# Testing for A Loss of Homozygosity and Compound Heterozygosity Using Human Standing Variation

by

Guang-jian Du

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Andrew S. Allen, Supervisor

_____
Elizabeth Hauser

_____
Raluca Gordan

_____
Paul Magwene

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2017

# Abstract

## Testing for A Loss of Homozygosity and Compound Heterozygosity Using Human Standing Variation

by

Guang-jian Du

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Andrew S. Allen, Supervisor

_____
Elizabeth Hauser

_____
Raluca Gordan

_____
Paul Magwene

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2017

# Abstract

Homozygosity indicates the state of possessing two identical alleles of a particular gene, one inherited from each parent. Homozygous genes are these where both copies share identical alleles at one specific site, this can result from identical mutations on one site or several sites on both copies. In contrast, heterozygous genes are those with different alleles at a given site, compound heterozygous genotype occurs when there is more than one mutation on either copy of the gene but at different sites.

Homozygosity plays a key role in the risk of recessive Mendelian diseases. Because for recessive diseases, it is only when dysfunctional mutations are expressed on both copies of an individual's genome that these variants cause genetic diseases such as cystic fibrosis and phenylketonuria. When a gene has two recessive alleles for the same gene, but the two alleles are different from each other, that is both copies have dysfunctional mutations at different locations, these genotypes are called compound heterozygosity. Both homozygosity and compound heterozygosity could end up in completely knocked out of the function for a selected gene. Therefore, homozygosity and heterozygosity are important risk factors for recessive genetic disorders. It is essential to understand which genes have recessive effects on phenotypes.

Some work has been done on ranking human genes based on their tolerance to functional genetic variants. These studies give a sense of how unusual functional

mutation is in the context of a particular gene. Other work like genetic constraints test for a depletion of rare singleton qualifying variation over expectation, based on mutation rate, using large population databases. This work has been useful for identifying genes with strong dominant effects. However, there is currently no method for identifying recessive intolerant gene.

Here, we propose a method for identifying recessive intolerant genes by looking for a deficit of homozygosity and compound heterozygosity using human standing variations. We first develop a novel computationally efficient and robust statistical model to evaluate the viability of individuals according to the number of copies of a selected gene harboring rare dysfunctional variants, using human standing variation data. Then, we build a general framework to assess whether there's evidence supporting a shift towards a deficit of homozygosity or compound heterozygosity from the distribution of expected genotypes. Third, we apply the statistical Score tests to evaluate the deficit probability of a given gene. Finally, we use a simulation model to further confirm the accuracy of our framework.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations and Symbols

## Symbols

Put general notes about symbol usage in text here. Notice this text is double-spaced, as required.

| | |
|---|---|
| $G$ | A blackboard bold $G$. Genotype. |
| $X$ | A capital $X$. Number of gene copies harboring LoF mutaitons. |
| $V$ | A capital $V$. The viability of a person selected from the population data. |
| $\rho_g$ | The probability of a given genotype, when the number of alleles and the corresponding allele frequencies are known. |
| $\pi_{x\|g}$ | The probability a selected person has $x$ copies of genes affected with LoF mutations, given the genotype is known. |
| $S_i$ | The score contribution for the $i_{th}$ individual's genome information on a certain gene. |
| $T$ | A capital $T$. The statistic score test of a selected gene. |

## Abbreviations

Long lines in the `symbollist` environment are single spaced, like in the other front matter tables.

| | |
|---|---|
| GWAS | Genome Wide Association Study. |
| SNP | Single Nucleotide Polymorphism. |
| NGS | Next Generation Sequencing. |

| | |
|---|---|
| WES | Whole Exome Sequence |
| ENCODE | Encyclopedia of DNA Elements |
| RVIS | Residual Variation Intolerance Score |
| ROH | Regions of Homozygousity |
| LoF | Loss of Function |
| RST | Rao's Score Test |
| LD | Linkage Disequilibrium |
| CDD | Conserved Domain Database |
| ExAC | Exome Aggregation Consortitum, an exome sequencing database from Broad Institute |

# Acknowledgements

First of all, I want to thank Prof. Andrew S. Allen for giving me the opportunity to work on this project and thank Prof. Elizabeth Hauser, Prof. Raluca Gordan, and Prof. Paul Magwene for their indispensable supports and feedbacks.

I also want to thank the members of my work group in Center for Statistical Genetics and Genomics, Meng-qi Zhang, Shuai-qi Zhang for inspiring discussions and excellent ideas. I want to thank Tom Milledge from Duke research computing for his support on high performance computing. Special thanks to Carolyn Zhang at Duke BME department, for proof reading and editing.

Finally, I thank my wife, Yan Tang, for all her support.

# 1

# Introduction

## 1.1 Homozygous and heterozygous genotypes

Homozygous genes are these where both copies share identical alleles at one specific site, this can result from identical mutations on one site or several sites on both copies. In contrast, heterozygous genes are those with different alleles at a given site, compound heterozygous genotype occurs when there are more than one mutation on the gene at different sites, these mutations could happen on either copy of the gene [1]. As shown in **Figure 2.1**, heterozygous genotype could either denotes multi mutations on only one copy of the gene, or these mutations could affect both copies of a given gene. For recessive Mendelian diseases, only when both copies of a gene get affected by dysfunctional mutations, the individual would have high risk to the disease. Under this scenario, both homozygous genotypes and compound heterozygous genotypes could result in a complete 'knock out' of gene function [2].

Different genes have different intolerant abilities to dysfunctional mutations. It is essential to understand which genes have recessive effect on phenotypes. With

the revolutionary development of next-generation sequencing (NGS) techniques, an increasing amount of population data is available for genetic studies, researchers are now able to study variations both within a given gene or domain and throughout the whole human genome [3]. This gives us the opportunity to focuses on associations between genotypes like single nucleotide polymorphisms (SNPs) and phenotypes with dominant, recessive, or additive Mendelian diseases [4, 5]. Most phenotypes of interest are results of multi-gene dysfunction, with more than one mutation influencing a set of genes, whole genome sequence (WGS) and whole exome sequence (WES) can capture these mutations at a low cost [6].

In the same way that different mutations on a gene would play different roles on the protein products, mutations related to dominant, recessive, and additive phenotypes of a gene also result in differential impacts on the biological function of the protein products. For variants correlated to Mendelian recessive disorders, individuals carrying only one copy of a selected gene affected by these mutations would not have diseases or their risk to some certain diseases would be significant low. When both copies of an intolerant gene got affected by dysfunctional mutations, the individual would have relative high risk to some certain diseases. It is essential to develop a framework which could highlights genes that are intolerant to having both copies of selected genes affected by dysfunctional mutations.

## 1.2 Genetic intolerance and human genetic constraints

Progress in scoring the functional impact of mutations along genes, sub domains of genes, and exomes has been made. Residual Variation Intolerance Score (RVIS) ranks genes based on their likelihood to impact diseases [7], similar work has been done on evaluating mutations on non-coding regions, and predicting the genes known to cause disease through up or down regulate expression level based on their depleted

of loss-of-function alleles in the general population [8], and subRVIS evaluates gene sub regions based on their intolerance score to dysfunctional mutations [9]. These methods evaluate the proportion of common functional and dysfunctional variations in each gene or sub region; as a result, identifying genes, sub regions, and non-coding segments appear to be intolerant to mutations. This work provides the opportunity to evaluate the impact of functional variations on certain diseases, (e.g., mutations down regulate or up regulate gene expression, and mutations can significantly change the protein products) [10]. In addition, these metrics allow us to systemically evaluate genes according to their impact on whole genome, as well as their intolerance to dysfunctional mutations. Although these methods worked well on ranking genes according to the intolerance to LoF mutations, and predicting the effects of *de novo* mutations, these methods lost power in ranking genes according to the mutations associated to recessive genetic disorders.

Methods have recently been developed for testing the association of compound heterozygosity with diseases in case-control studies [11, 12]. For example, some work has been done in cancer related homozygous mutations [13, 14] and genome-wide copy-number analysis through microarray to detect regions of homozygosity (ROH) in a genome wide manner [15, 16]. Additionally, recent studies have shown that CRISPR mediated heterozygous mutations play an important role in heterozygous gene repair [17]. A recent study tested the effect of rare compound heterozygous and recessive mutations in case-parent trios [18]. Another way to test a depletion of rare dysfunction variation over expectation using large population database, is to use the human genetic constraints [19].

Some early studies showed that population homozygosity is a powerful parameter to test a departure of alleles from neutrality to heterozygous advantage [20, 21].

Other studies showed that to test a departure of alleles from neutrality to heterozygous advantage, the population homozygosity is a powerful parameter [22]. Watterson demonstrated that the sample homozygosity has a distribution corresponding to the numbers of alleles and the proportion of alleles converages, and the homozygosity is determined by the departure of heterozygous disadvantage [20, 23]. However, this method lost power in testing homozygosity when the allele frequencies are rare, especially when rare variants are more likely to be of recent originated. Another limitation is that, there's no human population data to support these studies till recently. Genes correlated to recessive diseases have been implicated in a large number of studies, but there are few statistical methods for quantifying their impact on survival rate on the population scale, especially when such mutations are rare.

## 1.3 A method for identifying recessive intolerant genes

In this study, we propose a framework to test the loss of homozygosity and compound heterozygosity in recessive disease models on population data. We aim to test on coding regions of genes to assess if there is a loss of homozygosity and compound heterozygosity throughout the whole genome. This study not only develops a framework to evaluate the distribution of human standing variants but also estimate the viabilities for individuals with known number of variants on certain regions of genes. Based on population variants frequencies and viabilities, this work provides a pattern of natural selection on individuals with risk to recessive diseases.

The **core hypothesis** in this study is that there's a deficit of homozygosity and *trans* compound heterozygosity in population data. Under the null hypothesis, the odds ratio of getting some disease for an individual with both copies of a gene affected by a loss-of-function (LoF) mutation is identical to those from individuals

with only one or neither copy affected. Although a lot of work has been done on testing correlations between homozygous and heterozygous variants and diseases, few statistical methods to systematically analyze genes have been explored [7], no work has been done to evaluate genes' intolerance to harboring homozygous and compound heterozygous mutations. In this study, we focus on both coding regions of the genome. This frame work can also be applied to sub regions and sub domains of genes [24], as well as non-coding regions of the genome, which have been shown to play an important role in numerous diseases [25].

There is current no method for identify recessive intolerant genes, in this work, we first develop a novel computationally efficient and robust statistical model to evaluate the viability of individuals according to the number of copies of a given gene harboring rare dysfunctional variants using population data. Next, we build a general framework to assess whether there is evidence supporting a shift towards a deficit of homozygosity or compound *trans* heterozygosity from the distribution of expected genotypes. Under the null hypothesis, there is no difference between viabilities from individuals having less than 2 copies of a gene affected by dysfunctional mutations and individuals having both copies of the gene affected. To infer, the model is based on a shift in homozygosity from the expectation given the assumptions that variants are independent, there's no linkage equilibrium between rare mutations, and gene copies are randomly sampled. Third, we apply the statistical score tests to evaluate the deficit probability of a given gene. Rao's Score Test (RST) is the most powerful test when the true value of parameter $\beta$ is close to alternative parameter $\beta_0$. One of its main advantages is that the test does not require an estimate of the information under the alternative hypothesis or unconstrained maximum likelihood. So, we apply RST to further confirm the confidence and accuracy of our model.

Finally, we use a simulation model to further confirm the accuracy of our framework. To do this, we simulate a set of genotypes based on the allele frequencies from the population data, and then filter the genotypes according to the viabilities of their risk to recessive diseases. With sufficient samples of simulated genotypes, we could apply the developed framework to test the confidence region of viabilities for each gene given the variants frequencies in population data.

# 2

# A Framework to Test A Loss of Homozygosity and Compound Heterozygosity

In our work, we first develop a novel computationally efficient and robust statistical model to evaluate the viability of individuals according to the number of copies of a given gene harboring rare dysfunctional variants using population data. Next, we build a general framework to assess whether there is evidence supporting a shift towards a deficit of homozygosity or compound *trans* heterozygosity from the distribution of expected genotypes. Under the null hypothesis, there is no difference between viabilities from individuals having less than 2 copies of a gene affected by dysfunctional mutations and individuals having both copies of the gene affected.

To infer, the model is based on a shift in homozygosity from the expectation given the assumptions that variants are independent, there is no linkage equilibrium between rare mutations, and gene copies are randomly sampled. Third, we apply the statistical score test to evaluate the deficit probability of a given gene. Rao's Score Test is the most powerful test when the true value of parameter $\beta$ is close to alternative parameter $\beta_0$, one of its main advantages is that the test does not require an

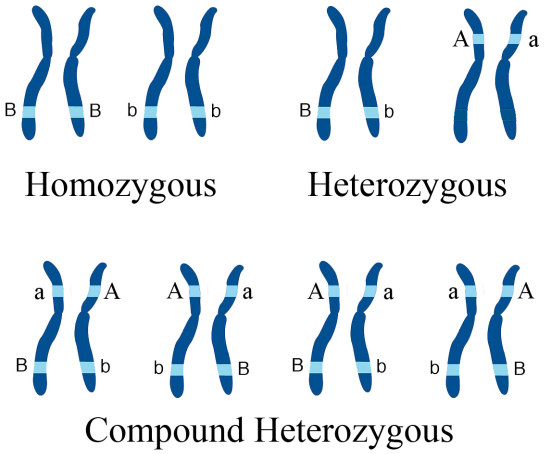


FIGURE 2.1: Example of Homozygous and heterozygous genotypes.

estimate of the information under the alternative hypothesis or unconstrained maximum likelihood. So, we apply RST to further confirm the confidence and accuracy of our model. Finally, we will use a simulation model to further confirm the accuracy of our framework. To do this, we simulate a set of genotypes based on the allele frequencies from the population data, and then filter the genotypes according to the viabilities of their risk to recessive diseases. With sufficient samples of simulated genotypes, we could apply the developed framework to test the confidence region of viabilities for each gene given the variants frequencies in population data.

## 2.1 Models to demonstrate homozygosity and heterozygosity

*Develop a new method to estimate loss of homozygosity and compound trans heterozygosity in the population.*

This study quantifies mutations defined by their predicted functional impact, and their rarity in the population. In order to investigate the correlation of validity and genotypes with both copies of a gene affected by LoF mutations. With more vari-

ants, the more susceptibility a gene would be to have both copies affected. The more important a gene is, the more intolerant that would be to dysfunctional mutations. The idea here is very straight forward, we assume that the population breeds randomly. And that genes are randomly mixed, according to Hardy-Weinberg Equilibrium (HWE) theory. As a result, there should be an increase of homozygous genotypes without natural selection, thus the risk of having recessive Mendelian diseases would increase among the population. When natural selection is involved, individuals harboring two mutated copies of sensitive genes would not likely to survive, resulting in a deficit of homozygosity to dysfunctional mutations in the population. We first develop a new method to estimate the viability ratio of individuals having $x$ copies of gene affected by dysfunctional mutations. Next we develop a statistical method to test the loss of homozygosity and compound trans heterozygosity in the population.

## 2.2 Genes intolerant to recessive mutations

*GWAS relies on linkage between variants and diseases, and lacks the power to test rare variants* GWAS relies on linkage disequilibrium (LD) at the population level, to link genomic loci to disease phenotypes or complex traits. Under the assumption that genetic markers play an important role in the disease risk. Traditionally, GWAS focuses on common variants to common disease (CVCD), recently a lot of work has been done on rare variants correlated to disease risk [26, 27]. For the complex traits, a lot of studies have shown that multiple loci have genome wide statistical significance, thus GWAS may lose the power to identify most of the loci related to the traits. This can occur because of a lack of statistical significance, and the genetic recombination would break the LD [28]. Under the natural selection and the evolutionary theory, variants correlated with disease risk would be eliminated and left at low frequency in

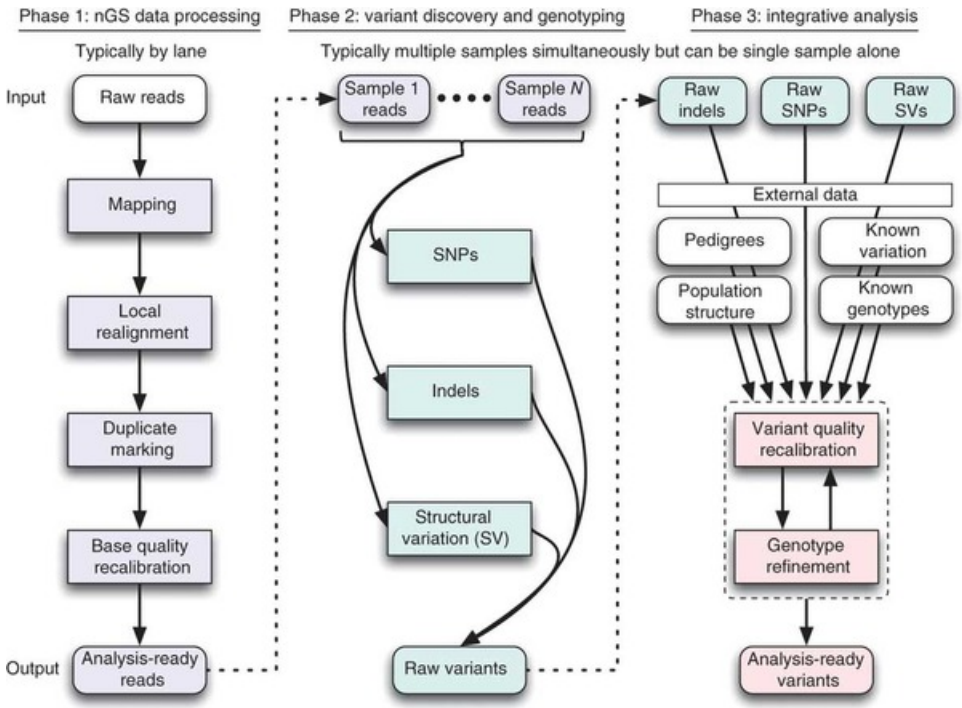


FIGURE 2.2: Framework for variation discovery and genotyping from next-generation DNA sequencing. Nature Genetics 43, 491498 (2011)

the population data, and such disease causing variants would lose LD in comparison to common variants [29]. In this research, we develop a novel framework to evaluate the intolerance of genes to rare mutations. This method focuses on the overall impact of variants on genes to individual viability. This is independent to any simple or multiple diseases models, thus our model is powerful in testing variants without known linkage to disease risk.

*SNP and genotype calling focuses on sets of variants, heterozygous identification have limitations in whole genome analysis* GWAS vary from identifying SNPs which directly change protein products to correlating SNPs related to disease risk [4]. In SNP calling and genotype calling, the first step is to prepare a DNA segment library. In this step, a DNA sequence, either a short gene or a whole chromosome, is digested

by restriction enzymes into short sequences (200-600 bases). This is followed by 3' and 5' end ligation with adapters, the segments are then amplified by PCR and purified. The second step is cluster generation, where the labeled amplified short sequences are loaded into sequencing chips through the hybridization of adapters. The third step is sequencing. Through fluorescence-based chain-termination methods. After sequencing, reads for these short DNA sequences are generated, and the final step is to analyze the reads data, either by *de novo* assembly or by mapping the reads back to a reference genome [30]. Following this mapping, SNP calling and allele identification methods can be used, whereas genotype calling determines the genotype of a specific gene for a person by that SNP site [31]. SNP calling relies on the quality of reads mapping to the reference sequence, which is quantified by Phred quality score for each base, this is defined by $Q_{Phred} = -10 * log_{10}P(error)$ [32, 33], where, Q20 means an error rate of 1%. Early methods to call genotypes depend on the proportion of the non-reference allele, when this lies between 20% and 80%, a heterozygous genotype is called for that specific site, otherwise a homozygous genotype is called [34]. The accuracy of this method depends on the depth of sequencing, it does not work well with low sequencing depths data, like when the coverage is less than 5. In contrast, probability methods based on fixed cutoffs could handle low sequencing data well [35]. Given a genotype $P(G = g)$, the probability of getting that genotype by observing a series of allels $P(Obs)$ on a specific site can be expressed as $P(Obs|G = g) = \frac{P(G = g|Obs) * P(Obs)}{\sum_{G=g'} P(G = g')P(Obs|G = g')}$, and the likelihood for the genotype is calculated according to Bayes' Theorem, which is the posterior probability of genotype $P(G = g)$ [36]. Using known SNP sites from allele frequency databases to calculate the prior probability of each genotype could significantly improve the accuracy of heterozygous site detection. Some work has been done in identifying homozygosity and heterozygosity in diseases like cancer and Hepatitis

C [10, 37, 38]. However, homozygous and heterozygous genotype calling is limited to a small set of SNPs within genes correlated with disease risk, and it is often restricted by familial data. To overcome these limitations, we develop a framework to test variants throughout the whole genome, and rank genes or regions intolerant to harboring both copies affected with these variants. As a result, our work does not rely on genotype calling, instead it relies on known allele frequencies to evaluate the viability of individuals carrying homozygous and compound *trans* heterozygous genotypes based on the prior probability of a given genotype.

*RVIS ranks genes in a systematic way, based on variant scoring metric, but loses statistic power in genes related to recessive disease risk.* Previous work focuses on ranking genes according to their intolerance to LoF mutations, models like RVIS, ncRVIS, and subRVIS can provide a sense of how unusual functional mutations are in the context of a particular gene, which is useful in interpreting family wide studies of genome *de novo* mutations [7, 8, 9]. Although some mutations could result in severe Mendelian diseases, many genes are able to tolerate such alleles as long as only one copy of the gene is affected. Take the Titan gene for example, it has over 10,000 known variants on both copies, of which over 400 are rare dysfunctional variants. Theoretically, there should be a lot of cases with homozygosity and compound *trans* heterozygosity from clinical sequencing data. However we did not see any homozygosity among 58 LoF mutations in Exome Variant Server (EVS), and no homozygosity among 24 LoF variants in 4365 individuals from the Center for Human Genome Variant (CHGV) control dataset. There is no homozygosity observed among offspring in 415 infantile spasms and Lennox-Gastaut Syndrome (IS/LGS) trios. These results show a deficit of homozygosity for this gene in the population. Additionally, little is known about how to quantify the rarity of functional homozygous or compound heterozygous mutations in a particular genetic context. Testing

ht

Table 2.1: Variants on Titan Gene

| Expected and Observed Variants on TTN gene | | | |
|---|---|---|---|
| Constraint from ExAC | Expected | Observed | Constraint Metric |
| Synonymous | 4148.1 | 4252 | z = -1.00 |
| Missense | 10523.4 | 11673 | z = -5.48 |
| LoF | 893.9 | 247 | pLI = 0.00 |

the present of rare mutations related to recessive diseases remains unexplored. As a result, we test the loss of homozygosity and compound heterozygosity in all genes, and rank those intolerant to having both copies affected by rare LoF mutations.

We propose the development of a novel method that highlights genes which are intolerant to having both gene copies affected by deleterious alleles **(Fig 2.1)**. Our approach looks for a deficit in the observed number of individuals in which mutations affect both copies of a gene from the expected population data of human standing variation. The assumption is developed under Mendelian law and Linkage equilibrium, when the allele frequencies for each rare LoF mutation are known and are independently distributed across the gene, any genotypes are randomly sampled. Our framework does not rely on disease related variants, and is powerful in testing rare mutations correlated to human viability. This model does not have the limitation of a certain set of prior selected genes or already know genes related to diseases risk, it considers variants throughout the whole genome. This will provide a systematic view about genes' intolerance to having both copies affected by rare dysfunctional mutations. And most importantly, to our knowledge, this model will be the first one to evaluate the loss of homozygosity and compound heterozygosity for all genes based on human population data.

## 2.3 Innovation, from prior selected genotype studies to systematic studies

Genetic risk factors underlying rare and common variants correlated to recessive disorders remain unexplored. Due to the complex traits and the heterogeneous nature of the correlation between genotypes and phenotypes. Genome wide sequencing and GWAS make it possible to plot a physical map of sequence and variants for each individual, which gives us a new way to study genes associated with disease instead of genotype-phenotype correlation studies based on priori evidence of biological relevance [39].

We introduce a novel framework for systematically studying the intolerance of genes based on population data, this work is independent from any genotype-phenotype correlations or known disease pathways. After applying the model to personal sequencing data, this model could transfer next-generation sequencing (NGS) data from individuals to predict their risk for recessive diseases. It is also powerful in its potential to predict the genetic risk of recessive genetic disorders of offspring given the sequencing data of parents.

# 3

# Hypothesis Testing

## 3.1 Develop a Bayesian model to estimate heterozygous advantage

Homozygosity and heterozygosity are terms used to describe the genotypes on a gene based on whether the alleles are identical at a single locus. A genotype is heterozygous at a gene locus when there are two different alleles, while the gene is homozygous when two alleles are identical at that locus. Since there might be many variants lie on a particular gene, heterozygous are divided into two categories, compound *cis* heterozygous when all variants are identified on only one copy of the gene, and compound *trans* heterozygous when different alleles are identified on both copies of the gene **(Fig 2.1)**. We test our recessive model on rare variants based on population data. For a given gene, let $G$ denote multilocus genotypes of qualifying (deleterious) mutations within a gene. $X$ is the number of gene copies harboring a qualifying variant ($X = 0, 1, 2$). And we refer to $V$ as an indicator for whether that individual was selected as viable in the cast-control sample. We present an algorithm for inferring homozygous and compound heterozygous genotypes from multilocus genotype population data, and develop modifications when both homozygous and compound

heterozygous genotypes are inferred in the below section regrading the genotype inference. In recessive disease models, when $X$ is known, the genotype could be inferred and the likelihood for that individual is be characterized by the probability of viability from observed homozygous and compound heterozygous genotypes. It could also be quantitatively characterized according to the number of gene copies affected by the LoF mutations, i.e., from $Pr(V = 1|X = x)$ to $Pr(G = g|V = 1)$. According to Bayesian theorem, $Pr(G = g|V = 1) = \frac{Pr(V = 1|G = g) * Pr(G = g)}{\sum_{g'} Pr(V = 1|G = g') * Pr(G = g')}$, and the total probability law, which assumes that all individuals in the population are independent. And the genes are randomly sampled, the likelihood contribution of one gene from one individual is written as

$$Pr(G = g|V = 1) = \frac{\sum_{x=0}^{x=2} Pr(V = 1|X = x) * Pr(X = x|G = g) * Pr(G = g)}{\sum_{g'} \sum_{x'=0}^{2} Pr(V = 1|X = x') * Pr(X = x'|G = g') * Pr(G = g')} \tag{3.1}$$

We use $r(x) = \frac{Pr(V = 1|X = x)}{Pr(V = 1|X = 0)}$ to represent the odds ratio of disease conferred by having $x$ gene copies affected relative to the risk of disease when no copies harbor qualifying LoF mutations. Since we focus on recessive genetic disorders (RGD), which could only result from inheriting two defective recessive alleles, we assume $r(0) = r(1) = 1$ and $r(2) = \beta$. Letting $\pi_{x|g} = Pr(X = x|G = g)$ and $\rho_g = Pr(G = g)$, we could simplify the probability equation:

$$Pr(G = g|V = 1) = \frac{\sum_{x=0}^{2} r(x) * \pi_{x|g} * \rho_g}{\sum_{g'} \sum_{x'=0}^{2} r(x') * \pi_{x'|g'} * \rho_{g'}} = \frac{(\pi_{0|g} + \pi_{1|g} + \beta * \pi_{2|g}) * \rho_g}{\sum_{g'} (\pi_{0|g'} + \pi_{1|g'} + \beta * \pi_{2|g'}) * \rho_{g'}} \tag{3.2}$$

## 3.2 Homozygous and compound heterozygous genotypes inferences

Homozygous genotypes are inferred by having both copies of a gene affected by LoF mutations at exactly one locus. This information can be directly taken from Amyotrophic Lateral Sclerosis (ALS) dataset **(from Duke clinical)**. The heterozygous genotypes can be inferred based on the known count of variants on a given gene for a selected person. This information could also be obtained from ALS dataset. Specifically, we follow two principles to infer homozygosity and compound *trans* heterozygosity when rare variants are observed: (1) if homozygous variants are found on that gene, then the genotype of that gene belongs to the case group, the probability that the person has both copies of the gene affected by LoF mutations is 100%; (2) if the number of homozygous variants for that gene is zero, and the number of heterozygous variants is greater than one, then we could calculate the probability that the person has both copies of the gene affected by the LoF mutations, through a heterozygous genotype estimation equation shown below. We note that the principle for case 2 is based on the assumption that the variants are in approximate linkage equilibrium and that, under the alternative $H_0$, any given mutation is equally likely to fall on any given genotype. We specify $\pi_{2|g} = Pr(X_i = x|G_i = g_i)$

$$\pi_{2|g} = \begin{cases} 0 & if\ n_1 \leqslant 1, n_2 = 0 \\ 1 - (\frac{1}{2})^{(n_1-1)} & if\ n_1 > 1, n_2 = 0 \\ 1 & if\ n_2 > 0 \end{cases} \tag{3.3}$$

where $n_1$ and $n_2$ are the observed number of heterozygous and homozygous variants from gene $g$ of a selected person respectively.

We specify the population structure $\rho_g$, which is the multi locus genotype distribution in the general population before case-control selection. We assume gene copies (i.e., genotype) are randomly sampled from general population, there's no

population structure. Since we only focus on rare variants, the variant on a specific locus would follow Bernoulli distribution, while all variants on a gene would follow Binomial distribution, and then the genotype frequencies can be approximated by the list of allele frequencies comprising them. Assume we are calculating $\rho_g$ for a gene with 3 variants, we will have frequencies for all genotype:

$$\rho_g = \begin{cases} Pr(0,0,0) -> & (1-p_1)^2 * (1-p_2)^2 * (1-p_3)^2 \\ Pr(0,0,1) -> & (1-p_1)^2 * (1-p_2)^2 * p_3(1-p_3) \\ Pr(0,0,2) -> & (1-p_1)^2 * (1-p_2)^2 * (p_3)^2 \\ \quad * * * \quad & \quad * * * \\ Pr(2,2,2) -> & (p_1)^2 * (p_2)^2 * (p_3)^2 \end{cases} \tag{3.4}$$

To verify homozygous probability equations, let $\xi_i$ be an estimate of the probability that the $j$th deleterious variant is found on a specific gene, assuming that there are $k$ variants lie on it, such that $j = 1, 2, ..., k$. Then an estimate of the probability that no variant is observed on one copy of the gene is evaluated by $1 - \prod_{j=1}^{k}(1-\xi_j)$. Applying this equation to both copies, we can have a estimate of the expectation of homozygosity and compound *trans* heterozygosity:

$$\sum_{g'} \pi_{2|g'} * \rho_{g'} = (1 - \prod_{j=1}^{k}(1-\xi_j))^2 \tag{3.5}$$

We apply this integration over all genotypes for both copies of the gene consistent with observed genotypes $g$ under the assumption that all variants are equally likely distributed on both copies.

## 3.3 Testing the loss of homozygosity and compound heterozygosity using population data.

Differentiating the log-likelihood with respect to $\beta$ and evaluating under the null hypothesis there's no departure of homozygosity and compound *trans* heterozygosity

among population data. Thus, all individuals with 0, 1, and 2 copies of genes affected with LoF mutations would have the same survival rate. As a result, the odds ratio of having 2 copies of a gene affected by LoF would be equivalent to having 0 copies affected (i.e., $\beta = 1$). Therefore, the score contribution for the $i^{th}$ individual's genome information is defined by

$$S_i = \frac{\pi_{2|g}}{\pi_{0|g} + \pi_{1|g} + \beta * \pi_{2|g}} - \frac{\sum_{g'} \pi_{2|g'} * \rho_{g'}}{\sum_{g'} (\pi_{0|g'} + \pi_{1|g'} + \beta * \pi_{2|g'}) * \rho_{g'}} \tag{3.6}$$

Under the null hypothesis, $\beta = 1$, $\pi_{0|g} + \pi_{1|g} + \beta * \pi_{2|g} = \pi_{0|g} + \pi_{1|g} + \pi_{2|g} = 1$, and $\sum_{g'} (\pi_{0|g'} + \pi_{1|g'} + \beta * \pi_{2|g'}) * \rho_{g'} = 1$, and the final score contribution is:

$$S_i = \pi_{2|g} - \sum_{g'} \pi_{2|g'} * \rho_{g'} \tag{3.7}$$

Here we assume that, in the presence of the homozygous and compound heterozygous genotype $X$, $G$ provides no further information concerning viability. We discuss how $\pi_{x|g}$ is estimated in detail below, but, as a preview, we assume a model where genotypes are randomly sampled and so any loss of homozygosity will be captured through the model relating viability to the number of dysfunctional gene copies. Under the null hypothesis $E(S_i) = 0$ and we have statistic test $T = \frac{(\sum_1^n S_i)^2}{n * Var(S_i)}$, this will asymptotically follow $\chi_1^2$ distribution, as the number of observed population becomes large. According to the definition, the $P$-value is the smallest $\alpha$-level at which $H_0$ can be rejected based on the observed value of the test statistic. Therefore, we compute the $P$-value by summing over the distribution of $S_i$. Specifically, we enumerate the likelihood for all possible combinations of $\sum_{g'} \pi_{2|g'} * \rho_{g'}$ by integrating over all genotypes consistent with observed genotypes under the assumption that all variants on a given gene are equally likely distributed.

In the statistical test, we test the $i_{th}$ person's contribution to the total score:

$\frac{S_i^2}{n * Var(S_i)} \sim \chi_1^2$. This works fine with long genes with over 400 LoF rare variants. However, some genes have a limited number of qualified variants, and thus the count for homozygous and compound heterozygous genotypes would be zero. Under this scenario, the $Var(S_i)$ would be zero, and the statistical test does not work in this case. We apply RST from Equation.2. The **Likelihood function** $L(\beta) = Pr(G = g|V = 1) = \frac{\rho_g * (\pi_{0|g} + \pi_{1|g} + \beta\pi_{2|g})}{\sum_{g'} \rho_{g'}(\pi_{0|g'} + \pi_{1|g'} + \beta\pi_{2|g'})}$ . $\frac{\partial log[L(\beta)]}{\partial \beta}$ denotes the first derivative of the log likelihood over $\beta$, taken at $\beta = 1$, which is our null hypothesis, we could have score function:

$$U(\beta) = \frac{\partial log[L(\beta)]}{\partial \beta}|(\beta = 1) = \pi_{2|g} - \sum_{g'} \pi_{2|g'} * \rho_{g'} \tag{3.8}$$

To get the Fisher Information, we take the second derivative of the log-likelihood function. The negative expectation of the 2nd derivative log-likelihood is $I(\beta) = -E(\frac{\partial^2 log[L(\beta)]}{\partial \beta^2})$. We calculate the 1st and 2nd derivative of the log-likelihood:

$$\frac{\partial^2 log[L(\beta)]}{\partial \beta^2} = \frac{\partial^2 log[\rho_g * (\pi_{0|g} + \pi_{1|g} + \beta\pi_{2|g})]}{\partial \beta^2} - \frac{\partial^2 log[\sum_{g'} \rho_{g'} * (\pi_{0|g'} + \pi_{1|g'} + \beta\pi_{2|g'}])}{\partial \beta^2} \tag{3.9a}$$

$$= -\frac{\pi_{2|g}^2}{(\pi_{0|g} + \pi_{1|g} + \beta\pi_{2|g})} + \frac{(\sum_{g'} \rho_{g'} * \pi_{2|g'})^2}{\sum_{g'} \rho_{g'}(\pi_{0|g'} + \pi_{1|g'} + \beta\pi_{2|g'})} \tag{3.9b}$$

Under the null hypothesis, $\beta = 1$, $\pi_{0|g}+\pi_{1|g}+\beta\pi_{2|g} = 1$, and $\sum_{g'}(\pi_{0|g'}+\pi_{1|g'}+\beta\pi_{2|g'}) = 1$. So

$$\frac{\partial^2 log[L(\beta, x)]}{\partial \beta^2} = (\sum_{g'} \rho_{g'} * \pi_{2|g'})^2 - \pi_{2|g_i}^2 \tag{3.9c}$$

The information matrix (vector) function is given by

$$I(\beta) = -E(\frac{\partial^2 log[L(\beta)]}{\partial \beta^2}) = E[\pi_{2|g}^2 - (\sum_{g'} \rho_{g'} * \pi_{2|g'})^2] \approx \frac{1}{n}\sum_{1}^{n}(\pi_{2|g}^2 - (\sum_{g'} \rho_{g'} * \pi_{2|g'})^2) \tag{3.10}$$

This is the new variance we will use in the Rao's score test. The statistic test under the null hypothesis can now be defined by

$$S(\beta = 1) = \frac{U(\beta = 1)^2}{I(\beta = 1)} \sim \chi_1^2 \tag{3.11}$$

To calculated the **Rao's Score Test** with 3047 observations, we will use $(\sum_1^{3047} S_i)^2 = (\sum_1^{3047} \pi_{2|g} - \sum_{g'} \pi_{2|g'} * \rho_{g'})^2$ as the score, and $V = \sum_1^{3047} I(\beta_0) = \sum_1^{3047}[\pi_{2|g_i}^2 - (\sum_{g'} \rho_{g'} * \pi_{2|g'})^2]$ as the estimated variance.

The test will follow Chi-square distribution with 1 degree of freedom.

$$\frac{(\sum_1^{3047} S_i)^2}{\sum_1^{3047} I(\beta_0)} = \frac{[\sum_1^{3047}(\pi_{2|g} - \sum_{g'} \pi_{2|g'} * \rho_{g'})]^2}{\sum_1^{3047}[\pi_{2|g}^2 - (\sum_{g'} \rho_{g'} * \pi_{2|g'})^2]} \sim \chi_1^2 \tag{3.12}$$

We can calculate $\pi_{2|g}$ base on the known homozygous variants $n_1$ and compound heterozygous variants $n_2$ counts, and evaluate $I(\beta)$ based on the clinical ALS dataset.

## 3.4 Simulating the viability of individuals given the population allele frequencies.

Since some genes do not have any observed homozygous and heterozygous variants in the clinical dataset, in the statistic t-test we result in $Var(S_i) = 0$; and in the RST, the $\pi_{2|g} = 0$ giving $(\sum(S_i))^2 / \sum(I(\beta)) = n$, losing the power to test our model. We simulate a set of genotype samples for a given gene's allele frequencies, with 200,000

genotypes per sample. We filter the genotypes based on corresponding viabilities, only keep the genotype when it is viable. From the simulation process, we could obtain the count numbers of homozygous SNPs and heterozygous SNPs distributed alone the gene, and the population frequencies for each allele in the simulated samples; And then we could pass the simulated $n_1$, $n_2$, $\rho_g$and $\pi_{2|g}$ to the model we developed, to further estimate the confidence we have to reject the null hypothesis. Further details about the simulation will be discussed in Chapter 5.

# 4

# Testing Loss of Homozygousity on Titan Gene

## 4.1 Methods

### *4.1.1 Data sources*

*Population level data* We use population data from Broad Institute; Exome Aggregation Consortitum (ExAC) is an exome sequencing database from a wide variety of large scale sequencing projects. The dataset spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies [40]. To get qualified allele frequencies of all variants on a given gene, we first obtain the gene IDs and exon frames for each gene from CCDS.r14, this file contains identical protein annotations on the reference human genomes according to a CCDS ID [41]. Next, we pull down information of all LoF variants for each gene from ExAC, with a java based net-crawling software. Third, we use the exon frame as a filter to screen the variants data, only keeping the variants on exons. Since we focus on rare mutations, we set the threshold to 10%. In this way, we will get all rare LoF variants on exons for each gene. The expected homozygous and compound *trans* heterozygous probability based on given variants could be estimated by *equation* 5.

*Personal level data* The personal variants data is obtained from the ALS dataset. It contains genotype data for 3047 individuals, with counts for homozygous and heterozygous variants in each gene. After obtaining the zygousity counts for each gene, we use *equation* 3 to calculate the observed $\pi_{2|g}$ in ALS data set.

### *4.1.2 Statistical analysis*

Using the population data obtained from ExAC and individual data obtained from ALS, and evaluating under the null hypothesis $\beta = 1$, we can calculate the marginal contribution of each individual to the score using *equation* 7, and then apply statistical test to evaluate the observed likelihood of being homozygous or compound *trans* heterozygous to the expected likelihood.

## 4.2 Testing loss of homozygosity and compound heterozygosity on Titan gene

We test the loss of homozygosity and compound heterozygosity in Titin (TTN) Gene. The TTN gene provides instructions for making a very large protein called Titin. This protein plays an important role in muscles the body uses for movement (skeletal muscles) and in cardiac muscles. Slightly different versions (called isoforms) of titin are made in different muscles.

Within muscle cells, Titin is an essential component of sarcomeres. Sarcomeres are the basic units of muscle contraction; they are made up of proteins that generate the mechanical force needed for muscles to contract. Titin has several functions within sarcomeres. One of its main functions is to provide structure, flexibility, and

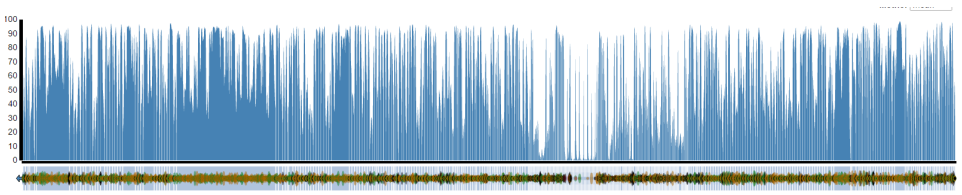

FIGURE 4.1: Variants along Titan gene, figure from ExAC website.

stability to cell structures. Titin interacts with other muscle proteins, including actin and myosin, to keep the sarcomere components in place as muscles contract and relax. Additionally, researchers have found that titin plays a role in chemical signaling and in assembling new sarcomeres.

Table 4.1: Score Tests on Titan Gene

| Test | Score | Variant | Value | P-Value |
|---|---|---|---|---|
| Stats T-Test | $\sum(Si)^2$=221.93 | Var = 0.0028 | 25.96 | 1.74E-7 |
| Rao's S-Test | $\sum(Si^2)$=8.6189 | $I(\beta)$ = 0.00011 | 25.73 | 1.96E-7 |

As shown in **Table 4.1**, TTN is a large gene, 105kb in length and flooded with over 10,000 missence variation. Because the variants are in large quantities, its RVIS score is in the 99th percentile of tolerance, which indicates the gene is sensitive to LoF mutations. However, we observed complete deficit of homozygous LOF mutations among the 58 LOF mutations in EVS, there is no homozygous variants among the 24 LOF mutations in 4365 CHGV control cases; Also, no homozygous or compound heterozygous variants found among offspring in 415 IS/LGS trios.

We apply the statistical test and the RST to TTN gene, given the rare LoF variants allele frequencies from population data and the zygousity counts from ALS data. We also simulate the viabilities for individuals with both copies of the gene affected and the viabilities for individuals with less then 2 copies affected using (*equation* 13). From the Statistic T-test, P-value = 1.9612e-06; and from the RST,

P-value = 1.7422e-07. The null hypothesis in our simulation is that the viabilities are equal for individuals with 0 or 2 copies affected. We simulated 1200 samples under the null hypothesis and applied our model developed in **Chapter 2**, and the distribution of $P$-values followed almost uniformly between 0 and 1 **Fig 5.2**. Based on these results, we can reject the null hypothesis, this implies a certain pattern of deficit homozygosity and compound heterozygosity in population data for the Titan gene.

# 5

# Viability Simulation under the Null Hypothesis and under the Alternative Hypothesis

Model simulation in biostatistics and population genetics could play an key role in helping us better understand the impact of various evolutionary and demographic scenarios on sequence mutations and diseases correlations to variants. A good simulation study could help investigators to develop better statistical models and test the hypothesis based on both null hypothesis and alternative hypothesis.

## 5.1 Methods

### *5.1.1 Simulating genotypes based on allele frequencies*

In this study, we develop a simulation frameworks, based on the allele frequencies of all known LoF alleles on a given gene, and the risk of potential recessive Mendelian diseases, as well as the viabilities correlated to an selected individual's genotype. Assuming the current simulation parameters in our model would fall under the framework we develop in **Chapter 2**, we adjust the viabilities for individuals with both copies of a given gene affected by at least one LoF mutations, and compare them with

respect to their evolutionary and demographic scenarios. Additionally, we address some limitations in our simulation algorithm and discussed future challenges in the development of more powerful simulation models.

Take one random gene with 3 variants for example, there are $3^3$ genotypes. Assuming these variants on the given gene G are randomly distributed and there is no LD among them. For each variant site, the probability of having mutations would follow Bernoulli distribution. While for the whole gene, the probability of genotypes would follow Binomial distribution:

$$\rho_g = \begin{cases} Pr(0,0,0) -> & (1-p_1)^2 * (1-p_2)^2 * (1-p_3)^2 \\ Pr(0,0,1) -> & (1-p_1)^2 * (1-p_2)^2 * p_3(1-p_3) \\ Pr(0,0,2) -> & (1-p_1)^2 * (1-p_2)^2 * (p_3)^2 \\ \quad * * * & \quad * ** \\ Pr(2,2,2) -> & (p_1)^2 * (p_2)^2 * (p_3)^2 \end{cases} \tag{5.1}$$

### *5.1.2 Algorithm of simulating genotypes based on allele frequencies*

For a selected gene with K variants, there are $3^K$ genotypes, we could randomly simulate one genotype based on the allele frequencies along the gene. We first download all allele frequencies for LoF mutations from ExAC website with a Java-based new crawler software (see appendix). The allele frequencies could be stored in an array: $a_1, a_2, a_3.....a_k$. After that, we generate a array of uniformly distributed random numbers (0-1), $u_1, u_2, u_3....u_k$. By comparing the random numbers generate to the allele frequencies one-by-one, we could simulate one copy of the gene, with detail information on which SNP site there's an allele or not. Repeat the random number generation and comparation, we could have another list of random numbers, $v_1, v_2, v_3....v_k$, and another simulated copy of the gene.

After generating two array lists of random numbers, we set the count numbers of homozygous SNPs (n2) and heterozygous SNPs (n1) as 0. Loop through the indics of array lists, compare between generated random numbers and allele frequencies at the same index, if $v_i > a_i$ *and* $u_i > a_i$ we get one homozygous SNPs on both copies, n2++. If $u_i > a_i$ *or* $v_i > a_i$, we get a heterozygous SNP on one copy, n1++. Meanwhile, we could specifically mark whether the genotype we simulated has both copies of the gene affected by LoF mutations or not. Thus, we could estimate the viability for that simulated genotype accordingly, if the genotype only has 1 or 0 copies affected, the viability could be 95%, otherwise, the viability could be a percentage less than 95%. In the end, we filter the genotype with viabilities corresponding to number of gene copies **X** harboring dysfunctional mutations. Only keep the gene under the viability threshold.

For each set we simulate 1000 samples, for each sample we simulate 200K genotypes. For each sample, after getting 200k genotypes simulated, we have 200K counts numbers of n1s and n2s. We calculate $\pi_{2|g}$ for each genotype, based the n1 and n2 (according to equation: 3). We also calculate the new allele frequencies based on all the total number of mutations and total number of genotypes simulated. Thus, a new $\sum_g \pi_{2|g} * \rho_g$ could be developed, based on the new allele frequencies **(Equation 4)**.

For each sample, we apply the Score test to get the P-values. Repeating this for 1000 times, we get 1000 P-values, and calculate the proportion of P-values less than 0.05. This is the proportion out of the 1000 datasets which we have confidence to reject the null hypothesis, which also indicates the proportion of datasets shown significant deficit of homozygosity and compound heterozygosity on simulated data.

The equation of viability for an individual with a known genotype could be expressed as below:

$$Pr(V = 1|x) = \frac{1}{1 + e^{\alpha+\beta*I_{(x=2)}}} \tag{5.2}$$

Where $I$ is an indicator factor, only when $x = 2$ can $I = 1$, otherwise $I = 0$. Only genes with $V = 1$ will be kept for next step. We repeat the procedure until 100,000 genotypes are obtained. We use these simulations to investigate the relationship between power and parameters under the hypothesis for a certain sample size. By calculating the simulated $\pi_{2|g}$ and perform the same statistic text, we could evaluate at what $\beta$ range we could get similar P-values for genes with sufficient homozygous and heterozygous counts in the clinical dataset, and then do the simulation for genes lacking the sufficient number of homozygous and heterozygous variants counts.

In our simulation study, we get a simulated pattern for the count of homozygous and compound heterozygous genotypes for the Titan gene from the simulation. This pattern is similar to what we observed in the Duke clinical ALS data set. As shown in **Equation 5.2**, we set $\alpha = -2.2$ and adjust $\beta$ ranging from 0 to 5.0, we get viability for people with 0 or only 1 gene copy affected with rare dysfunctional mutations of 95% and viability of people with both copies of gene affected ranging from 95% to 5%. As the power of our simulation increases, viability for individuals with both copies of a gene affected would decrease, the statistic test scores and proposition of P-values less than 0.05 increase significantly.

## 5.2 Simulation under the null hypothesis

Under the null hypothesis, all individuals would share the same viability no matter how many number of gene copies get affected by LoF mutations. We simulate 1200

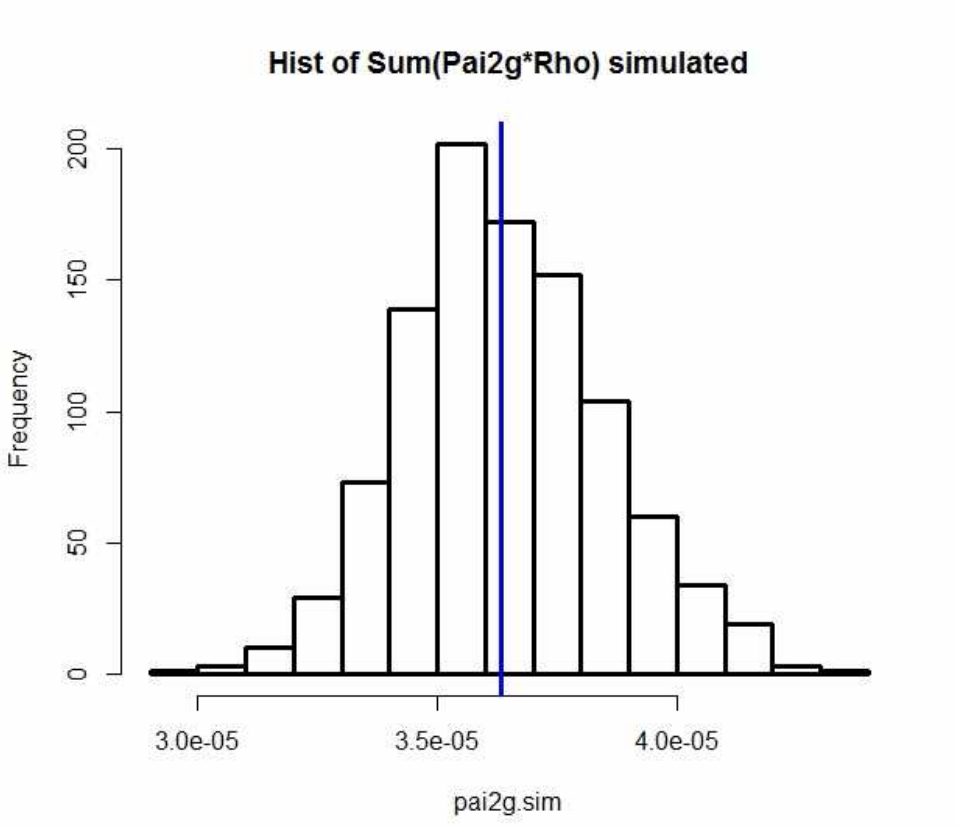


FIGURE 5.1: Distribution of $\pi_{2|g}$ from simulated genotype samples under the null hypothesis that there's no viable advantage for individuals with 0 or 1 copies of genes affected over individuals with 2 copies of genes affected by LoF.

samples of genotypes based on all 416 allele frequencies on Titan gene, with 200,000 genotypes for each sample. Each genotype simulation is based on the qualified total variants on exons and the corresponding population allele frequencies, these variants are sampled to get a random genotype; After the genotype is formed, the viability status of the gene could be generated by a Bernoulli random variable with 'success' probability given by the viability probability equation, $r(x) = \dfrac{Pr(V = 1|X = x)}{Pr(V = 1|X = 0)} = 1$. The viabilities for individuals have 0, 1, or 2 copies affected would be equal.

After getting 200k genotypes for each sample, we calculate the count $n1$, $n2$, and

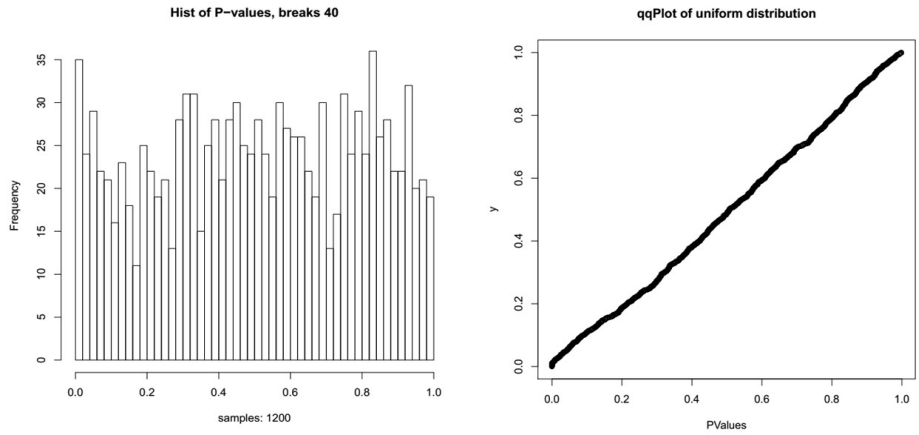


FIGURE 5.2: Distribution of P-values from simulated genotype sampels under the null hypothesis that there's no viable advantage for individuals with 0 or 1 copies of genes affected over individuals with 2 copies of genes affected by LoF.

the new expected $\pi_{2|g}$. As shown in **Fig 5.1**, the P-values we get from 1200 samples follow uniform distribution.

## 5.3 Simulation under the alternative hypothesis

Under the alternative hypothesis, individuals with less than 2 copies of the gene affected would have viability advantage over individuals with both copies of the gene affected by Lof mutations. The simulation procedure as shown in subsection 5.1, we simulate a set of samples, for each sample there are 500K genotypes. We calculate the count numbers n1 and n2 for each genotype, and the new $\pi_{2|g}$, the main difference is that, when filtering the viability of each genotype, we set the viability for genotypes with both copies of the gene affected at a lower percentage. Following the hypothesis, $r(x) < 1$, we set the viability for individuals with 0, or 1 copy affected at 0.95, for individuals with 2 copies of the selected gene affected, we test the viability from 0.95 to 0.6.

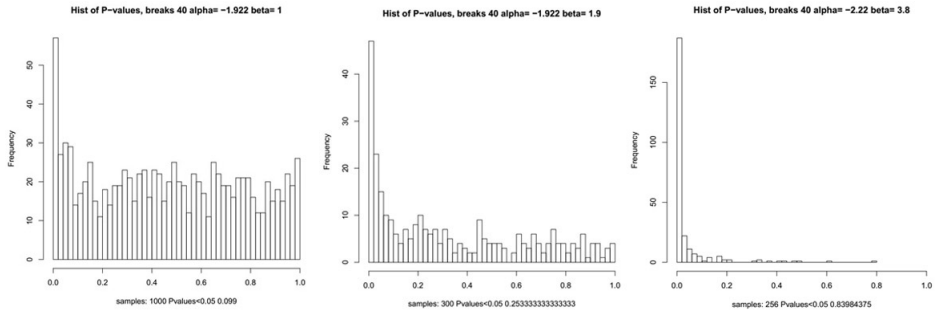


FIGURE 5.3: Distribution of P-values from simulated genotype samples under the alternative hypothesis that there's certain viable advantage for individuals with 0 or 1 copies of genes affected over individuals with 2 copies of genes affected by LoF.

Three simulation results are shown in **Fig 5.2**. As we increase the power of our simulation, and reduce the viabilities for genotypes with both copies affected, we observe a significant shift of the proportion of P-values from uniform distribution to right screwed distribution.

## 5.4 Results and conclusions

To facilitate our study in the deficit of homozygosity and compound heterozygosity, it is important to design and develop simulation model with the capability to accurately generate genotypes under various allele frequency data, and consider the viability under natural selection pressure, and then pass the viable genotypes to the model we developed to further test our hypothesis. In our simulation, a number of simulated genotypes under different natural selection pressure are simulated. Score tests are performed for each simulated sample, and the corresponding P-values for each sample set are calculated based on the new allele frequencies, homozygosity counts and heterozygosity counts from the simulated genotypes data. Each P-value

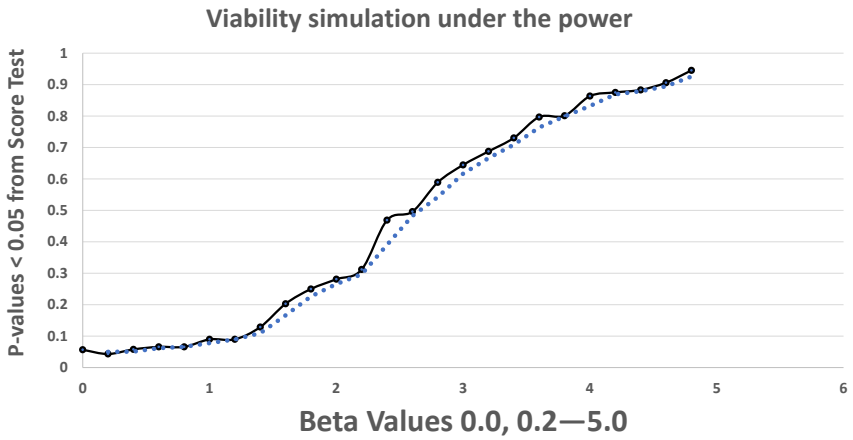


FIGURE 5.4: Distribution of P-values from simulated genotype samples under the alternative hypothesis that viable advantage for individuals with 0 or 1 copies of genes affected over individuals with 2 copies of genes affected by LoF.

indicates confidence in rejecting the null hypothesis, the proportion of these P-values would show whether or not there is a certain pattern of selection advantage for genotypes with less than 2 copies of gene affected.

We find there's a certain departure of homozygosity and compound heterozygosity in some intolerant genes (Titan gene for example). The viability simulation confirms there is a pattern of reduce for individuals with both copies affected by LoF variants, as shown in **Fig 5.3** and **Fig 5.4**.

It appears that many limitations remain in current simulation algorithms, and there are several main challenges for further simulation studies in population genetic and bioinformatics. First of all, natural selection and evolution is a very complicated process which happens in a extremely long historical category. Meanwhile, it involves inheritance from ancestries and *de-novo* mutations, both genetic pressures and en-

vironmental factors play key role in the development of human genome. All these impactors can only be approximated and most of them have not been considered in our simulations. More work will be done on genes with different number of alleles and genes with different range of allele frequencies.

# 6

# Discussion

Homozygous and heterozygous genotypes play key role in the risk of recessive Mendelian diseases, the viability of individuals without Mendelian diseases has advantage over that of individuals with Mendelian diseases, thus it is essential to develop a model to quantify the relationship between homozygosity and compound heterozygosity and the viabilities accordingly. Some pure theoretical studies had been done on testing the loss of homozygosity against heterozygosity when there's selection advantage of heterozygosity [20, 23]. However, these work was done 30 to 40 years ago, when there was no genetic data or population data, and the computing power was not as good as we have today.

Due to the natural selection, individuals with genes harboring LoF mutations are less likely to pass their genome information to next generation, as a result, the mutations which could cause serous disease are rare. Genes also show different tolerance abilities to mutations on them, some studies like RVIS has shown positive results to rank genes according to their intolerance to LoF mutations. Genetic constraints test for a depletion of rare (singleton) qualifying variation over expectation (based on

mutation rate) using large population data base (ExAC). Both RVIS and Constraint focus on priority dominant effects. Both work well for *de novo* mutations. However, for recessive Mendelian diseases, individuals have high risk only when both copies of gene got affected. Some clinical data showed that some genes could tolerate Loss-of-Function variants just fine, as long as these alleles only affect one copy of the gene.

In this study, we have developed a framework that highlights genes that are intolerant to having both gene copies impacted by Loss-of-Function mutation. Based on our model, we have tested a hypothesis that under the natural selection and evolutionary pressure, the individuals with both copies of some certain genes affected by LoF mutations would have lower viability than those with 0 or only 1 copies affected. We test our hypothesis on population data with Rao's Score Test. We find there's a deficit of homozygosity and compound heterozygosity in some certain genes (Titan gene for example). To further confirm our hypothesis, we simulate a series sets of genotype samples, perform the same RST for each sample, and calculate the P-values to reject the null hypothesis. The viability simulation results confirm that there is a pattern of population reduction for individuals with both copies affected by LoF variants.

This result is novel and notable because it is the first study focusing on deficit of homozygosity and compound heterozygosity with rare LoF mutations. It also demonstrates the practical applicability of evaluation on genes intolerant to having both copies affected by LoF, even in situations where the distribution of the alleles on the gene is unknown, our model could give an accurate estimation of the departure of such pattern.

Moving forward with our model, to fully realize the potential of this method, more

testing should be done to further determine if the hypothesis testing will remain efficient for other genes with different distributions of allele numbers and different allele frequencies. Since some genes are relatively short, with few known alleles on them, so from our clinical data there's no mutation observed on these genes, thus it is impossible to test the loss of homozygosity and compound heterozygosity for them. As more and more clinical data become available for research works, we will be able to perform our model and test on other genes in the future.

In conclusion, we have developed a framework to test the hypothesis that individuals with less than 2 copies affected by LoF mutations would have viable advantage over those with both copies affected. We hope that this work will make contribution to current studies like RVIS, which would give us predictions about how sensitive a gene would be to LoF mutations, in the categories of both dominate and recessive Mendelian diseases.

# Appendix A

## Appendix Code and Algorithms

### Appendix

**Verifying equation 3.4 and 3.5** The Java code for verifying two methods calculating the probability of homozygosity **Equation 3.4 and 3.5** could be found: `https://github.com/breezedu/LossHomozygosity/blob/master/JavaCodes/D0608_verify_homozysity_equations.java` could verify whether these two methods would give the same result or not. Both methods gave almost identical probabilities.

**Example: when there are twelve variants:**

According to method TWO,

The probability of having homozygous would be: 1.6551662611646306E-4

According to method ONE,

The probability of having homozygous would be: 1.655166261164481E-4

**Since all these test codes gave almost exactly the same homozygosity probability, we could say the result we got from equation 5 is reliable for TTN gene.**

**Building gene-exon frame objects from CCDS.r14** The Java code for getting gene-exon frames could be found: `https://github.com/breezedu/LossHomozygosity/blob/master/JavaCodes/D0724_AllGene_CSV_fit_CCDSHash.java` This code will get unique gene ID from CCDS and get exons frames for each gene.

**Downloading LoF allele frequencies for each gene from broad institute website.** The Java code for pulling down LoF variants frequencies from ExAC website could be found: `https://github.com/breezedu/LossHomozygosity/blob/master/JavaCodes/D0701_tryPullData_from_ExAC.java` This code will pull down all LoF variants frequencies, after being filtered by exon frames and 10% threhold, we can get qualified rare variants for each gene.

**Simulating viability advantage** The Java code for simulating viabilities under the null hypothesis that viabilities for individuals with 0 or 2 copies affected by LoF mutations are equal. `https://github.com/breezedu/LossHomozygosity/blob/master/JavaCodes/D20170315_Simulate_TTN_Viability2copiesBeta0.java` This code will get the variants allele frequencies from population data, and fit these variants to CCDS.r14 exon frames, and keep qualified rare LoF variants. Next, it will simulate a sample of genotypes based on these allele frequencies. And filter the genotypes based on the viability equations **equation 13** to decide whether to keep the genotype or not. Keep calling the simulation class till sufficient genotypes are simulated for testing our model.

**Rao's Score Test with and without Fisher Information.** The R code to perform both statistical T test and the Rao's Score Test on TTN gene. `https://github.com/breezedu/LossHomozygosity/blob/master/RSimulation/20170208_`

`chi_test_of_TTN_gene.R` The code first reads in homozygous variants count, and compound heterozygous variants counts, calculate $\pi_{2|g}$ and $S_i$, and perform $\chi^2_1$ test, calculate the score and P-value. And then perform RST with the same data.

The R code to perform the Rao's Score Test, and plot the P-values from all genotypes simulated. `https://github.com/breezedu/LossHomozygosity/blob/master/RSimulation/20170315PlotPValuesbeta0.R`

*Population data from ExAC*

As described in Part II, Supplemental Instructions for Preparing the ExAC Subjects Section of the Research Plan.

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community. The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed here.

*Clinical data from ALS*

The genes considered for amyotrophic lateral sclerosis (ALS), a "late-onset" severe neuronal disorder, were similarly extracted from OMIM: SOD1 (ALS1 - OMIM# 105400), ALS2 (ALS2 - OMIM# 205100). Of the 15 ALS genes, ALS2 and SIGMAR1 lacked OMIM annotation for a dominant model. OMIM susceptibility genes were not considered, and only genes with reported causal genetic variants were eligible.

*Non-coding sequencing data from ncRVIS paper*

Described in Supplemental Instructions for Preparing the Human Subjects Section of the Research Plan, Sections 4.4 and 5.7.

*Sub-region sequencing data from subRVIS paper*

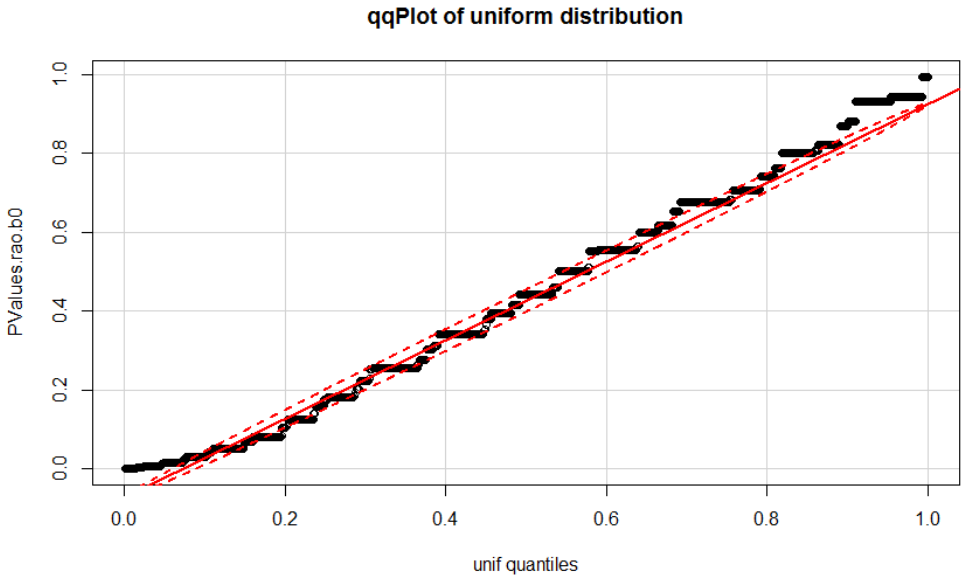


FIGURE A.1: The QQPlot of P-values from simulation under the null hypothesis that viabilities for individuals with 0 or 2 copies of TTN gene affected are equal.

# Bibliography


[1] H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell. *Molecular cell biology*, volume 3. Scientific American Books New York, 1995.

[2] A. J. Griffiths, S. R. Wessler, S. B. Carroll, and J. Doebley. *An Introduction to Genetic Analysis. 11th edition.* Freeman/Worth, W. H. Freeman, 2015.

[3] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17:333–351, 2016.

[4] M. Sharma, R. Kruger, and T. Gasser. From genome-wide association studies to next-generation sequencing: Lessons from the past and planning for the future. *JAMA Neurology*, 71(1):5–6, 2014.

[5] M. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(1):507–515, 2013.

[6] R. Bahareh, T. Mustafa, and M. Nejat. The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, 59:5–15, 2014.

[7] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*, 9(8):e1003709, 2013.

[8] S. Petrovski, A. B. Gussow, Q. Wang, M. Halvorsen, Y. Han, W. H. Weir, A. S. Allen, and D. B. Goldstein. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS genetics*, 11(9):e1005492, 2015.

[9] A. B. Gussow, S. Petrovski, Q. Wang, A. S. Allen, and D. B. Goldstein. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology*, 17(1):9, 2016.

[10] L. Yang, T. Feng, H. Zhenjun, and D. Charles. Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers. *Scientific Reports*, 5(10204):1–15, 2014.

[11] G. B. Christensen and C. G. Lambert. Search for compound heterozygous effects in exome sequence of unrelated subjects. *BMC Proceedings*, 5(9):S95, 2011.

[12] C. Fischer, S. Trajanoski, L. Papić, C. Windpassinger, G. Bernert, M. Freilinger, M. Schabhüttl, M. Arslan-Kirchner, P. Javaher-Haghighi, B. Plecko, J. Senderek, C. Rauscher, W. N. Löscher, T. R. Pieber, A. R. Janecke, and M. Auer-Grumbach. SNP array-based whole genome homozygosity mapping as the first step to a molecular diagnosis in patients with Charcot-Marie-Tooth disease. *Journal of Neurology*, 259(3):515–523, 2012.

[13] L. L. Chen, J. A. Holden, H. Choi, J. Zhu, E. F. Wu, K. A. Jones, J. H. Ward, R. H. Andtbacka, R. L. Randall, C. L. Scaife, K. K. Hunt, V. G. Prieto, A. K. Raymond, W. Zhang, J. C. Trent, R. S. Benjamin, and M. L. Frazier. Evolution from heterozygous to homozygous KIT mutation in gastrointestinal stromal tumor correlates with the mechanism of mitotic nondisjunction and significant tumor progression. *Modern Pathology*, 21:826–836, 2008.

[14] G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, et al. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–898, 2010.

[15] P. Papenhausen, S. Schwartz, H. Risheg, E. Keitges, I. Gadi, R. D. Burnside, V. Jaswaney, J. Pappas, R. Pasion, K. Friedman, and J. Tepperberg. UPD detection using homozygosity profiling with a SNP genotyping microarray. *American Journal of Medical Genetics Part A*, 155(4):757–768, 2011.

[16] K. L. Sund, S. L. Zimmerman, C. Thomas, A. L. Mitchell, C. E. Prada, L. Grote, L. Bal, L. J. Martin, and T. A. Smolarek. Regions of homozygosity identified by SNP microarray analysis aid in the diagnosis of autosomal recessive disease and incidentally detect parental blood relationships. *Genetics in Medicine*, 15(1):70–78, 2013.

[17] P. Dominik, K. Dylan, C. Antonia, A. S. Andrew, J. Samson, T. Shaun, O. Kimberly, G. Andrew, A. N. Scott, and T.-L. Marc. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature*, 533(7601):47–68, 2016.

[18] Y. Jiang, J. M. McCarthy, and A. S. Allen. Testing the Effect of Rare Compound-Heterozygous and Recessive Mutations in Case-Parent Sequencing Studies. *Genetic Epidemiology*, 39(3):166–172, 2015.

[19] K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–950, 2014.

[20] G. Watterson. Heterosis or neutrality? *Genetics*, 85(4):789–814, 1977.

[21] J. Perlow. The distribution of homozygosity for four alleles. *Theoretical population biology*, 30(2):161–165, 1986.

[22] W.-H. Li. Maintenance of genetic variability under mutation and selection pressures in a finite population. *Proceedings of the National Academy of Sciences*, 74(6):2509–2513, 1977.

[23] G. Watterson. The homozygosity test of neutrality. *Genetics*, 88(2):405–417, 1978.

[24] The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, 2004.

[25] M. Spielmann and S. Mundlos. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics*, 25(R2):R157, 2016.

[26] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[27] A. Korte and A. Farlow. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9(1):29, 2013.

[28] D. B. Hancock, M. Eijgelsheim, J. B. Wilk, S. A. Gharib, L. R. Loehr, K. D. Marciante, N. Franceschini, Y. M. Van Durme, T.-h. Chen, R. G. Barr, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics*, 42(1):45–52, 2010.

[29] A. Eyre-Walker. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1752–1756, 2010.

[30] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[31] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[32] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research*, 8(3):175–185, 1998.

[33] B. Ewing and P. Green. Base-calling of automated sequencer traces usingPhred. II. Error Probabilities. *Genome research*, 8(3):186–194, 1998.

[34] M. P. Hare and S. R. Palumbi. The accuracy of heterozygous base calling from diploid sequence and resolution of haplotypes using allele-specific sequencing. *Molecular Ecology*, 8(10):1750–1752, 1999.

[35] H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[36] R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang. SNP detection for massively parallel whole-genome resequencing. *Genome research*, 19(6):1124–1132, 2009.

[37] L. H. Erin, M. N. Benjamin, F. T. Stephen, S. A. Andrew, and B. G. David. The Genetics of Neuropsychiatric Diseases: Looking In and Beyond the Exome. *Annual Review of Neuroscience*, 38:47–68, 2015.

[38] B. C. Smith, J. Grove, M. A. Guzail, C. P. Day, A. K. Daly, A. D. Burt, and M. F. Bassendine. Heterozygosity for hereditary hemochromatosis is associated with more fibrosis in chronic hepatitis C. *Hepatology*, 27(6):1695–1699, 1998.

[39] A. Conesa and A. Mortazavi. The common ground of genomics and systems biology. *BMC Systems Biology*, 8(2):S1, 2014.

[40] E. A. Consortium et al. ExAC database. *Cambridge, MA: Broad Institute*, 2016.

[41] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome research*, 19(7):1316–1323, 2009.