Strategies to prioritize potentially disease-causing mutations in Mendelian disorders

Dissertation zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)
dem Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Na Zhu
aus Shandong, China

Berlin, 26. April, 2016

Die vorgestellten Arbeiten wurden durchgeführt am Institut für Medizinische Genetik und Humangenetik der Charité.

Betreuer:

Prof. Dr. Martin Vingron

Prof. Dr. Peter N. Robinson

Gutachter:

Prof. Dr. Martin Vingron

Prof. Dr. Peter N. Robinson

Prof. Dr. Rosario M. Piro

Dr. Alena Van Bommer

Datum der Disputation: 20. July, 2016

Preface

With high-throughput sequencing technology, the bottleneck that we are searching for disease-causing mutations has shifted from data generation to data interpretation. For many patients with rare Mendelian disorders genome data exist but a conclusive pathogenic mutation has not been identified. Besides the commonly used linkage analysis and intersection filters, the novel solutions are required. Genome-wide association study (GWAS) has successfully identified a large number of disease-associated variants, but it mainly conducts on common variants for common diseases or traits. With the development of sequencing technology and broad availability of high-throughput sequencing data, such association studies can be extended to rare variants. This will allow us to search for the missing heritability from rare variants in complex diseases and additionally to analyze cohorts with rare phenotypes. However, there are specific characteristics of rare variants so that new bioinformatics and statistical frameworks have to be developed. Especially the error rates of rare variants and their geographical distribution is different from common variants. Methods for population stratification between cases and controls thus have to be adapted to avoid spurious associations. Especially for rare disorders, the ethnicities of the affected individuals are often diverse. Such population substructure in the case group can cause substantial inflation of test statistics and can yield artifacts in case-control studies if not properly adjusted for. Existing techniques to correct for confounding effects were especially developed for common variants but do not properly work for rare variants.

I therefore analyzed the matching strategies to select suitable controls for cases that originate from different ethnicities. This work was published in Bioinformatics 2015. The algorithms of similarity metric and the generation of similarity matrix were done by Verena Heinrich. Based on the generated similarity matrix, I developed an approach to build up a control group that is most similar to the individuals in the case group with respect to ethnicity and data quality. I simulated different disease entities with real exome data and showed that similarity-based selection schemes can help to reduce false-positive associations and to optimize the performance of the statistical tests. Finally, I applied this method to analyze a case group of five individuals with Catel-Manzke syndrome, which is an ultra-rare autosomal recessive disorder, and identified *TGDS* as disease associated gene, this work is published in American journal of human genetics 2014. As the prospect of genomic matchmaking database which is a community to share patients, Prof. Peter N. Robinson and Dr. Peter M. Krawitz discussed the required size of the database and the potential impact factors in Human Mutation 2015. As it was built on the rare variants association tests, I joined the simulation in this project.

With my research, I contributed to the following publications:

- Na, Zhu, Verena Heinrich, Thorsten Dickhaus, Jochen Hecht, Peter N Robinson, Stefan Mundlos, Tom Kamphans and Peter M Krawitz. Strategies to improve the performance of rare variant association studies by optimizing the selection of controls. Bioinformatics (Oxford, England), August 2015.

- Peter M Krawitz, Orion Buske, Na Zhu, Michael Brudno, and Peter N Robinson. The Genomic Birthday Paradox: How Much Is Enough? Human mutation, 36 (10) : 989-97, October 2015

- Nadja Ehmke, Almuth Caliebe, Rainer Koenig, Sarina G Kant, Zornitza Stark, Valérie Cormier-Daire, Dagmar Wieczorek, Gabriele Gillessen-Kaesbach, Kirstin Hoff, Amit Kawalia, Holger Thiele, Janine Altmüller, Björn Fischer-Zirnsak, Alexej Knaus, Na Zhu, Verena Heinrich, Celine Huber, Izabela Harabula, Malte Spielmann, Denise Horn, Uwe Kornak, Jochen Hecht, Peter M Krawitz, Peter Nürnberg, Reiner Siebert, Hermann Manzke, Stefan Mundlos. Homozygous and Compound-Heterozygous Mutations in TGDS Cause Catel-Manzke Syndrome. American journal of human genetics, 95(6):76370, December 2014.

- Tom Kamphans, Peggy Sabri, Na Zhu, Verena Heinrich, Stefan Mundlos, Peter N Robinson, Dmitri Parkhomchuk, Peter M Krawitz. Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PloS one, 8(8):e70151, January 2013.

- Peter M Krawitz, Yoshiko Murakami, Angelika Rieß, Marja Hietala, Ulrike Krüger, Na, Zhu, Taroh Kinoshita, Stefan Mundlos, Jochen Hecht, Peter N Robinson, Denise Horn. PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. American journal of human genetics, 92(4):5849, April 2013.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Next generation sequencing

The introduction of dideoxynucleotides for chain termination by Sanger et al. [1] marked a milestone in the history of Deoxyribonucleic acid (DNA) sequencing. Automated Sanger sequencing [2, 3] was developed based on this concept, which supports simultaneous sequencing of 1000 base pairs (bp) per DNA fragment in 96 capillaries. Automated Sanger sequencing was the core technology of the Human Genome Project which took 13 years to map the entire human genome. Next generation sequencing (NGS) sets itself apart from conventional capillary-based sequencing, by the ability to process millions of sequence reads in parallel rather than 96 at a time, in a cost-effective manner (Figure 1.1).

The cost per reaction of DNA sequencing in Sanger sequencing followed Moore's Law [4] until January 2008. After that, the introduction of NGS resulted in a sudden and profound out-pacing of Moore's law. Due to miniaturization and parallelization, NGS platforms can generate millions of short sequence reads in a cost-effective manner.

In 2005, Roche 454 pyrosequencer was introduced. It only cost one-sixth to generate as much data as 50 capillary sequencers [5, 6]. In 2006, Illumina launched Solexa Genome Analyzer which uses a technique called sequencing by synthesis to generate tens of millions of short reads. Applied Biosystems made SOLiD available in 2007, which generate 3G data of 35 bp reads per run with a high accuracy. These three technologies have dominated the current sequencing market. Table 1.1 gives an overview of throughput of Illumina, 454 and Solid technologies.

Via real-time microscopic imaging, all these high-throughput sequencers made revolutions in detecting strand synthesis and in sequencing chemistry. Currently, it can obtain 40 GB data by a single instrument on a single day [7]. It only took a single investigator few days to sequence a human genome.

Platform	Read Length (bp)	Run Time (days)	Size/Run (Gb)	cost/Mb ($)	Error Rate (%)
Roche 454	400	0.42	$0.4 - 0.6$	7	1
Illumina	2×150	$0.3 - 11$	96-600	0.04	0.1
SOLiD	2×50	$4 - 7$	~ 150	0.07	≤ 0.1

Table 1.1: NGS technologies and their throughput until 2014. Data collected from sequencing company websites.

Recently, third-generation sequencing methods have started emerging [8]. Also called single molecule sequencing methods, they do not require a fragment amplification step but work on single DNA molecules. These methods are expected to deliver longer reads and lower costs per run. Currently, they are not widely adopted. However, the definite trend in DNA sequencing is decreasing costs with increasing throughput and data quality.

These new technologies have also increased the spectrum of applications of DNA sequencing to span a wide variety of research areas such as epidemiology, population genetics, phylogenetics or biodiversity and so on [9].

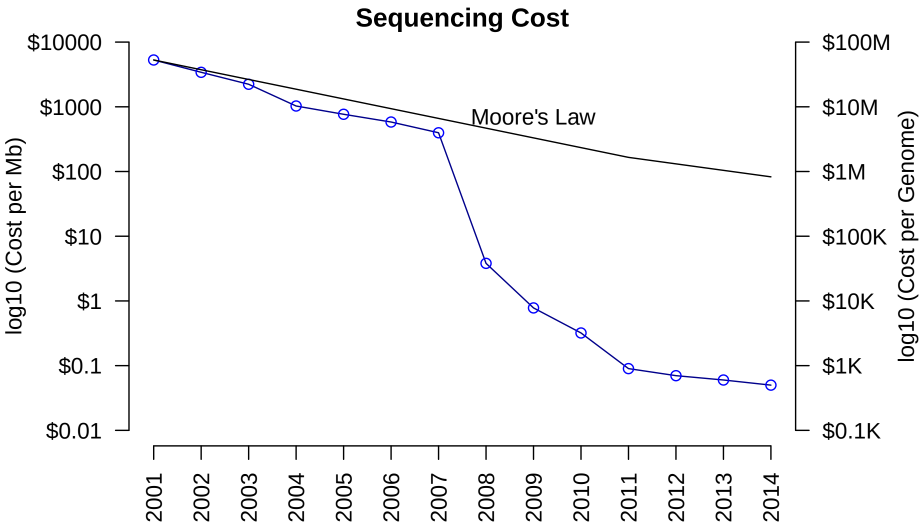


Figure 1.1: Costs associated with DNA sequencing. The data collected from the National Human Genome Research Institute (NRGHI) in 2014. The black line represents the cost of sequencing following the same pattern as Moores law. The blue line shows the declining cost of sequencing per human genome over time.

1.2 Sequencing strategies in human genetics

NGS technologies have revolutionized the study of human and medical genetics. The continually decreasing price of sequencing makes whole genome sequencing and whole exome sequencing studies of complex diseases feasible. However, the costs are still considerable under the scale with the number of individuals, the sequencing depth and the number of bases. Depending on the budget and the goal of the study, different sequencing strategies could be selected: deep Whole genome sequencing (WGS), low depth WGS, Whole exome sequencing (WES), target-region sequencing and custom genotyping arrays (Table 1.2).

Deep WGS is the most comprehensive dataset and has the highest probability of identifying the disease-causing mutation [10]. However, it is hampered by high costs and challenges of data interpretation, especially for non-coding variants.

Low depth WGS provides a cost-effective alternative to deep WGS. Although the genotyping error rates are higher per position and individual, low-depth WGS can detect shared variants effectively [11]. With low depth WGS, one can sequence more individuals compared to deep WGS at the same costs, which can increase the power in association studies [12].

WES aims to sequence the 1% - 2% of the genome that codes for protein [13]. WES usually comprises the consensus coding sequence (CCDS) which consists of about 30 million bases, but the precisely targeted regions may differ depending on the enrichment kit. The average depth of exome-sequencing is typically around 60X-80X. An exome dataset is usually regarded high quality if a fraction of more than 80 % of the target region is covered by more than 20X reads [14]. The proportion of reads that map to the target region reflects the efficiency of the enrichment. This enrichment factor is usually higher for larger target regions and exomes. The primary limitation of exome sequencing is that it only captures genetic variation in the exome and ignores the non-coding regions which might limit the diagnostic yield. However, before deep WGS becomes less costly, WES is a competitive approach that will probably become a standard routine for some clinical indications.

Another cost effective strategy is the enrichment of customized target regions. For molecular pathway diseases, a limited number of genes are involved. For GPI-anchor deficiencies, we designed, for instance, such a customized gene-panel [15]. On one hand, this allows a further reduction in sequencing costs. On the other hand, certain non-coding regions that contained pathogenic mutations may additionally be incorporated in the set of customized oligo baits.

The last approach is customized genotyping arrays. It may include common variants selected from Genome-wide association study (GWAS) and variants of low frequency that might be potentially relevant to a specific study. The exome chips developed by Illumina and Affymetrix provide an inexpensive array-based approach to exome sequencing [16]. The arrays collected data mainly from

12,000 sequenced exomes (mostly of European ancestry). It includes about 250,000 missense variants, 12,000 splicing variants, 7,000 stop-altering variants, and ancestry-informative markers. For the European population, the majority of variants with an allele frequency above 0.001 will be included in this array. However, family specific variants or de novo mutations are obviously not detectable with this approach.

Target specific resequencing and custom genotyping arrays make certain assumptions about the relevant mutations. Whereas, WGS is a hypothesis-free approach for disease gene identification.

Table 1.2: Array and sequencing platforms for variants analysis

	Advantage	Drawback
Deep WGS	identify genomic variants; high confidence	currently expensive; huge data amount
Low depth WGS	cost-effective	limited accuracy
WES	high detection rate in protein coding exons; cost-effective	limited to protein-coding exons
Target region sequencing	inexpensive	lower accuracy for imputed rare variants; limited region
Custom array	inexpensive	limited coverage for rare variants; currently specific for Europeans

1.3 Disease gene identification

NGS technology revolutionized medical genetics by making DNA sequence broadly available. As introduced above, the sequencing strategies are dependent on the study goal and the budget. In the following we will discuss the usual considerations for selecting individuals if the budget is limited. Most of these strategies originate from the analysis of Mendelian disorders.

1.3.1 Selection strategies of sequencing individuals

In a family with a Mendelian disorder, it is assumed that all affected family members share the same disease-causing mutation. The more distant the relationship, the smaller is the set of shared rare variants. When only a fixed number of family members can be sequenced, the best combination of individuals is the one with the largest number of meioses, which can minimize the number of variants[17].

When quantitative traits are analyzed, intuitively the samples with the extremes phenotype should be sequenced. By this selection of patients, it may increase the probability that differences in risk- or phenotype, and it may maximize the modifying alleles. The effect sizes estimated in phenotypic extremes are also systematically larger than those estimated in random samples [18, 19, 20].

1.3.2 Strategies for disease gene identification

All sequencing approaches mentioned previously would yield thousands of variants per individual. In this section, common strategies to filter for potentially pathogenic mutations or disease-linked loci will be discussed. Figure 1.2 shows three common scenarios encountered in rare Mendelian diseases. The ideal situation is a large pedigree with multiple affected family members in several generations. As shown in family A, the disorder is inherited in an autosomal dominant mode in a large family. All family members are informative for a linkage analysis and could be used to limit the genomic search space. In family B the parents are healthy while about a quarter of the children are affected, suggesting a recessive mode of inheritance. Depending on the degree of consanguinity a search for homozygous or compound heterozygous candidate mutations is the first line strategy. The scenario as shown in family C depicts some "sporadic" cases and filtering for *de novo* mutations is an effective analysis strategy for such phenotypes. Whenever the disease-causing mutations cannot be identified with the classical analysis strategies, phenotypically similar cases can be grouped and analyzed for gene associations.

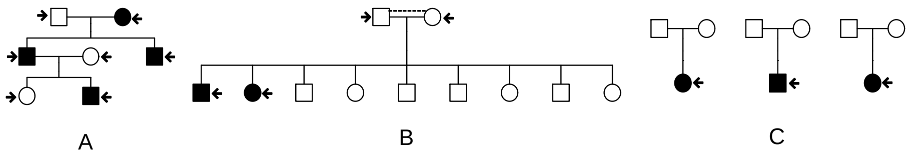


Figure 1.2: Common scenarios when analyzing rare disorders. Rectangles in pedigrees represent male and circles represent female family members; filled symbols represent affected individuals. A) Large pedigree with multiple affected family members, autosomal dominant mode of inheritance B) A recessive trait in a potentially non-consanguineous pedigree. C) Multiple "sporadic" cases in nuclear families.

1.3.2.1 Linkage analysis

Classical linkage analysis can be used in a pedigree with multiple affected family members to narrow the genomic search space. In a pedigree with a dominant or recessive disorder, LOD score (logarithm of odds) is calculated for single genomic position. We can use this score to determine if a loci is linked to a disorder. In a consanguineous family with a recessive disorder, the disease-causing mutation is rooted most likely in a common ancestor. The founder with the pathogenic mutation transmitted the pathogenic allele to both parents. The parents share the same haplotype with the pathogenic mutation but are only heterozygous for this variant. Rare variants can be prioritized by identifying large homozygous intervals in the genome of the affected individuals but not the healthy ones via homozygosity mapping [21]. An alternative strategy in large pedigrees is to sequence several distantly related affected family members and to filter for shared rare variants (see Section 1.3.1). Genotypes of sequenced unaffected individuals can additionally help to exclude benign family specific variants [17].

1.3.2.2 Filtering for compound heterozygotes

In non-consanguineous families with a recessive disorder, a possible combination of pathogenic mutations is compound heterozygotes. That means there are two different pathogenic alleles in the same gene. The parents transmit two same heterozygous mutations to all affected individuals. The disease locus can be narrowed down by identity by descent mapping that identifies shared haplotypes [22]. For exome data of multiple sequenced family members, direct filtering for rare compound heterozygous variants is very effective. We have developed such a filtering tool that was used successfully to identify several pathogenic mutations [23, 24]

1.3.2.3 *De novo* mutations

Many disorders such as intellectual disability (ID), often present as singular cases in a family. In a landmark paper for non-syndromic ID, it was shown that the majority of cases are due to *de novo* mutations [25]. In an exome there are about 0-3 new single nucleotide variants per individuals and nonsynonymous events are highly likely to be pathogenic. On a genome-wide level *de novo* mutations, notably structural variants, are much harder to detect and interpret and are a current challenge to bioinformatics.

1.4 Genome-wide association studies

Whenever the disease-causing mutation cannot be conclusively identified in a single pedigree, unrelated affected individuals can be combined to a case group and analyzed for gene associations. Although this approach has so far been mostly used for complex disorders, it also works for monogenic diseases. In the following, it shows some of the commonalities and key differences between association studies for Mendelian and common disorders. Association studies for Mendelian disorders are always based on rare variants, Rare variant association study (RVAS), whereas association studies for complex diseases usually deal with polymorphisms, common variant association study (CVAS). The power of an association study depends on many factors, such as case and control group sizes, the intended level of statistical significance, allele frequencies and effect size of the variants [26, 27]. Despite the many differences there are also challenges that are common to both approaches such as genetic heterogeneity of the disorder and spurious associations due to population substructure. In addition, not every sample is necessarily informative, such as the sample with incompleteness of exome sequencing data.

1.4.1 Common versus rare variant association studies

The first variant association studies were motivated by the common disease common variant (CDCV) hypothesis, that assumes that a small number of common variants have moderately small effects on the complex disease [28]. In CVAS, a variant is common if its minor allele frequency lies above 1% in the general populations. The odds ratios for the functional polymorphisms are assumed be modest (1.1-1.5). With these typical assumptions, a study with adequate power would require at least a thousand subjects [29]. With the advancements in single nucleotide polymorphism (SNP) genotyping technologies, CVAS have been conducted and revealed many new loci [30, 31]. However, the identified common variants can only explain about 30% of the heritability for numbers of diseases and the CDCV has thus to be challenged [32, 33]. Different strategies have been suggested to search for the "missing heritability". One can either extend the search for polymorphisms with an even lower effect size, requiring ever larger case groups, or one can include also rare variants, which makes different statistical tests necessary [34, 35].

1.4.2 RVAS on complex and rare diseases

The theory of evolution predicts that purifying selection may lead deleterious alleles rare. This should be particularly the case for loss of function variants in vital genes. Thus, many research groups turned to search for rare variants, commonly Minor allele frequency (MAF) below 1% [36]. The majority of identified rare variant associations to date have odds ratio greater than two, and the mean odds ratio is 3.74 [37]. Successful RVAS identified new gene associations in disorders such as type 1 diabetes, age-related macular degeneration and Alzheimer's disease [38, 39, 40, 41, 42, 43, 44]. However, the rare variant common diseases hypothesis doesn't seem to apply to all complex diseases [45]. For instance in type 2 diabetes [46], schizophrenia [47], epilepsy [48], autism [49] and autoimmune diseases [50], no significant associations with rare variants were found so far. Thus, the importance of rare variants seems to depend on genetic architecture of the disease.

In contrast to most common diseases with complex genetic interactions, many rare diseases are Mendelian disorders. In the USA, a disease is called rare if its prevalence is lower than $1/1,500$ according to the Rare Diseases Act of 2002, whereas the European Commission on Public Health choose a cutoff of $1/2,000$. The prevalence of rare diseases can vary between different populations, the geographic area and age. For instance, a collection of 40 rare diseases that are due to a founder effect are significantly more common in Finns than other populations [51]. Due to the low prevalence of these disorders, research funding is notoriously scarce, and the pathophysiology of many of them is not yet clear. However, the identification of disease genes in rare Mendelian disorders often deepens our understanding of related complex diseases and is thus a promising field of research [52]. Although rare disorders are expected to be monogenic, rare causal variants are difficult to identify due to the inherently small case group sizes, and such diseases can be heterogeneous though following Mendelian modes of inheritance. All above reason lead to the low performance of RVAS

The required number of cases are dependent on the relative risk, the disruptive allele frequency and the selection coefficient. Given specific statistical power (see Section 3.5) and false positive rate, the higher relative risk of pathogenic mutations can reduce the required effect size. The stronger selection on mutations can lead to lower disruptive allele frequency, and further increase the required sample size to achieve a specific power. The higher disruptive allele frequency requires fewer samples. Note that rare pathogenic variants associated with the rare disorders usually have small disruptive allele frequency and stronger selection coefficient.

Compared to CVAS, RVAS differs in two aspects. Firstly, as rare variants are so infrequent that it is impossible to conduct association tests for single marker. It is required to aggregate rare variants in a genomic region and to compare the accumulated frequency between groups. The aggregating strategy further makes the second difference to CVAS that rare variants association test is sensitive to variant filters and the aggregating bins. A good filter is the one that could gather more damaging alleles while ignoring more benign alleles in the particular genomic region, such as gene. Besides allele frequency, RVAS requires additional filters to enrich the deleterious mutations. Typically function in protein-coding region is further used to categorize the variants.

The pathogenic mutations of rare disorders are expected to have extremely high relative risk, as most of these mutations never occurred in controls or healthy populations. For a specific disorder, more strict filters can be applied, for instance, one can only test the nonsense mutations or highly conservative mutations. To amplify the signal of associations, one could also collapse the mutations on gene level or pathway level.

For the unrelated cohort with rare disorders, besides gene identification strategy (Section 1.3.2.3), rare variants association tests could be the alternative and more intuitive solution. Moreover, RVAS is advantageous to downgrade the highly variable genes, as the number of mutations in controls can balance the one in cases.

1.4.3 Population substructure

In genetic association study, a region (like a snp or a gene) with significant test statistic may indicate the enrichment of a risk factor. These significant regions could be true associations or spurious associations. The difference from data quality, population structure or genetic relatedness between case and control groups can cause spurious associations and inflate test statistics [53]. The same protocol for NGS technologies and bioinformatics procedure may resolve the difference in data quality between samples. However, the difference of population substructure or genetic relatedness is tricky, which cause the difference in allele frequency between groups due to systematic ancestry differences, as demonstrated in Figure 1.3. It could even exist among populations that were assumed to be relatively homogeneous such as Europeans [54, 55, 56]. Thus, accounting for population stratification in association study is a crucial issue, and is more challenging if family structure or cryptic relatedness present as well [57].

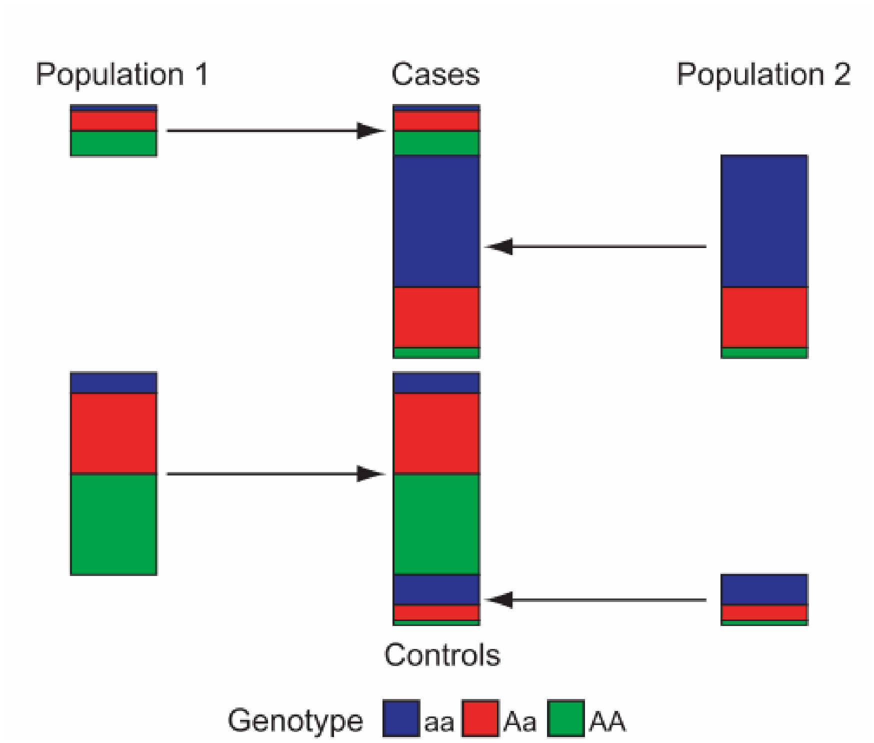


Figure 1.3: The demonstration of population structure at a SNP locus. Population 2 has a lower frequency of allele A than that of population 1. Case group and control group have different proportions of these two populations. The significant signal of association comes from difference of allele and genotype frequencies between cases and controls. The figure is adopted from Marchini et al.[58]

.

1.4.3.1 Population substructure in CVAS

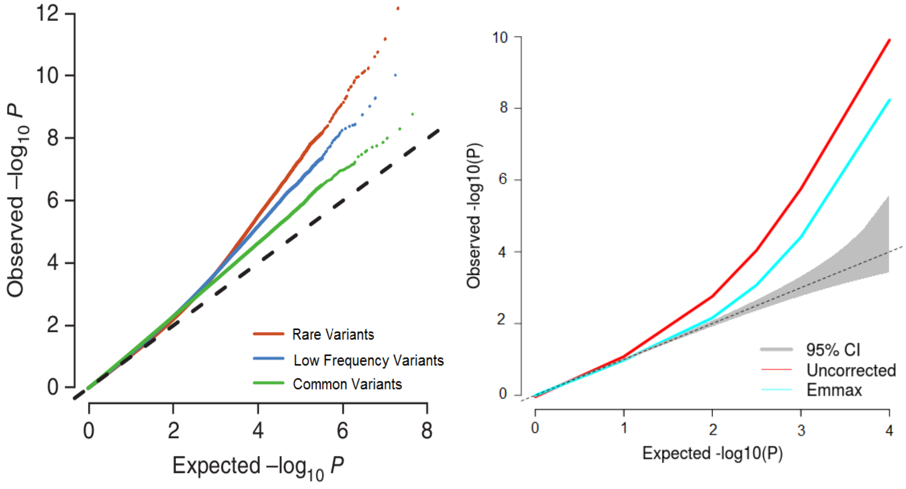
The reason for population stratification could be due to ancient population divergence or recent genetic drift [57]. Many methods have been developed to account for the population stratification due to common variants. There are three common strategies. The first one is genomic control which measures the extent of inflation from confounders. Genomic control could perform well if the stratification due to genetic drift while it is too conservative if the stratification from population divergence [59, 60]. The second method is to infer genetic ancestry, such as principal component analysis (PCA) [53] or structured association [61]. PCA assumes a small number of ancestral populations and admixture, so it can only partially capture the multiple levels of population structure and genetic relatedness. However, this method cannot account for cryptic relatedness and family structure while some studies showed that cryptic relatedness was common in many datasets [59, 62]. The third method is based on the linear mixed model (LMM), which can model population substructure, cryptic relatedness and family structure. The basic method is to model phenotypes as a mixture of fixed effects due to candiate SNP, and random effects due to confounders. The effect of confouders is assumed to be randomly distributed and can be inffered by the covariance of kinship matrix among samples [63]. Mixed model has been applied in methods Emmax [64], TASSEL [65], FaST-LMM [66] and GEMMA [67].

1.4.3.2 Population substructure in RVAS

The population stratification due to rare variants is more pronounced than with common variants (Figure 1.4a). The reason is following: The different frequency of rare individual alleles between populations may result from geographic localization and small number of shared rare variants [68]. There is a very low rate of sharing of rare alleles even between very closely related human populations [69]. Babron et al. investigated the stratification patterns in UK population in three different allele frequency categories. They found that the top principal component obtained from rare variants ($< 1\%$) did not correlate with any principal components from low frequency ($1\% < AF < 5\%$) or common variants ($> 5\%$) categories [70].

Furthermore, the total quantity of rare alleles is also different among populations because of differences in effective population sizes,demographic events, bottlenecks or selective pressures. This may also deteriorate the spurious associations in RVAS. The reason is that, in order to increase the statistical power, RVAS commonly use aggregation tests rather than single variant tests. In single marker tests, stratification is only dependent of different allele frequencies at individual sites. Whereas aggregation tests, which aggregate the number of alleles across multiple positions, have to tackle population differences in both individual allele frequencies and the total number of rare variants [68].

These non-genetic risks which may contribute to the population stratification may show a very specific distribution, such as the localized environment exposure. Typically, the more localized a risk factor is ,the less we are likely to know about it and the greater effect this lack of knowledge will have on rare variants, which results in the difficulty for accounting for the known non-genetic risk factors.



(a)

(b)

Figure 1.4: Quantile-quantile (QQ) plots of association tests with sharply and small spatial distributed risk. a) The inflation due to rare variants is higher than due to common variants. b) None of the correction methods developed for CVAS can account for the population stratification due to rare variants. The figure is adopted from the study of Mathieson and McVean [71]

.

The study of Mathieson and McVean showed that none of the existing methods for accounting for the population stratification cannot work properly in RVAS. Genomic control cannot work because most variants have no correlation with the nongenetic risk. PCA and mixed models assume a smooth distribution of minor allele frequency over ancestry space and all nongenetic risks are linear related with top components (Figure 1.4b). However, the small, sharp region of risks would require a highly nonlinear function to be expressed, but it cannot be achieved only by including the top components [71]. A new method based on linear mixed model, FaST-LMM-Select, selected a few of phenotype-selected variants to build the kinship matrix, instead of all SNPs in traditional LMM. Compared to traditional LMM, the performance of FaST-LMM-Select is that it can yield non-inflated test statistics. However, if the causal variants are spatially structured, the false positive rate could be under control but the statistical power decreases as well, as the causal variants are treated as confounders [14, 72, 73].

1.5 Matching strategies for correcting population stratification

The confounding due to population stratification is caused by the mismatched genetic ancestry between case and control groups. Thus, fine matching of cases and controls based on genetic ancestry may help accounting for confounding. Matching strategies try to set up case and control groups which share similar genetic ancestry. The matching strategy can be implemented in different approaches, such as GEM [74], SpectralGEM [75], stratification score matching [76] and GSM [77]. These approaches can be divided into two categories: An estimation of genetic similarity among individuals that is based on 1) the ancestry components from principal components or spectral-graphs (GEM, SpectralGEM) and 2) the average proportion of alleles shared identical-by-state over large number of SNPs (GSM).

Many GWAS of complex diseases, including studies of ulcerative colitis [78], asthma [79], and presenile dementia [80], have employed fine matching to deal with confounding due to population stratification. For RVAS, the performance of the matching strategies still needs to be investigated.

1.6 Aim of the study and structure of the thesis

As shown above, RVAS for rare disorders is still needed further study and the existing methods that account for population substructure cannnot correct the inflation sufficiently in RVAS [27]. Therefore, I studied the performance of RVAS in rare disorders and I also investigated the performance of 'matching strategy' in RVAS. In the second chapter, I outlined the data used in this work including the in-house data, data from 1000 genome project and the simulated disorders. I made an investigation of the features of variants and genes in the clinical data and non-clinical data, which served for the following chapter. The methods used in this work were described in the third chapter. I described the similarity metrics which were used for the 'matching strategy', the methods for test statistics which used for the association tests, the methods for accounting for the genetic relatedness and the workflow of simulations. I showed all results in the fourth chapter. It included the performance of RVAS with the 'matching strategy', the factors which affected the results and the application of RVAS in real cohorts. Finally, I summarized the implications of the project and gave an outlook for future research in the last chapter.

Chapter 2

Materials

2.1 Data-sets

2.1.1 In-House Exomes

In recent years, many patients with unknown genetic disorders were subjected to WES at Charité, University Hospital Berlin. These inhouse cohorts consisted of samples from multiple populations: European, Arabian, African and Asian. The majority had the European background. It was also heterogeneous cohorts, parts of exomes from patients with different diseases, such as Mabry syndrome, Catel-Manzke syndrome and Marfan syndrome [81, 15, 24, 82], parts from healthy parents and gathered controls. All exomes were enriched with Agilent Human All Exon SureSelect baits and sequenced on Illumina Genome Analyzer IIx and Hiseq. All sequences were mapped to human reference sequence GRCh37/hg19, and variants were called with GATK [83]. As it took many years to collect these cohorts, the data quality between samples varied with the developed sequencing technologies. I removed the data of the children in the trios to maximize the number of unrelated samples. I referred to this cohort as Cohorts sequenced in Charité - Universitätsmedizin Berlin (BER) in the following.

2.1.2 Data from 1000 Genomes Project

1000 Genome Project (1KGP) is the first international project to sequence the genomes of individuals from all over the world. One aim of the project was to analyze the variability of allele frequencies between populations from different continents. The allele frequencies for 26 populations from 2504 individuals in total were made publicly available.

The 1000 Genome Project proceeded in 3 phases: phase pilot, phase 1 and phase 3. Each phase analyzed through a combination of low-coverage WGS data and targeted deep WES data [84, 11]. This sequencing design is cost-effective in discovering genotypic variants. Phase pilot and phase 1 had a mixture of both read lengths 36bp to 160bp and used three sequencing platforms including Illumina [85, 86], ABI SOLiD and Roche 454 while phase 3 only used the Illumina sequencing platform and reads lengths of 70 bp+ [87]. The uniform sequence technology in phase 3 largely erased the difference in variants quality [88, 89, 90]. The employed bioinformatic tools were also improved in phase 3. Many variant callers were used in phase 3, such as GATK [83], Samtools [91], Delly [92] and Pindel [93]. It considered low coverage genome sequence and exome sequence together. 24 genotyping tools were used for calling short variants, structural variants and short tandem repeats. Phase 3 integrated multi allelic variants and complex events that were impossible in phase 1 (Figure 2.1). The sequencing data quality was high for all populations, but it varied in populations due to different sequencing centers (Figure 2.2).

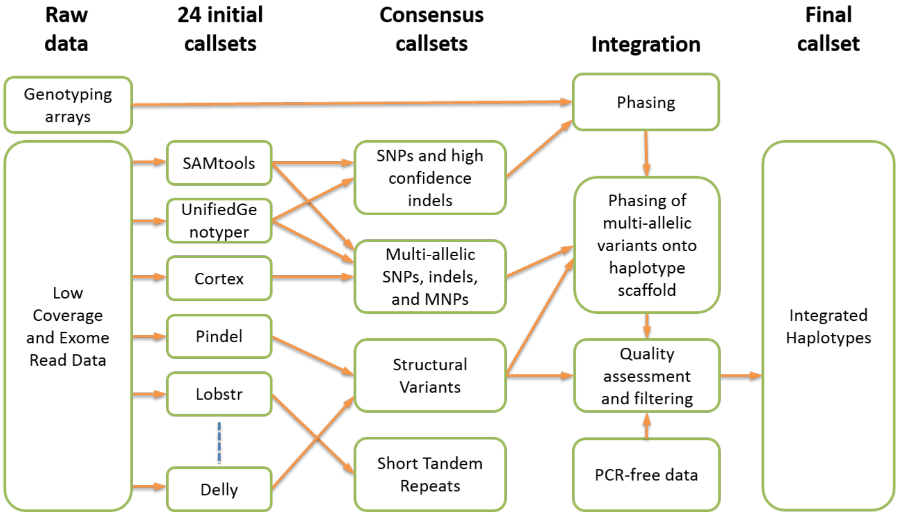


Figure 2.1: A combination of low-coverage WGS data and targeted deep WES data was performed in Phase 3 of 1KGP. Phased variants were the consensus results from 24 variant callers including 10 for calling short variants, two for calling short tandem repeats and 12 for calling structural variants. This figure was adopted from the 1000 Genomes Project Consortium.

As improvements in sequencing technology emerged, sequencing time and cost reduced significantly. Along the way, more and more populations were sequenced across these phases. Finally, phase 3 sequenced 26 populations across five continents, adding up to 2504 individuals in total. The populations are chosen based on scientific, ethical and practical considerations, with the expectation to obtain broadly representative genetic variation data for the vast majority of individuals within each continent [11]. All donors were over 18 years old and healthy at the time of collection.

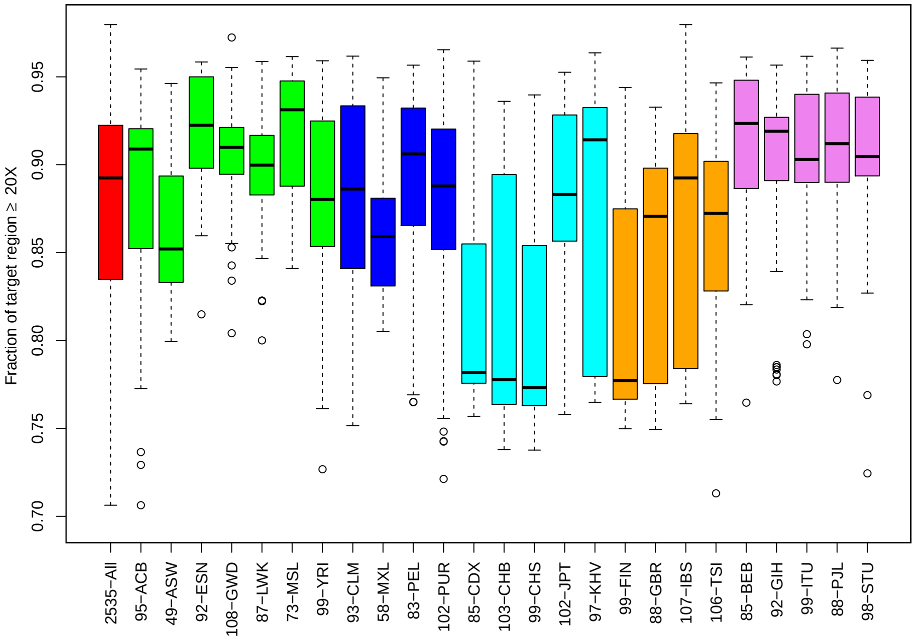


Figure 2.2: More than 70% of the target region are covered by at least 20 reads for all samples. Populations of the same continent are color-coded, and the number in front of the population ID indicates the size the cohort. There is substantial variability in the median coverage for different subpopulations, indicating different mean data qualities.

2.2 Simulated disorders

We selected eight known rare diseases with a prevalence lower than 1/1000 (Table 2.1). From the inheritance pattern point of view, some disorders are transmitted in the autosomal recessive pattern such as Hyperphosphatasia with mental retardation syndrome (HPMRS); some disorders are in the autosomal dominant pattern such as Noonan syndrome; some have several inheritance patterns, likewise Deafness, which could be autosomal recessive or X-linked or autosomal dominant pattern. Respecting the genetic heterogeneity, some diseases are heterogeneous, which means that several genes could contribute to the disorders, such as HPMRS. The others are homogeneous in that all pathogenic mutations are in the same gene. For example, gene *HEXA* is the only gene associated with Tay-Sachs syndrome. In the following, the disorders and their genetic mechanism are introduced.

Noonan Syndrome The typical features of Noonan Syndrome are typical facial dysmorphology, short stature and congenital heart defects. Its incidence lies between 1:1000 and 1:2500 in live births [94, 95]. It is an autosomal dominant disorder. Approximately 50% of cases are affected because of missense mutations in gene *PTPN11* on chromosome 12 which results in a gain of function of the non-receptor protein tyrosine phosphatase SHP-2 protein [96, 97]. Another 20% of patients possess missense mutations or gain-of-function mutations in the genes *KRAS* [98], *SOS1* [99], *RAF1* [100], *NRAS* [101] and *BRAF* [102, 100]. The genetic etiology for the remaining patients with Noonan Syndromes remains unknown.

Nonsyndromic deafness Nonsyndromic deafness is hearing loss that is not linked to abnormalities of the body. It has different patterns of inheritance. $75\% - 80\%$ patients inherit the disorder in an autosomal recessive pattern which is designated as DFNB. Another $20\% - 25\%$ of cases are in autosomal dominant pattern which is designated DFNA [103]. $1\% - 2\%$ of the remaining cases show an X-linked pattern of inheritance which is named as DFN [104]. 1% inherit mitochondrial nonsyndromic deafness where a mother passed the altered mitochondrial DNA to all of the children [105]. Different inheritance can share the same pathogenic gene, for instance, mutations on *TECTA* can cause deafness in the dominant and recessive model.

To simplify the simulation of deafness in the current work, i only tested DFNB Deafness. The approximate prevalence of DFNB in the general population is $\frac{1}{2000} \times 0.7 \times 0.8 = 14 : 50,000$, with a $1/2,000$ incidences of congenital hereditary hearing impairment in neonates, of which 70% have nonsyndromic hearing loss [106] and $75\% \sim 80\%$ of cases with nonsyndromic hearing loss are autosomal recessive [107, 108].

50% of patients with autosomal recessive nonsyndromic hearing loss have pathogenic mutations in *GJB2* [109, 110, 111]. Mutations in numerous genes make contributions to the other 50% patients, many of which have been found only in one or two families. For the sake of simplicity, we only selected nine reported genes and assumed that mutations in these genes contribute to the pathogenesis of 20% of patients [112, 113, 114].

Mabry syndrome Mabry syndrome, also known as Hyperphosphatasia with mental retardation syndrome (HPMRS), is a rare recessive genetic disorder that causes mental retardation, seizures and characteristic raised blood levels of the enzyme alkaline phosphatase. The incidence of Mabry syndrome is still unknown but likely to be rare, as less than 30 cases were reported by the end of 2014 [115, 116, 117, 118, 119, 15]. The inheritance model of Mabry syndrome is autosomal recessive. Mutations in *PIGV*, *PIGO*, *PGAP2* or *PGAP3* genes are the underlying cause. All of these genes are linked to the synthesis of the glycosylphosphophatidylinositol (GPI) anchor. Approximately 30% of patients with Marby syndrom are affected because of mutations in gene *PIGV* [82, 118]. Mutations in the *PIGO*, *PGAP2* and *PGAP3* genes contribute to a small proportion of cases with HPMRS [15, 24, 120].

Tay-Sachs disease Tay-Sachs disease is a neurodegenerative disorder caused by a deficiency of an enzyme called hexosaminidase A, *HEXA*. Lack of this enzyme causes rapid and progressive deterioration of the brain and nervous system. HEXA gene produces a protein which forms the alpha subunit of hexosaminidase A. More than 120 mutations in gene *HEXA* are linked to Tay-Sachs disease. The activity of the enzyme beta-hexosaminidase A is reduced or eliminated due to these mutations [121]. Tay-Sachs syndrome is inherited autosomal recessively. Its incidence is 1 in 3600 in the Ashkenazi Jewish Population and 1 in 360,000 in other populations [122, 123].

Cystic fibrosis Cystic fibrosis is a recessive monogenic disorder caused by mutations in cystic fibrosis transmembrane conductance regulator ($CFTR$) gene. It causes various dysfunction in different organs, including lung disease, meconium ileus, diabetes, and liver disease [124]. The incidence of cystic fibrosis is estimated at around $1/2500$ in Caucasians, $1/3500$ in Europe, $1/350,000$ in Asia and $1/15,000$ in Africa [125, 126]. It distributes across a broad age range. With the development of health policies such as newborn screening, the incidence has been lowered nowadays [127].

Neurofibromatosis type 1 Neurofibromatosis type 1 is multisystem disease mainly related with skin and nervous system. Its typical feature is changes in pigmentation and the growth of tumors along nerves in skin, brain, and other parts of the body. It is genetically a homogeneous disorder caused by mutations in the *NF1* gene. The *NF1* gene is related to protein neurofibromin which acts as a tumor suppressor. Mutations in the *NF1* gene result in its loss of function. Neurofibromatosis type 1 is an autosomal dominant disorder. Its incidence is about 1 in 3500 people worldwide [128, 129].

Catel-Manzke syndrome Catel-Manzke syndrome is depicted by a unique form of bilateral hyperphalangy causing a clinodactyly of the index finger. It is rare, as currently 28 cases with Catel-Manzke syndrome have been reported [81, 130]. Mutations in gene $TGDS$ cause this syndrome, which has a general effect on connective tissue. The $TGDS$ gene is related to either proteoglycan synthesis or sulfation. Catel-Manzke syndrome is inherited in a recessive pattern [81].

Kabuki makeup syndrome The phenotypes of Kabuki makeup syndrome are typical facial features, minor skeletal anomalies, the persistence of fetal fingertip pads, mild to moderate intellectual disability, and postnatal growth deficiency [131]. The incidence is about 1 out pf 32,000 newborns in Japan [132] and 1 in 86,000 in Australia and New Zealand [133]. Its incidence in other ethnic groups is estimated to be similar to that in the Japanese population. Mutations in gene *KMT2D* (or *MLL2*) [134] or gene *KDM6A* [135, 136] lead to this syndrome. 55 ∼ 80% of the Kabuki makeup syndrome cases result from mutations in gene *KMT2D*. 6% of cases possess mutations in the *KDM6A* gene. The cause of the disorder in the remaining cases is still unknown [137]. Mutations in *KMT2D* and *KDM6A* genes lead to the related functional enzyme absent and further result in the development abnormalities. Mutations in gene *KMT2D* are transmitted in an autosomal dominant pattern while mutations in gene *KDM6A* are transmitted in an X-linked dominant pattern [138]. As I ignored sex chromosomes in this project, I only tested mutations in *KMT2D* and set its prevalence as 70%.

Disease	Proportion of cases attributed to mutations in specific genes	Known pathogenic mutations
Noonan-Syndrome autosomal dominant	PTPN11 (50%)	74
	SOS1 (10%)	44
	RAF1 (5%)	18
	KRAS (2%)	14
	BRAF (2%)	4
	NRAS (1%)	3
Nonsyndromic hearing impairment autosomal recessive	GJB2 (50%)	56
	ATP2B2 (2%)	2
	CDH23 (2%)	5
	CLDN14 (2%)	2
	DFNB31 (2%)	1
	GJA1 (2%)	0
	MYO6 (2%)	2
	OTOA (2%)	1

	OTOF (2%)	8
	TECTA (2%)	1
	PIGV (30%)	9
HPMRS	PIGO (10%)	2
	PGAP2 (10%)	3
autosomal recessive	PGAP3(10%)	0
Tay-Sachs disease autosomal recessive	HEXA (100%)	109
Cystic Fibrosis autosomal recessive	CFTR(100%)	825
Neurofibromatosis type 1 autosomal dominant	NF1 (100%)	565
Catel-Manzke autosomal recessive	TGDS (100%)	5
Kabuki makeup syndrome autosomal dominant	KMT2D (70%)	10

Table 2.1: Eight rare monogenic disorders were simulated for rare variant association tests. This consists of four genetic homogeneous disorders and four genetic heterogeneous disorders. Three autosomal dominant and five autosomal recessive disorders were included from the perspective of inheritance mode. The prevalence of mutations in each gene in heterogeneous disorders varies and are obtained from literature.

2.3 Quality control

To obtain a set of genotype calls with high quality, I restricted the variants of all datasets in the consensus coding DNA sequence (CCDS) region of exome comprising 28Mb. As the INDELs and multiple nucleotide positions had lower accuracy [139], I removed insertions, deletions and the positions with multiple alternated alleles.

BER data included healthy samples and patients samples from many studies. Due to the potential intrinsic divergency to the simulated disorders [140], I removed the known pathogenic mutations from the variants list. To reduce the false positive calls in BER data, I also removed the site if less than 90% exomes detected it and eliminated the positions which frequently occurred (at least 10%) in BER, but never find in dbSNP database.

In this work, we made simulations for autosomal disorders, we thus ignored the variants in chromosome X and Y, which largely removed the bias from sex in the association tests.

2.4 Variant filters

RVAS requires aggregation of the variants in a genomic region, as rare variants are too infrequent to test on individual variant [27]. Aggregation is the critical step for RVAS; proper aggregation can increase the power of detecting associations in RVAS. In the attempt to enhance the proportion of the deleterious alleles to the benign alleles as much as possible, a proper filter is required [141]. In this work, I filtered variants from three classifications: the predicted effect of protein function, the sequence conservation and allele frequency. In order to choose a suitable cut-off for each filter, I firstly investigated the features between non-clinical variants and clinical variants based on public data. I took variants in Clinvar which had clinicalinvestigated significance ”pathogenic variants” or ”likely pathogenic” ([142]) as clinical variants. Non-clinical data were the variants in dbSNP137 [143] that had never been cited in PubMed and not known in the clinic context(no ”PM” in field ’INFO’).

2.4.1 Predicted effect on protein function

In protein coding regions, mutations can be categorized into three general categories: synonymous mutations, nonsynonymous mutations and stop-codon mutations. In a synonymous mutation or silent mutation, a change in one base pair has no effect on the protein produced by the gene. Certain codon may be more efficient than others in some cases [144, 145], but silent mutations are often assumed to be evolution neutral. Nonsynonymous mutations include missense mutations and nonsense mutations. A missense mutation changes the code for a single amino acid and further results in a different protein. For example, Cystic Fibrosis is caused by some missense mutations [146, 147]. Evolutionary studies and an analysis of mutations responsible for Mendelian diseases suggest that 20% of missense mutations are strongly deleterious; about 50% are weakly deleterious, and the remainders are essentially neutral [148, 149]. Nonsense mutations change a single base pair and create a stop codon, which makes the resulting protein nonfunctional. These mutations are so severely disruptive that they may cause a disease [150]. Stop-codon mutation is the opposite of nonsense mutation, in that it changes the stop codon into a codon for an amino acid and then leads to the protein being too large. Such mutations destroy the protein and can cause diseases too. A small part of Cystic Fibrosis patients are caused by stop-codon mutations [151]. Exome sequencing can also detect a small fraction of non-coding sites with high quality [152]. These variants include intronic mutations, intergenic mutations, splicing mutations and so on. Except splicing mutations, other non-codign mutations are little known.

I compared the distribution of mutations across different categories from two data sets: non-clinical SNVs from dbSNP [143] and clinical SNVs from Clinvar (Figure 2.3). It was found that about 70% of clinical mutations are nonsynonymous. The proportion is similar to that in the OMIM database [153, 154]. Only 0.6% of clinical mutations are synonymous. Some of these synonymous mutations are found to be deleterious [144, 145], but the small deleterious proportion indicates a large proportion of neural variants, which can dilute the effect of the accumulation of disease-causing mutations. Thus, synonymous mutations were ignored in this project. In the consideration of the severity of disrupting protein structure, I only kept nonsynonymous mutations.

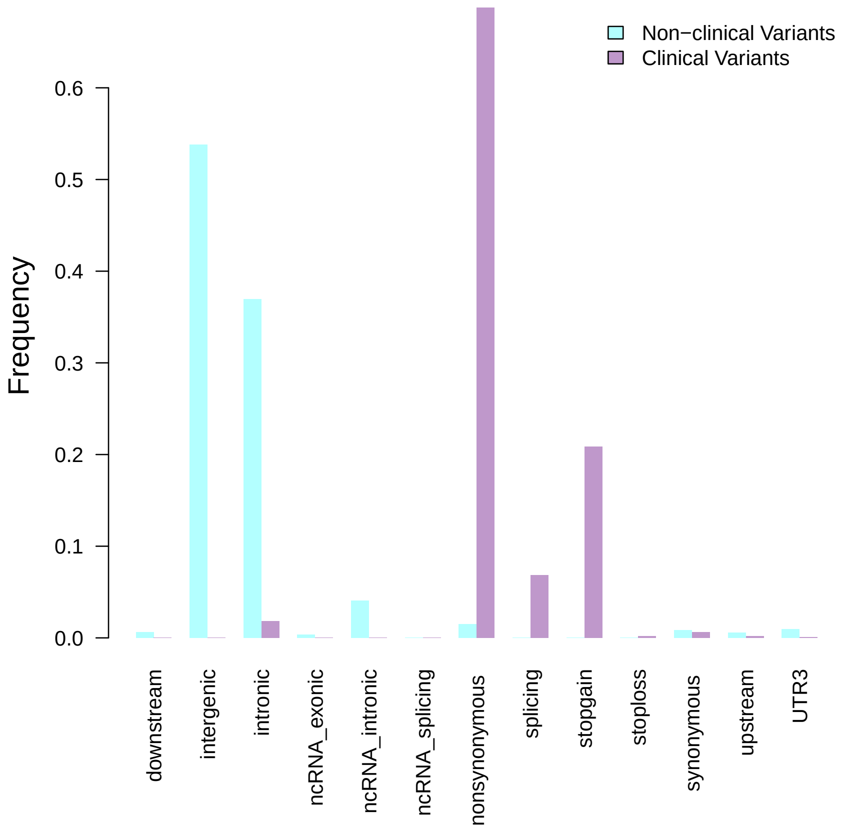


Figure 2.3: The protein-function distribution of mutations in non-clinical data and clinical data. The non-clinical data were the non-pathogenic variants in dbSNP137. The clinical data were the pathogenic or likely pathogenic variants in Clinvar. These variants were annotated with Jannovar [155, 156].

2.4.2 Sequence conservation

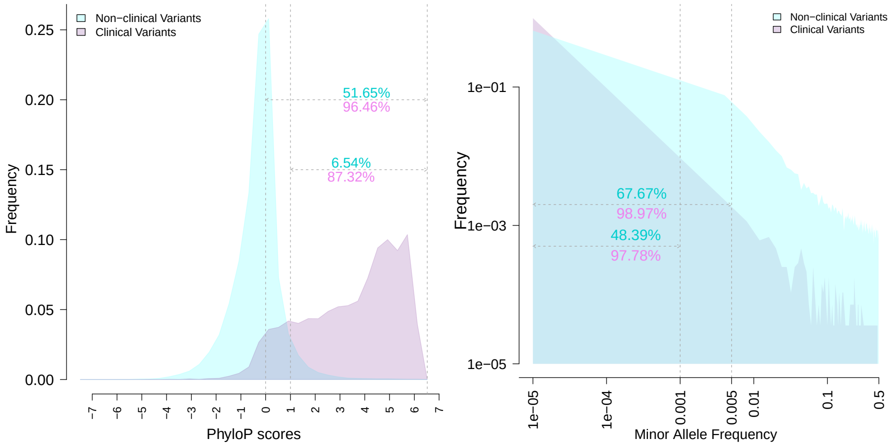
A typical human genome carries around 300-600 nonsynonymous mutations that are found in ¡ 1% of the population at large, and not all nonsynonymous mutations are deleterious. From the evolution point of view, nonsense mutations are null mutations. Missense mutations are the mixture of null and neutral mutations. The effect of missense mutations on molecular function, phenotype and organism fitness can be extremely diverse. Some missense mutations can be lethal or cause severe Mendelian disease. Some missense mutations can be mildly deleterious, neutral or beneficial. Relying on computational prediction programs, we can further quantify the functional significance of mutations [157]. The prediction program classifies variants into 'conservation' and 'acceleration', where 'acceleration' means the position is experiencing faster than neutral evolution, and 'conservation' means slower than neutral evolution. Most prediction methods can predict that $70\% - 90\%$ of the amino acid substitutions in HGMD [158], OMIM [153] and Swiss-Prot [159] are damaging [160, 154, 161, 162].

In this project, I used the phyloP score based on the alignments of the 44 ENCODE regions [163], which constituted the largest published comparative genomic data set for mammals [164, 165]. Variants with positive phyloP scores are conservative and indicate slower evolution than neutral drift. A higher score for a variant means that it is more conservative and deleterious. Variants are neutral if their phyloP scores are negative. Figure 2.4b a) showed that most of the clinical variants were conservative (score from 0 to 7). Around 1% clinical data were synonymous mutations and intronic mutations which had small phyloP scores. Therefore, I chose phyloP score = 1 as the threshold to include 88% clinical data.

2.4.3 Population allele frequencies

Allele frequency filter

Allele frequency is the most obvious filter for RVAS. It is the proportion of a particular allele occurring in a population. The incidence of rare disorders is commonly less than 0.001 [134, 15, 166, 167]. I investigated the allele frequency distribution for clinical variants and non-clinical variants. Figure 2.4b b) showed that the vast majority of clinical variants were rare ($< 0.1\%$) and less than half non-clinical variants passed the threshold. I chose an allele frequency cut-off of 0.1% in this project.



(a) (b)

Figure 2.4: a) PhyloP scores distribution in non-clinical and clinical data. b) Minor allele frequency distribution in non-clinical data and clinical data.

2.5 Residual variation Intolerance score

Petrovski et.al. introduced residual variation intolerance score (RVIS) score to rank genes according to the likelihood to affect disease based on Exome sequencing project (ESP) data. It predicted the expected amount of common functional variation based on the total amount of variants in each gene. Defining Y as the total number of common function variants in a gene and X as the total number of protein-coding variants. RVIS score was the studentized residual when regressing Y on X. A gene with a negative score was intolerant, whereas the gene with a positive score was tolerant [168]. In this work, I annotated genes with RVIS score.

Chapter 3

Methods and simulations

3.1 Similarity metric

Epidemiological studies involve large numbers of individuals. As the genetic background of individuals is relevant to disease-contributing variations, one concern of these studies is to identify and characterize the genetic backgrounds by their genomic profile. The admixture of populations or the cryptic relatedness in the studied data result in false positives and false negatives. The strategies for assessing the genetic backgrounds is to estimate the similarity score among samples by the great number of markers [169]

Similarity metric is a method to quantify the genetic similarity of a pair using a sets of markers. The simplest metric is to calculate the fraction of alleles shared Identity by state (IBS) over all the loci. we can use genetic similarity to infer the relatedness of individuals or to check a pedigree for correctness [170, 171, 172]. Moreover, we also can use it to estimate genotyping accuracy by calculating the distance to the reference set with high quality like 1000 genome data [173]. In the following, I will describe Identity by state (IBS) and its variations in detail.

3.1.1 Basic IBS metric

IBS metric assesses the genetic similarity by calculating the fraction of positions that shared identity-by-state. The more positions two subjects shared genotypes, the more similar the two subjects are.

IBS metric has many varieties by adapting the factors for weighting schemes, such as allele frequency [173, 174] or nucleotide conservation score [175, 176, 177].

Then one can set up an $N \times N$ similarity matrix S for N individuals with a similarity metric. Each element $S_{i,j}$ is the similarity score between individual i and individual j.

$$S_{i,j} = 1 - \frac{1}{C_{ij}} \sum_k I_{ij}(k) * W_{ij}(k) \tag{3.1}$$

where

$$I_{ij}(k) = \begin{cases} 1 & x_i(k) = x_j(k) \\ 0 & x_i(k) \neq x_j(k) \end{cases}$$

$W_{ij}(k)$ is the weight at position k and $C_{ij} = \sum_k W_{ij}(k)$ is used for normalization.

The underlying IBS metric calculates the fraction of alleles that any two individuals share purely by state. It is simple to determine how many alleles (0, 1 or 2) a pair of individuals shared. For any position k, the weight is:

$$W_{ij}(k) = 1 \tag{3.2}$$

In this thesis, IBS metric represented this metric. In the following, we introduce two varieties of IBS metric differed in the weighting schemes.

3.1.2 Weighted IBS - W^1

In the basic IBS metric, each position contributes equally to the distance. However, we can also weight each position differently. Due to the combined effects of exponential population growth and weak purifying selection, rare variants may excess in a population. The vast majority of protein-coding variations is evolutionarily recent and rare [178], they likely make a significant contribution to human phenotypes and disease susceptibility. Thus, it is reasonable to give higher weight to the rare variants for calculating similarity score. Each position was weighted by the inverse of genotype frequency, in which rare variants have higher weights [173]. The weight at each position shared between individual i and individual j:

$$W_{ij}k = \frac{1}{f(x_i(k))} \tag{3.3}$$

where $x_i(k)$ is the genotype of individual i at position k. $f(x_i(k))$ is the genotype frequency of $x_i(k)$, which is determined in a large population genetics studies such as 1KGP. This metric is designated as W^1 metric in this thesis.

3.1.3 Weighted IBS - W^2

In the W^1 metric, rare variants played an important role in estimating the genetic distance. As common SNPs can reflect a deep evolutionary history[179], we also studied the third metric, W^2, where common variants were given higher weight. The weighting scheme was built on Shannon's information theory. In this context, entropy H was a measure for the expected information content [180, 181]:

$$H = -\sum_{i=1}^{m} p_i log(p_i) \tag{3.4}$$

where m is the number of possible genotypes at this position and p_i is the probability for each genotype i.

We can generate the similarity matrix among samples with either of the three metrics IBS, or W^1, or W^2, and then apply it in matching strategies to find the similarity-matched neighbors, or in linear mixed model for accounting for population substructure. [64].

3.2 Davies-Bouldin Index

Davies-Bouldin Index (DB) is a clustering metric to evaluate how well two clusters are separated [182]. We used DB to estimate the level of separation between case and control group.

$$S_{cases} = \frac{2}{m * (m-1)} \sum_{i=1}^{i=m-1} \sum_{j=i+1}^{j=m} d_{ij}$$

$$S_{controls} = \frac{2}{n * (n-1)} \sum_{i=1}^{i=n-1} \sum_{j=i+1}^{j=n} d_{ij}$$

$$M = \frac{2}{n * m} \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} d_{ij}$$

$$DB = \frac{S_{controls} + S_{cases}}{M} \tag{3.5}$$

Where m is the size of case group, n is the size of control group, $d_{i,j}$ is the distance between samples i and j measured by similarity metric. Two clusters is well-separated if the DB score is low.

3.3 Rare variant association tests

CVAS commonly run single variant tests, which conduct test statistic, such as χ^2 test, for a single position. The typical significance threshold for single variant tests is 5×10^{-8} in CVAS, as one million common variants are expected in a large cohorts [183].

Single variant tests are theoretically also possible for low-frequency variants if the sample size is sufficiently large [184]. However, for rare disorders, it is usually not feasible to collect that many patients. Therefore, single variant tests does not work in RVAS.

Instead of testing each variant individually, RVAS usually conducts aggregation tests or burden tests, which evaluate cumulative effects of multiple genetic variants in a genomic region. Burden test collects information for multiple genetic variants in the same genomic region into a single genetic score and test the association between the score and the disorder. In this project, I used several different burden tests. Most of simulations were run by Cochran-Armitage test for trend (CATT), Combined Multivariate and Collapsing (CMC) tests and permutation tests, which were implemented in Java by myself. Variable threshold tests and composite likelihood tests contributed to a small part of results.

3.3.1 Cochran-Armitage test for trend

CATT tests are applied for categorical data analysis. It aims to test for the presence of an association between the responses and the ordered categories. In case-control association tests, the responses are the phenotype of individuals and the categories are different alleles or genotypes. Typically, we can set up a contingency table for genotypes. The affected and unaffected individuals are two responses while different genotypes (homozygous reference AA, heterozygous Aa and homozygous alternate aa) are three categories.

Instead of setting up contingency table for each position as in single variant tests, one can build the contingency table across a genomic region in aggregation tests, where each cell is the cumulative sum for a genotype in this region (Table 3.1).

Table 3.1: Contingency table for burden tests

Genotype	AA	Aa	aa	Total
Cases	O_{11}	O_{12}	O_{13}	R_1
Controls	O_{21}	O_{22}	O_{23}	R_2
Total	C_1	C_2	C_3	N

CATT is usuallt studied for the underlying trend. It emphasizes the importance of utilizing ordered categories in a contingency table [185, 186].

Hypothesis 1 *there is a linear trend in binomial proportions of cases across different genotypes.*

Null Hypothesis 1 *there is no linear trend in binomial proportions of cases across different genotypes.*

The linear regression model for CATT is:

$$y_i = \alpha + \beta * s_i, \tag{3.6}$$

where y_i is the real underlying proportion of cases in each genotype, and s_i is a score assigned to a genotype. s_i is suggested to be $\{0, 1, 1\}$ for the dominant model, $\{0, 0, 1\}$ for the recessive model and $\{0, 1, 2\}$ in the additive model [187, 188].

The null hypothesis can be written as:

$$H_0 : y_1 = y_2 = y_3$$

The alternative hypothesis is:

$$H_1 : y_1 \leq y_2 \leq y_3 \text{ , at least one strict inequality exists.}$$

To measure and test the significance of the trend in y_i, one can apply regression analysis of π on score s, The prediction equation under ordinary least squares fit is

$$\hat{y_i} = p + b(s_i - \bar{s}). \tag{3.7}$$

$$\bar{s} = \frac{\sum_{i=1}^{3} C_i * s_i}{N}, \tag{3.8}$$

$$\hat{y_i} = \frac{O_{1i}}{C_i}, \tag{3.9}$$

$$\bar{y} = \frac{R_1}{N}, \tag{3.10}$$

$$b = \frac{\sum_{i=1}^{3} C_i * (\hat{y_i} - \bar{y}) * (s_i - \hat{s})}{\sum_{i=1}^{3} C_i * (s_i - \hat{s})^2} \tag{3.11}$$

The test statistic for CATT is:

$$z^2 = \frac{b^2}{\bar{y} * (1 - \bar{y})} * \sum_{i=1}^{3} O_{1i} * (s_i - \hat{s})^2 \tag{3.12}$$

has an asymptotic chi-squared distribution with $df = 1$.

3.3.2 Combined Multivariate and Collapsing test

The collapsing method involves collapsing genotypes across variants in a region and then applying a univariate test on the collapsed contingency table. It is a powerful method for analyzing rare variants if the proportion of causal variants is high. However, power of collapsing methods may reduce significantly if the nonfunctional variants are misclassified. In contrast, a multivariate test is robust in the presence of mis-classification of non-causal variants, although it is not as powerful as collapsing methods. In order to integrate the merits of both collapsing and multiple-marker tests, Li et.al. proposed Combined Multivariate and Collapsing (CMC) method [189].

Collapsing Method Define an indicator variable X for the j^{th} case individuals as

$$X_j = \begin{cases} 1 & \text{rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

An individual rarely carry more than one variants in a region because of the rarity of variants. The way to collapse genotypes across all sites in a region is: the variable for an individual is one if a rare allele presents in this individual at any site and otherwise zero. Then one can check whether the proportions of individuals carried variants differ between groups via association tests.

Multivariate Test The multivariate test can test many variants simultaneously, such as Hotelling's T^2 test. In this project, I used Hotelling's T^2 test for the multivariate test. Following Xiong et.al [190], an indicator variable X_{ji} is defined for the i^{th} site for the j^{th} individual in the affected population:

$$X_{ji} = \begin{cases} 1 & \text{Genotype is AA} \\ 0 & \text{Genotype is Aa} \\ -1 & \text{Genotype is aa} \end{cases}$$

Similarly, Y_{ji} is for unaffected population. Let

$$X_j = (X_{j1}, ..., X_{jM})^T \tag{3.13}$$

$$Y_j = (Y_{j1}, ..., Y_{jM})^T \tag{3.14}$$

$$\bar{X}_i = \frac{1}{N_A} \sum_{j=1}^{N_A} X_{ji} \tag{3.15}$$

$$\bar{Y}_i = \frac{1}{N_{\bar{A}}} \sum_{j=1}^{N_{\bar{A}}} Y_{ji} \tag{3.16}$$

$$\bar{X} = (\bar{X_1}, ..., \bar{X_M})^T \tag{3.17}$$

$$\bar{Y} = (\bar{Y_1}, ..., \bar{Y_M})^T \tag{3.18}$$

Where M is the number of markers in this region, N_A is the number of affected individuals and $N_{\bar{A}}$ is the number of unaffected individuals. The covariance matrix of the case and control groups is

$$S = \frac{1}{N_A + N_{\bar{A}} - 2} \left\{ \sum_{j=1}^{N_A} (X_j - \bar{X})(X_j - \bar{X})^T + \sum_{j=1}^{N_{\bar{A}}} (Y_j - \bar{Y})(Y_j - \bar{Y})^T \right\} \tag{3.19}$$

Hotelling's T^2 statistic is denoted as

$$T^2 = \frac{N_A N_{\bar{A}}}{N_A + N_{\bar{A}}} (\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y}) \tag{3.20}$$

Under the null hypothesis that no variants is disease-associated,

$$\frac{N_A + N_{\bar{A}} - M - 1}{M(N_A + N_{\bar{A}} - 2)} T^2 \tag{3.21}$$

is asymptotically distributed as an F distribution, with M and $N_A + N_{\bar{A}} - M - 1$ as degrees of freedom for a large sample size of cases and controls.

$$\frac{N_A + N_{\bar{A}} - M - 1}{M(N_A + N_{\bar{A}} - 2)} T^2 \sim F(M, N_A + N_{\bar{A}} - M - 1) \tag{3.22}$$

Combined Multivariate and Collapsing Method The CMC method combines the collapsing strategy and the multivariate test. It firstly classifies makers into subgroups with predefined criteria, such as allele frequency or protein function. Then markers are collapsed into a single score within each group. In the end, the multivariate test is applied to all subgroups. In the exome-wide data, we took gene as a genomic region. Since we tested only the rare variants in this project, we did not further divide a genomic region into subgroups according to allele frequency. Therefore, the CMC method in this project only worked as a multivariate test.

3.3.3 Variable threshold test

The Variable threshold (VT) test is based on the intuition that some threshold T for which variants with a MAF below T is more likely to be functional than those variants with a MAF above T [191]. Test statistics can run on each allele frequency threshold T. Price et al. [191] used z-score test, defined z_{max} as the maximum Z-score across values of T, and assessed the statistical significance of z_{max} by permutations on phenotypes. The z-score test was calculated as follows:

$$z(T) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} {\xi_i}^T C_{ij} (\pi_j - \bar{\pi})}{\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} ({\xi_i}^T C_{ij})^2}} \tag{3.23}$$

where n was the total number of samples (including cases and controls). m was the number of variants in the tested genomic region. ${\xi_i}^T$ was an indicator variable that was equal to 1 if the frequency of SNP i was below the threshold T and otherwise zero. π_j was the phenotype of sample j. $\bar{\pi}$ was the mean value of π_j across samples. C_{ij} was the reference allele count of SNP i in sample j. z(T) was proportional to a standard normal variable.

3.3.4 Composite likelihood ratio test

The Composite likelihood ratio (CLR) test is designed to evaluate whether a gene or other feature contributes to disease risk. The statistic value of this test is the ratio of two likelihood functions based on the null hypothesis and the alternative hypothesis. In the burden test, we first collapsed variants in a genomic region, then calculated the maximum likelihood which was equal to the observed frequency of the minor allele [192]. So the log-likelihood ratio was as follows:

$$\begin{aligned} \lambda &= -2log(\frac{L_{Null}}{L_{Alt}}) \\ &= -2log\frac{(\frac{m}{n})^m(1-\frac{m}{n})^{(n-m)}}{(\frac{a}{n_a})^a(1-\frac{a}{n_a})^{(n_a-a)}(\frac{u}{n_u})^u(1-\frac{u}{n_u})^{(n_u-u)}} \end{aligned} \tag{3.24}$$

Where m was the total number of mutations in a given gene. n was the total observed genotypes in this gene among n_a cases and n_u controls. a was the number of mutations appeared in n_a cases and u was the number of mutations in n_u controls [193]. The probability distribution of the test statistic λ was approximately a chi-squared distribution with one degree of freedom.

3.3.5 Permutation test

Permutation tests do not make any assumption on the specific distribution of the underlying data, the basic assumption is only that it is possible that all of the treatment groups are equivalent. Thus, its null distribution is obtained by a random assignment of samples to case and control groups. The procedure of permutation test is following:

step 1: compute the observed test statistic (T_{obs}).

step 2: enumerate all permutations of the labels (N permutations).

step 3: read the fist permutation of the labels and assigned to each group

step 4: calculate the test statistic for the shuffled data.

step 5: go to next permutation of the labels until all permutations are tested and repeat step 3-4.

step 6: use all test statistics from step 3-5 to construct the null distribution.

step 7: find where the observed test statistic located in null distribution.
step 8: the permuted p-value ($T_{permut} = (M + 1)/(N + 1)$, where M is the number of test statistics not smaller than T_obs).

For large data, to enumerate all permutations is very computationally intensive. To balance the advantage of permutation test and its computational cost, Monte Carlo sampling is proposed. The trick is that it randomly generates a small number of permutations (like 1000) instead of all permutations at step 2 [194, 195].

3.4 Multiple testing corrections

When multiple hypotheses are tested in one experiment, the rate of false positives can significantly increase. In the random scenario, if m hypotheses are tested simultaneously, $m \times \alpha$ false positives are expected,where α is the significant level. In order to restrict the Family-wise error rate (FWER), a stricter significance for multiple tests is required.

3.4.1 Bonferroni corrections

Bonferroni adjustment is one approach for multiple-tests correction. The idea of Bonferroni correction is that: in order to control the expected significance level for the entire family of tests at most α, the significance level of each single test should be $\frac{\alpha}{m}$, where m is the number of tests.

Let $H_1, ..., H_m$ be a family of hypotheses and $p_1, ..., p_m$ be p-values of each hypothesis. The Bonferroni correction states that choosing all $p_i \leq \frac{\alpha}{m}$ will control the $FWER \leq \alpha$. The proof follows from Boole's inequality:

$$FWER = Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_o}\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m_0 \times \frac{\alpha}{m} \leq m \times \frac{\alpha}{m} = \alpha$$

where I_0 is the subset of the true null hypotheses, having m_0 members. This result does not require that the tests be independent. Although the Bonferroni correction restricts false positives, it become very conservative when the number of tests is large. In turn, it increases the risk of generating false negatives.

3.4.2 Experiment-wide significance

When the study is small, the p-values from association tests are usually not as significant as the expected under the null hypothesis. Bonferroni correction will be too conservative in this case, therefore another approach to correct for multiple testing is proposed. It calculates an experiment-wide significance level with permutation method [141, 196].

The procedure is quite similar as the way to generate the permuted p-values. Instead to construct a null distribution for each test, it construct a null distribution for the minimal p-value of all tests in each permutation.

Supposed that M tests (such as M genes to be tested) are included in the data. The steps are as follows:

step 1: run statistical test on the original to get the observed p-values for all M tests.

step 2: assuming that the categories of all samples are unknown, we re-sample the labels for all samples.

step 3: run the same test on the new label samples data and get the smallest empirical p-value.

step 4: repeat Step 2-3 for N times, such as $N = 1000$ or 10000....

step 5: construct the empirical distribution of the smallest p-values.

step 6: the value at the significance level(like 0.05) of this distribution is the empirical significance for observed tests.

Unlike permutation p-value which is independent of statistical tests, empirical significance depends on the statistical tests. Thus, the challenge of experiment-wide significance is to select an appropriate statistic test.

3.5 Readout of statistical tests

The classical approach to test hypothesis includes setting up a null hypothesis (H_0) and an alternative hypothesis (H_1), calculating a test statistic (T) in a statistical test from the observed data, then finally deciding whether to reject H_0 [197].

Power Statistical power is the probability that the test correctly rejects the null hypothesis (H_0) when the alternative hypothesis (H_1) is true.

$$power = P(reject \quad H_0|H_1 \quad is \quad true) \tag{3.25}$$

Family-wise error rate FWER is the probability of making one or more type I errors, where the test incorrectly rejects a null hypothesis, among all hypotheses tests.

$$FWER = P(V >= 1) \tag{3.26}$$

where V is the number of type I error.

Top-ranked rate Top ranked rate is the probability that a test has the lowest p-value when H_1 is true. The reason for using a top ranked rate is that: it is not possible to collect a large number of cohorts for rare disorder, so it is hard to achieve a significant p-values t reject H_0. Therefore, it is reasonable to use the 'top-ranked rate' to evaluate the performance in RVAS. In this thesis, I used an alternative term "top-ranked rate" or "disease causing gene is top ranked".

3.6 Accounting for confounders

In case-control association tests, population stratification due to individuals from the multiple source populations, and cryptic relatedness due to the relatedness among individuals, are confounding factors. they may lead to false association signals. It is therefore important to account for these confounders [198].

3.6.1 Accounting for population stratification

Population substructure between case and control groups is a major confounding factor in case-control association studies that can cause spurious associations. Some methods have been developed to correct these confoundings. The principle of the correction methods is to describe the effect of the confounding genetic structure as random effects and quantify the covariation regarding the degree of genetic relatedness among the samples. In GWAS, principal-component analysis (PCA) and linear mixed models are popularly used. The principal components in PCA are estimated from a genome-wide covariance matrix holding all genotyped markers for all case-control individuals. However, PCA is unlikely to correct for cryptic relatedness present in the data [53].

The linear mixed model also uses an empirical covariance matrix to account for both pedigree and population structure. It can correct the empirical relatedness matrix encoding a wide range of sample structures, including both hidden relatedness and population stratification. To make clear how the ’direct adjustment’ methods performed on rare variants, we tried EMMAX to correct the confoundings in this work. EMMAX is based on the linear mixed model [64] and is implemented in EPACTS package [199]. In linear mixed models, the phenotype is typically modeled as the sum of a fixed linear regression, including the effects of the marker to be tested and a random linear-additive term that accounts for unwanted confounding structure. The idea of confounder correction with linear mixed model is to assume that the effects of confounding genetic structure randomly exist. We can evaluate the covariation of these confounder’s effects according to the genetic relatedness between samples. Phenotype y is written as the mixed sum of a linear term in the fixed effects β and random effect μ, that The general variance component approach is as follows:

$$y = X\beta + G\mu + \epsilon \tag{3.27}$$

Where G is a $N \times S$ matrix holding S genotyped markers for N individuals. Each locus is assumed to have the equal effect of the total genetic variance σ^2. The S loci included in matrix G is assumed to have a mean of zero and unit variance. The realized relationship matrix (RPM) is defined as the empirical covariance matrix

$$K_{RPM} = \frac{1}{S} G G^T \tag{3.28}$$

the random genetic effect

$$G\mu \sim \mathcal{N}(0, \sigma^2 K_{RPM}) \tag{3.29}$$

RPM is used to capture the confounding variation in the phenotype. With marginal likelihood method where the random genetic effect was marginalized out, we could know the genetic variance ${\sigma_g}^2$.

3.6.2 Accounting for cryptic relatedness

Cryptic relatedness means that some of the individuals in case-controls cohorts may have close relatedness. This situation violates the assumption of the case-control association study that all genotypes are independent draws from the overall population frequencies. Thus, it may lead to a larger variance than expected and further result in the false positive association in the association tests [59, 62]. Due to DNA sample mix-ups, cryptic relatedness may exist between samples.

The sample-relatedness can be investigated using both Identity by descent (IBD) [200] and IBS estimations [201, 202]. In this thesis, we investigated sample-relatedness using IBD estimates in PLINK [200]. The downstream analysis in IBD infered the possible relationships between the set of four alleles of two individuals when assuming symmetry between maternal and paternal gametes [203].

With the *–genome* option in PLINK, it is easy to compute pairwise kinship estimates between any individuals. PLINK infers the relationship types, such as siblings, parent-child and unrelated, with the proportion of loci where individuals share zero, one and two alleles identical by descent. If the probability of a pair sharing two alleles IBD is around one, it means that this pair is monozygotic twins, or a pair is replicates of a single sample. If a pair shares zero alleles IBD at every locus, then they are unrelated. If a pair shares one allele IBD at every locus, then they are the parent-child relationship. The relatedness between samples can be inferred by the proportion that two individuals share identity by descent positions. For full siblings, they share respectively 25% zero allele IBD, 50% one allele IBD and 25% two alleles IBD in the genome. The proportion of identity by descent for full siblings is 25% in an infinitely large panmictic population.

3.7 Simulation

3.7.1 Case group setup

In reality, the patients with the same disorder (especially rare disorder) are disseminated all over the world. In the simulation, I randomly chose samples from a pool as case group. The number of samples in the simulated cases was from 5 to 60. A cohort of five patients was typical for rare disorders while a cohort of 60 patients was large enough for a rare disorder. I performed simulations on BER and 1KGP data.

3.7.2 Spike pathogenic mutations

To simulate real cases, I further spiked causal mutations of a rare disorder into cases. I simulated three dominant disorders and five recessive disorders. All pathogenic mutations in these disorders were from the HGMD database [158]. In these eight disorders, three of them were heterogeneous disorders. The other five disorders were homogeneous disorders. All disorders were monogenic (Table 2.1), although the disorder in different patients may be caused by different genes (details in Section 2.2).

Figure 3.1 shows the tree structure among disorders, genes and mutations. s_i, $i = (1, ..., 8)$, is the disorder of interest. g_{ij}, $j = 1, ..., G$, is the j^{th} disease-associated gene of disorder s_i, G is the total number of disease-associated genes for a disorder. w_{ij} is the prevalence that patients carried mutations on this disease-associated gene. g_{ij} of disorder s_i. V_{ijk} is the k^{th} pathogenic mutation in gene g_j for disorder s_i.

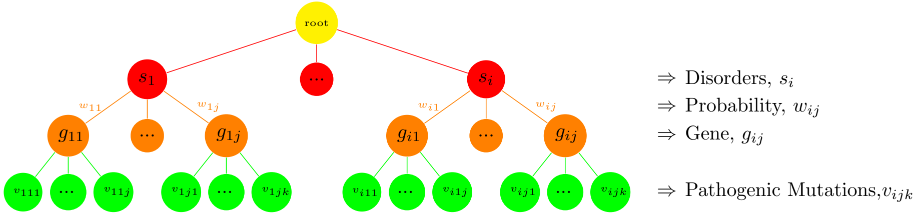


Figure 3.1: hierarchy of disorders, genes and mutations. Each disorder has one or more disease-linked genes, and each gene has its own prevalence. Each gene contains many pathogenic mutations (details in Section 2.2).

Figure 3.2 describes the process of selecting pathogenic mutations for case group. When selecting the mutations to spike into cases, all disease-associated genes of the disorder were obtained, then one of these genes was picked up following their prevalence. The list of pathogenic mutations in the selected gene was then read out. If the inheritance model of the simulated disorder was autosomal dominant, I randomly chose one mutation in the list a and set the genotype of the patient to be heterozygous. If the inheritance model was autosomal recessive, then I randomly chose one or two mutations from the list. If only one mutation was selected, then we set the genotype at this site for the patient as homozygous. Otherwise, if two mutations were selected, then I added two heterozygous mutations to the variants profile of this patient. All added positions were set to be reference homozygous for control individuals.

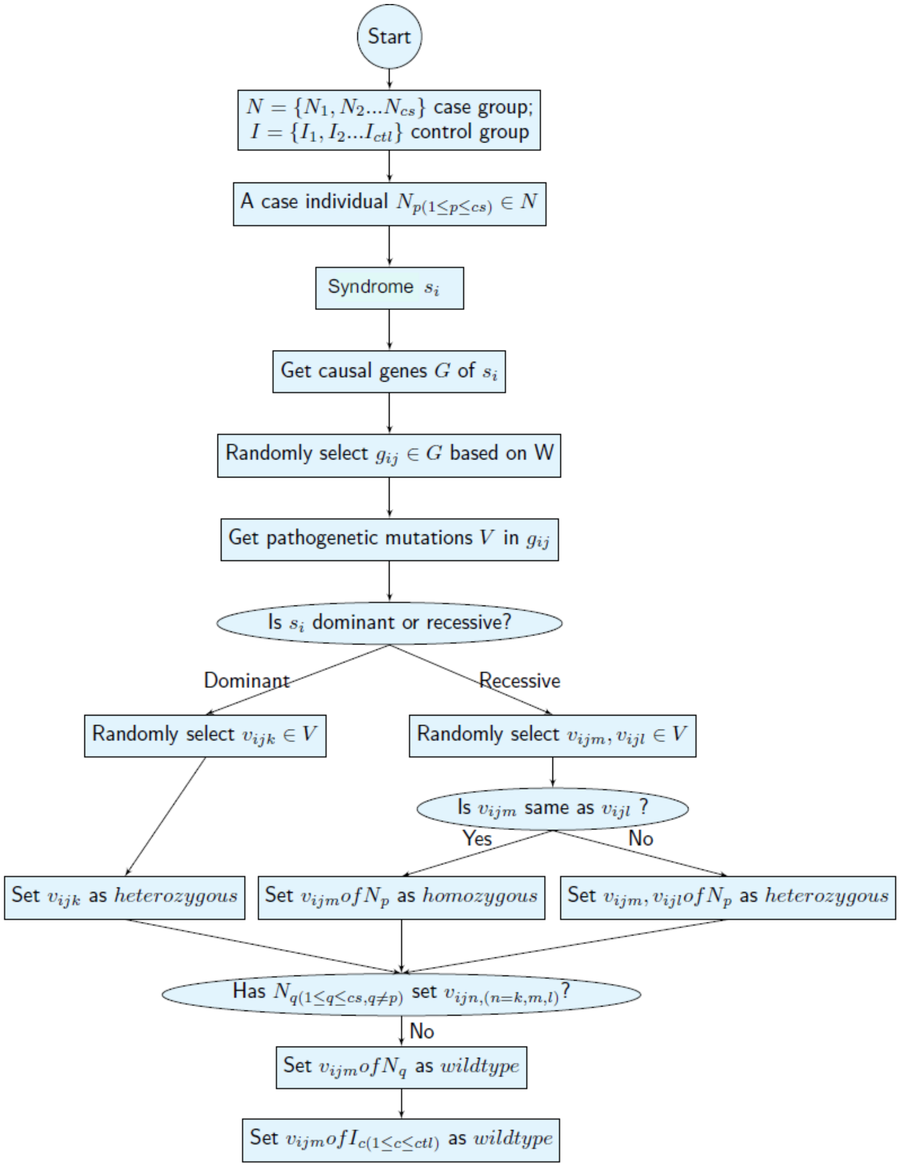

Start
$N = \{N_1, N_2...N_{cs}\}$ case group;
$I = \{I_1, I_2...I_{ctl}\}$ control group
A case individual $N_{p(1\leq p\leq cs)} \in N$
Syndrome s_i
Get causal genes G of s_i
Randomly select $g_{ij} \in G$ based on W
Get pathogenetic mutations V in g_{ij}
Is s_i dominant or recessive?
Dominant
Recessive
Randomly select $v_{ijk} \in V$
Randomly select $v_{ijm}, v_{ijl} \in V$
Is v_{ijm} same as v_{ijl} ?
Yes
No
Set v_{ijk} as $heterozygous$
Set $v_{ijm} of N_p$ as $homozygous$
Set $v_{ijm}, v_{ijl} of N_p$ as $heterozygous$
Has $N_{q(1\leq q\leq cs, q\neq p)}$ set $v_{ijn,(n=k,m,l)}$?
No
Set $v_{ijm} of N_q$ as $wildtype$
Set $v_{ijm} of I_{c(1\leq c\leq ctl)}$ as $wildtype$

Figure 3.2: Workflow of spiking causal mutations into cases. The number of mutations chosen for a patient depended on the inheritance of the disorders. The probability to a gene being selected was dependent on its prevalence.

3.7.3 Control Group Setup

The ideal situation in GWAS is that the individuals in controls groups have the same population background as case groups. This can minimize the likelihood of generating false positive associations due to various biases and confounding data. In practice, the population stratification between cases and controls is the well-known. The matched ethnicity between case and control groups can minimize the population stratification [204].

As the 'direct adjustment' methods cannot work in RVAS (see Section 1.4.3.2), the 'matching strategy' is a possible alternative to correct for the population stratification in RVAS, which reduces the potential stratification at the design stage.

In GWAS, the number of controls was commonly equal to the number of cases [46, 41, 38]. This is based on the reaseons that the correction methods for population substructure worked best if the group size of case and control was the same (according to the personal correspondence with EMMAX group). However, it has been proposed that including more controls can increase power. The study of Zondervan and Cardon [205] showed that the control:case ratio up to ~ 3 to 4 would get the best performance while including as many control as possible can maximize the power showd in the study of Zhuang et.al [140], although the rate of false positives increased as well.

In the current work, I initialized the size of the control group as the same as case group and then expanded the control group as large as possible. I used two strategies to set up the control group, either in a similarity-matched way or in a random way. In the similarity-matched method, I chose the individuals who were genetically closer to the cases. In a random way, control individuals were randomly chosen. As shown in Figure 3.3, I ranked the similarity score for a case to all individuals in the pool. If control group had the same size as case group, then for each case individual, its first available nearest neighbor was its control individual. If the size of the control group was larger than the size of case group, we chose the first possible nearest neighbor of all cases in the first round, the second possible nearest neighbor of all cases in next round, and so forth.

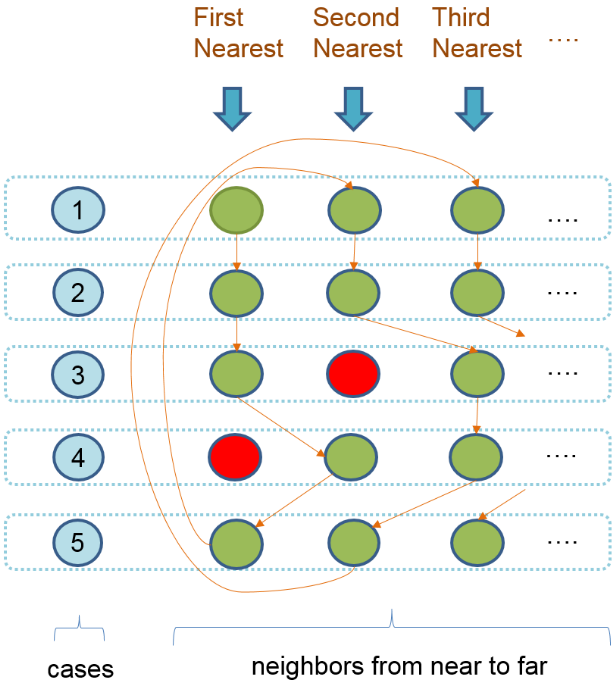


Figure 3.3: The strategy of choosing similarity-matched controls. The blue points represented cases, and the horizontal row represented the neighbors of each case from near to far. Green points represented the available individuals in the remaining pool. The red points meant that the individuals were not available in the remaining pool, as they were already in case group or control group.

3.8 Summary of the chapter

In this chapter, I described the methods and simulations used in this thesis. The similarity metrics are the basis of the 'matching strategy'. The 'matching strategy' chose the similarity-matched individuals as control group based on the genetic similarity among individuals. Several statistical tests were applied to detect the associations, such as tests (CATT, CMC test, VT test and CLR test). I implemented CATT, CMC test and permutation test in Java and used them in most simulations. CLR test was used in the genomic matchmaking study. VT test with and without EMMAX correction had been implemented in the EPACTS package and was used in the part of verifying whether EMMAX works in RVAS. I studied the performance of RVAS with matching strategy from the power and FWER, which were widely used statistical term in association studies. As the limited power observed in the simulations, I also took top-ranked rate as a readout. To control the FWER, I used the multiple tests corrections or the experiment-wide significance in the results. The cryptic relatedness in this work was accounted for by PLINK. RVIS was used to investigate the characteristic of the frequent false positive genes in the following section. Finally, the workflows of setting up simulations were introduced.

Chapter 4

Results

In this chapter, I firstly introduced the procedure to prepare the data used for simulation. I showed the data quality of BER and 1KGP cohorts and the genetic relatedness presented in these data. Then I compared the advantages and disadvantages of the 'matching strategy' in RVAS. I showed the improvement of RVAS with matching strategy achieved in several readouts, such as enhanced power and top-ranked rate and declined FWER. I also made a study of the frequent false positive genes. I further studied the factors which had impacts on the performance of RVAS, such as the data quality of samples, the inheritance model and heterogeneity of disorders, the number of controls, the statistical tests and the variants filters. I also showed the application of RVAS on identifying the disease-causing gene *TGDS* for Catel-Manzke Syndrome. Finally, I studied the challenge of recruiting the samples with significantly different quality and of collecting the cases with the same phenotypes from the genomic matching community.

4.1 Data quality of two cohorts

Rare variants are sensitive to data quality. The low data quality may result in large number of false rare variants. Here I compared the data quality of

BER and 1KGP cohorts from two aspects: read coverage and the genotyping accuracy.

4.1.1 Read coverage

To evaluate the data quality, I firstly compared the fraction of the exome target regions which were at least covered by 20 reads. The median of this fraction was higher in 1KGP than BER (Figure 4.1). I also calculated this fraction for the subsets of 1KGP samples (S_{IBS}, S_{W^2} and S_{W^1}) which were the most similarity-matched groups as BER cohorts (based on metric IBS, W^2, W^1). The mean coverage of the cohorts S_{IBS}, S_{W^1}, S_{W^2} differed from that of the entire 1KGP cohorts, indicating that the similarity metric is affected by data quality. The entire data quality of S_{W^1} was lower than S_{W^1} and S_{W^2}, due to W^1 metric put more weight on rare variants, which were sensitive to data quality.

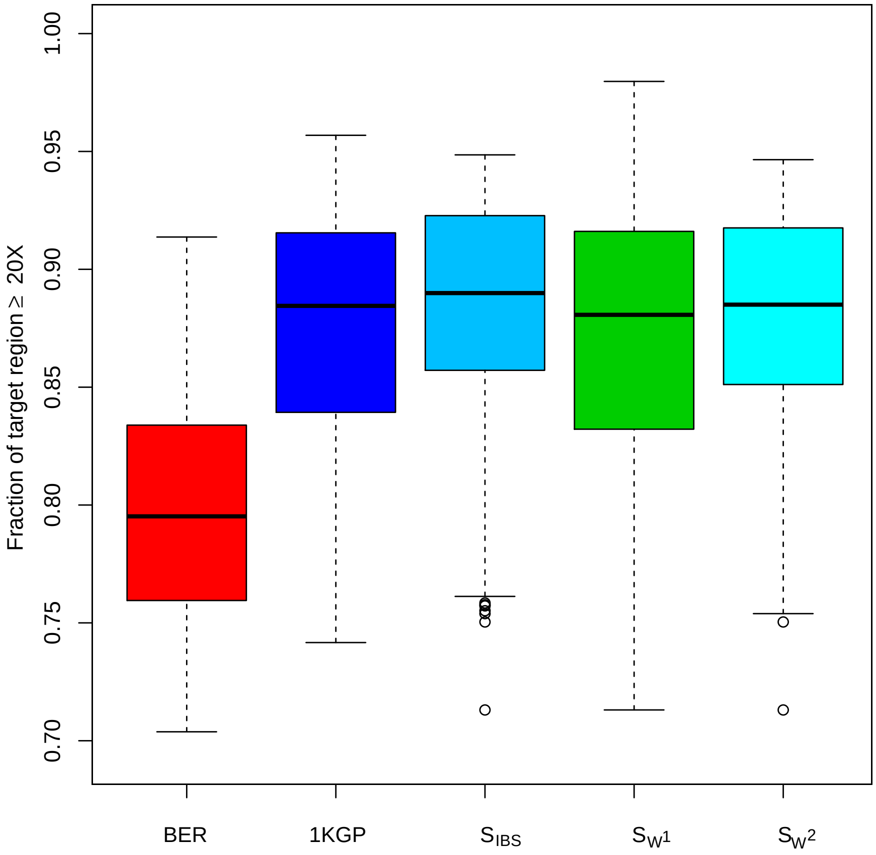


Figure 4.1: Exome data quality of BER and 1KGP cohorts. The fraction of the target regions that are covered by $\geq 20X$ correlate with the data quality of exomes. The mean coverage of the 1KGP data is higher than that of BER cohorts. The subset of 1KGP S_{IBS}, S_{W^1}, S_{W^2} are composed of the 1KGP individuals who are the first available closest individual for each BER individual in different similarity metrics (IBS, W^1, W^2). The mean coverage of the cohorts S_{W^1} differs from that of 1KGP cohorts, indicating that metric W^1 can match data quality too.

4.1.2 Genotyping quality

W^1 is especially sensitive to genotyping errors as it puts higher weight on rare variants. The distance to a reference dataset with high quality can be used to assess the data quality of a test sample [173]. A sample of high genotyping quality is closer to the reference set than a sample with comparable ethnicity but low genotyping quality. Normalization of this distance yields a dissimilarity score that can be used for a quantitative comparisons. The genotyping quality of samples can be evaluated by measuring the dissimilarity scores between the tested samples and their nearest neighbors in the 1KGP reference set. The distribution of the dissimilarity score is set up for BER cohorts as well as British in England and Scotland (GBR) and Finnish in Finland (FIN) cohorts which are two populations in 1KGP data. The median dissimilarity score for BER cohorts is considerably higher than that of GBR and FIN (Figure 4.2). It is known that the majority of BER cohorts have European and Middle East background. Therefore, there is no matched population in 1KGP data. To remove the impact of the reference population, GBR and FIN cohorts are excluded from the reference set and the dissimilarity scores are recalculated. The exclusion of the subpopulations itself makes the entire dissimilarity scores only slightly higher. This indicates that the higher dissimilarity score in BER cohorts resulted from the poor genotyping quality but not from the background of the reference sets.

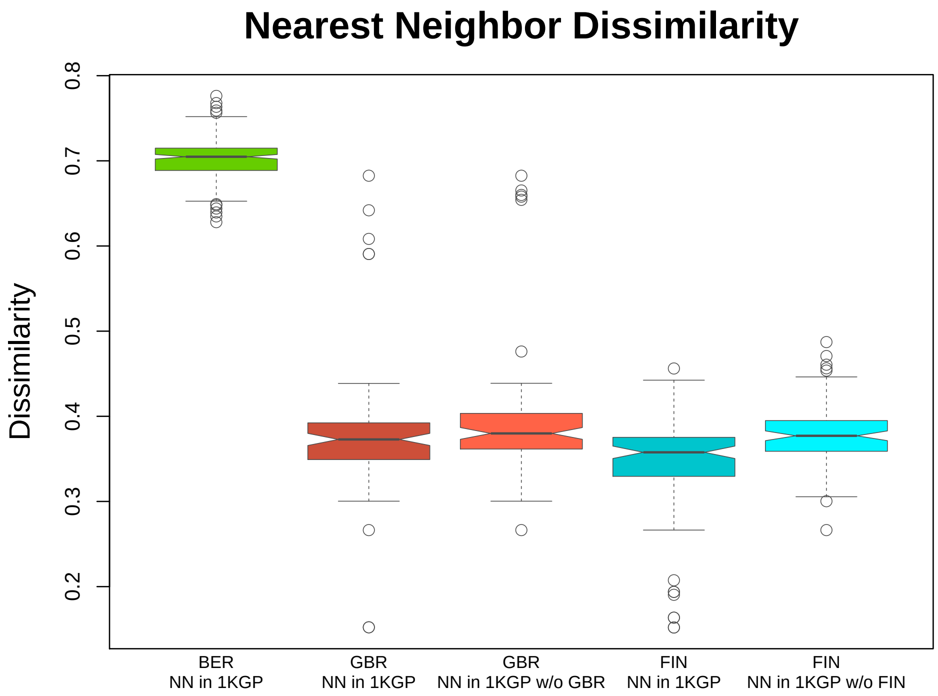


Figure 4.2: **Genotyping accuracy of BER data.** NN means the nearest neighbors. The genetic distance from the tested samples to a reference dataset of high genotyping accuracy correlates with the data quality. Comparing to GBR and FIN, the higher dissimilarity score in BER indicates its lower genotyping accuracy. Excluding GBR and FIN from reference sets subtly changed the distribution.

4.2 Cryptic relatedness in the data

In this work, I used PLINK to detect the cryptic relatedness hidden in the data. In PLINK, the relatedness can be estimated by $\theta = 0.5 * Z_2 + 0.25 * Z_1$, where Z_2 is the fraction that is identical in both copies, Z_1 is the fraction that is identical by descent in one copy. The cryptic relatedness is defined as $\theta > 0.1$. I iteratively excluded one sample in the related pairs. Finally, 33 individuals in

BER and nine individuals in 1KGP data were removed (Figure 4.3).

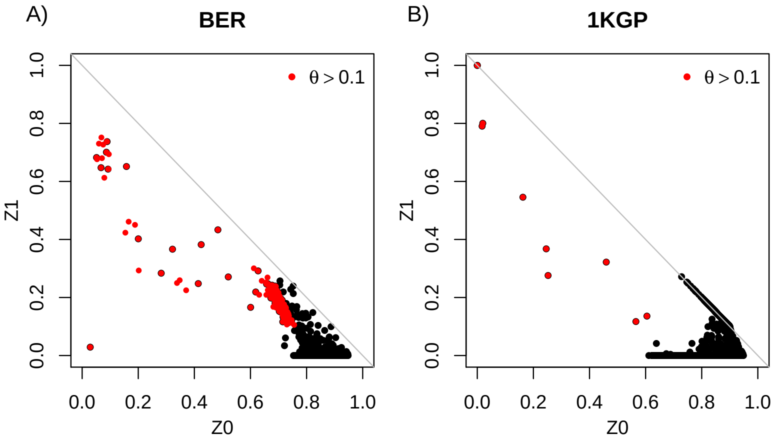


Figure 4.3: Identification of cryptic relatedness. PLINK generated the relatedness for BER (A) and 1KGP cohorts (B). Z_1 and Z_0 are the fraction of positions that are identical by one or zero copy for a pair of samples. $Z_2 = 1 - Z_0 - Z_1$ is the fraction of positions that are identical by two alleles. The relatedness of samples is estimated by $\theta = 0.5 * Z_2 + 0.25 * Z_1$. Red dots indicates pairs with $\theta > 0.1$ indicating cryptic relatedness.

4.3 Cluster analysis

4.3.1 Similarity metric in clustering population

Davies Boulding index (DB) evaluates intra-cluster similarity and inter-cluster differences. The low DB indicates a better separation between clusters and the tighter inside of a cluster (see detail in Chapter 3.2). The DB scores for pairs of populations were calculated to estimate the separation between them, where the similarities among individuals were based on different metrics IBS, W^1 and W^2.

There are 26 populations in 1KGP data, including East Asians, South Asians, Africans, Europeans and Americans. BER cohorts mainly consist of Europeans and Arabs, a few of Africans and Asians. As DB score is not a normalized metric, it cannot absolutely compare the level of separation among different metrics but it is clear to see the difference in separating clusters due to different metrics.

As shown in Figure 4.4, the highest DB score is for the test clusters to themselves. There are peaks for the tested samples to the populations from the same continents, which indicate the tested populations are close to the populations from the same continents. Comparing to metric W^2 and IBS, the tested population in 1KGP data have better separation to other populations in metric W^1 (as the peak is narrower). For instance, the distribution of DB scores for population FIN is sharper in metric W^1 than other metrics, that means, FIN is only close to GBR, while FIN is also close to population CEU, Iberian Population in Spain (IBS) and Toscani in Italia (TSI) in metric W^2 and IBS, thus the cluster of FIN is tighter with metric W^1. Moreover, for BER data, it is isolated to 1KGP populations if using metric W^1, because DB scores between BER and any 1KGP population are always zero without variability. Whereas, the distribution of DB score of BER data is rough in IBS and W^2, especially in metric W^2. The small peak for BER at region of European populations (GBR, FIN, CEU, IBS and TSI) in metric W^2 indicates that BER is close to European populations of 1KGP data.

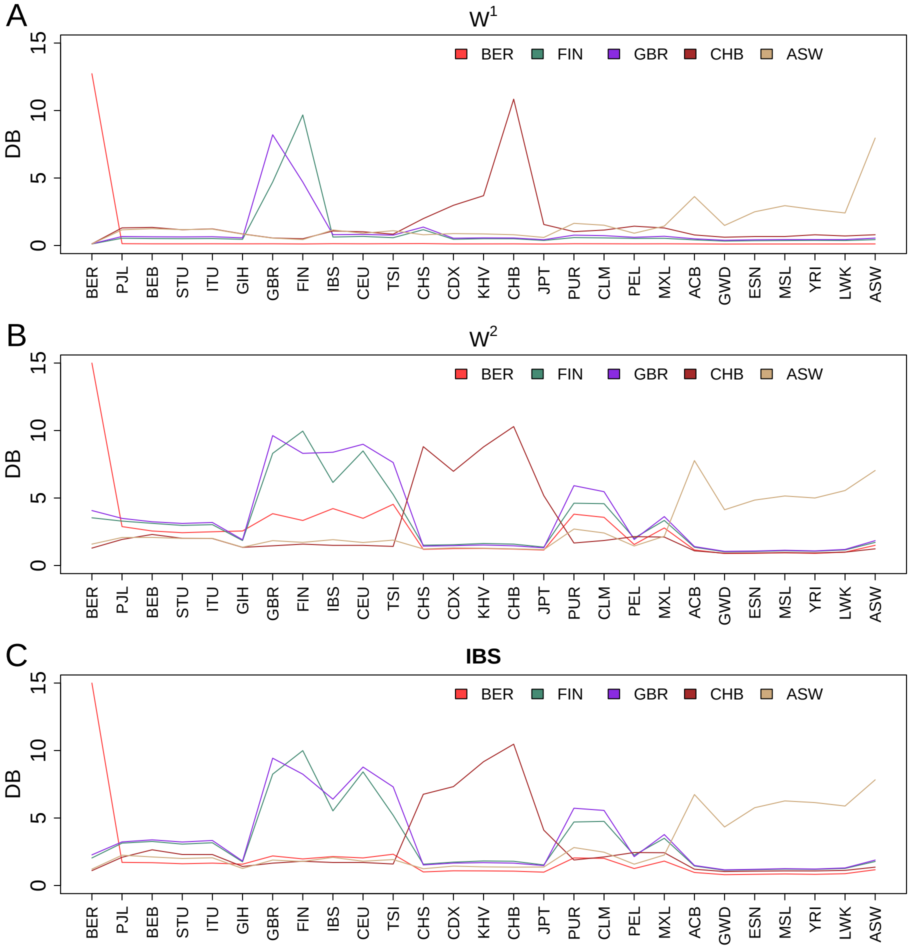


Figure 4.4: *DB score* between populations. The lower score means the better separation and tighter inside. A) the similarity used in calculating DB score was generated based on metric W^1. B) the similarity used for calculating DB score was generated based on metric W^2. C) the similarity used for calculating DB score was generated based on metric IBS.

4.3.2 Similarity metric in clustering case and control groups

I also used the Davies Boulding index (DB) [59] to quantify the separation between case and control groups. A low DB score (See Chapter 3.2) means that two clusters are better separated and the individuals in the two groups are genetically further apart. I chose case individuals from BER and chose the same number of controls from the remaining samples either randomly or matched by similarity. Here I presented the results for the similarity metric W^1, but the other metrics showed similar results. Compared to random selection, the similarity-matched control groups were more similar to the case group. This was especially prominent for small groups, as indicated by their higher DB scores (Figure 4.5). When group sizes increased, DB scores decreased in all selection strategies. Because the remainders in the infinite pool are limited, it leads to a large intersection between random controls and similarity-matched controls.

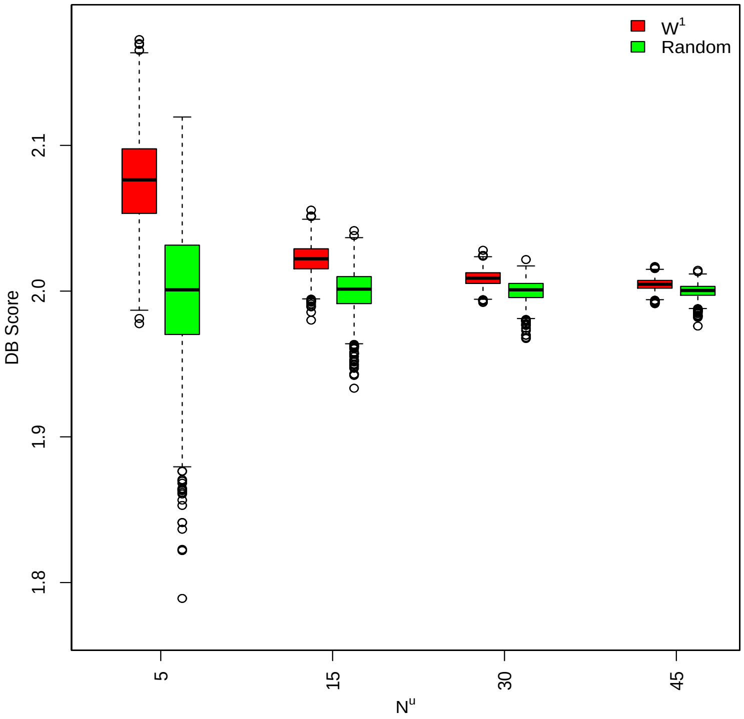


Figure 4.5: **DB score between case and control groups.** A lower score means control groups are genetically further from case groups. N^u means the number of samples in control group. Here case and control groups have the same number of samples (5, 15, 30 and 45) from BER. 1000 simulations were run for each setting.

4.4 The definition of allele frequency

The allele frequencies are usually based on population data of the 1KGP, Exome Aggregation Consortium (ExAC) or ESP. The profile of rare variants strongly depends on populations [206, 207]. Thus, the definition of allele frequency had a great impact on filtering variants. In addition, sequencing technologies and bioinformatics processing pipelines may result in artifacts as well [89]. Finding out the proper reference sets for the definition of allele frequency was firstly studied in this project.The methodology is to investigate the effect of excluding/including the population of the tested samples in the reference sets. The test samples were the cohorts FIN, Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) in 1KGP data and BER data. There was a initially comparable number of nonsynonymous variants in all cohorts. The variants were further filtered out if their allele frequency above 0.001, where the reference sets for the definition of allele frequency were 1KGP data after excluding FIN cohorts (Figure 4.6 A). For comparison, another allele frequency profile was built based on the reference sets including the entire 1KGP data as well as BER cohorts, and the variants were re-filtered based on this profile. With the new profile of allele frequency, the number of rare variants decreased dramatically, especially for BER and FIN cohorts (Figure 4.6 B).

The changes due to including or excluding the test samples in the definition of allele frequency showed that rare variants are population-specific. The inclusion of FIN removed a quarter of rare variants. The higher inbreeding coefficient in FIN may be responsible for the loss of the number of rare variants [208]. Furthermore, the dramatic change of the number of rare variants in BER caused by including BER in the reference sets indicated higher genotyping error rate in this cohort. A low data quality and many genotyping errors resulted in a substantially larger number of singletons in the BER cohort. Compared to BER and FIN, the number of rare variants in CEU were not affected much by the inclusion of BER and FIN. In general, the calling artifacts are randomly distributed over the genome and present as singletons. Such artifacts cannot be filtered out based on allele frequencies.

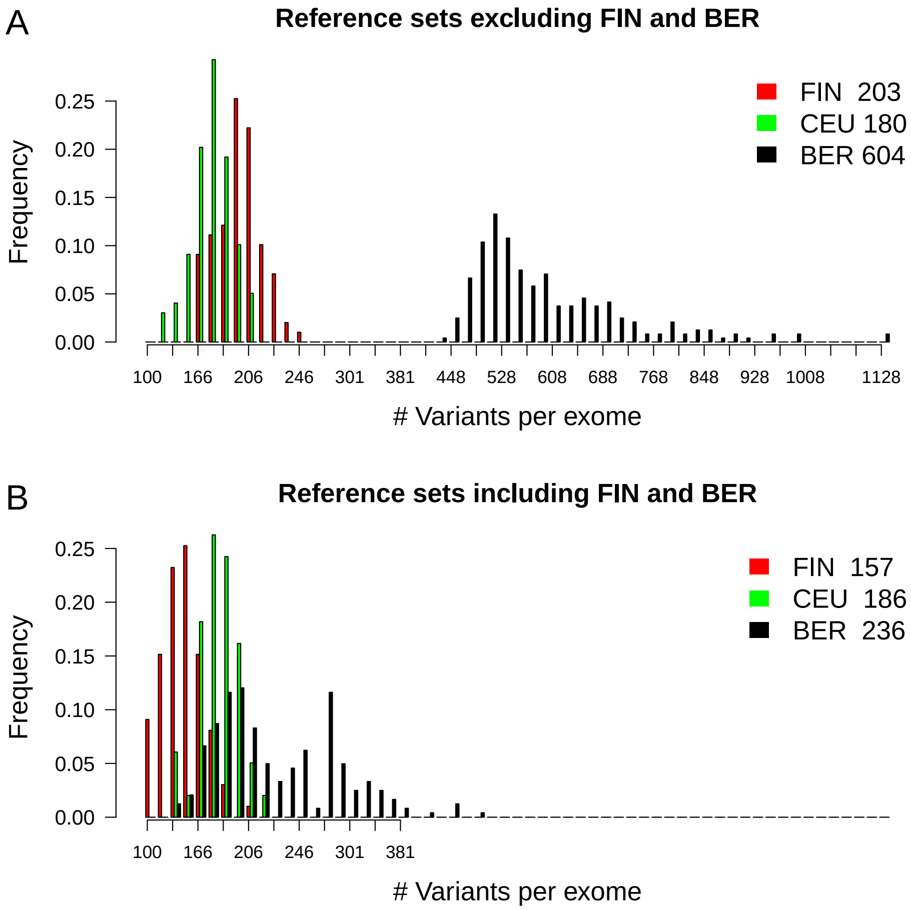


Figure 4.6: **Distribution of rare variants per exome for different populations and sequencing studies.** A) Excluding FIN and BER from the reference sets for calculating allele frequency. B) Including FIN and BER in the reference sets. The number of rare variants per exome decreases dramatically by including the tested population itself in reference sets, especially for BER.

4.5 ROC-like curves with different selection schemes

In statistics, a Receiver Operating Characteristic (ROC) curve demonstrates the performance of a binary classifier system when its discrimination threshold varies. Usually the sensitivity is plotted against specificity at various threshold settings. However, statistical power and FWER are more relevant in association studies. Therefore I showed a ROC-like curve with power against FWER. In Figure 4.7, power (see Equation 3.25) and FWER (see Equation 3.26) were computed for simulations of five cases and five controls from BER. Controls were chosen randomly or in a similarity-matched way. For each case sample, I spiked in either a homozygous or two heterozygous pathogenic mutations of HPMRS. I ran 1000 simulations and computed the power-FWER value pairs for the thresholds [0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9]. The p-values were computed by 10,000 permutations per test. Compared to random controls, RVAS with similarity-matched controls achieved lower FWER in any metric and a slightly higher power in the W^2 and IBS metrics. It indicated that performance was improved if population substructure was accounted for.

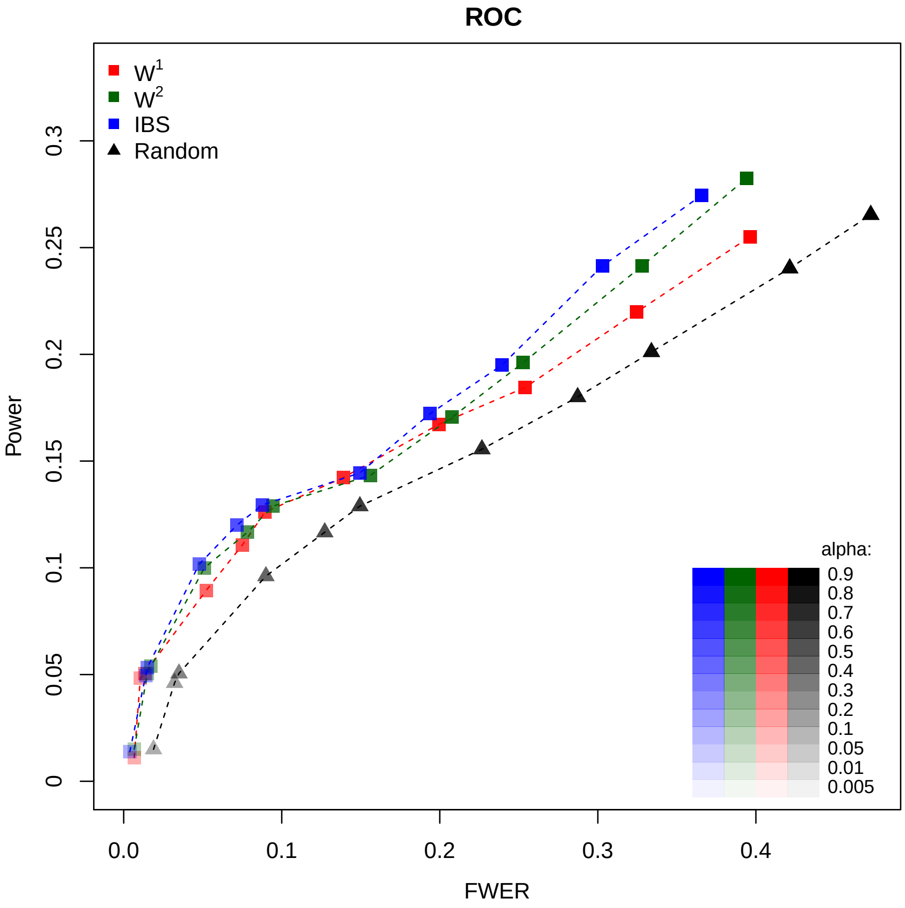


Figure 4.7: **ROC-like curves for power against FWER.** Case and control groups of size 5 were simulated by choosing individuals from the BER cohort either randomly or matched by their similarity. Three different metrics were used to infer kinship matrices: IBS, W^1 and W^2. Pathogenic mutations of Mabry syndrome (HPMRS) were spiked into the rare variant sets of individuals of the case group and permutation-based p-values from CATT were computed for every gene. The value pairs for power and FWER, were plotted for a range of significance levels ([0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9]), shown the color intensities at the lower right corner of the plot.

4.6 Power of RVAS in different disorders

I studied the power of detecting the disease-causing genes for different disorders. For all disorders, the power of the RVAS increases with the size of the case and control groups. In general, it was more likely to detect a true association in a disease with a recessive mode of inheritance, which was mainly due to the higher burden of pathogenic alleles. Moreover, associations in small disease genes with a highly conserved sequence like gene *TGDS* were easier to be detected. For the large and/or variable genes, the presence of many rare benign mutations diluted the contribution of the pathogenic mutations (Figure 4.8).

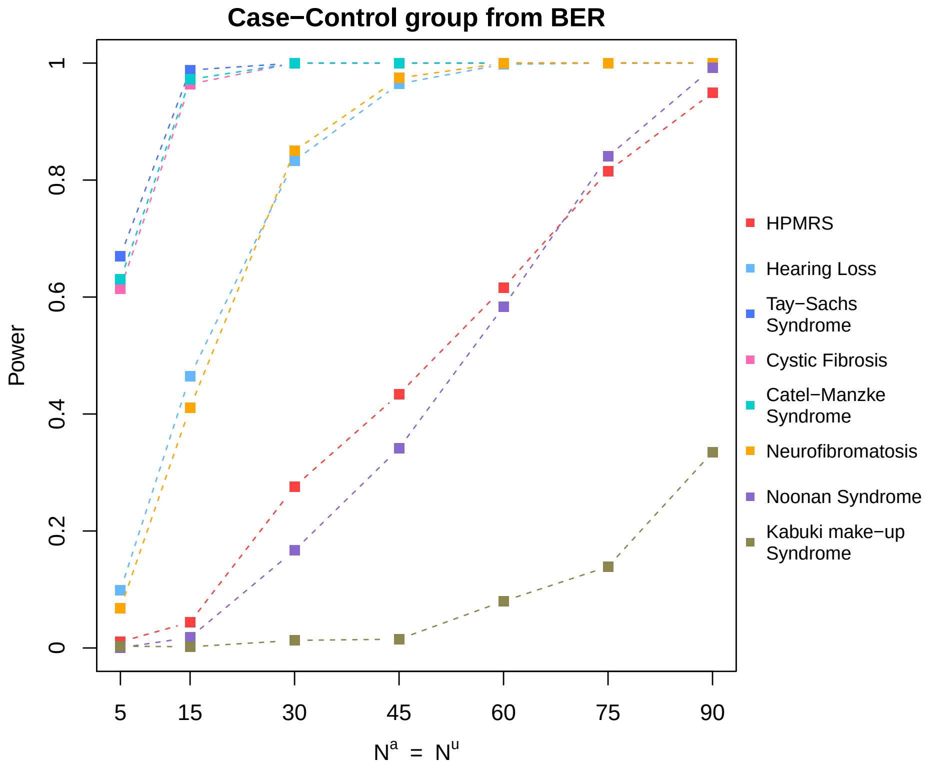


Figure 4.8: **Statistical power on different disorders.** Eight disorders are simulated with increasing group size. They consist of five single gene disorders (Tay-Sachs Syndrome, Cystic Fibrosis, Catel-Manzke Syndrome, Neurofibromatosis, Kabuki make-up Syndrome) and three heterogeneous disorders (HPMRS, autosomal recessive hearing loss, Noonan Syndrome). Five of them are recessive disorders (Tay-Sachs Syndrome, Cystic Fibrosis, Catel-Manzke Syndrome, HPMRS, autosomal recessive hearing loss) and the other three are dominant (Neurofibromatosis, Noonan Syndrome and Kabuki make-up Syndrome). N^a is the number of individuals in case group. N^u is the individuals in control group. The power of the RVAS is computed with permutation test. The significance cutoff is 0.05.

4.7 Top-ranked rate with different selection schemes

For rare disorders, it is not feasible to collect large cohorts, but it is hard to observe a very significant association with small cohorts if using multiple testing corrections. Permutation test can compensate for this shortcoming, but it is computationally intensive. Ideally, the p-values from the statistical tests at the causal locus should be smaller than any tests at neutral loci. Therefore, I used the rank of p-values as an alternative way to estimate the performance of the association test. I used the term "disease-causing gene is top ranked" in figures, which meant the disease causal gene had the lowest p-value.

Figure 4.9 showed that the probability of ranking the causal gene at the top increased with growing case group size for both BER and 1KGP subgroup which were the closest neighbors to BER samples in W^2 metric (see Figure 4.1). RVAS with any similarity-matched controls performed better than one with random controls in BER while only one with W^1 controls performed better in 1KGP data. I chose W^2 metric to evaluate the similarity between BER and 1KGP data, because data quality had tiny influence on W^2 metric (see Figure 4.4). The probability of ranking disease-causing genes at the top in BER was lower than in 1KGP, as the data quality of these two data had a large difference (Figure 4.1).

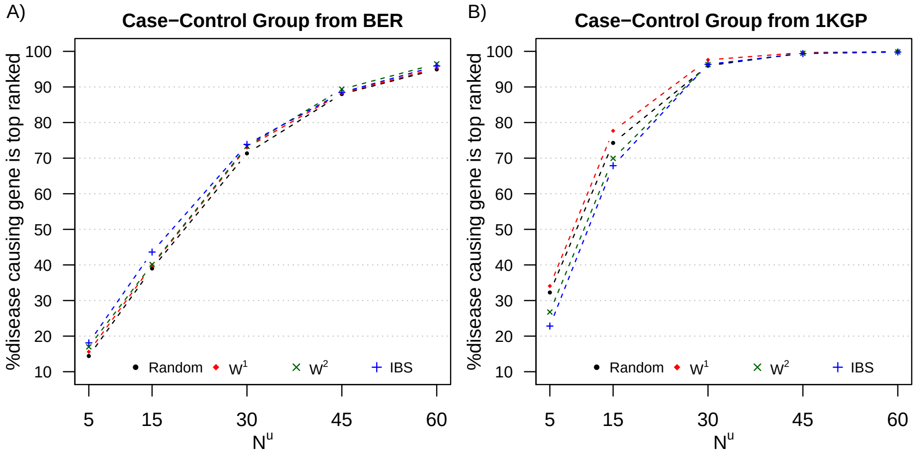


Figure 4.9: **Top-ranked rate of disease gene identification in 1KGP and BER.** The case and control groups have the same group size. All tests are calculated from CATT. The simulated disorder in this figure is HPMRS. The group sizes are increased from 5 to 60. (A) All case and control samples are selected from BER. (B) All cases and controls are selected from the subgroup of 1KGP. The subgroup of 1KGP is composed of samples which are the closest neighbors of BER individuals.

4.8 False positives genes in two cohorts

I have shown the top-ranked rate and power of detecting disease-associated genes, however, these two values are not high for small group size. Thus, I studied the false positive genes occurred in BER and 1KGP data. The false positive genes are the genes which had the lowest p-value in a simulation, but not disease-associated genes. In order to investigate what kind of genes were false positive genes, I collected the frequency of each gene to be false positive. Figure 4.10 showed the most ten frequent false positives observed in BER and 1KGP cohorts. For S_{W^2}, a subgroup of 1KGP cohorts (see Figure 4.1), these top 10 false positives were: *GPR98, LAMA5, MUC17, DNAH3, PLEC, SYNE2, AHNAK, FLG, OTOF, ABCA13.* For BER cohorts, these top 10 false positives were: *FLG, MUC17, SYNE2, AHNAK2, CUBN, MUC6, PLEC, HRNR, SYNE1, ABCA13.* Five genes were very frequent false positives in both cohorts. 12 false positive genes were observed in both cohorts. Gene *CUBN* and gene *HRNR* were frequent for BER while they were never observed in 1KGP. In contrast, gene *OTOF* was frequent in 1KGP while it was never observed in BER.

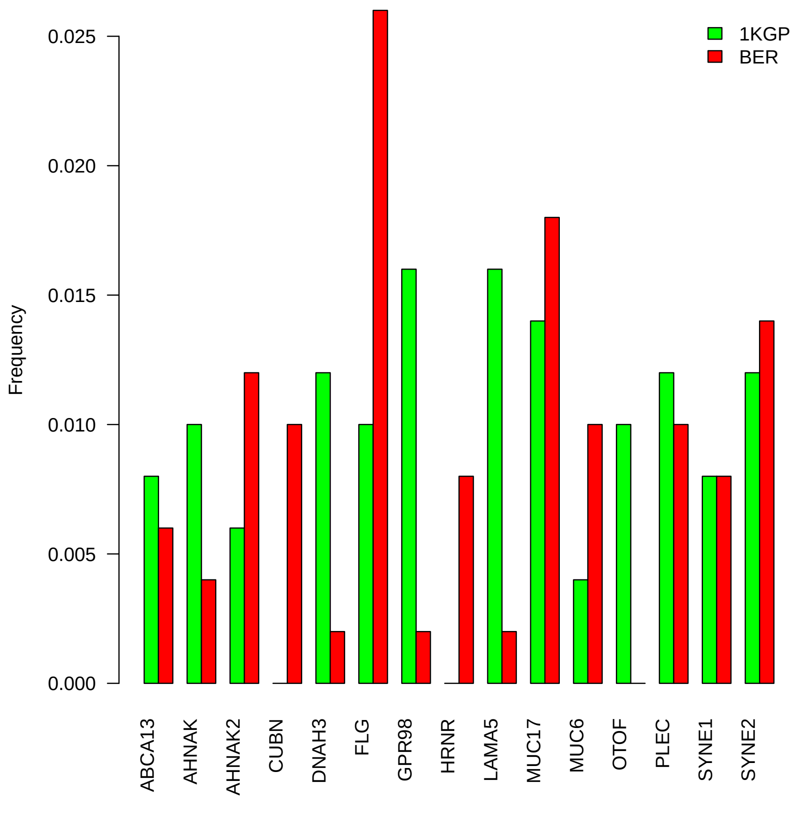


Figure 4.10: **The most frequent false positive genes in 1KGP and BER.** The ten most frequent false positives occurred in 1KGP data (*GPR98, LAMA5, MUC17, DNAH3, PLEC, SYNE2, AHNAK, FLG, OTOF and ABCA13*) and BER data (*FLG, MUC17, SYNE2, AHNAK2, CUBN, MUC6, PLEC, HRNR, SYNE1 and ABCA13*). The true positive genes are associated with HPMRS. The false positive genes are the gene which have the lowest p-value but not disease-associated genes. y-axis is the frequency of a false positive gene observed in 1000 simulations. Here 15 cases and 15 control are randomly selected from S_{W^2} of 1KGP and BER. Permutation tests are conducted here.

To classify these false positive genes, I learned the information of genes including: length, RVIS and number of variants in two cohorts. A negative RVIS score of a gene indicated purified selection while a positive score was likely to mean either the absence of purifying selection or the presence of balanced or positive selection [168]. Figure 4.11 showed nine of 14 frequent false positives had positive RVIS score. Gene *HRNR* did not have a RVIS score at present version (2016-03-12, http://genic-intolerance.org/). The length of all false positives was large. To further study the reason of these genes to be false positives, I collected the number of variants of a gene in two cohorts. As shown in Figure 4.11 B, all false positive genes had a large number of variants in each cohort. Due to the small number of variants in BER, gene *OTOF* was frequently observed in 1KGP but not in BER. It was true for gene *CUBN* and *HRNR* as well. Genes with large RVIS, such as *MUC16*, were not frequently observed due to few variants in two cohorts after filtering. Genes with a large number of variants in any cohorts, such as gene *PKD1* in 1KGP and gene *OR4A16* in BER, were observed false positives in the simulations, but not the ten most frequent ones.

Therefore, the large genes and the tolerant genes were more likely to be false positive findings in RVAS. It was also dependent on different data sets. Due to sequencing center, filter definition or the population components, the number of the variants in each gene varied among data sets.

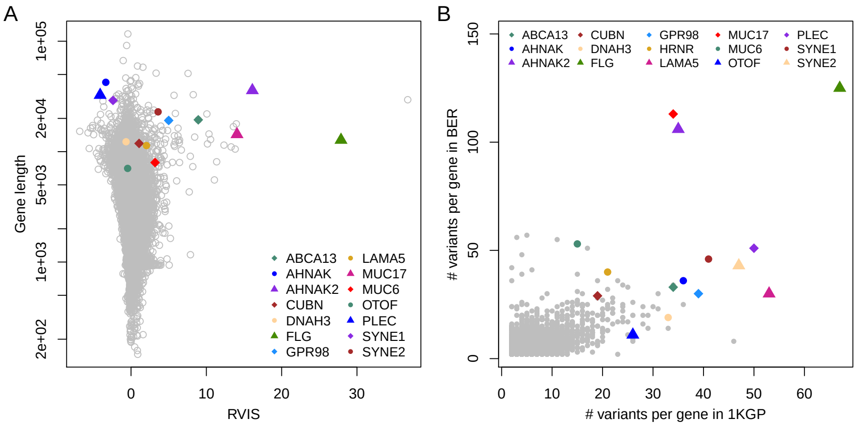


Figure 4.11: **The classifiers of false positive genes** A) Genes are classified by gene length and RVIS. Gene length is the cumulative length of exons in a gene. RVIS is the data published in 2016-03-12 (http://genic-intolerance.org/). B) The number of rare functional variants per gene in BER and 1KGP data.

4.9 Effect of data quality in RVAS

To carefully study the effect of the data quality and population background on the performance of RVAS, I conducted simulations on BER and two 1KGP sub-populations, FIN and GBR. FIN and GBR were homogeneous populations while BER was heterogeneous in their nature. RVAS for GBR and FIN data had higher probability of ranking the disease-associated gene at the top than one for BER. Besides the difference in population background, the difference of data quality had an impact on the performance of BER too. GBR and FIN had a larger fraction of the target region with above 20 reads compared to BER (Figure 4.12 A). I therefore tested the performance of RVAS with variable quality in the same population.

For this purpose, I only kept the samples which were at least 80% of the target region had more than 20 reads (*_H). The samples with higher data quality, such as GBR_H, FIN_H and BER_H, could increase the probability of ranking pathogenic genes. If comparing the performance between GBR and FIN, FIN had a higher probability of ranking the disease gene at the top than GBR, because FIN was a more homogeneous population than GBR [11]. Therefore, the high data quality and lesser population stratification could improve the performance in RVAS.

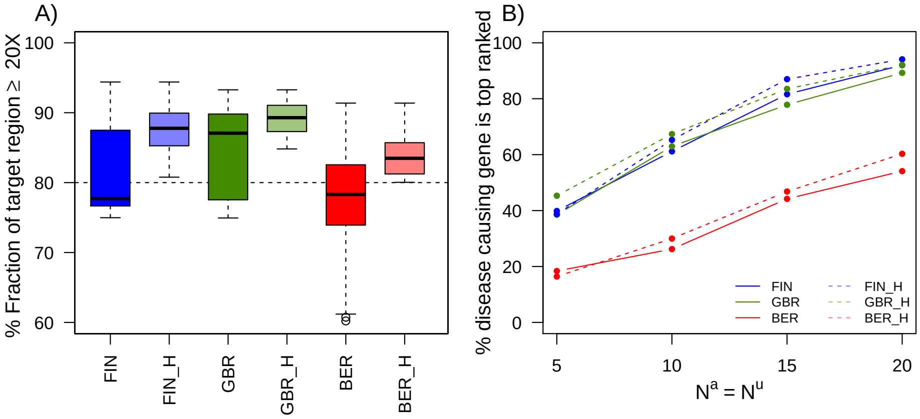


Figure 4.12: **Influence of population substructure and data quality on disease gene identification.** (A) The fraction of the target region in a sample with coverage above 20 reads was used to estimate the data quality. The higher the fraction a sample had, the lower the expected false positive and false negative genotyping error rates. FIN and GBR were all unrelated 1KGP individuals from these two populations. BER are the unrelated in-house samples. "_H" were the samples which had at least 80% target region with ≥ 20 reads in these three populations. (B) The probability of ranking the disease-causing gene at the top increased with an expanding case group. N^a was the number of individuals in case group. N^u was the individuals in control group. Here controls were the closest neighbors of cases with similarity metric W^1. The simulated disorder was HPMRS. BER showed the worst performance. For all three populations, the performance improved with higher data quality.

4.10 Effect of extended control group in RVAS

To achieve a significant true association or rank the disease gene at the top, high data quality and matched population stratification are required. Are there any other ways to further improve the performance of RVAS? While the size of the case group with a rare disorder is small due to the rareness of the disease in Mendelian disorders, there is theoretically no size limitation for the control group.

Expanded control cohorts was suggested to increase power in CVAS. When stratification was present, larger controls were preferred to decrease the chance of matching errors [77]. The optimal ratio of case:control was 1:4 proposed by Zondervan et al. [204]. The exact ratio of diminishing return may vary according to disease risk and allele frequency. If additional genotypes were effectively free, power can be maximized by including as many controls as possible. [140].

I therefore analyzed the performance of RVAS when gradually increasing the control group size but keeping case group fixed. I made simulations on BER data and the subgroup of 1KGP S_{W^2} (see Figure 4.1). I randomly selected a small number of cases (5 or 15) from a pool, then I chose controls at random or in a similarity-matched way. I expanded the control group from as large as case group to all available samples in each pool.

As shown in Figure 4.13, the probability of ranking the disease gene at the top increased as the control group size increased, which was consistent with the findings in CVAS. However, the optimal was neither four nor infinite. In general, the optimal control group size for five cases was 60 for BER and 120 for the subgroup of 1KGP (Figure 4.13 A, B); When case group size was fixed at 15, the optimum moved to 90 for BER and 120 for 1KGP (Figure 4.13 C, D). Compared to random selection, the similarity-matched selections improved the performance of RVAS prominently when the case group was small (size 5). Around 10% higher probability at the optimal points was gained by matching methods for both pools. Compared to the performance of the small control group, the optimum controls increased the probability of ranking disease gene at the top by $20\% \sim 40\%$.

RVAS with similarity-matched controls based on W^1 metric obtained the highest probability of ranking disease-causing genes top for BER data, while the performance with any similarity-matched controls was the same for 1KGP data. The reason is that BER cohorts have more heterogeneous data quality than that of 1KGP data. W^1 metric is the most sensitive to data quality, as it gives high weights to rare variants. It could match the data quality among samples, which could improve the performance of RVAS. From this analysis, we concluded that the large control group could increase the performance in RVAS. However, the size of the optimal control group was highly dependent on the characteristic of case group such as the group size, the patients' individual ethnicities and the carried disease, the characteristic of the pool such as the population structure and the data quality. In contrast, it was independent of the similarity metrics. If there was explicit population stratification, the optimum control was crucial for RVAS. If there was no population stratification, it may be beneficial to include as many controls as possible, as the finding in CVAS [140]. However, this did not necessarily apply to ultra-rare variant disorders, as the patients may be from different population across the world. A recent example was the patients with Catel-Manzke syndrome [81]. Three patients came from northern Germany, one patient was of British descent and the fifth was from Cameroon. Thus, controls had to be chosen carefully with consideration for the population compositions in cases and candidates controls.

Moreover, the size of the case group affected the optimal number in controls too. This may be due to the alteration of population composition in the large case group.

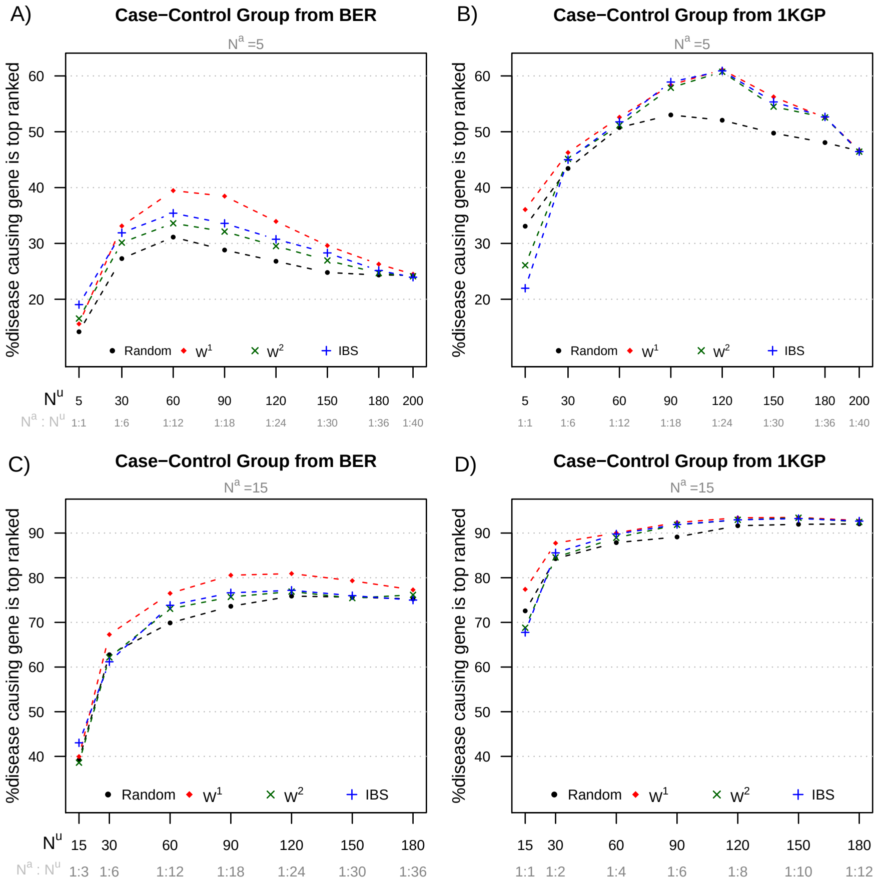


Figure 4.13: **Performance of RVAS with expanding control group.** The simulations are performed on exome data from BER cohort (A, C) and the subgroup of 1KGP, S_{W^2}, (B, D). The size of the case group is kept fixed at 5 (A, B) or 15 (C, D) individuals with pathogenic mutations for HPMRS. The probability of ranking disease-causing genes top increases if including more controls.

4.11 Comparision of statistical tests in RVAS

The basic idea of burden tests is to collapse information for multiple genetic variants into a single genetic score. They are powerful when a large proportion of variants are causal and effects of these variants are in the same direction.

In this work, I applied two burden tests, an univariate test CATT and a multivariate test CMC test for eight rare disorders (Figure 4.14). In this simulation, I randomly chose five cases from 1KGP, then chose another 100 controls from remaining 1KGP with different selection strategies (W^1, W^2, IBS or random). Neurofibromatosis and Kabuki make-up syndrome were caused by highly variable genes *NF1* and *KMT2D*. Due to many non-pathogenic mutations in these genes, the probability of detecting such genes was low. For instance, the probability of detecting *KMT2D* was low in all tests. It was even harder to be detected than gene *PTPN11* for Noonan syndrome. For these highly variable genes, CMC performed better than CATT, such as *NF1* and *KMT2D*. For genes with high sequence conservation, CATT performed better than CMC. This observation coincided with the conclusion of Li et.al that the inclusion of non-causal variants in a genomic region had a smaller influence in multivariate tests than univariate tests [189]. The merit of the similarity-matched strategies was independent of the statistical tests. For all disorders, it was more likely to rank the disease gene at the top if the controls were chosen based on similarity metric W^1 that was the most sensitive to data quality.

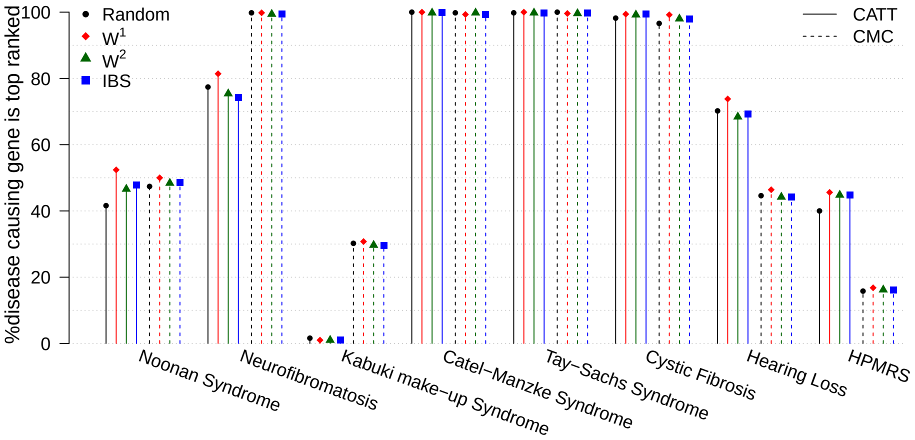


Figure 4.14: **The performance of disease gene identification for different disorders.** The case group is composed of five random individuals from 1KGP data. 100 controls are chosen at random or in a similarity-matched way from the remaining 1KGP data. The solid line shows the results from CATT tests and the dotted line shows the results from CMC test.

4.12 Effect of variant filter in RVAS

As the number of rare alleles at a single position is too numerous to run single marker tests, aggregation tests, which aggregate rare variants across a genomic region, are suggested for RVAS to increase the power. To avoid enriching noise in a region, a proper filter can aggregate the damaging alleles and ignore benign alleles. In this section, I studied the effect of variant filters on the performance of RVAS.

Here I took HPMRS as the simulated disorder. The variants were firstly filtered with allele frequency and protein function. Subsequently, I only kept the variants which were nonsynonymous and had minor allele frequency not greater than *0.001.* For comparison purposes, I further filtered the variants with phyloP score greater than 1 (Section 2.4.2). All cases and controls were from BER data. The case and control groups had the same group size. Control individuals were randomly selected from BER.

As shown in Figure 4.15, the stricter filter, nonsynonymous and conservative rare variants, increased the probability to identify the disease-associated genes in RVAS. Because the stricter filters removed more background mutations in the disease-associated gene when using the stricter filter.

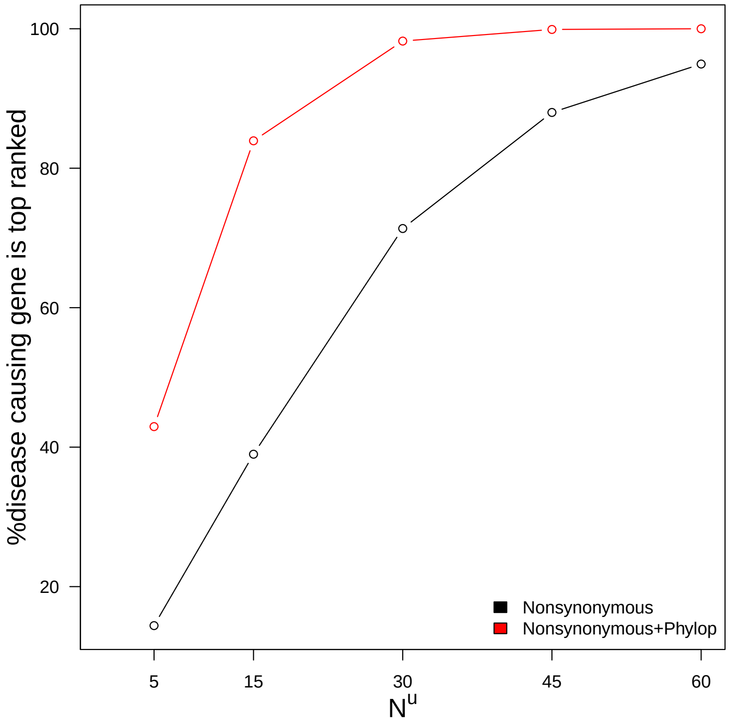


Figure 4.15: **The effect of Variants filter on RVAS**. Besides filtering variants by allele frequency, variants are further filtered with the protein function of variants or the conservative score. Black line: the filter includes allele frequency less than 0.1% and nonsense, missense and splicing. Red line: the filter includes allele frequency less than 0.1%, nonsynonymous mutations and phyloP score above one. N^u is the number of samples in control group.

4.13 Direct adjustment approach in RVAS

As introduced in the Chapter 1, the 'direct adjustment' method for population substructure cannot work in RVAS. To verify the conclusion, I applied EMMAX [64] to account for the substructure induced by rare variants (Figure 4.16). EMMAX has already been implemented in EPACTS package [199] which offered several statistical tests with correction methods. I chose VT to check whether EMMAX worked on RVAS, in this simulation, as VT had already been implemented with and without EMMAX model in the EPACTS package.

To make sure that the population substructure existed between case and control groups, I chose cases from GBR and controls from other 1KGP populations except GBR. The control individuals were selected at a random or in a similarity-matched way with W^1 metric. Figure 4.16 showed that the matching strategy improved RVAS before/after correction. However, EMMAX cannot improve the performance of RVAS, the performance with EMMAX was worse than one without correction, which was consistent with the study of Mathieson et al. [73].

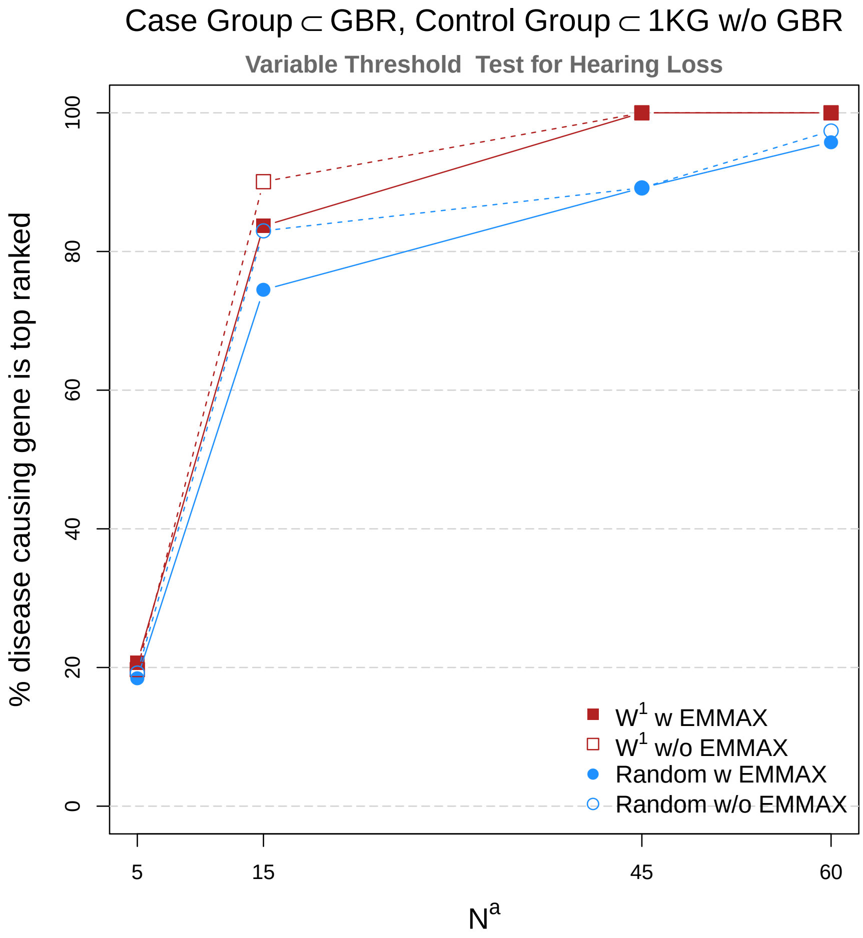


Figure 4.16: EMMAX performed in RVAS. EMMAX was used to account for the population stratification due to rare variants. Cases were chosen from GBR of 1KGP cohorts and controls were from other populations in 1KGP data. Metric W^1 was used to calculate the similarity score among individuals.

4.14 Disease gene identification in real studies

I tested the performance of RVAS in the disease gene identification of three monogenic disorders which were recently resolved, Kabuki make-up syndrome

[134], HPMRS [15] and Catel-Manzke syndrome [81]. There were ten unrelated affected individuals with Kabuki make-up syndrome, 13 samples of HPMRS cohorts and seven unrelated cohorts for Catel-Manzke syndrome. All cohorts were resolved with intersection filtering, by identifying the intersection genes among most of the unrelated and affected individuals [25, 209].

The example of studying of Kabuki make-up syndrome showed the limitations of the conventional intersection approach. Ten unrelated individuals were studied, seven of European ancestry, two of Hispanic ancestry and one of mixed European and Haitian ancestry. Kabuki make-up syndrome was a dominant disorder. Therefore Ng et al. firstly considered the gene for which all cases had at least one previously unidentified nonsynonymous variant, splice acceptor and donor site mutation or coding indel variant on it. With this filter, only *MUC16* was shared across ten exomes, which was highly likely to be a false positive gene due to its extremely large size. Then they conducted a less stringent analysis by looking for candidate genes shared among subsets of affected individuals. Several groups of candidate genes were obtained. Finally, they prioritized the candidate genes with genotypic and phenotypic stratification. They found only nonsense mutations in *KMT2D*/*MLL2* shared by four highest-ranked cases and found another three cases with loss-of-function mutations in this gene.

As RVAS was more straightforward for disease gene identification, I simulated the Kabuki make-up cohorts by subsequently increasing the number of cases with pathogenic variants from one to ten. I randomly chose ten cases and forty similarity matched controls from 1KGP. Further, I randomly spiked the pathogenic variants into the case group. Finally, I tested the relationship between the number of cases with pathogenic mutations and the probability of detecting the disease-caused gene at the top. RVAS ranked the disease gene at the top position in almost 100% of the instances when at least six out of ten individuals had pathogenic nonsynonymous mutations (Figure 4.17 A, B).

For Catel-Manzke syndrome and HPMRS, the identification of disease-associated gene was highly effective even if the number of samples with pathogenic mutations was smaller than that of the initial study. Especially the disease gene *TGDS* in Catel-Manzke syndrome had such a low variability that it can readily be identified with as little as four affected samples (Figure 4.17 C, D). As HPMRS was a heterogeneous disorder, patients may be affected due to mutations in different genes, thus more cases were needed to detect the disease gene for HPMRS (Figure 4.17 E, F).

Spurious associations often occurred for highly variable genes, such as genes from the mucin family or genes that show a higher rate of calling artifacts, such as the pseudogene *KRT1* (Figure 4.17 D). The false positive error resulting from such genes can also be reduced by using a similarity-matched setup of the control group.

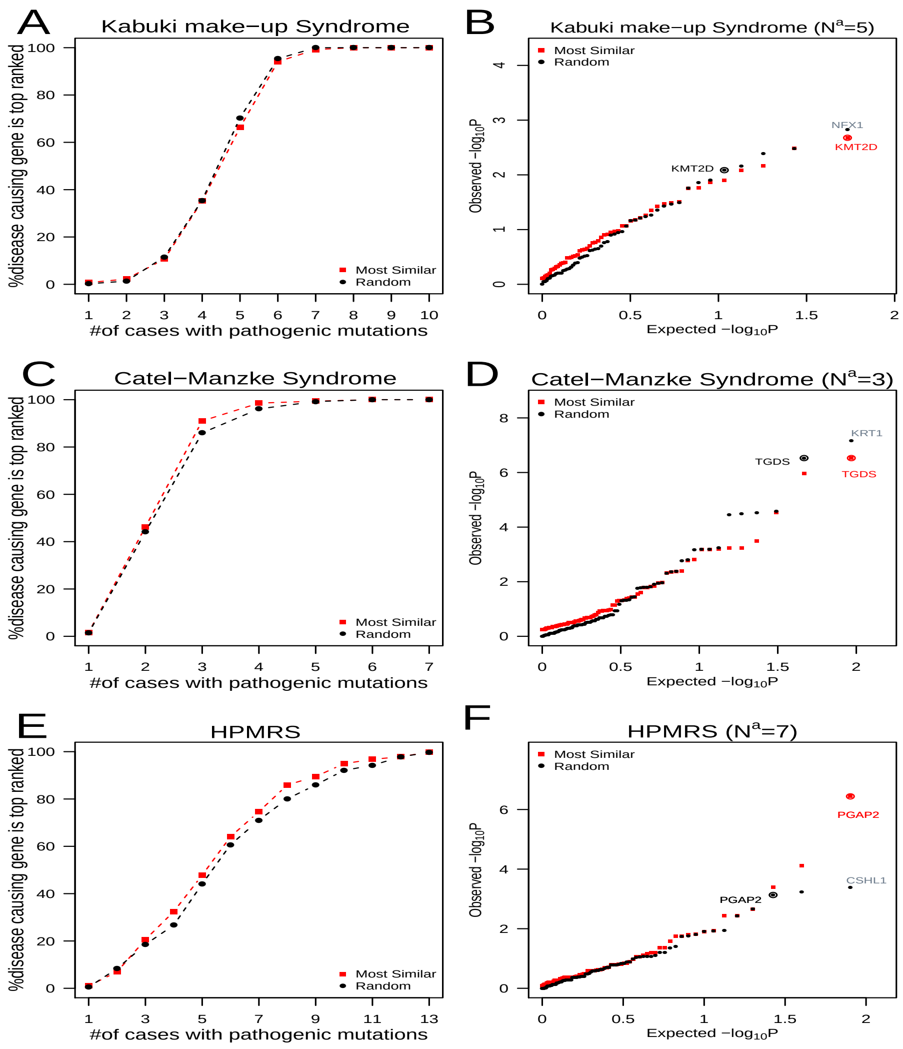

A
Kabuki make−up Syndrome
%disease causing gene is top ranked
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
#of cases with pathogenic mutations
Most Similar
Random
B
Kabuki make−up Syndrome (N^a=5)
Most Similar
Random
Observed $-\log_{10}P$
0
1
2
3
4
0
0.5
1
1.5
2
Expected $-\log_{10}P$
NFX1
KMT2D
KMT2D
C
Catel−Manzke Syndrome
%disease causing gene is top ranked
0
20
40
60
80
100
1
2
3
4
5
6
7
#of cases with pathogenic mutations
Most Similar
Random
D
Catel−Manzke Syndrome (N^a=3)
Most Similar
Random
Observed $-\log_{10}P$
0
2
4
6
8
0
0.5
1
1.5
2
Expected $-\log_{10}P$
KRT1
TGDS
TGDS
E
HPMRS
%disease causing gene is top ranked
0
20
40
60
80
100
1
3
5
7
9
11
13
#of cases with pathogenic mutations
Most Similar
Random
F
HPMRS (N^a=7)
Most Similar
Random
Observed $-\log_{10}P$
0
2
4
6
0
0.5
1
1.5
2
Expected $-\log_{10}P$
PGAP2
CSHL1
PGAP2

Figure 4.17: **RVAS for three resolved monogenic disorders.** The pathogenic mutations of Kabuki make-up syndrome, HPMRS and Catel-Manzke syndrome have been identified in new disease genes by intersections of variant candidates in case groups of unrelated, affected individuals comprising 10, 7 or 13 cases. Via RVAS approach with 40 controls, the probability of identifying the disease gene in such cohorts was still considerable even when the number of cases with pathogenic mutations in these cohorts was reduced markedly (A,C, E). Additionally, a selection of similarity-matched controls may also help to reduce spurious associations effectively: the QQ plots (B,D,F) showed the instances of the RVAS simulations where 5, 3 and 7 individuals had pathogenic mutations in the disease genes *KMT2D*, *TGDS* and *PGAP2*. It also showed that the disease gene got the lowest p-value only if similarity-matched controls were used.

4.15 RVAS on Catel-Manzke syndrome

From the simulated data in Section 4.14, the disease gene *TGDS* could be identified in the Catel-Manzke syndrome cohorts with four cases. The homozygous or compound heterozygous mutations in gene *TGDS* caused Catel-Manzke syndrome [81]. In our clinic, we collected seven families from all over the world (Figure 4.18). Family 1 was from Cameroon, patient 2 was of mixed British and South American descent, patients 3, 4 and 5 were of German descent, family 6 were Dutch and family 7 was from northern France.

Individuals in family 1 (proband and parents), family 2 (proband and parents), family 4 (proband and mother), family 5 (proband and children) and the affected individual of family 7 were subjected to exome sequencing.

The other four samples, the parents and the affected child from family 1 and the affected child from family 7, were sequenced with Illumina HiSeq system with paired-end 2 × 100 bp protocol. Sequence reads were mapped to human genome reference hg19 using Novoalign [210]. Single nucleotide variant (SNV) and short Insertion and deletion (INDEL) were also called by GATK toolkit.

All variants were annotated at the functional level with Jannovar [155].

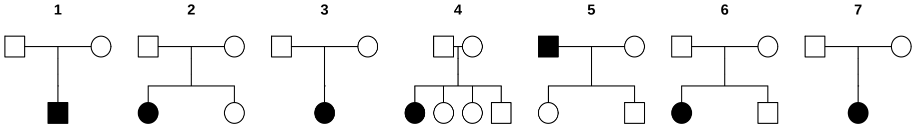


Figure 4.18: **Pedigree structures in Catel-Manzke cohorts.** Family 1 was from Cameroon, patient 2 was of mixed British and South American descent, patients 3, 4 and 5 were of German descent, family 6 was Dutch and family 7 was from northern France.

As no candidate gene was detected via the separate analysis of the families, all affected individuals were collected to be a case group. In this pedigree (Figure 4.18), the recessive inheritance model and the *de novo* dominant inheritance model were considered. Under the hypothesis of *de novo* dominant mode, no candidate gene was reliable. For the hypothesis of recessive inheritance mode, only the singleton homozygous variants, and the heterozygous variants with lower than 0.01 frequency were kept in cases, where allele frequency was based on large population studies, such as 1KGP, ExAC `http://exac.broadinstitute.org/` or ESP (`http://evs.gs.washington.edu/EVS/`). The analysis of the autosomal-recessive mode of inheritance yielded three candidate genes: *MUC4*, *MUC6* and *TGDS*. Mucin genes are highly variable, thus gene *TGDS* was the most likely candidate. Its pathogenesis was established by bioinformatic predication tools and further biological function analysis. Based on the intersection strategy, *TGDS* gene was identified.

I also ran RVAS on this cohort. Due to the limited components of in-house data, I took 1KGP and BER as control group. Metric W^1 is sensitive to data quality, it matched data quality rather than population background when the quality among samples varies. Thus I used W^2 to estimate the similarity between individuals and chose their first ten nearest neighbors as a control group from

BER and 1KGP. In the similarity-matched controls, some controls were from BER and some from 1KGP. The cases were close to their controls as shown in the Multidimensional scaling (MDS) plot (only the five nearest neighbors were plotted, Figure 4.19).

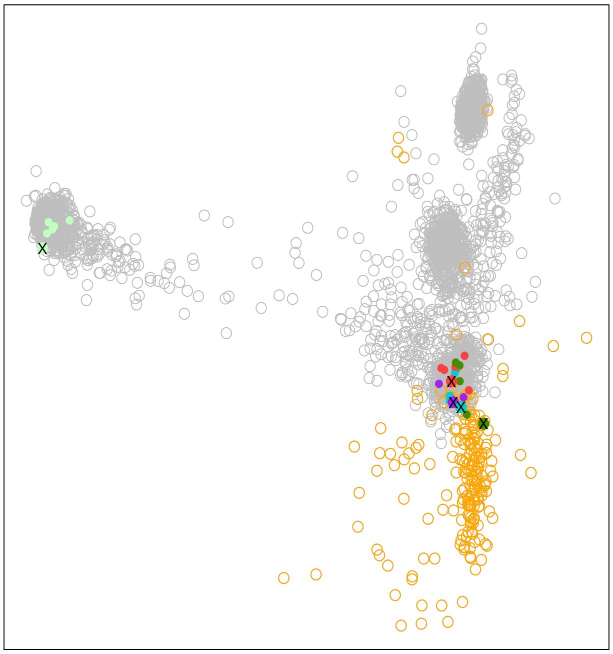


Figure 4.19: The MDS plot illustrated the similarity between cases and all 1KGP and BER data. The similarity matrix was calculated based on metric W^2. The colored dots with ’X’ are cases. The colored dots without ’X’ are the selected closest controls. The dots of the same color indicate the first five closest controls for a patient (with ’X’). The orange circles are individuals from BER. The gray circles are the individuals from 1KGP. Some of the selected controls seem far from the patients; it may be due to the visualization angle.

I only tested the missense or nonsense variants and further filtered out the variants with MAF above 0.001. The variants appeared more than three times in our in-house data were also filtered out. The variants which occurred in case and control individuals simultaneously were filtered out further. I then aggregated variants in each gene and did CATT on these genes. Table 4.1 showed the genes ranked among the top ten. With similarity-matched controls, gene *TGDS* was ranked at the top and mucin genes were ranked at lower positions. With the random controls, gene *MUC*4 still distracted analysis. The balanced distribution of rare variants in the case and similarity-matched controls degraded the disturbed genes, especially the highly variable genes.

	Matched Controls		Random Controls	
Rank	Gene	p-value	Gene	p-value
1	TGDS	$2.32E^{-5}$	MUC4	$7.04E^{-8}$
2	ANKRD20A4	$1.08E^{-4}$	TGDS	$2.32E^{-5}$
3	ANKRD36B	$1.08E^{-4}$	ANKRD20A4	$1.08E^{-4}$
4	FRG2C	$1.06E^{-3}$	ANKRD36B	$1.08E^{-4}$
5	KIR3DL1	$1.06E^{-3}$	MUC16	$2.92E^{-4}$
6	POTED	$1.06E^{-3}$	FRG2C	$1.06E^{-3}$
7	SPATA20	$2.38E^{-3}$	POTED	$1.06E^{-3}$
8	DNAH5	$3.41E^{-3}$	MUC6	$1.41E^{-3}$
9	MUC4	$3.41E^{-3}$	SPATA20	$2.38E^{-3}$
10	PRR21	$3.41E^{-3}$	DNAH5	$3.41E^{-3}$

Table 4.1: **Genes of the ten lowest p-values**. The rank of p-values was calculated in CATT with similarity-matched controls or random controls. The similarity among individuals was obtained based on metric W^2. The 'matching strategy' improved the ranks. Gene *MUC*4 ranked at the top with random controls while it ranked lower with the similarity-matched controls.

In this case, gene *TGDS* ranked at the top for several reasons. Because *TGDS* was a gene with high sequence conservation, it had few rare mutations in the control group. In addition, Catel-Manzke was a homogeneous and recessive disorder, all patients were affected by one or two pathogenic mutations on *TGDS*.

4.16 Genomic matchmaking database

Since many Mendelian disorders have not been elucidated, it is difficult for a clinical center to obtain a sufficient number of patients. To circumvent this difficulty, Genomic Matchmaking databases (GMD) was proposed, which allowed participants to submit genomic and phenotypic data in order to identify unreported disease-associated genes by matching them with other comparable cases. At least 3000 such genes are expected in Mendelian disorders [211]. However, what is the sufficient data needed to ensure two or more individuals shared the same disorders and caused by the same disease-caused gene? This problem resembled the birthday paradox. In that scenario, the probability p of a matched pair in a group of r individuals is:

$$p \cong 1 - e^{\frac{-r(r-1)}{2*365}} \tag{4.1}$$

If 23 individuals were investigated, there was a 50% chance of finding a matched pair. However, if there was an expected 50% chance of three individuals having the same birthday, at least 88 individuals belonged to this group [212]. In the similar calculation for 3000 disease-associate genes, the expected probability of some individuals sharing the same disease-associated gene varied with the group size (Figure 4.20 A). Moreover, there were more factors that affected the efficiency of identifying the disease causal gene in Mendelian diseases with RVAS, such as the detection rate of mutations, d, the inheritance model of the disorder, the genetic heterogeneity $\mathbf{h}$ and the neutral rare variants λ.

This study showed that the lower proportion of rare neutral variants in the control group and the higher detection rate could increase the chance of identifying the disease-associated genes (Figure 4.20 B, C). From the standpoint of inheritance models, the disease-associated genes were more difficult to detect if more autosomal dominant diseases and heterogeneous disorders were involved in the database (Figure 4.20 C and D).

Supposed that half of unsolved Mendelian diseases were dominant and half were recessive, then a third of these diseases had homogeneity and another third had heterogeneity with ten genes and the remaining third had heterogeneity with 30 genes. Furthermore, a 70% detection rate and a 0.02 rare neutral rate were assumed. Under these assumptions, GMD would require approximately 80,000 patients in order to identify all disease-associated genes [193].

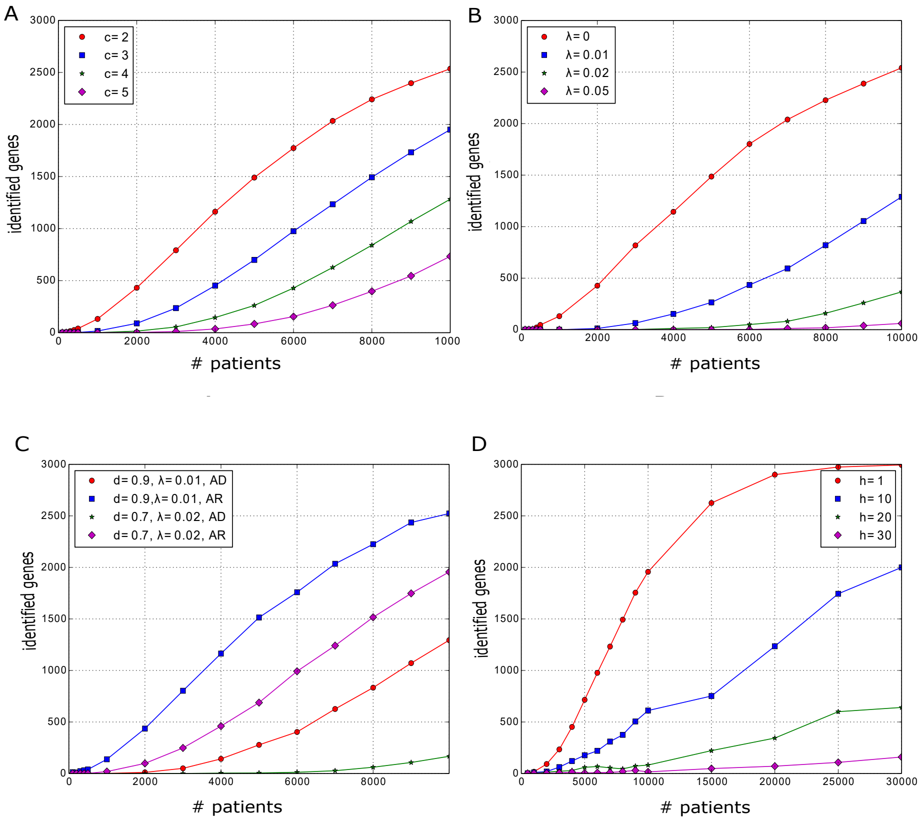


Figure 4.20: **The impact factors on GMD.** Multinomial distribution was used for simulating the relationship between the number of patients in GMD and the identified disease genes. A) The influence of the size of the dataset. Large data in the database increased the number of patients who shared the same disease gene (c) and further increased the chance of solving the disease. B) The influence from the background mutations. The rarer the neutral mutations λ which appeared in the control group, the greater the difficulty it had to identify the disease genes. C) The influence of the inheritance model of disorders. The disease associate genes of autosomal dominant disorders were more difficult to detect than those of the recessive disorders. The lower rate of mutation detection also decreased the chance of identifying the disease genes. D) The influence of the heterogeneity of the disease-linked genes. Genetic heterogeneity affected the chance of significantly identifying the disease-associated genes. The greater the heterogeneity of the disorders, the harder it was to detect the disease associated genes.

Chapter 5

Discussion

GWAS have successfully identified hundreds of thousands of SNPs that contribute to complex clinical conditions and phenotypic traits [213]. However, the associations due to these SNPs can only explain a certain fraction of overall heritability [33, 32]. Therefore the assumption that rare variants play a significant role in explaining this 'missing heritability' came up.

Due to small cohorts and the intrinsic relationship inside the cohorts, rare disease genes can be traditionally resolved by linkage analysis and the intersection filter among patients. However, researchers may be distracted by the artifacts from large genes or the low complexity region, such as MUC gene or pseudogenes, because these genes are likely to pass the intersection filters. Besides the traditional approaches, RVAS compares the patient cohorts to the healthy cohorts. It can decrease false positive genes and prioritize the disease-associated genes.

It is well known that population stratification can lead to spurious associations in GWAS. In CVAS, many methods have developed for accounting for the stratification. Generally, these methods can be divided into two classifiers. The first cluster is the 'matching strategy' at the design stage, which corrects for population stratification by involving tight matching of cases and controls. The other cluster is 'direct adjustment' after the design, which adjusts for the confounding by using the ancestry components as covariates of association tests. A comparison of the performance of these two approaches in CVAS reveals that: when population stratification is small, the 'matching strategy' approaches perform comparably to the 'direct adjustment' approaches. However, when the stratification becomes large, the 'matching strategy' approaches perform stably while the 'direct adjustment' approaches perform variably depending on the algorithms. For instance, genomic control [59] became too conservative, but PCA approach [53] performed still well [77].

Comparing to CVAS, population stratification is more pronounced in RVAS. The reasons are three-fold. Firstly, as many rare variants typically evolve recently, it is more population-specific [71]. Secondly, as the demand for GMD increases, the patients with same disorders are dispersed all over the world. This further exacerbates the stratification. Finally, RVAS commonly uses burden tests, which aggregate information across multiple sites. RVAS has to tackle the population stratification in both individual allele frequencies and the total quantities of rare variants [68]. All in all, to solve the problem of population stratification in RVAS becomes a necessary and urgent task.

Unfortunately, the existing methods cannot correct the confounding of the stratification due to rare variants. Therefore, I set up this study for searching the strategy to account for the stratification due to rare variants.

In general, similar genetic backgrounds in case and control group can even up the stratification, which is also the baseline of the 'matching strategy'. Therefore, in this work, I worked on 'matching strategy' to select the genetically similar individuals to construct control group in the design stage of RVAS.

To achieve this goal, three similarity metrics (W^1, W^2 and IBS) were studied. From the perspective of evolution, common variants reflect older evolutionary history and contribute more to the population background. Whereas rare variants are evolutionarily recent and have significant impact on human phenotypes and disease susceptibility. Thus, common variants and rare variants are weighted differently in these three metrics. Rare variants are given higher weights in metric W^1 and common variants get higher weights in metric W^2.

Basic IBS metric gives the same weight to all variants.

Via simulations, I found that the 'matching strategy' with all three metrics can considerably account for the confounding of the population stratification. RVAS with the matching strategy improves the statistical power, reduced the FWER and also increased the probability of ranking the disease-associated genes at the top.

Despite the consistent improvement of RVAS with the matching strategy, different performance among similarity metrics was observed. Metric W^1 can separate the populations by considering the population background and the data quality, whereas W^2 and basic IBS only consider the population background. When the data quality among samples is comparable, W^1 separates the cluster more clearly comparing to metric W^2 and basic IBS. In this scenario, the matching strategy with metric W^1 improves the performance of RVAS more than the other two metrics.

In the CVAS study, a larger control group has been proposed to increase the power of the association tests. The optimal control:case ratio is suggested differently in several study [205, 140, 77]. In this work, I found that there was also an optimal ratio of case:control in RVAS. More surprisingly, RVAS with the 'matching strategy' plus the optimal ratio could maximize improvement. In addition, the similarity metric for the 'matching strategy' affects the performance of RVAS, but it cannot affect the optimal ratio. The optimal ratio is dependent on the population structure and data quality in all samples, the inheritance mode of disorders, the heterogeneity of disease-linked genes and the tolerance of background variants of the statistical test. Moreover, the fraction of components of ancestry correlated with disease risk is also likely to have effect on the optimal ratio [77]. Guan et.al proposed the minimize a cost algorithm to find the optimal control group in CVAS with inputted group size [77]. This algorithm needs to be tested in RVAS. As the optimal ratio is study-specific in RVAS, individual efforts are also needed for each study design separately.

In the application of case studies, I showed the benefits due to a larger control group and the 'matching strategy'. In the real-case study of Catel-Manzke syndrome, RVAS with the 'matching strategy' and with a larger control group ranked the disease-caused gene *TGDS* at the top while suspending the frequent false positive genes out of the hit list. As another example, RVAS was able to rank the disease-caused gene *KMT2D* at the top for Kabuki make-up syndrome for a cohort of as small as six cases. Compared to the traditional linkage analysis and the intersection filter among patients, RVAS resolves the disorder with smaller cohorts, and accelerates the progress for identifying disease-associated genes.

As the demand of GMD increases, the patients with the same disorders are dispersed all over the world. The population stratification between case and control group will become even worse. It further highlights the importance of the 'matching strategy'.

As many factors may affect the performance of RVAS, there is still considerable space to optimize these methods used in this work. In the following, I discuss these factors and the limitation of this work, and suggest possible improvements in further research of RVAS.

As seen in this work, when the data quality among individuals significantly varies, the influence of data quality in the similarity metric differed. Metric W^1 clustered samples by data quality rather than ethnicity while the other two metrics matched samples by ethnicity. However, none of the metrics can work properly for the 'matching strategy'. This emphasizes the importance of good data quality, but it also calls for a matching strategy which can overcome the data quality difference to identify the right population background for further study.

CVAS study has shown that choosing a small informative set of genetic variants can estimate the genetic similarity as exact as involving all available markers. This could be achieved, for instance, by selecting markers based on the Hardy-Weinberg disequilibrium tests [77]. This idea also offers a hint for optimizing the calculation of similarity matrix in the current work and improving the similarity metric for the further research.

This study employed the individual-based 'matching strategy' to select the tight matching controls. An alternative approach in CVAS for the 'matching strategy' is cluster-based analysis, such as GEM [74] and spectralGEM [214]. In cluster-based approach, multiple ancestry components are summarized into a single scalar measure. Then the score is used to assign subjects to a small number of strata. Regarding CVAS study, it has been reported that these two approaches perform similarly. Whether the cluster-based approach will also work for RVAS demands further investigations.

The 'matching strategy' with the expanded control group could maximize the performance of RVAS when the number of patients is fixed. However, due to the extremely low prevalence of rare disorders, collecting sufficiently large cohorts of unsolved phenotype is a major challenge for a single clinical center. Community effort may solve the problem of small cohorts, that is, many geneticists contribute their cohorts to a large database GMD. Expanded GMD can increase the possibility to detect the disease-associated genes for the cohorts with similar phenotypes. However, it also increases the possibility that the matched cases may carry imprecise phenotypes, which will increase the potential for false positive associations as well [215]. Generally, genic tolerant genes or large genes are prone to be false positive genes. The 'matching strategy' can help to reduce the false discoveries, but more strategies are required to prioritize and interpret the interesting candidates. Some studies have developed scoring systems to indicate how likely a gene is genic intolerant based on known large datasets, such as RVIS based on ESP or pLI score based on ExAC [168, 216, 217]. These scores, especially pLI score, only perform well in identifying potential dominant disease genes. For recessive disorders, several millions of healthy individuals in a random mating population may be needed to detect a depletion of homozygous LoF mutations. Alternatively individuals from consanguineous marriages represent an alternative to detecting viable or lethal recessive gene via human knockouts [218]. This kind of research will help to interpret recessive disease genes.

Apart from the data quality and population components of samples, the number of individuals in case and control groups, the features of disorders and the similarity metrics used in the 'matching strategy', variants filter also have a strong influence on the performance of RVAS. The goal of variants filter in RVAS is to aggregate the damaging alleles and ignore benign ones. A filter-fixed approach is common in RVAS. For instance, the non-synonymous variants in the protein-coding region and variants with a frequency below a specific threshold T are commonly used. The profiles of rare variants varied considerably between different populations and different sequencing studies. Here again, lower data quality which consequently caused more genotyping errors and more diverse population backgrounds resulted in more singletons for in-house data. Therefore, including in-house samples for allele frequency calculation can reduce artifacts.

Besides allele frequency filter, one can also filter the variants with other criteria, such as phylogenic score and mutation function. The stricter filter may increase the probability of detecting the disease-linked genes in RVAS, which is especially true for the causal genes have high sequence conservation. However, it may exclude the true disease-causing mutations, for instance, the pathogenic variants of TAR syndrome consist of one rare variant and one polymorphism in most cases [219]. Apparently, thorough understanding and careful analysis of the disorders are required before choosing filters.

This work has focused on the association study for unrelated individuals. Beyond this method, the family-based association test (Family-based association test (FBAT)) is also widely used. Both designs have advantages and disadvantages. The disadvantage of the unrelated case-control study is that the significant association may be due to the population stratification. The family-based study designs are robust against population substructure, as the family members have similar population background. However, it takes much more time and money to gather the probands and their relatives in the family-based association studies. The association test of unrelated individuals performs worse than the family-based design if all trios data are available, whereas, the population-based association study is more efficient than family-based association study if limiting to the same expense [220, 221]. However, the power of population-based association tests highly depends on the number of patients, and the significant association is hard to be observed in small cohorts.

Considering the merits and limitations of population-based association tests and family-based association tests, a general framework unified both designs is proposed. The integrative approach builds a connection between population-based association tests and family-based association tests. Its test statistics includes the statistics from population-based tests and correction factor for considering the population structure and pefigree. This design improves the power and decreases type I error. [222, 223, 224].

In addition to the association tests for the population-based and family-based data, a method for association tests without the control group was proposed [216]. It particularly worked on estimating the enrichment of *de novo* mutations in genes [225, 216]. It firstly estimated the expected per-gene probabilities of *de novo* mutations for each mutation type (synonymous, missense, nonsense, splice sites) from the public data. Secondly, it evaluated whether the observed mutations in cases exceeded the expected number. Compared to RVAS, on one hand, this method focused on the *de novo* mutations, which had a much stricter filter than for rare variants. On the other hand, this approach calculated the association without control groups. It saved the cost of sequencing large control cohorts, and it was efficient. This method offered the ability to evaluate the rate of rare variation from learning large databases, such as ESP or ExAC. This method can work in individual genes where burden test would fail. To extend this method to rare variants on a broad scale needs further study.

This study mainly discussed RVAS in the coding regions, as the coding sequence are the main functional and medically relevant part in the genome. By contrast, the function of non-coding DNA cannot be deciphered only with the sequence. Fortunately, a growing number of non-coding transcripts in gene regulation and RNA processing have been confirmed, such as cis-regulatory elements: enhancers, silencers, promoters [226, 227]. Furthermore, many SNPs in non-coding regions are significantly associated with disease in GWAS [228, 229]. Therefore, extending RVAS to non-coding regions is an obvious next step. In order to make RVAS feasible in the non-coding region, several major challenges must be overcome.

The first challenge is how to filter the rare variants in the non-coding region. Because rare variants in non-coding regions are likely to have smaller effects and have an overwhelming number, the true signal will immerse in lots of false positive findings. Thus, RVAS may require a much larger number of samples to detect a comparable effect in the non-coding region compared to coding regions. It also highlights the need for advanced annotation tools for rare variants in whole genome to filter out as many disease-irrelevant variants as possible [230, 231, 232, 233].

To aggregate the variants in a gene or pathway is the most intuitive way in coding regions. However, non-coding regions have more complicated regulatory mechanism. There are many choices for aggregating variants in the whole genome, such as genomic physical locations like window size or biological function units like topological association domains (TADs) [234, 235]. A good understanding of the studied disorders will help to choose the aggregation unit. For example, for a disease of little previous research, we may test the aggregation unit from TAD regions to small interesting regions; For a disease with lots of previous research, one could restrict the investigated regions to the known regions related to the phenotypes. For instance, the interested HPO terms related to the disorders can be generated with Phenomizer and the interested regions (like TAD) can be further generated with known HPO terms [236, 237, 238, 239]. In the suspected regions, we could further divide the vast regions into small bins and test all possible combinations of bins.

All in all, RVAS is a straightforward and efficient method to prioritize the disease-caused genes in Mendelian disorders. Although there is still much work to do in future, there is no doubt that RVAS will make a significant contribution to the identification of disease genes. In January 2016, NIH reported their plan of genomic research for human disease. With increasing technical capabilities and theoretical know-how, the endeavor to comprehensively understand the genetic disease has just begun.

Appendices

Summary

It is well known that population substructure can lead to spurious associations in GWAS. Two strategies,'direct-adjustment' and 'matching strategy', have been developed to account for such population stratification in CVAS. However, the population stratification behaves differently in RVAS and CVAS. It results that the existing methods based on 'direct adjustment' strategy cannot work in RVAS. However, whether 'matching strategy' would work in RVAS is still unclear.

Therefore, in this work I studied the matching strategy at the design stage of RVAS. Three similarity metrics with different weighting schemes were set up for the matching strategy. I evaluated the performance of RVAS by power, FWER and top-ranked rate. In addition, I also studied the impact factors for RVAS performance, such as the data quality of samples, number of samples, the inheritance model of disorders and the heterogeneity of disease-caused genes. I also studied the existing problems in RVAS and also suggested the solutions, such as the bad quality samples and the small number of cohorts. Finally, I applied RVAS approach in the Catel-Manzke cohorts, RVAS identified the disease-associated gene *TGDS*.

Thus, RVAS is a comprehensive approach to prioritize the causal genes in Mendelian disorders. The 'matching strategy' for RVAS could account for the population stratification. RVAS with matching strategy could increase the statistical power and reduce the FWER.

Zusammenfassung

In genomweiten Assoziationsstudien, GWAS, können Unterschiede in der ethnischen Herkunft der Individuen in den Fall- und Kontrollgruppen zu Assoziationen führen, die nicht auf den eigentlich zu untersuchenden Phänotyp zurückzuführen sind. Diese Signale stellen damit unerwünschte Artefakte dar. Zur Vermeidung dieser fehlerhafter Assoziationen wurden Strategien entwickelt, die entweder eine Korrektur auf zuvor definierten Gruppen vornehmen, oder aber Kontrollen passend zu den betroffenen Individuen auswählen. Neuerdings sind aufgrund moderner Sequenziertechnologien auch Assoziationsstudien für seltene genetische Varianten, RVAS, möglich. Es zeigte sich jedoch, dass hierbei eine nachträgliche Korrektur nicht möglich ist, da seltene Varianten ein dafür ungeeignetes Verteilungsmuster aufweisen. In meiner Arbeit wurde untersucht, inwieweit eine Auswahl passender Kontrollen Fehlerraten reduzieren kann und welche Metriken zur Ähnlichkeitsberechnung geeignet sein könnten. Zur Auswahl der Kontrollen wurden unterschiedliche Distanzmetriken analysiert, die eine Gewichtung anhand von Allelfrequenzen vornehmen. Die Güte dieser Auswahlverfahren wurde anhand von simulierten Fall-Kontrollgruppen bewertet. Bei der Zusammensetzung der Fallgruppen wurde neben unterschiedlicher Herkunft der Individuen auch eine hohe Variabilität in der Datenqualität untersucht. Es zeigte sich, dass eine Ähnlichkeitsmetrik, die eine stärkere Gewichtung seltener Varianten vornimmt besonders gut geeignet ist, um fehlerhafte Assoziationen zu reduzieren. Bei einer kleinen Fallgruppengröe, wie sie für die meisten Studienkohorten Mendelscher Erkrankungen typisch sind, konnten die erwünschten Krankheitsgene leichter identifiziert werden, wenn es sich um rezessive Erkrankungen handelte. Eine hohe Heterogenität der Erkrankung und Variabilität der Zielgene erschwerte die Detektion. Mit einer Vergrößerung der Kontrollgruppe konnten Verbesserungen in der Detektionsrate erzielt werden. Die erarbeiteten Auswahlstrategien wurden schließlich angewendet, um eine Fallsammlung von Patienten mit Catel-Manzke Syndrom zu analysieren. In den betroffenen Individuen konnte eine signifikante Anreicherung seltener Mutationen im Gen *TGDS* identifiziert werden, die eine Auswirkung auf die Proteinstruktur haben. Die entwickelten Analyseverfahren können damit eingesetzt werden, um die Identifikation einer Anreicherung klinisch relevanter Mutationen in Patientenkollektien zu erleichtern.

Acknowledgement

First and foremost, I would thank my supervisor Dr. Peter Krawitz for instructing my research and for allowing me to grow as a research scientist slowly. I appreciate all his contributions of time, ideas, and patience. I also thank him to save me from lots of confused moments. It is my greatest luck to be his student.

I would like to thank Prof. Peter Nick Robinson for allowing me to study in the great group, tolerating my ordinary in the research, and giving me big supports on the research and thesis.

I would also like to thank Prof. Martin Vingron for giving me the chance to study in Free University Berlin and reviewing my thesis.

I deeply thank Dr. Lei Mao. She is not only my friend but also a second instructor. She helped me to understand the genetic and biology theory. She instructed and pushed me to write this thesis. Without her, I am not able to finish the thesis. She is such great person with many high-quality characters: happy, confident, optimistic, diligent ... I also thank her to give me energy when I felt tired. Her philosophy is always the best medicine for me.

I would like to thank Verena Heinrich for helping me on the research, accompanying me during the Ph.D. time and solving many problems on the paperwork.

I would like to thank Prof. Stefan Mundlos and Prof. Heiko Krude for supporting me at the last two years. I also thank them for giving me the chance to join their projects where I applied the theoretical knowledge to practice and broadened my vision on genetics.

I would like to thank my colleagues too. Thanks my office roommate Sebastian Köhler for helping me to check the thesis and the procedure of promotion. I also appreciate his other helps and his funny jokes to make me laugh. I am thankful to Martin Jäger and Manuel Holtgrewe to give me NGS technology support. I also thank Max Schubach, Anne Thorwarth, Tom Kamphans, Malte Spielmann and so on to corporate with me on the projects. I thank all colleagues for taking care of me in these years.

I also appreciate the supports from all friends here. I thank Floyd Douglas for helping me to correct the grammar of my thesis and teaching me a lot of politics and economics. I am thankful

Geng Chen for listening my complaint, sharing his experience with me. I thank Yujiao Qu for accompanying me at the depressive moments and cooking tasty food for me. I thank all friends here. We travel together, do sports together, join social events together and have a lot of enjoyable parties. I appreciate these lovely people never allowing me to be alone at all important festivals.

I am also thankful my friends in China for encouraging me, understanding me and offering me help at any time.

I appreciate my country China. I thank China scholarship council for awarding me three-year PhD scholarship and giving me a chance for abroad study. I also thank for the great care from China Embassy for Chinese students.

As the second longest city I stay, Berlin is already my second hometown. I thank Berlin for giving me a great European life and offering me the chance to meet so many great people here. I have already missed here.

In the end, I must thank my dearest families. Whenever for whatever, they are always there to give me the biggest support. Thanks my brothers to take care of mother and thanks mother understanding my absence in her life for these years. The final words to my father are: if you could see everything, how great it would be.

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe. Die Arbeit wurde in keinem früheren Promotionsverfahren eingereicht.

Berlin, April 2016

Acronyms

1KGP	1000 Genome Project. 20, 21, 36, 49, 55–58, 60, 61, 64, 69–79, 81–83, 87, 88
BER	Cohorts sequenced in Charité - Universitätsmedizin Berlin. 19, 28, 49, 55–61, 63–66, 69–77, 79, 87, 88
CATT	Cochran-Armitage test for trend. 38, 39, 54, 67, 70, 78, 79, 89
CDCV	common disease common variant. 10
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry. 61, 64, 65
CLR	Composite likelihood ratio. 42, 54
CMC	Combined Multivariate and Collapsing. 38, 40–42, 54, 78, 79
CVAS	common variant association study. 10, 12, 16, 37, 75, 76, 93–97, 102
DB	Davies-Bouldin Index. 36, 37, 60–64
DNA	Deoxyribonucleic acid. 1, 2, 6, 23, 28, 48, 99
ESP	Exome sequencing project. 33, 64, 87, 97, 99
ExAC	Exome Aggregation Consortium. 64, 87, 97, 99

FBAT	Family-based association test. 98
FIN	Finnish in Finland. 58, 59, 61, 64–66, 73, 74
FWER	Family-wise error rate. 44, 46, 54, 55, 66, 67, 102
GBR	British in England and Scotland. 58, 59, 61, 73, 74, 81, 82
GMD	Genomic Matchmaking databases. 90, 92, 94, 96, 97
GWAS	Genome-wide association study. 4, 17, 47, 52, 93, 99, 102
HPMRS	Hyperphosphatasia with mental retardation syndrome. 22–24, 27, 66, 69, 70, 72, 77, 79, 83, 84, 86
IBD	Identity by descent. 48
IBS	Iberian Population in Spain. 61
IBS	Identity by state. 34–36, 48, 56, 57, 60–62, 66, 67, 78, 94, 95
ID	intellectual disability. 9
INDEL	Insertion and deletion. 86
MAF	Minor allele frequency. 11, 42
MDS	Multidimensional scaling. 88
NGS	Next generation sequencing. 1–3, 6, 13
ROC	Receiver Operating Characteristic. 66, 67

RVAS	Rare variant association study. 9, 11, 12, 16–18, 28, 32, 37, 46, 52, 54, 55, 66, 69, 72–77, 79–83, 86, 87, 90, 93–100, 102
RVIS	residual variation intolerance score. 33, 54, 72, 73, 97
SNP	single nucleotide polymorphism. 10
SNV	Single nucleotide variant. 86
TSI	Toscani in Italia. 61
VT	Variable threshold. 42, 54, 81
WES	Whole exome sequencing. 3, 4, 6, 19–21
WGS	Whole genome sequencing. 3–6, 20, 21

Bibliography

[1] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, December 1977.

[2] W Ansorge, B Sproat, J Stegemann, C Schwager, and M Zenke. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic acids research*, 15(11):4593–602, June 1987.

[3] L M Smith, J Z Sanders, R J Kaiser, et al. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–9, January.

[4] Gordon E. Moore. Cramming more components onto integrated circuits.

[5] Marcel Margulies, Michael Egholm, William E Altman, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, September 2005.

[6] Susanne Balzer, Ketil Malde, Anders Lanzén, Animesh Sharma, and Inge Jonassen. Characteristics of 454 pyrosequencing data–enabling realistic simulation with flowsim. *Bioinformatics (Oxford, England)*, 26(18):i420–5, September 2010.

[7] Scott D Kahn. On the future of genomic data. *Science (New York, N.Y.)*, 331(6018):728–9, February 2011.

[8] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into

third-generation sequencing. *Human molecular genetics*, 19(R2):R227–40, October 2010.

[9] Antonis Rokas and Patrick Abbot. Harnessing genomics for evolutionary insights. *Trends in ecology & evolution*, 24(4):192–200, April 2009.

[10] Christian Gilissen, Jayne Y Hehir-Kwa, Djie Tjwan Thung, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347, July 2014.

[11] The 1000 Genome Project Consortium. An integrated map of genetic variation from 1092 human genomes. *Nature*, 491(56), 2012.

[12] Yun Li, Carlo Sidore, Hyun Min Kang, Michael Boehnke, and Gonçalo R Abecasis. Low-coverage sequencing: implications for design of complex trait association studies. *Genome research*, 21(6):940–51, June 2011.

[13] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*, 12(11):745–55, November 2011.

[14] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American journal of human genetics*, 95(1):5–23, July 2014.

[15] Peter M Krawitz, Yoshiko Murakami, Angelika Rieß, et al. PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. *American journal of human genetics*, 92(4):584–9, April 2013.

[16] Ron Do, Sekar Kathiresan, and Gonçalo R Abecasis. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human molecular genetics*, 21(R1):R1–9, October 2012.

[17] Elizabeth T Cirulli and David B Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics*, 11(6):415–425, June 2010.

[18] Lin T Guey, Jasmina Kravic, Olle Melander, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic epidemiology*, 35(4):236–46, May 2011.

[19] Ian J Barnett, Seunggeun Lee, and Xihong Lin. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic epidemiology*, 37(2):142–51, February 2013.

[20] N Risch and H Zhang. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science (New York, N.Y.)*, 268(5217):1584–9, June 1995.

[21] D. Seelow, M. Schuelke, F. Hildebrandt, and P. Nurnberg. HomozygosityMapper–an interactive approach to homozygosity mapping. *Nucleic Acids Research*, 37(Web Server):W593–W599, May 2009.

[22] Christian Rödelsperger, Peter Krawitz, Sebastian Bauer, et al. Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. *Bioinformatics (Oxford, England)*, 27(6):829–36, March 2011.

[23] Tom Kamphans, Peggy Sabri, Na Zhu, et al. Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PloS one*, 8(8):e70151, January 2013.

[24] Peter M Krawitz, Yoshiko Murakami, Jochen Hecht, et al. Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. *American journal of human genetics*, 91(1):146–51, July 2012.

[25] Christian Gilissen, Alexander Hoischen, Han G Brunner, and Joris a Veltman. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, 20(5):490–497, 2012.

[26] Sham P C Purcell S Cherny SS. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–150, 2003.

[27] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, et al. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111:E455—-64, 2014.

[28] J. K. Pritchard. The allelic architecture of human disease genes: common disease-common variant... or not? *Human Molecular Genetics*, 11(20):2417–2423, October 2002.

[29] John P A Ioannidis. Genetic associations: false or true? *Trends in molecular medicine*, 9(4):135–8, April 2003.

[30] Pim van der Harst, Weihua Zhang, Irene Mateo Leach, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–75, December 2012.

[31] Hana Lango Allen, Karol Estrada, Guillaume Lettre, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–8, October 2010.

[32] Jian Yang, Beben Benyamin, Brian P McEvoy, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–9, July 2010.

[33] S Hong Lee, Teresa R DeCandia, Stephan Ripke, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature genetics*, 44(3):247–50, March 2012.

[34] Evan E Eichler, Jonathan Flint, Greg Gibson, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11(6):446–50, June 2010.

[35] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8, January 2012.

[36] Greg Gibson. Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–45, February 2011.

[37] J Asimit and E Zeggini. Rare Variant Association Analysis Methods for Complex Traits. *Annual Review of Genetics*, 44:293–308, August 2010.

[38] Sergey Nejentsev, Neil Walker, David Riches, Michael Egholm, and John A Todd. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.)*, 324(5925):387–9, April 2009.

[39] Weizhen Ji, Jia Nee Foo, Brian J O'Roak, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature genetics*, 40:592–599, 2008.

[40] Thorlakur Jonsson, Jasvinder K Atwal, Stacy Steinberg, et al. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature*, 488(7409):96–9, August 2012.

[41] Manuel A Rivas, Mélissa Beaudoin, Agnes Gardet, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*, 43(11):1066–73, November 2011.

[42] Xiaowei Zhan, David E Larson, Chaolong Wang, et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nature genetics*, 45(11):1375–9, November 2013.

[43] Johanna M Seddon, Yi Yu, Elizabeth C Miller, et al. Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. *Nature Genetics*, 45(11):1366–1370, September 2013.

[44] Hannes Helgason, Patrick Sulem, Maheswara R Duvvari, et al. A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. *Nature Genetics*, 45(11):1371–1374, September 2013.

[45] Martin Ladouceur, Zari Dastani, Yurii S Aulchenko, Celia M T Greenwood, and J Brent Richards. The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS genetics*, 8(2):e1002496, feb 2012.

[46] Kirk E. Lohmueller, Thomas Sparsø, Qibin Li, et al. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *American Journal of Human Genetics*, 93:1072–1086, 2013.

[47] Anna C Need, Joseph P McEvoy, Massimo Gennarelli, et al. Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *American journal of human genetics*, 91(2):303–12, August 2012.

[48] Erin L. Heinzen, Chantal Depondt, Gianpiero L. Cavalleri, et al. Exome Sequencing Followed by Large-Scale Genotyping Fails to Identify Single Rare Variants of Large Effect in Idiopathic Generalized Epilepsy. *The American Journal of Human Genetics*, 91(2):293–302, August 2012.

[49] Li Liu, Aniko Sabo, Benjamin M. Neale, et al. Analysis of Rare, Exonic Variation amongst Subjects with Autism Spectrum Disorders and Population Controls. *PLoS Genetics*, 9(4):e1003443, April 2013.

[50] Karen A Hunt, Vanisha Mistry, Nicholas A Bockett, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*, 498(7453):232–5, June 2013.

[51] Norio R. The Finnish Disease Heitage III: the individual diseases. *Human Mutation*, 112(5-6):470–526, 2003.

[52] David R Blair, Christopher S Lyttle, Jonathan M Mortensen, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell*, 155(1):70–80, September 2013.

[53] A L Price, N J Patterson, R M Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

[54] Chao Tian, Robert M Plenge, Michael Ransom, et al. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS genetics*, 4(1):e4, January 2008.

[55] Marc Bauchet, Brian McEvoy, Laurel N Pearson, et al. Measuring European population stratification with microarray genotype data. *American journal of human genetics*, 80(5):948–56, May 2007.

[56] John Novembre, Toby Johnson, Katarzyna Bryc, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, November 2008.

[57] Alkes L Price, Agnar Helgason, Snaebjorn Palsson, et al. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS genetics*, 5(6):e1000505, June 2009.

[58] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517, May 2004.

[59] Devlin B. and Roeder K. Genomic control for association study. *Biometrics*, 55(4):997–1004, 1999.

[60] S a Bacanu, B Devlin, and K Roeder. The power of genomic control. *American journal of human genetics*, 66:1933–1944, 2000.

[61] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, June 2000.

[62] Benjamin F. Voight and Jonathan K. Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS genetics*, 1(3), 2005.

[63] Jianming Yu, Gael Pressoir, William H Briggs, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–8, February 2006.

[64] Hyun Min Kang, Jae Hoon Sul, Susan K Service, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42:348–354, 2010.

[65] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355–60, April 2010.

[66] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, et al. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6):525–6, June 2012.

[67] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–4, July 2012.

[68] Matthew Zawistowski, Mark Reppell, Daniel Wegmann, et al. Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *European journal of human genetics : EJHG*, 22(9):1137–44, sep 2014.

[69] Carlos D Bustamante, Esteban González Burchard, and Francisco M De la Vega. Genomics for the world. *Nature*, 475(7355):163–5, jul 2011.

[70] Marie-Claude Babron, Marie de Tayrac, Douglas N Rutledge, Eleftheria Zeggini, and Emmanuelle Génin. Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PloS one*, 7(10):e46519, January 2012.

[71] I Mathieson and G McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44:243–246.

[72] Jennifer Listgarten, Christoph Lippert, and David Heckerman. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature genetics*, 45(5):470–1, May 2013.

[73] Iain Mathieson and Gil McVean. Reply to: "FaST-LMM-Select for addressing confounding from spatial structure and rare variants". *Nature genetics*, 45(5):471, May 2013.

[74] Diana Luca, Steven Ringquist, Lambertus Klei, et al. On the Use of General Control Samples for Genome-wide Association Studies: Genetic Matching Highlights Causal Variants. *The American Journal of Human Genetics*, 82(2):453–463, February 2008.

[75] Ann B Lee, Diana Luca, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Discovering genetic ancestry using spectral graph theory. *Genetic epidemiology*, 34(1):51–9, January 2010.

[76] Michael P Epstein, Richard Duncan, K Alaine Broadaway, et al. Stratification-score matching improves correction for confounding by population stratification in case-control association studies. *Genetic epidemiology*, 36(3):195–205, April 2012.

[77] Weihua Guan, Liming Liang, Michael Boehnke, and Gonçalo R Abecasis. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genetic epidemiology*, 33(6):508–17, September 2009.

[78] Mark S Silverberg, Judy H Cho, John D Rioux, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nature genetics*, 41(2):216–20, March 2009.

[79] Blanca E Himes, Gary M Hunninghake, James W Baurley, et al. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *American journal of human genetics*, 84(5):581–93, May 2009.

[80] Vivianna M Van Deerlin, Patrick M A Sleiman, Maria Martinez-Lage, et al. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nature genetics*, 42(3):234–9, March 2010.

[81] Nadja Ehmke, Almuth Caliebe, Rainer Koenig, et al. Homozygous and Compound-Heterozygous Mutations in TGDS Cause Catel-Manzke Syndrome. *American journal of human genetics*, 95(6):763–70, December 2014.

[82] Peter M Krawitz, Michal R Schweiger, Christian Rödelsperger, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nature genetics*, 42(10):827–9, October 2010.

[83] Mark a DePristo, Eric Banks, Ryan Poplin, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, May 2011.

[84] The 1000 Genome Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467, 2011.

[85] Illumina. Illumina — Sequencing and array-based solutions for genetic research.

[86] Yan Guo, Jing He, Shilin Zhao, et al. Illumina human exome genotyping array clustering and quality control. *Nature protocols*, 9(11):2643–62, November 2014.

[87] The 1000 Genome Project Consortium. About — 1000 Genomes.

[88] Carrie B. Moore, John R. Wallace, Daniel J. Wolfe, et al. Low Frequency Variants, Collapsed Based on Biological Knowledge, Uncover Complexity of Population Stratification in 1000 Genomes Project Data. *PLoS Genetics*, 9(12), 2013.

[89] Michael Nothnagel, Alexander Herrmann, Andreas Wolf, et al. Technology-specific error signatures in the 1000 Genomes Project data. *Human genetics*, 130(4):505–16, October 2011.

[90] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, sep 2015.

[91] Heng Li, Bob Handsaker, Alec Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009.

[92] Tobias Rausch, Thomas Zichner, Andreas Schlattl, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, 28(18):i333–i339, September 2012.

[93] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21):2865–71, November 2009.

[94] J E Allanson. Noonan syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 145C(3):274–279, August 2007.

[95] M Sharland, M Burch, W M McKenna, and M A Paton. A clinical study of Noonan syndrome. *Archives of disease in childhood*, 67(2):178–83, February 1992.

[96] Judith E Allanson and Amy E Roberts. Noonan Syndrome, August 2011.

[97] Ineke van der Burgt. Noonan syndrome. *Orphanet journal of rare diseases*, 2:4, January 2007.

[98] Md Abdur Razzaque, Yuta Komoike, Tsutomu Nishizawa, et al. Characterization of a novel KRAS mutation identified in Noonan syndrome. *American journal of medical genetics. Part A*, 158A(3):524–32, March 2012.

[99] Amy E Roberts, Toshiyuki Araki, Kenneth D Swanson, et al. Germline gain-of-function mutations in SOS1 cause Noonan syndrome. *Nature genetics*, 39(1):70–4, January 2007.

[100] M Abdur Razzaque, Tsutomu Nishizawa, Yuta Komoike, et al. Germline gain-of-function mutations in RAF1 cause Noonan syndrome. *Nature genetics*, 39(8):1013–7, August 2007.

[101] E Denayer, H Peeters, L Sevenants, et al. NRAS Mutations in Noonan Syndrome. *Molecular syndromology*, 3(1):34–38, June 2012.

[102] Anna Sarkozy, Claudio Carta, Sonia Moretti, et al. Germline BRAF mutations in Noonan, LEOPARD, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum. *Human mutation*, 30(4):695–702, April 2009.

[103] Richard JH Smith and Michael Hildebrand. DFNA2 Nonsyndromic Hearing Loss, June 2013.

[104] Simone Rost, Elisa Bach, Cordula Neuner, et al. Novel form of X-linked nonsyndromic hearing loss with cochlear malformation caused by a mutation in the type IV collagen gene COL4A6. *European journal of human genetics : EJHG*, 22(2):208–15, February 2014.

[105] Arti Pandya. Nonsyndromic Hearing Loss and Deafness, Mitochondrial, July 2014.

[106] G Van Camp, P J Willems, and R J Smith. Nonsyndromic hearing impairment: unparalleled heterogeneity. *American journal of human genetics*, 60(4):758–64, April 1997.

[107] Richard JH Smith, A Eliot Shearer, Michael S Hildebrand, and Guy Van Camp. Deafness and Hereditary Hearing Loss Overview, January 2014.

[108] Richard JH Smith and Guy Van Camp. Nonsyndromic Hearing Loss and Deafness, DFNB1, January 2014.

[109] P M Kelley, D J Harris, B C Comer, et al. Novel mutations in the connexin 26 gene (GJB2) that cause autosomal recessive (DFNB1) hearing loss. *American journal of human genetics*, 62(4):792–9, April 1998.

[110] X Estivill, P Fortina, S Surrey, et al. Connexin-26 mutations in sporadic and inherited sensorineural deafness. *Lancet*, 351(9100):394–8, February 1998.

[111] L Zelante, P Gasparini, X Estivill, et al. Connexin26 mutations associated with the most common form of non-syndromic neurosensory autosomal recessive deafness (DFNB1) in Mediterraneans. *Human molecular genetics*, 6(9):1605–9, September 1997.

[112] Nele Hilgert, Richard J H Smith, and Guy Van Camp. Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics? *Mutation research*, 681(2-3):189–96, January 2009.

[113] J M Bork, L M Peters, S Riazuddin, et al. Usher syndrome 1D and nonsyndromic autosomal recessive deafness DFNB12 are caused by allelic mutations of the novel cadherin-like gene CDH23. *American journal of human genetics*, 68(1):26–37, January 2001.

[114] X. Z. Liu. Mutations in GJA1 (connexin 43) are associated with non-syndromic autosomal recessive deafness. *Human Molecular Genetics*, 10(25):2945–2951, December 2001.

[115] C C Mabry, A Bautista, R F Kirk, et al. Familial hyperphosphatase with mental retardation, seizures, and neurologic deficits. *The Journal of pediatrics*, 77(1):74–85, July 1970.

[116] K Kruse, F Hanefeld, A Kohlschütter, R Rosskamp, and G Gross-Selbeck. Hyperphosphatasia with mental retardation. *The Journal of pediatrics*, 112(3):436–9, March 1988.

[117] Carlo L Marcelis, Paul Rieu, Frits Beemer, and Han G Brunner. Severe mental retardation, epilepsy, anal anomalies, and distal phalangeal hypoplasia in siblings. *Clinical dysmorphology*, 16(2):73–6, April 2007.

[118] Denise Horn, Peter Krawitz, Anca Mannhardt, Georg Christoph Korenke, and Peter Meinecke. Hyperphosphatasia-mental retardation syndrome due to PIGV mutations: expanded clinical spectrum. *American journal of medical genetics. Part A*, 155A(8):1917–22, August 2011.

[119] Denise Horn, Gudrun Schottmann, and Peter Meinecke. Hyperphosphatasia with mental retardation, brachytelephalangy, and a distinct facial gestalt: Delineation of a recognizable syndrome. *European journal of medical genetics*, 53(2):85–8, January.

[120] Malcolm F Howard, Yoshiko Murakami, Alistair T Pagnamenta, et al. Mutations in PGAP3 impair GPI-anchor maturation, causing a subtype of hyperphosphatasia with mental retardation. *American journal of human genetics*, 94(2):278–87, February 2014.

[121] Stuart A Scott, Lisa Edelmann, Liu Liu, et al. Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. *Human mutation*, 31(11):1240–50, November 2010.

[122] D J Boles and R L Proia. The molecular basis of HEXA mRNA deficiency caused by the most common Tay-Sachs disease mutation. *American journal of human genetics*, 56(3):716–24, March 1995.

[123] R Rozenberg and L da V Pereira. The frequency of Tay-Sachs disease causing mutations in the Brazilian Jewish population justifies a carrier screening program. *São Paulo medical journal = Revista paulista de medicina*, 119(4):146–9, July 2001.

[124] Mohamad R Chaaban, Alexandra Kejner, Steven M Rowe, and Bradford A Woodworth. Cystic fibrosis chronic rhinosinusitis: a comprehensive review. *American journal of rhinology & allergy*, 27(5):387–95, January 2013.

[125] Kevin W Southern, Anne Munck, Rodney Pollitt, et al. A survey of newborn screening for cystic fibrosis in Europe. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society*, 6(1):57–65, January 2007.

[126] Brian P O'Sullivan and Steven D Freedman. Cystic fibrosis. *Lancet*, 373(9678):1891–904, May 2009.

[127] Virginie Scotet, Ingrid Duguépéroux, Philippe Saliou, et al. Evidence for decline in the incidence of cystic fibrosis: a 35-year observational study in Brittany, France. *Orphanet journal of rare diseases*, 7:14, January 2012.

[128] Rosalie E Ferner, Susan M Huson, Nick Thomas, et al. Guidelines for the diagnosis and management of individuals with neurofibromatosis 1. *Journal of medical genetics*, 44(2):81–8, February 2007.

[129] G F Xu, P O'Connell, D Viskochil, et al. The neurofibromatosis type 1 gene encodes a protein related to GAP. *Cell*, 62(3):599–608, August 1990.

[130] Hermann Manzke, Katarina Lehmann, Eva Klopocki, and Almuth Caliebe. Catel-Manzke syndrome: two new patients and a critical review of the literature. *European journal of medical genetics*, 51(5):452–65, January.

[131] Yoshikazu Kuroki, Yasuyuki Suzuki, Hiroyuki Chyo, Akira Hata, and Ichiro Matsui. A new malformation syndrome of long palpebralfissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. *The Journal of Pediatrics*, 99(4):570–573, October 1981.

[132] N Niikawa, Y Kuroki, T Kajii, et al. Kabuki make-up (Niikawa-Kuroki) syndrome: a study of 62 patients. *American journal of medical genetics*, 31(3):565–89, November 1988.

[133] S M White, E M Thompson, A Kidd, et al. Growth, behavior, and clinical findings in 27 patients with Kabuki (Niikawa-Kuroki) syndrome. *American journal of medical genetics. Part A*, 127A(2):118–27, June 2004.

[134] Sarah B Ng, Abigail W Bigham, Kati J Buckingham, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics*, 42(9):790–793, 2010.

[135] Damien Lederer, Bernard Grisart, Maria Cristina Digilio, et al. Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *American journal of human genetics*, 90(1):119–24, January 2012.

[136] Noriko Miyake, Seiji Mizuno, Nobuhiko Okamoto, et al. KDM6A point mutations cause Kabuki syndrome. *Human mutation*, 34(1):108–10, January 2013.

[137] N Bögershausen and B Wollnik. Unmasking Kabuki syndrome. *Clinical genetics*, 83(3):201–11, March 2013.

[138] Margaret P Adam, Louanne Hudgins, and Mark Hannibal. Kabuki Syndrome, May 2013.

[139] Shunichi Kosugi, Satoshi Natsume, Kentaro Yoshida, et al. Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data. *PloS one*, 8(10):e75402, January 2013.

[140] J J Zhuang, K Zondervan, F Nyberg, et al. Optimizing the power of genome-wide association studies by using publicly available reference samples to expand the control group. *Genetic Epidemiology*, 34:319–326, 2010.

[141] Adam Kiezun, Kiran Garimella, Ron Do, et al. Exome sequencing and the genetic basis of complex traits. *Nature Genetics*, 44(6):623–630, 2012.

[142] Melissa J Landrum, Jennifer M Lee, Mark Benson, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–8, nov 2015.

[143] S. T. Sherry, M H Ward, M Kholodov, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, January 2001.

[144] David B. Carlini and Wolfgang Stephan. In Vivo Introduction of Unpreferred Synonymous Codons Into the Drosophila Adh Gene Results in Reduced Levels of ADH Protein. *Genetics*, 163(1):239–243, January 2003.

[145] Franco Pagani, Michela Raponi, and Francisco E Baralle. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18):6368–72, May 2005.

[146] Kristina V Krasnov, Maria Tzetis, Jie Cheng, William B Guggino, and Garry R Cutting. Localization studies of rare missense mutations in cystic fibrosis transmembrane conductance regulator (CFTR) facilitate interpretation of genotype-phenotype relationships. *Human mutation*, 29(11):1364–72, November 2008.

[147] C Lázaro, R de Cid, J Sunyer, et al. Missense mutations in the cystic fibrosis gene in adult patients with asthma. *Human mutation*, 14(6):510–9, January 1999.

[148] Ehm M G Kessner D et.al Nelson M.R Wegmann D and Ehm M G Kessner D et.al Nelson M.R Wegmann D. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337, 2012.

[149] Gregory V Kryukov, Len A Pennacchio, and Shamil R Sunyaev. Most rare missense alleles are deleterious in humans: implications for complex

disease and association studies. *American journal of human genetics*, 80(4):727–39, April 2007.

[150] A Hamosh, B J Rosenstein, and G R Cutting. CFTR nonsense mutations G542X and W1282X associated with severe reduction of CFTR mRNA in nasal epithelial cells. *Human molecular genetics*, 1(7):542–4, October 1992.

[151] David M. Bedwell, Anisa Kaenjak, Dale J. Benos, et al. Suppression of a CFTR premature stop mutation in a bronchial epithelial cell line. *Nature Medicine*, 3(11):1280–1284, November 1997.

[152] Yan Guo, Jirong Long, Jing He, et al. Exome sequencing generates high quality data in non-target regions. *BMC genomics*, 13(1):194, January 2012.

[153] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(Database issue):D514–7, January 2005.

[154] Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics*, 7:61–80, January 2006.

[155] Marten Jäger, Kai Wang, Sebastian Bauer, et al. Jannovar: a java library for exome annotation. *Human mutation*, 35(5):548–55, May 2014.

[156] K Wang, M Li, and H Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(e164), 2010.

[157] Shamil R Sunyaev. Inferring causality and functional significance of human coding DNA variants. *Human molecular genetics*, 21(R1):R10–7, October 2012.

[158] Peter D Stenson, Edward V Ball, Matthew Mort, et al. Human Gene Mutation Database (HGMD): 2003 update. *Human mutation*, 21(6):577–81, June 2003.

[159] Amos Bairoch, Rolf Apweiler, Cathy H Wu, et al. The Universal Protein Resource (UniProt). *Nucleic acids research*, 33(Database issue):D154–9, January 2005.

[160] Paul D Thomas and Anish Kejariwal. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15398–403, October 2004.

[161] Jonathan C Cohen, Robert S Kiss, Alexander Pertsemlidis, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science (New York, N.Y.)*, 305(5685):869–72, August 2004.

[162] Stephanie Hicks, David A Wheeler, Sharon E Plon, and Marek Kimmel. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human mutation*, 32(6):661–8, June 2011.

[163] Elliott H Margulies, Gregory M Cooper, George Asimenos, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome research*, 17(6):760–74, June 2007.

[164] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[165] A Siepel, G Bejerano, J S Pedersen, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, August 2005.

[166] Miles D Thompson, Tony Roscioli, Carlo Marcelis, et al. Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). *American journal of medical genetics. Part A*, 158A(3):553–8, March 2012.

[167] Orphanet: Hyperphosphatasia intellectual disability syndrome Mabry syndrome.

[168] Slavé Petrovski, Quanli Wang, Erin L Heinzen, Andrew S Allen, and David B Goldstein. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics*, 9(8):e1003709, jan 2013.

[169] Caroline M Nievergelt, Ondrej Libiger, and Nicholas J Schork. Generalized analysis of molecular variance. *PLoS genetics*, 3(4):e51, April 2007.

[170] M S McPeek and L Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. *American journal of human genetics*, 66(3):1076–94, March 2000.

[171] Michael S. Blouin. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, 18(10):503–511, October 2003.

[172] K. B., CASRIC V, and BoNhomme F. IDENTIX, a software to test for relatedness in a population using permutation methods. *Molecular Ecology Notes*, (2):611–614, 2002.

[173] V Heinrich, T Kamphans, J Stange, et al. Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Medicine*, 5(69), 2013.

[174] M Lynch and K Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152(4):1753–66, August 1999.

[175] Michael Brudno, Alexander Poliakov, Asaf Salamov, et al. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome research*, 14(4):685–92, April 2004.

[176] Kelly A Frazer, Lior Pachter, Alexander Poliakov, Edward M Rubin, and Inna Dubchak. VISTA: computational tools for comparative genomics. *Nucleic acids research*, 32(Web Server issue):W273–9, July 2004.

[177] Nameeta Shah, Olivier Couronne, Len A Pennacchio, et al. Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics (Oxford, England)*, 20(5):636–43, March 2004.

[178] Jacob A Tennessen, Abigail W Bigham, Timothy D O'Connor, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)*, 337(6090):64–9, July 2012.

[179] Ananyo Choudhury, Scott Hazelhurst, Ayton Meintjes, et al. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC genomics*, 15(1):437, jan 2014.

[180] Thomas D Schneider. A brief review of molecular information theory. *Nano communication networks*, 1(3):173–180, September 2010.

[181] Shannon C. E. The Mathematical Theory of Communication. *University of Illinois Press*, 1949.

[182] D L Davies and D W Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, February 1979.

[183] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, October 2005.

[184] Jeroen R Huyghe, Anne U Jackson, Marie P Fogarty, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature genetics*, 45(2):197–201, February 2013.

[185] Armitage P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, 11(3):375–386, 1955.

[186] W G Cochran. Some methods for strengthening the common chi-squared tests. *Biometrics*, 10(4):417–451, 1954.

[187] Geraldine M Clarke, Carl A Anderson, Fredrik H Pettersson, et al. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–33, February 2011.

[188] B Freidlin, G Zheng, Z Li, and J L Gastwirth. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*, 53:146–152, 2002.

[189] Li B., Sm M Leal, Bingshan Li, and Sm M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics*, 83:311–321, 2008.

[190] Momiao Xiong, Jinying Zhao, and Eric Boerwinkle. Generalized T2 test for genome association studies. *American journal of human genetics*, 70(5):1257–68, May 2002.

[191] Alkes L. Price, Gregory V. Kryukov, Paul I W de Bakker, et al. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *American Journal of Human Genetics*, 86(6):832–838, 2010.

[192] M. Yandell, C. Huff, H. Hu, et al. A probabilistic disease-gene finder for personal genomes. *Genome Research*, 21(9):1529–1542, June 2011.

[193] Peter Krawitz, Orion Buske, Na Zhu, Michael Brudno, and Peter N Robinson. The Genomic Birthday Paradox: How Much Is Enough? *Human mutation*, August 2015.

[194] Meyer Dwass. Modified Randomization Tests for Nonparametric Hypotheses on JSTOR. *The Annals of Mathematical Statistics*, 28(1):181–187, 1957.

[195] B V North, D Curtis, and P C Sham. A note on the calculation of empirical P values from Monte Carlo procedures. *American journal of human genetics*, 71(2):439–41, August 2002.

[196] Pak C Sham and Shaun M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature reviews. Genetics*, 15(5):335–46, 2014.

[197] Neyman and Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses on JSTOR, 1933.

[198] M J Sillanpää. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, 106(4):511–9, apr 2011.

[199] Hyun Min Kang. EPACTS - Genome Analysis Wiki, 2015.

[200] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75, September 2007.

[201] G. R. Abecasis, S. S. Cherny, W. O. C. Cookson, and L. R. Cardon. GRR: graphical representation of relationship errors. *Bioinformatics*, 17(8):742–743, August 2001.

[202] Elisha D O Roberson and Jonathan Pevsner. Visualization of shared genomic regions and meiotic recombination in high-density SNP data. *PloS one*, 4(8):e6711, January 2009.

[203] A Jacquard. Genetic information given by a relative. *Biometrics*, 28(4):1101–14, December 1972.

[204] Krina T. Zondervan. *Analysis of Complex Disease Association Studies.* Elsevier, 2011.

[205] Krina T Zondervan and Lon R Cardon. Designing candidate gene and genome-wide case-control association studies. *Nature protocols*, 2(10):2492–501, January 2007.

[206] Carrie C Buchanan, Eric S Torstenson, William S Bush, and Marylyn D Ritchie. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association : JAMIA*, 19(2):289–94, January.

[207] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):11983–8, July 2011.

[208] Steven Gazal, Mourad Sahbatou, Marie-Claude Babron, Emmanuelle Génin, and Anne-Louise Leutenegger. High level of inbreeding in final phase of 1000 Genomes Project. *Scientific reports*, 5:17453, jan 2015.

[209] Peter N Robinson, P Krawitz, and S Mundlos. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clinical genetics*, 80(2):127–32, August 2011.

[210] C Hercus. Novoalign V2.07. \url{http://www.novocraft.com}, 2011.

[211] Kym M Boycott, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature reviews. Genetics*, 14(10):681–91, October 2013.

[212] Anirban DasGupta. The matching, birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, 130(1-2):377–389, March 2005.

[213] Danielle Welter, Jacqueline MacArthur, Joannella Morales, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(Database issue):D1001–6, jan 2014.

[214] Ann B. Lee, Diana Luca, and Kathryn Roeder. A spectral graph approach to discovering genetic ancestry. *The Annals of Applied Statistics*, 4(1):179–202, March 2010.

[215] Sebastian Akle, Sung Chun, Daniel M Jordan, and Christopher A Cassa. Mitigating false-positive associations in rare disease gene discovery. *Human mutation*, 36(10):998–1003, oct 2015.

[216] Kaitlin E Samocha, Elise B Robinson, Stephan J Sanders, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–50, sep 2014.

[217] Exome Aggregation Consortium, Monkol Lek, Konrad Karczewski, et al. Analysis of protein-coding genetic variation in 60,706 humans. Technical report, oct 2015.

[218] V. M. Narasimhan, K. A. Hunt, D. Mason, et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, page aac8624, mar 2016.

[219] Cornelis A Albers, Dirk S Paul, Harald Schulze, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature genetics*, 44(4):435–9, S1–2, April 2012.

[220] Nan M Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature reviews. Genetics*, 7(5):385–94, May 2006.

[221] Paul L Auer and Guillaume Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome medicine*, 7(1):16, jan 2015.

[222] W C Stewart and J Cerise. Increasing the power of association studies with affected families, unrelated cases and controls. *Frontiers in genetics*, 4, 2013.

[223] H Putter, J Houwing-Duistermaat J., and N J D Nagelkerke. Combining evidence for association from transmission disequilibrium and case-control studies using single-nucleotide polymorphisms. *BMC Genetics*, 6(Suppl 1), 2005.

[224] Yun Zhu and Momiao Xiong. Family-based association studies for next-generation sequencing. *American journal of human genetics*, 90(6):1028–45, June 2012.

[225] J. Homsy, S. Zaidi, Y. Shen, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science*, 350(6265):1262–1266, dec 2015.

[226] Megha Ghildiyal and Phillip D Zamore. Small silencing RNAs: an expanding universe. *Nature reviews. Genetics*, 10(2):94–108, February 2009.

[227] Malte Spielmann and Stefan Mundlos. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 35(6):533–43, June 2013.

[228] Axel Visel, Edward M Rubin, and Len A Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, September 2009.

[229] Dirk A Kleinjan and Veronica van Heyningen. Long-range control of gene expression: emerging mechanisms and disruption in disease. *American journal of human genetics*, 76(1):8–32, January 2005.

[230] Qiongshi Lu, Yiming Hu, Jiehuan Sun, et al. A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. *Scientific Reports*, 5:10576, June 2015.

[231] Yao Fu, Zhu Liu, Shaoke Lou, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15(10):480, October 2014.

[232] Roger P Alexander, Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B Gerstein. Annotating non-coding regions of the genome. *Nature reviews. Genetics*, 11(8):559–71, August 2010.

[233] Brad Gulko, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, 47(3):276–283, January 2015.

[234] Tim R Mercer and John S Mattick. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome research*, 23(7):1081–8, July 2013.

[235] DaríoG. Lupiáñez, Katerina Kraft, Verena Heinrich, et al. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012–25, May 2015.

[236] Sebastian Köhler. The Phenomizer - Clinical Diagnostics with Similarity Searches in Ontologies.

[237] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–5, November 2008.

[238] Sebastian Köhler, Marcel H Schulz, Peter Krawitz, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*, 85(4):457–64, October 2009.

[239] Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(Database issue):D966–74, January 2014.