Analysis of the Relationship Between Functional Divergence and the Propensity of Duplicated Genes to be Involved in Human Mendelian Diseases

By

Katherina Radan

A Thesis Submitted to
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Bachelors of Science.

April, 2017, Halifax, Nova Scotia

© Copyright Katherina Radan, 2017

Approved:	Dr. Daniel Gaston External supervisor
Approved:	Dr. Zhongmin Dong Internal supervisor
Approved:	Dr. Timothy Frasier Reader
Date:	April 21, 2017

Analysis of the Relationship Between Functional Divergence and the Propensity of Duplicated Genes to be Involved in Human Mendelian Diseases

by

Katherina Radan

Abstract

In recent studies, it has been observed that genes that have been duplicated during the course of vertebrate evolution are overrepresented among those genes that cause Mendelian diseases. My objective was to determine whether measures of functional divergence are correlated with the propensity of duplicated genes to be involved in Mendelian disease. To test this, I used a phylogeny-based maximum-likelihood mixture-model prediction program, FunDi, that accounts for functional divergence in phylogenetic trees. I then conducted a statistical analysis of the data, measuring the Rho value of functional divergence weight and branch lengths values, using a Pearson correlation test and two-sided Wilcoxon-Mann-Whitney U-test. Statistically significant correlation was found for the relationship between the length of the branch in the phylogenetic tree separating disease-associated genes and its orthologs from the rest of the gene family and the propensity for a gene to be involved in autosomal recessive disorders. Optimization with FunDi, which accounts for functional divergence in its model, resulted in shorter branch lengths. Unfortunately, no statistical significance was found between the analyzed gene categories for the Rho value. Therefore, I conclude that while some measures of evolution and functional divergence, such as the internal branch length between groups, may be correlated with disease-association, direct measures of functional divergence measured in this study do not explain the propensity of duplicated genes to be involved in Mendelian diseases.

April 21, 2017

TABLE OF CONTENTS

List of Abbreviations and Symbols Used

Δ	Delta
AD	Autosomal Dominant
AR	Autosomal Recessive
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
DMD	Dystrophin Gene
DNA	Deoxyribonucleic Acid
EMF	Enhanced MetaFile
ETE	Environment for Tree Exploration
FASTA	FAST-All
FD	Functional divergence/functionally divergent
FMRP	Fragile X Mental Retardation Protein
FMR1	Fragile X Mental Retardation 1 Gene
FunDi	Phylogeny-based Functional Divergence Prediction Program
HTT	Huntingtin Gene
IQ-Tree	Efficient Phylogenomic Software by Maximum Likelihood
JTT	Jones-Taylor-Thornton
LG	Le and Gascuel
MD	Monogenic / Mendelian Disease
ND	Non-Disease
Rho	Functional Divergence Weight Parameter
WAG	Whelan and Goldman

Introduction

Deoxyribonucleic acid (DNA) is a polymer of nucleic acids that carries the genetic instructions for the growth, function, development, and reproduction of all living organisms. Damage to DNA can cause genetic alterations, known as mutations, which can lead to hereditary diseases. It is important to note that mutations are not always harmful. In most cases, they are neutral, and in some instances, can even be beneficial (Ohta, 1973; Ohta & Gillespie, 1996). Human hereditary diseases are broadly classified as either monogenic or complex disease, based on whether they are caused by mutations in a single gene, or multiple genes. Monogenic diseases (also known as Mendelian diseases) are caused by mutations that lead to changes in the function of a single gene, including the complete loss of function. Sickle cell anemia, Marfan syndrome, Huntington's disease, and cystic fibrosis are all examples of monogenic diseases (Carter, 1977; Rees *et al*., 2010; Dietz & Cutting, 1991; Roos, 2010; Zielenski *et al.*, 1991). Some Mendelian diseases are more common in certain areas and populations. For example, Fabry's disease and Niemann-Pick disease are particularly common in Nova Scotia (Greer *et al.*, 1998; Kirkilionos *et al.,* 1991). The Finnish Heritage diseases, which are a collection of approximately 40 rare genetic diseases (Norio, 2003), are common among ethnic Finns. In addition, many monogenic diseases are more commonly found in Ashkenazi Jewish populations, such as Canavan disease, Gaucher disease, familial dysautonomia, Bloom syndrome, and Fanconi anemia (Scott *et al*., 2010). Complex genetic diseases (also known as polygenic diseases) are responsible for the vast majority of human genetic-related diseases, and are caused by mutations in multiple genes that must be inherited together (Wink, 2006), and may also be influenced by the environment (Caspi *et al.,* 2010). Many different mutations, with smaller effects, all contribute to the risk for a polygenic disease. Examples of polygenic diseases include coronary heart disease, many cancers, Type 2 diabetes, and a number of birth defects and psychiatric disorders (Weeks & Lathrop, 1995; Taylor, 1999; Swerdlow *et al.*, 2012). Related to complex diseases are those that are caused by acquired genetic mutations, called acquired genetic disorders (Risch & Merikangas, 1996). These develop during one's lifetime and include cancer and "cancer-like" diseases such as myelodysplastic syndrome and thrombocytopenia (Mijović & Mufti 1998; Drachman, 2004).

Deleterious mutations can be either "loss-of-function" or "gain-of-function". Loss-of-function (inactivating/null) mutations result in a gene product that has reduced or no function, while gain-of-function (activating) mutations result in a gene product that has a new or enhanced function, pattern of gene expression, or regulation. Some genes can be essential to the viability of organisms because they can acquire deleterious mutations that cause loss-of-function or null mutations in genes. These mutations do not affect the phenotype and in turn, the organism's viability. Other gene families have been described as "dangerous", due to their tendency to acquire deleterious gain-of-function mutations, which increases their susceptibility to genetic diseases (Singh *et al.*, 2012; Singh *et al.*, 2014). In both cases, functional divergence can play an important role in gene evolution and functionality in the organism, and is a driving force in the evolution of genes involved in genetic diseases, including Mendelian disease.

1.1 Mendelian Diseases

Mendelian disease genes follow four main patterns of inheritance: autosomal dominant, autosomal recessive, X-linked recessive, and X-linked dominant (Chial, 2008). In the autosomal dominant pattern, a disease occurs when one copy of an allele is mutated and the disease will typically appear in every generation of the family. An example of this type of Mendelian inheritance is Huntington's disease, which is a progressive disorder of motor, cognitive, and psychiatric changes. This disease is caused by a genetic defect that alters the Huntingtin (HTT) gene on chromosome 4 (Roos, 2010), which changes the number of C-A-G trinucleotide repeats in the gene (Ross & Tabrizi, 2011). Generally, the number of repeats is between 10 and 35, but Huntington's disease occurs when these increase to 36 or more, which produces a longer and unstable Huntingtin protein (HTT) that ultimately leads to neurodegeneration (Finkbeiner, 2011).

In autosomal recessive diseases, two copies of the harmful allele must be present for the individual to express the disease. The disease will not appear in every generation of the family, but carriers will be present in every generation. A typical example of a Mendelian disease with autosomal recessive inheritance is cystic fibrosis (Zielenski *et al.*, 1991). This relatively common genetic disease occurs in about 1 in 3,500 individuals of European descent (Ratjen, 2009), and is caused by the loss-of-function of a chloride channel, which is coded for by the cystic fibrosis transmembrane conductance regulator (CFTR) gene on chromosome 7 (Zielenski *et al.*, 1991). The loss-of-function mutations in the protein leads to an accumulation of thick mucus in the digestive, reproductive, and respiratory systems, leading to an increase in inflammation, infection, and respiratory problems (Zielenski *et al.*, 1991; Welsh & Smith, 1993).

X-linked diseases operate the same way as diseases on autosomes, except that males only have one copy of the X chromosome. This means that if a mutation appears on the X chromosome, the male will be affected. Since females have two copies of the X chromosome, X-linked recessive diseases are more common among men. An example of a disease with an X-linked recessive pattern of inheritance is Duchenne muscular dystrophy. The disease is caused by a mutation of the dystrophin gene (DMD) located at the short arm of the X chromosome (Blake & Kröger, 2000). Mutation of the DMD gene prevents the creation of the protein dystrophin, which leads to an excess of calcium in the cell membrane, alters cell signaling pathways, and ultimately leads to a progressive muscular disorder (Dobyns *et al*., 2004). Fragile X syndrome is a disease with an X-linked dominant pattern of inheritance, which is caused by mutation in the fragile X mental retardation 1 (FMR1) gene, located on the X chromosome. Similar to Huntington's disease, the mutation alters the length of the gene by expending the C-G-G trinucleotide repeats in the gene. The number of repeats is increased from between 5 and 44 to 45 or more, resulting in failure to express a normal protein (FMRP) which leads to abnormal neural development (Hagerman, 2005; Garber *et al*., 2008).

1.2 Mendelian Diseases and Gene Duplications

It has been observed that genes that have been duplicated are overrepresented among genes that cause Mendelian diseases (Chen *et al.,* 2013; Chen *et al.,* 2014; Singh *et al.*, 2012; Singh *et al.,* 2014). Duplicated genes result from gene or genome duplications, which are important mechanisms for creating genetic variation and novelty in organisms (Stephens, 1951; Magadum *et al*., 2013) by providing new genetic material for mutation, drift, and selection to act upon. Many new gene functions have evolved through the process of gene duplication (Alberts *et al.,* 2002; Wolfe & Li, 2003). These duplications are divided into two main categories: small-scale duplication and whole-genome duplication (Dehal & Boore, 2005). Small-scale duplication occurs when a specific region of the genome, containing a single gene or a few genes located close together, is duplicated. Genes that originate from small-scale duplication are highly diverse in their function and are thought to be more essential (Hakes *et al*., 2007). Whole-genome duplications are large-scale events where the entire genome is duplicated, resulting in additional copies of the genome (polyploidy) (Dehal & Boore, 2005). Polyploid cells contain more than two paired (homologous) chromosome sets. Genomes that have been duplicated in this fashion eventually return to a diploid state (diploidization), and some gene copies are lost or gained during this process (Conrad & Antonarakis, 2007). Because of this, genes that originate from whole-genome duplication are less diverse in their functions, are thought to be less essential, and are more likely to be members of a protein complex (Hakes *et al.*, 2007). Since small-scale duplication provides no immediate benefit, they will have low probability to be retained and will be rapidly lost following the duplication. Whole-genome duplication, on the other hand, can provide immediate benefit thus selection will act stronger to retain these duplicates (Hakes *et al.*, 2007). During the course of vertebrate evolution, two rounds of whole-genome duplication have occurred, and this is hypothesized to be a driving force behind increases in organismal complexity (Brunet *et al.*, 2006; Acharya & Ghosh, 2016). This hypothesis argues that gene duplication created genetic redundancy, which allowed for novel genes and gene functions to develop (Chen *et al.*, 2013; Chen *et al.*, 2014). This redundancy reduces or changes the functional constraints that otherwise operated on the original single gene (Qian & Zhang, 2014).

Duplicated genes are called paralogs. These paralogs typically perform the same role as the original single gene initially, but can diverge over time (Innan & Kondrasov, 2010). After gene duplication, most duplicated genes acquire mutations that render them nonfunctional quickly (in evolutionary time) (Hughes, 1994; Hurles, 2004). These non-functional paralogs are called “pseudogenes”. Over time, many of these pseudogenes are completely lost from the genome, along with its functionality (Hughes, 1994; Hurles, 2004; Ohno, 1970; Hufton & Panopoulou, 2009). However, over the course of evolution, mutation and selection can act independently on the duplicate copies, leading to functional divergence between paralogs. The main types of functional divergence are neofunctionalization, an adaptive process where one paralog acquires a new function that was not present in the pre-duplicated gene, and subfunctionalization, which can be either an adaptive or neutral process where both paralogs in a duplicated gene partition the ancestral function (Rastogi & Liberles, 2005; Innan & Kondrasov, 2010). For example, the creation of hemoglobin, myoglobin, and cytoglobin from one ancestral gene (Hoffmann *et al.,* 2011; Ebner *et al.,* 2003), is an example of gene duplications followed by functional divergence. Currently, all three proteins perform a similar role in binding oxygen, but their specific function and tissue-specific expression differs.

As stated previously, genes that cause Mendelian disease are enriched in duplicated genes (Dickerson & Robertson, 2011; Chen *et al.*, 2013; Chen *et al.*, 2014; Singh *et al.*, 2012; Singh *et al.*, 2014). This observation was unexpected, as it was previously hypothesized that singletons (genes without duplicates) were more likely to be functionally critical as paralogous genes could potentially compensate for one another and mask the effects of deleterious mutations (Brunet *et al.*, 2006; Gu *et al.*, 2003; Dickerson & Robertson, 2011; Chen *et al.*, 2012; Chen *et al.*, 2013; Chen *et al.*, 2014). Several proposals have been put forward to explain this observation. First, the age of the duplicates may play a major role in their ability to functionally compensate for one another when there is a mutation in one of the paralogs. In the case of ancient duplications, due to the amount of time that has passed, functional divergence is more likely to have occurred. This functional divergence prevents the paralogous genes from functionally compensating for one another (Chen *et al.*, 2013; Chen *et al.*, 2014). In contrast to older duplication events, paralogs that result from more recent duplication events might still be able to functionally compensate for mutations in their gene duplicate. This means that if a mutation or damage occurs in one copy of the gene, the other copy will still produce a normal and functional gene product. In the absence of strong functional divergence, both paralogs can perform the same functions and are typically still expressed in the same tissues. Second, whole-genome duplication created the potential for more complex systems, as all genes were duplicated at the same time, allowing more opportunities for functional divergence. Gene divergence would then expose the deleterious effect of genetic alterations (mutations) and lead to disease. Existing studies have shown that genes that are prone to dominant deleterious mutations are considered to be more "dangerous" (Singh *et al.*, 2012; Singh *et al.*, 2014). In addition, functional compensation by duplication of genes masks the phenotypic effects of deleterious mutations and reduces the probability of purging the defective genes from human population (Chen *et al.*, 2013; Chen *et al.*, 2014).

Mendelian diseases can have a large burden on human health. This includes both loss of life and decreased quality of life, because most aren't fatal (Costa *et al.,* 1985; Botstein & Risch, 2003). Though relatively rare individually, it is estimated that over 10,000 human diseases are known to be Mendelian, and the global prevalence is approximately 10/1000 individuals at birth (World Health Organization, 2016; Chakravarti, 2011). In this project, the focus is on genes that can lead to Mendelian diseases when mutated. Understanding the mechanism of genes and diseases has long been a point of interest in genetic research. There are still many Mendelian diseases where the causal mutation and gene are not yet known. Discovering and analyzing the genetic basis of known Mendelian diseases will contribute to our knowledge of gene function and regulation and will also allow us to develop better treatment methods in the future (Chong *et al.*, 2015). Today genome-scale analyses are incredibly useful for identifying genetic mutations; however, the small number of rare mutations found in a typical genome means we need to develop methods that will prioritize genes likely to be involved in Mendelian disease.

1.3 Thesis Objectives

In this project, I hypothesize that direct measures of functional divergence between paralogs are associated with the propensity of duplicated genes to be involved in Mendelian disease. I test my hypothesis by using a phylogeny-based, functional divergence prediction program, FunDi (Gaston *et al.*, 2011), to analyze ~9000 gene families. These families include both genes involved in Mendelian disease and non-disease genes. I predict that when comparing gene families that contain genes that cause Mendelian disease to gene families that do not, measures of functional divergence produced by FunDi, particularly the functional divergence score, will be higher in the gene families that are involved in Mendelian diseases. This is due to the compensation hypothesis where genes become too diverse (=higher functional divergence) and cannot compensate each other, leading to Mendelian disease when a mutation is acquired. Additionally, I predict that when comparing the two main patterns of Mendelian inheritance in autosomal (non-sex chromosome) genes, autosomal dominant and autosomal recessive, the functional divergence score will be the highest for genes involved in autosomal dominant disorders. This is due to the strength of the pattern of inheritance, where in the dominant pattern a defect in one allele can lead to disease, which also based on the compensation hypothesis. I propose that having a tool that can analyze genes, and output a significant functional divergence score, will aid in identifying new disease-causing genes; helping us gain a better understanding of our genome, its evolution, and disease-causing potential.

Methods

2.1 Data Acquisition, Cleaning, and Integration

To construct a dataset of human gene families, I downloaded 15,570 gene sequence alignments and phylogenetic trees from TreeFam (v9) (Schreiber *et al*., 2012). This approach was based on the method used by Chen *et al*. (2013); however, only alignments and phylogenetic trees of gene families (where at least two paralogous human sequences are present in the alignment and tree) were retained, for a total of 8,166 sequence alignments and their respective phylogenetic trees. Genes were then linked to extra information, particularly their disease categorization (non-disease, recessive, dominant, etc.), from the Chen *et al.* (2013) supplementary data by using the TreeFam group identifier, Ensembl identifiers and a custom python script.

In this analysis, I was specifically interested in the relationship between functional divergence and the disease-causing potential of duplicated genes. Additionally, because FunDi in this analysis uses the program IQ-TREE to perform the maximum-likelihood phylogenetic analyses, a minimum of three taxa in each defined subgroup of the phylogenetic tree is required (IQ-TREE, 2016) as this is how the program was built. Therefore, 250 of the multiple sequence alignment files that had fewer than three taxa in each defined subgroup of the phylogenetic tree were removed from the analysis. Due to polytomies (unresolved evolutionary relationships in which three or more branches originate from the same node, particularly at the root of the tree (Olmstead, 1996, Lin *et al.*, 2011)) in some phylogenetic trees, an additional 1,500 multiple sequence alignment files were removed from my dataset. After filtering and cleaning the dataset of human gene families from the mentioned datasets, a total of 6,416 gene family alignments and phylogenetic trees were analyzed using FunDi. Because the phylogenetic tree information was encoded in the EMF (Enhanced MetaFile) file format, and multiple sequence alignments in FASTA (text-based format for representing peptide sequences) files, custom python scripts were written to extract newick-format phylogenetic trees from EMF files and to convert FASTA-format sequence alignments to the phylip-format used by FunDi.

Having only the Ensembl gene ID's of interest from Chen *et al.*'s (2013) dataset as a starting point, a datasheet that contained all human gene families, as our dataset and their associated paralog proteins, was created (Fig. 1).

2.2 Subtree Definition File Creation

The subtree definition files define two different subgroups within multi-member gene families. FunDi separates the tree into its constituent subtrees, using the subtree definition file as a map that identified which sequences will be in each of the two subgroups to be considered in the analysis. Knowing only the protein ID's as our starting point, a connection to its' paralogs was required. To split the tree files into subgroups, an algorithm was written in python that uses the ETE v3 (Environment for Tree Exploration v.3.0.0b35) package (Huerta-Cepas *et al.*, 2016; ETE Toolkit - Analysis and Visualization of (phylogenetic) trees, 2016). This toolkit allows programmatic access to, and manipulation of, a phylogenetic tree. Specifically, the subtree definition files were generated using the following algorithm (Fig. 2):

1. Iterate through the terminal (leaf) nodes of the tree until the node corresponding with the protein of interest is found.
2. Starting from the terminal node containing the identified protein in 1, move to the

parental node. Check all terminal nodes that are a descendant of this node to see if they contain a human protein identifier, excluding the original protein of interest. If a human paralog is identified here, then stop and proceed to step 4. Otherwise continue with step 3.

3. Repeat step 2 by continuing to traverse through ancestral nodes, towards the root of the tree, until a terminal node with a descendant terminal node corresponding to a human protein other than the original protein of interest is found.
4. Once an internal node of the tree that has a descendant node corresponding to a human protein other than the original protein of interest is identified, go to the previous internal node tested, as that was the last node that did not contain this paralog as a descendant.
5. Create a list that contains the sequence identifiers of all terminal nodes descended from the internal node identified in 4.
6. Iterate over all the terminal nodes in the tree and create a second list of all sequence identifiers that do not appear in the list created in 5.
7. Create a subtree definition file whose first line is list of nodes from step 5 and a second line that has the list of nodes from step 6.

2.3 FunDi Analysis

FunDi is a phylogenetic maximum-likelihood mixture-model prediction program that identifies functionally divergent sites among protein families, using specified models of amino acid substitution (Gaston *et al.*, 2011). A proper FunDi run of each gene family in the dataset uses the following files: a multiple sequence alignment file (phylip), a phylogenetic tree file (newick), and a subtree definition file. FunDi compares two subgroups in each gene family: a subgroup defined by a human protein sequence of interest along with its orthologs and the rest of the sequences in the gene family. In the case where a gene family contains three or more human paralogs FunDi will conduct one comparison per paralog.

FunDi uses a two-component mixture model where sites in the multiple sequence alignment are modelled using a dependent component (standard evolutionary model) and an independent model, where the specified subtrees of gene family's phylogenetic tree are treated as completely independent trees. After maximum-likelihood optimization, a site-wise posterior probability of functional divergence for each site in the alignment is calculated, where the independent model approximates functional divergence. In the dependent component (non-functionally divergent), the maximum-likelihood evaluation of the tree as a whole reflects normal evolutionary models, where all of the evolutionary parameters (i.e.: evolutionary rate, and amino acid frequency) are the same across the phylogenetic tree. The independent (functionally divergent) component models the two subtrees as independent of one another; therefore, the subtrees can be optimized to different evolutionary rates amino acid frequencies, and other evolutionary parameters between the two parts of the phylogenetic tree. FunDi optimizes the overall ratio between the independent and dependent components, the branch length between the two groups, and finally estimates a functional divergence value. The determination of whether the site is functionally divergent or not is dependent on a set cut-off threshold (standard threshold is 0.5). In our analysis, we set different thresholds (0.75, 0.9, and 0.95) to test the weight of functional divergence of the different sites in the alignment files. The use of the two components is an attempt to statistically and computationally model the process of molecular evolution when functional divergence might be occurring.

FunDi analysis used the LG model of amino acid substitution, a gamma distribution of evolutionary rates among sites in the sequence alignment, and an empirically estimated frequency of individual amino acids in the multiple sequence alignment. The LG model is one of many amino acid replacement matrices (such as WAG (Whelan & Goldman, 2001), JTT (Jones *et al.*, 1992), and PAM (Dayhoff *et al.*, 1978)) that are used in protein phylogenetic inference (Le & Gascuel, 2008). These models are used to calculate probabilities of amino acid substitution along branches of phylogenetic trees. The dataset of multiple sequence alignments that was used to construct the LG substitution matrix was larger and contained a more diverse set of sequences (Le & Gascuel, 2008). The gamma distribution is used to model multiple rates of evolution at sites in the multiple sequence alignment, allowing for faster or slower evolving sites (Yang, 1994; Yang, 1996). The shape of this gamma distribution of site-rate categories is controlled by the shape parameter alpha, which is optimized during the maximum-likelihood optimization process.

FunDi uses the site log-likelihood values estimated by the program IQ-TREE (v.1.5.0) (Nguyen *et al.*, 2015; IQ-TREE, 2016). IQ-TREE takes the provided phylogenetic tree and multiple sequence alignment and re-optimizes branch lengths, the gamma shape-parameter alpha, and other aspects of the phylogenetic tree, with the exception of the tree topology, given the provided evolutionary model as described previously.

Branch length optimization was used in FunDi for each phylogenetic tree, in order to properly model functional divergence. In some cases, this results in the shortening of the internal branch length, which can be artificially inflated under standard evolutionary models that do not account for functional divergence. This optimization allowed me to estimate the optimal maximum-likelihood tree, with optimized branch lengths. FunDi output the following values:

fraction of functionally divergent sites, optimal branch length of the internal branch separating the defined subgroups, and the optimized weight parameter, Rho, for the independent (functional divergence) model component.

The branch lengths in some of the analyzed tree files were theoretically very long and before using FunDi program, are referred to as "un-optimized". They are referenced in the following figures as the "Pre-FunDi Branch Length". As mentioned, FunDi uses the IQ-TREE (v.1.5.0) algorithm (Nguyen *et al*., 2015; IQ-TREE, 2016) that takes the provided phylogenetic tree file and re-optimizes the branch lengths. After optimizing the phylogenetic trees with FunDi, the branch length we considered "optimized" and were described as "Post-FunDi Branch Length".

2.4 Statistical Analysis and Data Visualization

A number of statistical tests and data visualization methods were used to examine the association between functional divergence and the disease-causing status of genes.

Comparisons of the functional divergence weight (Rho) were made between gene-families that cause Mendelian-disease (MD) vs non-disease (ND) gene-families. In addition, we specifically compared autosomal dominant (AD) and autosomal recessive (AR) genes.

While Rho is the most direct measure of functional divergence output by FunDi, additional comparisons were made between the fraction of functionally divergent sites for genes in the AD and AR disease categories (that represent the subcategories of MD) when compared to those in the ND category, using different cut-offs (0.5, 0.75, 0.9, and 0.95) for the site-wise posterior probability of functional divergence output by FunDi.

The various comparisons were plotted as box plots or scatter plots, and analyzed using Minitab v.17 (Minitab 17, 2016), with focus on internal branch length (branch length separating a subgroup of interest from the rest of a phylogenetic tree) and Rho (weight of the independent component of the FunDi mixture model) values estimated by FunDi.

Genes in our dataset were also separated into two-member and three or more-member gene families. Gene families with more than two paralogs are more difficult to accurately characterize and feature more complex, and overlapping, comparisons.

For the scatter plots, the Pearson correlation test was used to measure the strength of the relationship between two variables. For the box plots, two-sided Wilcoxon-Mann-Whitney U-tests were used to evaluate the existence of significant differences between two independent groups. For both tests, the obtained P-values measured the significance of the tested relationships. In addition, Bonferroni correction (also known as the Bonferroni type adjustment (Bonferroni, 1936; Dunn, 1959; Armstrong, 2014)) was made on the P-values in order to reduce the chance for a false positive error; rejecting the null hypothesis when I should not.

In order to better visualize the correlation of numerical data whose ranges differ significantly in magnitude, the Rho values were re-scaled using the following function:

$$\text{Logit} = \text{Ln}\,[\text{Rho} / (1\text{-Rho})] \tag{1}$$

The internal branch lengths were rescaled simply by taking the log of the branch length.

Results

All reported P-values were corrected for multiple comparisons using the Bonferroni method in order to reduce false-positive errors in the analysis.

3.1 Branch Length Results from Phylogenetic Trees

The relationship between internal branch length values and the FunDi functional divergence score, Rho, for each gene family are shown in figure 3 for the two-member only (A&B) gene family groups, both before (A) and after (B) FunDi optimization. Similarly, the results for three or more-member (3+) gene families are also shown (C&D). For the two-member gene families there were 78, 125, and 1145 sets of alignments and phylogenetic trees for the AD, AR, and ND categories respectively. For the three or more-member gene families there were 217, 210, and 3172 alignments and phylogenetic trees for the AD, AR, and ND categories respectively.

For the group of two-member gene families, the Pearson correlation coefficient in the ND and AR groups of genes, both before and after internal branch length optimization with FunDi, had relatively weak positive (before optimization: ND: r = 0.613, P = 0.0002 and AR: r = 0.660, P = 0.0002; Fig. 3A; after optimization: ND: r = 0.126, P = 0.0002, AR: r = 0.147; P = 0.2060; Fig. 3B) correlation between the functional divergence weight parameter, Rho, and the internal branch length. However, this relationship was not statistically significance in the ND category after optimization of the internal branch length. The AD subcategory did not have a statistically significant correlation between Rho and the internal branch length before or after FunDi optimization (before optimization: AD: r = 0.187, P = 0.2080; Fig. 3A; after optimization: AD: r = 0.001, P = 1; Fig. 3B).

For the 3+ gene family groups, the Pearson correlation coefficient had a strong positive measure between in the AR subcategory, and a weak positive correlation in the ND category and AD subcategory. There was statistical significance for all three subcategories before optimization with FunDi (before optimization: ND: r = 0.613, P = 0.0002, AR: r = 0.726, P = 0.0002, and AD: r = 0.650, P = 0.0002; Fig. 3C), and after FunDi analysis all three gene categories had relatively weak positive relationships, although it was not significant for the AD group (after optimization: ND: r = 0.067, P = 0.0002, AR: r = 0.188, P = 0.0120, and AD: r = 0.036, P = 1; Fig. 3D).

The distribution of internal branch length values, before and after FunDi optimization, for each disease-gene category are shown in figure 4 for the two-member only (A) gene family groups. Similarly, the results for three or more-member (3+) gene families are also shown (B). We tested for differences in the distribution of branch length between the AD and AR gene categories with those of ND genes as described in the methods.

For the two-member gene family group (A), a significant difference was only seen between the ND and AR gene categories before FunDi optimization (before optimization: ND vs. AR: P = 0.0256, and ND vs. AD: P = 1; after optimization: ND vs. AR: P = 0.2110, and ND vs. AD: P = 0.7706; Fig. 4A).

For the 3+ group of gene families (B), no significant difference was found between any of the gene categories before FunDi optimization, but a significant difference was seen between the distribution of internal branch lengths when comparing the ND gene category and the AR category after branch length optimization with FunDi (before optimization: ND vs. AR: P = 0.2482, and ND vs. AD: P = 1; after optimization: ND vs. AR: P = 0.0032, and ND vs. AD: P = 0.4552; Fig. 4B).

We also analyzed the distribution of the difference between pre- and post-FunDi optimization of the internal branch length (Fig. 5), for both the two-member (A) and 3+ member (B) gene families. Here the difference in branch lengths is a measure of improvement in model fit when FunDi is used versus a standard evolutionary model. No statistically significant difference was found for either comparison (ND vs. AR: P = 0.0552, and ND vs. AD: P = 1; Fig. 5A). For the 3+ gene family group (B), no significant result was found for any comparison (before and after optimization difference: ND vs. AR: P = 1, and ND vs. AD: P = 1; Fig. 5B).

3.2 Rho Value - Functional Divergence Results

Boxplots of the Rho values obtained for each group are shown in figure 6. For our analysis of the functional divergence score, Rho, no significant differences were seen between either the AD or AR category when compared to the ND category in either the two-member (ND vs. AD: P = 1, and ND vs. AR: P = 0.4422; Fig. 6A) or 3+ member (ND vs. AD: P = 0.8436, and ND vs. AR: P = 0.4712; Fig. 6B) groups.

While Rho is the most direct measure of functional divergence output by FunDi, we also compared the fraction of sites considered to be functionally divergent when using different cut-offs for the site-wise posterior probability of functional divergence output by FunDi. Boxplots of the fraction of functionally divergent sites for each gene category are shown in figure 7. For our analysis of the fractions of the functionally divergent sites, no significant differences between either the AD or AR categories when compared to the ND genes were seen for any category in the two-member group (fraction 0.5: ND vs. AD: P =1, and ND vs. AR: P = 0.1814; fraction 0.75: ND vs. AD: P = 1, and ND vs. AR: P = 0.3522; fraction 0.9: ND vs. AD: P = 1, and ND vs. AR: P = 0.9050; fraction 0.95: ND vs. AD: P = 1, and ND vs. AR: P = 0.8908; Fig. 7A).

For the 3+ member group, a statistically significant difference was only seen when comparing AR genes to ND genes when using the most conservative (0.9 and 0.95) cutoff for defining functionally divergent sites in the multiple sequence alignment comparing (fraction 0.5: ND vs. AD: P = 0.9808, and ND vs. AR: P = 0.7532; fraction 0.75: ND vs. AD: P = 1, and ND vs. AR: P = 0.1452; fraction 0.9: ND vs. AD: P = 1, and ND vs. AR: P = 0.0222; fraction 0.95: ND vs. AD: P = 1, and ND vs. AR: P = 0.0062; Fig. 7B). The results for figure 7 are illustrated in table 1.

Discussion

4.1 Observation and Predictions

Previous studies have shown that genes that cause Mendelian disease are overrepresented among genes that have been duplicated in the course of evolution (Chen *et al.*, 2013; Chen *et al.*, 2014; Singh *et al.*, 2012; Singh *et al.*, 2014). Our approach focused on analyzing the relationship between functional divergence and the propensity of gene families to be involved in Mendelian disease using three measures of functional divergence (Rho, the difference between pre- and post-optimized internal branch length, and the fraction of functionally divergent sites in the multiple sequence alignment) as well as a measure of the evolutionary distance between paralogs (internal branch length), which functions as a more indirect, and nonspecific, measure of functional divergence.

As a prediction, we expected to see some significant difference between gene families that cause Mendelian disease (MD) or specific categories of disease (AD and AR) when compared to non-disease causing gene families (ND). The reasoning behind this prediction was the assumption that duplicated genes that cause Mendelian disease evolved differently than non-disease genes. As previously stated, evolutionary pressures such as mutation and selection act independently on duplicated genes, releasing them from the constraints of their original function (Qian & Zhang, 2014; Chen *et al.*, 2013; Chen *et al.*, 2014; Innan & Kondrasov, 2010). Previous studies have shown that disease-associated genes, when compared to all other genes, were more conserved at the protein level (López-Bigas & Ouzounis, 2004; Huang *et al.*, 2004; Smith & Eyre-Walker, 2003; Tu *et al.*, 2006). Therefore, some significance between the two categories must be present at some level.

In addition to the previous prediction, when comparing the two Mendelian disease subgroups, we hypothesized that gene families that cause autosomal dominant diseases specifically (AD) may show more functional divergence than those that cause autosomal recessive diseases (AR). This would indicate that evolutionary forces acted strongly on gene families in that category, resulting in a greater degree of functional diversity between paralogs in these genes compared to either AD or ND genes (Singh *et al.*, 2014). In addition, this may provide evidence that may support or reject the compensation hypothesis (Chen *et al.*, 2013; Chen *et al.*, 2014). The rationale behind this prediction is that in dominant disorders we do not expect functional compensation, as only one copy of the disease-associated allele is required to cause disease, whereas in the recessive case, both alleles need to be affected to cause disease (Chial, 2008). In addition, due to the fact that a single copy of the normal gene cannot compensate in a dominant disease, we do not expect a paralog to be able to compensate either. Therefore, there won’t be any selective pressures to maintain a compensatory copy, allowing for greater functional divergence.

4.2 Results of the Analyzed Datasets - Proposed Explanations

Despite my predictions, my results mainly showed no significant differences in the most direct signature of functional divergence, the Rho weight parameter, between disease-causing and non-disease causing gene categories. The results were plotted using the FunDi output values, and included the internal branch lengths, Rho value, and fractions of functionally divergent sites.

For my prediction, I expected Rho values and pre-optimized internal branch length values to correlate with one another when comparing the results before and after FunDi optimization, as I assumed that genes that are more divergent tend to have longer branch lengths (Gu, 2001). My results showed that before optimizing with FunDi, there was relatively strong correlation between the categories, except the AD category in the two-member gene family group (Fig. 3). Since FunDi accounts for functional divergence in its model, it tends to result in a shortening of the branch lengths, which would break the correlation between Rho values and internal branch length values, as Rho stays constant between the two graphs. Therefore, my observation in general is exactly what I expect to see.

When I tested for differences in the distribution of internal branch lengths between the AD versus ND category and AR versus ND category, statistical significance was seen only between AR and ND categories but only when looking at the two-member gene family pre-FunDi optimization and three or more-member gene family post-FunDi optimization (Fig. 4). As expected, the results showed shortening of the internal branch lengths after optimizing the values with FunDi, which showed that FunDis' optimization tended to results in shorter internal branch lengths than those of the starting pre-optimization tree. The statistically significant difference observed between AR and ND categories in the two-member gene family group was no longer seen after FunDi optimization, as was expected, but it was interesting to see that in the three or more-member gene family group a statistical significance was gained after FunDi optimization. This could represent a false result as FunDi was not specifically designed to handle groups such as the three or more-member gene families as well as the two-member families.

In order to see how effective FunDi was in optimizing and shortening the internal branch length values for the two groups, I calculated the difference (Δ-delta) between pre- and post-optimized internal branch lengths (Fig. 5). With that, I wanted to see if FunDi's optimization has significantly shortened the internal branch lengths. Before the correction, the results showed that there was a statistical significance when comparing the AR and ND categories in the two-member gene family group, but after the correction it was no longer significant, although it was relatively close to the threshold. The three or more-member gene family group did not show any statistically significant results in any of the analyzed categories. One possible explanation for this observed difference when comparing between the two-member gene family group versus the three or more-member gene family group is the difference in the relationships between the two. The two-member gene family group has relatively simple relationships, while the three or more-member group has more complex relationships where there are potentially multiple subgroups within the tree undergoing functional divergence. FunDi is designed to explicitly model functional divergence that occurs along a single internal branch. While it is probably a better model than the standard evolutionary model, it still has shortcomings in some of these situations.

For my prediction, I expected to see significant differences in Rho values between the categories for both groups. My results did not see eye to eye with my logic and showed no significance in any category, for any of the analyzed groups (Fig. 6). As mentioned, Rho is the optimized weight of the independent component of the FunDi mixture model, which is a measure of the functional divergence between paralogs. I did not detect any statistically significant difference between either category, the disease-associated genes and non-disease causing genes. Either functional divergence does not explain the difference between disease-causing and non-disease causing paralogs, which also calls into question the functional compensation hypothesis, or my method did not detect the right signal of functional divergence. An additional possibility lies in the nature of the training set itself. Because the underlying genetic mutation causing many Mendelian diseases is still not known, the non-disease category does contain a significant number of false negative classifications. However, given the size of the datasets involved, and the size of the non-disease category compared to the disease category, I would expect this impact to be moderate.

As mentioned, Rho is the most direct measure of functional divergence output by FunDi, but also considered to be a broad measure of functional divergence, that reflects the total signal of functional divergence within the protein alignment file. This includes sites with both weak and strong signals of divergence. Strongly divergent sites are expected to be the most functionally important when looking at the difference in functions between paralogs. I predicted that a refinement of the functional divergent sites will provide a different answer. As mentioned, to determine whether the site is functionally divergent or not is dependent on a set cut-off threshold (standard threshold is 0.5), which is based on the posterior probability of functional divergence for individual sites. I set different thresholds (0.75, 0.9, and 0.95) and tested the fraction of functionally divergent sites in the alignment files for each of these cut-offs. The cut-offs created restrictions which allowed us to look progressively at only the fraction of sites with the strongest signals of functional divergence (Fig. 7). Results from the analyzed fractions did not shed new light on the stated prediction of Rho.

4.3 Conclusions and Further Research

Measures of functional divergence did not explain whether they are correlated with the propensity of duplicated genes to be involved in Mendelian Disease. Rho value, that represents the functional divergence weight, was not the factor that explained the overrepresentation of Mendelian Disease genes among the duplicated genes in my dataset.

Further research is required in order to provide an explanation for the overrepresentation of the Mendelian Disease gene among the duplicated genes, as Rho failed to explain the observation. One approach will be to test alignment files of each dataset for # of individual amino acid positions. This will allow us to focus on the genome positions that actually differ between genes, instead of looking at the whole genome. Another approach will be calculation of the maximum likelihood values for each of the dataset's branch lengths in the phylogenetic tree. This will provide an additional parameter that can be a factor that explains the observation. Third approach will be to try to look at other measures of functional divergence, like sequence entropy (Schmitt & Herzel, 1997), which is a mathematical approach that measures diversity.

In addition, it is important to note that only one database was analyzed, and it had its own specific categorization. Expanding the research to other databases might provide different results. Furthermore, there is still a lot we don’t know about the disease genes and there is a possibility that some disease genes weren’t categorized as such and are still considered as non-disease ones. This might give us a false-positive error which will affect the results. Aggregation of data from multiple databases and creation of one big database, with defined categorization might also provide more information on the analyzed genes and give better results. Lastly, optimizing FunDi algorithm in order to better handle the more complex relationships between the internal branch lengths- thus improving the program in general.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my honours supervisor, Dr. Daniel Gaston, who introduced me to Bioinformatics and Computational Biology, provided me with an interesting research thesis, and whose support and expertise greatly assisted the research.

I would also like to thank Dr. Andrew J. Roger and Dr. Laura Eme from Roger Lab for their insight on computational phylogenetics and genomics.

In addition, I would like to thank Saint Mary's University for allowing me to conduct this research. Thank you, Dr. Anne Dalziel and Dr. Laura Weir for all your support and positive feedback.

Lastly, I would like to thank my husband, Vladislav Radan, for thousands cups of coffee and a positive attitude.

Figures

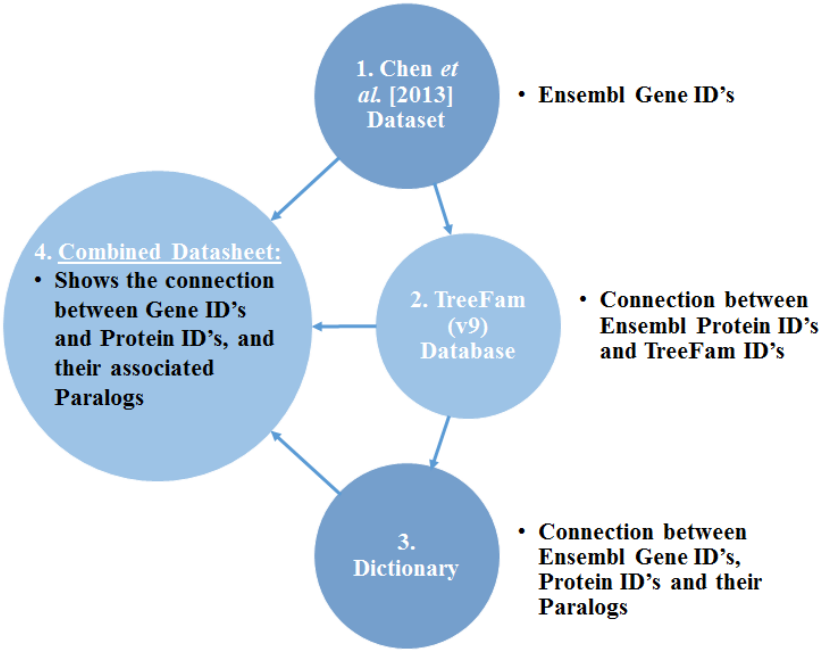

1. Chen *et al.* [2013] Dataset
• Ensembl Gene ID's
2. TreeFam (v9) Database
• Connection between Ensembl Protein ID's and TreeFam ID's
3. Dictionary
• Connection between Ensembl Gene ID's, Protein ID's and their Paralogs
4. Combined Datasheet:
• Shows the connection between Gene ID's and Protein ID's, and their associated Paralogs

Figure 1. Steps taken to create the datasheet of datasets from the original *Chen et al.*, (2013) data, that contained human genes, separated into two categories: disease genes and non-disease genes.

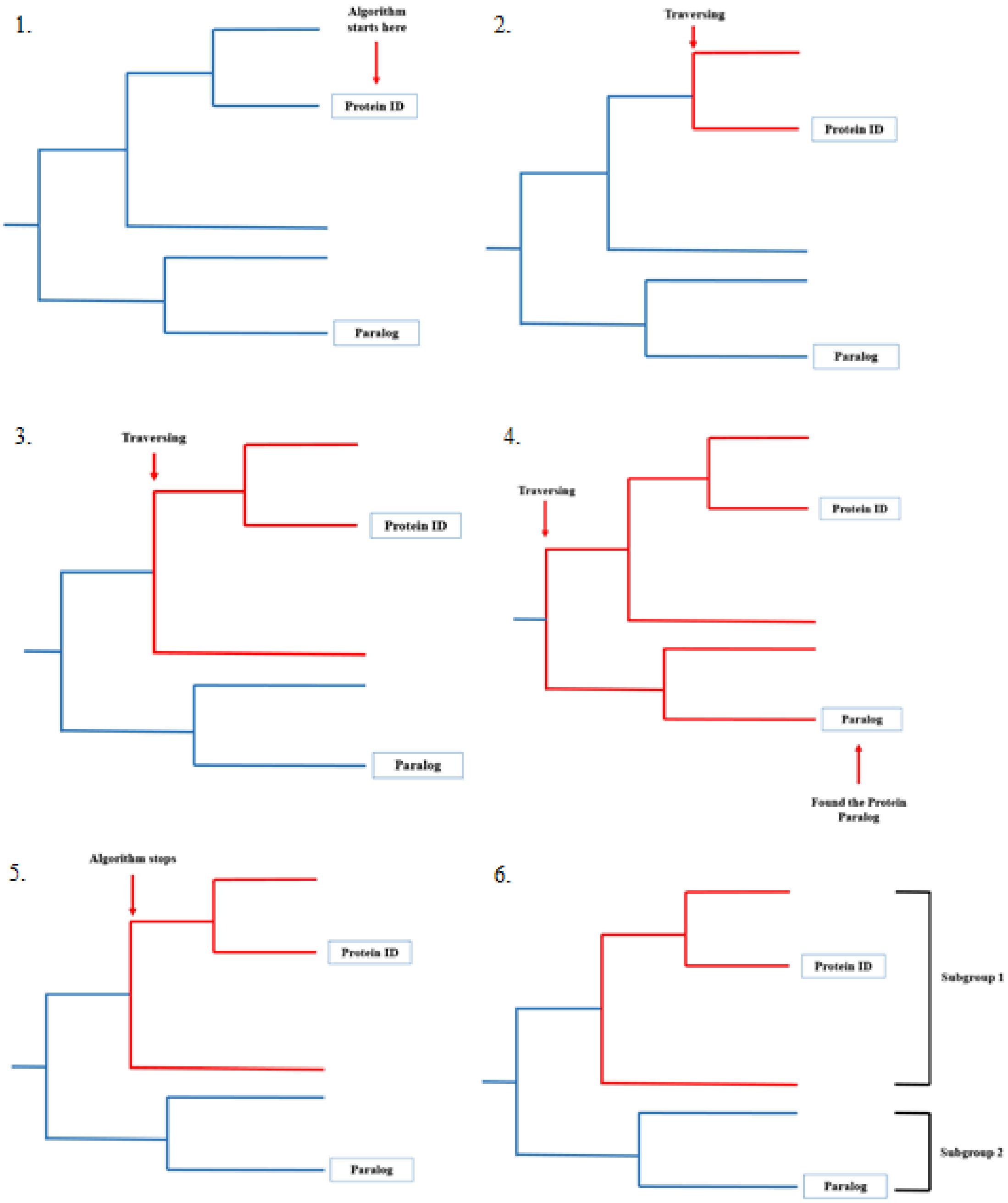


Figure 2. Visualization of the algorithm: (1.) The algorithm starts from a terminal node that contains the protein ID. (2.) Traversing through the tree to an inner node. (3.) Traversing. (4.) Traversing until finding a node that contains a paralog that is associated with the protein ID from step 1. (5.) Go to the previous node. (6.) Separation into subgroups.

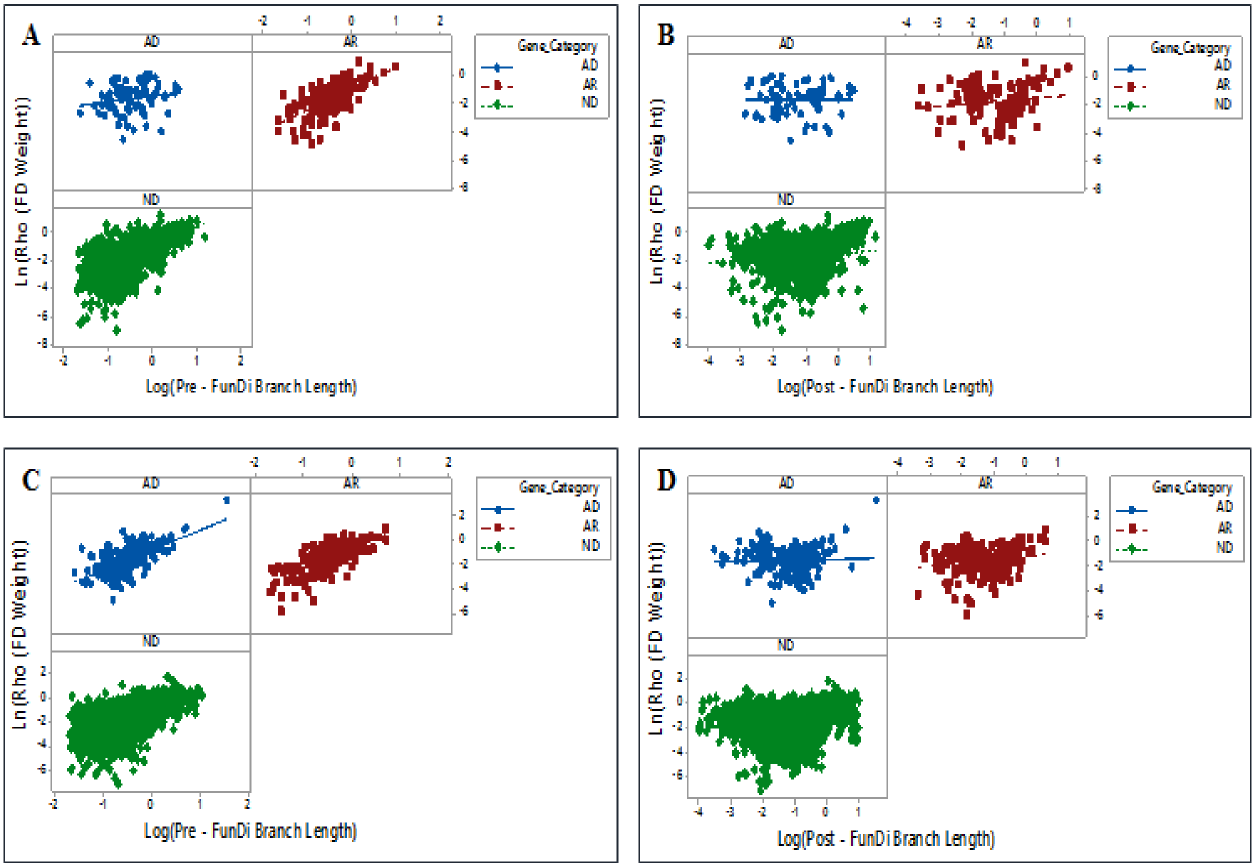


Figure 3. Scatterplots showing the correlation between Rho value (FD Weight) versus Branch Length values of each gene families' phylogenetic tree, before using FunDi and after for all gene categories: AD- Autosomal Dominant, AR- Autosomal Recessive, and ND- Non-Disease genes. (A) Two-member gene family scatterplot results before using FunDi. (B) Two-member gene family scatterplot results after using FunDi. (C) Three or more-member gene family scatterplot results before using FunDi. (D) Three or more-member gene family scatterplot results after using FunDi.

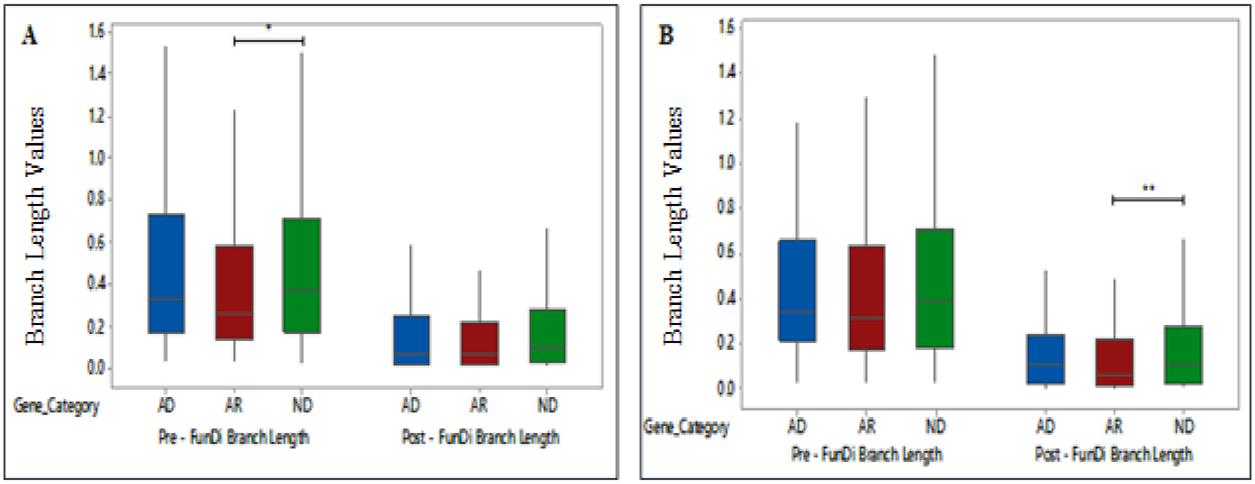


Figure 4. Boxplots showing the comparison between Branch Length values of each gene families' phylogenetic tree, before using FunDi and after for all gene categories: AD- Autosomal Dominant, AR- Autosomal Recessive, and ND- Non-Disease genes. (A) Two-member gene family scatterplot results before and after using FunDi. (B) Three or more-member gene family scatterplot results before and after using FunDi. P-values obtained from a two-sided Wilcoxon-Mann-Whitney U-test for the indicated comparisons as follows: *$P \leq 0.05$; ** $P \leq 0.01$.

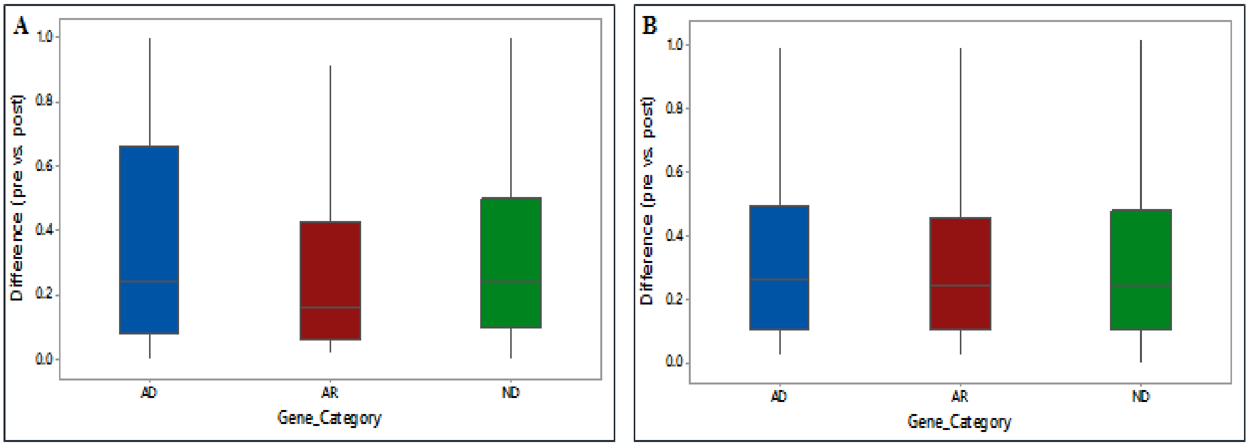


Figure 5. Boxplots showing the difference between the values of the branch lengths (pre vs. post) before using FunDi and after for all gene categories: AD- Autosomal Dominant, AR- Autosomal Recessive, and ND- Non-Disease genes. (A) Two-member gene family boxplot results. (B) Three or more-member gene family boxplot results.

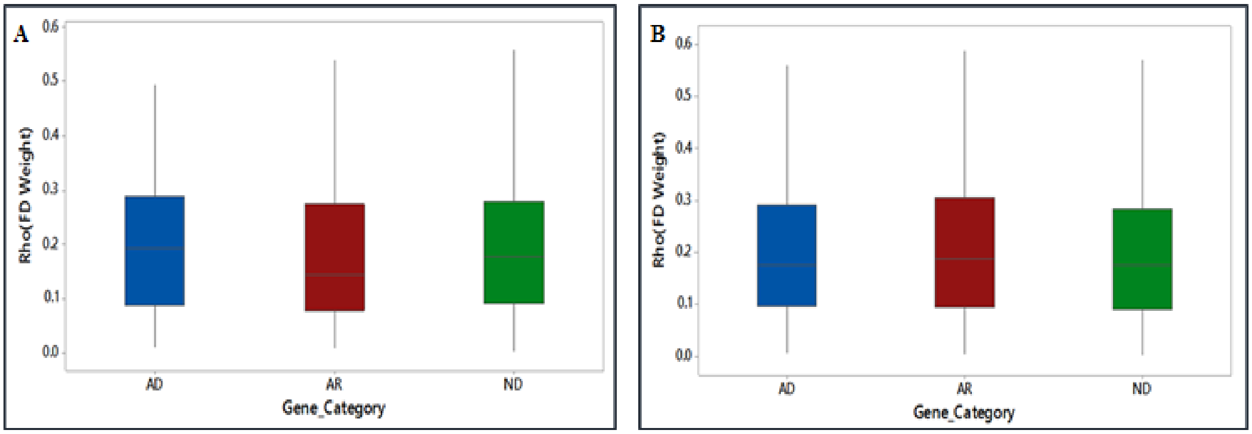


Figure 6. Boxplots showing the comparison of Rho (FD Weight) values for all gene categories: AD- Autosomal Dominant, AR- Autosomal Recessive, and ND- Non-Disease genes. (A) Two-member gene family boxplot results. (B) Three or more-member gene family boxplot results.

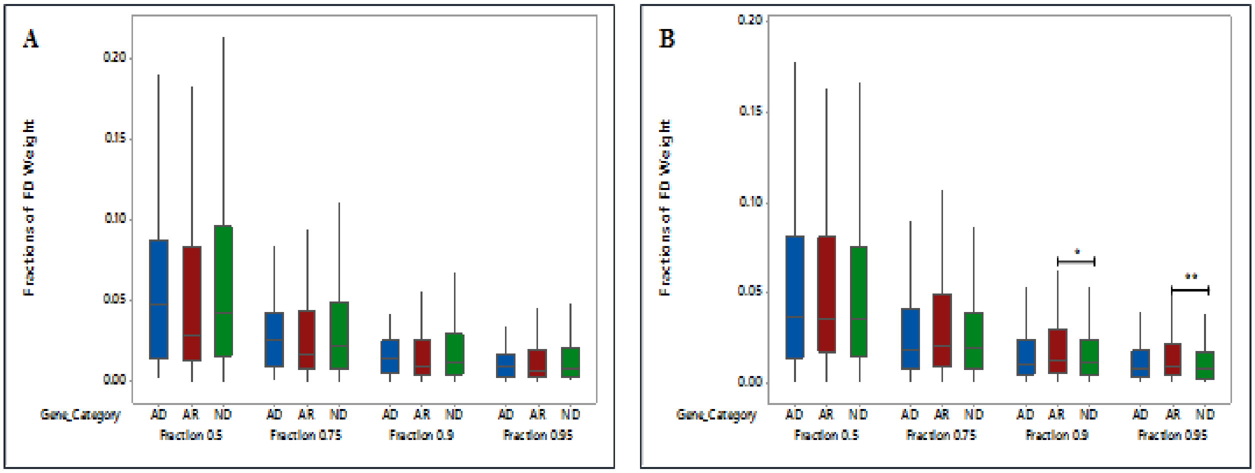
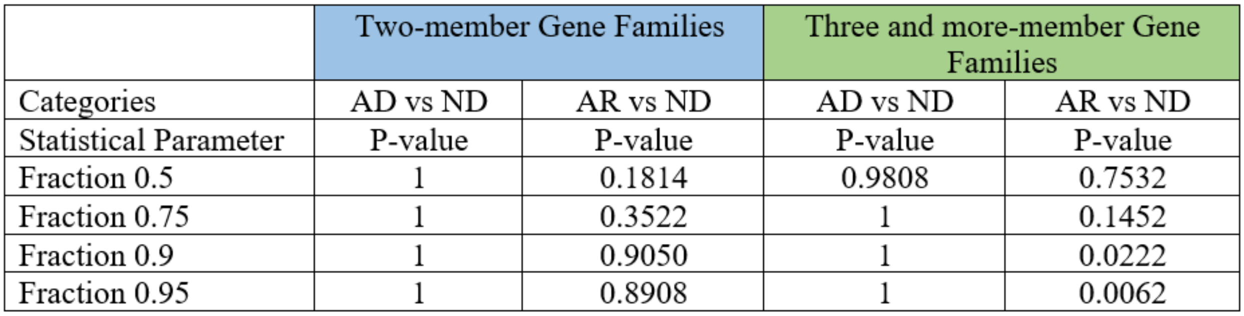


Figure 7. Boxplots showing the comparison between fractions of Rho (FD Weight) values for all gene categories: AD- Autosomal Dominant, AR- Autosomal Recessive, and ND- Non-Disease genes. (A) Two-member gene family boxplot results. (B) Three or more-member gene family boxplot results. P-values obtained from a two-sided Wilcoxon-Mann-Whitney U-test for the indicated comparisons as follows: $*P \leq 0.05$; $** P \leq 0.01$.

Table 1. Visualization of fractions of Rho value comparison between the analyzed categories after Bonferroni correction.



	Two-member Gene Families		Three and more-member Gene Families	
Categories	AD vs ND	AR vs ND	AD vs ND	AR vs ND
Statistical Parameter	P-value	P-value	P-value	P-value
Fraction 0.5	1	0.1814	0.9808	0.7532
Fraction 0.75	1	0.3522	1	0.1452
Fraction 0.9	1	0.9050	1	0.0222
Fraction 0.95	1	0.8908	1	0.0062

* Represents statistically significant P-value (*$P \leq 0.01$)*.

References

Acharya, D., & Ghosh, T. C. (2016). Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC genomics, 17(1), 1.

Alberts, B., Johnson, A., Lewis, J., Walter, P., Raff, M., & Roberts, K. (2002). Molecular Biology of the Cell 4th Edition: International Student Edition.

Armstrong, R. A. (2014). When to use the Bonferroni correction. Ophthalmic and Physiological Optics, 34(5), 502-508.

Blake, D. J., & Kröger, S. (2000). The neurobiology of duchenne muscular dystrophy: learning lessons from muscle?. Trends in neurosciences, 23(3), 92-99.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber.

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics, 33, 228-237.

Chicago

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., ... & Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Molecular biology and evolution, 23(9), 1808-1816.

Carter, C. O. (1977). Monogenic disorders. Journal of medical genetics, 14(5), 316.

Caspi, A., Hariri, A. R., Holmes, A., Uher, R., & Moffitt, T. E. (2010). Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. Focus, 8(3), 398-416.

Chakravarti, A. (2011). Genomic contributions to Mendelian disease. Genome research, 21(5), 643-644.

Chen, W. H., Trachana, K., Lercher, M. J., & Bork, P. (2012). Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Molecular biology and evolution, mss014.

Chen, W. H., Zhao, X. M., van Noort, V., & Bork, P. (2013). Human monogenic disease genes have frequently functionally redundant paralogs. PLoS Comput Biol, 9(5), e1003073.

Chen, W. H., Zhao, X. M., van Noort, V., & Bork, P. (2014). Comments on" Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication". PLOS Comput Biol, 10(7), e1003758.

Chial, H. (2008). Mendelian genetics: patterns of inheritance and single-gene disorders. Nature Education, 1(1), 63.

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., ... & Akdemir, Z. H. C. (2015). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. The American Journal of Human Genetics, 97(2), 199-215.

Conrad, B., & Antonarakis, S. E. (2007). Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu. Rev. Genomics Hum. Genet., 8, 17-35.

Costa, T., Scriver, C. R., Childs, B., Opitz, J. M., & Reynolds, J. F. (1985). The effect of Mendelian disease on human health: a measurement. American Journal of Medical Genetics Part A, 21(2), 231-242.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). 22 A Model of Evolutionary Change in Proteins. In Atlas of protein sequence and structure (Vol. 5, pp. 345-352). National Biomedical Research Foundation Silver Spring, MD.

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol, 3(10), e314.

Dickerson, J. E., & Robertson, D. L. (2011). On the origins of Mendelian disease genes in man: the impact of gene duplication. Molecular biology and evolution, msr111.

Dietz, H. C., & Cutting, G. R. (1991). Marfan syndrome caused by a recurrent de novo missense mutation in the fibrillin gene. Nature, 352(6333), 337.

Dobyns, W. B., Filauro, A., Tomson, B. N., Chan, A. S., Ho, A. W., Ting, N. T., ... & Ober, C. (2004). Inheritance of most X-linked traits is not dominant or recessive, just X-linked. American journal of medical genetics Part A, 129(2), 136-143.

Drachman, J. G. (2004). Inherited thrombocytopenia: when a low platelet count does not mean ITP. Blood, 103(2), 390-398.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. The Annals of Mathematical Statistics, 192-197.

Ebner, B., Burmester, T., & Hankeln, T. (2003). Globin genes are present in Ciona intestinalis. Molecular biology and evolution, 20(9), 1521-1525.

ETE Toolkit - Analysis and Visualization of (phylogenetic) trees. (2016, December 27). Retrieved from http://etetoolkit.org/

Finkbeiner, S. (2011). Huntington's disease. Cold Spring Harbor perspectives in biology, 3(6), a007476.

Garber, K. B., Visootsak, J., & Warren, S. T. (2008). Fragile X syndrome. European Journal of Human Genetics, 16(6), 666-672.

Gaston, D., Susko, E., & Roger, A. J. (2011). A phylogenetic mixture model for the identification of functionally divergent protein residues. Bioinformatics, 27(19), 2655-2663.

Greer, W. L., Riddell, D. C., Gillan, T. L., Girouard, G. S., Sparrow, S. M., Byers, D. M., ... & Neumann, P. E. (1998). The Nova Scotia (type D) form of Niemann-Pick disease is caused by a G 3097→ T transversion in NPC1. The American Journal of Human Genetics, 63(1), 52-54.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Molecular biology and evolution, 18(4), 453-464.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., & Li, W. H. (2003). Role of duplicate genes in genetic robustness against null mutations. Nature, 421(6918), 63-66.

Hagerman, R. J. (2005). Fragile X syndrome. Management of Genetic Syndromes.

Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., & Robertson, D. L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. Genome biology, 8(10), 1.

Hoffmann, F. G., Opazo, J. C., & Storz, J. F. (2011). Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome biology and evolution, 3, 588-600.

Hoffmann, F. G., Opazo, J. C., & Storz, J. F. (2011). Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Molecular biology and evolution, msr207.

Huang, H., Winter, E. E., Wang, H., Weinstock, K. G., Xing, H., Goodstadt, L., ... & Ponting, C. P. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome biology, 5(7), R47.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular biology and evolution, 33(6), 1635-1638.

Hufton, A. L., & Panopoulou, G. (2009). Polyploidy and genome restructuring: a variety of outcomes. Current opinion in genetics & development, 19(6), 600-606.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. Proceedings of the Royal Society of London B: Biological Sciences, 256(1346), 119-124.

Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. PLoS biology, 2(7), e206.

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics, 11(2), 97-108.

IQ-TREE. (2016, December 29). Retrieved from http://www. IQ-TREE.org/

Jill Harrison, C., & Langdale, J. A. (2006). A step by step guide to phylogeny reconstruction. The Plant Journal, 45(4), 561-572.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences: CABIOS, 8(3), 275-282.

Kirkilionis, A. J., Riddell, D. C., Spence, M. W., & Fenwick, R. G. (1991). Fabry disease in a large Nova Scotia kindred: carrier detection using leucocyte alpha-galactosidase activity and an NcoI polymorphism detected by an alpha-galactosidase cDNA clone. Journal of medical genetics, 28(4), 232-240.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. Molecular biology and evolution, 25(7), 1307-1320.

Lin, G. N., Zhang, C., & Xu, D. (2011). Polytomy identification in microbial phylogenetic reconstruction. BMC systems biology, 5(3), 1.

López-Bigas, N., & Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic acids research, 32(10), 3108-3114.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. Journal of genetics, 92(1), 155-161.

Mijović, A., & Mufti, G. J. (1998). The myelodysplastic syndromes: towards a functional classification. Blood reviews, 12(2), 73-83.

Minitab 17. (2016, December 29). Retrieved from http://www.minitab.com/en-us/products/minitab/

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution, 32(1), 268-274.

Norio, R. (2003). Finnish disease heritage I. Human genetics, 112(5-6), 441-456.

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

Ohta, T. (1973). Slightly deleterious mutant substitutions in evoluhypothesis by DNA polymorphism. Genetics, 123, 585-595.

Ohta, T., & Gillespie, J. H. (1996). Development of neutral and nearly neutral theories. Theoretical population biology, 49(2), 128-142.

Olmstead, R. G. (1996). Molecular systematics (Vol. 23). D. M. Hillis, C. Moritz, & B. K. Mable (Eds.). Sunderland, MA: Sinauer Associates.

Qian, W., & Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. Genome research, 24(8), 1356-1362.

Rambaut, A. (2007). FigTree, a graphical viewer of phylogenetic trees. See http://tree. bio. ed. ac. uk/software/figtree.

Rastogi, S., & Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC evolutionary biology, 5(1), 1.

Ratjen, F. A. (2009). Cystic fibrosis: pathogenesis and future treatment strategies. Respiratory care, 54(5), 595-605.

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). Sickle-cell disease. The Lancet, 376(9757), 2018-2031.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science, 273(5281), 1516-1517.

Roos, R. A. (2010). Huntington's disease: a clinical review. Orphanet journal of rare diseases, 5(1), 1.

Ross, C. A., & Tabrizi, S. J. (2011). Huntington's disease: from molecular pathogenesis to clinical treatment. The Lancet Neurology, 10(1), 83-98.

Schmitt, A. O., & Herzel, H. (1997). Estimating the entropy of DNA sequences. Journal of theoretical biology, 188(3), 369-377.

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., & Bateman, A. (2013). TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic acids research, gkt1055.

Scott, S. A., Edelmann, L., Liu, L., Luo, M., Desnick, R. J., & Kornreich, R. (2010). Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. Human mutation, 31(11), 1240-1250.

Singh, P. P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., & Isambert, H. (2012). On the expansion of “dangerous” gene repertoires by whole-genome duplications in early vertebrates. Cell reports, 2(5), 1387-1398.

Singh, P. P., Affeldt, S., Malaguti, G., & Isambert, H. (2014). Human dominant disease genes are enriched in paralogs originating from whole genome duplication. PLoS Comput Biol, 10(7), e1003754.

Smith, N. G., & Eyre-Walker, A. (2003). Human disease genes: patterns and predictions. Gene, 318, 169-175.

Stephens, S. G. (1951). Possible significance of duplication in evolution. Advances in genetics, 4, 247-265.

Swerdlow, D. I., Holmes, M. V., Harrison, S., & Humphries, S. E. (2012). The genetics of coronary heart disease. British medical bulletin, 102(1), 59-77.

Taylor, S. I. (1999). Deconstructing type 2 diabetes. Cell, 97(1), 9-12.

Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., & Sun, F. (2006). Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC genomics, 7(1), 31.

Weeks, D. E., & Lathrop, G. M. (1995). Polygenic disease: methods for mapping complex disease traits. Trends in Genetics, 11(12), 513-519.

Welsh, M. J., & Smith, A. E. (1993). Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. Cell, 73(7), 1251-1254.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution, 18(5), 691-699.

WHO | Genes and human disease. (2016, December 07). Retrieved from http://www.who.int/genomics/public/geneticdiseases/en/index2.html

Wink, M. (2006). An introduction to molecular biotechnology. Wiley-VCH.

Wolfe, K. H., & Li, W. H. (2003). Molecular evolution meets the genomics revolution. nature genetics, 33, 255-265.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular evolution, 39(3), 306-314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution, 11(9), 367-372.

Zielenski, J., Rozmahel, R., Bozon, D., Kerem, B. S., Grzelczak, Z., Riordan, J. R., ... & Tsui, L. C. (1991). Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Genomics, 10(1), 214-228.