

Finally, our minimum length choices for the two datasets collected were both high bars to clear and it is very likely that we would have been able to identify many more RFSD events if we had chosen to lower those thresholds. However, without any biologically meaningful limit we could have chosen, we elected to use the large minimum lengths so that any matches identified could be unequivocally called a true event. It is possible that follow up studies could use a lower requirement in order to identify more RFSD pairs and develop a workaround for any false positives introduced.

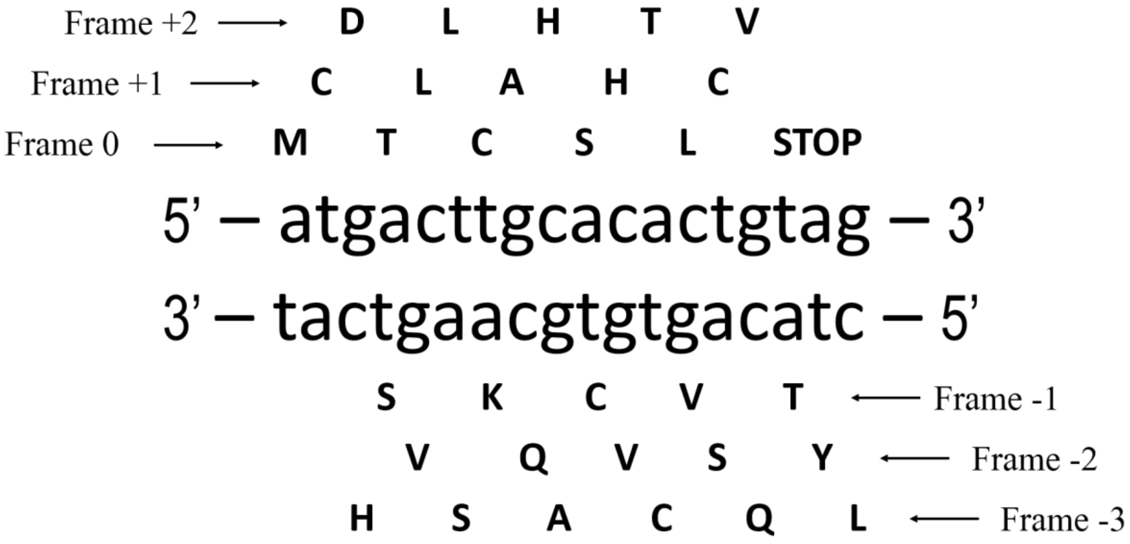


Fig. 1 Visual representation of the method used to translate and match 6 different frames for each input cDNA.

Dating RFSD events

The bioconductor package "biomaRt" [75] [76] was used to connect to BioMart [73] and identify homologs of human genes, involved in RFSD events, in other species. The species used were human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), orangutan (*Pongo abelii*), rhesus macaque (*Macaca mulatta*), marmoset (*Callithrix jacchus*), mouse (*Mus musculus*), guinea pig (*Cavia porcellus*), dog (*Canis familiaris*), cow (*Bos taurus*), armadillo (*Dasypus novemcinctus*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), anole lizard (*Anolis carolinensis*), frog (*Xenopus tropicalis*), coelacanth (*Latimeria chalumnae*), fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), hagfish (*Eptatretus burgeri*), lamprey (*Petromyzon marinus*), nematode (*Caenorhabditis elegans*), fly (*Drosophila melanogaster*) and yeast (*Saccharomyces cerevisiae*).

A branch number was assigned at each point where one or more of these species diverged from the human lineage. Species that share a common ancestor after they diverged from the human lineage are on the same branch. The oldest branch (Branch 0) corresponds to the common ancestor of human and yeast. The youngest branch (Branch 16) is the branch representing the human species after it diverged from chimpanzees. The parsimonious principle was used to determine gene gain or loss. The age of each gene was assigned by the average length of each branch in millions of years [77] [78] [79] [80] [81]. Each RFSD event was assigned to the branch number, and by extension the age, of the younger of the two genes in each pair. If both genes in a pair belong to the same branch the event was assigned to that branch. A z-score was calculated for the number of RFSD genes on each branch normalized by the length of the branch.

The length of each phylogenetic branch was determined using www.timetree.org [82] which uses published peer reviewed studies on the divergence time between species to produce current estimates [77] [78] [79] [80] [81]. The length used was the mean of the lengths found in the studies cited on the website. Though some of the species divergence estimates are variable, the mean of the proposed branch lengths is representative enough to serve as a reasonable approximation. The branch lengths in millions of years are:

Branch 16 – 7

Branch 15 – 7

Branch 14 – 16

Branch 13 – 29

Branch 12 – 43

Branch 11 – 90

Branch 10 – 96

Branch 9 – 105

Branch 8 – 159

Branch 7 – 177

Branch 6 – 312

Branch 5 – 352

Branch 4 – 413

Branch 3 – 435

Branch 2 – 615

Branch 1 – 797

Branch 0 – 1105

Gene Ontology enrichment analyses

Gene ontology (GO) enrichment analysis was performed using the clusterProfiler tool in R [83]. A list of all genes from the Standard dataset was used as an input into the ***enrichGO()*** function using a p-value cutoff of 0.05, a q-value cutoff of 0.10, calculated using the Benjamini-Hochberg procedure, and using the ***org.Hs.eg.db*** library from Bioconductor as the human genome annotation.

Genotype-Tissue Expression analysis

The data used for the analyses described in this manuscript were obtained from the Genotype-Tissue Expression (GTEx) consortium in 2018. Median transcripts per million (TPM) values of version 7 (2016-01-05) were downloaded and used to assess tissue expression [84]. A TPM threshold of 2 or greater was used as a cutoff to categorize a gene as "expressed." The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. A z-score was calculated for the number of RFSD genes expressed in each tissue.

Determining the location of each gene

A list of the genes in the Standard dataset was uploaded to BioMart and used to query their chromosomal location. The chromosome numbers were downloaded and the number of genes on each chromosome was normalized by the average number of genes on each chromosome according to the NIH U.S. National Library of Medicine [85]. A z-score was calculated for the number of RFSD genes on each chromosome.

Identifying nuclear-encoded genes which localize to the mitochondria

The RFSD genes in each dataset were sorted using MitoMiner v4.0 available at www.mitominer.mrc-mbu.cam.ac.uk [86] [87] [88] [89]. Based on published databases, genes were classified as mitochondrial if they were known or predicted to localize to the mitochondria and classified as non-mitochondrial if they were known or predicted not to localize to the mitochondria. A two sided Fischer's Exact Test was performed in R to determine whether the proportion of genes localizing to the mitochondria in the Standard dataset was significantly different from the genome average.

Calculating the proportions of each gene that are frameshifted

The proportion of each gene that was frameshifted was calculated by dividing the alignment length of each match by the length of the cDNA for each gene. A histogram was used to visualize the data produced. A z-score was calculated for the number of genes in each bin of the histogram.

Identifying shared domains between RFSD genes

The Ensembl sequence identifiers for each gene pair were used to search the Simple Modular Architecture Research Tool (SMART) database, available at http://smart.embl-heidelberg.de/. This tool produced a visual representation of each gene's domain architecture and order. The output was then examined and compared to the matching RFSD gene's output. The number of identifiable conserved domains in each gene were determined, as well as how many conserved domains were shared between the gene pairs. In order for a domain to be considered shared it had to be present in both genes in the gene pair in the same approximate location. Shared domains also had to be in the same order in each gene with approximately the same relative distance between them, unless the distance was increased by an identified frameshift mutation.

Determining whether RFSD genes are implicated in human disease

The Online Mendelian Inheritance in Man (OMIM) database available at https://omim.org/, was searched using each gene name, to determine if any of the RFSD genes are known to cause disease in humans. The genes which have a known mutant phenotype were collected and all the phenotypes associated with them were recorded.

The phenotypes were inputted into TagCrowd, a tag cloud generation tool available at https://tagcrowd.com/. The algorithm used to produce the tag cloud treats all words equally but permits omission of a subset of words. For this reason we chose to omit 16 words that are either extremely common, words associated with general disease or words that have a direct relationship to genetics. The omitted words are the following: abnormal, anomalies, autosomal, complex, congenital, deficiency, disease, disorder, dna, dominant, due, poor, recessive, susceptibility, syndrome, type.

We subsequently examined the OMIM phenotype data for diseases known to be associated with signaling defects. This was done by searching the phenotypes for 33 diseases that are known to result from inaccurate signaling. These diseases are the following: Age-related macular degeneration (AMD), Alzheimer’s disease, Asthma, Cirrhosis of the liver, Cushing’s syndrome, Diabetes, Diabetes insipidus (DI), Diabetic nephropathy, Diarrhea, Drug addiction, Ejaculatory dysfunction, Endotoxic shock, End-stage renal disease (ESRD), Epilepsy, Erectile dysfunction, Heart disease, Humoral hypercalcaemia of malignancy (HHM), Hypertension, Irritable bowel syndrome, Metabolic syndrome, Migraine, Multiple sclerosis, Nausea, Obesity, Osteoporosis, Pain, Hyperparathyroidism, Manic-depressive illness, Premature labour, Rheumatoid arthritis, Schizophrenia, Sudden infant death syndrome (SIDS) and Zollinger--Ellison syndrome.

Determining whether knockout data is available for RFSD gene mouse homologs

The KnockOut Mouse Project (KOMP) available at https://www.kompphenotype.org/, was queried using gene names for any homologs of the RFSD genes in our datasets. Specific examples have been highlighted to illustrate and support other sections of this study.

Results

RFSD events in the human genome

We generated two datasets of RFSD events using a customized tblastx function and filtering the results according to stringent criteria. For each dataset the paired genes must be different and must be in different frames. In addition, we required minimum a percentage identity of 80% and a maximum e-value of 1^e-10. In addition, the Standard dataset criteria required a minimum alignment length of 50 amino acids and the Conservative dataset criteria require a minimum alignment length of 100 amino acids. We created two datasets, one of which has been manually examined, in order to confirm that the results are genuine and comparisons between the two sets show consistent results. Notably the Conservative dataset requires a longer minimum alignment match for a gene pair to be included. Selecting for larger frameshifts meant the Conservative dataset is enriched for larger genes and by extension larger unframeshifted regions as well.

Almost all analysis of the Standard and Conservative datasets was consistent between datasets. Throughout this dissertation we will only mention the Standard dataset unless the results of the Conservative dataset differ from the Standard results. Results that differ will be specifically noted where appropriate. After filtering the results of the tblastx search we identified 315 RFSD events (630 genes) which fit the criteria for the Conservative Dataset and 628 RFSD events (1055 genes) which fit the criteria for the Standard Dataset (Table 1). These findings are consistent with the Okamura et al. paper that helped inspire this project. We can identify six types of frameshifts in the data as shown below. The insertion of 1, 2 or 3 nucleotides, or conversely the deletion of 2, 1 or 3 nucleotides respectively, into the original frame would result in three possible positive frames. In the case of duplications into the opposite strand, the deletion of 1, 2 or 3 nucleotides, or the addition of 2, 1 or 3 nucleotides, would result in three possible negative frames. Larger insertions and deletions can also shift the reading frame but the frame is always shifted by the remainder when dividing the size of the insertion or deletion by 3. We can identify +3 events when multiple frameshifts have occurred during the divergent evolution of the matched genes. For example one gene in the pair may have undergone a -2 frameshift and the other may have undergone a +1. This would lead to a different of +3 or -3 depending on which gene is the query. We found four types of events in the data: tandem duplications, regional duplications, interspersed duplications or overlapping duplications (Figure 2).

Table 1 Summary of identified RFSD events.

			Frameshifts					
	Events	**Genes**	**+1**	**+2**	**+3**	**-1**	**-2**	**-3**
Standard	628	1055	351	189	42	354	183	45
Conservative	315	630	96	49	11	86	53	20

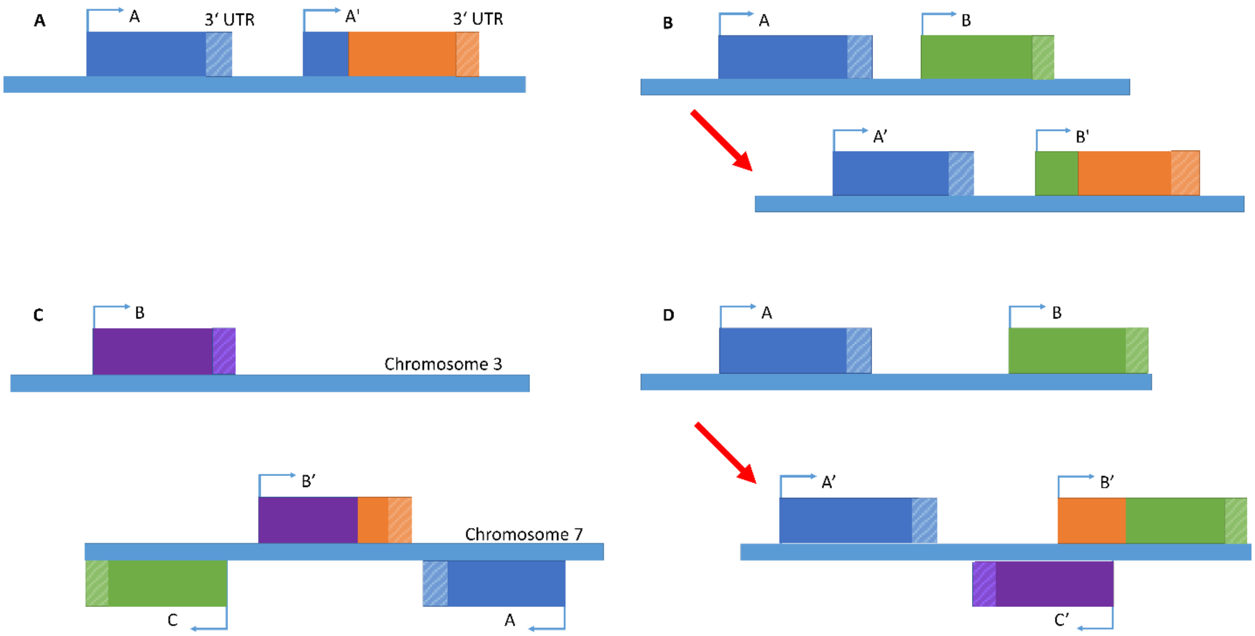


Fig. 2 Examples of RFSD events. (A) A tandem RFSD event. **(B)** A region RFSD event. **(C)** An interspersed RFSD event. **(D)** An overlapping RFSD event. Duplication of gene C into the region created a reading frame shift and a new start site.

Ages of human RFSD events

We then proceeded to establish the age of the events by identifying homologs of the genes involved in 22 other species (Figure 3A). The age of each gene was determined by the divergence time of the most divergent species with an ortholog. The age of each event was called as the age of the younger of the genes in each pair. We calculated the number of events per million years on each branch so we can compare the number of events between branches. (Figures 3B and 3C). The average rate was 0.405 events per million years for the Conservative dataset and 0.940 events per million years for the Standard dataset. Our method of identifying the ages of each branch produces an average age taken from multiple peer reviewed sources investigating the divergence time between each species and humans. This places the ages we have used in the center of the range of estimates available for this. This was the most effective way to select a specific divergence time, which we needed to conduct our analysis, but this estimate is influenced by the sources chosen and by the sources used when the analysis was done. Including or removing sources from the tool used to determine the age of each branch can have a significant impact on the mean age estimate, particularly if those sources include outlier estimates.

Our approach allowed us to assign an age to each event and whether RFSD events are tied to specific points in time or an occurrence that can be observed throughout our evolutionary history. The numbers of events at each branch we considered support the idea that RFSD events are a common occurrence (Figures 3B and 3C). One important note is that the age of the RFS event may not be the same age as the duplication event. A limitation of this method is that the RFS event and the duplication are assigned to the same branch even though we do not know when the RFS actually occurred. The data represent the youngest possible age of the duplication event by assigning to the age of the younger of the two genes involved and the oldest possible age of the

RFS mutation by assigning it to the age of the duplicated gene's appearance. One way to resolve this limitation is to reproduce this work in other species and determine if the homologous genes are frameshifted there as well. This would allow accurate dating of the frameshift as well as the duplication but is beyond the scope of this project. It would be interesting to determine whether the frameshifting observed in humans is a phenomenon specific to human evolutionary history or a common occurrence across several species. Based on the age distribution of RFSD genes across all branches examined, we would expect it to be ubiquitous in all included species. However, if it is a common occurrence it would also be important to determine whether the rate of frameshifting is constant across species and time or whether it varies. From the data collected we can determine that the rate of RFSD genes in humans varies across time and this may suggest that frameshifted new genes have been more likely to survive at specific points in our evolutionary history.

In both the Standard and Conservative datasets we observed a significant excess of events on branch 10 which represents the divergence time between humans and the clade including dogs and cows, approximately 96 million years (z-scores of 3.30 for Standard and 2.79 for Conservative) (Figures 3B and 3C). This divergence time represents some of the earliest mammalian lineages and these genes may have arisen at the time of the mammalian radiation. We can also note that younger genes tend to be smaller and so it is possible there is a slight bias towards undercounting young genes. However, we do not feel this is a significant concern because the Standard and Conservative datasets show similar distributions, meaning when the size barrier to being included in the data is lowered we do not see an increase in the numbers of younger genes identified. If we were significantly undercounting the youngest genes we would expect to see a commensurate increase in numbers at the most recent branches when we include smaller genes.

The mammalian radiation is estimated to have occurred approximately 75 – 100 million years ago and led to the extant diversity in eutherians [90] [91] [92]. Although older paleontological techniques led to the original hypothesis that the rapid expansion in mammals occurred after the Cretaceous/Tertiary (K/T) extinction event about 65 million years ago, molecular techniques have shown that the initial diversification occurred much earlier [93] [94] [91] [95]. The most supported current hypotheses suggest the adaptive radiation that occurred was more closely tied to the continental breakups that occurred in the Mesozoic [91] [96]. These land fragmentation events would have produced three factors which may have encouraged the rapid diversification.

Firstly, the geographical division of an early mammalian population would have allowed multiple subpopulations to diverge according to classic allopatric speciation models. These divisions were caused by shifting tectonic plates breaking up landmasses and high sea levels flooding and isolating areas [96]. Secondly, the rapid changes in habitat that occurred around this time would have exerted a strong selective pressure against the maintenance of the existing phenotypes and in favor of optimizing for the species' new environment. Finally, the changes in the planet during this period resulted in several mass extinction events [97] [98]. Although these never reached the planetary scale like the K/T event, these would have vacated several niches and functional roles which could be filled by early mammals.

These factors would have promoted the rapid diversification of what was previously a few relatively homogenous species and RFSD events that occurred at the time might have been maintained as part of that process. This could have been accomplished by providing the raw genetic material which could be co-opted to accomplish some of these phenotypic changes. On the other hand genes produced by RFSD could also have been near enough another beneficial locus to be swept to a high enough frequency to be resistant to elimination. RFSD genes that were swept to higher frequencies would have had more time to develop a useful function to the cell or species and thus be maintained long term. Without evidence supporting one possibility over the other it is impossible to say which scenario had a greater impact at that time but the true events are almost definitely some combination of both.

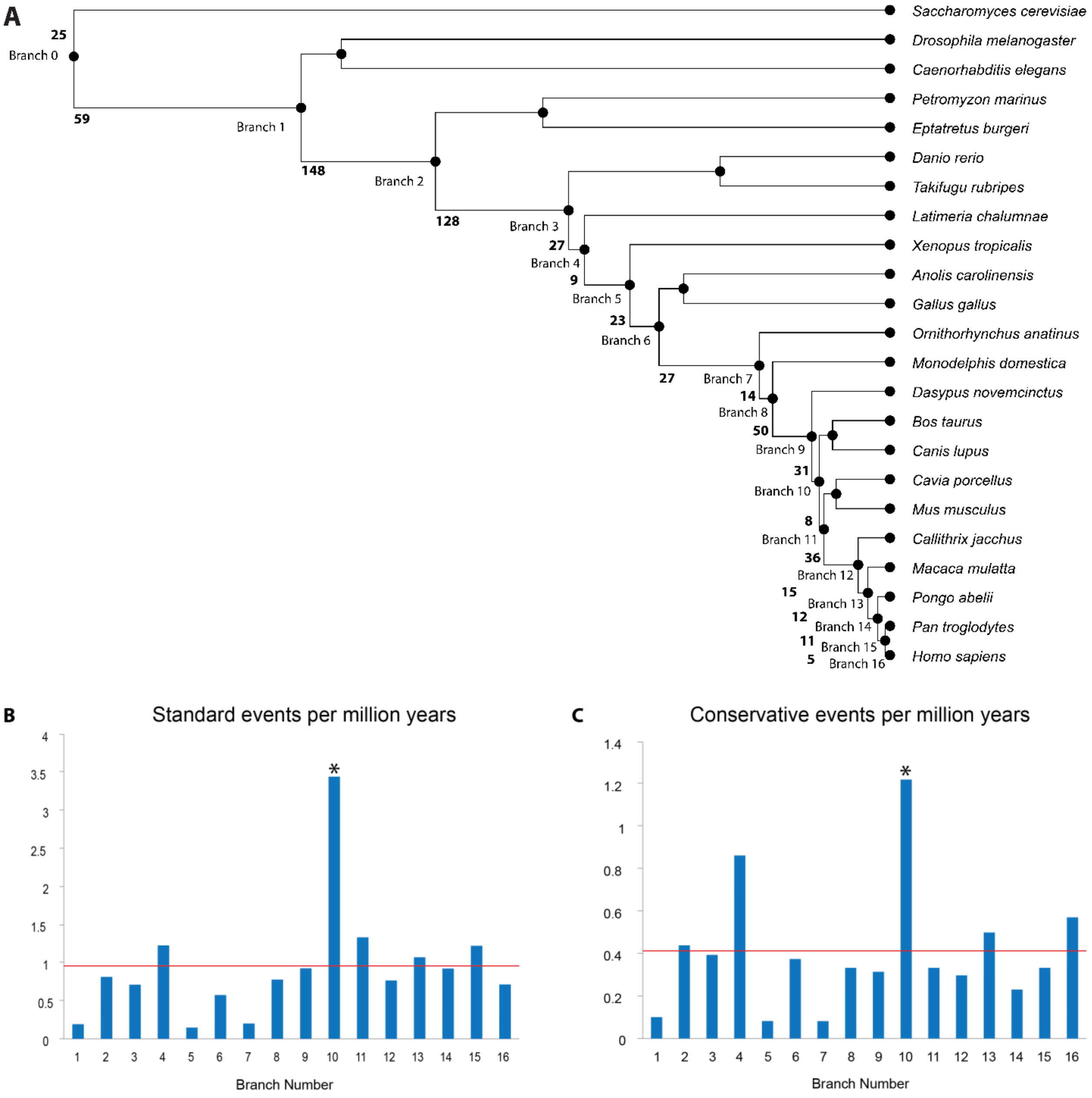


Fig. 3 Ages of identified RFSD events. (A) Phylogenetic tree indicating the species used to determine the ages of the RFSD events. **(B)** The number of identified events in the Standard dataset normalized by the length of each branch. The numbers on the x axis represent the branches on the tree. The red line represents the mean number of events. The asterisk represents a significant bar. **(C)** The number of identified events in the Conservative dataset normalized by the length of each branch. The numbers on the x axis represent the branches on the tree. The red line represents the mean number of events. The asterisk represents a significant bar.

Proportions of human RFSD genes that are affected by RFS mutations

To investigate how impactful the RFS mutations would be on a RFSD gene, we determined the proportion of each gene that is frameshifted. The mean proportion is 0.45 and the standard deviation is 0.28. The data suggest that the impact of the RFS mutations is significant as, on average, almost half of each RFSD gene is frameshifted. This does not guarantee that any specific RFS is significant but the effect of an RFS on the portion of a gene it occurs in is extreme. By using the proportions of the RFSD genes that are frameshifted as a proxy for the effect of the RFS we can gauge the significance of the mutations.

A histogram of the data reveals there are two significant columns (Figure 4A). The 0 - 0.05 column and the 0.1 - 0.15 column have z-scores of -2.09 and 2.16 respectively. The fact that the first column is underrepresented is expected as our filtering criteria in generating our datasets eliminate any small frameshifts detected. This has a depressive effect on the probability of identifying RFSD genes with only a small portion of the gene that is frameshifted. However, the finding that 10% - 15% of a RFSD gene being frameshifted is significantly overrepresented is a surprising and important finding. It is possible that frameshifting this proportion of a gene allows a RFS mutation the greatest chance of survival, at least within the context of human evolution. Nevertheless, a more likely interpretation is that this proportion of a gene is the most likely proportion to be sufficiently retained, after frameshifting, to be detected by our method. This explanation should not be dismissed as a curiosity intrinsic to the detection process. By determining that this range of proportions is the most likely to be maintained and recognizable, we are convinced that 10% - 15% of a gene is significantly more likely to be the most commonly useful proportion of an inherited gene and the most likely proportion to be adapted as part of a frameshift.

For example, the gene TRIO is a GDP to GTP exchange factor, which is involved in cell migration and growth by promoting the reorganization of the actin cytoskeleton [99]. It accomplishes this by being a part of the GPCR and NGF signaling pathways [99]. Homologs can be found in species as distant as lampreys and hagfish. The TRIO gene was duplicated in a common ancestor of humans and bony fish and then frameshifted to form the gene KALRN. KALRN has lost the last 5 domains that can be identified in TRIO and is a shorter gene (Figure 28 in Chapter 4). However, KALRN has an additional Spectrin domain that TRIO does not have. We are unable to determine whether TRIO also lost this domain or whether KALRN evolved this independently without examining an outgroup species. As a result we are unable to definitively state whether this is an example of neofunctionalization or subfunctionalization. Spectrin repeats are commonly found in proteins involved in cytoskeletal structure [100]. This is supported by the known function of KALRN which is a GDP to GTP exchange factor like TRIO [101]. KALRN is also involved in the GPCR signaling pathway but has not been identified in the NGF pathway [102]. Instead it has been found in the RET signaling pathway [103]. In this case, just over 11% of TRIO matches KALRN in a +1 frameshift. Given the significant domain loss it is unlikely that this was the proportion of the novel TRIO duplicate that was initially frameshifted. As a result this case would seem to support the probability that 11% of TRIO was repurposed for KALRN after a frameshift. Nevertheless, we cannot rule out the first option because of the possibility that the duplication was not a complete duplication.

A scatterplot of the frameshifted proportions of the RFSD genes assigned to their respective branches illustrates two main points (Figure 4B). Firstly, the range of proportions is relatively evenly dispersed across all branches indicating that the proportion of a gene that is frameshifted is not dramatically affected by the time period in which the RFS mutation occurred. Secondly, the plot clearly indicates visually that the process of RFSD gene origination is ongoing and can be found at all points in human evolutionary history.

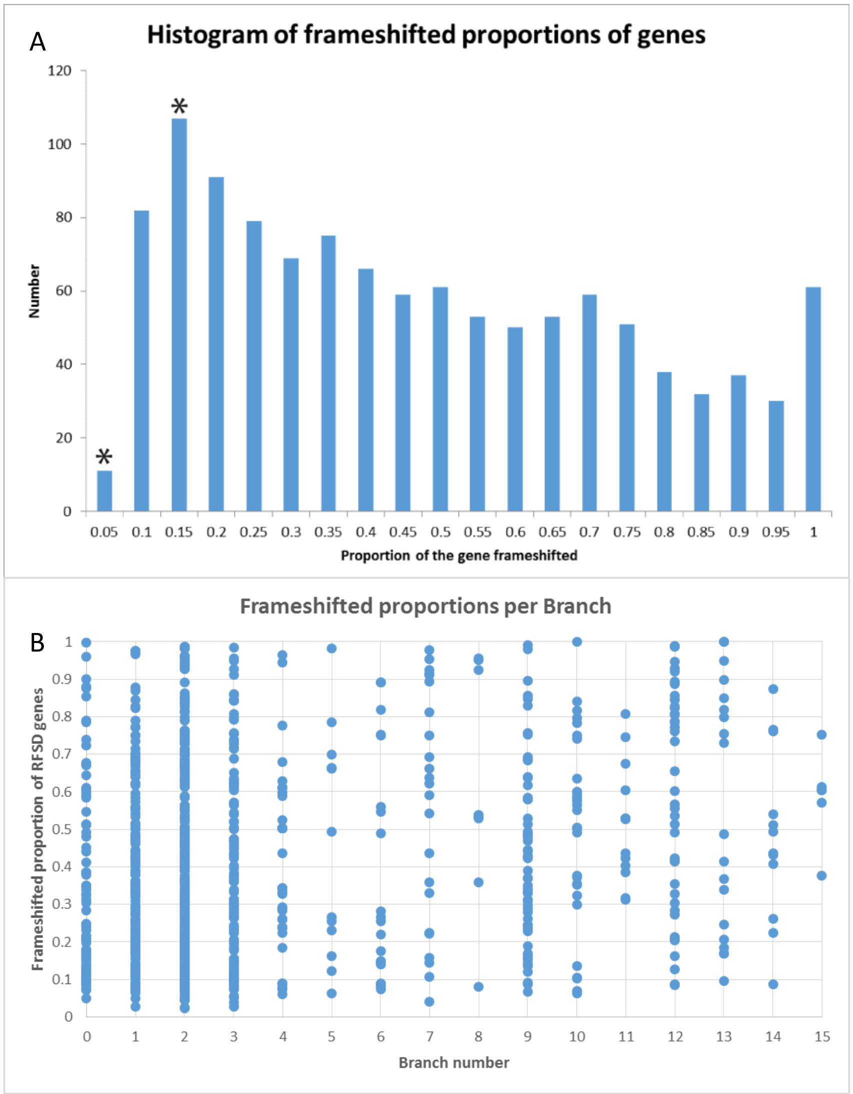


Fig. 4 Plots of frameshifted proportions of RFSD genes.
(A) Histogram of the frameshifted proportions of human RFSD genes The Mean proportion is 0.45 and the Standard Deviation is 0.28. The Mean frequency is 58.2 and the Standard Deviation is 22.6. The two asterisks represent the two significant columns. The 0 - 0.05 column has a z-score of -2.09 and the 0.1 - 0.15 column has a z-score of 2.16. **(B)** Scatterplot of the RFSD frameshifted proportions classified by species tree branch number

Inherited function between RFSD gene pairs

In order to ascertain whether the identified relationship between RFSD gene pairs is functionally meaningful, we examined each gene for known conserved functional domains. The relationship between the RFSD gene pair functions was inferred by determining which gene pairs shared conserved domains and which ones had different conserved domains. For each RFSD gene pair, shared domains had to be present in both genes in the same approximate locations, in the same order with a similar relative distance between the domains, unless a frameshift mutation disrupted the distance between domains.

This method allowed us to determine the true number of shared domains between each gene pair while enabling us to detect instances where similar domains had appeared independently or where an older domain was supplanted by a new one. RFS mutations can have an extreme effect on a peptide and change the previously conserved protein sequence dramatically. As a result, we identified 290 paired RFSD genes had lost domains or had evolved very different domains at the same relative position.

The relationship between parent and offspring genes connected by an RFSD event is clearly meaningful as approximately 60% of RFSD genes for which domain data is available share a conserved domain. This indicates that the RFSD gene pair association is consequential as the offspring gene has inherited and maintained a functional portion of the gene. In addition, 126 RFSD genes have evolved a novel conserved domain that is not shared with their parent genes. This suggests that the frameshifted portion of the gene has evolved to serve a novel purpose, reinforcing the importance of the relationship between the parent and offspring genes.

This relationship can be clearly seen in the case of the RFSD gene pair PDZD8 and SLC18A2. PDZD8 is an ancient gene (homologs can be found in *D. melanogaster* and *C. elegans*) that tethers the mitochondrial and endoplasmic reticulum membranes and is essential for Calcium ion transfer [104]. PDZD8 was duplicated in a common ancestor of humans and lampreys and was subsequently frameshifted to form SLC18A2, an essential ATP-dependent vesicular transport molecule, which transfers neurotransmitters across the cell membrane in human neurons [105]. PDZD8 and SLC18A2 have greatly diverged in their functions but they share a transmembrane domain (Figure 29 in Chapter 4). While PDZD8 is specialized for tethering two organelles, SLC18A2 has accumulated transmembrane domains and become an efficient and essential cell membrane protein [104] [105]. The transmembrane domain and novel peptide inherited together by a nascent SCL18A2 may have allowed the cell to effectively shortcut the evolution of a new transmembrane function.

The availability of a novel peptide attached to a functional one is a unique mechanism by which large adaptive steps can occur. This allows the cell and/or organism to co-opt a functional peptide to adapt to new circumstances or optimize for the current selective pressures it is experiencing. Even a marginally functional peptide can be a significant shortcut to escaping from adaptive constraint or neofunctionalizing. When combined with the inferences that can be made regarding the proportions of the RFSD genes that are frameshifted, we can infer that the genes generated by RFSD events inherit a significant proportion of their parent genes' sequence and domains. This strongly suggests that the relationships identified by this study are both meaningful and impactful.

Another example of such a relationship is the LPA and PLG gene pair. The LPA gene encodes another branch 1 protein that has protease activity and is responsible for inhibiting a plasminogen activator [106]. It is part of the lipoprotein metabolism and integrin signaling networks [106]. The branch 2 PLG gene encodes plasminogen and when activated is cleaved into plasmin, a protease which cleaves fibrin in blood clots, and angiostatin, an inhibitor of angiogenesis [107]. PLG is also known to participate in syndecan-4-mediated signaling [108]. Whereas LPA comprises of 16 Kringle domains and a protease domain, PLG has retained the protease domain and 5 Kringle domains but then evolved an APPLE-like binding domain (Figure 30 in Chapter 4). The frameshifted portion of the gene allowed the nascent plasminogen protein to evolve a secondary function and integrate into the same pathway as its parent gene, something likely aided by inheriting regulatory elements which controlled the nascent protein's expression pattern.

Table 2 Summary of identified functional domains for RFSD gene pairs in the Standard dataset.

Functional domain data	**Number of genes**
Domain data available	1042
At least 1 shared conserved domain	621
At least 1 different conserved domain	290
At least 1 shared domain and 1 different domain	126

Characteristics of human RFSD events

To determine whether genes involved in RFSD events are prone to having specific characteristics, we performed GO analyses. We observed that genes involved in RFSD events are enriched for molecular functions related to signaling activity or transcriptional activation (Figure 5A). These are functions where high molecular activity is often required. It is possible that a mutation leading to an increase in the activity of gene products with such functions could provide a benefit to the cell and the organism, even if the duplicated product is imperfect due to a frameshift. This could ameliorate the negative selective pressure on a frameshifted gene or even provide a positive pressure.

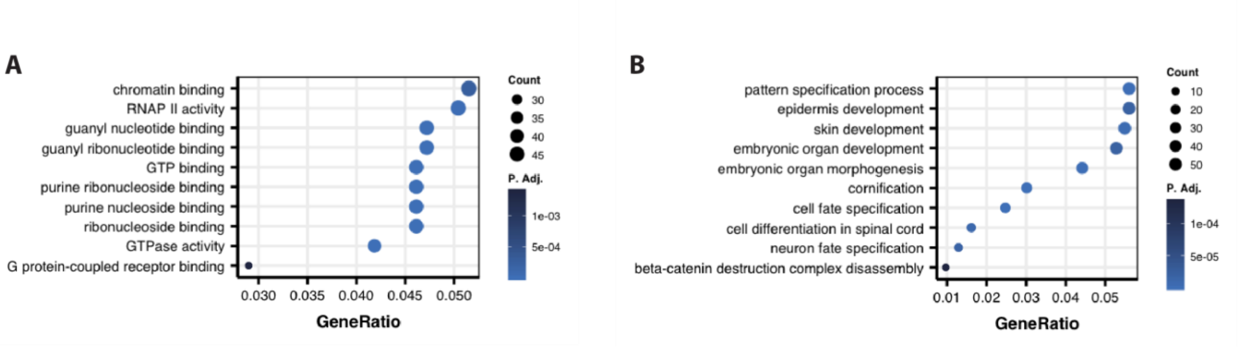
In particular, signaling peptides often have multiple discrete functions, including recognition of a target or ligand, localization to a part of the cell or having a specific number of transmembrane domains [109] [110] [111]. The duplication of a signaling gene followed by a frameshift mutation can generate proteins that can still perform one part of these functions while being tethered to a novel peptide that is expressed and free to adapt to the cell's needs. This can rapidly diversify a pathway's signaling targets and receptors to allow that pathway to take on a greater role in regulating the cell or the species. This in turn can lead to an increase in complexity for the species and can greatly shorten the amount of time which it takes for a species to react to selective pressures.

A good example of this is the abrupt appearance of many members of the nicotinic acetylcholine receptor family in a common ancestor of humans and lampreys. Several members of the CHRN gene family were duplicated and frameshifted in order to produce a diverse set of receptors. This family encodes for proteins that primarily transport neurotransmitters across synapses but have been a target of a number of drugs, as well as nicotine, and are important to human health [112]. Mutations in this family of receptors often leads to epilepsy and a variety of other neurological disorders. RFSD gene pairs that arose at that time include CHRNA2 and CHRNA4, CHRNB2 and CHRNB4, and CHRNA7 and CHRFAM7A, a fusion gene between CHRNA7 and FAM7A with a critical role in inflammation, immunity, neurodegeneration, and cognitive function [113]. Interestingly, frameshifts in the extant CHRFAM7A have been associated with epilepsy and other disease states [114].

We also observed that these genes show an enrichment for involvement in developmental and patterning processes (Figure 5B). It is possible that organisms have co-opted the genetic novelty provided by these frameshifts to adapt and evolve on a morphological level. This may also be related to the enrichment observed in signaling functions as an increase in signaling complexity is often tied to an increase in developmental complexity. Novel developmental processes can involve highly specific signaling and the excess of developmental GO terms would be observed if a RFSD event is the most efficient way to achieve that. This finding may also tie in to the excess of RFSD events found to have occurred around the time of the increase in mammalian biodiversity. The adaptive radiation that happened would have involved many novel morphological changes and the RFSD events connected to this could have enabled those changes.

We then performed a GTEx analysis to examine the expression characteristics of these genes and observed that approximately a third of the genes show expression in each of 52 tissues examined while about ten percent show expression in none of the tissues. Over half the genes examined show non-constitutive tissue-specific expression. We also observed a significant enrichment for testes expression in these genes (z-score of 3.106 for the Standard dataset, Figure 5C). This could be a result of one or more of three possibilities. Firstly, the testes are known to have extremely open chromatin, which is conducive to the expression of a large percentage of the genome, not all of which is functional in the testes. Secondly, the enrichment could be the result of the dataset containing significant numbers of younger genes, which commonly show testes expression [115]. This, however, is unlikely to be the primary factor causing tis result. Finally, the RFSD genes show some evidence of being enrichment in the sex chromosomes. Although this will be discussed later on it is important to note that sex chromosome specific expression and function could result in an enrichment in testes expression.

We also found a significant underrepresentation of three very different and unrelated tissue types. Muscle – Skeletal had a z-score of -2.162, Heart – Left Ventricle had a z-score of -2.344 and Whole Blood had a z-score of -2.912. There are two likely explanations for this reduction. Some tissues could be slightly less amenable to a frameshift's survival than others leading to fewer RFSD genes being expressed in those tissues. The other possibility is that some tissues have lower expression of genes in general and so the underrepresentation is not specific to RFSD genes e.g. whole blood.



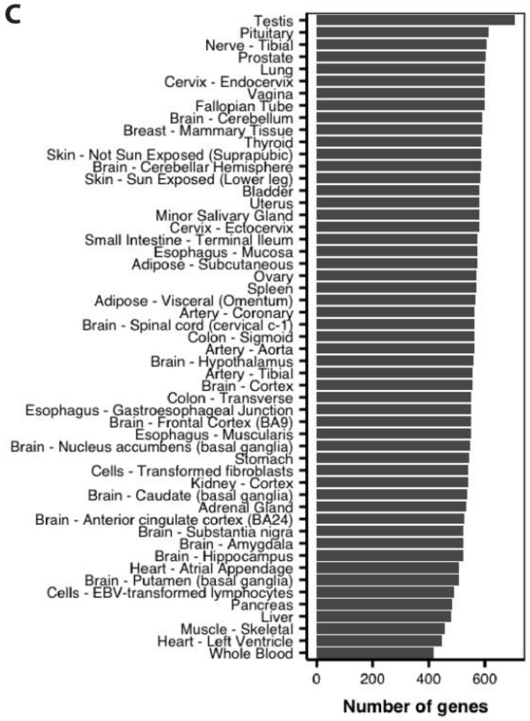


Fig. 5 Characteristics of human RFSD genes. **(A)** Plot of GO enrichment analysis data indicating the most common molecular functions of RFSD genes. **(B)** Plot of GO enrichment analysis data indicating the most common biological processes RFSD genes are involved in. **(C)** Bar chart of GTEx data showing the number of RFSD genes expressed in 52 examined tissues. Over 400 of these genes are expressed ubiquitously. RFSD genes are enriched for testis expression (z-score 3.106). They are also underrepresented in Muscle – Skeletal, Heart – Left Ventricle and Whole Blood (Respective z-scores of -2.162, -2.344 and -2.912).

Excess novel genes generated by RFSD on the X and Y chromosomes in humans

When we examined the locations of these genes we found a significant enrichment on the sex chromosomes, X and Y, when normalized by gene density (75 genes in total, Table 8 in Chapter 4). The sex chromosomes have a z-score of 2.36 for the Standard dataset. The highest excess was observed on the Y chromosome with 6 RFSD genes (z-score of 3.13 for the Standard dataset, Figure 6). As mentioned previously, this excess on the sex chromosomes may be partially responsible for the excess of RFSD gene expression detected in the testes. Duplications have been shown to be proportional to the size of the chromosomes they are on so we considered normalizing RFSD events by chromosome length as well [116]. However, when normalized by chromosome length the most gene dense autosomes show significant enrichment suggesting this is an artefact.

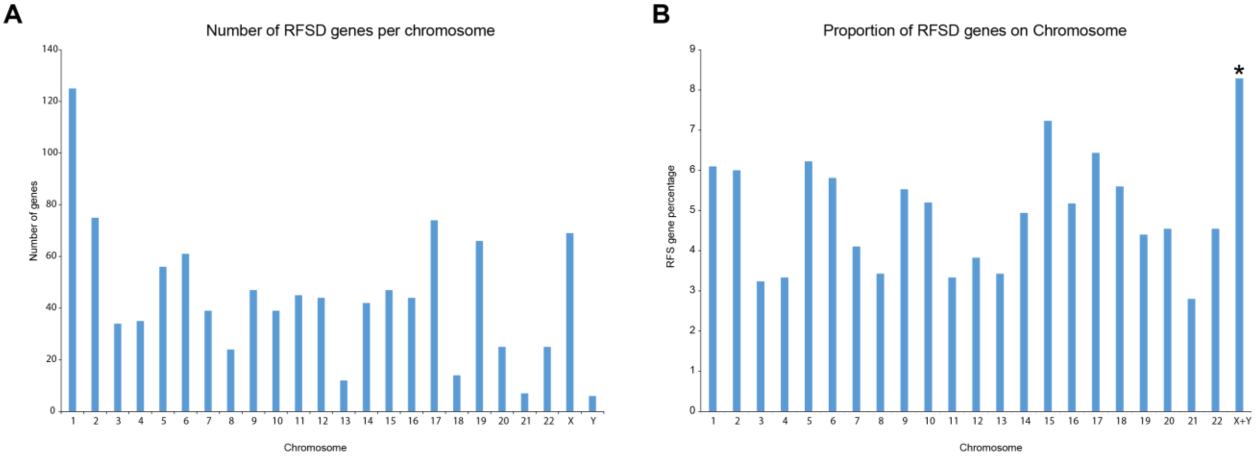


Fig. 6 RFSD genes on the human sex chromosomes. (A) Bar chart of the number of RFSD genes on each chromosome. **(B)** Bar chart of the proportion of RFSD genes on each chromosome pair normalized by gene density. The asterisk represents a significant bar (z-score of 2.36).

The human X chromosome is often sorted into various strata or clusters which demarcate segments of different age [117] [118] [119]. We used the 12 cluster method of Pandey et al. to place the Standard dataset RFSD X chromosome genes into their respective clusters as this method provides the most defined boundaries and the greatest resolution (Figure 7 and Table 8 in chapter 4) [119]. This method denotes cluster 1 as the oldest and cluster 12 as the youngest with clusters 10, 11 and 12 making up PAR1 on the X chromosome. Interestingly none of the RFSD genes are located in those clusters. In order to understand the context in which these clusters arose, we can identify which species branches are most similar in age to the X chromosome's subdivisions. The strata traditionally used to divide the X chromosome are overlaid on the current 12 cluster system which does not use an age based method of determination [119]. The youngest strata are 5, which appeared 29-32 Mya and 4, which appeared 38-44 Mya [120]. These would correspond to Branches 13 and 12 respectively. Stratum 3 appeared approximately 50 Mya, dating it to Branch 11 [117]. The oldest strata, 2 and 1, date to the proto-sex chromosomes that predate the mammalian X and Y [117] [120]. This most likely dates them to approximately 100-150 Mya and 100 – 350 Mya respectively around the time marsupials and eutherians were starting to diverge (Branches 8 and 6/7) [117] [120]. Overlaying these ages and relating them to the evolutionary history that was occurring at the time can give us a clearer idea of when exactly these dramatic rearrangements occurred and what the evolution of the sex chromosomes entailed.

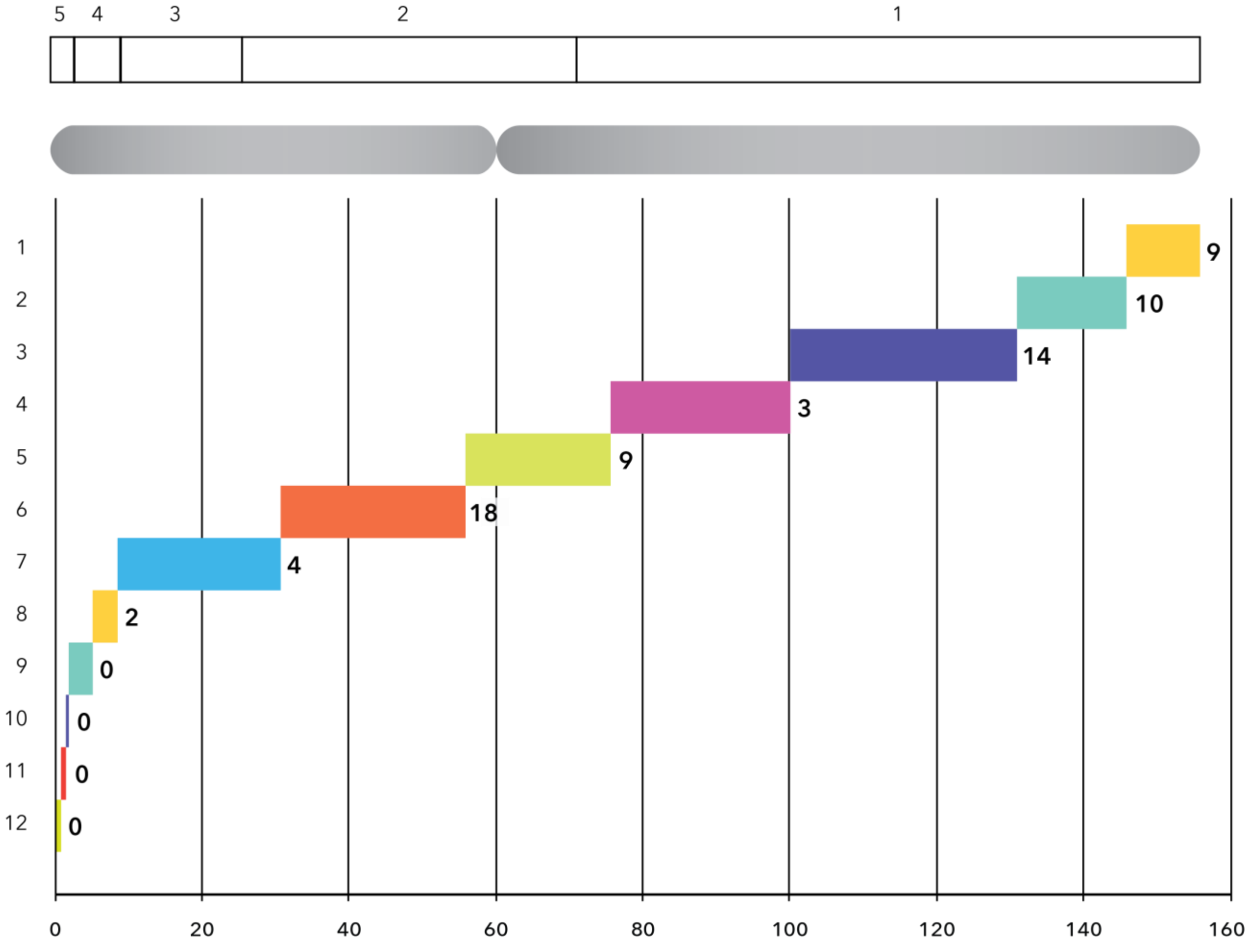


Fig. 7 RFSD genes in the human X chromosome clusters. The chromosome is divided into 12 clusters of varying age (millions of years ago) and the number of RFSD genes in each one is indicated. The 5 strata used previously are shown at the top of the figure.

Shared characteristics between RFSD pairs

We compared the characteristics of the pairs in each RFSD event in order to determine whether they shared their molecular functions, biological processes or expression patterns. We expected that these gene pairs would be least likely to share molecular functions due to the frameshift mutations, which would potentially change the protein products significantly. However, we did expect them to share expression patterns and perhaps biological processes as well. We observed that the Standard dataset pairs are unlikely to share a molecular function (Figure 8A). It is important to note that the proportion of Standard dataset pairs that share a molecular function is slightly lower than the proportion of pairs that share domains and haven't evolved an additional conserved domain. We identified 495 gene pairs that share domains out of the 1164 identified RFSD gene pairs. This discrepancy can be addressed by taking into account that numerous domains exist in many types of proteins and so sharing domains is indicative but not proof of shared function, as it is of inherited DNA. Thus the probability of a gene pair sharing a molecular function is shown to be smaller than the proportion of pairs that share domains. However, the chances of RFSD pairs sharing a function if we only look at the data in the Conservative dataset is far greater (Figure 18 in Chapter 4). This is likely because the genes in the Conservative dataset are enriched for larger unframeshifted portions as a result of being selected for having larger frameshifted portions. This increases the probability of the RFSD gene pairs sharing domains, and by extension, functions.

We also observed that these gene pairs commonly share a biological process (Figure 8B), which was a consistent finding across both datasets. This could be explained by our third finding which was that the gene pairs commonly share an expression pattern (Figure 8C). The figure below indicates that the tissues in which parent and offspring genes are expressed are positively correlated. The correlation means that parent and offspring genes that are connected via a RFSD event are likely to have similar expression, suggesting that the expression pattern was inherited. This is expected because regulatory signals are unlikely to be affected by a frameshift mutation and can be maintained by the offspring gene. If the parent and offspring genes are expressed in the same tissues or at the same times they are also more likely than random to be involved in similar biological processes which is borne out in our data.

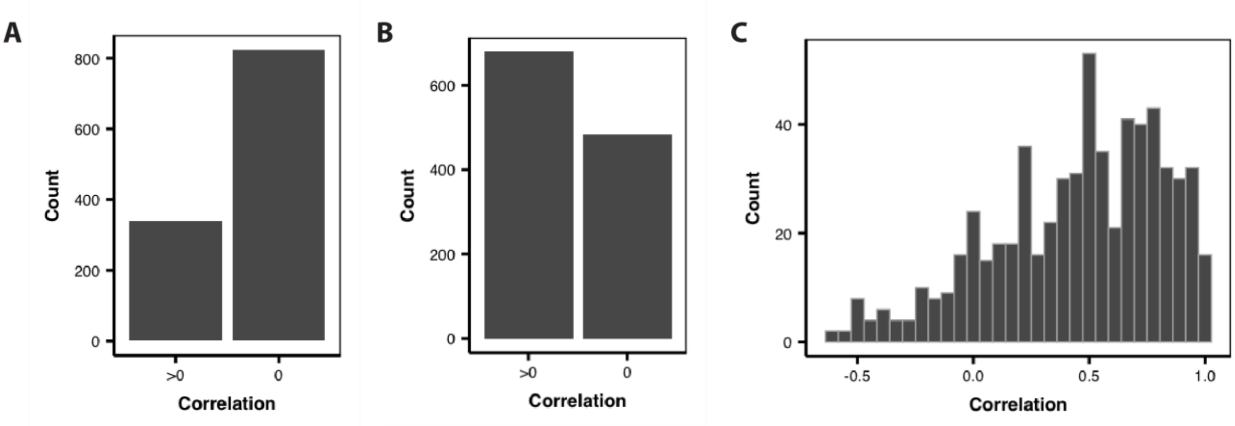


Fig. 8 Shared characteristics between parent and offspring RFSD genes. (A) Bar chart showing the number of gene pairs that share a molecular function**. (B)** Bar chart showing the number of gene pairs that share a biological process**. (C)** Histogram showing the degree of similarity in expression pattern between RFSD pairs.

RFSD genes localize to the mitochondria

We searched a database of proteins that localize to the mitochondria for the genes in the Standard dataset. We discovered 73 genes that were either known or predicted to localize to the mitochondria (Table 9 in Chapter 4). This comprises 6.9% of our dataset which was higher than estimates of the genome average of 5% (Figure 9) [86] [121] [122]. When we performed a two-tailed Fischer’s Exact test we found that the difference was highly significant, p = 0.0046. This suggests that novel RFSD human genes can be co-opted into the mitochondrial proteome.

Common mitochondrial functions associated with RFSD genes in our datasets include oxygen binding, solute transfer and ATP synthesis.

There is some overlap between RFSD generated proteins that localize and function in the mitochondria and those that are involved in signaling pathways. PDZD8 tethers the mitochondria to the endoplasmic reticulum and regulates calcium ion uptake. This plays a critical role in the regulation of the intra- and extracellular signaling dynamics in neuronal dendrites [104]. PDZD8 is an essential gene that has multiple diverse roles, a trait which is not uncommon in the RFSD genes we have identified.

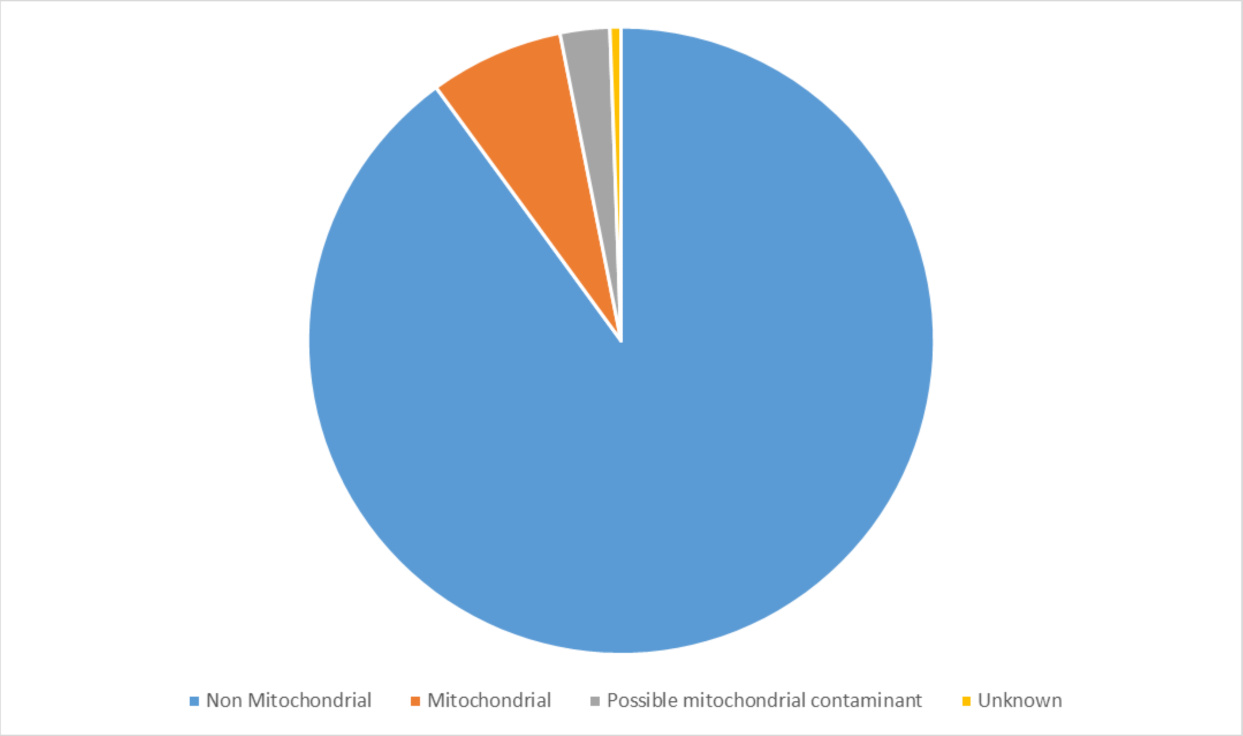


Fig. 9 RFSD genes localize to the human mitochondria. Pie chart indicating the percentage of the Standard dataset found to localize to the mitochondria. 6.9% of the dataset is known or predicted to localize to the mitochondria.

RFSD genes are found to have direct effects on human health

An outstanding question is the phenotype of these RFSD genes and what meaningful role they play. To determine this we queried the effects of RFSD genes on human health by searching the OMIM database for any information on these genes. We identified 248 genes with at least one associated disease phenotype and 72 with two or more. This suggests that RFSD genes can play a significant role in human health as 24% of RFSD genes are shown to present a disease phenotype when mutated. This method allowed us to determine the types of phenotypes caused by mutations to the RFSD genes including, but not limited to, a frameshift mutation.

There are three significant points that must be noted here. Firstly, the RFSD genes identified in our study had all been duplicated before they underwent a frameshift, which would likely alleviate any phenotype displayed by a mutant gene. This limits the inferences we can make about the effect of any particular frameshift mutation involved in a RFSD event, without preventing us from predicting what a RFS or other mutation would do to the extant genes. Secondly, OMIM is an incomplete database and it is likely that several of the genes in our dataset are implicated in human disease but their phenotypes have not been attributed to them and recorded yet.

Finally, a class of genes which will be overwhelmingly absent from the OMIM database are essential genes. The mutant phenotypes of the RFSD genes which are essential are usually not included if the phenotypes are a variation of lethality during development. These phenotypes are hard to identify if the mutations result in a miscarriage. The only essential genes commonly included in OMIM are the ones that show infertility phenotypes when mutated. These alleles also prevent transmission to the next generation but can be phenotyped and recorded with much greater ease. As a result, it is likely that our reported proportion of RFSD genes that are implicated in human disease is an underestimate and that there are more genes in our datasets which have a significant impact on human health.

Table 3 Summary of OMIM database search for RFSD genes in the Standard dataset.

OMIM entries found	Number of genes
no disease data available	807
one associated disease	176
two or more associated diseases	72

To search the data collected from OMIM for patterns we generated a tag cloud using the words in the names of the diseases associated with the RFSD genes. This allowed us to identify trends in the phenotypes recorded and ascertain the likeliest effects of mutant RFSD genes. The algorithm used to produce the tag cloud treats all words equally but permits omission of a subset of words. For this reason we chose to omit 16 words that are either extremely common, words associated with general disease or words that have a direct relationship to genetics. These words don’t give us any additional information on the RFSD genes other than that their phenotypes are known to be genetic disorders, which is an assumption we can make to begin with. The full list of omitted words can be found in the methods section.

The tag cloud could also be used to validate our previous inferences by checking if the most common end phenotypes matched our previous inferences. Interestingly, the tag cloud shows almost all previously identified patterns in the data, increasing our confidence in these results. The descriptions of the phenotypes caused by mutations to RFSD genes include mitochondrial, supporting the excess of RFSD genes found in the mitochondrial genome, and x-linkage, supporting the excess of RFSD genes found on the sex chromosomes. In addition, the tag cloud highlights several words relating to neurological disorders and developmental defects, supporting the data produced by our GO and GTEx analyses.

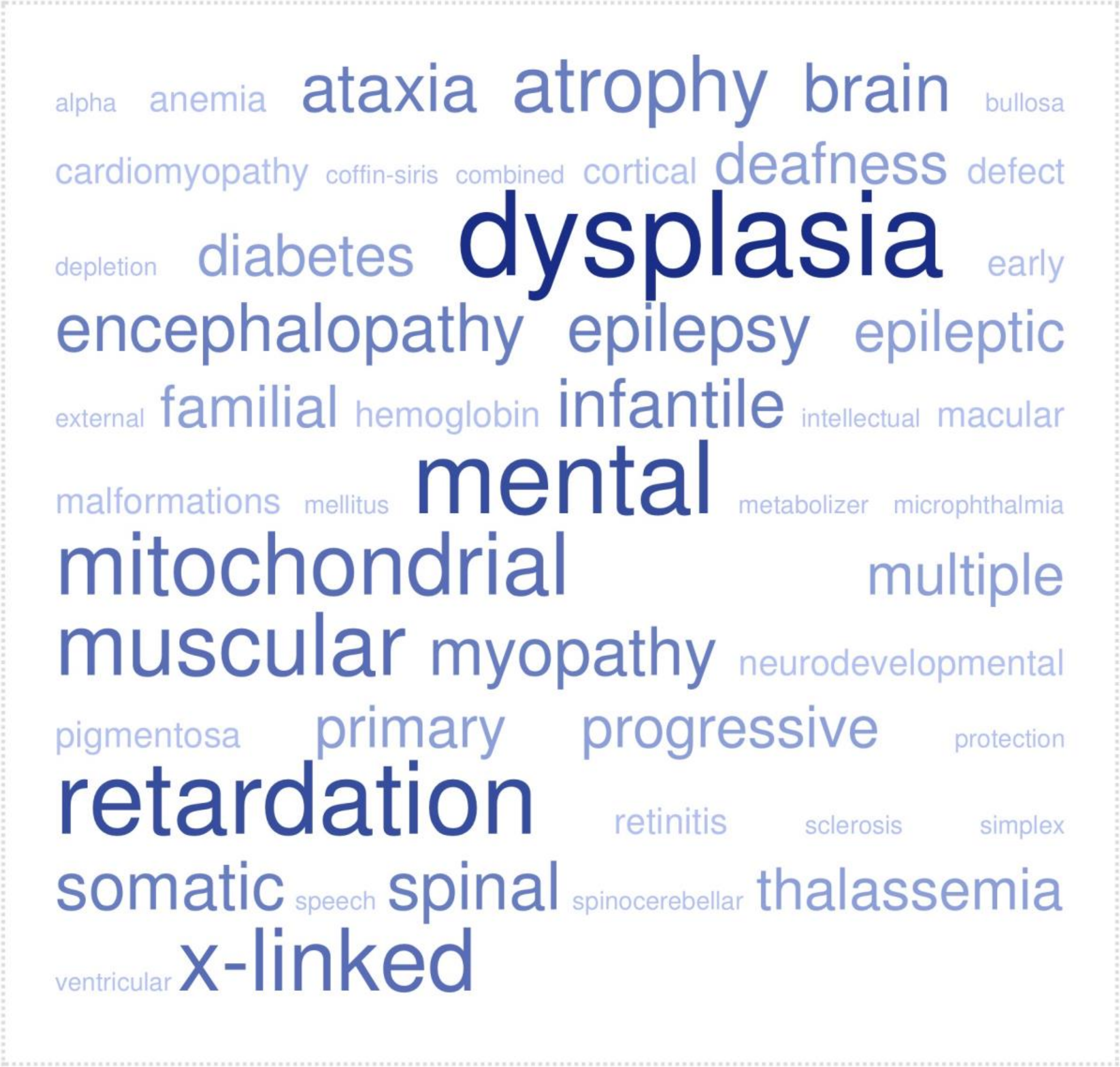


Fig. 10 Tag Cloud of disease phenotype terms associated with RFSD genes. The size of each word corresponds to the frequency with which it appeared in the RFSD phenotypes found in OMIM. Common words or words directly related to genetics have been omitted.

Finally, we determined that the several RFSD genes, when mutated, cause diseases known to be caused by signaling defects. This was done by examining the OMIM phenotype data for the

33 diseases most associated with signaling defects [123]. We determined that the Standard RFSD dataset contains genes that relate to 22 signaling disease phenotypes, a finding that supports the GO analysis inferences made previously.

It is noteworthy that RFSD genes that are associated with these disease states include members of classic cell signaling pathways such as WNT2B, a member of the WNT signaling family, CD209 and CLEC4M, transmembrane receptors that can initiate an intracellular signal cascade as part of the C-type lectin receptor signaling pathway, and PDGFRA and –B, cell surface receptors and kinases that can activate the RET, BCR or MAP kinase signaling pathways [124] [125] [126] [127] [128]. The mechanism of RFSD could have allowed these genes to arise and diversify relatively quickly but the drawback of increased complexity is the increased probability of a critical failure in a longer and more intricate pathway. This theory of conflict inherent to genomic and phenotypic changes has also been proposed by Gunter Wagner [129]. The theory holds that every advantageous modification to the genome is accompanied by the negative effects of disrupting the status quo [129]. Nevertheless, if the sum of the systemic changes results in at least a slightly positive fitness difference, the mutations can be fixed in a natural population.

We should also note that, although these genes have been associated with specific disease phenotypes and those disease phenotypes have been associated with signaling defects, this is not direct evidence that the disease states associated with these RFSD are the result of a signaling failure. Some of the disease included above have multiple causes and several of the genes have multiple functions so there is a small possibility of a spurious result when only considering signaling as a cause. We don’t believe this is likely and do not think it is a significant drawback as, regardless of the specific cause of the disease state in each individual gene study, the phenotypes recorded have been directly linked to the genes in question.

Table 4 Summary of signaling disease phenotypes associated with the RFSD genes in the Standard dataset.

Signaling related disease	Number of disease phenotypes found
Age-related macular degeneration (AMD)	3
Diabetes insipidus (DI)	1
Diarrhea	1
Epilepsy	7
Heart disease	1
Hypertension	2
Metabolic syndrome	1
Multiple sclerosis	1
Obesity	3
Hyperparathyroidism	1
Rheumatoid arthritis	1

Knockout mouse studies of RFSD gene homologs

To supplement our human phenotypic data, we searched for knockout data from RFSD gene homologs in mice so we could gain some additional insights into the functional consequences of the genes in our datasets. We searched the KnockOut Mouse Project (KOMP) database for the genes in our study and determined that 18 of the RFSD genes' homologs had been knocked out and the mutants phenotyped in mice. These 18 genes are PLXNA2, POMGNT1, KMO, SLAMF8, ITLN1, HMGN2, STRN, LYPD1, ATP13A5, MED28, SLC22A4, CUZD1, CALCB, C1QTNF9, DHRS4, ACOT1, STARD5 and KRT16. Resources are available for a further 97 RFSD genes in knockout mice but the mutant mice have not been phenotyped yet. This however would be a very useful tool going forward and these 97 genes would be good candidates for follow-up studies. The list of 97 genes can be found in the Supplemental Results section. It should be noted that many of the RFSD genes are highly conserved, very old and/or are involved in developmental or signaling processes. Consequently, we estimate there is a non-negligible possibility of any given RFSD gene being essential. This would preclude any knockout mouse lines being generated and the effect of a deleterious mutation can only be interrogated by knock down or partial loss of function experiments.

The mouse data available for the 18 genes that have been phenotyped are all very divergent and inconsistent between them as they are composite results of individual studies that have been performed on mice knocked out for these genes. The data is divided into two types, each with 9 genes representing it. It is presented as either statistical confidence in a knockout mouse presenting a quantitatively measured phenotype that diverges from a wild type expectation or LacZ staining data on various tissues indicating an expression pattern. The quantitative information available for the 9 genes that were measured includes blood chemistry, cardiology, growth, weight and histology, immunology, neurology and behavior, and physical traits. The specific data collected within those categories is quite variable, unlike the expression data for the other 9 genes which is much more consistent. The LacZ staining patterns for those 9 genes suggests they are quite narrowly expressed, with the majority showing no expression in most tissues interrogated and low to moderate expression in a few. The low number of genes in each group greatly reduces our ability to make a broad claim about RFSD gene homologs.

Nevertheless, on a case by case basis useful inferences can be made. For example, PLXNA2 is a semaphorin co-receptor, which are secreted or membrane-bound proteins that mediate nervous system development [130]. Significant deviations from expected values were observed in bone mineral density, bone mineral content, and growth curves in both heterozygous and homozygous knock out mice for the homologous gene. The homozygous mice, however, also exhibited significantly abnormal hemoglobin levels, body and organ tissue weights and aberrant behavior. In particular, the behavior changes observed such as an abnormal gait could be indicative of an incorrectly developed nervous system but it is difficult to state that definitively without follow up studies. Another interesting observation was that each of these phenotypes was more likely to be observed in male mice than female, suggesting there might be a partial sex bias with regard to this gene.

Similarly, useful information can be drawn from the knock out studies for which LacZ data was collected. KRT16 was formed via RFSD from KRT14 in a common ancestor of humans and cows (Branch 10). They both belong to the keratin gene family and are responsible for the structural integrity of epithelial cells [131]. As expected KRT16 shows narrow expression in epithelial tissues including esophagus, foot, skin, and tongue, as well as the vagina in female mice and preputial gland in males. Unexpectedly all mice tested show expression of the gene in the thymus, suggesting that KRT16 may have a larger role in the immune system than previously suspected.

Table 5 Summary of KOMP database search for RFSD genes in the Standard dataset.

KOMP entries found	Number of genes
Data available	18
Resources available but no data	97
Not studied	940

Discussion

Frequency and patterns of reading frame-shift rates in evolution of the human genome

As Jackson and Loeb report, most frameshift mutations have been thought to be inactivating [132]. Previous studies have suggested that frameshifted genes are usually eliminated [28] [41], become pseudogenes [40] or acquire compensatory mutations to restore their original frame [29], yet our study suggests that duplicated genes which are then frameshifted survive far more frequently. The previously unexpected rate of functional frameshifted genes is actually an underestimate, due to the conservative criteria we implemented in the pipelines for identification, for example, excluding shorter reading frames (<50 aa in the Standard dataset or <100 aa in the Conservative dataset) or exon duplications that are frameshifted.

When identifying the ages of these RFSD events we can see a clear increase in their frequency on Branch 10 representing the divergence time between humans and cows or dogs. This divergence time coincides with the beginning of the evolutionary radiation leading to the diversity of extant mammals we can observe [133] [134]. The mammalian radiation began in the late Cretaceous and ended in the early Cenozoic [133] [134]. The dramatic upheavals to the global environment would have radically changed mammalian habitats and selective pressures [133] [134]. RFSD events likely provided the basis for some of the large evolutionary steps necessary to survive and thrive at that time and resulted in the diversity of species that arose.

Initially after a RFS, the frameshifted portion of the gene is unlikely to have functional properties that are meaningfully useful for the cell. However, it will be attached to the non-frameshifted portion of the gene, which is likely to retain its function. If the non-frameshifted fragment has a function where high activity is useful to the cell it may be selected for, even if the remainder of the peptide is non-functional [42]. Marginally functional mutant genes have been observed to amplify and proliferate under selective pressure [62]. Given that we uncovered an excess of RFSD genes with high activity functions, it is likely that this initial advantage greatly increases the odds of the new gene's fixation. Our findings also indicate that RFSD pairs have a very high likelihood of sharing their expression patterns with close duplicates, suggesting the newly generated gene products would work in same pathways of parental genes.

RFSD genes as a proxy for de novo genes

If a gene formed via a RFSD mechanism is almost entirely frameshifted then it can be used as a proxy for a de novo gene as the peptide it forms is completely novel. Due to the limited numbers of de novo genes in the human genome and the lack of information about the conditions under which they arose it is difficult to draw conclusions about de novo gene origination. In fact the details of de novo gene evolution have been described as an unanswerable question [135]. Using RFSD derived genes that are overwhelmingly frameshifted as a proxy for de novo genes can help us make inferences about early de novo gene evolution due to the context we have about RFSD genes which arise from duplications.

When examining RFSD genes that have been entirely frameshifted and we can observe what they have evolved into and we can make inferences about the circumstances influencing their early evolutionary history. For example, SERPINB4 matches SERPINB3 with a nearly complete +1 frameshift (Figure 11). They both arose in the common ancestor of humans and frogs. Given their close genomic proximity to each other it is likely the pair are related via a tandem duplication event followed by a RFS mutation. Although they both have specific expression patterns, SERPINB4 has more specialized expression (2 tissues) than SERPINB3 (5 tissues). Interestingly, they both share a SERPIN domain which represents their core function. This could be a case of convergent evolution as they have closely related but distinct functions. It is possible that the newly duplicated gene evolved a similar structure and function to its parent gene, despite the frameshift, due to an inherited regulatory region which provided the framework for its early evolution. Tandemly duplicated genes often inherit regulatory elements as they are in close proximity to their parent gene's regulatory region. This convergent evolution suggests that the initial expression pattern of a de novo gene plays a large role in directing its evolutionary path.

Another example we can draw inferences from is the match between RAET1L and ULBP2 (Figure 11). Both genes arose in the common ancestor of humans and opossums and have similar functions. Based on their proximity, they also likely arose via a tandem duplication event followed by a frameshift. Their functions also suggest that their shared context guided their evolutionary trajectories. However, ULBP2 encodes for a transmembrane domain at its 3' end whereas RAET1L encodes for two transmembrane domains, one at each end. RAET1L also has a more specific expression pattern than ULBP2, 4 vs 8 tissues. This suggests that RAET1L and ULBP2 are specialized for two distinct niches and fulfill distinct roles despite the similarities they share in molecular function and biological process. This highlights both the power RFSD has to diversify matched genes and that the shared regulatory regions two genes may have are not deterministic but instead a quasi de novo gene can successfully respond to selective pressures to suit the host organism's needs.

One important caveat to take into consideration is that due to the inherited context that often comes with RFSD events, the genes that have very large frameshifts are not true de novo genes although they have many of their characteristics. They may have fully functional enhancers or repressing controlling their expression. They may also have inherited regulation done at the

RNA level such as non-coding RNA regulation which could still recognize the mRNA of a frameshifted gene. They are not quite proto-genes either as those usually encode for a marginally functional peptide while also having a marginally functional regulatory region. Instead these genes were likely expressed in complex and specific expression patterns when they arose while simultaneously encoding for a novel peptide after the RFS mutation. The closest analogy to a true de novo situation is the case of a de novo gene arising near a preexisting enhancer or suite of enhancers which is co-opted into regulating the novel gene.

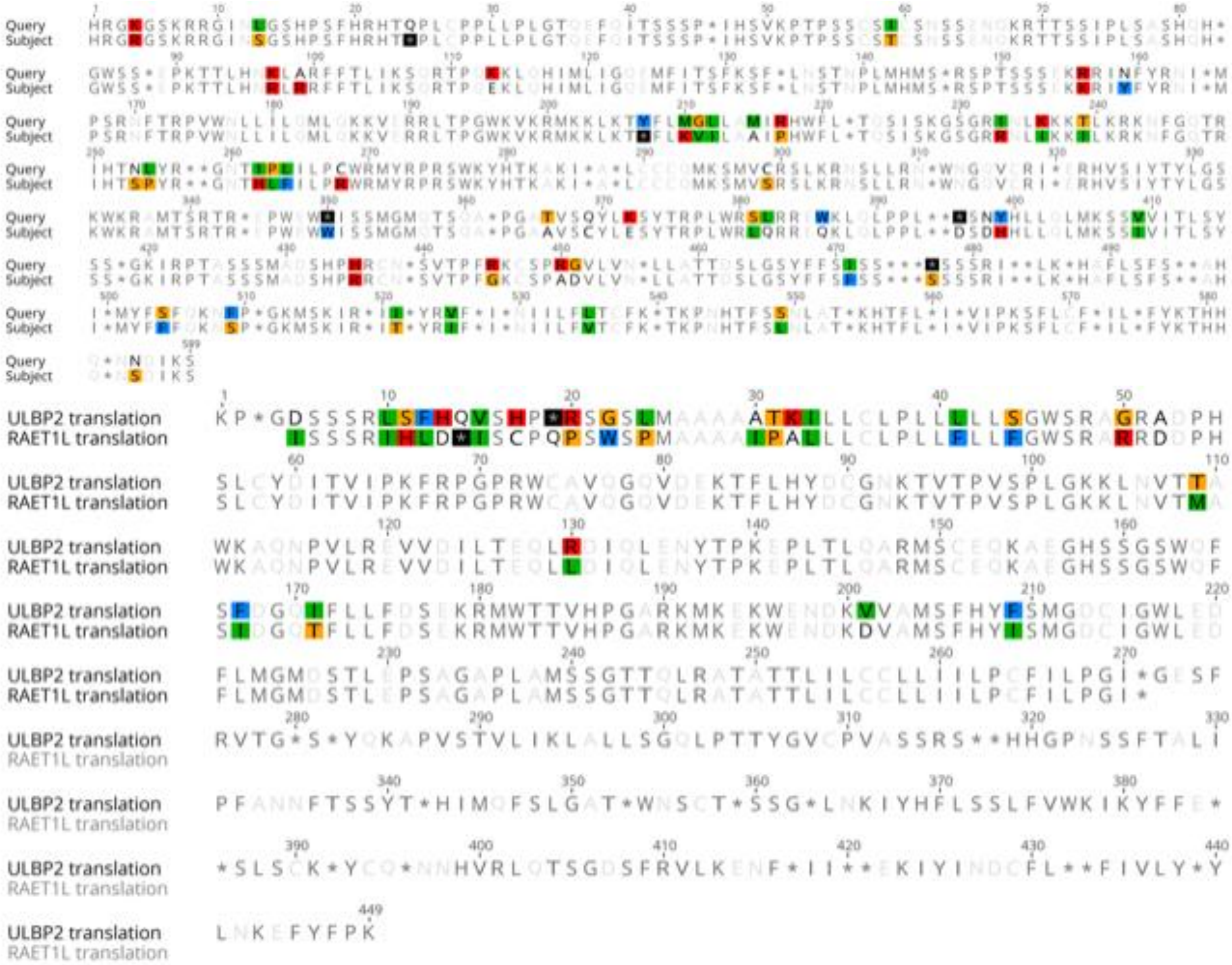


Fig. 11 Examples of matches between RFSD genes with >95% frameshifts. Translated frameshift match between SERPINB3 (indicated as query) and SERPINB4 (indicated as subject) and ULBP2 and RAET1L. SERPINB3 and RAET1L are in frame 1 whereas SERPINb4 and RAET1L are in frame 2 indicating a +1 frameshift.

Human RFSD genes are enriched on the sex chromosomes

Our research reveals an excess of very old RFSD genes on the Y chromosome. The majority of the genes on the Y chromosome date to Branch 2 or older and are paired with genes on the X chromosome. It is likely that most of them evolved prior to the evolution of sex chromosomes as we know them in extant species [136] [137] [138] [139] (the mammalian Y chromosome is approximately 300 million years old [140]). It is possible that some of these genes are not duplicates of their related genes but homologs which diverged from each other at the same time as the chromosomes they were on did. For example, our method identifies RPS4X, found on the X chromosome, and RPS4Y1, found on the Y chromosome, as being connected by a RFSD event. Each of them encodes a version of ribosomal protein S4, a component of the 40S ribosomal subunit [141]. S4 is the only ribosomal protein that is known to be encoded by multiple distinct genes and does not undergo X-inactivation [141]. Each isoform is not identical to the other but is functionally interchangeable [141]. The RFS mutation could have occurred in either but, given that both these genes have orthologs in *S. cerevisiae*, it is likely they were ancestral homologs. Of all the RFSD genes found on the Y chromosome all have a similar relationship to an X-linked gene other than two that are paired with each other and were likely the result of a tandem duplication.

Regardless of the manner in which these genes originally evolved, once the homologs were decoupled from each other, they effectively functioned as duplicate genes. The Y chromosome has been rapidly shrinking due to near constant gene decay caused by silencing and subsequent pseudogenization [140] [142] [143] and yet these ancient genes have been retained. This suggests that a frameshift mutation may be a possible mechanism of diversification and adoption of an essential function which can ensure survival. As all the Y genes but one date to branch 2 or older, they predate the mammalian sex chromosomes and could have survived via this mechanism. None of the RFSD genes are located in the pseudoautosomal region, PAR1, which supports this hypothesis.

The X and Y chromosomes have been shown to have evolved far more rapidly than autosomes in several species, including humans [116] [144]. This was driven by several evolutionary forces, including positive selection leading to the degeneration of the Y chromosome, remodeling of sex chromosomes since their origination from autosomal ancestors and acquisition of new sex-related functions [116] [144] [145] [146]. In this study, we discovered an excess of extreme protein novelties created by frameshifting duplicated genes, or ancestral homologs, on the pair of sex chromosomes, revealing the powerful impact of these evolutionary forces on the protein diversity that shapes who we are and underlies sexual dimorphisms.

RFSD genes are likely to be involved in mammalian signaling pathways

Our study suggests that molecular signaling functions are likely to be inherited or acquired by RFSD genes. This could be a result of the modular nature of signaling molecules, as they often include multiple domains that are each responsible for discrete functions [109] [110] [111]. A RFS which disrupts part of the protein could allow the unaffected portion of the peptide to function as it did previously [109] [147]. This is an excellent example of a case where a frameshift domain can tether a “novel” peptide to a functional one. This would allow the cell to shortcut the process of developing a peptide to recognize a new target or a new transmembrane receptor. The benefit of increased signaling complexity compounded with the flexibility of a RFSD protein and a permissive environment for rapid adaptation are an extremely powerful combination. This conjunction of unlikely elements could possibly explain the diversity of extant signaling in mammals.

Signaling pathways have grown in complexity in the mammalian lineage, a development that likely started early on in mammalian evolution [147] [148]. This increase in complexity has grown with the increasing levels of biodiversity that evolved in the mammalian lineage [147] [148]. The identification of an increase in RFSD events during a time of rapid mammalian diversification may be evidence of fortuitous timing leading to fixation of a lot of signaling genes generated by RFSD or may instead be evidence of an underlying phenomenon that was leveraged by the ancestral mammalian genomes to adapt when selective pressures changed dramatically. In either case, our method can detect the resulting RFSD generated expansion in signaling pathways, both intra- and extracellular.

We propose that RFSD events and the genes involved in them are directly involved in the evolution of mammalian signaling. The data produced by our GO analyses have suggested that RFSD genes may often have signaling functions. However, GO analysis alone is not sufficient evidence to conclude that this is the primary role played by these RFSD genes as even a minor connection is sufficient to associate a GO term with a gene. When combined with the independent inference of an excess of RFSD genes at the base of the mammalian radiation and the supporting evidence collected on mutant RFSD genes causing human signaling defect phenotypes, our GO analysis strongly suggests that RFSD genes played a significant role in developing mammalian signaling.

Mitochondrial proteins are more likely to be encoded by a RFSD gene

Our study suggests that an excess of RFSD genes localize to the mitochondria when compared to the genome average of 5% [149] [150] [151]. Given that the mitochondria are the energy centers of the cell and their proteome is enriched for signaling, metabolic, transport and other high activity functions [152] [153], it is possible that even frameshifted duplicate proteins are more likely to be positively selected for if they localize to the mitochondria and retain some of the parent gene’s activity. We have identified examples of polymorphic genes, such as PDZD8, which fulfill multiple functions and have a modular domain architecture. This results in the duplicated and frameshifted genes produced from them being able to maintain their localization signals or catalytic activity in a way the cell or species can utilize. The mitochondria are a subcellular location where this likely to have an increased effect due to the selective pressure for increased activity. In the data we can identify several cases where paired genes both localize to the mitochondria, although gain and loss of mitochondrial localization can also be observed.

This is supported by our previous conclusions based on the identified molecular functions and shared characteristics of RFSD genes. We can conclude that RFSD genes are enriched for peptides in high activity functional classes and genes formed by a RFSD mechanism are likely to be involved in the same biological process as their parent gene. This strongly suggests the mitochondria as the ideal subcellular location to benefit from this mechanism. The mitochondria has the increased disadvantage of needing a localization signal and transport into the organelle, as well as the devastating effect a disruption can have on the complex cycles within it. A partially functional peptide may confer just enough advantage to survive.

Author Contributions

Alexander Advani (AA) and Manyuan Long (ML) collaboratively conceived this paper. AA designed, interpreted and wrote this paper with edits and assistance from ML. AA assembled the unfiltered datasets and determined the filtering criteria. Filtering was done by Philipp Ross (PR). AA and PR identified the ages of RFSD genes. PR performed the GO and GTEx analysis. All other work done by AA.

Chapter 3: Discussion and Summary

Understanding sources of genetic novelty is a key process in advancing many diverse fields tackling everything from basic questions on evolution to unraveling the genetic underpinnings of human health. In this dissertation I have shown that RFSD derived genes are prevalent throughout the human genome and retain characteristics of their parent genes while introducing novel peptide sequence. The data shown above suggest that RFSD mechanisms are a widespread phenomenon in the human genome and that the following conclusions can be drawn.

RFSD genes are a potential source of significant genetic novelty

Duplicated proteins are known to be maintained if their increased presence or activity is advantageous [55] [56] [57]. However, as described by various models of gene evolution, there are multiple scenarios where novel activity is selected for to allow the organism to adapt [62] [63] [64] [65]. A redundant protein that is part conserved domain and part novel peptide could allow for major adaptive changes on a relatively short timescale. The partially novel protein is also likely to have regulatory elements duplicated with it as suggested by the data in this dissertation. This combination can result in a mature regulatory region governing a partial peptide with the potential of a proto-gene. This suggests that the organism in which it arose had the opportunity to co-opt the nascent function in a new tissue or biological process [42] or to improve the original biological process the unframeshifted protein was involved in with a new specialized protein [72]. The outcome is the RFSD gene may rapidly produce a useful and functional new protein via a relatively small number of mutations. Such opportunities to take large adaptive steps are extremely rare and may be essential to rapid diversification [42] [72].

This potential may be even more significant for functions or processes that are not easy to evolve *de novo* or incrementally. Given that I observed an excess of RFSD genes in the mitochondria, one example of this could be genes that produce cellular components with localization signals found at their N-termini. This would allow novel peptides to be localized to subcellular locations which would be a faster way for an organism to adapt than a series of mutations in an unrelated polypeptide occurring in such a way that a localization signal is formed. This would also be the case for processes that are critical to the cell or organism and hence do not tolerate disruption easily. For example, we might have greater expectations for a novel ribosomal protein to successfully evolve via an RFSD event than entirely *de novo*. This expectation is borne out in our data as we see multiple ribosomal protein in our datasets.

Another class of genes that could benefit from this mechanism are genes that contain multiple independent domains. An example commonly seen in our data are the RFSD genes involved in signaling. Signaling peptides are usually modular as most of them perform two or more discrete functions. This modularity could allow a functional portion of the gene to survive a RFS mutation and increase the probability of the new gene surviving long enough to gain a novel function.

One caveat to this hypothesis is that the identification of an RFSD pair does not account for the proportion of the genes that are frameshifted. I set minimum length requirements for the matches in order to avoid false negatives but I did not set a minimum proportion for two reasons. Firstly, I felt the minimum length requirement was conservative enough and so sufficient to convince me that the identified events were real. Secondly, the proportion of the gene that is frameshifted is something that I would have less power to detect over time as both the genes in each pair could subfunctionalize or even just passively accumulate mutations. In addition, the frameshifted portion of the new gene was likely to adapt to the selective pressures on it and develop further function. This would likely result in significant changes to the gene, maybe including a new stop codon or intron. This in turn would mean portions of the original gene that were duplicated and frameshifted could be lost entirely over time. If that occurred the current proportion of the gene that is frameshifted might not be representative of the size or proportion of the original frameshift.

There is also another hypothesis that can explain the integration of RFSD derived genes in the genome and that is complementary evolution of the biological system the RFSD genes arise in. It is possible that the network of proteins that interact with a newly arisen RFSD gene undergoes several compensatory mutations which allow the assimilation of the new peptide. This would speak to the malleability of biological networks and not the adaptive potential of frameshifted genes. However, the resulting increased complexity of the biological system still depends on the introduction of novel genetic material via a RFSD mechanism, regardless of whether the RFSD gene adapts to its new circumstances or the network adapts to incorporate the new gene.

Inherited expression and function in RFSDs

A frameshift mutation is unlikely to affect regulatory signals in a gene as the cell doesn’t interpret these in-frame. As a result we expect genes related by a RFSD to have a better than random chance of sharing their expression pattern, particularly as most duplication events are tandem duplications or copy regulatory elements as well [18] [68]. It also follows then that if these genes share an expression pattern and are at least partially expressed in the same places they have a better than random chance of being involved in the same biological process. Our findings support this expectation and indicate that RFSD pairs have a very high likelihood of sharing their expression patterns and a good likelihood of being involved in at least one shared biological process.

Conversely due to the drastic change caused by a frameshift mutation we would expect a far smaller proportion of RFSD pairs to share a molecular function. This expectation is borne out by the data collected on domains shared between RFSD gene pairs, as only 495 out of the 1164 share domains. This is supported by the GO analysis of the Standard dataset which actually shows an even lower proportion. This is likely due to peptide domains having multiple uses and being able to contribute to multiple overall molecular functions. For example, a transmembrane domain will control the location of a protein but will not control its function. Similarly, DNA binding domains are found in transcription factors, endonucleases and polymerases which all have distinct functions. A partially frameshifted protein that retains a modular domain could be selected for, regardless of whether the function the protein will ultimately fulfill matches the function of its parent gene. The difference in the proportions can be explained by this flexibility in domain function and architecture.

The Conservative dataset, in this one instance, produces a different result. Our observation that the pairs in the Conservative dataset, with frameshift matches of greater than 100 amino acids, are far more likely to share a molecular function than the complete set of pairs in the Standard dataset is possibly because the Conservative dataset is enriched for very large genes and as a result could be enriched for genes with larger non-frameshifted domains as well. The larger the unframeshifted portion of a gene the more domains it is likely to share with its parent gene. Sharing more domains between two genes increases the chance of these genes sharing a function as well. The more peptide sequence a new protein shares with its parent, the likelier it is to adopt a similar role. Genes with smaller unframeshifted segments have a far greater chance of diverging over time from the ancestral function they shared with their parent genes. Understanding the frequency with which genes formed by a RFSD inherit the characteristics of their parent genes will allow us to infer their evolutionary history and gain a better understanding of how genetic novelty is used in adaptation.

RFSD events gave rise to conserved sex chromosome genes

The Y chromosome has experienced massive gene loss over its evolutionary history [142] [154] [155]. We still do not fully understand the lineage and evolutionary history of many surviving Y chromosome genes [142] [154] [156]. I identified an excess of RFSD genes on the Y chromosome which are paired with X chromosome genes. This association indicates a direct relationship which is usually interpreted as a parent-offspring connection but in this case it may be a result of these pairs being divergent homologs. This is almost certainly the case with RPS4X and RPS4Y1, which each encode for a functionally equivalent but structurally different component of the 40S ribosomal subunit [141]. Both of these genes have orthologs in *S. cerevisiae*, making it extremely unlikely that one of these genes lost its ancestral homolog and was duplicated into the same space during the evolution of the mammalian sex chromosomes. In conjunction with the excess of Y chromosome genes, this suggests that frameshifting redundant homologs could have been a survival strategy for the genes on the Y.

This hypothesis is supported by the existence of the linked X-Y gene pairs solely outside the pseudoautosomal regions. Given that most duplication events are tandem, it would be possible for a gene in one of the pseudoautosomal regions to duplicate and then frameshift, essentially the same scenario that occurs on any autosome. This is slightly less feasible on the sex chromosomes due to the constant selective pressure for the Y chromosome to remain as small as possible.

Nevertheless we see a significant excess of RFSD genes on the sex chromosomes, especially driven by an excess on the Y. This suggests that the RFSD mechanism played a role in the survival of the Y chromosome genes.

The survival of the RFSD genes, while the majority of the ancestral Y chromosome genes degraded and were lost, could have been the result of acquiring novel function or becoming different enough to avoid redundancy. This could have led to the observed significant increase in the frequency of frameshifted Y genes. Most RFSD genes identified on the Y are very old (Branches 0 - 2) and predate the evolution of the sex chromosomes as we currently know them. This strategy of avoiding pseudogenization and degradation can be extrapolated to other classes of redundant genes. Examining more classes of formerly redundant genes for evidence of frameshifts in their history might reveal that frameshifting is a way of leveraging redundancy to adapt or avoid elimination.

RFSD genes can take various adaptive paths to function

As shown above, many of the genes involved in RFSDs were found to have molecular functions that are involved in many high activity biological processes and contexts, such as GTP binding or transcriptional activity. It is possible that these are the most likely functions to survive a RFSD event because the functions the genes retain from the unframeshifted portion are useful in the new environment and are maintained. In addition, all these genes are reasonably large (at least 50 amino acids match in a different frame, over 100 amino acids match for more than half of identified events) and are more likely to have multiple functions. There are three possible interpretations of these results.

It is possible that these high activity functions identified for these genes are the functions they retained from their inherited unframeshifted portions. Even though the proportion of gene pairs that share a molecular function is relatively small, it's large enough to contribute a significant amount to the results of the GO analyses. This would suggest these are the functions that best pair with a frameshifted domain or that are the most advantageous in a new context.

The second explanation would be that the frameshifted parts of the duplicated genes evolved convergent functions. This would only be possible if the functions identified were the easiest functions to evolve from a more or less random peptide in a stepwise fashion or if there is a very strong selective pressure for an increase in the prevalence of these functions in the human lineage.

Finally, it's possible that the inherited regulatory elements and genetic context from the parent genes contribute to convergence on a specific functions. If the regulatory region governing the parent gene is inherited and the new gene is expressed during development it is probably going to acquire a developmental function. If the new protein is expressed in the central nervous system it might obtain signaling activity.

The true situation is probably a combination of all three to varying degrees, depending on the needs and selective pressures being faced by the species when an RFSD event occurs. If the new gene is not advantageous, even if it is only a few small steps away from independent function, it will not be maintained and the species will have to wait for a different mutation or die out. The RFSD genes are most likely a mix of all possible methods of generating a new useful peptide and the specific context of their origination are probably optimized for the needs of the cell at the time.

Competing models can explain observed frequency of RFSDs

There are two competing models that can explain the results detailed above. Firstly, duplicated and frameshifted genes may have been surviving at much higher rates than previously thought throughout evolutionary history. Alternatively, frameshift mutations occur relatively frequently and frameshifted genes have been surviving with much greater frequency in the human lineage. These competing models are mutually exclusive but without further data I cannot definitively state that one is true.

The possibility of much greater frequency of RFSD events occurring than previously thought is the simplest explanation and is plausible given that few studies have thoroughly investigated the prevalence of RFSDs thus far. For this to be true two distinct events have to occur regularly, DNA duplication in some form and frameshift mutations in the newly duplicated DNA. We know that gene duplication through replication errors, retrotransposition of RNA or TE duplication occurs relatively frequently [157] [158] [159]. Although high rates of survival for frameshifted genes is contrary to our prior expectations, the data presented in this thesis and some reports discussed previously suggest those expectations may be based on at least one invalid assumption. Older studies on frameshifted proteins have mainly focused on clinical settings where a frameshift is causing a disease state. When outside the clinical arena frameshift studies have not taken the benefits of generating a frameshifted and quasi-novel peptide into account until recently. If frameshift tolerance is higher than previously believed, particularly when the frameshifted protein is partially or completely redundant, it would explain the high frequency of RFSD genes observed throughout the human lineage.

The second scenario of frameshift mutations occurring and surviving with much greater frequency in the human lineage is also possible. A limitation of the method I used to identify RFSD events is that I used only the human genome to identify frameshifted genes. Although we can tell the age of each gene by identifying homologs in other species, we do not know if the homologous genes are also frameshifted. This leaves open the possibility that these genes are duplicated in many species but the frameshifts we have identified are a human specific phenomenon. Given the number of RFSD genes identified, this is a less likely model than the first one because it would require frameshift mutations to occur quite frequently and the human species would have to be exceptionally tolerant of frameshifted genes, something we have no evidence of so far. However, based on the data available this possibility cannot be excluded.

Additional and future work

The implications of the work I have completed are quite broad and there are inferences that can be made which would affect a diverse number of fields. As a result there are many potential follow up studies that can be done but I have focused on four of the most promising or direct studies can be done to further this field or expand our knowledge of the topics raised in this dissertation.

The most apparent and significant follow up study which can be done is to repeat the study done above for other focal species. Ideally, if the species I used to determine the ages of the RFSD events were studied for RFSD derived genes in their own genomes we would gain two extremely valuable pieces of information. Firstly, we would know whether the human RFSD genes I identified are frameshifted in other species as well. This would answer the pressing question of whether the phenomenon described is intrinsic to humans in some way or whether it can be observed more broadly across the tree of life. Secondly, taking a human centric view, we would be able to more accurately date the RFS mutation in each gene by determining which species have frameshifted orthologs and which have retained unframeshifted ones. This would essentially allow us to decouple the duplication event from the frameshift mutation. The benefits of doing this include more accuracy regarding the rate of RFSD occurrence, finer resolution on the circumstances under which RFS mutations can survive and by extension better inferences about the utility of frameshifting in adaptation, and more detailed interpretations of the evolutionary history of our species.

Another, study which directly arises out of this thesis is the direct identification of the function of frameshifted portions of RFSD genes. To fully investigate this I would recommend two parallel approaches. Firstly, I would suggest isolating the extant sequence that corresponds to the frameshifted portion of an RFSD gene and run it through a database of Position Weight Matrices for known or predicted domains. This would allow computational identification of any functional sequence and allow designation of a putative function. I would then biochemically test that sequence, in cell lines if possible or with purified protein if not, for the suspected function and validate the assigned presumed function. This would directly answer several questions about the value of a frameshifted peptide and the type of function it can acquire.

A third study that stems from this thesis is to conduct knock out experiments in human cell lines for a subset of the RFSD genes I have identified. I would focus on genes that are not already well characterized, as some of the genes' phenotypes and functions have been well documented and can be collected by a simple, albeit time consuming, literature search. The knock out experiments would allow us to answer questions about essentiality of RFSD genes and the functions they have evolved. I would be particularly interested to learn more about the functions of the genes in my datasets which have no known conserved domains. They have been maintained for millions of years and have been shown to be translated in humans so these genes are presumably performing a useful function for the cell or organism. I would also recommend performing these experiments in neuronal cells as well, as we have shown a large number of RFSD genes are involved in neurological pathways.

Finally, I would suggest investigating the pathways the identified RFSD genes are involved in. I have proposed the hypothesis that the RFSD mechanism can be co-opted to rapidly diversify pathways, in particular intra- and extracellular signaling cascades. This could be investigated by searching for data on the known pathways each gene is a part of and grouping the RFSD genes by shared pathway. As many genes are polymorphic and/or involved in multiple pathways, it might be possible to create a gene network of pathways linked by RFSD genes. Determining which pathways genes formed by RFSD have integrated into can help us learn how frameshift mutations are picked up by existing biological networks. If compared with their corresponding parent genes we can also distinguish between networks that take up a RFSD derived gene because the new gene's inherited context made it convenient and networks that have a true ability to absorb and utilize new frameshifted proteins.

Completion of these four studies will give us a much clearer picture of the role RFSD genes play in our evolutionary history. There are many other projects that could be undertaken however, such as frameshifting a protein and tracking its evolution in bacteria or *in silico*, searching model populations for naturally occurring frameshifts and determining current rates of frameshifting, or inserting duplicated genes into a model population and determining how long it would take for a new frameshift mutation to appear in a redundant protein. These are only a few of the potential ideas that can be sparked by this project and they primarily focus on the field of evolutionary genetics. The implications for medicine, biochemistry and molecular biology are profound as well and there are numerous projects that can be carried out in those fields and others depending on the particular interest of the person who will undertake them.

Conclusions

In this dissertation I have described a dataset of 1055 genes involved in 628 RFSD events based on extremely conservative criteria. By cross-referencing the dataset with other available data I have determined the patterns present in their molecular functions, biological processes, expressions and locations. Furthermore, I have ascertained which frameshifted proportions, domains and characteristics are likely to be inherited from parent gene to offspring, which can allow us to infer the evolution of RFSD genes. I have determined whether RFSD genes are associated with human diseases and which types, supported by some data from available mouse knock out information. I have also described the existence of linkage via this mechanism between ancient homologs on the human sex chromosomes and identified signaling pathways that may have taken up RFSD derived genes.

In addition, I have shown evidence to support RFSD mechanisms as a significant source of genetic novelty. There are few other known mechanisms that have such potential to rapidly diversify a gene’s or gene family’s functions and permit large adaptive steps. The combination of a parental regulatory region with the instant and dramatic partial divergence of the protein product places RFSD genes in the unique position of simultaneously having a redundant function which can be selected for immediately after the frameshift mutation and a random sequence similar to a proto-gene. This potential appears to have been used repeatedly by genes throughout the evolutionary history of our species to avoid degradation, amplify and diversify, shortcut subcellular innovation or bring together previously disparate functions.

This mechanism of introducing genetic novelty definitely warrants further study as there remain many unanswered questions about the details by which this process operates in humans and more broadly all DNA-based organisms. Using molecular and bioinformatic approaches to better understand the nature of RFSD events and expanding my datasets to include other species or types of data will grant us a clearer idea of the scope of this mechanism of gene origination and genetic novelty. A comprehensive database of genes formed by RFSDs will allow the identification of previously unknown relationships between genes and give us a better understanding of how genetic novelty can be co-opted and reconciled with preexisting biological networks.

Chapter 4: RFSD Gene Pair Datasets and Supplemental Results

RFSD Gene Pair Datasets

The RFSD gene pair datasets were collected by identifying all expressed human cDNAs which matched other expressed human cDNAs in a different frame and filtering them by the criteria listed in the Methods section. The final datasets produced are listed below along with summaries of the criteria used to filter them. The e-value is the e-value associated with the match, the alignment length is the length of the match in amino acids and the percentage identity refers to the proportion of the match that is identical between query gene and subject gene. The Standard dataset summary lists all matches. The Conservative dataset summary lists only one match out of every pair because all matches are identically reciprocal. It is important to note that due to the higher alignment length criteria of the Conservative dataset, the dataset is enriched for larger frameshifts and by extension much larger genes. This overall increase in average gene size means the Conservative dataset is enriched for genes with larger unframeshifted portions as well as frameshifted portions. The Standard dataset represents all RFSD gene pairs and the frames in which they matched.

Standard dataset

Table 6 Summary of identified RFSD gene pairs in the Standard dataset.

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000005073	2	ENSG00000128713	12	3.68E-71	86	81.395	1	0.10955414
ENSG00000006451	2	ENSG00000144118	2	1.07E-71	126	89.683	-2	0.136363636
ENSG00000004975	2	ENSG00000161202	3	0	131	91.603	1	0.137316562
ENSG00000004975	2	ENSG00000107404	2	0	131	81.679	1	0.137316562
ENSG00000006116	2	ENSG00000166862	1	5.47E-132	175	84.571	-1	0.193584071
ENSG00000006059	7	ENSG00000131738	2	0	386	93.264	2	0.925659472
ENSG00000006059	7	ENSG00000094796	7	0	373	93.834	2	0.894484412
ENSG00000010017	2	ENSG00000141084	7	2.65E-154	110	83.636	1	0.138539043
ENSG00000015568	2	ENSG00000183054	2	0	946	100	1	0.535977337
ENSG00000015568	2	ENSG00000169629	2	0	946	99.683	1	0.535977337
ENSG00000088256	2	ENSG00000156052	1	0	360	90.278	-1	0.669144981
ENSG00000088256	2	ENSG00000156049	1	0	356	81.742	-1	0.661710037
ENSG00000050327	2	ENSG00000213214	4	0	389	99.486	1	0.239237392
ENSG00000019549	1	ENSG00000124216	2	4.92E-73	113	85.841	-1	0.168656716
ENSG00000112852	9	ENSG00000120327	0	0	124	95.161	1	0.134929271
ENSG00000186847	8	ENSG00000128422	3	0	312	89.423	2	0.537931034
ENSG00000186847	8	ENSG00000186832	2	0	312	88.141	2	0.537931034
ENSG00000068976	2	ENSG00000100994	2	0	833	84.154	2	0.83718593

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000083720	1	ENSG00000198754	1	0	221	80.543	-2	0.201275046
ENSG00000197208	4	ENSG00000197375	1	0	322	86.025	2	0.436314363
ENSG00000100490	2	ENSG00000125375	1	2.67E-175	258	99.225	-2	0.684350133
ENSG00000101162	3	ENSG00000124172	2	0	395	100	1	0.339055794
ENSG00000101162	3	ENSG00000137285	2	0	433	80.6	-2	0.37167382
ENSG00000099804	2	ENSG00000107341	3	8.22E-122	198	86.869	1	0.406570842
ENSG00000083812	11	ENSG00000249471	1	0	516	90.891	1	0.529774127
ENSG00000087303	2	ENSG00000087302	1	0	236	98.729	-3	0.191403082
ENSG00000183741	3	ENSG00000141570	0	2.85E-40	79	86.076	-1	0.073080481
ENSG00000095917	13	ENSG00000172236	9	0	282	85.106	-1	0.84939759
ENSG00000099974	4	ENSG00000099977	0	4.27E-61	104	96.154	1	0.776119403
ENSG00000100994	2	ENSG00000068976	2	0	697	83.07	-2	0.503612717
ENSG00000100450	9	ENSG00000100453	0	4.38E-108	129	86.047	2	0.369627507
ENSG00000100564	2	ENSG00000054690	3	3.88E-65	100	100	-1	0.215053763
ENSG00000100314	3	ENSG00000100319	2	5.87E-83	139	97.122	1	0.137080868
ENSG00000100554	0	ENSG00000134001	2	1.70E-51	84	100	-3	0.161538462
ENSG00000101292	3	ENSG00000169618	2	1.10E-152	241	84.232	-1	0.572446556
ENSG00000101405	2	ENSG00000101200	2	2.33E-54	107	80.374	1	0.633136095
ENSG00000100528	1	ENSG00000143786	2	3.92E-49	51	82.353	-2	0.106918239
ENSG00000090581	2	ENSG00000059145	2	0	302	100	-3	0.743842365

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000088356	1	ENSG00000131044	1	2.97E-133	207	100	1	0.473684211
ENSG00000086232	2	ENSG00000106305	2	7.52E-106	178	100	-2	0.188559322
ENSG00000100453	10	ENSG00000100450	3	7.72E-110	152	84.211	-2	0.501650165
ENSG00000092607	3	ENSG00000112837	5	7.74E-157	234	85.47	2	0.396610169
ENSG00000243811	9	ENSG00000128394	0	0	246	88.211	-2	0.637305699
ENSG00000243811	9	ENSG00000244509	3	3.73E-152	119	82.353	-1	0.308290155
ENSG00000100030	1	ENSG00000102882	7	0	346	88.15	1	0.198167239
ENSG00000111981	8	ENSG00000131019	2	1.48E-77	85	89.412	1	0.080721747
ENSG00000109061	7	ENSG00000264424	1	0	630	95.873	-2	0.32208589
ENSG00000113211	10	ENSG00000113209	0	0	299	83.612	1	0.376574307
ENSG00000088038	0	ENSG00000105617	3	1.05E-74	68	100	-2	0.072110286
ENSG00000113209	7	ENSG00000113211	2	0	104	91.346	2	0.107106076
ENSG00000016082	1	ENSG00000159556	2	0	188	81.383	-1	0.230392157
ENSG00000105664	7	ENSG00000113296	0	0	271	87.823	1	0.330487805
ENSG00000111725	1	ENSG00000131791	2	1.37E-107	95	88.421	1	0.117428925
ENSG00000102128	12	ENSG00000172476	3	0	254	98.031	-1	0.740524781
ENSG00000105649	2	ENSG00000152932	2	1.68E-123	194	88.66	-1	0.390342052
ENSG00000104863	1	ENSG00000148943	1	2.56E-101	204	82.353	-2	0.822580645
ENSG00000104129	1	ENSG00000137880	2	9.50E-128	161	100	3	0.473529412
ENSG00000108773	3	ENSG00000114166	2	0	239	83.682	2	0.230028874

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000104888	3	ENSG00000091664	1	0	512	82.422	1	0.519269777
ENSG00000105254	1	ENSG00000105258	1	3.44E-79	120	100	-3	0.355029586
ENSG00000114853	2	ENSG00000198740	1	0	260	91.538	-1	0.315917375
ENSG00000113205	6	ENSG00000113248	2	0	73	83.562	2	0.087845969
ENSG00000105258	0	ENSG00000105254	9	1.70E-77	120	100	1	0.38585209
ENSG00000108379	2	ENSG00000154342	1	0	347	85.014	-1	0.321296296
ENSG00000108590	0	ENSG00000129235	4	0	408	100	-1	0.739130435
ENSG00000111615	0	ENSG00000139278	2	0	379	99.736	3	0.338695264
ENSG00000113248	9	ENSG00000113205	1	0	158	89.241	1	0.168443497
ENSG00000109272	12	ENSG00000163737	0	2.29E-47	102	87.255	-1	0.822580645
ENSG00000106305	1	ENSG00000086232	11	1.88E-106	178	100	2	0.446115288
ENSG00000108759	8	ENSG00000197079	1	1.25E-179	329	80.851	-1	0.734375
ENSG00000091010	1	ENSG00000152192	3	2.80E-119	165	89.091	-2	0.418781726
ENSG00000039123	0	ENSG00000067113	1	1.23E-100	160	100	-2	0.153550864
ENSG00000108417	10	ENSG00000171360	2	0	252	87.302	-1	0.561247216
ENSG00000106004	2	ENSG00000120075	2	3.49E-78	80	93.75	-1	0.207253886
ENSG00000106004	2	ENSG00000172789	2	2.90E-41	65	87.692	-1	0.168393782
ENSG00000060138	1	ENSG00000065978	1	7.86E-56	91	94.505	1	0.244623656
ENSG00000111639	1	ENSG00000010292	2	6.78E-42	73	100	1	0.32735426
ENSG00000114349	2	ENSG00000134183	2	0	326	83.129	1	0.926136364

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000103064	3	ENSG00000103061	0	0	977	100	1	0.467911877
ENSG00000107796	2	ENSG00000159251	2	0	388	96.907	2	0.816842105
ENSG00000108468	1	ENSG00000122565	2	1.30E-59	67	86.567	2	0.090053763
ENSG00000113722	6	ENSG00000165556	1	2.31E-37	64	93.75	-1	0.146788991
ENSG00000113722	6	ENSG00000131264	1	1.22E-32	63	88.889	-1	0.144495413
ENSG00000105617	1	ENSG00000088038	2	3.83E-75	68	100	2	0.225165563
ENSG00000112309	2	ENSG00000112305	1	0	519	100	-3	0.25
ENSG00000113212	10	ENSG00000120324	2	0	106	90.566	2	0.104228122
ENSG00000113212	10	ENSG00000177839	9	6.60E-151	65	87.692	-1	0.063913471
ENSG00000107018	5	ENSG00000107014	1	5.52E-139	260	83.462	1	0.785498489
ENSG00000109208	12	ENSG00000171201	2	9.67E-97	59	94.915	1	0.375796178
ENSG00000123908	1	ENSG00000092847	3	0	838	83.652	-2	0.82480315
ENSG00000109132	2	ENSG00000165462	3	1.75E-55	92	82.609	1	0.184368737
ENSG00000129514	3	ENSG00000125798	1	6.21E-85	109	89.908	2	0.113778706
ENSG00000127720	1	ENSG00000133773	14	1.31E-58	97	100	-3	0.139568345
ENSG00000132207	0	ENSG00000181625	14	0	244	100	-1	0.887272727
ENSG00000105613	3	ENSG00000086015	14	0	557	80.969	1	0.29517753
ENSG00000171360	10	ENSG00000108417	13	2.17E-169	159	90.566	1	0.18233945
ENSG00000254245	9	ENSG00000081853	1	0	412	100	-1	0.260924636
ENSG00000059145	2	ENSG00000090581	1	0	302	100	3	0.363855422

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000130544	13	ENSG00000167785	3	0	170	81.765	1	0.367965368
ENSG00000134590	9	ENSG00000203950	2	1.66E-171	179	86.592	1	0.428229665
ENSG00000134590	9	ENSG00000212747	3	1.28E-158	179	87.709	2	0.428229665
ENSG00000122696	9	ENSG00000141437	3	0	306	95.425	1	0.4896
ENSG00000120094	2	ENSG00000105991	4	5.20E-38	54	85.185	1	0.113924051
ENSG00000134108	2	ENSG00000143862	4	2.85E-115	183	91.257	1	0.181727905
ENSG00000130810	0	ENSG00000243207	1	0	337	100	-2	0.602862254
ENSG00000094796	14	ENSG00000006059	2	0	359	95.822	-2	0.76059322
ENSG00000094796	14	ENSG00000131738	1	2.05E-177	192	93.229	1	0.406779661
ENSG00000121068	3	ENSG00000135111	2	4.79E-169	239	87.448	-1	0.23454367
ENSG00000130449	3	ENSG00000162415	2	0	532	80.263	2	0.437860082
ENSG00000282608	3	ENSG00000121933	3	9.93E-121	163	100	1	0.26986755
ENSG00000123143	3	ENSG00000065243	1	0	331	80.363	2	0.281223449
ENSG00000196757	12	ENSG00000171291	2	0	112	88.393	-1	0.127562642
ENSG00000119638	1	ENSG00000160602	9	4.07E-74	71	87.324	-1	0.069744597
ENSG00000120324	9	ENSG00000177839	10	0	672	92.411	-2	0.616513761
ENSG00000120324	9	ENSG00000113212	2	0	101	90.099	-2	0.09266055
ENSG00000124140	2	ENSG00000113504	1	0	256	89.453	1	0.128902316
ENSG00000122543	3	ENSG00000135175	0	2.30E-153	231	98.268	2	0.995689655
ENSG00000131094	3	ENSG00000165985	2	4.67E-87	133	87.218	-1	0.263366337

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000131094	3	ENSG00000186897	0	7.33E-87	134	88.06	-2	0.265346535
ENSG00000130733	1	ENSG00000142453	2	1.09E-136	245	100	-2	0.472972973
ENSG00000134365	4	ENSG00000116785	3	0	191	80.105	2	0.342908438
ENSG00000115042	2	ENSG00000144199	2	0	329	96.657	-2	0.740990991
ENSG00000124172	6	ENSG00000101162	2	0	395	100	1	0.75095057
ENSG00000160145	3	ENSG00000038382	3	0	182	86.264	-1	0.10944077
ENSG00000131462	1	ENSG00000037042	1	0	454	97.577	-1	0.828467153
ENSG00000127780	10	ENSG00000180016	2	5.45E-164	192	91.667	1	0.590769231
ENSG00000120322	13	ENSG00000187372	2	0	421	93.587	-1	0.466223699
ENSG00000120327	10	ENSG00000112852	2	0	110	83.636	-1	0.134969325
ENSG00000115486	1	ENSG00000168906	2	3.62E-126	185	100	-2	0.1716141
ENSG00000120329	9	ENSG00000102743	1	0	315	86.667	2	0.667372881
ENSG00000129204	2	ENSG00000170832	1	0	779	92.94	-1	0.554054054
ENSG00000132915	4	ENSG00000133256	3	0	326	82.209	1	0.333333333
ENSG00000121281	3	ENSG00000166164	12	0	725	100	2	0.43622142
ENSG00000134250	2	ENSG00000264343	0	2.59E-159	238	97.479	-2	0.09631728
ENSG00000126778	1	ENSG00000170577	3	1.59E-128	188	95.213	2	0.412280702
ENSG00000128383	13	ENSG00000179750	2	7.76E-173	229	93.886	-2	0.898039216
ENSG00000213366	9	ENSG00000134184	2	0	385	87.532	1	0.9697733
ENSG00000059122	7	ENSG00000162076	0	6.51E-49	67	85.075	1	0.040167866

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000124216	1	ENSG00000019549	3	1.92E-75	113	85.841	1	0.199294533
ENSG00000120952	10	ENSG00000116721	2	0	547	94.333	1	1
ENSG00000120952	10	ENSG00000204481	3	0	547	92.505	1	1
ENSG00000116132	1	ENSG00000167157	3	7.01E-60	71	92.958	1	0.257246377
ENSG00000130950	9	ENSG00000188152	3	0	377	97.347	2	0.498677249
ENSG00000119778	0	ENSG00000156802	1	0	335	81.791	-2	0.229766804
ENSG00000133773	1	ENSG00000127720	2	3.76E-64	97	100	2	0.27247191
ENSG00000058262	0	ENSG00000065665	1	0	475	93.684	1	0.391591096
ENSG00000121297	1	ENSG00000179981	2	0	148	85.135	-1	0.125636672
ENSG00000105705	1	ENSG00000129933	1	4.67E-32	54	100	3	0.07703281
ENSG00000185479	9	ENSG00000170465	1	0	190	100	-1	0.244845361
ENSG00000185479	9	ENSG00000205420	3	0	190	97.895	-1	0.244845361
ENSG00000120903	2	ENSG00000101204	5	0	307	80.13	-1	0.433615819
ENSG00000131791	3	ENSG00000111725	14	2.94E-107	95	88.421	-1	0.052486188
ENSG00000197375	4	ENSG00000197208	1	0	322	86.025	-2	0.343649947
ENSG00000109805	0	ENSG00000178177	0	0	474	100	2	0.440111421
ENSG00000028839	1	ENSG00000146411	0	0	463	100	-2	0.749190939
ENSG00000125398	2	ENSG00000100146	3	7.06E-129	121	90.909	1	0.140534262
ENSG00000083307	2	ENSG00000134317	3	0	53	92.453	-1	0.051158301
ENSG00000119669	2	ENSG00000170604	3	9.54E-134	84	89.286	-1	0.105527638

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000119669	2	ENSG00000168264	1	1.62E-75	61	85.246	1	0.076633166
ENSG00000124657	7	ENSG00000168131	0	5.33E-171	310	81.613	2	0.981012658
ENSG00000119729	2	ENSG00000151665	5	2.05E-143	219	100	-2	0.337442219
ENSG00000119729	2	ENSG00000126785	2	4.20E-100	190	80	-1	0.292758089
ENSG00000116035	1	ENSG00000148704	7	8.92E-57	97	90.722	-2	0.255263158
ENSG00000129824	0	ENSG00000198034	1	1.15E-171	265	92.83	-1	0.85483871
ENSG00000116455	2	ENSG00000116459	0	3.08E-70	112	97.321	1	0.138442522
ENSG00000187545	10	ENSG00000204479	0	0	489	86.912	1	0.964497041
ENSG00000264424	7	ENSG00000109061	2	0	862	95.592	2	0.444559051
ENSG00000099822	3	ENSG00000138622	2	0	550	90	-1	0.618672666
ENSG00000099822	3	ENSG00000164588	2	0	545	85.138	-1	0.613048369
ENSG00000170465	9	ENSG00000185479	1	0	190	100	1	0.243277849
ENSG00000125966	2	ENSG00000156103	2	0	176	85.227	-1	0.272868217
ENSG00000125966	2	ENSG00000157227	2	0	112	81.25	-1	0.173643411
ENSG00000118579	1	ENSG00000047662	5	0	1107	100	-3	0.976190476
ENSG00000105819	0	ENSG00000105821	5	2.50E-48	58	91.379	-1	0.118609407
ENSG00000187272	11	ENSG00000241595	2	4.92E-135	172	83.14	-1	0.982857143
ENSG00000124766	2	ENSG00000176887	2	1.16E-68	88	93.182	1	0.125356125
ENSG00000133256	2	ENSG00000132915	2	5.22E-32	71	80.282	2	0.068532819
ENSG00000131668	2	ENSG00000043039	2	8.03E-46	72	80.556	-1	0.140350877

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000138286	2	ENSG00000213551	2	7.60E-176	232	100	-3	0.39862543
ENSG00000129235	1	ENSG00000108590	2	0	408	100	-1	0.542553191
ENSG00000129235	1	ENSG00000198920	4	1.77E-33	56	100	3	0.074468085
ENSG00000037042	9	ENSG00000131462	0	0	465	96.344	1	0.752427184
ENSG00000120328	10	ENSG00000197479	11	0	672	89.583	1	0.588957055
ENSG00000120328	10	ENSG00000120327	11	0	117	89.744	1	0.10254163
ENSG00000128245	2	ENSG00000170027	2	4.00E-142	246	86.992	1	0.420512821
ENSG00000132475	1	ENSG00000188375	10	0	129	96.899	-1	0.345844504
ENSG00000134001	0	ENSG00000100554	1	4.34E-51	84	100	3	0.084507042
ENSG00000119673	9	ENSG00000184227	0	0	304	98.355	1	0.513513514
ENSG00000131738	10	ENSG00000006059	8	0	386	93.264	-2	0.714814815
ENSG00000131738	10	ENSG00000094796	3	0	183	93.443	-1	0.338888889
ENSG00000120075	2	ENSG00000106004	7	2.33E-68	80	93.75	1	0.211640212
ENSG00000128713	2	ENSG00000005073	2	5.86E-68	77	89.61	-1	0.227810651
ENSG00000131459	1	ENSG00000198380	3	0	446	81.166	-2	0.438544739
ENSG00000115386	9	ENSG00000172023	15	2.78E-108	217	81.106	1	0.84765625
ENSG00000112659	2	ENSG00000044090	12	0	210	85.714	2	0.083432658
ENSG00000188536	3	ENSG00000206172	1	5.22E-112	178	95.506	1	0.843601896
ENSG00000128340	1	ENSG00000169750	12	9.70E-120	203	85.714	-2	0.400394477
ENSG00000125629	0	ENSG00000186480	3	3.09E-110	185	84.324	-1	0.213872832

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000133243	2	ENSG00000064726	0	0	379	84.697	2	0.827510917
ENSG00000134072	1	ENSG00000183049	10	0	323	82.353	1	0.700650759
ENSG00000105821	0	ENSG00000105819	1	3.70E-69	57	100	2	0.09178744
ENSG00000182187	7	ENSG00000163254	0	1.19E-98	83	95.181	2	0.399038462
ENSG00000182187	7	ENSG00000168582	1	4.43E-95	86	80.233	2	0.413461538
ENSG00000103769	1	ENSG00000185236	2	3.58E-61	98	80.612	-1	0.119366626
ENSG00000136231	2	ENSG00000159217	1	0	188	80.319	1	0.254397835
ENSG00000144118	2	ENSG00000006451	8	1.34E-69	126	89.683	2	0.173553719
ENSG00000141570	3	ENSG00000183741	1	8.94E-38	50	88	1	0.099403579
ENSG00000129933	1	ENSG00000105705	1	7.82E-29	50	100	-3	0.081566069
ENSG00000116017	3	ENSG00000179361	2	7.22E-100	138	83.333	-2	0.146652497
ENSG00000121454	2	ENSG00000107187	2	1.23E-138	127	80.315	1	0.220103986
ENSG00000125492	3	ENSG00000143032	2	7.08E-37	84	82.143	1	0.196261682
ENSG00000139648	9	ENSG00000186049	2	0	365	89.315	1	0.5703125
ENSG00000139648	9	ENSG00000170484	1	0	364	87.088	-1	0.56875
ENSG00000005339	2	ENSG00000100393	9	0	471	90.446	1	0.192874693
ENSG00000083750	0	ENSG00000155876	6	0	273	97.802	-1	0.383966245
ENSG00000138109	12	ENSG00000165841	6	0	513	88.499	-2	0.946494465
ENSG00000138109	12	ENSG00000108242	7	0	484	81.818	1	0.89298893
ENSG00000141437	9	ENSG00000122696	10	0	311	93.248	-1	0.896253602

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000067113	2	ENSG00000039123	10	4.27E-101	160	100	2	0.561403509
ENSG00000240403	6	ENSG00000242019	10	0	70	87.143	-1	0.140280561
ENSG00000109181	15	ENSG00000171234	9	0	556	83.273	-2	0.604347826
ENSG00000101624	2	ENSG00000128789	2	2.08E-35	85	92.941	1	0.088449532
ENSG00000047457	2	ENSG00000163755	2	0	411	100	-2	0.3425
ENSG00000128881	1	ENSG00000146216	1	0	314	81.529	1	0.231392778
ENSG00000136240	0	ENSG00000105438	3	1.05E-123	213	83.568	-2	0.547557841
ENSG00000091664	2	ENSG00000104888	6	0	512	82.422	-1	0.76077266
ENSG00000100393	2	ENSG00000005339	6	0	490	88.163	-1	0.202982601
ENSG00000126934	1	ENSG00000169032	2	0	228	92.544	-2	0.389078498
ENSG00000139797	0	ENSG00000125352	2	2.21E-94	52	84.615	1	0.112068966
ENSG00000146216	4	ENSG00000128881	2	0	317	81.073	-1	0.23996972
ENSG00000044090	2	ENSG00000112659	2	0	223	84.305	-2	0.121592148
ENSG00000137285	1	ENSG00000137267	2	0	428	98.832	2	0.670846395
ENSG00000115808	2	ENSG00000196792	2	0	99	83.838	2	0.126923077
ENSG00000137273	6	ENSG00000103241	2	4.21E-66	113	96.46	1	0.254504505
ENSG00000139133	0	ENSG00000175548	1	4.09E-95	187	85.027	2	0.31270903
ENSG00000103241	3	ENSG00000137273	9	1.65E-68	127	90.551	-1	0.15356711
ENSG00000113504	2	ENSG00000124140	1	0	548	81.204	-1	0.311010216
ENSG00000127412	0	ENSG00000165125	2	1.25E-148	193	84.974	2	0.264746228

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000138083	2	ENSG00000184302	3	2.24E-131	210	86.667	1	0.63253012
ENSG00000136379	2	ENSG00000129968	1	2.42E-152	240	82.917	1	0.729483283
ENSG00000134853	2	ENSG00000113721	1	0	152	81.579	1	0.139577594
ENSG00000243709	2	ENSG00000143768	3	0	399	94.486	-1	0.72810219
ENSG00000144119	2	ENSG00000165985	2	1.20E-92	132	88.636	-1	0.459930314
ENSG00000103740	3	ENSG00000166411	0	0	870	100	1	0.410571024
ENSG00000133026	2	ENSG00000133392	3	0	981	81.957	1	0.457983193
ENSG00000135100	2	ENSG00000157895	10	0	474	100	-3	0.712781955
ENSG00000107341	2	ENSG00000099804	10	5.82E-120	196	86.735	-1	0.307692308
ENSG00000047662	2	ENSG00000118579	12	0	1107	100	3	0.862821512
ENSG00000141965	2	ENSG00000145780	2	0	115	83.478	1	0.090125392
ENSG00000114166	2	ENSG00000108773	3	0	252	81.349	-2	0.167776298
ENSG00000140521	0	ENSG00000140525	0	2.30E-83	155	94.194	-1	0.10326449
ENSG00000109220	1	ENSG00000204116	13	3.18E-73	148	83.108	2	0.407713499
ENSG00000139266	3	ENSG00000144583	2	9.52E-99	159	90.566	-2	0.240181269
ENSG00000143862	1	ENSG00000134108	3	1.64E-115	183	91.257	1	0.62244898
ENSG00000136842	1	ENSG00000136925	2	0	555	100	3	0.940677966
ENSG00000089558	2	ENSG00000183960	1	0	224	81.696	2	0.220255654
ENSG00000064726	3	ENSG00000133243	2	0	379	84.697	-2	0.520604396
ENSG00000137880	1	ENSG00000104129	2	6.83E-129	161	100	-3	0.735159817

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000135945	0	ENSG00000158417	2	3.60E-180	284	100	-3	0.179633144
ENSG00000081019	1	ENSG00000187257	2	0	363	82.369	-1	0.288324067
ENSG00000126814	0	ENSG00000198830	1	1.41E-156	183	85.246	-1	0.117760618
ENSG00000113721	3	ENSG00000134853	2	0	154	80.519	-1	0.08045977
ENSG00000116254	3	ENSG00000111642	2	0	349	86.533	1	0.178607984
ENSG00000181381	12	ENSG00000137628	9	0	150	80.667	-1	0.087924971
ENSG00000084731	1	ENSG00000101350	1	0	115	82.609	1	0.078231293
ENSG00000139725	2	ENSG00000188735	1	0	581	100	3	0.642699115
ENSG00000087302	1	ENSG00000087303	0	0	236	98.729	3	0.967213115
ENSG00000138622	2	ENSG00000099822	2	0	562	88.612	1	0.46716542
ENSG00000135018	1	ENSG00000188021	2	0	163	90.184	1	0.125868726
ENSG00000113712	1	ENSG00000180138	0	0	335	81.194	2	0.485507246
ENSG00000141429	2	ENSG00000144278	2	0	505	85.941	-1	0.811897106
ENSG00000109787	3	ENSG00000118922	2	1.92E-59	91	92.308	-2	0.095088819
ENSG00000138685	4	ENSG00000170917	3	5.26E-138	196	98.98	-3	0.680555556
ENSG00000136682	2	ENSG00000172785	13	0	597	98.66	-1	0.981907895
ENSG00000136682	2	ENSG00000196873	9	0	597	97.99	-1	0.981907895
ENSG00000117971	2	ENSG00000160716	9	0	283	80.212	-1	0.503558719
ENSG00000141232	1	ENSG00000183864	10	1.74E-82	121	80.165	2	0.196110211
ENSG00000135175	3	ENSG00000122543	10	2.30E-153	231	98.268	-2	0.995689655

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000104903	5	ENSG00000162367	10	9.37E-31	62	85.484	-1	0.123260437
ENSG00000118620	2	ENSG00000197020	9	0	591	80.88	-1	0.957860616
ENSG00000137801	2	ENSG00000186340	1	0	261	83.142	2	0.157990315
ENSG00000143786	1	ENSG00000100528	0	4.79E-58	51	82.353	2	0.077625571
ENSG00000139112	4	ENSG00000170296	2	1.39E-68	116	87.069	1	0.184713376
ENSG00000160602	2	ENSG00000119638	10	5.74E-156	71	87.324	1	0.097260274
ENSG00000143933	2	ENSG00000160014	10	4.47E-98	153	99.346	-1	0.394329897
ENSG00000099308	3	ENSG00000086015	3	0	359	83.844	1	0.274255157
ENSG00000099308	3	ENSG00000105613	3	0	202	83.168	1	0.154316272
ENSG00000100764	0	ENSG00000119720	1	0	567	99.471	-2	0.786407767
ENSG00000102882	3	ENSG00000100030	2	0	345	88.116	-1	0.571192053
ENSG00000109158	3	ENSG00000145863	2	0	334	83.533	-1	0.268273092
ENSG00000103061	2	ENSG00000103064	4	0	977	100	-1	0.656586022
ENSG00000105464	3	ENSG00000161509	6	0	418	83.493	-1	0.312874251
ENSG00000136698	5	ENSG00000152093	2	0	250	99.6	-2	0.796178344
ENSG00000116489	3	ENSG00000198898	2	1.19E-164	288	86.806	1	0.362720403
ENSG00000116489	3	ENSG00000007341	2	2.66E-50	82	100	-2	0.103274559
ENSG00000165055	2	ENSG00000087995	3	0	458	95.633	2	0.629120879
ENSG00000112246	2	ENSG00000159263	3	0	359	86.072	-1	0.468668407
ENSG00000102753	0	ENSG00000186432	3	0	523	85.851	-1	0.351006711

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000139946	2	ENSG00000197329	1	0	411	81.752	1	0.68159204
ENSG00000139278	4	ENSG00000111615	9	0	379	99.736	-3	0.287338893
ENSG00000151615	2	ENSG00000152192	12	8.77E-105	177	85.876	-2	0.220423412
ENSG00000221866	3	ENSG00000076356	0	0	726	81.818	1	0.383315734
ENSG00000163755	1	ENSG00000047457	3	0	411	100	2	0.320843091
ENSG00000065665	0	ENSG00000058262	4	0	477	93.501	-1	0.644594595
ENSG00000008226	4	ENSG00000060971	1	1.28E-87	143	97.902	1	0.07788671
ENSG00000155760	2	ENSG00000180340	7	0	368	82.88	1	0.360078278
ENSG00000155760	2	ENSG00000157240	2	0	134	85.075	1	0.13111546
ENSG00000163286	0	ENSG00000163283	11	0	488	97.746	1	0.587951807
ENSG00000163286	0	ENSG00000163295	6	0	516	85.659	-2	0.621686747
ENSG00000154174	1	ENSG00000206535	1	2.65E-41	68	100	-1	0.111842105
ENSG00000156103	3	ENSG00000125966	11	0	176	85.227	1	0.26707132
ENSG00000120438	0	ENSG00000120437	2	1.15E-86	136	98.529	-1	0.212832551
ENSG00000181826	3	ENSG00000154274	6	2.82E-82	142	99.296	-2	0.31277533
ENSG00000123427	4	ENSG00000037897	7	4.94E-40	64	100	3	0.283185841
ENSG00000128394	9	ENSG00000243811	2	0	360	83.889	2	0.4400978
ENSG00000128394	9	ENSG00000244509	7	0	155	81.29	-2	0.189486553
ENSG00000160882	12	ENSG00000179142	0	0	407	89.681	-2	0.534822602
ENSG00000164933	0	ENSG00000164934	0	2.34E-49	108	99.074	2	0.114164905

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000152977	2	ENSG00000156925	1	5.18E-136	186	89.785	1	0.368316832
ENSG00000148943	1	ENSG00000104863	1	1.89E-100	204	82.353	2	0.129441624
ENSG00000153779	10	ENSG00000176679	2	1.50E-124	197	92.893	1	0.740601504
ENSG00000147274	6	ENSG00000213516	9	0	95	83.158	2	0.140740741
ENSG00000173451	2	ENSG00000133858	3	0	320	99.688	-3	0.737327189
ENSG00000043039	3	ENSG00000131668	2	9.65E-46	72	80.556	1	0.230031949
ENSG00000145780	2	ENSG00000141965	2	0	115	83.478	-1	0.086466165
ENSG00000181541	2	ENSG00000180660	0	0	386	92.746	-1	0.507894737
ENSG00000181789	0	ENSG00000158623	1	0	596	83.725	-1	0.584887144
ENSG00000173898	2	ENSG00000115306	2	0	592	81.588	2	0.247698745
ENSG00000180596	10	ENSG00000158373	2	6.36E-57	84	94.048	-1	0.371681416
ENSG00000145736	0	ENSG00000183474	1	0	571	98.949	1	0.875766871
ENSG00000112210	1	ENSG00000112208	2	0	341	100	-1	0.429471033
ENSG00000148377	0	ENSG00000107937	0	0	311	100	1	0.673160173
ENSG00000075886	1	ENSG00000152086	1	0	330	97.576	1	0.647058824
ENSG00000198034	0	ENSG00000129824	0	1.37E-171	265	92.83	1	0.880398671
ENSG00000177971	0	ENSG00000173548	2	1.11E-155	128	100	1	0.329048843
ENSG00000167191	2	ENSG00000174628	0	0	592	100	-1	0.622502629
ENSG00000145863	2	ENSG00000109158	1	0	334	83.533	1	0.651072125
ENSG00000005022	3	ENSG00000151729	10	1.67E-176	299	88.963	-2	0.717026379

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000160014	2	ENSG00000198668	2	6.41E-99	173	90.751	1	0.237964237
ENSG00000160014	2	ENSG00000143933	2	8.60E-98	153	99.346	1	0.21045392
ENSG00000167553	3	ENSG00000167552	1	0	463	88.553	1	0.860594796
ENSG00000122194	2	ENSG00000198670	13	0	250	85.6	-1	0.274725275
ENSG00000176679	10	ENSG00000153779	9	9.45E-127	197	92.893	-1	0.841880342
ENSG00000148672	0	ENSG00000182890	9	4.75E-68	99	81.818	-2	0.096303502
ENSG00000188612	1	ENSG00000177688	7	4.39E-127	73	87.671	-1	0.148373984
ENSG00000272617	0	ENSG00000258429	6	2.75E-126	189	98.413	2	0.288109756
ENSG00000213516	6	ENSG00000147274	2	0	91	89.011	-2	0.074225122
ENSG00000172345	2	ENSG00000172349	0	0	1381	100	-1	0.852469136
ENSG00000155428	14	ENSG00000178809	1	0	389	99.743	-1	0.874157303
ENSG00000174233	3	ENSG00000173175	7	0	359	82.73	-1	0.307363014
ENSG00000173404	2	ENSG00000168348	2	9.62E-76	93	81.72	1	0.182352941
ENSG00000156925	3	ENSG00000152977	1	7.45E-158	185	89.73	-1	0.362035225
ENSG00000166800	12	ENSG00000171989	1	0	332	82.229	-1	0.929971989
ENSG00000104043	2	ENSG00000143515	3	0	189	81.481	-1	0.158557047
ENSG00000198077	7	ENSG00000255974	0	0	490	94.082	-2	0.914179104
ENSG00000198077	7	ENSG00000197838	1	0	489	91.207	-2	0.912313433
ENSG00000162972	1	ENSG00000162971	1	4.11E-49	80	100	1	0.274914089
ENSG00000170549	2	ENSG00000177508	1	2.39E-67	96	80.208	1	0.2

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000171132	2	ENSG00000027075	3	0	67	88.06	1	0.068859198
ENSG00000159251	2	ENSG00000143632	13	0	378	98.942	-1	0.736842105
ENSG00000159251	2	ENSG00000107796	1	0	388	96.907	-2	0.756335283
ENSG00000186832	10	ENSG00000186847	0	0	312	88.141	-2	0.55026455
ENSG00000128789	0	ENSG00000101624	3	1.15E-35	85	92.941	-1	0.240112994
ENSG00000164900	1	ENSG00000168505	9	2.45E-69	108	87.037	2	0.297520661
ENSG00000170296	6	ENSG00000139112	8	1.00E-66	116	87.069	-1	0.489451477
ENSG00000128422	7	ENSG00000186847	8	0	320	88.75	-2	0.622568093
ENSG00000152270	2	ENSG00000172572	1	0	100	84	-1	0.072674419
ENSG00000140632	1	ENSG00000163735	3	6.12E-114	107	81.308	2	0.085805934
ENSG00000164934	0	ENSG00000164933	3	2.13E-49	108	99.074	-3	0.166153846
ENSG00000177879	3	ENSG00000157823	6	9.06E-109	192	84.375	-2	0.449648712
ENSG00000134184	11	ENSG00000213366	1	0	183	92.896	2	0.4575
ENSG00000146083	3	ENSG00000137075	1	4.86E-147	118	84.746	2	0.085198556
ENSG00000163295	2	ENSG00000163286	2	0	508	87.992	-1	0.674634794
ENSG00000163295	2	ENSG00000163283	2	0	313	88.818	2	0.415670651
ENSG00000173908	7	ENSG00000171446	12	0	331	85.196	2	0.713362069
ENSG00000173908	7	ENSG00000204897	10	0	331	84.894	1	0.713362069
ENSG00000156269	10	ENSG00000102030	10	1.65E-115	172	91.86	-1	0.751091703
ENSG00000167977	3	ENSG00000180901	3	3.13E-106	162	83.951	-2	0.2

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000180016	10	ENSG00000127780	0	1.80E-162	188	92.021	-1	0.598726115
ENSG00000171234	13	ENSG00000109181	3	0	558	87.455	-1	0.948979592
ENSG00000180233	1	ENSG00000186187	1	1.01E-48	87	80.46	-1	0.359504132
ENSG00000146049	14	ENSG00000146038	4	6.53E-124	199	99.497	1	0.432608696
ENSG00000163064	2	ENSG00000164778	2	1.34E-65	125	84.8	-1	0.318877551
ENSG00000170442	12	ENSG00000205426	2	0	419	99.045	-1	0.807321773
ENSG00000164736	2	ENSG00000171056	2	2.55E-42	71	90.141	1	0.171497585
ENSG00000164736	2	ENSG00000203883	2	7.64E-42	71	90.141	2	0.171497585
ENSG00000180818	2	ENSG00000253293	1	1.03E-20	55	80	-1	0.082956259
ENSG00000152932	2	ENSG00000105649	2	3.65E-118	194	88.66	1	0.565597668
ENSG00000181693	9	ENSG00000181767	0	0	345	85.797	1	0.997109827
ENSG00000154767	0	ENSG00000170860	1	1.80E-58	92	100	3	0.075471698
ENSG00000151729	2	ENSG00000005022	2	6.34E-176	299	88.963	2	0.681093394
ENSG00000167552	2	ENSG00000167553	2	0	454	89.427	-1	0.799295775
ENSG00000167785	10	ENSG00000130544	11	0	67	86.567	1	0.069791667
ENSG00000267631	1	ENSG00000104818	3	1.67E-112	169	98.817	1	0.645038168
ENSG00000142789	2	ENSG00000219073	2	2.07E-180	295	93.898	2	0.951612903
ENSG00000087995	2	ENSG00000165055	3	0	452	96.239	-2	0.852830189
ENSG00000166794	2	ENSG00000157734	4	5.17E-92	111	100	2	0.327433628
ENSG00000171103	0	ENSG00000163806	9	2.43E-96	133	99.248	-2	0.215909091

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000165516	2	ENSG00000165525	9	0	550	99.818	-1	0.936967632
ENSG00000256713	4	ENSG00000229183	2	0	506	99.407	-1	0.502482622
ENSG00000256713	4	ENSG00000229859	2	0	506	99.012	-1	0.502482622
ENSG00000172058	5	ENSG00000205572	1	5.94E-136	211	100	-1	0.338683788
ENSG00000156802	0	ENSG00000119778	2	0	335	81.791	2	0.216129032
ENSG00000198670	1	ENSG00000122194	2	0	311	80.064	1	0.15245098
ENSG00000159556	1	ENSG00000016082	3	2.71E-174	188	81.383	1	0.307189542
ENSG00000169618	4	ENSG00000169621	9	0	241	100	1	0.611675127
ENSG00000169618	4	ENSG00000101292	0	1.74E-152	241	84.232	1	0.611675127
ENSG00000189306	0	ENSG00000183569	2	3.19E-78	128	95.312	1	0.101265823
ENSG00000173020	4	ENSG00000100077	0	0	685	84.088	-2	0.597731239
ENSG00000165462	1	ENSG00000109132	2	2.73E-59	99	84.848	-1	0.174911661
ENSG00000175077	7	ENSG00000198471	3	7.41E-100	169	89.349	1	0.222955145
ENSG00000149929	2	ENSG00000169592	3	7.39E-171	284	99.648	-3	0.35813367
ENSG00000244414	9	ENSG00000080910	2	0	172	98.256	2	0.390909091
ENSG00000244414	9	ENSG00000000971	5	1.33E-169	217	94.47	-1	0.493181818
ENSG00000168505	1	ENSG00000164900	0	1.17E-69	100	90	-2	0.236406619
ENSG00000172519	5	ENSG00000186723	1	0	318	92.767	-1	0.843501326
ENSG00000175344	2	ENSG00000166664	2	0	450	99.778	2	0.706436421
ENSG00000161509	2	ENSG00000105464	2	0	418	83.493	1	0.314049587

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000057149	5	ENSG00000206073	2	0	383	91.906	-1	0.639398998
ENSG00000065978	2	ENSG00000060138	1	6.48E-56	91	94.505	-1	0.170731707
ENSG00000179981	3	ENSG00000121297	6	0	148	85.135	-1	0.143410853
ENSG00000164588	3	ENSG00000099822	1	0	545	85.138	1	0.612359551
ENSG00000159263	3	ENSG00000112246	1	0	359	86.072	-1	0.473614776
ENSG00000170950	0	ENSG00000102144	4	0	405	87.16	2	0.72450805
ENSG00000175564	6	ENSG00000175567	2	6.82E-139	58	81.034	-1	0.087613293
ENSG00000092871	1	ENSG00000005156	1	0	867	100	2	0.640798226
ENSG00000172023	12	ENSG00000115386	2	1.41E-101	217	81.106	-1	0.844357977
ENSG00000166363	6	ENSG00000170790	12	0	342	92.105	-1	0.974358974
ENSG00000170790	6	ENSG00000166363	12	0	338	92.012	1	0.962962963
ENSG00000169710	1	ENSG00000122224	6	2.95E-93	77	80.519	-2	0.027266289
ENSG00000169469	9	ENSG00000169474	2	6.62E-45	57	92.982	1	0.640449438
ENSG00000157227	2	ENSG00000125966	2	0	112	81.25	1	0.115463918
ENSG00000166823	1	ENSG00000188095	4	2.52E-62	102	85.294	-1	0.259541985
ENSG00000187048	12	ENSG00000162365	4	0	479	95.407	-1	0.513948498
ENSG00000258429	1	ENSG00000272617	0	6.82E-127	189	98.413	-2	0.490909091
ENSG00000168872	2	ENSG00000157349	3	0	489	97.751	-1	0.50308642
ENSG00000153922	2	ENSG00000173575	4	0	499	85.972	-1	0.25949038
ENSG00000170484	7	ENSG00000139648	9	0	361	86.981	1	0.68241966

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000170484	7	ENSG00000186049	9	0	361	86.427	2	0.68241966
ENSG00000163464	3	ENSG00000180871	2	1.99E-153	292	84.589	-1	0.432592593
ENSG00000179142	12	ENSG00000160882	0	0	416	92.788	-1	0.827037773
ENSG00000157240	3	ENSG00000155760	2	0	133	85.714	-1	0.155192532
ENSG00000157240	3	ENSG00000180340	0	0	309	85.113	1	0.360560093
ENSG00000106714	2	ENSG00000154529	1	0	1216	98.273	-2	0.930374904
ENSG00000170255	4	ENSG00000179817	1	8.12E-97	56	80.357	1	0.138271605
ENSG00000154342	3	ENSG00000108379	1	0	344	85.465	1	0.754385965
ENSG00000155918	8	ENSG00000131015	3	1.08E-141	235	88.936	2	0.951417004
ENSG00000178338	9	ENSG00000169131	1	0	439	87.927	-1	0.684867395
ENSG00000168671	2	ENSG00000145626	2	0	85	90.588	1	0.11659808
ENSG00000159217	6	ENSG00000136231	9	0	163	82.209	-1	0.282495667
ENSG00000184492	9	ENSG00000170122	1	0	358	94.413	-1	0.58496732
ENSG00000184492	9	ENSG00000187559	3	0	273	95.971	-2	0.446078431
ENSG00000168928	1	ENSG00000168925	2	5.29E-158	154	96.104	-2	0.48125
ENSG00000170262	10	ENSG00000142207	0	3.86E-139	203	100	-1	0.636363636
ENSG00000162924	2	ENSG00000196911	2	0	395	93.418	1	0.49375
ENSG00000048828	2	ENSG00000184083	3	0	220	80	-2	0.13986014
ENSG00000166947	5	ENSG00000166946	2	0	361	100	-3	0.494520548
ENSG00000151693	2	ENSG00000119185	2	0	736	100	2	0.398268398

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000157895	2	ENSG00000135100	4	0	477	100	3	0.7453125
ENSG00000146281	3	ENSG00000166200	4	7.57E-86	132	82.576	2	0.083756345
ENSG00000164729	0	ENSG00000177710	3	0	380	93.684	-1	0.607028754
ENSG00000158941	1	ENSG00000147439	4	0	356	100	1	0.385698808
ENSG00000168961	10	ENSG00000171916	3	0	322	94.72	-1	0.575
ENSG00000176887	2	ENSG00000124766	2	6.21E-48	86	81.395	1	0.195011338
ENSG00000176887	2	ENSG00000177732	2	2.95E-51	74	83.784	-1	0.167800454
ENSG00000169385	9	ENSG00000169397	1	1.10E-102	75	88	1	0.330396476
ENSG00000154274	6	ENSG00000181826	1	4.31E-82	142	99.296	2	0.220496894
ENSG00000175548	0	ENSG00000139133	2	0	131	89.313	-1	0.247169811
ENSG00000010292	0	ENSG00000111639	2	3.36E-41	73	100	-1	0.048537234
ENSG00000158417	0	ENSG00000135945	2	4.40E-180	284	100	1	0.205350687
ENSG00000175567	2	ENSG00000175564	7	5.42E-139	58	81.034	1	0.106032907
ENSG00000145626	12	ENSG00000168671	2	0	389	80.206	-1	0.423286181
ENSG00000186049	7	ENSG00000139648	2	0	365	89.315	-1	0.675925926
ENSG00000186049	7	ENSG00000170484	2	0	364	86.538	-2	0.674074074
ENSG00000170577	2	ENSG00000126778	12	2.17E-121	174	95.977	-2	0.245070423
ENSG00000182793	1	ENSG00000243955	0	9.08E-143	253	84.585	-1	0.900355872
ENSG00000171295	14	ENSG00000197054	2	0	300	83.333	-1	0.223380491
ENSG00000171295	14	ENSG00000171291	2	0	353	82.153	1	0.262844378

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000171291	14	ENSG00000171295	3	0	353	82.153	-1	0.436341162
ENSG00000171291	14	ENSG00000196757	14	0	70	81.429	1	0.086526576
ENSG00000169397	9	ENSG00000169385	3	5.91E-111	77	89.61	-1	0.340707965
ENSG00000171056	3	ENSG00000164736	4	9.31E-53	89	85.393	-1	0.118508655
ENSG00000166503	3	ENSG00000136404	3	0	312	99.359	1	0.472727273
ENSG00000169032	1	ENSG00000126934	3	0	225	92.444	2	0.426944972
ENSG00000163322	2	ENSG00000163319	2	0	394	99.746	-3	0.963325183
ENSG00000154025	2	ENSG00000154016	3	0	494	99.595	-3	0.665768194
ENSG00000129968	1	ENSG00000136379	6	1.27E-140	249	80.723	-1	0.689750693
ENSG00000173175	1	ENSG00000174233	9	0	317	83.281	2	0.251387787
ENSG00000196778	8	ENSG00000181963	0	2.47E-172	324	83.951	1	0.944606414
ENSG00000164855	1	ENSG00000198517	3	4.09E-178	269	100	3	0.388167388
ENSG00000123064	0	ENSG00000186710	3	2.17E-94	125	100	3	0.141723356
ENSG00000144583	2	ENSG00000139266	3	2.45E-122	185	86.486	2	0.190525232
ENSG00000173349	0	ENSG00000136709	5	0	692	100	-2	0.701825558
ENSG00000128655	6	ENSG00000284741	0	0	578	100	-1	0.489830508
ENSG00000160339	12	ENSG00000085265	12	1.03E-162	219	84.018	1	0.655688623
ENSG00000151665	0	ENSG00000119729	0	8.88E-144	219	100	2	0.610027855
ENSG00000162391	1	ENSG00000162390	1	9.79E-180	255	99.608	3	0.381736527
ENSG00000171201	13	ENSG00000109208	3	4.14E-89	71	91.549	1	0.290983607

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000140525	1	ENSG00000140521	3	3.89E-81	153	94.118	3	0.108587651
ENSG00000126785	2	ENSG00000119729	3	4.69E-100	190	80	1	0.295489891
ENSG00000163623	2	ENSG00000148826	1	6.92E-86	113	83.186	2	0.307901907
ENSG00000165646	2	ENSG00000165650	4	0	718	100	-3	0.556589147
ENSG00000172789	3	ENSG00000106004	4	7.75E-38	54	88.889	1	0.243243243
ENSG00000149289	4	ENSG00000102053	2	0	258	83.333	1	0.29218573
ENSG00000149289	4	ENSG00000163874	8	1.17E-138	209	82.775	2	0.236693092
ENSG00000174628	2	ENSG00000167191	5	0	592	100	1	0.696470588
ENSG00000108242	11	ENSG00000138109	2	0	484	81.818	-1	0.806666667
ENSG00000152093	5	ENSG00000136698	5	0	250	99.6	2	0.796178344
ENSG00000167721	0	ENSG00000167720	2	0	243	99.588	3	0.172953737
ENSG00000173548	3	ENSG00000177971	0	4.63E-155	128	100	-1	0.118190212
ENSG00000171989	11	ENSG00000166800	2	0	341	82.698	-1	0.603539823
ENSG00000146707	2	ENSG00000188372	2	2.13E-126	111	98.198	-2	0.123333333
ENSG00000172476	12	ENSG00000102128	1	0	242	98.347	2	0.733333333
ENSG00000165985	3	ENSG00000131094	1	7.50E-91	141	86.525	1	0.505376344
ENSG00000165985	3	ENSG00000144119	5	1.58E-86	132	88.636	1	0.47311828
ENSG00000173480	12	ENSG00000198466	2	0	724	95.58	2	0.601328904
ENSG00000166439	1	ENSG00000166435	2	0	727	100	1	0.846332945
ENSG00000177551	7	ENSG00000171786	2	1.87E-16	55	98.182	2	0.157142857

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000167395	5	ENSG00000151006	2	1.44E-89	128	100	-1	0.061865636
ENSG00000270316	16	ENSG00000214435	2	0	753	99.602	1	0.867511521
ENSG00000270316	16	ENSG00000166275	2	3.34E-93	115	99.13	1	0.132488479
ENSG00000168569	1	ENSG00000185475	2	1.84E-85	145	100	1	0.622317597
ENSG00000163737	12	ENSG00000109272	2	2.59E-60	86	89.535	1	0.761061947
ENSG00000111196	1	ENSG00000162385	10	1.95E-96	148	98.649	-1	0.573643411
ENSG00000163735	12	ENSG00000140632	2	4.35E-99	102	86.275	1	0.204
ENSG00000136709	0	ENSG00000173349	9	0	692	100	2	0.517964072
ENSG00000166946	2	ENSG00000166947	3	0	361	100	3	0.703703704
ENSG00000174125	9	ENSG00000174130	2	0	326	90.184	2	0.349785408
ENSG00000144199	7	ENSG00000115042	1	0	321	98.131	-2	0.75
ENSG00000152086	1	ENSG00000075886	6	0	288	96.181	-1	0.555984556
ENSG00000170832	2	ENSG00000129204	3	0	784	92.73	-1	0.488778055
ENSG00000157326	6	ENSG00000187630	2	0	235	91.064	-2	0.560859189
ENSG00000186787	3	ENSG00000147059	1	0	350	96.857	-1	0.868486352
ENSG00000248871	6	ENSG00000161955	0	0	431	98.144	-1	0.833655706
ENSG00000162076	7	ENSG00000059122	14	3.43E-34	67	85.075	-1	0.224080268
ENSG00000174990	7	ENSG00000169239	2	9.83E-34	50	80	3	0.145772595
ENSG00000165584	9	ENSG00000241476	2	0	50	96	-1	0.181818182
ENSG00000180340	2	ENSG00000155760	0	0	387	80.362	-1	0.612341772

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000180340	2	ENSG00000157240	8	0	309	85.113	-1	0.488924051
ENSG00000171446	7	ENSG00000204897	3	0	269	91.45	-1	0.586056645
ENSG00000171446	7	ENSG00000173908	3	0	269	85.13	-2	0.586056645
ENSG00000169750	1	ENSG00000128340	3	5.00E-116	200	86.5	2	0.552486188
ENSG00000161202	2	ENSG00000004975	10	0	131	91.603	-1	0.162732919
ENSG00000161202	2	ENSG00000107404	0	0	132	82.576	-2	0.163975155
ENSG00000253293	4	ENSG00000180818	2	1.01E-46	78	93.59	2	0.089449541
ENSG00000173273	2	ENSG00000107854	3	0	615	85.041	-1	0.441176471
ENSG00000167702	0	ENSG00000160973	2	2.34E-166	263	100	-2	0.238440617
ENSG00000128886	7	ENSG00000167004	3	2.38E-124	210	100	3	0.358361775
ENSG00000153976	9	ENSG00000125430	3	2.56E-180	273	95.238	1	0.579617834
ENSG00000170860	1	ENSG00000154767	6	1.79E-58	92	100	-3	0.486772487
ENSG00000160683	3	ENSG00000186174	2	0	882	100	-3	0.911157025
ENSG00000172939	4	ENSG00000198648	4	0	338	83.728	1	0.224286662
ENSG00000255974	7	ENSG00000198077	2	0	535	91.402	2	0.978062157
ENSG00000255974	7	ENSG00000197838	3	0	323	89.783	-2	0.590493601
ENSG00000169840	1	ENSG00000180613	2	1.11E-43	71	88.732	2	0.268939394
ENSG00000204897	7	ENSG00000171446	2	0	269	91.45	1	0.597777778
ENSG00000204897	7	ENSG00000173908	7	0	269	85.13	-1	0.597777778
ENSG00000141084	3	ENSG00000010017	0	3.07E-54	68	89.706	-1	0.038812785

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000133392	2	ENSG00000133026	0	0	896	82.478	-1	0.452753916
ENSG00000166351	12	ENSG00000183206	11	0	536	96.455	-1	0.881578947
ENSG00000163254	7	ENSG00000182187	10	1.16E-98	83	95.181	-2	0.417085427
ENSG00000168131	7	ENSG00000124657	1	7.12E-171	310	81.613	-2	0.767326733
ENSG00000156052	3	ENSG00000088256	6	0	315	91.111	1	0.432098765
ENSG00000157827	2	ENSG00000161791	2	0	130	86.154	-1	0.072747622
ENSG00000169239	2	ENSG00000174990	2	1.34E-63	90	82.222	3	0.283911672
ENSG00000170917	1	ENSG00000138685	3	8.38E-139	196	98.98	3	0.5
ENSG00000166377	2	ENSG00000054793	10	0	570	82.281	-1	0.484282073
ENSG00000146411	3	ENSG00000028839	10	0	463	100	2	0.624831309
ENSG00000148136	5	ENSG00000204246	1	5.88E-176	317	83.912	1	0.996855346
ENSG00000165525	0	ENSG00000165516	2	0	550	99.818	1	0.511152416
ENSG00000168298	3	ENSG00000187837	2	3.82E-48	71	100	-1	0.288617886
ENSG00000168298	3	ENSG00000184357	2	3.84E-35	71	95.775	-1	0.288617886
ENSG00000153443	2	ENSG00000185262	3	3.86E-33	89	91.011	1	0.18697479
ENSG00000149968	8	ENSG00000166670	12	0	217	85.714	-1	0.454926625
ENSG00000171478	1	ENSG00000171489	2	5.78E-119	170	100	2	0.696721311
ENSG00000152518	2	ENSG00000185650	3	2.55E-55	74	90.541	1	0.116719243
ENSG00000168906	0	ENSG00000115486	2	1.30E-126	185	100	2	0.197018104
ENSG00000181767	9	ENSG00000181693	2	0	351	85.755	1	1

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000151006	5	ENSG00000167395	2	4.21E-90	128	100	1	0.231464738
ENSG00000158714	11	ENSG00000188004	12	0	327	99.388	2	0.862796834
ENSG00000180871	3	ENSG00000163464	2	6.51E-146	305	82.623	1	0.3125
ENSG00000206172	3	ENSG00000188536	2	4.91E-112	178	95.506	-1	0.843601896
ENSG00000125375	1	ENSG00000100490	2	0	260	100	2	0.461811723
ENSG00000171786	4	ENSG00000177551	2	2.07E-13	51	100	-1	0.060570071
ENSG00000166670	9	ENSG00000149968	3	0	390	80.256	1	0.663265306
ENSG00000242019	6	ENSG00000240403	6	0	123	82.114	-2	0.280821918
ENSG00000168582	8	ENSG00000182187	7	4.16E-95	86	80.233	-2	0.387387387
ENSG00000166664	2	ENSG00000175344	12	0	463	99.784	-2	0.726844584
ENSG00000147439	0	ENSG00000158941	2	0	360	100	2	0.601001669
ENSG00000163888	3	ENSG00000145194	1	2.48E-84	124	100	-2	0.984126984
ENSG00000168348	2	ENSG00000173404	2	2.21E-82	96	82.292	-1	0.157635468
ENSG00000180901	2	ENSG00000167977	1	6.85E-96	162	83.951	2	0.615969582
ENSG00000164778	2	ENSG00000163064	1	1.42E-52	85	85.882	1	0.196759259
ENSG00000168916	1	ENSG00000180357	2	3.90E-176	96	82.292	1	0.050632911
ENSG00000170604	3	ENSG00000119669	2	9.19E-139	87	88.506	1	0.099315068
ENSG00000170604	3	ENSG00000168264	12	2.12E-81	83	87.952	1	0.094748858
ENSG00000163319	8	ENSG00000163322	2	0	394	99.746	3	0.530282638
ENSG00000184227	9	ENSG00000119673	2	0	475	96	-1	0.83041958

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000177839	9	ENSG00000120324	1	0	669	92.377	2	0.691830403
ENSG00000177839	9	ENSG00000113212	2	1.19E-127	65	87.692	1	0.067218201
ENSG00000179817	4	ENSG00000170255	12	3.59E-151	196	81.122	1	0.402464066
ENSG00000171489	1	ENSG00000171478	15	6.43E-119	170	100	-2	0.696721311
ENSG00000151746	1	ENSG00000185963	7	0	201	85.572	1	0.189443921
ENSG00000213512	13	ENSG00000162654	2	0	482	82.365	1	0.755485893
ENSG00000167004	2	ENSG00000128886	2	3.86E-124	210	100	-3	0.303907381
ENSG00000170027	2	ENSG00000128245	0	4.92E-135	235	87.66	-1	0.18875502
ENSG00000153147	2	ENSG00000102038	0	0	750	86.4	1	0.622406639
ENSG00000169629	2	ENSG00000183054	2	0	802	99.751	2	0.454390935
ENSG00000169629	2	ENSG00000015568	2	0	946	99.683	-1	0.535977337
ENSG00000175711	1	ENSG00000141556	8	2.16E-58	102	89.216	-1	0.127659574
ENSG00000154016	3	ENSG00000154025	2	0	494	99.595	3	0.734026746
ENSG00000169592	1	ENSG00000149929	3	2.74E-144	226	99.558	2	0.583979328
ENSG00000105618	0	ENSG00000105619	2	1.44E-43	73	100	3	0.118699187
ENSG00000060971	0	ENSG00000008226	12	2.85E-88	143	97.902	-2	0.2495637
ENSG00000166411	0	ENSG00000103740	12	0	869	100	-1	0.960220994
ENSG00000166862	2	ENSG00000006116	1	9.66E-130	167	86.228	1	0.442970822
ENSG00000242220	1	ENSG00000166984	2	5.62E-132	169	80.473	2	0.342105263
ENSG00000171017	2	ENSG00000171488	2	0	90	84.444	-2	0.074812968

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000178934	6	ENSG00000205076	3	1.69E-103	155	100	-1	0.890804598
ENSG00000179361	2	ENSG00000116017	3	8.92E-97	135	84.444	2	0.096153846
ENSG00000160868	2	ENSG00000160870	0	0	505	88.515	-1	0.838870432
ENSG00000189132	13	ENSG00000185448	2	0	242	80.579	2	0.33988764
ENSG00000189132	13	ENSG00000198173	2	0	175	81.143	-1	0.245786517
ENSG00000027075	3	ENSG00000171132	2	0	67	88.06	-1	0.055417701
ENSG00000181625	0	ENSG00000132207	1	0	244	100	1	0.62086514
ENSG00000186564	1	ENSG00000187140	1	1.61E-62	104	91.346	-1	0.21010101
ENSG00000177732	6	ENSG00000176887	0	3.60E-56	78	87.179	1	0.080495356
ENSG00000187527	6	ENSG00000127249	0	0	108	80.556	1	0.088669951
ENSG00000102743	2	ENSG00000120329	1	0	315	86.667	-2	0.234549516
ENSG00000185127	2	ENSG00000130024	2	6.54E-39	64	96.875	2	0.045519203
ENSG00000177688	1	ENSG00000188612	0	2.90E-167	137	86.131	-1	0.541501976
ENSG00000157349	2	ENSG00000168872	1	0	475	96.842	-2	0.790349418
ENSG00000178802	0	ENSG00000178761	2	0	744	100	3	0.789808917
ENSG00000188735	1	ENSG00000139725	0	0	295	99.322	-3	0.790884718
ENSG00000186710	3	ENSG00000123064	2	8.45E-94	130	100	-3	0.422077922
ENSG00000100023	1	ENSG00000100027	9	0	569	100	-3	0.440402477
ENSG00000186187	3	ENSG00000180233	2	1.36E-48	87	80.46	1	0.207142857
ENSG00000104177	2	ENSG00000188467	1	0	222	97.297	-2	0.145956607

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000229859	4	ENSG00000256713	12	0	506	99.012	1	0.502482622
ENSG00000177182	3	ENSG00000198363	12	0	363	100	-3	0.306846999
ENSG00000213218	3	ENSG00000136487	9	0	293	83.618	-1	0.864306785
ENSG00000054793	3	ENSG00000166377	3	0	570	82.281	1	0.544412607
ENSG00000142453	2	ENSG00000130733	3	4.97E-138	245	100	2	0.402960526
ENSG00000161955	6	ENSG00000248871	3	0	431	98.144	1	0.883196721
ENSG00000206181	1	ENSG00000183791	3	0	269	82.156	-1	0.357237716
ENSG00000183281	2	ENSG00000125551	0	0	388	99.742	1	0.987277354
ENSG00000183281	2	ENSG00000122194	1	0	179	88.827	1	0.455470738
ENSG00000100027	2	ENSG00000100023	12	0	569	100	3	0.405559515
ENSG00000122565	1	ENSG00000108468	6	6.56E-60	68	85.294	-1	0.111292962
ENSG00000185650	3	ENSG00000152518	2	2.84E-54	82	87.805	-1	0.120234604
ENSG00000198517	2	ENSG00000164855	0	2.15E-178	269	100	-3	0.282266527
ENSG00000137267	1	ENSG00000137285	3	0	225	100	-2	0.415896488
ENSG00000114374	2	ENSG00000124486	1	0	473	88.795	2	0.298611111
ENSG00000156049	3	ENSG00000088256	3	0	352	82.386	1	0.540706605
ENSG00000172236	13	ENSG00000095917	2	0	294	85.374	1	0.731343284
ENSG00000197172	13	ENSG00000221867	9	0	466	96.996	-1	0.818980668
ENSG00000243772	9	ENSG00000240403	2	0	414	82.85	-1	0.756855576
ENSG00000243772	9	ENSG00000167633	12	0	377	86.472	-1	0.689213894

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000182315	1	ENSG00000237247	1	1.92E-177	257	98.444	2	0.992277992
ENSG00000127955	2	ENSG00000065135	2	0	362	92.541	1	0.322638146
ENSG00000070808	2	ENSG00000058404	2	0	476	85.714	1	0.297128589
ENSG00000105991	2	ENSG00000120094	5	1.33E-44	65	86.154	-2	0.077105575
ENSG00000215252	1	ENSG00000175265	1	0	258	100	-2	0.332474227
ENSG00000182931	1	ENSG00000180305	2	7.98E-59	58	81.034	-1	0.302083333
ENSG00000185236	1	ENSG00000103769	0	1.16E-103	155	95.484	-2	0.668103448
ENSG00000183303	6	ENSG00000182334	2	2.81E-118	113	83.186	1	0.31741573
ENSG00000104938	14	ENSG00000090659	2	5.27E-174	348	81.609	-1	0.541213064
ENSG00000197021	4	ENSG00000197620	5	0	355	97.183	-1	0.628318584
ENSG00000163806	2	ENSG00000171103	2	4.26E-99	132	100	-2	0.225641026
ENSG00000188467	1	ENSG00000104177	3	0	222	97.297	2	0.424474187
ENSG00000177508	2	ENSG00000170549	9	2.16E-63	88	85.227	-1	0.129032258
ENSG00000196735	11	ENSG00000237541	9	0	323	85.449	1	0.675732218
ENSG00000132356	3	ENSG00000162409	0	0	243	82.716	-2	0.143447462
ENSG00000131584	2	ENSG00000114331	1	1.81E-67	122	81.148	-2	0.097211155
ENSG00000205426	12	ENSG00000170442	0	0	455	96.484	1	0.784482759
ENSG00000105438	0	ENSG00000136240	10	1.53E-108	184	84.239	2	0.35047619
ENSG00000141556	1	ENSG00000175711	3	3.18E-45	106	84.906	3	0.080181543
ENSG00000187082	9	ENSG00000186579	3	3.44E-60	100	100	-1	0.952380952

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000166435	3	ENSG00000166439	7	0	727	100	-1	0.846332945
ENSG00000122824	12	ENSG00000196368	3	8.89E-136	177	96.045	1	0.272727273
ENSG00000161791	3	ENSG00000157827	3	0	391	82.353	1	0.104545455
ENSG00000089057	1	ENSG00000170482	3	0	92	82.609	2	0.141538462
ENSG00000187630	6	ENSG00000157326	3	0	235	91.064	2	0.547785548
ENSG00000186510	10	ENSG00000184908	2	0	687	91.266	-1	0.778911565
ENSG00000186599	10	ENSG00000186562	1	4.36E-74	110	100	-1	0.990990991
ENSG00000185448	13	ENSG00000189132	1	0	156	83.333	1	0.184834123
ENSG00000185448	13	ENSG00000198173	8	0	174	82.759	-1	0.206161137
ENSG00000183054	2	ENSG00000015568	2	0	946	100	-1	0.535977337
ENSG00000183054	2	ENSG00000169629	2	0	802	99.751	-2	0.454390935
ENSG00000119723	0	ENSG00000187097	0	1.22E-153	235	99.574	-3	0.451055662
ENSG00000134317	3	ENSG00000083307	1	0	53	92.453	-1	0.077259475
ENSG00000117713	3	ENSG00000049618	1	0	65	80	-2	0.028446389
ENSG00000187243	1	ENSG00000154545	1	0	612	100	1	0.714953271
ENSG00000249471	11	ENSG00000083812	10	0	333	93.093	-1	0.422053232
ENSG00000135111	2	ENSG00000121068	1	1.35E-168	239	87.448	1	0.204099061
ENSG00000115306	1	ENSG00000173898	12	0	428	87.383	1	0.181049069
ENSG00000171916	10	ENSG00000168961	9	0	322	94.72	1	0.564912281
ENSG00000198754	1	ENSG00000083720	2	0	218	81.193	2	0.356792144

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000179914	2	ENSG00000158764	9	0	283	88.693	-1	0.702233251
ENSG00000187372	10	ENSG00000120322	9	0	325	96.308	-2	0.352112676
ENSG00000187372	10	ENSG00000113205	1	0	298	90.94	-1	0.322860238
ENSG00000241595	11	ENSG00000187272	9	1.53E-75	151	83.444	1	0.629166667
ENSG00000162654	12	ENSG00000213512	9	0	482	82.365	-1	0.304677623
ENSG00000037897	0	ENSG00000123427	1	3.93E-40	64	100	-3	0.133611691
ENSG00000240654	2	ENSG00000205863	1	0	226	95.133	1	0.366883117
ENSG00000058404	3	ENSG00000070808	4	0	476	85.714	-1	0.714714715
ENSG00000186119	8	ENSG00000205029	4	3.26E-148	304	81.25	-2	0.921212121
ENSG00000188095	1	ENSG00000166823	1	6.42E-60	97	88.66	1	0.233173077
ENSG00000237247	1	ENSG00000182315	2	1.92E-177	257	98.444	-2	0.992277992
ENSG00000162971	2	ENSG00000162972	2	2.21E-47	83	97.59	-3	0.215025907
ENSG00000180613	1	ENSG00000169840	1	1.01E-45	78	82.051	-2	0.242990654
ENSG00000175868	3	ENSG00000110680	1	2.15E-72	76	89.474	-2	0.217765043
ENSG00000185262	3	ENSG00000153443	1	1.02E-47	103	83.495	-1	0.62804878
ENSG00000185100	0	ENSG00000035687	1	0	91	85.714	1	0.153456998
ENSG00000162390	2	ENSG00000162391	2	1.55E-179	255	99.608	-3	0.411290323
ENSG00000188375	1	ENSG00000132475	9	0	129	96.899	1	0.362359551
ENSG00000157823	3	ENSG00000177879	4	3.28E-108	192	84.375	2	0.223515716
ENSG00000198466	12	ENSG00000173480	10	0	745	91.812	1	0.856321839

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000158623	0	ENSG00000181789	6	0	585	83.761	1	0.671641791
ENSG00000206073	5	ENSG00000057149	12	0	583	91.767	1	0.989813243
ENSG00000054690	2	ENSG00000100564	2	1.96E-64	100	100	1	0.045977011
ENSG00000241119	1	ENSG00000242366	9	0	778	96.53	-1	0.964064436
ENSG00000241119	1	ENSG00000244122	1	0	777	95.495	-1	0.962825279
ENSG00000186579	9	ENSG00000187082	7	3.21E-60	100	100	1	0.952380952
ENSG00000196565	8	ENSG00000213934	7	1.30E-124	195	97.949	2	0.924170616
ENSG00000049618	2	ENSG00000117713	2	0	110	83.636	-1	0.048910627
ENSG00000240403	6	ENSG00000167633	2	0	56	80.357	1	0.112224449
ENSG00000140526	1	ENSG00000101558	3	2.03E-162	121	86.777	1	0.16005291
ENSG00000181396	3	ENSG00000169660	6	1.02E-18	55	100	-2	0.112016293
ENSG00000182968	1	ENSG00000134595	2	4.70E-59	69	86.957	2	0.176470588
ENSG00000182968	1	ENSG00000181449	1	1.00E-50	71	91.549	2	0.181585678
ENSG00000188021	9	ENSG00000135018	1	0	153	90.85	1	0.141535615
ENSG00000184357	1	ENSG00000168298	0	6.79E-29	71	95.775	1	0.269961977
ENSG00000186723	5	ENSG00000172519	2	0	325	91.077	1	0.868983957
ENSG00000181963	8	ENSG00000196778	3	1.17E-170	321	90.654	1	0.904225352
ENSG00000205420	9	ENSG00000185479	3	0	226	91.15	-2	0.289372599
ENSG00000185009	1	ENSG00000035403	6	0	369	100	3	0.29193038
ENSG00000116721	10	ENSG00000204481	0	0	486	96.091	2	0.888482633

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000116721	10	ENSG00000120952	2	0	547	94.333	-1	1
ENSG00000102053	3	ENSG00000149289	13	0	258	83.333	-1	0.185478073
ENSG00000177613	2	ENSG00000101811	13	0	208	87.5	1	0.262958281
ENSG00000182334	6	ENSG00000183303	14	4.39E-61	65	84.615	-2	0.209003215
ENSG00000178177	2	ENSG00000109805	14	0	474	100	-2	0.408973253
ENSG00000124486	2	ENSG00000114374	14	0	332	89.458	-2	0.129182879
ENSG00000186480	0	ENSG00000125629	14	1.29E-89	185	84.324	1	0.378323108
ENSG00000137075	2	ENSG00000146083	3	2.01E-136	118	84.746	-2	0.191558442
ENSG00000197838	7	ENSG00000255974	10	0	307	91.205	-1	0.62145749
ENSG00000197838	7	ENSG00000198077	7	0	489	91.207	2	0.989878543
ENSG00000182255	2	ENSG00000177272	7	6.96E-172	133	91.729	1	0.203675345
ENSG00000127249	2	ENSG00000187527	1	0	108	80.556	-1	0.090301003
ENSG00000241484	2	ENSG00000248405	3	0	269	100	2	0.579741379
ENSG00000186562	10	ENSG00000186599	1	4.36E-74	110	100	1	0.990990991
ENSG00000182890	0	ENSG00000148672	4	6.35E-48	99	81.818	2	0.126598465
ENSG00000099977	4	ENSG00000099974	10	1.61E-61	104	96.154	-1	0.525252525
ENSG00000184486	2	ENSG00000196767	11	2.28E-71	144	80.556	2	0.104423495
ENSG00000113296	2	ENSG00000105664	12	0	277	87.004	-1	0.247100803
ENSG00000160870	2	ENSG00000160868	5	0	506	88.538	1	0.892416226
ENSG00000186174	2	ENSG00000160683	1	0	882	100	3	0.588392262

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000183682	3	ENSG00000116985	13	0	325	95.692	2	0.808457711
ENSG00000181449	1	ENSG00000182968	0	4.13E-72	99	90.909	-2	0.253846154
ENSG00000176490	3	ENSG00000165023	2	2.83E-101	126	86.508	-1	0.111504425
ENSG00000196911	2	ENSG00000025800	2	0	308	92.532	-1	0.427184466
ENSG00000134183	2	ENSG00000114349	0	0	344	82.558	-1	0.764444444
ENSG00000108242	11	ENSG00000165841	1	0	231	81.818	1	0.385
ENSG00000187634	2	ENSG00000188976	12	4.90E-72	124	100	-2	0.173669468
ENSG00000136487	3	ENSG00000213218	5	1.88E-179	206	81.068	-2	0.727915194
ENSG00000076685	2	ENSG00000148842	3	0	528	100	1	0.471008029
ENSG00000100319	1	ENSG00000100314	1	1.42E-83	139	97.122	-1	0.463333333
ENSG00000099290	1	ENSG00000172661	2	0	958	98.225	2	0.714392245
ENSG00000169594	1	ENSG00000173068	3	0	80	87.5	-1	0.080482897
ENSG00000119711	1	ENSG00000119636	4	6.16E-45	65	93.846	3	0.092724679
ENSG00000114331	2	ENSG00000131584	4	1.50E-81	122	81.148	2	0.051542036
ENSG00000170482	1	ENSG00000089057	1	0	100	81	-2	0.127226463
ENSG00000136925	2	ENSG00000136842	1	0	555	100	-3	0.406295754
ENSG00000187010	1	ENSG00000188672	2	0	276	91.667	-2	0.559837728
ENSG00000188152	9	ENSG00000130950	0	0	307	98.371	-1	0.414304993
ENSG00000241476	9	ENSG00000165584	2	0	50	96	-2	0.105042017
ENSG00000169660	1	ENSG00000181396	2	1.14E-15	67	100	3	0.102918587

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000183960	2	ENSG00000089558	12	0	224	81.696	-1	0.20234869
ENSG00000038382	2	ENSG00000160145	3	0	364	85.165	1	0.117533097
ENSG00000177143	1	ENSG00000147400	2	1.09E-38	52	90.385	-2	0.121495327
ENSG00000188372	2	ENSG00000146707	0	2.12E-126	111	98.198	2	0.218074656
ENSG00000187257	1	ENSG00000081019	2	0	363	82.369	1	0.352427184
ENSG00000184302	3	ENSG00000138083	7	1.39E-127	209	86.603	-1	0.6875
ENSG00000205358	7	ENSG00000125144	7	4.16E-53	99	84.848	-2	0.692307692
ENSG00000186897	2	ENSG00000165985	9	3.59E-87	133	82.707	1	0.186797753
ENSG00000186897	2	ENSG00000131094	2	1.10E-81	134	88.06	2	0.188202247
ENSG00000184014	2	ENSG00000170456	3	0	96	84.375	1	0.060225847
ENSG00000197054	15	ENSG00000171295	2	0	289	85.467	1	0.571146245
ENSG00000197054	15	ENSG00000196757	0	0	191	86.387	2	0.377470356
ENSG00000130024	1	ENSG00000185127	7	4.09E-39	64	96.875	-3	0.115523466
ENSG00000183840	1	ENSG00000150551	3	0	319	100	2	0.702643172
ENSG00000179750	13	ENSG00000128383	7	8.34E-173	229	93.886	2	0.488272921
ENSG00000203883	3	ENSG00000164736	1	1.14E-42	74	86.486	-2	0.128249567
ENSG00000203883	3	ENSG00000171056	2	5.37E-34	70	80	-1	0.121317158
ENSG00000150551	2	ENSG00000183840	0	1.96E-113	165	100	1	0.528846154
ENSG00000213934	8	ENSG00000196565	6	3.80E-125	196	97.959	-2	0.951456311
ENSG00000183569	2	ENSG00000189306	2	3.30E-91	132	90.152	-3	0.328358209

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000122483	3	ENSG00000117500	1	2.71E-60	110	99.091	1	0.084680523
ENSG00000159899	2	ENSG00000169418	3	0	368	80.435	1	0.35148042
ENSG00000197479	10	ENSG00000120328	3	0	672	89.583	-1	0.815533981
ENSG00000186432	0	ENSG00000102753	10	0	523	85.851	1	0.411163522
ENSG00000183864	1	ENSG00000141232	2	1.45E-68	121	80.165	-2	0.083218707
ENSG00000244734	12	ENSG00000223609	2	1.73E-87	161	85.093	2	0.735159817
ENSG00000167720	1	ENSG00000167721	2	0	243	99.588	-3	0.684507042
ENSG00000187559	8	ENSG00000184492	1	0	272	95.956	-1	0.652278177
ENSG00000184083	2	ENSG00000048828	3	0	130	80.769	2	0.094614265
ENSG00000124593	6	ENSG00000278224	2	0	279	100	-2	0.7265625
ENSG00000204481	10	ENSG00000116721	2	0	486	96.091	-2	0.694285714
ENSG00000204481	10	ENSG00000120952	7	0	547	92.505	-1	0.781428571
ENSG00000185475	2	ENSG00000168569	2	1.41E-85	145	100	-1	0.494880546
ENSG00000187097	3	ENSG00000119723	1	3.51E-153	235	99.574	3	0.40239726
ENSG00000177138	2	ENSG00000183304	0	4.27E-124	87	94.253	1	0.266871166
ENSG00000180357	1	ENSG00000168916	6	0	95	82.105	-1	0.067328136
ENSG00000157734	2	ENSG00000166794	9	8.98E-115	111	100	3	0.247216036
ENSG00000188672	1	ENSG00000187010	9	0	341	92.962	-1	0.652007648
ENSG00000184814	9	ENSG00000206260	3	1.50E-121	141	88.652	2	0.298097252
ENSG00000100077	2	ENSG00000173020	0	0	687	83.988	2	0.709710744

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000178809	14	ENSG00000155428	2	0	345	99.42	-1	0.766666667
ENSG00000135702	9	ENSG00000183196	9	0	364	86.538	2	0.685499058
ENSG00000188976	0	ENSG00000187634	2	5.40E-72	124	100	2	0.132904609
ENSG00000119720	1	ENSG00000100764	0	0	567	99.471	2	0.487113402
ENSG00000243955	1	ENSG00000182793	14	4.00E-153	250	84	1	0.637755102
ENSG00000101558	2	ENSG00000140526	6	1.16E-149	106	95.283	2	0.197761194
ENSG00000196981	1	ENSG00000196363	12	0	301	89.369	-1	0.273139746
ENSG00000198648	1	ENSG00000172939	2	0	368	83.424	-1	0.675229358
ENSG00000204880	11	ENSG00000212721	13	4.56E-170	123	88.618	-1	0.664864865
ENSG00000204880	11	ENSG00000212722	4	3.73E-128	110	80	1	0.594594595
ENSG00000145194	2	ENSG00000163888	2	3.42E-84	124	100	2	0.117647059
ENSG00000169131	9	ENSG00000178338	3	0	433	81.062	1	0.52998776
ENSG00000187837	9	ENSG00000168298	10	1.94E-37	87	96.552	1	0.311827957
ENSG00000175029	3	ENSG00000019995	12	0	622	100	-3	0.631472081
ENSG00000219073	2	ENSG00000142789	1	4.02E-166	261	90.421	1	0.847402597
ENSG00000183791	1	ENSG00000206181	1	0	258	80.62	-2	0.472527473
ENSG00000165650	1	ENSG00000165646	2	0	718	100	3	0.558754864
ENSG00000117009	0	ENSG00000054277	2	0	504	100	3	0.9
ENSG00000196873	3	ENSG00000215126	1	0	560	99.286	-1	0.68627451
ENSG00000196873	3	ENSG00000147996	10	0	597	99.33	1	0.731617647

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000198889	13	ENSG00000198354	1	0	471	86.837	2	0.415343915
ENSG00000196767	3	ENSG00000184486	2	2.06E-69	132	81.061	-2	0.36565097
ENSG00000196767	3	ENSG00000198914	12	8.25E-65	135	82.222	-2	0.373961219
ENSG00000025039	2	ENSG00000116954	2	0	336	87.798	-1	0.740088106
ENSG00000065135	2	ENSG00000127955	3	0	319	91.85	-1	0.395781638
ENSG00000213551	1	ENSG00000138286	2	2.27E-169	235	99.149	-2	0.773026316
ENSG00000198668	2	ENSG00000160014	9	1.25E-98	173	90.751	-1	0.332692308
ENSG00000185966	12	ENSG00000163202	12	6.60E-60	63	95.238	-1	0.4921875
ENSG00000085265	3	ENSG00000160339	3	2.43E-169	234	80.769	-1	0.565217391
ENSG00000131721	12	ENSG00000203989	1	0	401	100	-1	0.987684729
ENSG00000198830	6	ENSG00000126814	2	0	177	85.876	-1	0.266165414
ENSG00000168925	1	ENSG00000168928	0	4.86E-158	154	96.104	2	0.504918033
ENSG00000147400	9	ENSG00000177143	9	5.87E-31	87	85.057	2	0.235772358
ENSG00000172288	0	ENSG00000172352	9	0	725	100	-1	0.920050761
ENSG00000137193	2	ENSG00000198355	0	2.69E-134	194	80.928	-1	0.217002237
ENSG00000264343	1	ENSG00000134250	3	2.86E-171	246	97.561	2	0.388625592
ENSG00000204116	1	ENSG00000109220	8	2.24E-68	68	85.294	-2	0.065827686
ENSG00000138161	6	ENSG00000213185	7	6.60E-74	119	95.798	1	0.174743025
ENSG00000168118	3	ENSG00000213029	7	2.39E-145	215	100	1	0.346774194
ENSG00000054277	1	ENSG00000117009	2	0	504	100	-3	0.687585266

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000280987	3	ENSG00000015479	1	0	727	100	-2	0.574703557
ENSG00000131019	8	ENSG00000111981	6	1.28E-71	88	87.5	-1	0.359183673
ENSG00000101811	2	ENSG00000177613	12	0	216	86.111	-1	0.334365325
ENSG00000168264	3	ENSG00000119669	0	1.36E-73	56	85.714	-1	0.095400341
ENSG00000150337	5	ENSG00000198019	1	0	198	98.99	1	0.266487214
ENSG00000107187	2	ENSG00000121454	0	9.39E-141	126	80.952	-1	0.313432836
ENSG00000124103	7	ENSG00000213714	1	2.73E-115	228	90.351	-2	0.636871508
ENSG00000035687	0	ENSG00000185100	1	0	91	85.714	-1	0.105691057
ENSG00000198355	2	ENSG00000137193	1	1.10E-129	183	81.967	1	0.273952096
ENSG00000244509	9	ENSG00000243811	1	9.25E-162	125	82.4	1	0.333333333
ENSG00000244509	9	ENSG00000128394	2	0	159	81.761	2	0.424
ENSG00000165125	0	ENSG00000127412	2	0	523	81.262	1	0.683660131
ENSG00000186340	2	ENSG00000137801	9	0	259	83.012	-1	0.146245059
ENSG00000203785	15	ENSG00000241794	12	1.35E-124	166	90.964	1	0.603636364
ENSG00000203989	12	ENSG00000131721	2	0	401	100	1	0.919724771
ENSG00000203859	12	ENSG00000203857	2	0	440	84.773	-1	0.771929825
ENSG00000112208	1	ENSG00000112210	6	0	790	100	-2	0.534144692
ENSG00000156875	2	ENSG00000148110	2	0	374	84.225	1	0.763265306
ENSG00000148704	1	ENSG00000116035	0	2.62E-55	105	85.714	2	0.314371257
ENSG00000131015	8	ENSG00000155918	2	1.23E-154	244	90.164	1	0.543429844

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000085998	1	ENSG00000171357	3	1.26E-137	199	100	1	0.217724289
ENSG00000172572	3	ENSG00000152270	3	0	128	82.031	1	0.108566582
ENSG00000125356	1	ENSG00000125352	1	1.90E-73	113	100	-2	0.869230769
ENSG00000076356	1	ENSG00000221866	1	0	634	84.227	-2	0.319073981
ENSG00000203811	1	ENSG00000278828	1	2.31E-86	137	99.27	1	0.85625
ENSG00000183206	12	ENSG00000166351	2	0	536	96.455	1	0.71849866
ENSG00000171357	3	ENSG00000085998	2	7.68E-138	199	100	-1	0.621875
ENSG00000143355	3	ENSG00000106689	2	1.21E-146	84	91.667	1	0.125937031
ENSG00000198626	2	ENSG00000196218	9	0	129	87.597	-2	0.025971411
ENSG00000143556	3	ENSG00000184330	3	8.59E-81	145	88.276	-1	0.953947368
ENSG00000068383	1	ENSG00000148826	3	3.88E-63	100	100	-1	0.181488203
ENSG00000096080	1	ENSG00000172426	1	4.84E-152	215	95.349	-2	0.595567867
ENSG00000121933	4	ENSG00000282608	3	2.39E-98	164	98.78	1	0.472622478
ENSG00000214435	3	ENSG00000270316	2	0	753	99.602	-1	0.916058394
ENSG00000163202	12	ENSG00000185966	9	9.95E-61	53	94.34	-2	0.417322835
ENSG00000196792	0	ENSG00000115808	9	0	99	83.838	-2	0.124215809
ENSG00000198173	13	ENSG00000189132	8	0	175	81.143	1	0.169082126
ENSG00000198173	13	ENSG00000185448	9	0	174	84.483	1	0.168115942
ENSG00000198740	3	ENSG00000114853	10	1.82E-176	245	93.878	1	0.134986226
ENSG00000080910	9	ENSG00000244414	1	0	172	98.256	-2	0.480446927

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000196911	2	ENSG00000162924	0	0	396	93.434	-1	0.549237171
ENSG00000119185	2	ENSG00000151693	2	0	736	100	-2	0.929292929
ENSG00000181274	9	ENSG00000165879	1	7.16E-108	119	90.756	-1	0.163461538
ENSG00000102038	1	ENSG00000153147	3	0	753	86.321	-1	0.616707617
ENSG00000184388	3	ENSG00000186288	13	0	269	99.257	2	0.366984993
ENSG00000035403	1	ENSG00000185009	13	0	369	100	2	0.215537383
ENSG00000100146	2	ENSG00000125398	2	1.02E-127	128	89.062	-1	0.133333333
ENSG00000166275	1	ENSG00000270316	9	1.42E-68	115	100	-1	0.30104712
ENSG00000143632	2	ENSG00000159251	9	0	378	98.942	1	0.670212766
ENSG00000143632	2	ENSG00000107796	3	0	378	97.884	-1	0.670212766
ENSG00000007341	1	ENSG00000134245	3	0	668	99.551	1	0.772254335
ENSG00000197329	1	ENSG00000139946	12	0	411	81.752	-1	0.345378151
ENSG00000125352	0	ENSG00000139797	7	7.86E-91	71	90.141	-1	0.177944862
ENSG00000125352	0	ENSG00000125356	7	3.67E-73	113	100	2	0.28320802
ENSG00000065243	2	ENSG00000123143	8	0	332	80.422	-1	0.337398374
ENSG00000134245	3	ENSG00000007341	2	0	668	99.551	-1	0.797136038
ENSG00000169474	9	ENSG00000169469	3	9.74E-33	60	93.333	-1	0.48
ENSG00000121481	2	ENSG00000204227	3	1.69E-123	151	84.106	1	0.239302694
ENSG00000196787	5	ENSG00000196747	2	1.19E-75	123	92.683	2	0.16356383
ENSG00000213714	7	ENSG00000124103	0	1.96E-127	228	90.351	2	0.953974895

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000167136	0	ENSG00000198917	0	1.98E-95	122	96.721	1	0.304239401
ENSG00000213185	10	ENSG00000138161	10	1.45E-74	119	95.798	-1	0.504237288
ENSG00000160716	2	ENSG00000117971	10	1.71E-58	65	81.538	-2	0.124282983
ENSG00000196406	15	ENSG00000198021	1	3.08E-97	214	83.645	-1	0.990740741
ENSG00000196406	15	ENSG00000203926	9	3.08E-97	214	83.645	-1	0.990740741
ENSG00000102144	0	ENSG00000170950	10	0	421	87.173	-2	0.676848875
ENSG00000163216	15	ENSG00000241794	3	4.26E-108	160	89.375	-1	0.61302682
ENSG00000167157	2	ENSG00000116132	5	9.67E-54	71	92.958	-1	0.160633484
ENSG00000130827	3	ENSG00000114554	3	0	623	83.949	-1	0.332977018
ENSG00000162367	2	ENSG00000104903	10	3.03E-30	62	85.484	1	0.187311178
ENSG00000165879	9	ENSG00000181274	8	6.85E-80	60	80	1	0.155844156
ENSG00000174876	0	ENSG00000187733	8	0	538	100	2	0.887788779
ENSG00000196475	3	ENSG00000198814	2	0	556	88.309	2	0.858024691
ENSG00000148513	12	ENSG00000180777	1	0	286	83.566	-2	0.213273676
ENSG00000124157	11	ENSG00000124233	12	0	275	83.273	2	0.409836066
ENSG00000178761	2	ENSG00000178802	9	0	744	100	-3	0.96124031
ENSG00000116985	3	ENSG00000183682	3	0	331	95.77	-2	0.635316699
ENSG00000203926	15	ENSG00000196406	11	3.13E-109	214	83.645	1	0.990740741
ENSG00000086015	2	ENSG00000105613	2	0	540	82.222	-1	0.281984334
ENSG00000086015	2	ENSG00000099308	13	0	376	80.851	-1	0.196344648

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000187140	2	ENSG00000186564	10	6.91E-63	104	91.346	1	0.217573222
ENSG00000143184	4	ENSG00000143185	10	5.96E-119	188	95.745	1	0.413186813
ENSG00000143032	2	ENSG00000125492	10	3.37E-42	73	82.192	-1	0.17016317
ENSG00000172426	1	ENSG00000096080	10	1.26E-151	215	95.349	2	0.693548387
ENSG00000000971	12	ENSG00000244414	1	2.04E-169	217	94.47	1	0.162303665
ENSG00000125551	2	ENSG00000183281	1	0	388	99.742	-1	0.987277354
ENSG00000125551	2	ENSG00000122194	11	0	179	89.385	1	0.455470738
ENSG00000127125	0	ENSG00000066185	10	1.83E-29	50	100	-1	0.101832994
ENSG00000221867	13	ENSG00000197172	6	0	466	96.996	1	0.797945205
ENSG00000162365	12	ENSG00000187048	10	0	528	94.886	1	0.566523605
ENSG00000240224	1	ENSG00000244474	8	0	781	95.519	1	0.967781908
ENSG00000184330	3	ENSG00000143556	8	8.50E-81	145	88.276	1	0.099793531
ENSG00000221864	13	ENSG00000187175	8	1.04E-85	52	84.615	1	0.305882353
ENSG00000143768	2	ENSG00000243709	6	0	335	95.821	1	0.481321839
ENSG00000142615	3	ENSG00000215704	2	1.05E-164	258	90.31	-1	0.791411043
ENSG00000092847	3	ENSG00000123908	0	0	856	82.477	2	0.343499197
ENSG00000131264	6	ENSG00000113722	9	5.57E-44	72	84.722	1	0.252631579
ENSG00000122224	3	ENSG00000169710	16	7.45E-73	78	80.769	2	0.109859155
ENSG00000244474	1	ENSG00000240224	9	0	781	95.519	-1	0.967781908
ENSG00000110680	10	ENSG00000175868	1	1.71E-68	80	88.75	2	0.327868852

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000198380	2	ENSG00000131459	9	0	446	81.166	2	0.423551757
ENSG00000025800	2	ENSG00000196911	2	0	260	93.462	1	0.19667171
ENSG00000163874	2	ENSG00000149289	1	2.16E-138	208	83.173	-2	0.234498309
ENSG00000198917	0	ENSG00000167136	3	3.44E-92	122	96.721	-1	0.324468085
ENSG00000187180	10	ENSG00000187223	1	8.46E-65	64	93.75	-1	0.581818182
ENSG00000187180	10	ENSG00000159455	3	3.07E-62	54	85.185	-2	0.490909091
ENSG00000278828	0	ENSG00000203811	1	1.67E-86	137	99.27	-1	0.872611465
ENSG00000112837	1	ENSG00000092607	1	6.33E-159	252	80.952	-1	0.172013652
ENSG00000120437	1	ENSG00000120438	1	1.74E-92	135	98.519	-1	0.280665281
ENSG00000188004	16	ENSG00000158714	1	0	327	99.388	-2	0.297543221
ENSG00000101204	2	ENSG00000120903	0	0	340	80.588	-1	0.465753425
ENSG00000162409	1	ENSG00000132356	13	0	243	82.716	2	0.343220339
ENSG00000197020	2	ENSG00000118620	13	0	591	80.88	1	0.308777429
ENSG00000198471	9	ENSG00000175077	0	2.46E-96	127	92.126	-1	0.279735683
ENSG00000284741	6	ENSG00000128655	7	0	582	99.828	1	0.623794212
ENSG00000019995	1	ENSG00000175029	2	0	622	100	3	0.878531073
ENSG00000185963	3	ENSG00000151746	1	0	185	89.189	-1	0.085687818
ENSG00000198354	2	ENSG00000198889	9	0	437	86.041	-2	0.752151463
ENSG00000180305	1	ENSG00000182931	15	9.21E-59	58	81.034	1	0.436090226
ENSG00000242366	1	ENSG00000244122	15	0	426	92.019	2	0.531835206

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000242366	1	ENSG00000241119	3	0	646	98.452	1	0.806491885
ENSG00000111642	2	ENSG00000116254	8	0	529	88.658	-1	0.241883859
ENSG00000203950	11	ENSG00000134590	11	1.10E-139	186	93.548	2	0.403470716
ENSG00000203950	11	ENSG00000212747	5	1.68E-137	201	91.045	1	0.436008677
ENSG00000117226	13	ENSG00000162645	12	0	75	89.333	-2	0.096525097
ENSG00000125430	9	ENSG00000153976	12	4.09E-140	274	93.066	1	0.154453213
ENSG00000158764	10	ENSG00000179914	3	0	283	88.693	1	0.797183099
ENSG00000116954	2	ENSG00000025039	3	0	336	87.798	1	0.7
ENSG00000196126	7	ENSG00000198502	8	0	332	87.952	1	0.811735941
ENSG00000186288	3	ENSG00000184388	5	0	269	99.257	-2	0.370523416
ENSG00000177272	2	ENSG00000182255	0	5.55E-173	147	89.796	-1	0.170533643
ENSG00000187733	0	ENSG00000174876	3	0	537	100	-2	0.886138614
ENSG00000134595	2	ENSG00000182968	3	1.74E-39	68	88.235	-2	0.142557652
ENSG00000175265	1	ENSG00000215252	3	0	258	100	2	0.332046332
ENSG00000102030	1	ENSG00000156269	3	2.22E-95	117	94.017	2	0.383606557
ENSG00000162645	2	ENSG00000117226	2	2.11E-174	94	88.298	2	0.129120879
ENSG00000198920	2	ENSG00000129235	2	4.03E-33	56	100	-3	0.047822374
ENSG00000198898	3	ENSG00000116489	1	2.09E-164	288	86.806	-1	0.5
ENSG00000104818	1	ENSG00000267631	2	4.30E-113	169	98.817	-1	0.98255814
ENSG00000244122	1	ENSG00000242366	4	0	426	92.019	-2	0.618287373

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000214827	6	ENSG00000182712	15	5.09E-90	152	99.342	-2	0.45508982
ENSG00000171488	3	ENSG00000171017	8	0	90	84.444	1	0.108043217
ENSG00000143185	4	ENSG00000143184	13	2.08E-119	188	95.745	-1	0.964102564
ENSG00000198914	5	ENSG00000196767	4	5.80E-58	81	80.247	2	0.162
ENSG00000213029	16	ENSG00000168118	1	4.29E-146	215	100	-1	0.846456693
ENSG00000198019	5	ENSG00000150337	4	5.79E-179	196	98.98	2	0.7
ENSG00000116459	0	ENSG00000116455	10	1.47E-74	126	88.095	-3	0.4921875
ENSG00000148842	3	ENSG00000076685	4	0	528	100	-2	0.603428571
ENSG00000196218	2	ENSG00000198626	7	0	122	87.705	2	0.024215959
ENSG00000196363	1	ENSG00000196981	7	0	320	87.188	1	0.512820513
ENSG00000117500	2	ENSG00000122483	5	2.91E-66	126	97.619	2	0.095238095
ENSG00000203923	13	ENSG00000204363	15	0	206	89.806	-1	1
ENSG00000196747	5	ENSG00000196787	0	2.06E-76	123	92.683	-2	0.694915254
ENSG00000169418	1	ENSG00000159899	7	0	368	80.435	-1	0.260070671
ENSG00000187223	10	ENSG00000187180	7	2.62E-38	64	93.75	1	0.35359116
ENSG00000187223	10	ENSG00000159455	13	2.83E-43	54	83.333	-1	0.298342541
ENSG00000166984	1	ENSG00000242220	13	5.63E-95	92	81.522	-2	0.260623229
ENSG00000143515	4	ENSG00000104043	2	0	110	88.182	1	0.089942764
ENSG00000066185	3	ENSG00000127125	2	3.61E-29	50	100	1	0.08912656
ENSG00000122194	2	ENSG00000125551	2	0	179	89.385	-1	0.196703297

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000122194	2	ENSG00000183281	2	0	179	88.827	-1	0.196703297
ENSG00000107854	4	ENSG00000173273	12	0	383	86.162	1	0.328473413
ENSG00000122136	6	ENSG00000171102	9	1.79E-142	229	94.323	1	0.817857143
ENSG00000102359	3	ENSG00000102362	7	0	432	100	-3	0.65158371
ENSG00000162385	1	ENSG00000111196	2	4.83E-97	148	98.649	1	0.657777778
ENSG00000213648	1	ENSG00000261052	12	0	426	99.531	-1	0.957303371
ENSG00000112305	2	ENSG00000112309	2	0	519	100	1	0.483690587
ENSG00000102362	2	ENSG00000102359	1	0	432	100	-3	0.331288344
ENSG00000171102	6	ENSG00000122136	2	1.80E-142	229	94.323	-1	0.753289474
ENSG00000248405	2	ENSG00000241484	1	0	269	100	-2	0.377279102
ENSG00000198692	0	ENSG00000173674	0	6.26E-48	91	90.11	2	0.34469697
ENSG00000198021	15	ENSG00000196406	3	3.13E-109	214	83.645	1	0.942731278
ENSG00000159455	10	ENSG00000187180	1	2.31E-72	66	84.848	2	0.6
ENSG00000159455	10	ENSG00000187223	3	2.13E-70	66	83.333	1	0.6
ENSG00000203857	12	ENSG00000203859	6	0	439	83.827	2	0.805504587
ENSG00000162415	2	ENSG00000130449	13	0	193	86.01	-2	0.162869198
ENSG00000180777	12	ENSG00000148513	3	0	286	85.664	-1	0.20545977
ENSG00000116785	12	ENSG00000134365	5	0	154	90.909	1	0.284658041
ENSG00000165841	12	ENSG00000138109	0	0	518	90.541	-1	0.556390977
ENSG00000165841	12	ENSG00000108242	2	0	80	87.5	2	0.085929108

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000107937	0	ENSG00000148377	15	0	311	100	-1	0.385856079
ENSG00000124233	11	ENSG00000124157	1	0	409	80.44	1	0.834693878
ENSG00000106689	2	ENSG00000143355	12	4.66E-137	142	85.915	1	0.324942792
ENSG00000182712	0	ENSG00000214827	12	2.31E-90	152	99.342	2	0.45508982
ENSG00000148826	3	ENSG00000163623	3	7.97E-90	122	82.787	-2	0.36746988
ENSG00000148826	3	ENSG00000068383	13	3.47E-63	100	100	1	0.301204819
ENSG00000118922	2	ENSG00000109787	15	2.94E-59	92	92.391	2	0.180746562
ENSG00000136404	3	ENSG00000166503	11	0	312	99.359	-1	0.570383912
ENSG00000204147	1	ENSG00000188611	11	0	741	98.516	-1	0.666966697
ENSG00000183196	9	ENSG00000135702	12	0	360	86.944	-2	0.857142857
ENSG00000204479	10	ENSG00000187545	1	0	489	86.912	-1	0.929657795
ENSG00000172352	0	ENSG00000172288	1	0	725	100	1	0.920050761
ENSG00000197620	4	ENSG00000197021	4	0	263	96.578	-1	0.58836689
ENSG00000107014	5	ENSG00000107018	5	6.30E-149	260	84.231	-1	0.663265306
ENSG00000173068	1	ENSG00000169594	13	0	80	87.5	1	0.072793449
ENSG00000131044	1	ENSG00000088356	9	1.66E-142	206	100	-1	0.332258065
ENSG00000146038	3	ENSG00000146049	9	2.41E-123	199	99.497	-1	0.282670455
ENSG00000119636	2	ENSG00000119711	3	1.88E-74	65	93.846	-2	0.122873346
ENSG00000278224	5	ENSG00000124593	3	0	279	100	2	0.398571429
ENSG00000204246	5	ENSG00000148136	10	4.33E-166	317	83.912	-1	0.856756757

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000183304	2	ENSG00000177138	10	1.11E-151	92	94.565	-1	0.229426434
ENSG00000147059	3	ENSG00000186787	10	0	321	97.819	2	0.713333333
ENSG00000133858	2	ENSG00000173451	8	0	320	99.688	3	0.160884867
ENSG00000147996	3	ENSG00000196873	11	0	597	99.33	-1	0.731617647
ENSG00000147996	3	ENSG00000215126	7	0	597	98.827	-1	0.731617647
ENSG00000172785	2	ENSG00000136682	7	0	597	98.66	1	0.966019417
ENSG00000172785	2	ENSG00000196873	8	0	560	98.036	1	0.906148867
ENSG00000183049	1	ENSG00000134072	6	0	309	84.79	-1	0.515
ENSG00000182583	12	ENSG00000169059	7	4.65E-113	92	98.913	-2	0.328571429
ENSG00000173575	2	ENSG00000153922	9	0	531	85.687	1	0.2904814
ENSG00000081853	4	ENSG00000254245	12	0	412	100	1	0.260924636
ENSG00000205497	10	ENSG00000205496	2	0	314	96.815	-2	0.94011976
ENSG00000172062	1	ENSG00000205571	10	0	192	100	2	0.381709742
ENSG00000154529	2	ENSG00000106714	10	0	1216	98.273	2	0.944099379
ENSG00000198363	2	ENSG00000177182	1	0	363	100	3	0.478891821
ENSG00000206260	9	ENSG00000184814	5	2.94E-119	125	88.8	-2	0.469924812
ENSG00000204388	3	ENSG00000204389	2	0	616	100	2	0.728994083
ENSG00000167633	9	ENSG00000240403	5	0	56	80.357	-1	0.086153846
ENSG00000167633	9	ENSG00000243772	3	0	377	86.472	1	0.58
ENSG00000243207	0	ENSG00000130810	1	0	312	100	1	0.392947103

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000205476	2	ENSG00000090061	9	0	447	100	1	0.433980583
ENSG00000165023	1	ENSG00000176490	6	4.64E-93	129	84.496	1	0.207395498
ENSG00000205029	8	ENSG00000186119	11	4.44E-160	304	81.25	2	0.926829268
ENSG00000180660	2	ENSG00000181541	11	0	374	94.118	1	0.455542022
ENSG00000196368	9	ENSG00000122824	11	8.63E-139	180	93.333	-1	0.228136882
ENSG00000205496	10	ENSG00000205497	11	0	314	96.815	2	1
ENSG00000166164	3	ENSG00000121281	11	0	726	100	-2	0.956521739
ENSG00000188611	1	ENSG00000204147	11	0	741	98.516	1	0.95
ENSG00000114554	2	ENSG00000130827	16	0	635	82.677	1	0.334915612
ENSG00000114554	2	ENSG00000076356	10	0	287	81.533	2	0.151371308
ENSG00000237541	11	ENSG00000196735	1	0	323	85.449	-1	0.74595843
ENSG00000205076	6	ENSG00000178934	3	1.58E-103	155	100	1	0.890804598
ENSG00000120235	3	ENSG00000188379	9	8.16E-105	86	81.395	1	0.452631579
ENSG00000174130	2	ENSG00000174125	13	0	315	86.349	-1	0.395728643
ENSG00000241978	16	ENSG00000157654	6	0	907	100	2	0.618690314
ENSG00000212722	11	ENSG00000212721	1	2.61E-130	132	84.848	1	0.628571429
ENSG00000184908	10	ENSG00000186510	1	0	687	91.266	1	0.790563867
ENSG00000090659	9	ENSG00000104938	7	4.67E-157	58	82.759	-1	0.121085595
ENSG00000170456	2	ENSG00000184014	8	0	101	82.178	-1	0.079277865
ENSG00000152192	2	ENSG00000091010	3	1.32E-117	163	89.571	2	0.38902148

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000152192	2	ENSG00000151615	6	1.53E-102	158	93.038	2	0.377088305
ENSG00000157654	4	ENSG00000241978	3	0	907	100	-2	0.618690314
ENSG00000144278	2	ENSG00000141429	3	0	505	85.941	1	0.585168019
ENSG00000198502	7	ENSG00000196126	1	0	332	87.952	-1	0.661354582
ENSG00000137628	4	ENSG00000181381	3	0	150	80.667	1	0.074626866
ENSG00000148110	3	ENSG00000156875	2	0	374	84.225	-1	0.739130435
ENSG00000215704	3	ENSG00000142615	13	1.22E-174	300	88.667	1	0.949367089
ENSG00000173674	0	ENSG00000198692	3	2.07E-43	82	91.463	-2	0.207594937
ENSG00000197079	8	ENSG00000108759	13	2.86E-151	51	80.392	1	0.112087912
ENSG00000197079	8	ENSG00000094796	15	2.58E-150	126	80.952	1	0.276923077
ENSG00000154545	1	ENSG00000187243	4	0	612	100	-1	0.714953271
ENSG00000160973	2	ENSG00000167702	4	1.57E-166	263	100	2	0.720547945
ENSG00000215126	3	ENSG00000196873	1	0	560	99.286	1	0.68627451
ENSG00000215126	3	ENSG00000147996	11	0	597	98.827	1	0.731617647
ENSG00000241794	14	ENSG00000163216	1	4.41E-126	160	89.375	1	0.49382716
ENSG00000241794	14	ENSG00000203785	6	1.43E-124	166	90.964	-1	0.512345679
ENSG00000166200	0	ENSG00000146281	6	1.07E-85	132	82.576	-2	0.213592233
ENSG00000155876	0	ENSG00000083750	6	0	273	97.802	1	0.513157895
ENSG00000204363	13	ENSG00000203923	2	0	206	89.806	1	1
ENSG00000198814	3	ENSG00000196475	1	0	457	85.339	-2	0.375513558

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000198610	12	ENSG00000187134	9	0	360	82.222	-1	0.886699507
ENSG00000163283	2	ENSG00000163286	2	0	493	97.566	-1	0.520591341
ENSG00000163283	2	ENSG00000163295	11	0	504	85.913	-2	0.532206969
ENSG00000165556	2	ENSG00000113722	14	4.16E-41	71	90.141	1	0.101139601
ENSG00000184659	8	ENSG00000204779	14	0	407	99.017	1	0.978365385
ENSG00000205572	5	ENSG00000172058	16	4.06E-136	211	100	1	0.512135922
ENSG00000169621	1	ENSG00000169618	6	0	241	100	-1	0.187111801
ENSG00000204227	4	ENSG00000121481	1	3.56E-123	151	84.106	-1	0.259005146
ENSG00000204779	8	ENSG00000184659	1	0	407	99.017	-1	0.676079734
ENSG00000204779	8	ENSG00000187559	1	0	165	98.788	-1	0.274086379
ENSG00000172349	2	ENSG00000172345	0	0	1381	100	3	0.852469136
ENSG00000125798	1	ENSG00000129514	2	1.42E-41	87	81.609	1	0.134883721
ENSG00000170122	9	ENSG00000184492	9	0	358	94.413	1	0.485753053
ENSG00000170122	9	ENSG00000187559	9	0	273	95.604	-1	0.370420624
ENSG00000177710	0	ENSG00000164729	9	0	184	92.935	1	0.481675393
ENSG00000187175	13	ENSG00000221864	1	6.13E-87	108	83.333	1	0.9
ENSG00000158373	7	ENSG00000180596	1	2.05E-65	118	92.373	1	0.435424354
ENSG00000204389	3	ENSG00000204388	1	0	616	100	-2	0.755828221
ENSG00000261052	1	ENSG00000213648	9	0	426	99.531	1	0.957303371
ENSG00000188379	3	ENSG00000120235	9	1.59E-124	86	81.395	-1	0.320895522

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000015479	3	ENSG00000280987	1	0	727	100	1	0.574703557
ENSG00000187134	11	ENSG00000198610	9	0	360	82.222	1	0.528634361
ENSG00000206535	6	ENSG00000154174	9	1.28E-41	68	100	1	0.150110375
ENSG00000142207	0	ENSG00000170262	12	5.20E-138	203	100	3	0.089387935
ENSG00000090061	1	ENSG00000205476	2	0	447	100	-1	0.697347894
ENSG00000204382	9	ENSG00000204379	6	1.29E-96	80	98.75	1	0.547945205
ENSG00000169059	12	ENSG00000182583	11	1.39E-113	92	98.913	2	0.353846154
ENSG00000180138	12	ENSG00000113712	4	0	334	81.138	1	0.413878563
ENSG00000101200	2	ENSG00000101405	9	2.95E-54	107	80.374	-1	0.537688442
ENSG00000172661	1	ENSG00000099290	7	0	554	98.917	-2	0.41969697
ENSG00000212721	11	ENSG00000204880	7	0	145	93.103	-2	0.419075145
ENSG00000212721	11	ENSG00000212722	4	2.23E-121	132	84.848	-1	0.38150289
ENSG00000101350	1	ENSG00000084731	4	0	114	83.333	-2	0.117163412
ENSG00000105619	5	ENSG00000105618	1	7.72E-44	73	100	-3	0.255244755
ENSG00000205863	2	ENSG00000240654	1	0	403	97.022	-1	0.671666667
ENSG00000107404	2	ENSG00000161202	1	0	132	82.576	2	0.171206226
ENSG00000229183	4	ENSG00000256713	7	0	506	99.407	1	0.502482622
ENSG00000223609	15	ENSG00000244734	1	2.29E-87	161	85.093	-2	0.752336449
ENSG00000213214	1	ENSG00000050327	16	0	389	99.486	-1	0.68006993
ENSG00000205571	1	ENSG00000172062	16	0	192	100	-2	0.377952756

Table 6, continued

Query	Query Branch	Subject	Subject Branch	e-value	Alignment Length	Percent Identity	RFS	Query RFS proportion
ENSG00000005156	1	ENSG00000092871	0	0	867	100	-2	0.749351772
ENSG00000125144	7	ENSG00000205358	5	3.96E-53	78	80.769	1	0.541666667
ENSG00000204379	9	ENSG00000204382	0	9.21E-97	80	98.75	-1	0.327868852
ENSG00000183474	0	ENSG00000145736	3	0	392	99.49	-1	0.590361446
ENSG00000212747	11	ENSG00000134590	3	2.18E-171	216	85.648	-2	0.316715543
ENSG00000212747	11	ENSG00000203950	6	3.07E-166	213	86.385	-1	0.312316716

Conservative Dataset

Table 7 Summary of identified RFSD gene pairs in the Conservative dataset. All matches are reciprocal.

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000006451	ENSG00000144118	1.07E-71	126	89.683	-2
ENSG00000004975	ENSG00000161202	0	131	91.603	1
ENSG00000006116	ENSG00000166862	5.47E-132	175	84.571	-1
ENSG00000015568	ENSG00000183054	0	946	100	1
ENSG00000088256	ENSG00000156052	0	360	90.278	-1
ENSG00000050327	ENSG00000213214	0	389	99.486	1
ENSG00000019549	ENSG00000124216	4.92E-73	113	85.841	-1
ENSG00000186847	ENSG00000128422	0	312	89.423	2
ENSG00000068976	ENSG00000100994	0	833	84.154	2
ENSG00000083720	ENSG00000198754	0	221	80.543	-2
ENSG00000197208	ENSG00000197375	0	322	86.025	2
ENSG00000100490	ENSG00000125375	2.67E-175	258	99.225	-2
ENSG00000101162	ENSG00000124172	0	395	100	1
ENSG00000099804	ENSG00000107341	8.22E-122	198	86.869	1
ENSG00000083812	ENSG00000249471	0	516	90.891	1
ENSG00000087303	ENSG00000087302	0	236	98.729	-3
ENSG00000095917	ENSG00000172236	0	282	85.106	-1
ENSG00000099974	ENSG00000099977	4.27E-61	104	96.154	1
ENSG00000100450	ENSG00000100453	4.38E-108	129	86.047	2
ENSG00000100564	ENSG00000054690	3.88E-65	100	100	-1
ENSG00000100314	ENSG00000100319	5.87E-83	139	97.122	1
ENSG00000101405	ENSG00000101200	2.33E-54	107	80.374	1
ENSG00000090581	ENSG00000059145	0	302	100	-3
ENSG00000086232	ENSG00000106305	7.52E-106	178	100	-2

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000092607	ENSG00000112837	7.74E-157	234	85.47	2
ENSG00000243811	ENSG00000128394	0	246	88.211	-2
ENSG00000100030	ENSG00000102882	0	346	88.15	1
ENSG00000109061	ENSG00000264424	0	630	95.873	-2
ENSG00000016082	ENSG00000159556	0	188	81.383	-1
ENSG00000105664	ENSG00000113296	0	271	87.823	1
ENSG00000102128	ENSG00000172476	0	254	98.031	-1
ENSG00000105649	ENSG00000152932	1.68E-123	194	88.66	-1
ENSG00000104863	ENSG00000148943	2.56E-101	204	82.353	-2
ENSG00000104129	ENSG00000137880	9.50E-128	161	100	3
ENSG00000108773	ENSG00000114166	0	239	83.682	2
ENSG00000104888	ENSG00000091664	0	512	82.422	1
ENSG00000105254	ENSG00000105258	3.44E-79	120	100	-3
ENSG00000114853	ENSG00000198740	0	260	91.538	-1
ENSG00000108379	ENSG00000154342	0	347	85.014	-1
ENSG00000108590	ENSG00000129235	0	408	100	-1
ENSG00000111615	ENSG00000139278	0	379	99.736	3
ENSG00000039123	ENSG00000067113	1.23E-100	160	100	-2
ENSG00000108417	ENSG00000171360	0	252	87.302	-1
ENSG00000114349	ENSG00000134183	0	326	83.129	1
ENSG00000103064	ENSG00000103061	0	977	100	1
ENSG00000112309	ENSG00000112305	0	519	100	-3
ENSG00000107018	ENSG00000107014	5.52E-139	260	83.462	1
ENSG00000123908	ENSG00000092847	0	838	83.652	-2
ENSG00000132207	ENSG00000181625	0	244	100	-1
ENSG00000254245	ENSG00000081853	0	412	100	-1
ENSG00000121068	ENSG00000135111	4.79E-169	239	87.448	-1
ENSG00000130449	ENSG00000162415	0	532	80.263	2
ENSG00000282608	ENSG00000121933	9.93E-121	163	100	1
ENSG00000123143	ENSG00000065243	0	331	80.363	2
ENSG00000120324	ENSG00000177839	0	672	92.411	-2

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000124140	ENSG00000113504	0	256	89.453	1
ENSG00000122543	ENSG00000135175	2.30E-153	231	98.268	2
ENSG00000131094	ENSG00000186897	7.33E-87	134	88.06	-2
ENSG00000130733	ENSG00000142453	1.09E-136	245	100	-2
ENSG00000115042	ENSG00000144199	0	329	96.657	-2
ENSG00000160145	ENSG00000038382	0	182	86.264	-1
ENSG00000131462	ENSG00000037042	0	454	97.577	-1
ENSG00000127780	ENSG00000180016	5.45E-164	192	91.667	1
ENSG00000115486	ENSG00000168906	3.62E-126	185	100	-2
ENSG00000120329	ENSG00000102743	0	315	86.667	2
ENSG00000129204	ENSG00000170832	0	779	92.94	-1
ENSG00000121281	ENSG00000166164	0	725	100	2
ENSG00000134250	ENSG00000264343	2.59E-159	238	97.479	-2
ENSG00000126778	ENSG00000170577	1.59E-128	188	95.213	2
ENSG00000128383	ENSG00000179750	7.76E-173	229	93.886	-2
ENSG00000119778	ENSG00000156802	0	335	81.791	-2
ENSG00000058262	ENSG00000065665	0	475	93.684	1
ENSG00000121297	ENSG00000179981	0	148	85.135	-1
ENSG00000109805	ENSG00000178177	0	474	100	2
ENSG00000028839	ENSG00000146411	0	463	100	-2
ENSG00000125398	ENSG00000100146	7.06E-129	121	90.909	1
ENSG00000124657	ENSG00000168131	5.33E-171	310	81.613	2
ENSG00000119729	ENSG00000151665	2.05E-143	219	100	-2
ENSG00000187545	ENSG00000204479	0	489	86.912	1
ENSG00000099822	ENSG00000138622	0	550	90	-1
ENSG00000125966	ENSG00000156103	0	176	85.227	-1
ENSG00000118579	ENSG00000047662	0	1107	100	-3
ENSG00000187272	ENSG00000241595	4.92E-135	172	83.14	-1

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000128245	ENSG00000170027	4.00E-142	246	86.992	1
ENSG00000132475	ENSG00000188375	0	129	96.899	-1
ENSG00000119673	ENSG00000184227	0	304	98.355	1
ENSG00000131459	ENSG00000198380	0	446	81.166	-2
ENSG00000115386	ENSG00000172023	2.78E-108	217	81.106	1
ENSG00000112659	ENSG00000044090	0	210	85.714	2
ENSG00000188536	ENSG00000206172	5.22E-112	178	95.506	1
ENSG00000125629	ENSG00000186480	3.09E-110	185	84.324	-1
ENSG00000133243	ENSG00000064726	0	379	84.697	2
ENSG00000134072	ENSG00000183049	0	323	82.353	1
ENSG00000136231	ENSG00000159217	0	188	80.319	1
ENSG00000121454	ENSG00000107187	1.23E-138	127	80.315	1
ENSG00000139648	ENSG00000186049	0	365	89.315	1
ENSG00000005339	ENSG00000100393	0	471	90.446	1
ENSG00000083750	ENSG00000155876	0	273	97.802	-1
ENSG00000047457	ENSG00000163755	0	411	100	-2
ENSG00000128881	ENSG00000146216	0	314	81.529	1
ENSG00000136240	ENSG00000105438	1.05E-123	213	83.568	-2
ENSG00000126934	ENSG00000169032	0	228	92.544	-2
ENSG00000137273	ENSG00000103241	4.21E-66	113	96.46	1
ENSG00000138083	ENSG00000184302	2.24E-131	210	86.667	1
ENSG00000136379	ENSG00000129968	2.42E-152	240	82.917	1
ENSG00000134853	ENSG00000113721	0	152	81.579	1
ENSG00000243709	ENSG00000143768	0	399	94.486	-1
ENSG00000144119	ENSG00000165985	1.20E-92	132	88.636	-1
ENSG00000103740	ENSG00000166411	0	870	100	1
ENSG00000133026	ENSG00000133392	0	981	81.957	1
ENSG00000135100	ENSG00000157895	0	474	100	-3
ENSG00000141965	ENSG00000145780	0	115	83.478	1

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000139266	ENSG00000144583	9.52E-99	159	90.566	-2
ENSG00000136842	ENSG00000136925	0	555	100	3
ENSG00000135945	ENSG00000158417	3.60E-180	284	100	-3
ENSG00000081019	ENSG00000187257	0	363	82.369	-1
ENSG00000126814	ENSG00000198830	1.41E-156	183	85.246	-1
ENSG00000116254	ENSG00000111642	0	349	86.533	1
ENSG00000181381	ENSG00000137628	0	150	80.667	-1
ENSG00000084731	ENSG00000101350	0	115	82.609	1
ENSG00000135018	ENSG00000188021	0	163	90.184	1
ENSG00000113712	ENSG00000180138	0	335	81.194	2
ENSG00000141429	ENSG00000144278	0	505	85.941	-1
ENSG00000138685	ENSG00000170917	5.26E-138	196	98.98	-3
ENSG00000136682	ENSG00000172785	0	597	98.66	-1
ENSG00000141232	ENSG00000183864	1.74E-82	121	80.165	2
ENSG00000137801	ENSG00000186340	0	261	83.142	2
ENSG00000139112	ENSG00000170296	1.39E-68	116	87.069	1
ENSG00000143933	ENSG00000160014	4.47E-98	153	99.346	-1
ENSG00000100764	ENSG00000119720	0	567	99.471	-2
ENSG00000109158	ENSG00000145863	0	334	83.533	-1
ENSG00000105464	ENSG00000161509	0	418	83.493	-1
ENSG00000136698	ENSG00000152093	0	250	99.6	-2
ENSG00000116489	ENSG00000198898	1.19E-164	288	86.806	1
ENSG00000165055	ENSG00000087995	0	458	95.633	2
ENSG00000112246	ENSG00000159263	0	359	86.072	-1
ENSG00000102753	ENSG00000186432	0	523	85.851	-1
ENSG00000139946	ENSG00000197329	0	411	81.752	1
ENSG00000221866	ENSG00000076356	0	726	81.818	1
ENSG00000008226	ENSG00000060971	1.28E-87	143	97.902	1
ENSG00000146707	ENSG00000188372	2.13E-126	111	98.198	-2
ENSG00000163286	ENSG00000163283	0	488	97.746	1

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000181826	ENSG00000154274	2.82E-82	142	99.296	-2
ENSG00000164933	ENSG00000164934	2.34E-49	108	99.074	2
ENSG00000152977	ENSG00000156925	5.18E-136	186	89.785	1
ENSG00000153779	ENSG00000176679	1.50E-124	197	92.893	1
ENSG00000173451	ENSG00000133858	0	320	99.688	-3
ENSG00000181541	ENSG00000180660	0	386	92.746	-1
ENSG00000181789	ENSG00000158623	0	596	83.725	-1
ENSG00000148377	ENSG00000107937	0	311	100	1
ENSG00000075886	ENSG00000152086	0	330	97.576	1
ENSG00000177971	ENSG00000173548	1.11E-155	128	100	1
ENSG00000167191	ENSG00000174628	0	592	100	-1
ENSG00000005022	ENSG00000151729	1.67E-176	299	88.963	-2
ENSG00000167553	ENSG00000167552	0	463	88.553	1
ENSG00000272617	ENSG00000258429	2.75E-126	189	98.413	2
ENSG00000172345	ENSG00000172349	0	1381	100	-3
ENSG00000155428	ENSG00000178809	0	389	99.743	-1
ENSG00000166800	ENSG00000171989	0	332	82.229	-1
ENSG00000104043	ENSG00000143515	0	189	81.481	-1
ENSG00000198077	ENSG00000255974	0	490	94.082	-2
ENSG00000164900	ENSG00000168505	2.45E-69	108	87.037	2
ENSG00000152270	ENSG00000172572	0	100	84	-1
ENSG00000177879	ENSG00000157823	9.06E-109	192	84.375	-2
ENSG00000146083	ENSG00000137075	4.86E-147	118	84.746	2
ENSG00000167977	ENSG00000180901	3.13E-106	162	83.951	-2
ENSG00000146049	ENSG00000146038	6.53E-124	199	99.497	1
ENSG00000181693	ENSG00000181767	0	345	85.797	1
ENSG00000166794	ENSG00000157734	5.17E-92	111	100	2

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000171103	ENSG00000163806	2.43E-96	133	99.248	-2
ENSG00000165516	ENSG00000165525	0	550	99.818	-1
ENSG00000256713	ENSG00000229183	0	506	99.407	-1
ENSG00000172058	ENSG00000205572	5.94E-136	211	100	-1
ENSG00000169618	ENSG00000169621	0	241	100	1
ENSG00000173020	ENSG00000100077	0	685	84.088	-2
ENSG00000175077	ENSG00000198471	7.41E-100	169	89.349	1
ENSG00000244414	ENSG00000080910	0	172	98.256	2
ENSG00000172519	ENSG00000186723	0	318	92.767	-1
ENSG00000175344	ENSG00000166664	0	450	99.778	2
ENSG00000057149	ENSG00000206073	0	383	91.906	-1
ENSG00000170950	ENSG00000102144	0	405	87.16	2
ENSG00000092871	ENSG00000005156	0	867	100	2
ENSG00000166363	ENSG00000170790	0	342	92.105	-1
ENSG00000187048	ENSG00000162365	0	479	95.407	-1
ENSG00000153922	ENSG00000173575	0	499	85.972	-1
ENSG00000163464	ENSG00000180871	1.99E-153	292	84.589	-1
ENSG00000106714	ENSG00000154529	0	1216	98.273	-2
ENSG00000168928	ENSG00000168925	5.29E-158	154	96.104	-2
ENSG00000170262	ENSG00000142207	3.86E-139	203	100	-1
ENSG00000166947	ENSG00000166946	0	361	100	-3
ENSG00000151693	ENSG00000119185	0	736	100	2
ENSG00000146281	ENSG00000166200	7.57E-86	132	82.576	2
ENSG00000164729	ENSG00000177710	0	380	93.684	-1
ENSG00000168961	ENSG00000171916	0	322	94.72	-1
ENSG00000166503	ENSG00000136404	0	312	99.359	1
ENSG00000163322	ENSG00000163319	0	394	99.746	-3
ENSG00000154025	ENSG00000154016	0	494	99.595	-3
ENSG00000196778	ENSG00000181963	2.47E-172	324	83.951	1
ENSG00000164855	ENSG00000198517	4.09E-178	269	100	3

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000123064	ENSG00000186710	2.17E-94	125	100	3
ENSG00000173349	ENSG00000136709	0	692	100	-2
ENSG00000160339	ENSG00000085265	1.03E-162	219	84.018	1
ENSG00000162391	ENSG00000162390	9.79E-180	255	99.608	3
ENSG00000165646	ENSG00000165650	0	718	100	-3
ENSG00000149289	ENSG00000102053	0	258	83.333	1
ENSG00000167721	ENSG00000167720	0	243	99.588	3
ENSG00000166439	ENSG00000166435	0	727	100	1
ENSG00000167395	ENSG00000151006	1.44E-89	128	100	-1
ENSG00000168569	ENSG00000185475	1.84E-85	145	100	1
ENSG00000111196	ENSG00000162385	1.95E-96	148	98.649	-1
ENSG00000157326	ENSG00000187630	0	235	91.064	-2
ENSG00000248871	ENSG00000161955	0	431	98.144	-1
ENSG00000171446	ENSG00000204897	0	269	91.45	-1
ENSG00000173273	ENSG00000107854	0	615	85.041	-1
ENSG00000167702	ENSG00000160973	2.34E-166	263	100	-2
ENSG00000128886	ENSG00000167004	2.38E-124	210	100	3
ENSG00000160683	ENSG00000186174	0	882	100	-3
ENSG00000172939	ENSG00000198648	0	338	83.728	1
ENSG00000166351	ENSG00000183206	0	536	96.455	-1
ENSG00000157827	ENSG00000161791	0	130	86.154	-1
ENSG00000166377	ENSG00000054793	0	570	82.281	-1
ENSG00000148136	ENSG00000204246	5.88E-176	317	83.912	1
ENSG00000149968	ENSG00000166670	0	217	85.714	-1
ENSG00000171478	ENSG00000171489	5.78E-119	170	100	2
ENSG00000158714	ENSG00000188004	0	327	99.388	2
ENSG00000163888	ENSG00000145194	2.48E-84	124	100	-2
ENSG00000151746	ENSG00000185963	0	201	85.572	1
ENSG00000213512	ENSG00000162654	0	482	82.365	1

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000153147	ENSG00000102038	0	750	86.4	1
ENSG00000178934	ENSG00000205076	1.69E-103	155	100	-1
ENSG00000187527	ENSG00000127249	0	108	80.556	1
ENSG00000178802	ENSG00000178761	0	744	100	3
ENSG00000100023	ENSG00000100027	0	569	100	-3
ENSG00000104177	ENSG00000188467	0	222	97.297	-2
ENSG00000177182	ENSG00000198363	0	363	100	-3
ENSG00000183281	ENSG00000125551	0	388	99.742	1
ENSG00000114374	ENSG00000124486	0	473	88.795	2
ENSG00000197172	ENSG00000221867	0	466	96.996	-1
ENSG00000127955	ENSG00000065135	0	362	92.541	1
ENSG00000070808	ENSG00000058404	0	476	85.714	1
ENSG00000215252	ENSG00000175265	0	258	100	-2
ENSG00000197021	ENSG00000197620	0	355	97.183	-1
ENSG00000196735	ENSG00000237541	0	323	85.449	1
ENSG00000132356	ENSG00000162409	0	243	82.716	-2
ENSG00000131584	ENSG00000114331	1.81E-67	122	81.148	-2
ENSG00000187082	ENSG00000186579	3.44E-60	100	100	-1
ENSG00000122824	ENSG00000196368	8.89E-136	177	96.045	1
ENSG00000186510	ENSG00000184908	0	687	91.266	-1
ENSG00000186599	ENSG00000186562	4.36E-74	110	100	-1
ENSG00000119723	ENSG00000187097	1.22E-153	235	99.574	-3
ENSG00000187243	ENSG00000154545	0	612	100	1
ENSG00000179914	ENSG00000158764	0	283	88.693	-1
ENSG00000240654	ENSG00000205863	0	226	95.133	1
ENSG00000186119	ENSG00000205029	3.26E-148	304	81.25	-2
ENSG00000241119	ENSG00000242366	0	778	96.53	-1
ENSG00000196565	ENSG00000213934	1.30E-124	195	97.949	2
ENSG00000185009	ENSG00000035403	0	369	100	3
ENSG00000116721	ENSG00000204481	0	486	96.091	2
ENSG00000177613	ENSG00000101811	0	208	87.5	1

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000182255	ENSG00000177272	6.96E-172	133	91.729	1
ENSG00000241484	ENSG00000248405	0	269	100	2
ENSG00000183682	ENSG00000116985	0	325	95.692	2
ENSG00000176490	ENSG00000165023	2.83E-101	126	86.508	-1
ENSG00000187634	ENSG00000188976	4.90E-72	124	100	-2
ENSG00000076685	ENSG00000148842	0	528	100	1
ENSG00000099290	ENSG00000172661	0	958	98.225	2
ENSG00000159899	ENSG00000169418	0	368	80.435	1
ENSG00000244734	ENSG00000223609	1.73E-87	161	85.093	2
ENSG00000124593	ENSG00000278224	0	279	100	-2
ENSG00000184814	ENSG00000206260	1.50E-121	141	88.652	2
ENSG00000135702	ENSG00000183196	0	364	86.538	2
ENSG00000196981	ENSG00000196363	0	301	89.369	-1
ENSG00000175029	ENSG00000019995	0	622	100	-3
ENSG00000117009	ENSG00000054277	0	504	100	3
ENSG00000196873	ENSG00000147996	0	597	99.33	1
ENSG00000198889	ENSG00000198354	0	471	86.837	2
ENSG00000025039	ENSG00000116954	0	336	87.798	-1
ENSG00000131721	ENSG00000203989	0	401	100	-1
ENSG00000172288	ENSG00000172352	0	725	100	-1
ENSG00000137193	ENSG00000198355	2.69E-134	194	80.928	-1
ENSG00000138161	ENSG00000213185	6.60E-74	119	95.798	1
ENSG00000168118	ENSG00000213029	2.39E-145	215	100	1
ENSG00000124103	ENSG00000213714	2.73E-115	228	90.351	-2
ENSG00000203785	ENSG00000241794	1.35E-124	166	90.964	1
ENSG00000156875	ENSG00000148110	0	374	84.225	1
ENSG00000085998	ENSG00000171357	1.26E-137	199	100	1
ENSG00000125356	ENSG00000125352	1.90E-73	113	100	-2

Table 7, continued

Query	Subject	e-value	Alignment Length	Percentage Identity	Frameshift
ENSG00000203811	ENSG00000278828	2.31E-86	137	99.27	1
ENSG00000198626	ENSG00000196218	0	129	87.597	-2
ENSG00000143556	ENSG00000184330	8.59E-81	145	88.276	-1
ENSG00000068383	ENSG00000148826	3.88E-63	100	100	-1
ENSG00000096080	ENSG00000172426	4.84E-152	215	95.349	-2
ENSG00000184388	ENSG00000186288	0	269	99.257	2
ENSG00000007341	ENSG00000134245	0	668	99.551	1
ENSG00000121481	ENSG00000204227	1.69E-123	151	84.106	1
ENSG00000167136	ENSG00000198917	1.98E-95	122	96.721	1
ENSG00000130827	ENSG00000114554	0	623	83.949	-1
ENSG00000174876	ENSG00000187733	0	538	100	2
ENSG00000196475	ENSG00000198814	0	556	88.309	2
ENSG00000143184	ENSG00000143185	5.96E-119	188	95.745	1
ENSG00000142615	ENSG00000215704	1.05E-164	258	90.31	-1
ENSG00000196126	ENSG00000198502	0	332	87.952	1
ENSG00000214827	ENSG00000182712	5.09E-90	152	99.342	-2
ENSG00000203923	ENSG00000204363	0	206	89.806	-1
ENSG00000122136	ENSG00000171102	1.79E-142	229	94.323	1
ENSG00000102359	ENSG00000102362	0	432	100	-3
ENSG00000213648	ENSG00000261052	0	426	99.531	-1
ENSG00000204147	ENSG00000188611	0	741	98.516	-1
ENSG00000205497	ENSG00000205496	0	314	96.815	-2
ENSG00000172062	ENSG00000205571	0	192	100	2
ENSG00000204388	ENSG00000204389	0	616	100	2
ENSG00000205476	ENSG00000090061	0	447	100	1
ENSG00000241978	ENSG00000157654	0	907	100	2
ENSG00000212722	ENSG00000212721	2.61E-130	132	84.848	1
ENSG00000184659	ENSG00000204779	0	407	99.017	1

Supplemental Results

The analysis of the Conservative dataset revealed results consistent with the analysis of the Standard dataset. To increase readability the Conservative results were removed from Chapter 2 and are listed below with a brief description of their analysis. All analysis was done identically to that done for the Standard dataset unless specifically reported otherwise here. Analysis was done for the Conservative dataset in cases where we wanted to confirm that the larger gene and frameshift sizes it had would not skew the results, for example some molecular functions require larger proteins so we wanted to determine whether this would affect the most common molecular functions in each dataset.

Analysis of Conservative Dataset RFSD genes' molecular functions

To determine the characteristics shared by RFSD genes in the Conservative dataset we performed Gene Ontology (GO) enrichment analysis. Similar to the Standard dataset we observed that RFSD genes are enriched for molecular functions related to signaling activity or transcriptional activation (Figure 12).

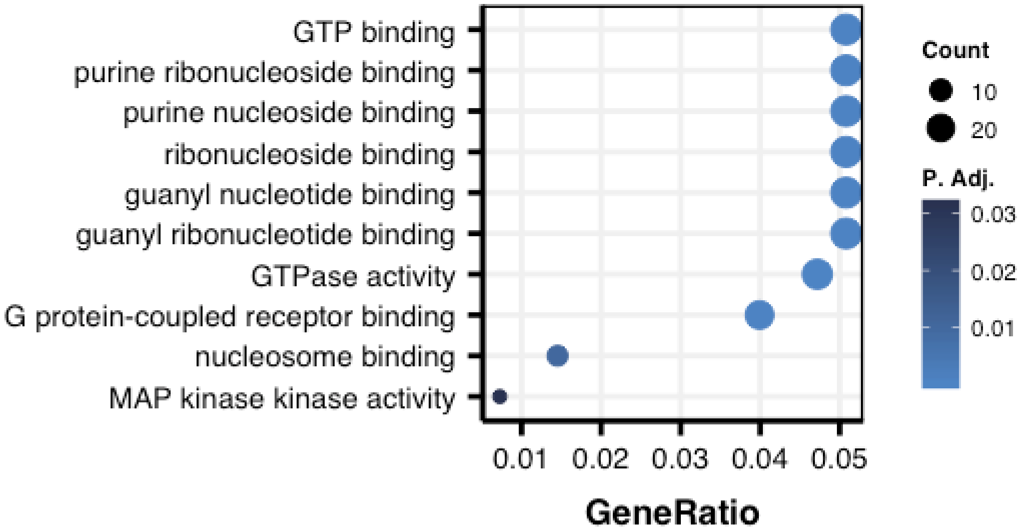


Fig. 12 Gene Ontology enrichment analysis results of the molecular functions of the RFSD genes in the Conservative dataset

Analysis of Conservative Dataset RFSD genes' biological processes

When performing a GO enrichment analysis for biological processes RFSD genes are involved in we observed that these genes show an enrichment for involvement in developmental and patterning processes (Figure 13). This was also consistent with the results for the Standard dataset.

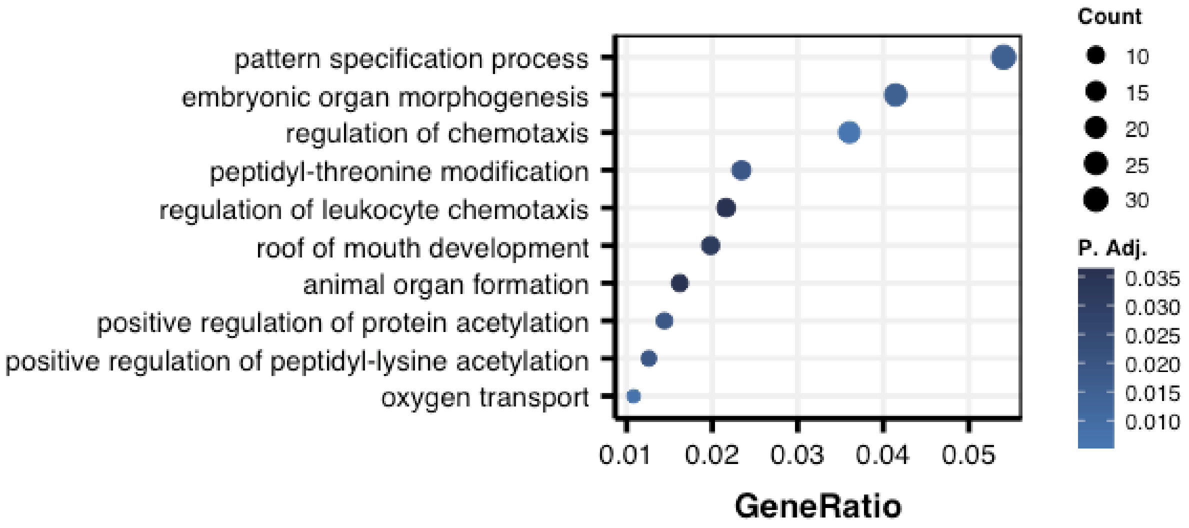


Fig. 13 Gene Ontology enrichment analysis results of the biological processes RFSD genes participate in for the Conservative dataset

Genotype-Tissue Expression analysis of Conservative Dataset RFSD genes

A GTEx analysis revealed that the Conservative dataset and Standard datasets show the same expression patterns. Approximately a third of the Conservative RFSD genes show expression in each of 52 tissues examined while about ten percent show expression in none of the tissues. Over half the genes examined show non-constitutive tissue-specific expression. We observed a significant enrichment for testes expression in RFSD genes with a z-score of 2.974 (Figure 14). We also found a significant underrepresentation of the same three tissues as in the Standard dataset: Muscle – Skeletal, Heart – Left Ventricle and Whole Blood. The z-scores were -2.129, -2.168 and -3.134 respectively. For our interpretation of these results please see Chapter 2.

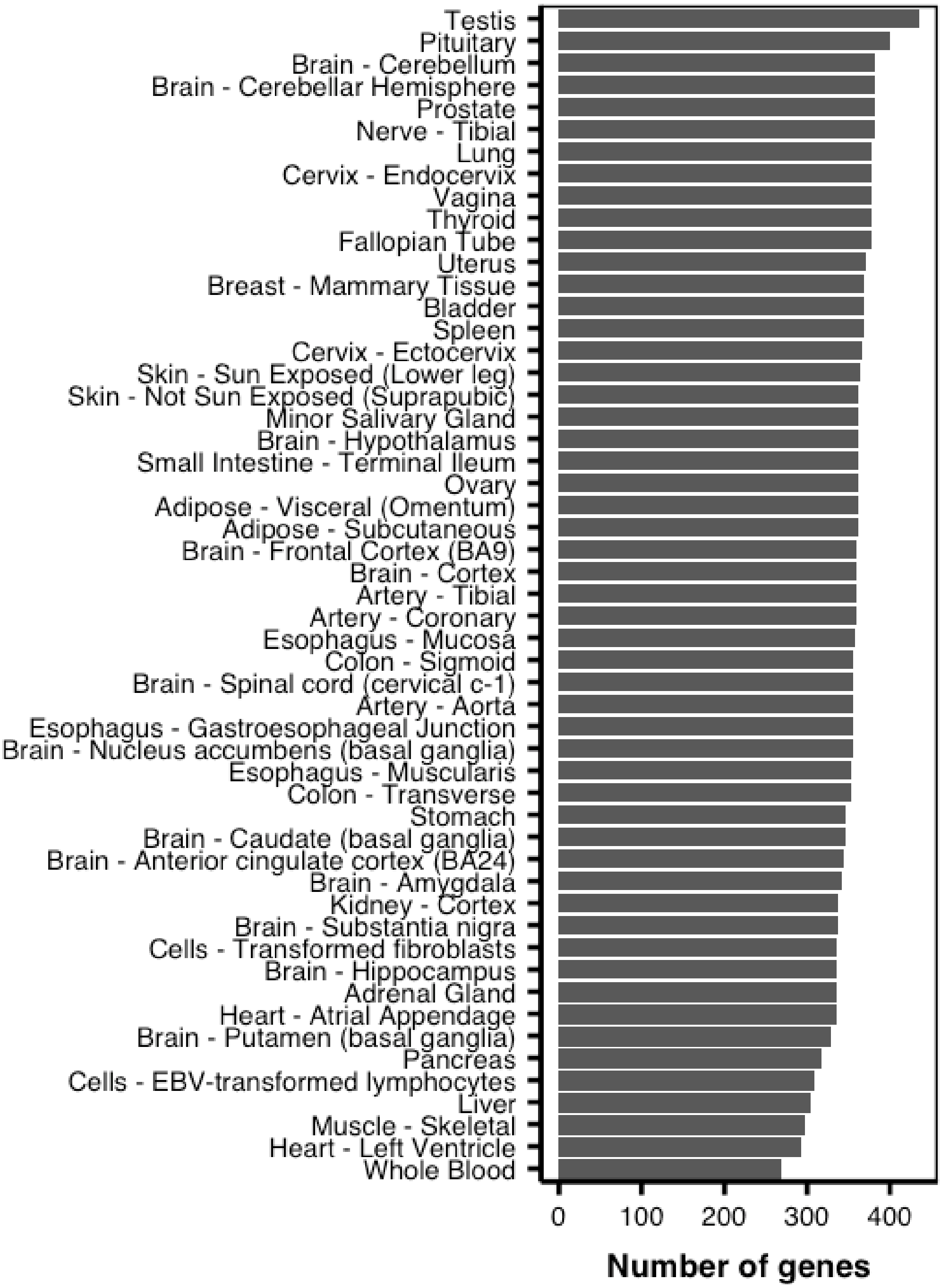


Fig. 14 Genotype-Tissue Expression analysis results of the RFSD genes in the Conservative dataset. Testis expression is significantly overrepresented (z-score of 2.974). Muscle – Skeletal, Heart – Left Ventricle and Whole Blood showed significant underrepresentation of RFSD genes (z-scores of -2.129, -2.168 and -3.134 respectively).

An excess of RFSD genes are found on the sex chromosomes

Although the Standard dataset produced a highly significant result, for the Conservative dataset I identified a marginally significant enrichment on the sex chromosomes, X and Y, when normalized by gene density (Figure 15). The sex chromosomes have a z-score of 1.68 for the Conservative dataset. The highest excess was observed on the Y chromosome with 4 RFSD genes (z-score of 3.16 for the Conservative dataset). Chromosome 15 is also marginally enriched for the Conservative dataset with a z-score of 1.84. The Y chromosome is significantly enriched, even though there aren't many Y chromosome RFSD genes because the Y chromosome is extremely gene sparse. The survival of so many Y chromosome genes could be the result of a RFSD mediated survival strategy.

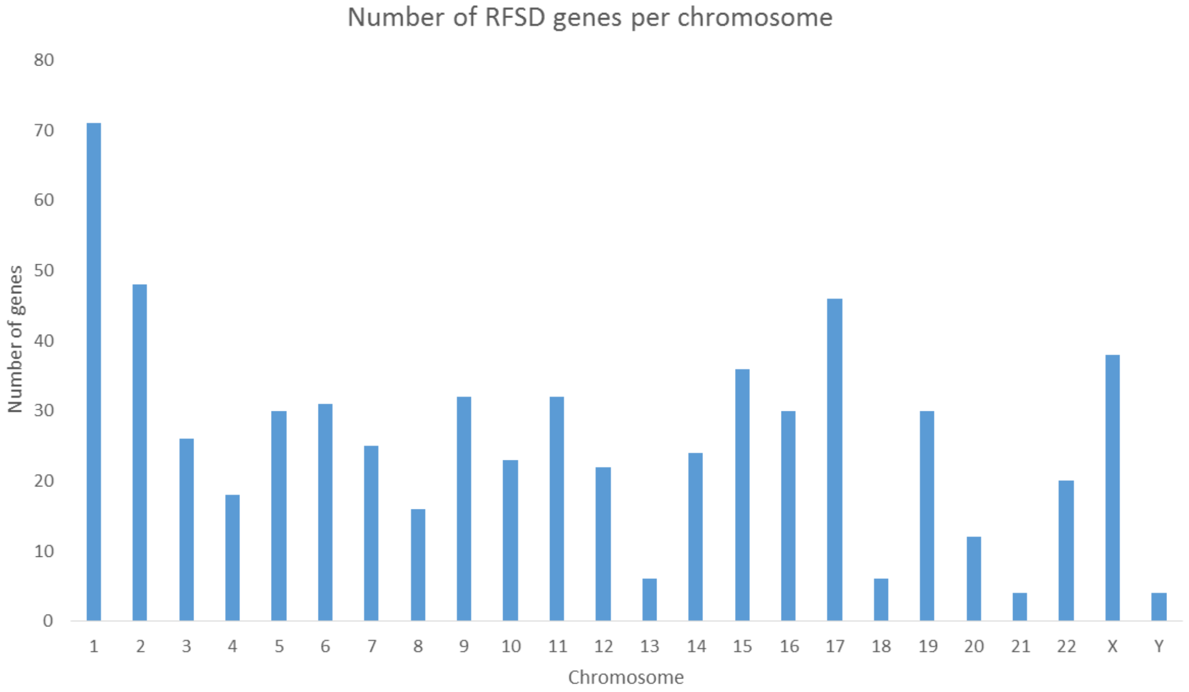


Fig. 15 Bar chart of the number of RFSD genes on each chromosome for the Conservative dataset. Chromosomes 15 and the Y chromosome are marginally significant. The X chromosome is not significant independently of the Y but it is barely below the significance cutoff.

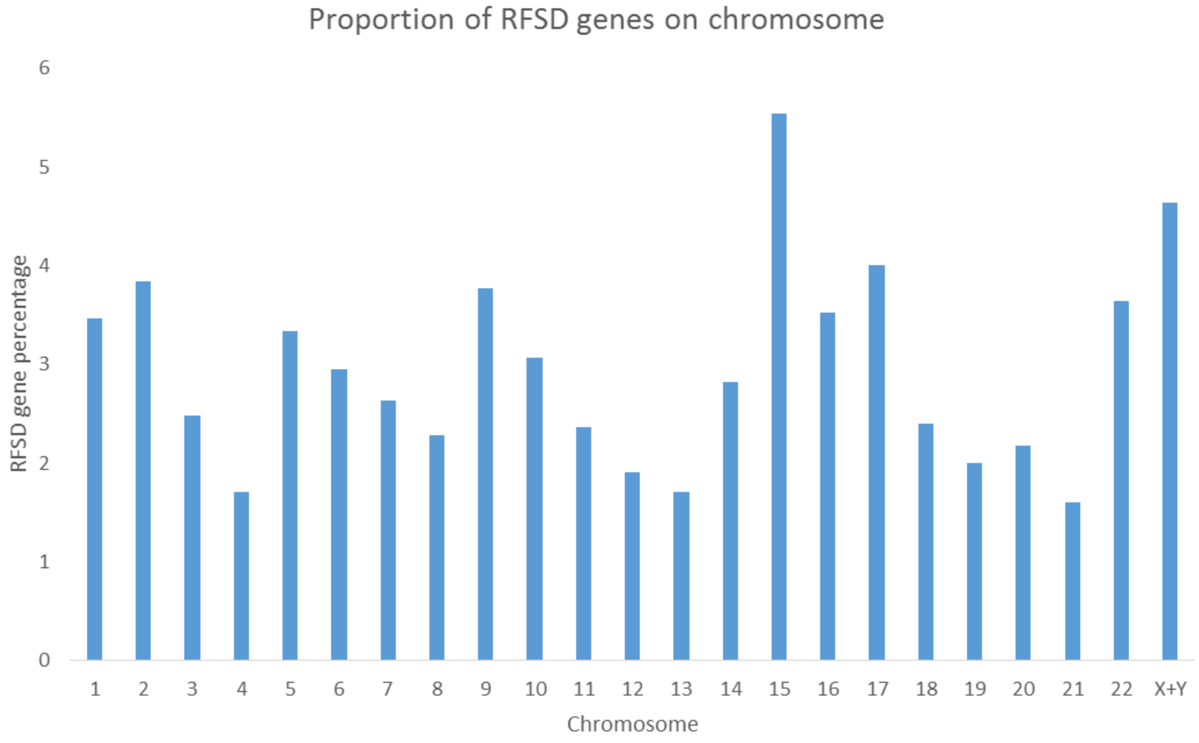


Fig. 16 Bar chart of the proportion of RFSD genes on each chromosome for the Standard dataset normalized by gene density. Chromosomes 15 and the sex chromosomes are marginally significant.

RFSD gene frequency normalized by chromosome length

I ran an analysis where I normalized the RFSD gene frequency by the chromosome length for each chromosome. Gene duplications are known to be proportional to the size of the chromosomes they occur on [116]. When normalized by chromosome length chromosomes 17 and 19 show significant enrichment with z-scores of 2.23 and 3.23 respectively for the Standard dataset (Figure 17). They are also enriched with z-scores of 2.54 and 2.23 respectively for the Conservative dataset. However, chromosomes 17 and 19 are some of the most gene dense autosomes suggesting this is an artefact.

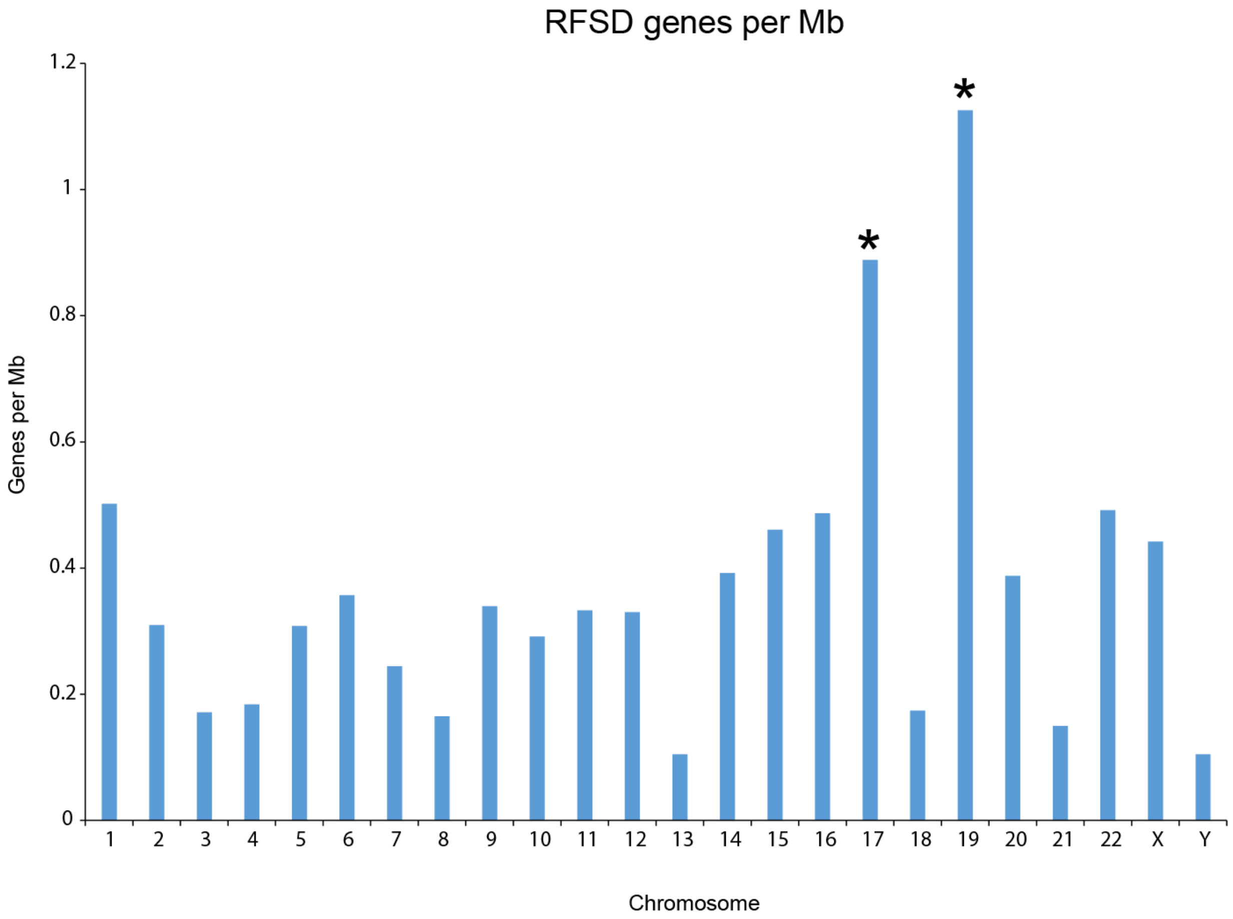


Fig. 17 Bar chart of the number of RFSD genes on each chromosome for the Standard dataset normalized by chromosome length. The asterisks represent significant bars. Chromosomes 17 and 19 are gene dense.

RFSD genes found on the sex chromosomes

A table of all RFSD genes on the sex chromosomes has been compiled below with a summary of their ages, represented by the branch they appear on, and their most common molecular functions. The X chromosome genes are also categorized by the cluster they appear in on the chromosome using the method described by Pandey et al. which uses 12 clusters [119]. The Y chromosome genes are all paired with X chromosome genes apart from two which are paired with each other. The most likely explanation is that the paired X-Y genes are ancestral homologs that predate the evolution of the mammalian sex chromosomes.

Table 8 Summary of identified RFSD genes on the human sex chromosomes.

Gene	Chr.	Branch	Molecular function	X-cluster
ENSG00000129824	Y	0	RNA binding	NA
ENSG00000176679	Y	10	DNA-binding transcription factor activity, RNA polymerase II-specific	NA
ENSG00000114374	Y	2	peptidase activity	NA
ENSG00000172288	Y	2	histone acetyltransferase activity	NA
ENSG00000172352	Y	2	histone acetyltransferase activity	NA
ENSG00000198692	Y	0	RNA binding	NA
ENSG00000005022	X	3	ATP:ADP antiporter activity	3
ENSG00000083750	X	0	GTPase activity	6
ENSG00000102128	X	12	GTPase activity	3
ENSG00000134590	X	9	protein binding	2
ENSG00000147274	X	6	RNA binding	2
ENSG00000153779	X	10	DNA-binding transcription factor activity, RNA polymerase II-specific	4
ENSG00000156925	X	3	DNA-binding transcription factor activity, RNA polymerase II-specific	2
ENSG00000165584	X	9	nucleic acid binding	6
ENSG00000172476	X	12	GTPase activity	3
ENSG00000186787	X	3	methylated histone binding	5
ENSG00000198034	X	0	RNA binding	5
ENSG00000101811	X	2	RNA binding	3
ENSG00000102030	X	1	N-acetyltransferase activity	1

Table 8, continued

Gene	Chr.	Branch	Molecular function	X-cluster
ENSG00000102038	X	1	DNA-binding transcription factor activity, RNA polymerase II-specific	3
ENSG00000102053	X	3	endonuclease activity	5
ENSG00000102144	X	0	kinase activity	4
ENSG00000102359	X	3	signaling receptor binding	3
ENSG00000102362	X	2	Rab GTPase binding	3
ENSG00000122824	X	12	endopolyphosphatase activity	6
ENSG00000124486	X	2	cysteine-type endopeptidase activity	6
ENSG00000125352	X	0	ubiquitin-protein transferase activity	3
ENSG00000125356	X	1	NADH dehydrogenase (ubiquinone) activity	3
ENSG00000130827	X	3	transmembrane signaling receptor activity	1
ENSG00000131264	X	6	DNA-binding transcription factor activity, RNA polymerase II-specific	5
ENSG00000131721	X	12	DNA-binding transcription factor activity, RNA polymerase II-specific	3
ENSG00000134595	X	2	DNA-binding transcription factor activity, RNA polymerase II-specific	2
ENSG00000147059	X	3	methylated histone binding	5
ENSG00000147400	X	9	G-protein beta/gamma-subunit complex binding	1
ENSG00000169239	X	2	carbonate dehydratase activity	7
ENSG00000171478	X	7	hydrolase activity	6
ENSG00000171489	X	7	hydrolase activity	6
ENSG00000177138	X	2	protein binding	7
ENSG00000182583	X	12	chromatin binding	8
ENSG00000182712	X	0	unknown	1
ENSG00000182890	X	0	glutamate dehydrogenase (NAD+) activity	3
ENSG00000183304	X	2	protein binding	7
ENSG00000184083	X	2	RNA binding	6
ENSG00000184388	X	3	mRNA 3'-UTR binding	5
ENSG00000185448	X	13	unknown	6
ENSG00000186288	X	3	mRNA 3'-UTR binding	5
ENSG00000187243	X	9	unknown	6
ENSG00000188021	X	9	polyubiquitin modification-dependent protein binding	5
ENSG00000189132	X	13	unknown	6
ENSG00000196368	X	9	diphosphoinositol-polyphosphate diphosphatase activity	6
ENSG00000196406	X	15	protein binding	2

Table 8, continued

Gene	Chr.	Branch	Molecular function	X-cluster
ENSG00000196767	X	3	RNA polymerase II regulatory region sequence-specific DNA binding	4
ENSG00000197021	X	4	unknown	1
ENSG00000197172	X	13	protein binding	1
ENSG00000197620	X	4	protein binding	1
ENSG00000198021	X	15	protein binding	2
ENSG00000198173	X	13	unknown	6
ENSG00000198354	X	2	protein binding	3
ENSG00000198889	X	13	protein binding	3
ENSG00000203923	X	13	unknown	2
ENSG00000203926	X	15	protein binding	2
ENSG00000203950	X	11	protein binding	2
ENSG00000203989	X	12	DNA-binding transcription factor activity, RNA polymerase II-specific	3
ENSG00000204116	X	1	unknown	5
ENSG00000214827	X	6	protein serine/threonine kinase activator activity	1
ENSG00000221867	X	13	caspase binding	1
ENSG00000241476	X	9	nucleic acid binding	6
ENSG00000154545	X	9	Unknown	6
ENSG00000169059	X	12	Unknown	8
ENSG00000173674	X	0	translation factor activity, RNA binding	7
ENSG00000198814	X	3	ATP binding	6
ENSG00000204363	X	13	Unknown	6
ENSG00000204379	X	9	protein binding	6
ENSG00000204382	X	9	protein binding	6
ENSG00000212747	X	11	protein binding	2

Shared molecular functions between RFSD pairs in the Conservative Dataset

In order to ascertain whether RFSD derived genes inherit their characteristics we compared the molecular functions, biological processes or expression patterns of the genes in each pair. As reported previously, Standard dataset pairs are unlikely to share a molecular function (Figure 8A in Chapter 2). However, for the Conservative dataset the pairs were approximately as likely to share a molecular function as not (Figure 18). This is likely because the Conservative dataset is enriched for larger frameshifts and by extension larger genes. If the dataset is mostly comprised of large genes this could potentially enrich the number of genes with large unframeshifted domains as well, which would produce this result.

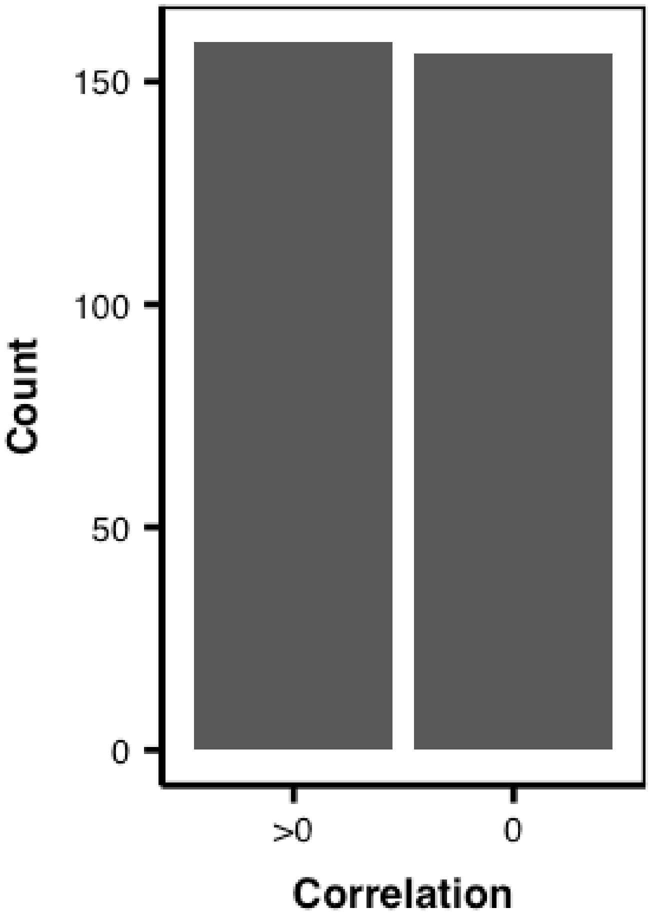


Fig. 18 Bar chart showing the number of RFSD gene pairs that share a molecular function for the Conservative dataset.

Shared biological processes between RFSD pairs in the Conservative Dataset

As expected and consistent with the Standard dataset analysis, we observed that Conservative dataset RFSD gene pairs commonly share a biological process (Figure 19). This is most likely related to the finding that RFSD gene pairs share expression patterns.

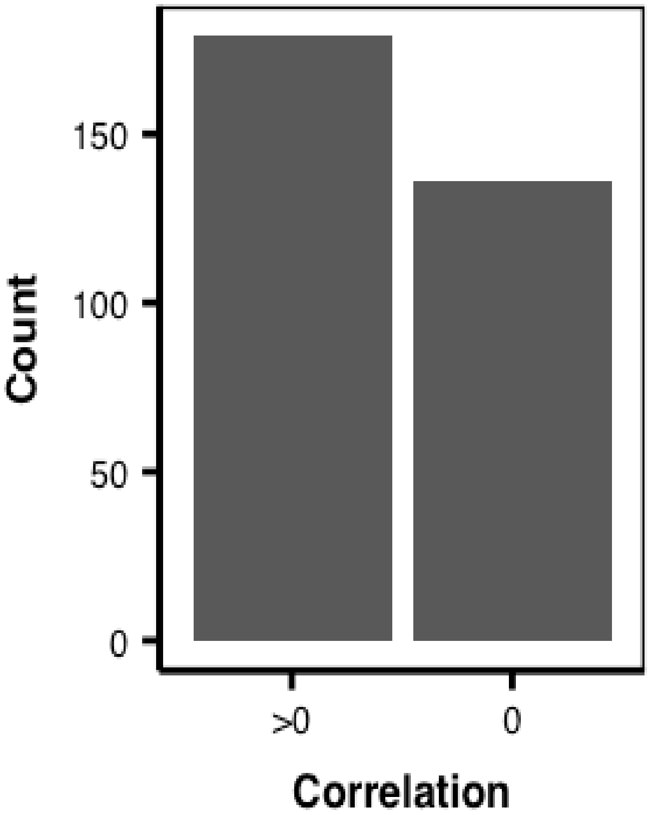


Fig. 19 Bar chart showing the number of RFSD gene pairs that share a biological process for the Conservative dataset.

Shared expression patterns between RFSD pairs in the Conservative Dataset

Conservative dataset RFSD gene pairs also commonly share an expression pattern (Figure 20). This is supported by the Standard dataset analysis and our prior expectations, as regulatory elements are likely to be inherited and function the same way regardless of the frameshift mutation.

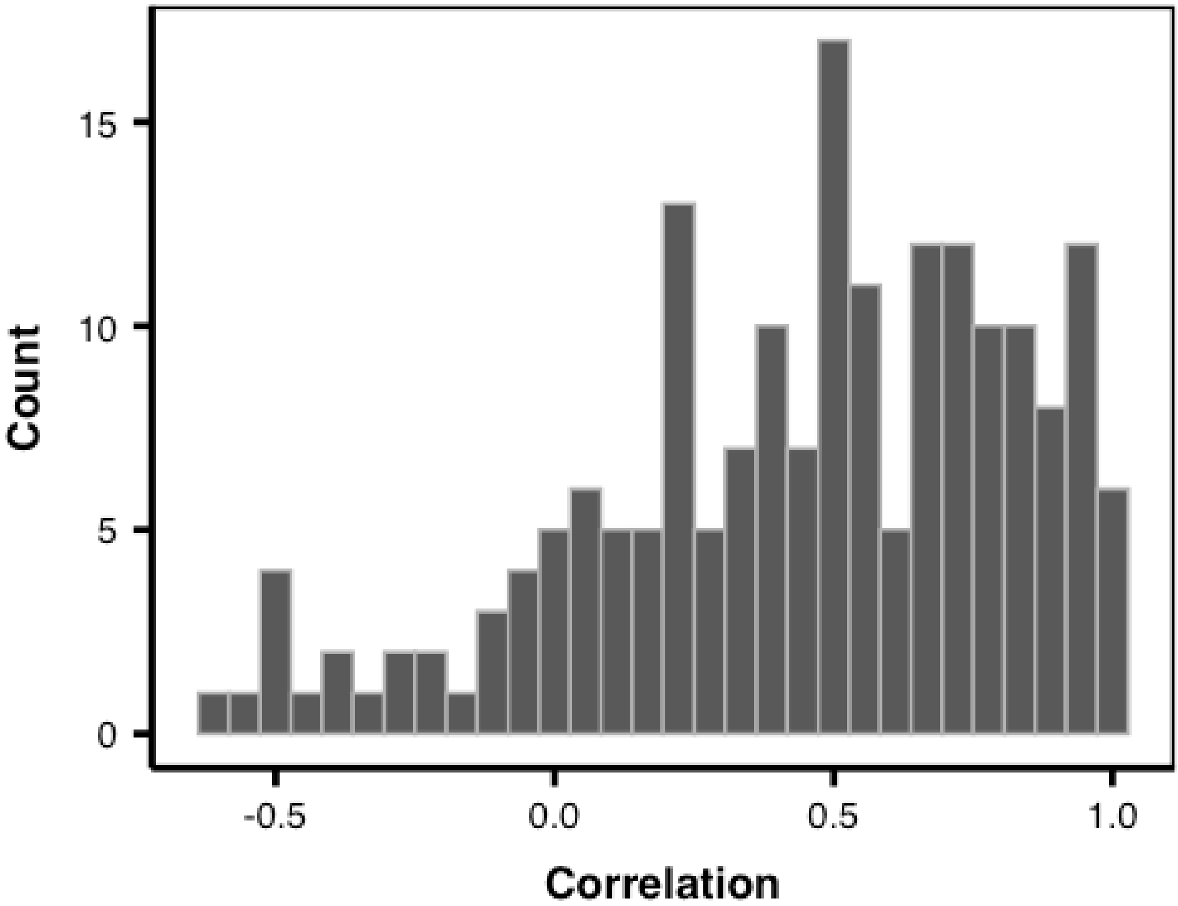


Fig. 20 Histogram showing the degree of similarity in expression pattern between Conservative dataset RFSD pairs.

Combined RFSD gene properties

Observations of the RFSD gene properties were examined pairwise to search for patterns that could inform our understanding of RFS biology and the mechanism of RFSD (Figures 21-27). The resulting figures illustrate two main points. Firstly, the properties of RFSD genes are consistent over time. The spread of all the RFSD gene properties when classified by species branches did not change significantly with the exception of frameshift size. As seen in Figure 22, the size of frameshifted portions of genes is larger in older branches. However, this is likely a result of younger genes being smaller on average than older genes. The smaller younger genes will have smaller frameshifted proportions. The evidence of larger frameshifts is primarily concentrated on genes that existed before the common ancestors of humans and bony fish diverged.

The second finding is that RFSD genes often maintain the domains they inherit from their parent genes as seen in Figures 24 and 27. This is indicated by the grouping of most RFSD pairs around the y=x line in the figures. This may be simply because inheriting more functional domains from a parent gene would give a new gene a greater chance of surviving. The outlier cases may warrant more study as they have clearly diverged from their paired genes.

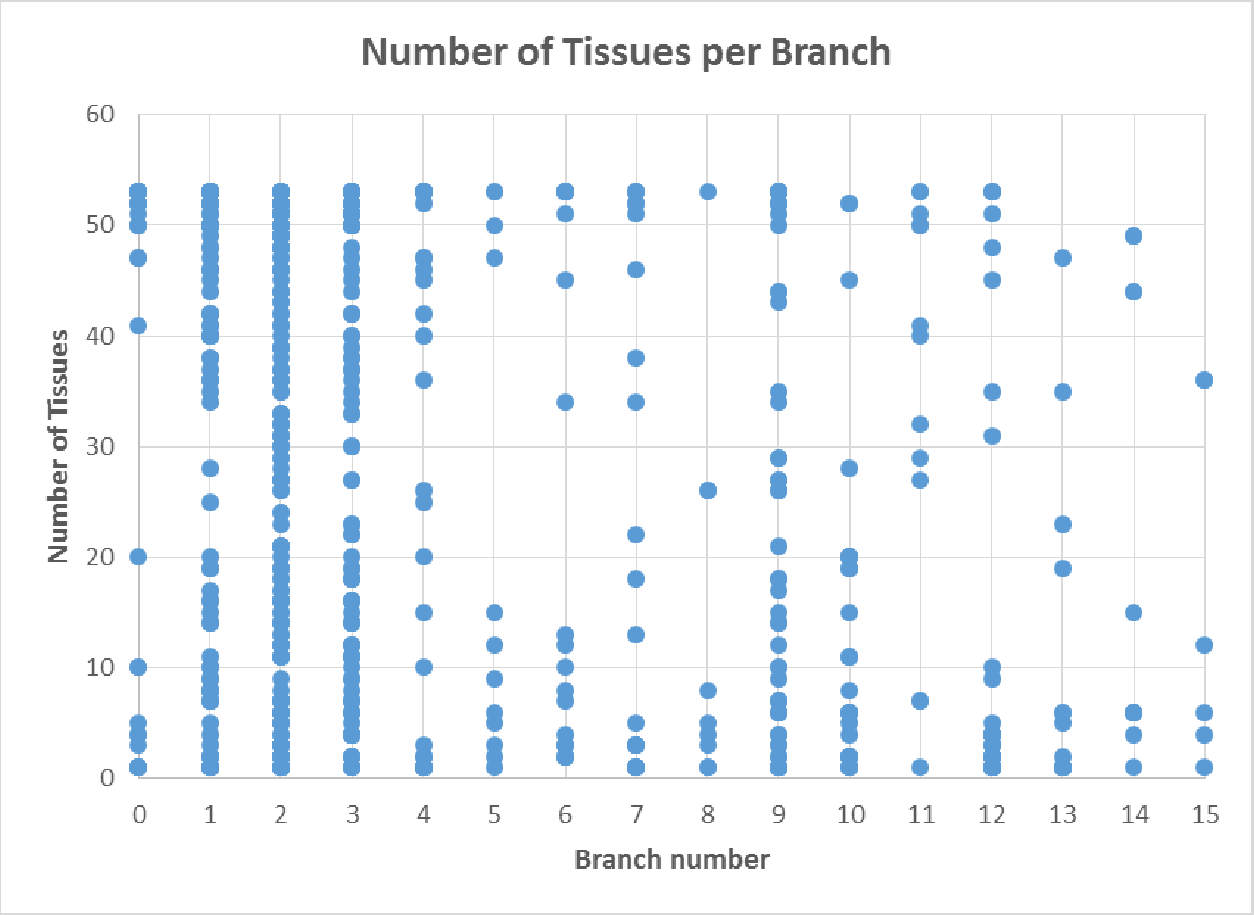


Fig. 21 Scatterplot showing the number of tissues each RFSD gene is expressed in, classified by Branch number.

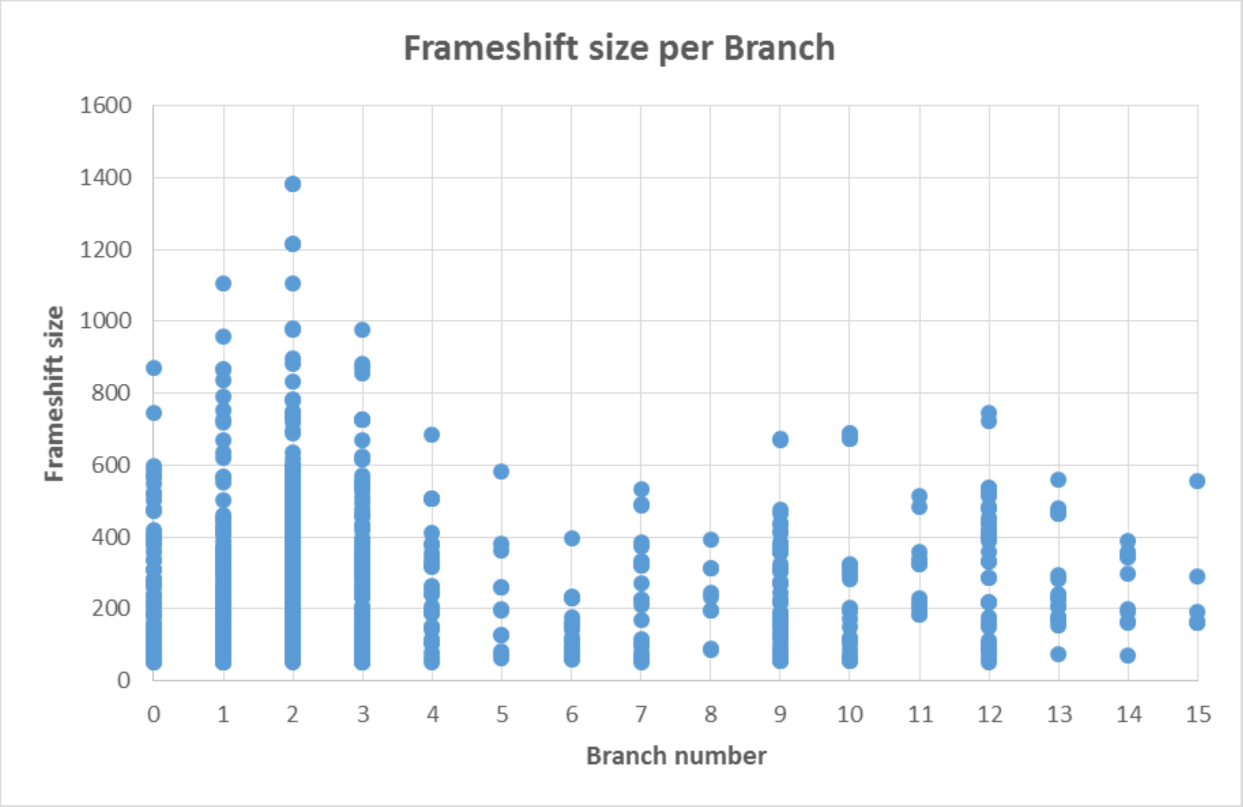


Fig. 22 Scatterplot showing the size of the frameshift for each RFSD gene, classified by Branch number.

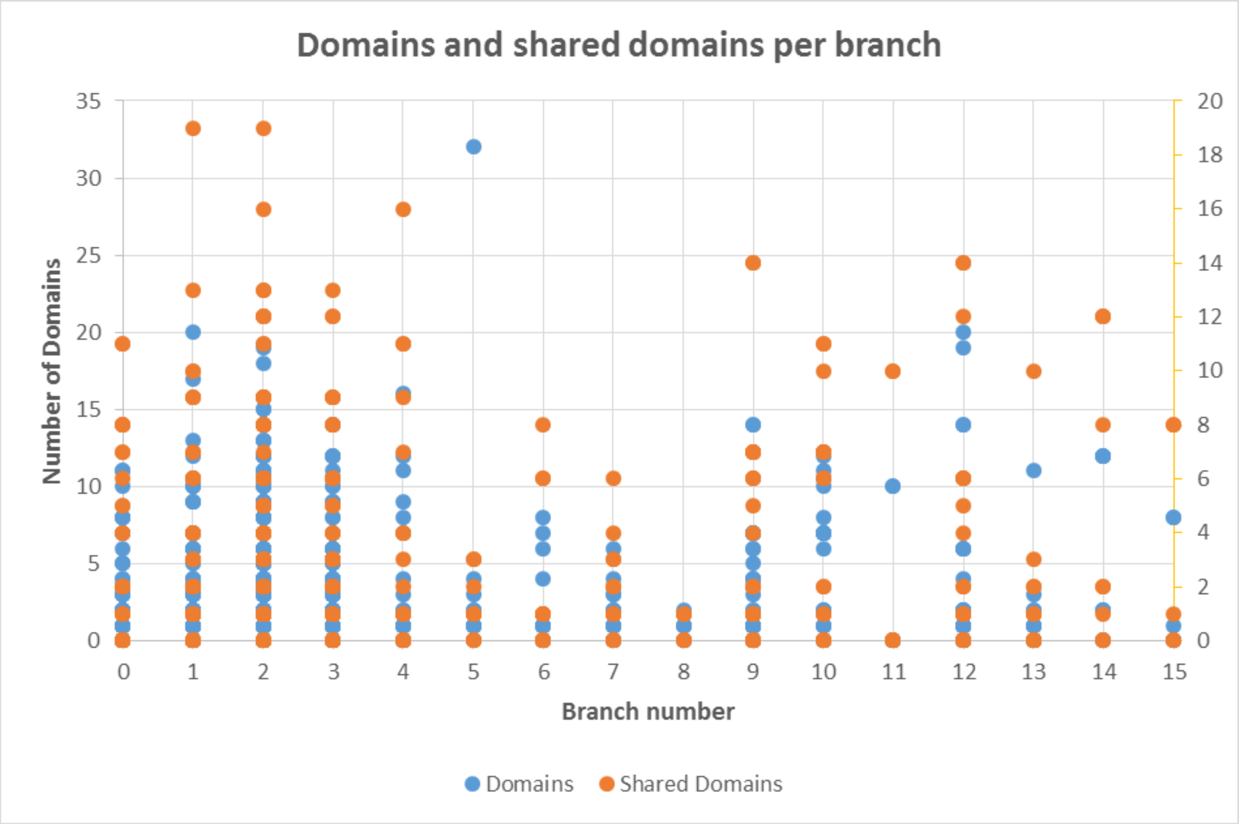


Fig. 23 Scatterplot showing the number of domains for each RFSD gene and shared domains between RFSD gene pairs, classified by Branch number.

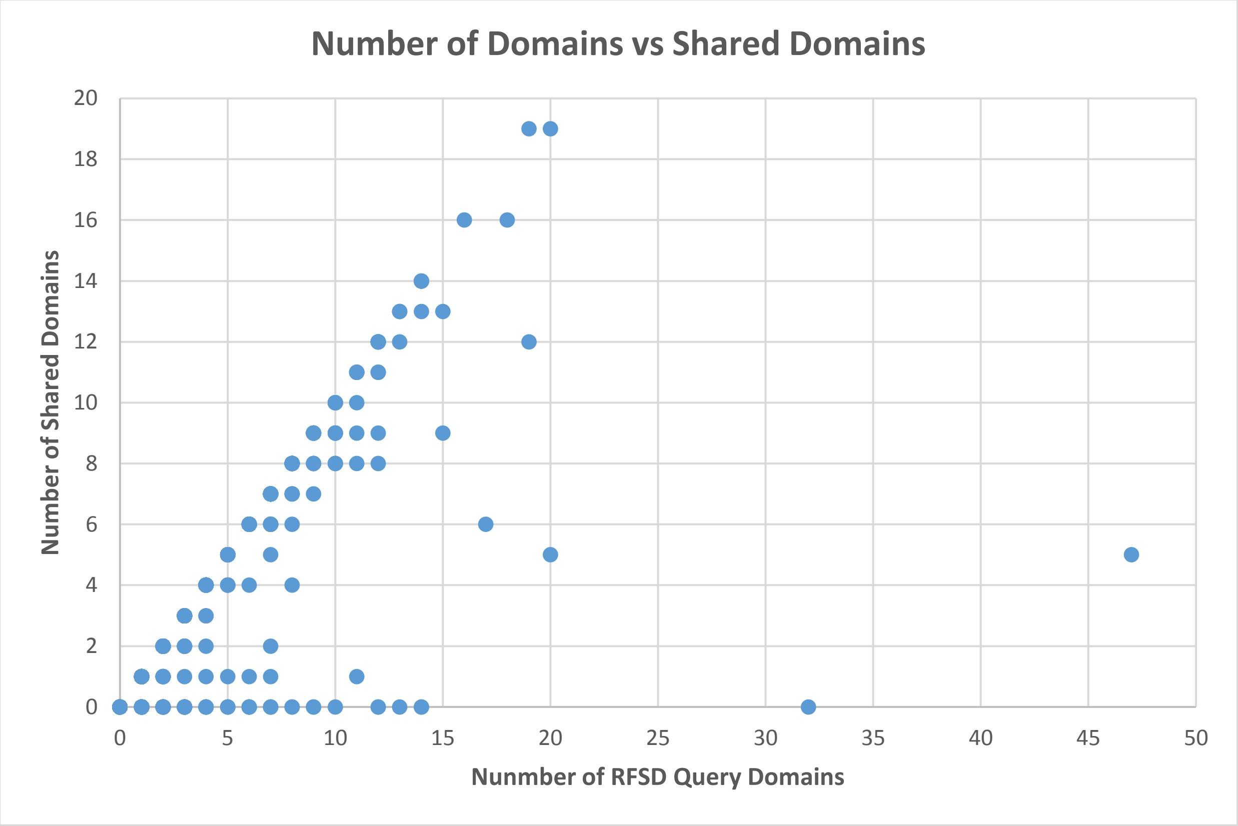


Fig. 24 Scatterplot showing the number of domains for each RFSD gene by the number of shared domains between RFSD gene pairs.

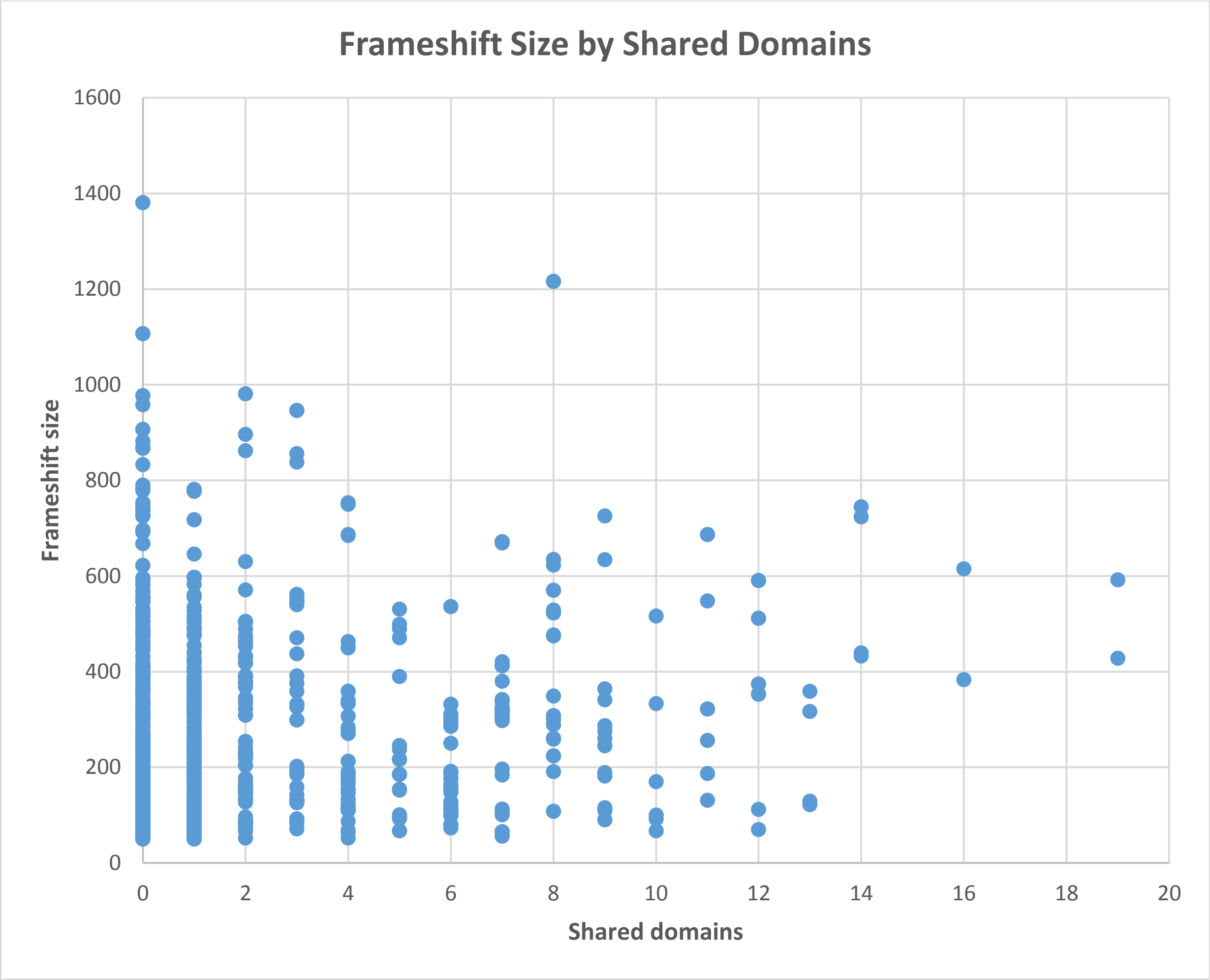


Fig. 25 Scatterplot showing the size of the frameshift for each RFSD gene by the number of shared domains between RFSD gene pairs.

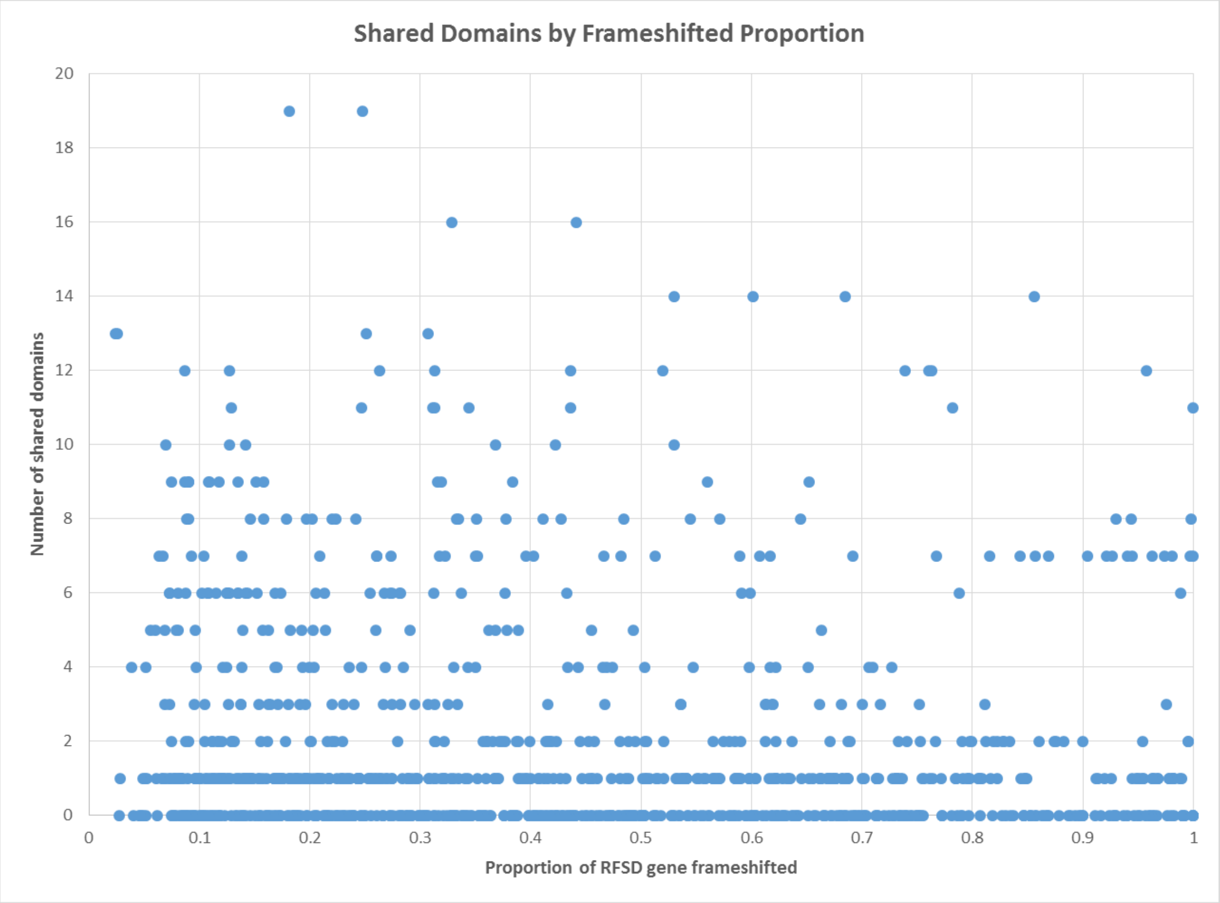


Fig. 26 Scatterplot showing the number of shared domains between each RFSD gene pair by the proportion of the RFSD genes that are frameshifted.

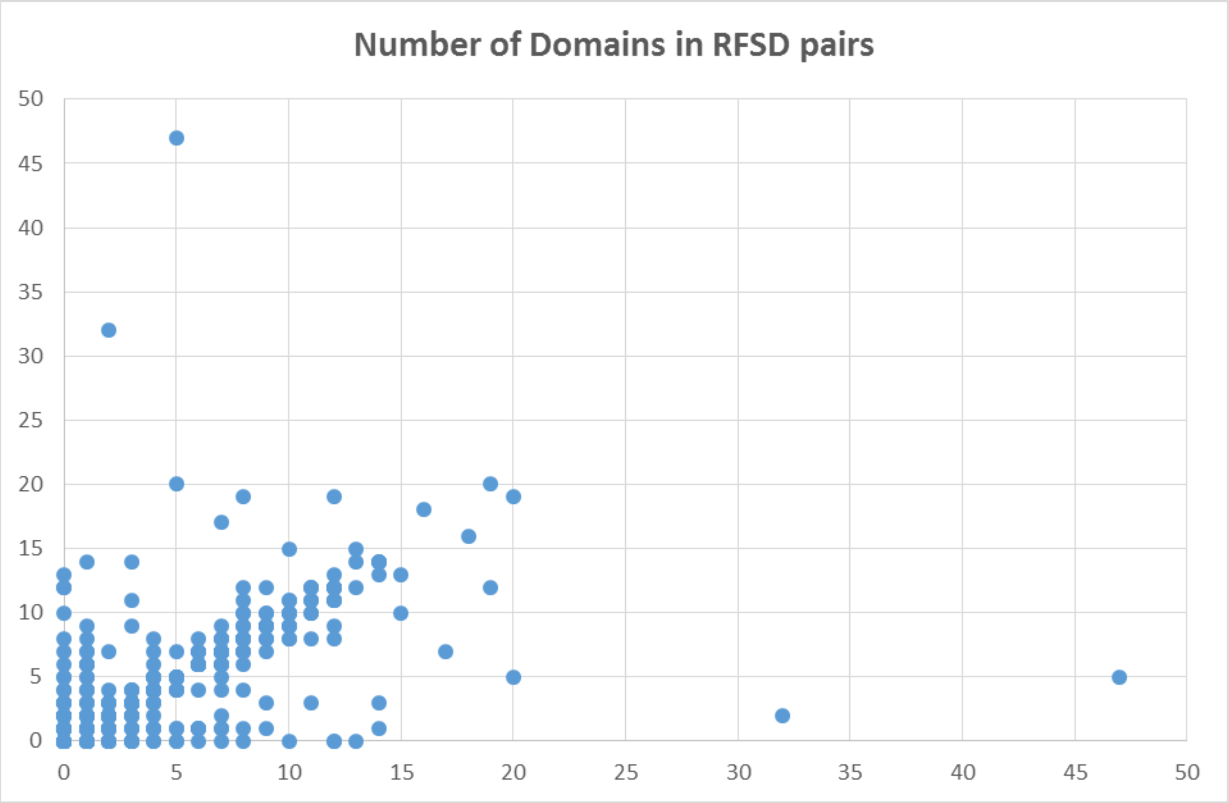


Fig. 27 Scatterplot showing the number of domains for each RFSD gene pair. Since the matched gene pairs are all reciprocal the plot is symmetrical around the y=x line.

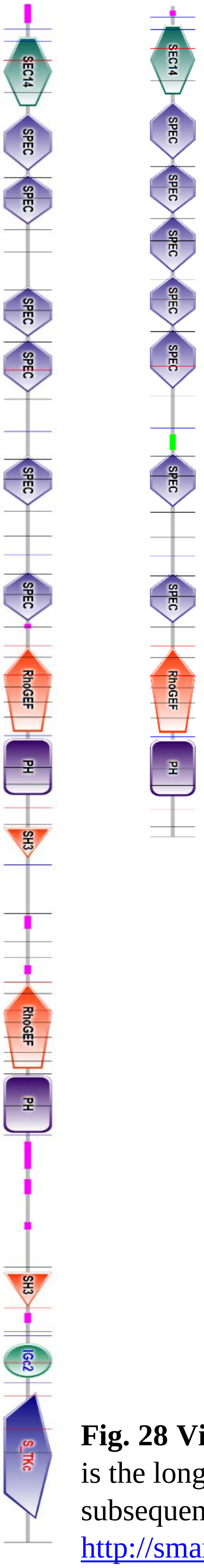


Fig. 28 Visualization of the domains encoded by the genes TRIO and KALRN. TRIO is the longer gene and KALRN is shorter. KALRN was formed via RFSD from TRIO and subsequently diverged. The figure was produced via the SMART tool available at http://smart.embl-heidelberg.de/.

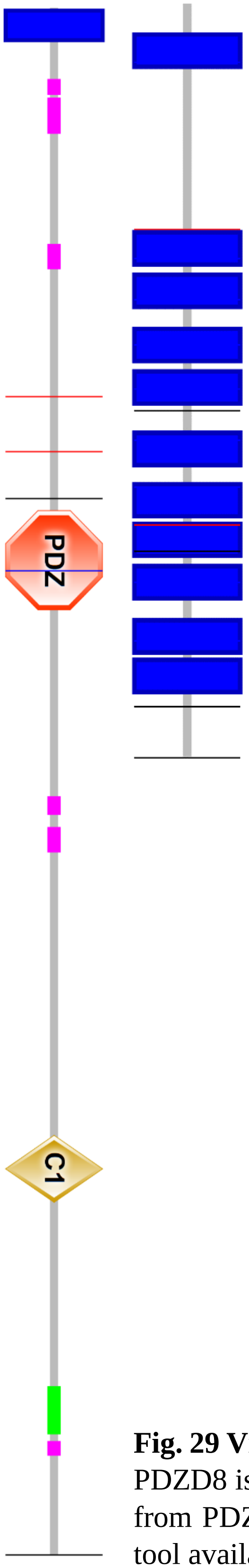


Fig. 29 Visualization of the domains encoded by the genes PDZD8 and SLC18A2. PDZD8 is the longer gene and SLC18A2 is shorter. SLC18A2 was formed via RFSD from PDZD8 and subsequently diverged. The figure was produced via the SMART tool available at http://smart.embl-heidelberg.de/.

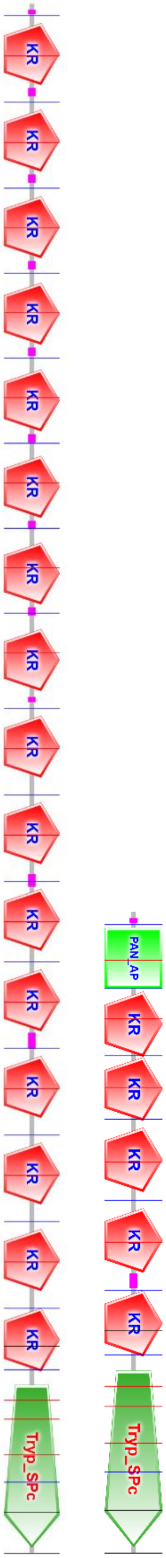


Fig. 30 Visualization of the domains encoded by the genes LPA and PLG. LPA is the longer gene and PLG is shorter. PLG was formed via RFSD from LPA and subsequently diverged. The figure was produced via the SMART tool available at http://smart.embl-heidelberg.de/.

Mitochondrial localization of RFSD genes

Observing functions such as oxygen binding and ATP related functions led me to search the Standard dataset of proteins that localize to the mitochondria, as mentioned in Chapter 2. The 73 genes that were identified as mitochondrial are summarized below in Table 9. For each RFSD gene the table includes the Gene ID, the gene name, the chromosome it is located on and a description of the gene.

Table 9 Summary of identified RFSD genes that are found in the mitochondrial proteome.

Gene Ensembl ID	Gene Name	Chromosome	Gene description
ENSG00000005022	SLC25A5	X	solute carrier family 25 member 5
ENSG00000005156	LIG3	17	DNA ligase 3
ENSG00000083720	OXCT1	5	3-oxoacid CoA-transferase 1
ENSG00000096080	MRPS18A	6	mitochondrial ribosomal protein S18A
ENSG00000102128	RAB40AL	X	RAB40A like
ENSG00000102144	PGK1	X	phosphoglycerate kinase 1
ENSG00000102743	SLC25A15	13	solute carrier family 25 member 15
ENSG00000105649	RAB3A	19	RAB3A, member RAS oncogene family
ENSG00000105819	PMPCB	7	peptidase, mitochondrial processing beta subunit
ENSG00000111639	MRPL51	12	mitochondrial ribosomal protein L51
ENSG00000115042	FAHD2A	2	fumarylacetoacetate hydrolase domain containing 2A
ENSG00000116459	ATP5PB	1	ATP synthase peripheral stalk-membrane subunit b
ENSG00000117009	KMO	1	kynurenine 3-monooxygenase
ENSG00000119673	ACOT2	14	acyl-CoA thioesterase 2
ENSG00000119711	ALDH6A1	14	aldehyde dehydrogenase 6 family member A1
ENSG00000119723	COQ6	14	coenzyme Q6, monooxygenase
ENSG00000120329	SLC25A2	5	solute carrier family 25 member 2
ENSG00000122696	SLC25A51	9	solute carrier family 25 member 51
ENSG00000124172	ATP5F1E	20	ATP synthase F1 subunit epsilon
ENSG00000125356	NDUFA1	X	NADH:ubiquinone oxidoreductase subunit A1

Table 9, continued

ENSG00000125375	ATP5S	14	ATP synthase, H+ transporting, mitochondrial Fo complex subunit s (factor B)
ENSG00000126814	TRMT5	14	tRNA methyltransferase 5
ENSG00000133026	MYH10	17	myosin heavy chain 10
ENSG00000140521	POLG	15	DNA polymerase gamma, catalytic subunit
ENSG00000141437	SLC25A52	18	solute carrier family 25 member 52
ENSG00000144199	FAHD2B	2	fumarylacetoacetate hydrolase domain containing 2B
ENSG00000148672	GLUD1	10	glutamate dehydrogenase 1
ENSG00000151729	SLC25A4	4	solute carrier family 25 member 4
ENSG00000154174	TOMM70	3	translocase of outer mitochondrial membrane 70
ENSG00000160882	CYP11B1	8	cytochrome P450 family 11 subfamily B member 1
ENSG00000162972	MAIP1	2	matrix AAA peptidase interacting protein 1
ENSG00000163319	MRPS18C	4	mitochondrial ribosomal protein S18C
ENSG00000164933	SLC25A32	8	solute carrier family 25 member 32
ENSG00000166411	IDH3A	15	isocitrate dehydrogenase 3 (NAD(+)) alpha
ENSG00000167136	ENDOG	9	endonuclease G
ENSG00000169239	CA5B	X	carbonic anhydrase 5B
ENSG00000170917	NUDT6	4	nudix hydrolase 6
ENSG00000171103	TRMT61B	2	tRNA methyltransferase 61B
ENSG00000171132	PRKCE	2	protein kinase C epsilon
ENSG00000174990	CA5A	16	carbonic anhydrase 5A
ENSG00000175564	UCP3	11	uncoupling protein 3
ENSG00000175567	UCP2	11	uncoupling protein 2
ENSG00000179142	CYP11B2	8	cytochrome P450 family 11 subfamily B member 2
ENSG00000182890	GLUD2	X	glutamate dehydrogenase 2
ENSG00000185236	RAB11B	19	RAB11B, member RAS oncogene family
ENSG00000188611	ASAH2	10	N-acylsphingosine amidohydrolase 2
ENSG00000196475	GK2	4	glycerol kinase 2
ENSG00000197208	SLC22A4	5	solute carrier family 22 member 4
ENSG00000198754	OXCT2	1	3-oxoacid CoA-transferase 2
ENSG00000198814	GK	X	glycerol kinase
ENSG00000204389	HSPA1A	6	heat shock protein family A (Hsp70) member 1A

Table 9, continued

ENSG00000258429	PDF	16	peptide deformylase, mitochondrial
ENSG00000006451	RALA	7	RAS like proto-oncogene A
ENSG00000060971	ACAA1	3	acetyl-CoA acyltransferase 1
ENSG00000099977	DDT	22	D-dopachrome tautomerase
ENSG00000100030	MAPK1	22	mitogen-activated protein kinase 1
ENSG00000102882	MAPK3	16	mitogen-activated protein kinase 3
ENSG00000112208	BAG2	6	BCL2 associated athanogene 2
ENSG00000114374	USP9Y	Y	ubiquitin specific peptidase 9, Y-linked
ENSG00000128245	YWHAH	22	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein eta
ENSG00000134108	ARL8B	3	ADP ribosylation factor like GTPase 8B
ENSG00000136682	CBWD2	2	COBW domain containing 2
ENSG00000157326	DHRS4	14	dehydrogenase/reductase 4
ENSG00000166794	PPIB	15	peptidylprolyl isomerase B
ENSG00000167004	PDIA3	15	protein disulfide isomerase family A member 3
ENSG00000167552	TUBA1A	12	tubulin alpha 1a
ENSG00000168569	TMEM223	11	transmembrane protein 223
ENSG00000182712	CMC4	X	C-X9-C motif containing 4
ENSG00000183569	SERHL2	22	serine hydrolase like 2
ENSG00000184227	ACOT1	14	acyl-CoA thioesterase 1
ENSG00000198610	AKR1C4	10	aldo-keto reductase family 1 member C4
ENSG00000203859	HSD3B2	1	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2
ENSG00000214827	MTCP1	X	mature T cell proliferation 1

Table 10 Names of identified RFSD genes for which mouse knockout resources are available.

ADCY5	EN2	IGF2BP1	PDE6B	SUGP1
ADCY6	ENDOG	IL16	PDF	TBPL1
ARID1A	FAM184B	IRX3	PELI1	THBS1
BCL9L	FASN	KCNH8	PLXNA3	THBS2
BICD1	FOXD4	KPNA4	PRKAB1	TRIO
BNC2	FZD2	KPNA6	RAC2	TSHZ1
BTBD2	GLUD1	KRT14	RAC3	TUBA1A
C1QL3	GNAI3	LIN7C	RANBP10	TUBB1
CACNG2	GNPTG	LRRC8C	RBMX	TUBB2B
CAMK1D	GPR39	MAB21L2	RHOF	UCP2
CELA2A	GPRC5B	MAGOHB	RHOJ	VAX1
CLCNKB	GRHL1	MRGPRX1	RHOQ	WDR5
CNNM2	GRIN2C	NAA10	RPS4X	WNT2B
CRYGC	GRK3	NHLH1	RSPH9	ZIC1
CSNK1A1	H3F3B	NHLH2	SLC22A5	ZNRF1
CSTF2T	HIRIP3	OPN3	SLC23A1	ZNRF2
CYP11B2	HOXA10	PCDHB2	SMR3A	ZSWIM5
DDT	HOXB1	PDE11A	SNAI1	
DDX19B	HS3ST3A1	PDE11A	SNAI2	
DDX54	HS3ST3B1	PDE6A	SNX22	

References

[1] C. Darwin, On the Origin of Species, London: John Murray, 1859.

[2] A. R. Wallace, "On the law which has regulated the introduction of new species," *Annals and Magazine of Natural History,* vol. 16, p. 184–196, 1855.

[3] S. Ohno, Evolution by Gene Duplication, New York: Springer-Verlag New York Inc., 1970.

[4] K. D. Rose and T. M. Bown, "Gradual phyletic evolution at the generic level in early Eocene omomyid primates," *Nature,* vol. 309, pp. 250-252, 1984.

[5] B. A. Malmgren and J. P. Kennett, "Phyletic gradualism in a Late Cenozoic planktonic foraminiferal lineage; DSDP Site 284, southwest Pacific," *Paleobiology,* vol. 7, no. 2, pp. 230-240, 1981.

[6] N. Eldredge and S. J. Gould, "Punctuated equilibria: an alternative to phyletic gradualism," *Models in Paleobiology* , pp. 82-115, 1972.

[7] S. J. Gould and N. Eldredge, "Punctuated equilibrium comes of age," *Nature,* vol. 366, p. 223–227, 1993.

[8] S. M. Stanley, "A theory of evolution above the species level," *PNAS,* vol. 72, no. 2, pp. 646-650, 1975.

[9] S. M. Stanley, "Rates of evolution," *Paleobiology,* vol. 11, no. 1, pp. 13-26, 1985.

[10] B. Charlesworth, R. Lande and M. Slatkin, "A Neo-Darwinian Commentary on Macroevolution," *Evolution,* vol. 36, no. 3, pp. 474-498, 1982.

[11] C. M. Newman, J. E. Cohen and C. Kipnis, "Neo-darwinian evolution implies punctuated equilibria," *Nature,* vol. 315, no. 6018, pp. 400-401, 1985.

[12] J. S. Levinton and M. S. Chris, "A Critique of the Punctuated Equilibria Model and Implications for the Detection of Speciation in the Fossil Record," *Systematic Biology,* vol. 29, no. 2, p. 130–142, 1980.

[13] G. L. Stebbins and F. J. Ayala, "Is a New Evolutionary Synthesis Necessary?," *Science,* vol. 213, no. 4511, pp. 967-971, 1981.

[14] D. Noble, "Evolution beyond neo-Darwinism: a new conceptual framework," *Journal of Experimental Biology,* vol. 218, pp. 7-13, 2015.

[15] W. C. H. Cross, T. A. Graham and N. A. Wright, "New paradigms in clonal evolution: punctuated equilibrium in cancer," *Journal of Pathology,* vol. 240, no. 2, pp. 126-136, 2016.

[16] M. Ricci, V. Peona, E. Guichard, C. Taccioli and A. Boattini, "Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals," *Journal of Molecular Evolution,* vol. 86, no. 5, p. 303–310, 2018.

[17] R. Lande, "Evolution of phenotypic plasticity in colonizing species," *Molecular Ecology,* vol. 24, no. 9, p. 2038–2045, 2015.

[18] S. Chen, B. H. Krinsky and M. Long, "New genes as drivers of phenotypic evolution.," *Nature Reviews Genetics,* vol. 14, p. 645–660, 2013.

[19] J. A. Lee, C. M. B. Carvalho and J. R. Lupski, "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders," *Cell,* vol. 131, no. 7, pp. 1235-1247, 2007.

[20] P. W. Messer and P. F. Arndt, "The Majority of Recent Short DNA Insertions in the Human Genome Are Tandem Duplications," *Molecular Biology and Evolution,* vol. 24, no. 5, p. 1190–1197, 2007.

[21] S. Newman, K. E. Hermetz, B. Weckselblatt and M. K. Rudd, "Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints," *American Journal of Human Genetics,* vol. 96, no. 2, pp. 208-220, 2015.

[22] A. B. Reams and J. R. Roth, "Mechanisms of Gene Duplication and Amplification," *Cold Spring Harbor Perspectives in Biology,* vol. 7, no. 2, 2015.

[23] F. N. Carelli, T. Hayakawa, Y. Go, H. Imai, M. Warnefors and H. Kaessmann, "The life history of retrocopies illuminates the evolution of new mammalian genes," *Genome Research,* vol. 26, pp. 301-314, 2016.

[24] F. C. P. Navarro and P. A. F. Galante, "A Genome-Wide Landscape of Retrocopies in Primate Genomes," *Genome Biology and Evolution,* vol. 7, no. 8, p. 2265–2275, 2015.

[25] D. Jangam, C. Feschotte and E. Betrán, "Transposable Element Domestication As an Adaptation to Evolutionary Conflicts," *Trends in Genetics,* vol. 33, no. 11, pp. 817-831, 2017.

[26] Huff, J. T., D. Zilberman and S. W. Roy, "Mechanism for DNA transposons to generate introns on genomic scales," *Nature,* vol. 538, p. 533–536 , 2016.

[27] R. A. Elbarbary, B. A. Lucas and L. E. Maquat, "Retrotransposons as regulators of gene expression," *Science,* vol. 351, no. 6274, p. aac7247, 2016.

[28] J. R. Roth, "Frameshift mutations," *Annual Review of Genetics,* vol. 8, pp. 319-346, 1974.

[29] J. Hu and P. C. Ng, "Predicting the effects of frameshifting indels," *Genome Biology,* vol. 13, no. 2, p. R9, 2012.

[30] K. A. Geiler-Samerotte, M. F. Dion, B. A. Budnik, S. M. Wang, D. L. Hartl and D. A. Drummond, "Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast," *Proc. Nat'l. Acad. Sci. USA,* vol. 108, no. 2, pp. 680-685, 2011.

[31] D. A. Drummond and C. O. Wilke, "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution," *Cell,* vol. 134, no. 2, pp. 341-352, 2008.

[32] N. M. Wills and J. F. Atkins, "The potential role of ribosomal frameshifting in generating aberrant proteins implicated in neurodegenerative diseases," *RNA,* vol. 12, pp. 1149-1153, 2006.

[33] R. K. Yuen, B. Adhami-Mojarad, I. Backstrom, A. Yin and T. Soman, "Whole-genome sequencing identified a frameshift mutation at LMNB1 in a family with early-onset dystonia," *Canadian Journal of Neurological Sciences,* vol. 46, no. s1, p. s28, 2019.

[34] C. B. Jackson, D. Hahn, B. Schröter, U. Richter, B. J. Battersby, T. Schmitt-Mechelke, P. Marttinen, J.-M. Nuoffer and A. Schaller, "A novel mitochondrial ATP6 frameshift mutation causing isolated complex V deficiency, ataxia and encephalomyopathy," *European Journal of Medical Genetics,* vol. 60, no. 6, pp. 345-351, 2017.

[35] U. Schwarze, T. Cundy, Y. J. Liu, P. L. Hofman and P. H. Byers, "Compound heterozygosity for a frameshift mutation and an upstream deletion that reduces expression of SERPINH1 in siblings with a moderate form of osteogenesis imperfecta," *American Journal of Medical Genetics,* vol. 179, no. 8, pp. 1466-1475, 2019.

[36] X. Wang, H. Jin, F. Han, Y. Cui, J. Chen, C. Yang, P. Zhu, W. Wang, G. Jiao, W. Wang, C. Hao and Z. Gao, "Homozygous DNAH1 frameshift mutation causes multiple morphological anomalies of the sperm flagella in Chinese," *Clinical Genetics,* vol. 91, no. 2, p. 313–321., 2017.

[37] X. Wang, Y. Liang, Z. Zhu, Y. Zhu, P. Li, J. Cao, Q. Zhang, Q. Liu and Z. Li, "A de novo frameshift mutation of the SRY gene leading to a patient with 46,XY complete gonadal dysgenesis," *Asian Journal of Andrology,* vol. 21, no. 5, p. 522–524., 2019.

[38] D. L. Huseby, G. Brandis, L. P. Alzrigat and D. Hughes, "Antibiotic resistance by high-level intrinsic suppression of a frameshift mutation in an essential gene," *PNAS,* vol. 117, no. 6, pp. 3185-3191, 2020.

[39] M. Kondo, H. Hirai, T. Furukawa, Y. Yoshida, A. Suzuki, T. Kawaguchi and F.-S. Che, "Frameshift Mutation Confers Function as Virulence Factor to Leucine-Rich Repeat Protein from Acidovorax avenae," *Frontiers in Plant Science,* vol. 7, p. 1988, 2017.

[40] C. Chandrasekaran and E. Betrán, "Origins of New Genes and Pseudogenes," *Nature Education,* vol. 1, no. 1, p. 181, 2008.

[41] G. Streisinger, Y. Okada, J. Emrich, J. Newton, E. T. A. Tsugita1 and M. Inouye, "Frameshift Mutations and the Genetic Code," *Cold Spring Harbor Symposia on Quantitative Biology,* vol. 31, pp. 77-84, 1966.

[42] J. Raes and Y. Van de Peer, "Functional divergence of proteins through frameshift mutations," *Trends in Genetics,* vol. 21, no. 8, pp. 428-431, 2005.

[43] K. Okamura, L. Feuk, T. Marques-Bonet, A. Navarro and S. W. Scherer, "Frequent appearance of novel protein-coding sequences by frameshift translation.," *Genomics,* p. 690=697, 2006.

[44] K. K. Q. Nguyen, Y. K. Gomez, M. Bakhom, A. Radcliffe, P. La, D. Rochelle, J. W. Lee and E. J. Sorin, "Ensemble simulations: folding, unfolding and misfolding of a high-efficiency frameshifting RNA pseudoknot.," *Nucleic Acids Research,* vol. 45, no. 8, pp. 4893 - 4904, 2017.

[45] X. Wang, Q. Dong, G. Chen, J. Zhang and Y. Liu, "Frameshifts and wild-type protein sequences are always similar because the genetic code is nearly optimal for frameshift tolerance," *bioRxiv,* 2019.

[46] X. Wang et al., "The universal genetic code, protein coding genes and genomes of all species were optimized for frameshift tolerance," *bioRxiv,* 2016.

[47] D. O. S. Hatfield, "The where, what and how of ribosomal frameshifting in retroviral protein synthesis," *Trends in Biochemical Sciences,* vol. 15, no. 5, pp. 186 - 190, 1990.

[48] D. Brégeon, V. Colot, M. Radman and F. Taddei, "Translational misreading: a tRNA modification counteracts a +2 ribosomal frameshift," *Genes & Development,* vol. 15, pp. 2295 - 2306 , 2001.

[49] O. Namy, S. J. Moran, D. I. Stuart, R. J. C. Gilbert and I. Brierley, "A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting," *Nature,* vol. 441, p. 244–247, 2006.

[50] I. Brierley, S. Pennell and R. J. C. Gilbert, "Viral RNA pseudoknots: versatile motifs in gene expression and replication," *Nature Reviews Microbiology,* vol. 5, p. 598–610, 2007.

[51] W. F. Waas, Z. Druzina, M. Hanan and P. Schimmel, "Role of a tRNA Base Modification and Its Precursors in Frameshifting in Eukaryotes," *The Journal of Biological Chemistry,* vol. 282, no. 36, p. 26026 – 26034, 2007.

[52] J. F. Atkins, G. Loughran, P. R. Bhatt, A. E. Firth and P. V. Baranov, "Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use," *Nucleic Acids Research,* vol. 44, no. 15, p. 7007–7078, 2016.

[53] F. Tuorto and F. Lyko, "Genome recoding by tRNA modification," *Open Biology,* vol. 6, no. 12, 2016.

[54] H. J. Muller, "Bar duplication," *Science,* vol. 83, no. 2161, pp. 528-530, 1936.

[55] L. Sandegren and D. I. Andersson, "Bacterial gene amplification: Implications for the evolution of antibiotic resistance," *Nature Reviews Microbiology,* vol. 7, p. 578–588, 2009.

[56] D. Romero and R. Palacios, "Gene amplification and genomic plasticity in prokaryotes," *Annual Review of Genetics,* vol. 31, p. 91–111, 1997.

[57] D. I. Andersson and D. Hughes, "Gene amplification and adaptive evolution in bacteria," *Annual Review of Genetics,* vol. 43, pp. 167-195, 2009.

[58] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics,* vol. 151, no. 4, pp. 1531-1545, 1999.

[59] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science,* vol. 290, pp. 1151-1155, 2000.

[60] M. Lynch and A. Force, "The probability of duplicate gene preservation by subfunctionalization," *Genetics,* vol. 154, p. 459–473, 2000.

[61] U. Bergthorsson, D. I. Andersson and J. R. Roth, "Ohno's dilemma: Evolution of new genes under continuous selection," *Proceedings of the National Academy of Sciences,* vol. 104, p. 17004–17009, 2007.

[62] M. E. Pettersson, D. I. Andersson, J. R. Roth and O. G. Berg, "The Amplification Model for Adaptive Mutation," *Genetics,* vol. 169, no. 2, pp. 1105-1115, 2005.

[63] D. I. Andersson, J. Jerlström-Hultqvist and J. Näsvall, "Evolution of New Functions De Novo and from Preexisting Genes," *Cold Spring Harbor Perspectives in Biology,* 2015.

[64] M. P. Francino, "An adaptive radiation model for the origin of new gene functions," *Nature Genetics,* vol. 37, p. 573–577, 2005.

[65] C. Deng, C.-H. Cheng, H. Ye, X. He and L. Chen, "Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict," *Proceedings of the National Academy of Sciences,* vol. 107, no. 50, p. 21593–21598, 2010.

[66] M. Long, E. Betrán, K. Thornton and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics,* vol. 4, no. 11, pp. 865-875, 2003.

[67] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Research,* vol. 20, no. 10, p. 1313–1326, 2010.

[68] M. Long, N. W. VanKuren, S. Chen and M. D. Vibranovski, "New Gene Evolution: Little Did We Know," *Annual Review of Genetics,* vol. 47, p. 307–333, 2013.

[69] W. Zhang, P. Landback, A. R. Gschwend, B. Shen and M. Long, "New genes drive the evolution of gene interaction networks in the human and mouse genomes," *Genome Biology,* 2015.

[70] J. T. Marinko, H. Huang, W. D. Penn, J. A. Capra, J. P. Schlebach and C. R. Sanders, "Folding and Misfolding of Human Membrane Proteins in Health and Disease: From Single Molecules to Cellular Proteostasis," *Chemical Reviews,* vol. 119, no. 9, pp. 5537 - 5606, 2019.

[71] S. Xue, M. D. Jones, Q. Lu, J. M. Middeldorp and B. E. Griffin, "Genetic Diversity : Frameshift Mechanisms Alter Coding of a Gene (Epstein-Barr Virus LF3 Gene) That Contains Multiple 102-Base-Pair Direct Sequence Repeats.," *Molecular and Cellular Biology.,* vol. 23, no. 6, p. 2192–2201, 2003.

[72] M. Vandenbussche, G. Theissen, Y. Van de Peer and T. Gerats, "Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations," *Nucleic Acids Research,* vol. 31, no. 15, p. 4401–4409, 2003.

[73] Ballester, B. et al, "Biomart," Biomart, [Online]. Available: http://www.biomart.org/.

[74] B. Haferkamp, H. Zhang, S. Kissinger, X. Wang, Y. Lin, M. Schultz and J. Xiang, "BaxΔ2 Family Alternative Splicing Salvages Bax Microsatellite-Frameshift Mutations," *Genes & Cancer,* vol. 4, no. 11-12, pp. 501-512, 2013.

[75] S. Durinck, P. Spellman, E. Birney and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nature Protocols,* vol. 4, p. 1184–1191. , 2009.

[76] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma and W. Huber, "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis," *Bioinformatics,* vol. 21, p. 3439–3440, 2005.

[77] L. W. Parfrey, D. J. G. Lahr, A. H. Knoll and L. A. Katz, "Estimating the timing of early eukaryotic diversification with multigene molecular clocks," *PNAS,* vol. 108, no. 33, pp. 13624-13629, 2011.

[78] M. dos Reis, Y. Thawornwattana, K. Angelis, M. J. Telford, P. Donoghue and Z. Yang, "Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales.," *Current Biology,* vol. 25, no. 22, pp. 2939-2950, 2015.

[79] M. L. Berbee and J. W. Taylor, "Dating the molecular clock in fungi – how close are we?," *Fungal Biology Reviews,* vol. 24, no. 1-2, pp. 1-16, 2010.

[80] Betancur-R. R. et al, "The tree of life and a new classification of bony fishes.," *PLoS Currents,* vol. 5, 2013.

[81] M. dos Reis, J. Inoue, M. Hasegawa, R. J. Asher, P. C. Donoghue and Z. Yang, "Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny.," *Proceedings. Biological Sciences.,* vol. 279, no. 1742, pp. 3491-3500, 2012.

[82] S. Kumar, G. Stecher, M. Suleski and S. Hedges, "TimeTree: a resource for timelines, timetrees, and divergence times," *Molecular Biology and Evolution,* vol. 34, pp. 1812-1819, 2017.

[83] G. Yu, L. Wang, Y. Han and Q. He, "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters," *OMICS,* vol. 16, no. 5, p. 284–287, 2012.

[84] Genotype-Tissue Expression Project, "GTEx Portal," GTEx Consortium, [Online]. Available: https://gtexportal.org/home/datasets.

[85] "NIH U.S. National Library of Medicine: Genetics Home Reference," NIH, 6 August 2019. [Online]. Available: https://ghr.nlm.nih.gov/chromosome.

[86] A. C. Smith and A. J. Robinson, "MitoMiner v3.1, an update on the mitochondrial proteomics database," *Nucleic Acids Research,* vol. 44, pp. 1258-1261, 2016.

[87] A. C. Smith and A. J. Robinson, "MitoMiner, an integrated database for the storage and analysis of mitochondrial proteomics data," *Molecular & Cellular Proteomics,* vol. 8, no. 6, pp. 1324-1337, 2009.

[88] A. C. Smith, J. A. Blackshaw and A. J. Robinson, "MitoMiner: a data warehouse for mitochondrial proteomics data," *Nucleic Acids Research*, vol. 40, pp. 1160-1167, 2012.

[89] A. C. Smith and A. J. Robinson, "MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases," *Nucleic Acids Research,* vol. 47, p. 1225–1228, 2018.

[90] W. J. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder and S. J. O'Brien, "Molecular phylogenetics and the origins of placental mammals," *Nature,* vol. 409, pp. 614-618, 2001.

[91] Meredith, R. W. et al., "Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification," *Science,* vol. 334, no. 6055, pp. 521-524, 2011.

[92] Z. Luo, "Transformation and diversification in early mammal evolution," *Nature,* vol. 450, p. 1011–1019, 2007.

[93] S. Kumar and S. B. Hedges, "A molecular timescale for vertebrate evolution," *Nature,* vol. 392, p. 917–920, 1998.

[94] S. B. Hedges, P. H. Parker, C. G. Sibley and S. Kumar, "Continental breakup and the ordinal diversification of birds and mammals," *Nature,* vol. 381, p. 226–229, 1996.

[95] O'Leary et al., "The Placental Mammal Ancestor and the Post–K-Pg Radiation of Placentals," *Science,* vol. 339, no. 6120, pp. 662-667, 2013.

[96] M. J. Vavrek, "The fragmentation of Pangaea and Mesozoic terrestrial vertebrate biodiversity," *Biology Letters,* vol. 12, no. 9, 2016.

[97] T. L. B. M. J. Stubbs, "Ecomorphological diversifications of Mesozoic marine reptiles: the roles of ecological opportunity and extinction," *Paleobiology,* vol. 42, no. 4, p. 547–573, 2016.

[98] A. M. Dunhill, W. J. Foster, S. Azaele, J. Sciberras and R. J. Twichett, "Modelling determinants of extinction across two Mesozoic hyperthermal events," *Proceedings of the Royal Society B,* vol. 285, no. 1889, 2018.

[99] S. Estrach, S. Schmidt, S. Diriong, A. Penna, A. Blangy, P. Fort and A. Debant, "The Human Rho-GEF Trio and Its Target GTPase RhoG Are Involved in the NGF Pathway, Leading to Neurite Outgrowth," *Current Biology,* vol. 12, no. 4, pp. 307-312, 2002.

[100] Y. Yan, E. Winograd, A. Viel, T. Cronin, S. C. Harrison and D. Branton, "Crystal structure of the repetitive segments of spectrin," *Science,* vol. 5142, no. 262, pp. 2027-2030, 1993.

[101] T. A. Russell, K. D. Blizinsky, D. J. Cobia, M. E. Cahill, Z. Xie, R. A. Sweet, J. Duan, P. V. Gejman, L. Wang, J. G. Csernansky and P. Penzes, "A sequence variant in human KALRN impairs protein function and coincides with reduced cortical thickness," *Nature Communications,* vol. 5, p. 4858, 2014.

[102] A. C. Magalhaes, H. Dunn and S. S. G. Ferguson, "Regulation of GPCR activity, trafficking and localization by GPCR-interacting proteins," *British Journal of Pharmacology,* vol. 165, no. 6, pp. 1717-1736, 2012.

[103] N. Sosonkina, S.-K. Hong, D. Starenki and J.-I. Park, "Kinome sequencing reveals RET G691S polymorphism in human neuroendocrine lung cancer cell lines," *Genes & Genomics,* vol. 36, p. 829–841, 2014.

[104] Y. Hirabayashi, S.-K. Kwon, H. Paek, W. Pernice, M. A. Paul, J. Lee, P. Erfani, A. Raczkowski, D. S. Petrey, L. A. Pon and F. Polleux, "ER-mitochondria tethering by PDZD8 regulates Ca2+ dynamics in mammalian neurons," *Science,* vol. 358, no. 6363, pp. 623-630, 2017.

[105] Z. Lin, Y. Zhao, C. Y. Chung, Y. Zhou, N. Xiong, C. E. Glatt and O. Isacson, "High regulatability favors genetic selection in SLC18A2, a vesicular monoamine transporter essential for life," *The FASEB Journal,* vol. 24, no. 7, pp. 2191-2200, 2010.

[106] F. Kronenberg and G. Utermann, "Lipoprotein(a): resurrected by genetics," *Journal of Internal Medicine,* vol. 273, no. 1, pp. 6-30, 2013.

[107] F. J. Castellino, "Biochemistry of Human Plasminogen," *Seminars in Thrombosis and Hemostasis,* vol. 10, no. 1, pp. 18-23, 1984.

[108] K. Elenius, M. Salmivirta, P. Inki, M. Mali and M. Jalkanen, "Binding of human syndecan to extracellular matrix proteins," *Journal of Biological Chemistry,* vol. 265, no. 29, pp. 17837-17843, 1990.

[109] R. P. Bhattacharyya, A. Reményi, B. J. Yeh and W. A. Lim, "Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits," *Annual Review of Biochemistry,* vol. 75, pp. 655-680 , 2006.

[110] K. Hofmann, "The modular nature of apoptotic signaling proteins," *Cellular and Molecular Life Sciences,* vol. 55, p. 1113–1128, 1999.

[111] Z. Songyang, "Recognition and regulation of primary-sequence motifs by signaling modular domains," *Progress in Biophysics & Molecular Biology,* vol. 71, pp. 359-372, 1999.

[112] C. Gotti, D. Fornasari and F. Clementi, "Human neuronal nicotinic receptors," *Progress in Neurobiology,* vol. 53, no. 2, pp. 199-237, 1997.

[113] T. W. Costantini, T. W. Chan, O. Cohen, S. Langness, S. Treadwell, E. Williams, B. P. Eliceiri and A. Baird, "Uniquely human CHRFAM7A gene increases the hematopoietic stem cell reservoir in mice and amplifies their inflammatory response," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 116, no. 16, pp. 7932-7940, 2019.

[114] A. Rozycka, J. Dorszewska, B. Steinborn, M. Lianeri, A. Winczewska-Wiktor, A. Sniezawska, K. Wisniewska and P. P. Jagodzinski, "Association study of the 2-bp deletion polymorphism in exon 6 of the CHRFAM7A gene with idiopathic generalized epilepsy.," *DNA Cell Biology,* vol. 32, no. 11, pp. 640-647, 2013.

[115] J. Zhang and Q. Zhou, "On the Regulatory Evolution of New Genes Throughout Their Life History," *Molecular Biology and Evolution,* vol. 36, no. 1, p. 15–27, 2019.

[116] J. J. Emerson, H. Kaessmann, E. Betrán and M. Long, "Extensive Gene Traffic on the Mammalian X Chromosome," *Science,* vol. 303, no. 5657, pp. 537-540, 2004.

[117] B. T. Lahn and D. C. Page, "Four Evolutionary Strata on the Human X Chromosome," *Science,* vol. 286, no. 5441, pp. 964-967, 1999.

[118] Y. E. Zhang, M. D. Vibranovski, P. Landback, G. A. B. Marais and M. Long, "Chromosomal Redistribution of Male-Biased Genes in Mammalian Evolution with Two Bursts of Gene Gain on the X Chromosome," *Plos Biology,* vol. 8, no. 10, p. e1000494, 2010.

[119] R. S. Pandey, M. A. W. Sayres and R. K. Azad, "Detecting Evolutionary Strata on the Human X Chromosome in the Absence of Gametologous Y-Linked Sequences," *Genome Biology Evolution,* vol. 5, no. 10, pp. 1863-1871, 2013.

[120] Ross, M. T. et al., "The DNA sequence of the human X chromosome," *Nature,* vol. 434, no. 7031, p. 325–337, 2005.

[121] Morgenstern, M. et al, "Definition of a High-Confidence Mitochondrial Proteome at Quantitative Scale," *Cell Reports,* vol. 19, no. 13, pp. 2836-2852, 2017.

[122] S. E. Calvo, K. R. Clauser and V. K. Mootha, "MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins," *Nucleic Acids Research,* vol. 44, pp. D1251-D1257, 2016.

[123] M. J. Berridge, Cell Signalling Biology, Portland Press Limited, 2012.

[124] M. Katoh and M. Katoh, "Transcriptional regulation of WNT2B based on the balance of Hedgehog, Notch, BMP and WNT signals," *International Journal of Oncology,* vol. 34, no. 5, pp. 1411-1415, 2009.

[125] A. A. Bashirova, L. Wu, J. Cheng, T. D. Martin, M. P. Martin, R. E. Benveniste, J. D. Lifson, V. N. KewalRamani, A. Hughes and M. Carrington, "Novel Member of the CD209 (DC-SIGN) Gene Family in Primates," *Journal of Virology,* vol. 77, no. 1, pp. 217-227, 2003.

[126] H. Li, J.-X. Wang, D.-D. Wu, H.-W. Wang, N. L.-S. Tang and Y.-P. Zhang, "The Origin and Evolution of Variable Number Tandem Repeat of CLEC4M Gene in the Global Human Population," *PLoS One,* vol. 7, no. 1, p. e30268, 2012.

[127] C. V. Ustach, W. Huang, M. K. Conley-LaComb, C.-Y. Lin, M. Che, J. Abrams and H.-R. C. Kim, "A Novel Signaling Axis of Matriptase/PDGF-D/β-PDGFR in Human Prostate Cancer," *Cancer Research,* vol. 70, no. 23, pp. 9631-9640, 2010.

[128] F. Perrone, L. DaRiva, M. Orsenigo, M. Losa, G. Jocollè, C. Millefanti, E. Pastore, A. Gronchi, M. A. Pierotti and S. Pilotti, "PDGFRA, PDGFRB, EGFR, and downstream signaling activation in malignant peripheral nerve sheath tumor," *Neuro-Oncology,* vol. 11, no. 6, p. 725–736, 2009.

[129] M. Pavlicev and G. P. Wagner, "A model of developmental evolution: selection, pleiotropy and compensation.," *Trends Ecol Evol,* vol. 27, no. 6, pp. 316-322, 2012.

[130] A. Ebert, S. J. Childs, C. L. Hehr, P. B. Cechmanek and S. McFarlane, "Sema6a and Plxna2 mediate spatially regulated repulsion within the developing eye to promote eye vesicle cohesion," *Development,* vol. 141, pp. 2473-2482, 2014.

[131] J. C. Lessard and P. A. Coulombe, "Keratin 16–Null Mice Develop Palmoplantar Keratoderma, a Hallmark Feature of Pachyonychia Congenita and Related Disorders," *Journal of Investigative Dermatology,* vol. 132, no. 5, pp. 1384-1391, 2012.

[132] A. L. Jackson and L. A. Loeb, "The Mutation Rate and Cancer," *Genetics,* vol. 148, no. 4, pp. 1483-1490, 1998.

[133] D. M. Grossnickle and E. Newham, "Therian mammals experience an ecomorphological radiation during the Late Cretaceous and selective extinction at the K–Pg boundary," *Proceedings of the Royal Society B, Biological Sciences,* vol. 283, no. 1832, 2016.

[134] G. P. Wilson, A. R. Evans, I. J. Corfe, P. D. Smits, M. Fortelius and J. Jernvall, "Adaptive radiation of multituberculate mammals before the extinction of dinosaurs," *Nature,* vol. 483, p. 457–460, 2012.

[135] D. Tautz, "The discovery of de novo gene evolution," *Perspectives in Biology and Medicine,* vol. 57, no. 1, pp. 149-161, 2014.

[136] B. Charlesworth, "The evolution of sex chromosomes," *Science,* vol. 251, no. 4997, pp. 1030-1033, 1991.

[137] S. Ohno, "Evolution of Sex Chromosomes in Mammals," *Annual Review of Genetics,* vol. 3, pp. 495-524, 1969.

[138] W. R. Rice, "Sex Chromosomes and the Evolution of Sexual Dimorphism," *Evolution,* vol. 38, no. 4, pp. 735-742, 1984.

[139] J. K. Abbott, A. K. Nordén and B. Hansson, "Sex chromosome evolution: historical insights and future perspectives," *Proceedings of the Royal Society B: Biological Sciences,* vol. 284, no. 1854, 2017.

[140] R. Ming and P. H. Moore, "Genomics of sex chromosomes," *Current Opinion in Plant Biology,* vol. 10, no. 2, pp. 123-130, 2007.

[141] M. Watanabe, A. R. Zinn, D. C. Page and T. Nishimoto, "Functional equivalence of human X– and Y–encoded isoforms of ribosomal protein S4 consistent with a role in Turner syndrome," *Nature Genetics,* vol. 4, p. 268–271, 1993.

[142] D. Bachtrog, "Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration," *Nature Reviews Genetics,* vol. 14, p. 113–124, 2013.

[143] Q. Zhou and D. Bachtrog, "Sex-specific adaptation drives early sex chromosome evolution in Drosophila," *Science,* vol. 337, p. 341–345, 2012.

[144] K. Thornton and M. Long, "Excess of Amino Acid Substitutions Relative to Polymorphism Between X-Linked Duplications in Drosophila melanogaster," *Molecular Biology and Evolution,* vol. 22, no. 2, p. 273–284, 2005.

[145] D. Charlesworth, B. Charlesworth and G. Marais, "Steps in the evolution of heteromorphic sex chromosomes," *Heredity,* vol. 95, p. 118–128, 2005.

[146] J. A. M. Graves, "Sex Chromosome Specialization and Degeneration in Mammals," *Cell*, vol. 124, no. 5, pp. 901-914, 2006.

[147] J. Kuriyan and D. Cowburn, "Modular Peptide Recognition Domains in Eukaryotic Signaling," *Annual Review of Biophysics and Biomolecular Structure* , vol. 26, pp. 259-288, 1997.

[148] K. Mori, "Signalling Pathways in the Unfolded Protein Response: Development from Yeast to Mammals," *The Journal of Biochemistry*, vol. 146, no. 6, pp. 743–750, , 2009.

[149] S. E. Calvo and V. K. Mootha, "The Mitochondrial Proteome and Human Disease," *Annual Review of Genomics and Human Genetics* , vol. 11, pp. 25 - 44, 2010.

[150] J. R. Friedman and J. Nunnari, "Mitochondrial form and function," *Nature*, vol. 505, p. 335 – 343, 2014.

[151] H. Rhee, P. Zou, N. D. Udeshi, J. D. Martell, V. K. Mootha, S. A. Carr and A. Y. Ting, "Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging," *Science*, vol. 339, no. 6125, pp. 1328 - 1331, 2013.

[152] Taylor, S. W. et al, "Characterization of the human heart mitochondrial proteome," *Nature Biotechnology*, vol. 21, p. 281–286, 2003.

[153] Sickman, A. et al., "The proteome of Saccharomyces cerevisiae mitochondria," *PNAS*, vol. 100, no. 23, pp. 13207-13212, 2003.

[154] Hughes, J. F. et al., "Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes," *Nature*, vol. 483, p. 82–86, 2012.

[155] G. H. Perry, R. Y. Tito and B. C. Verrelli, "The Evolutionary History of Human and Chimpanzee Y-Chromosome Gene Loss," *Molecular Biology and Evolution*, vol. 24, no. 3, p. 853–859, 2007.

[156] M. A. Jobling and C. Tyler-Smith, "Human Y-chromosome variation in the genome-sequencing era," *Nature Reviews Genetics*, vol. 18, p. 485–497, 2017.

[157] R. Pastor-Satorras, E. Smith and R. V. Solé, "Evolving protein interaction networks through gene duplication," *Journal of Theoretical Biology*, vol. 222, no. 2, pp. 199-210, 2003.

[158] K. J. Lipinski, J. C. Farslow, K. A. Fitzpatrick, M. Lynch, V. Katju and U. Bergthorsson, "High Spontaneous Rate of Gene Duplication in Caenorhabditis elegans," *Current Biology*, vol. 21, no. 4, pp. 306-310, 2011.

[159] J. A. Cotton and D. M. Page, "Rates and patterns of gene duplication and loss in the human genome," *Proceedings of the Royal Society B*, vol. 272, no. 1560, p. 277–283, 2005.

[160] Y. E. Zhang, P. Landback, M. D. Vibranovski and M. Long, "Accelerated Recruitment of New Brain Development Genes into the Human Genome," *PLoS Biology,* 2011.

[161] L. S. Ripley, "Frameshift mutation: Determinants of specificity," *Annual Review of Genetics,* vol. 24, pp. 189-213, 1990.