Methods for Sequence Based Studies of Complex Traits

by

Sara Rachel Rashkin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2015

Doctoral Committee:

Professor Goncalo Abecasis, Chair
Professor Michael Lee Boehnke
Assistant Professor Hyun Min Kang
Professor Peter Xuekun Song
Assistant Professor Cristen J. Willer

DEDICATION

To my family – Mom, Dad, Lee, and Hannah.

ACKNOWLEDGEMENTS

The research for this dissertation was done with the help of many people, particularly Professor Goncalo Abecasis, Tom Blackwell, Sai Chen, Goo Jun, Alan Kwong, and Mary Kate Wing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Thousands of loci have been associated with complex diseases and traits. However, there is still much we do not know about the biology of disease. Many reasons for this are possible, including the strong focus of genetic association studies over the past 5-10 years on common single nucleotide polymorphisms. In this dissertation, we focus on methods for the design and analysis of sequence based studies, to enable the assessment of other types of variants, particularly rare variants and copy number variants.

In the first chapter, we attempt to resolve the debate regarding the best strategy for studying very rare, disease-associated variants, particularly singleton variants appearing only once in a sample. We estimated the sensitivity to detect singleton variants using both simulations and analysis of real data. We extended this to determine the power of an association study for discrete traits, evaluating the burden of singletons under a variety of situations. We found that sensitivity to detect singletons increases with sequencing depth, plateauing when depth reaches ~25x. For a fixed sequencing capacity, we estimated that power is maximized when samples are sequenced at 15-20x coverage, which produced an optimal trade-off of singleton discovery and sample size. In our assessments, increasing coverage beyond 15-20x (and decreasing sample size) results in reduced power.

In the second chapter, we extend this analysis to a quantitative trait framework. Despite the different disease model, the results for quantitative traits are remarkably similar to those for binary traits. For constant sequencing effort, power is maximized at 11-16x coverage, for a variety of parameter values examined. Increasing coverage further for reduced sample size results in decreased power.

In the final chapter, we turn to another type of variant that may aid in understanding the etiology of disease: copy number variation. Copy number variants (CNVs) are associated with many diseases, especially psychiatric disorders, and there is great interest in methods to accurately detect and genotype CNVs. We developed a method that uses read depth information to estimate copy number for a set of sequenced individuals.

CHAPTER I. Introduction

One aim of genetic association studies is to further our mechanistic understanding of genetic diseases: knowing which genes are associated with disease can be used to determine functional pathways leading to disease.[1] Much work has been done to study the impact of genetic variation on human disease. In recent years, genome-wide association studies (GWAS) have allowed for testing of millions of variants simultaneously.[2; 3] This has made it possible to identify thousands of loci related to diverse diseases.[3; 4] Ultimately, we want to translate association signals to function, and studies have resulted in some key insights.[4] Despite this progress, however, there is still much we do not understand about the functional mechanism of disease.[1; 4]

It is believed that rare variants may prove particularly useful in further elucidating etiology of disease.[5; 6] A strong focus of genetic association studies in the past has been on common single nucleotide polymorphisms (SNPs).[3; 7; 8] While many common variants have been associated with diseases, most convey only a small increase in disease susceptibility – not enough to fully explain the proportion of phenotypic variance explained by genetic factors.[2; 3; 7; 8] Rare variants are more likely to be functional[4-7; 9; 10] and have large effect sizes[7; 9; 10], which point to key locations necessary for proper function. Therefore, rare variant association studies have been expanding rapidly. Examples of recent discoveries include: four low-frequency or rare variants associated with risk of type 2 diabetes, including a rare variant with minor allele frequency (MAF) of 0.20% (odds ratio of 2.27)[11]; a rare variant associated with increased risk of prostate

cancer (odds ratio of 2.90)[12]; and a very rare mutation with MAF 0.0076% associated with increased risk of schizophrenia (odds ratio 7.1).[13]

While it is now clear that rare variants increase disease susceptibility for many diseases, there is still much debate regarding the best way to identify these variants.[4] In particular, the popular theory is that deep sequencing is necessary to obtain a complete catalog of variation.[4; 10] However, some studies have shown that, especially for large sample sizes, many rare variants can be detected at low sequencing depth.[3; 4; 14] Additionally, in most studies, there is a trade-off between sample size and sequencing depth so that, to maintain constant cost, if one increases, the other must decrease. Therefore, deep sequencing often requires reducing sample size.[3; 4] While sequencing at greater depth may increase sensitivity to detect variants, association study power tends to increase with sample size. As a result, it is still unclear what is the ideal sequencing strategy for studying rare variants.

We focused on singleton variants (those that appear only once in a sample), as designs that work well for them will likely provide an upper bound of needed sequencing depth for more common variants, for which low pass data can be analyzed more effectively.[3; 15] We examined the balance between the sensitivity to detect singletons with association study power under a variety of situations in order to identify optimal study designs. We calculated the sensitivity to detect singletons at different sample sizes and read depths using both simulations and real data. We then used these estimates to assess association study power for a discrete trait under various conditions. We found that association study power is maximized at 15-20x coverage for a variety of trait models and study designs examined.

Another area of intense focus in recent years is studies of quantitative traits. Quantitative traits are frequently associated with disease (ie: cholesterol levels with heart disease, glucose levels with diabetes) and can, therefore, be used as markers for disease. This allows for a more flexible method of analysis. Additionally, knowledge of how genetic variation affects quantitative traits can aid in our understanding the mechanistic cause of disease.[16] For instance, knowing a variant is associated with increased cholesterol, which increases the risk of heart disease, tells us how that variant leads to increased risk of heart disease. We extended our analysis to consider the association of singleton variants with quantitative traits. We found that 11-16x coverage results in maximal power for the trait models and study designs we considered.

In addition to rare variants, a better functional understanding of the biology of disease may also lay beyond single nucleotide polymorphisms (SNPs).[2; 8; 17] Copy number variants (CNVs) have been associated with a number of diseases, particularly psychiatric disorders such as schizophrenia and autism.[18-20] It is necessary to have a reliable way of detecting such variants.

Historically, array-based methods have been used to locate CNVs – specifically array Comparative Genomic Hybridization (aCGH) – but these methods do not allow for other types of analyses.[20] As sequencing becomes cheaper, using sequence data to determine copy number becomes an attractive approach, as this allows for saving the cost of aCGH and allows for other types of analyses to be performed simultaneously. A number of methods have been devised to perform such an analysis, but most analyze one sample at a time.[21] We developed a quick, easy

to use, and accurate method that determines copy number within a particular, well-defined region of interest for a sample of individuals.

In this dissertation, we describe methods that will aid in the design and analysis of complex traits using sequence data. We begin by determining optimal designs for studying singleton variants related to discrete traits. We then extend this to ascertain the ideal strategy for studying singletons associated with quantitative traits. Finally, we propose a novel method for detecting copy number variants from sequence data.

CHAPTER II. Optimal Sequencing Strategies for Identifying Disease-Associated Singletons[1]

Abstract

With the increasing focus of genetic association on the identification of trait-associated rare variants, it is important to identify the most cost-effective sequencing strategies for these studies. Deep sequencing will accurately detect and genotype the most rare variants per individual, but may limit sample size. Low pass sequencing will miss some variants in each individual but has been shown to provide a cost effective alternative for studies of common variants. Here, we investigate the impact of sequencing depth on studies of rare variants, focusing on singletons - the variants that are sampled in a single individual and are hardest to detect at low sequencing depths. We first estimate the sensitivity to detect singleton variants in both simulated data and in down-sampled deep genome and exome sequence data. We then explore the power of association studies comparing burden of singleton variants in cases and controls under a variety of conditions. We show that the power to detect singletons increases with coverage, typically plateauing for coverage > ~25x. Next, we show that, when total sequencing capacity is fixed, the power of association studies focused on singletons is typically maximized for coverage of 15-20x, independent of relative risk, disease prevalence, singleton burden, and case-control ratio.

---

[1] Rashkin, S., Jun, G., Chen, S., Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), and Abecasis, G. (In preparation). Optimal Sequencing Strategies for Identifying Disease-Associated Singletons.

Our results suggest sequencing depth of 15-20x as an appropriate compromise of singleton detection power and sample size for studies of rare variants in complex disease.

Introduction

New sequencing technologies are shifting the focus of genetic association studies to rare variants. Rare variants may explain much of the heritability of common, complex diseases.[6; 10] Importantly, rare variants are more likely to severely disrupt gene function[5; 6], and identifying trait associated rare variants may accelerate progress from genetic association signals to mechanistic understanding of disease.

It is frequently asserted that the study of rare variants requires deep sequencing, which provides the highest power for variant discovery in any single genome.[4; 10] The alternative of low-pass sequencing has been advocated for studies of common variation,[3; 14] supported by empirical studies.[22] Low pass sequencing allows for larger sample sizes but misses some variants in each individual and reduces genotyping accuracy.[3; 4] The 1000 Genomes Project used low pass sequencing of ~2,500 individuals to catalog genetic variation and haplotypes across 26 populations.[14] The project has produced a near complete catalog of common variation but also identified many rare variants, including singletons, in sequenced individuals.

We speculated that low pass or intermediate depth sequencing could be useful even for studies of very rare variants. It is now clear that these studies often require large sample sizes, totaling thousands of individuals. Examples of successful rare variant association studies include several studies implicating rare variants in the complement genes (*CFH, C3, CFI, C9*) in the risk of age-

related macular degeneration,[23-26] a study showing rare *IFIH1* variants protect against type 1 diabetes,[27] and a study that found that rare variants that inactivate *NPC1L1* reduce the risk of coronary heart disease.[28]

Here, we attempt a more nuanced view of the optimal strategies for sequence based rare variant association studies and explore and compare the power of rare variant association studies that use low, intermediate, or deep sequencing strategies. Since common variants can be detected and genotyped efficiently by analyzing sequence data for many individuals jointly, we focused our analysis on singletons. Any sequencing depth that works well for singletons should provide an upper bound of needed sequencing depth for more common variants, for which low pass data can be analyzed more effectively.[3; 15] In this paper, we examine discrete traits and seek to maximize association study power for a fixed sequencing effort (in a companion manuscript, we explore quantitative traits[29]). We consider the balance between power to identify very rare variants (which increases with sequencing depth) and power to identify disease association (which increases with sample size). We explore both simulated data and actual sequence data. We estimate the power of association tests for study designs employing deep sequencing, low pass sequencing, and intermediate strategies across a range of sample sizes, singleton frequencies, disease relative risks, and disease prevalence. Our results show that, for fixed cost, power to detect association is maximized at a read depth of 15-20x and decreases rapidly as coverage is increased beyond this threshold.

Methods

Overview

To determine the ideal study design for studying disease-related singletons, we first used simulations to explore the ability to detect singleton variants present in sequenced samples. Next, we validated simulations by comparing results to down-sampled sequence data. Finally, we extended these analyses of variant discovery power to examine association study power.

Definitions

We consider sequencing studies that assess N individuals, sequencing each to an average depth d. We assume that the cost of the study is proportional to the cost of preparing samples for sequencing, Nc, and the total sequencing depth, Nd. (Here, c is a constant that places sequencing depth and per sample cost in the same scale).

Thus, we estimate the total cost of a sequencing study as:

$$\begin{aligned} Cost &= N \cdot \left(cost\ per\ sample\right) \\ &= N \cdot d \cdot \left(cost\ per\ depth\right) + N \cdot \left(cost\ of\ library\ and\ sample\ preparation\right) \\ &= N \cdot \left(cost\ per\ depth\right) \cdot \left(d + c\right) \\ &\propto N \cdot \left(c + d\right) \end{aligned}$$

where $N(c+d)$ is the total sequencing capacity. In this simple model, to keep total cost constant, sample size and depth must vary inversely of each other (*i.e.*, if sample size increases, coverage decreases). In our simulations, we first considered c=0. We then expanded our analyses to also consider c=5 and c=20 for total budgets of *sequencing capacity* = 50,000x, 100,000x, and 200,000x. With current genome sequencing costs of $1,000 - $3,000 per 30x genome, c=5 and

c=20 correspond to costs of ~$250 to $600 and of ~$700 - $2,000 for sample collection and preparation, respectively.

Sensitivity to Detect Singleton Variants

We first used simulations to estimate the sensitivity of singleton discovery. We ran these simulations using different read depths (d, ranging from 2x to 50x), sample sizes (N, ranging from 100 to 5,000), sequencing error rates (e, ranging from 0.001 to 0.02) and singleton discovery false positive rates (γ, 0.00001, 0.0001, 0.001, 0.01, and 0.05). For each combination of parameters, we generated 200,000 replicate samples, each with a single individual carrying the simulated singleton variant (when estimating sensitivity) or no individuals carrying the variant (when estimating false positive rates).

We assumed read depth followed a Poisson distribution. For each individual, we track the total number of sequencing reads as well as the number of reads in which a variant base was observed. For each individual, depth was generated from *Poisson(d),* and each read was generated with a sequencing error occurring with probability e. Briefly, our simulation proceeded as follows. We first sampled a total read depth for each individual and set the number of variant reads to zero. Then, for each read, we select a strand at random. Then, we increased the number of variant reads whenever the simulated read was assigned to the individual and strand that has the singleton variant (without a sequencing error, probability 1 - e) or when the simulated read was assigned to a different individual (but with a sequencing error, probability e). We then estimate the likelihood of the observed count of variant reads, conditional on depth and sequencing error rates and an allele frequency; first, assuming that all individuals match the reference so that the

number of variant reads in each individual is distributed as Binomial(Prob = *e*, Count = *depth*); next assuming that a rare variant is segregating with frequency *1/2N*, so that the number of variant reads is distributed as (1 – *1/2N*) x Binomial(Prob = *e*, Count = *depth*) + *1/2N* x Binomial(Prob = 0.5, Count = *depth*). Finally, we take the ratio of these two likelihoods. The sensitivity for detecting a variant with false positive rate γ was computed as the fraction of simulations with a simulated singleton for which the likelihood ratio was greater than the $(1-\gamma)^{th}$ percentile of null simulations using the same sample size, depth, and sequencing error rate parameters.

We validated these estimates by down-sampling on chromosome 20 from deep genome and exome samples and assessing the sensitivity to detect singletons called when all available sequence data was analyzed. Exome samples are from the NHLBI Exome Sequencing Project[6; 30] (original depth of exome sequenced regions in chromosome 20 averaging 106.80x, range 27.10–515.25x). We excluded samples with an average depth <50x to allow for down-sampling to depths 2-50x. Whole genome samples are from the Genetics and Epidemiology of Colorectal Cancer Consortium[31] (original depth averaging 35.79x, range 30.35-42.59x). For whole genome samples, we considered down-sampled depths 2-30x. In each down-sampling analysis, we sampled reads from each individual to create a new sample with the desired average depth. For instance, for an individual with an original depth of 100x, we would retain each of the original reads with a probability of 10% to achieve depth 10x.

After down-sampling, we performed variant calling using SAMtools mpileup.[32] Sensitivity for each subsample was computed as the proportion of singletons in the original deep sequence data

that were called in the down-sampled data. The false positive rate was estimated as the proportion of sites where variants were called in the down-sampled data but not in the original deep sequence data. We averaged the results of 100 replications for 100 individuals at each depth (each exome replicate examined 788,942 bases on chromosome 20 and included 126-231 singletons; each whole genome replicate examined 63,025,520 bases on chromosome 20 and included 75,269-79,849 singletons).

We chose parameter settings for computational simulations so that results closely mimicked those for analysis of down-sampled real data. We then used these values in the analysis of association study power across a broad range of sample sizes, sequencing depths, and cost models.

Power to Detect Association

We estimated association study power analytically by comparing the burden of singletons in a region between cases and controls, at significance level $\alpha = 2.5 \times 10^{-6}$, corresponding to the analysis of ~20,000 independent gene regions. Power for a two-sample t-test can be estimated using a non-central t-distribution to model test statistics as a function of sample size, the frequency of singleton variants per gene per person, the increased risk of disease conveyed by a singleton variant, and the sensitivity to detect each singleton (which is a function of sequencing depth).

Modeling this non-central t-distribution requires estimates of a non-centrality parameter λ, which describes the expected value of the statistic for a given disease model and experimental design. The requisite non-centrality parameter λ for a two-sample t-test can be expressed as:

$$\lambda = \frac{\mu_A - \mu_U}{\sqrt{\sigma_A^2/N_A + \sigma_U^2/N_U}}$$

where μ_A is the mean number of singletons per gene per person in affected individuals, and σ_A^2 is the corresponding variance. Similarly, μ_U and σ_U^2 are the mean and variance for unaffected individuals. N_A and N_U are the number of affected and unaffected individuals, respectively.

We assume that the number of singletons occurring per gene per person follows a Poisson distribution with rate parameter equal to the product of gene length (L) and frequency of singleton occurrence per site per person for cases (p_A) or controls (p_U). For a subset of simulations, we compared results of a two-sample t-test and a Wilcoxon rank-sum test. Since both gave similar results, we proceeded with the two-sample t-test.

The non-centrality parameter can be expanded as:

$$\lambda = \frac{L \cdot p_A - L \cdot p_U}{\sqrt{L \cdot p_A/N_A + L \cdot p_U/N_U}}$$

Here, p_A and p_U are the cumulative frequencies of singletons among cases and controls (for deeply sequenced samples). For low and intermediate sequencing depths, we replace these with p_A^* and p_U^*, which are the frequency of detected singletons in cases and controls at a given sequencing depth. These quantities can be defined as:

$$p_A^* = P(detect\ singleton \mid d, N, case)$$
$$= P(detect\ singleton \mid singleton, d, N) \cdot P(singleton \mid case)$$
$$+ P(detect\ singleton \mid no\ singleton, d, N) \cdot P(no\ singleton \mid case)$$

$$p_U^* = P(detect\ singleton \mid d, N, control)$$
$$= P(detect\ singleton \mid singleton, d, N) \cdot P(singleton \mid control)$$
$$+ P(detect\ singleton \mid no\ singleton, d, N) \cdot P(no\ singleton \mid control)$$

where *P(detect singleton | singleton, d, N)* is the sensitivity to detect singleton variants for a given read depth and sample size, and *P(detect singleton | no singleton, d, N)* is the corresponding false positive rate. The frequency of singleton occurrence in cases and controls can be expressed as:

$$P(singleton \mid case) = \frac{prf/L}{prf/L + (1 - p/L)f} = \frac{pr/L}{pr/L + (1 - p/L)}$$

$$P(singleton \mid control) = \frac{p/L(rp/L + 1 - p/L - rf)}{(1 - f)(rp/L + 1 - p/L)}$$

where r is relative risk of disease (ranging from 2 to 20 in our simulations), p is population frequency of singletons (ranging from 0.001% to 1% per gene per person in our simulations), L is gene length (ranging from 1,000 to 50,000 bps in our simulations), and f is the background prevalence of disease (ranging from 0.1% to 20% in our simulations). We considered different case-control ratios (1:1, 2:1, 3:1, and 1:2 in our simulations).

We considered prevalences from 0.1% to 20% to explore scenarios for studying different diseases. Such diseases include very complex diseases such as cardiovascular disease, which has a prevalence of 33% in American adults[33], intermediate frequency diseases such as age-related macular degeneration, which has an estimated prevalence of 1.47% in Americans 40 years and

older[34], but also less common diseases such as type 1 diabetes, which has an approximate prevalence of 0.33% in Americans 18 years and younger[35].

From our exome samples, we estimated that singletons occur at a rate of 0.79% per gene per person for the entire genome (assuming approximately 20,000 genes). This rate fluctuates when looking at specific genes. Longer genes are more likely to include singletons than shorter genes. For instance, *AVP* (a gene that provides instructions for making the hormone vasopressin) is 2,169 bp long[36] and has an estimated frequency of singletons of 0.058% per person in the coding region, while *TTN* (encoding a giant muscle protein that plays a key role in muscle assembly and one of the longest genes in the human genome) is 281,435 bp long[37] and has a singleton frequency of 42% per person in the coding region. Genes of different functions might have different frequencies of singletons. We varied population frequency of singletons from 0.001% to 1% per gene per person to account for this wide set of possibilities.

Once the non-centrality parameter is computed, the power of an association test can be estimated by:

$$Power = P\left(|X| > t_{\alpha/2}(\nu, 0) \mid X \sim t(\nu, \lambda)\right)$$

where

$$t_{\alpha/2}(\nu, 0) = 100(1-\alpha/2)\ \textit{percentile of the central } t \textit{ with } \nu \textit{ degrees of freedom}$$

$$t(\nu, \lambda) = \textit{Non-central } t \textit{ with } \nu \textit{ degrees of freedom and non-centrality parameter } \lambda$$

Results

Sensitivity to Detect Singletons

Our simulations show that, for a fixed sample size, sensitivity to detect singletons increases rapidly as coverage increases until ~25x (see Figure II-1a). After this point, increasing coverage has little effect on sensitivity. As sample size increases for a fixed depth, sensitivity decreases only slightly (see Figure II-1b), implying that coverage at a site has more impact than sample size in the overall ability to detect singletons. For constant depth and sample size, an increase in sequencing error rate reduces sensitivity (see Figure II-1c). At higher false positive rates, sensitivity is greater (see Figure II-1), although the number of incorrectly called singletons increases as well. Among the settings we considered, by 25x, sensitivity reaches 98.6% for a sequencing error rate of 0.005, 96.2% for an error rate of 0.01, and 89.9% for a sequencing error rate of 0.02, regardless of sample size or false positive rate. Increasing depth to 30x resulted in sensitivity of 99.6% for a sequencing error rate of 0.005, 98.8% for an error rate of 0.01, and 95.9% for an error rate of 0.02. Further increase in depth to 50x, resulted in 100% sensitivity, regardless of error rate.

As shown in Figure II-1, variant detection sensitivity changes rapidly with read depth but only very slowly with sample size (sensitivity decreases slightly with increased sample size because, when depth and total false positive rate are fixed, the caller must become gradually more stringent as more samples are sequenced so as to maintain a fixed false-positive rate). We next explored variant discovery power in experiments with constant cost, where sample size and read depth vary in opposite directions. We first considered a simplified case with no additional cost for library and sample preparation, so that read depth and sample size are inversely proportional.

In this case, as coverage increased, sensitivity increased until 20-25x, after which increasing read depth had little effect on sensitivity (see Figure II-2). When we varied the total sequencing capacity, there was little difference between the sensitivity to detect singletons at a fixed read depth, emphasizing that read depth has more effect than sample size on sensitivity. For instance, at 10x coverage, sequencing 5,000 samples provides 64% sensitivity, sequencing 10,000 samples provides 60.8% sensitivity, and sequencing 20,000 samples provides 57.9% sensitivity; whereas, at 20x coverage, sequencing 5,000 samples provides 97.4% sensitivity, sequencing 10,000 samples provides 96.8% sensitivity, and sequencing 20,000 samples provides 96.1% sensitivity.

Analyses of down-sampled data validated our computational simulations (see Figure II-3). For a fixed sample size of 100 individuals, empirical estimates of sensitivity closely resemble simulations that assume a sequencing error rate of 0.01 and a false positive rate of 0.001, though the simulations were slightly conservative at lower depth – detecting a lower proportion of variants than in the down-sampled data – and slightly anti-conservative at higher depth – detecting a larger fraction of variants than in the down-sampled data with the same total depth (see Figure II-3). We thus used $e = 0.01$ and $\gamma = 0.001$ to estimate variant detection sensitivity and assess association study power for a broad range of cost models and sequencing capacities.

Power to Detect Association

We first considered a situation of fixed cost (sample size and read depth vary inversely) with no extra cost of library/sample preparation (c=0) for equal numbers of affected and unaffected individuals. As depth increases, association study power quickly reaches a maximum and then rapidly decreases (see Figure II-4). For example, sequencing 20,000 samples at 5x provides only

1.19% power, sequencing 6,666 samples at 15x provides 91.08% power, and sequencing 2,000 samples at 50x provides 17.71% power for a relative risk of 15, population frequency of singletons 0.01 per person per gene, and a prevalence of 20%. Maximum power increases with relative risk, population frequency of singletons, or prevalence (see Figure II-4a-c). For unequal numbers of cases and controls, power decreases as the case-control ratio moves further away from 1:1 (see Figure II-4d).

When the NCP is large, a change in the NCP might not be reflected in power if the power is already 1. Therefore, we examined the read depth/sample size pair where the maximum NCP, rather than power, was attained. The depth at which the NCP is maximized occurs between 15-20x depending on study cost and the relative cost of library/sample preparation. As available sequencing capacity and total study cost increase, the maximum NCP increases (see Figure II-5). As relative cost of library/sample preparation (c) increases, NCP decreases slightly (see Figure II-6). When either total study cost or c increases, the point at which NCP is maximized shifts to a higher depth. For c=0, NCP is maximized at 15-16x; for c=5, NCP is maximized at 16-18x; and for c=20, NCP is maximized at 18-19x. For sequencing capacity=50,000x, NCP is maximized at 15-18x; for sequencing capacity=100,000x, NCP is maximized at 16-18x; and, for sequencing capacity=200,000x, NCP is maximized at 16-19x. This point is not affected much by relative risk, prevalence, population frequency of singletons, or gene length. The overall pattern is easy to understand intuitively: with increasing per sample preparation costs, it is advantageous to sequence fewer samples at higher depth; with increasing total sequencing capacity, the overall sample size increases and a slight increase in sequencing depth is needed to accommodate the greater stringency needed to maintain low false positive rates in variant calling.

For fixed sample size, increasing coverage is never harmful. At lower depth, increasing coverage increases association study power; at high depths, power eventually plateaus (see Figure II-7). For instance, for relative risk of 15, population frequency of singletons 0.008, and prevalence 0.2, sequencing 10,000 samples provides 30.45% power at 10x, 97.99% power at 25x, 98.27% power at 35x, and 98.27% power at 50x. For increased sample size, relative risk, population frequency of singletons, or prevalence, the magnitude of power increases. Regardless of the parameter values, NCP is maximized by 35x, with 99% of maximal NCP occurring by 25x.

Discussion

We set out to identify ideal sequencing strategies, in terms of read depth and sample size, focusing on studies exploring the association of singleton variants and discrete traits. We found that association study power is never large unless frequency of singletons or relative risk is large. When cost is fixed so sample size varies inversely of read depth, power decreased as coverage increased beyond 15-20x. Even for fixed sample size, increasing coverage beyond 25x had only a small impact on power. Therefore, we believe it will often be better to sequence larger samples at lower coverage rather than smaller samples at increased coverage when searching for disease associated singletons. We recommend that coverage should only be increased beyond 20x if sample numbers are limited or if applications other than genetic association studies (such as genetic counseling and diagnosis) can justify the advantages of more complete sequencing of each individual at the cost of reduced sample sizes.

While varying prevalence, singleton frequency, and relative risk varies association study power, the combination of read depth and sample size combination that maximizes power (for a fixed cost) remained constant. For example, consider two scenarios. The first has a relative risk of 10, a prevalence of 20%, and a population frequency of singletons of 0.8%. The other has a relative risk of 12.5, a prevalence of 20%, and a population frequency of singletons of 1%. Both scenarios have a sequencing capacity of 200,000x with intermediate library and sample preparation costs (c=5) and a gene length of 1000 bp. The first scenario attains a maximum power of 19.87% (NCP = 3.86); the second reaches a maximum power of 92.07% (NCP = 6.12). However, in both cases, NCP is maximized at a depth of 18x (with 8,695 samples). This suggests that deep sequencing is not justified for detecting rare variant association, irrespective of the disease or gene under study. While power may be low, increasing coverage beyond a threshold at 15-20x will not increase power if it requires a decrease in sample size. For more common variants, there is likely even higher sensitivity for detecting variants at lower depths, so we expect association study power will be maximized at even lower coverage.[3]

Of the parameters we considered, only the cost of library and sample preparation changed the depth required to maximize association study power, though this optimal depth remained between 15-20x. For larger library and sample preparation costs, the optimal depth increases slightly. For no extra cost of library/sample preparation (c=0), the ideal depth is 15-16x; for a moderate cost of library/sample preparation (c=5), the ideal depth is 16-18x; and when the cost of library/sample preparation is high (c=20), the ideal depth is 18-19x. When there is an increased penalty for increasing sample size rather than increasing coverage in the current sample, a slightly deeper coverage is ideal to compensate for this penalty. For very large studies,

the ideal depth shifts to slightly larger depths. For example, when the total sequencing capacity is 50,000x, the ideal depth is 15-18x per genome; when this increases to 100,000x, the ideal depth is 16-18x per genome; and, for sequencing capacity of 200,000x, the ideal depth is 16-19x per genome. This increase in per genome depth allows variant calling to become more stringent as sample size increases (there are more opportunities for false positive calls as more genomes are sequenced).

In summary, we have shown that, while deep sequencing is appealing for detecting a complete catalog of variants, increasing sample size at the cost of lower depth of coverage per sample results in higher power for association studies. An avenue of future work is to conduct a similar analysis for quantitative traits.[29] Additionally, our simulations focus on SNPs, but other types of variants such as insertions and deletions are of interest. These non-SNP variants are typically more difficult to detect than SNPs,[38] and may justify higher coverage to enable accurate genotyping followed by detection of association.

Figures and Tables

Figure II-1. Sensitivity to detect singletons by read depth, sample size, and sequencing error rate

Sensitivity vs. (a) read depth for N=1000 and e=0.01, (b) sample size for d=10x and e=0.01, and (c) sequencing error rate for N=1000 and d=10x at different false positive rates.

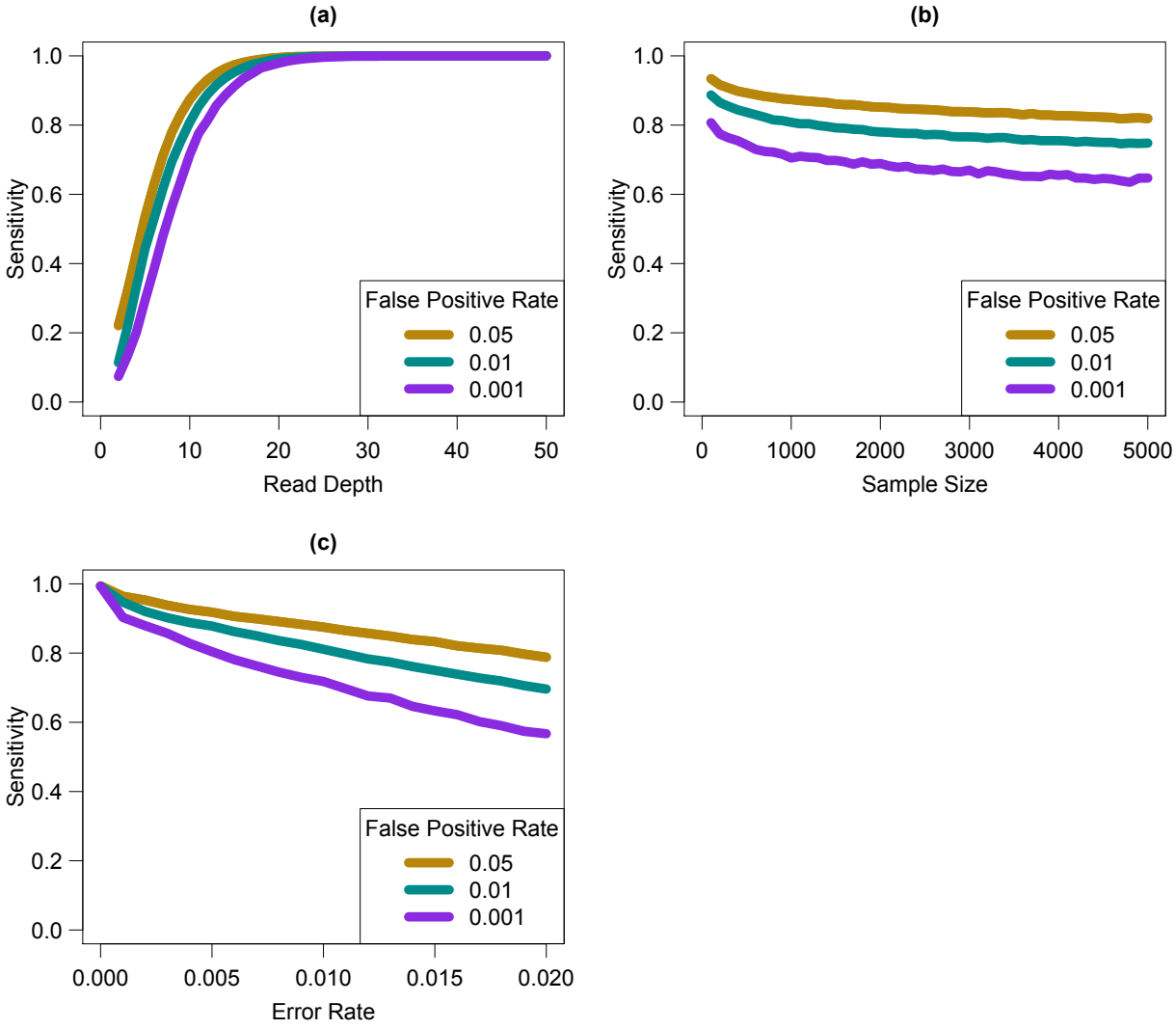


Figure II-2. Sensitivity to detect singletons by read depth for constant cost
Comparing computational simulations (for a sequencing capacity of 50,000x) for sensitivity to detect singletons for different false positive rates shows that power increases until 25-30x, exact threshold increasing with increased error rate or decreased false positive rate. (Sample size = cost/depth, assuming no cost of library/sample preparation, c=0)

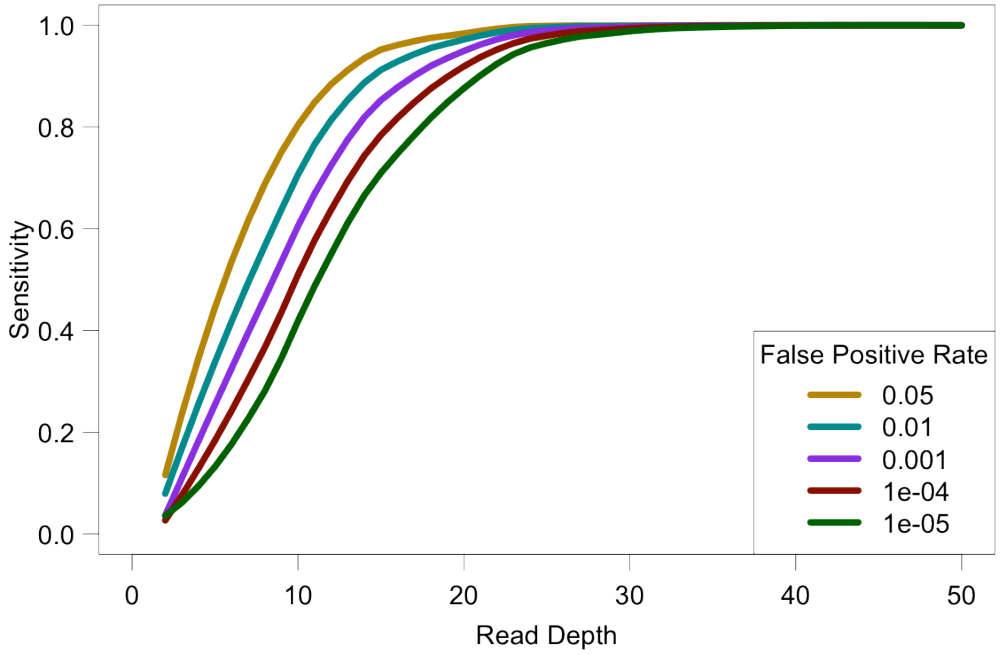


Figure II-3. Comparison of empirical sensitivity to detect singletons with computational estimates

For a sample size of 100, sequencing error rate of 0.01 with a false positive rate of 0.001, empirical and computational estimates are similar.

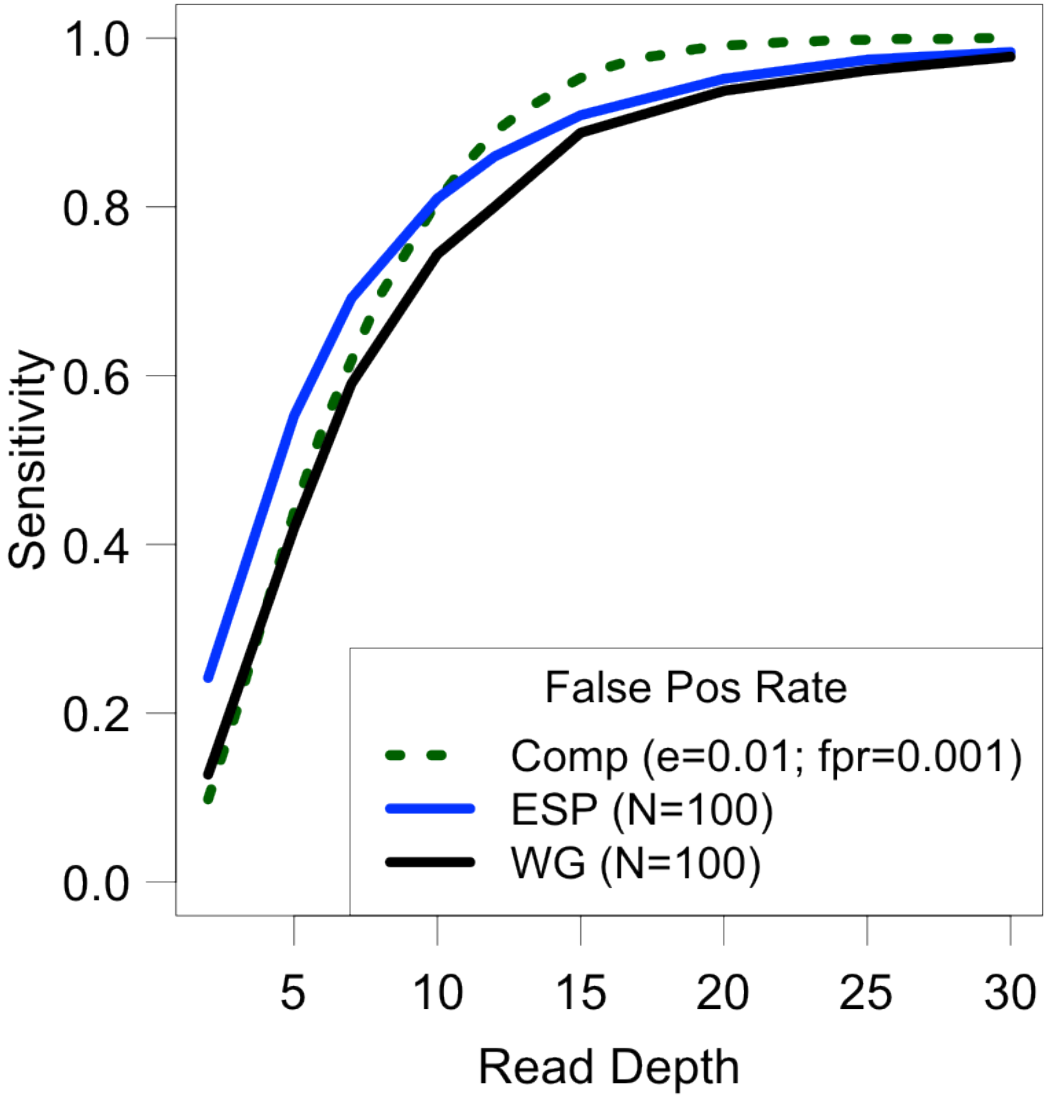


Figure II-4. Association study power by read depth for constant cost
Power of an association study increases with relative risk (a), population frequency of singletons (b), prevalence (c), and ratio of cases to controls (d) for a fixed sequencing capacity of 100,000x with no extra cost for library/sample preparation (c=0).

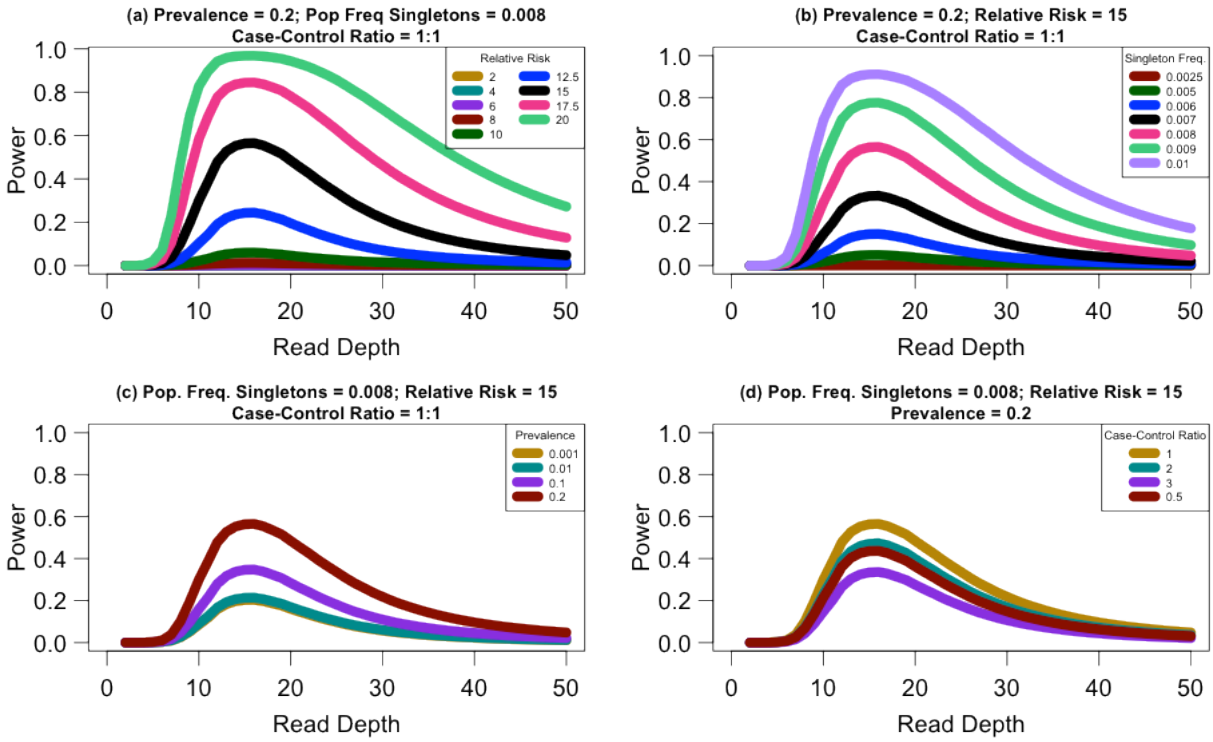


Figure II-5. Association study power and NCP by read depth for constant different sequencing capacities

Library/sample preparation costs low (c=0), relative risk 15, population frequency of singletons 0.01, prevalence 20%, case-control ratio 1:1

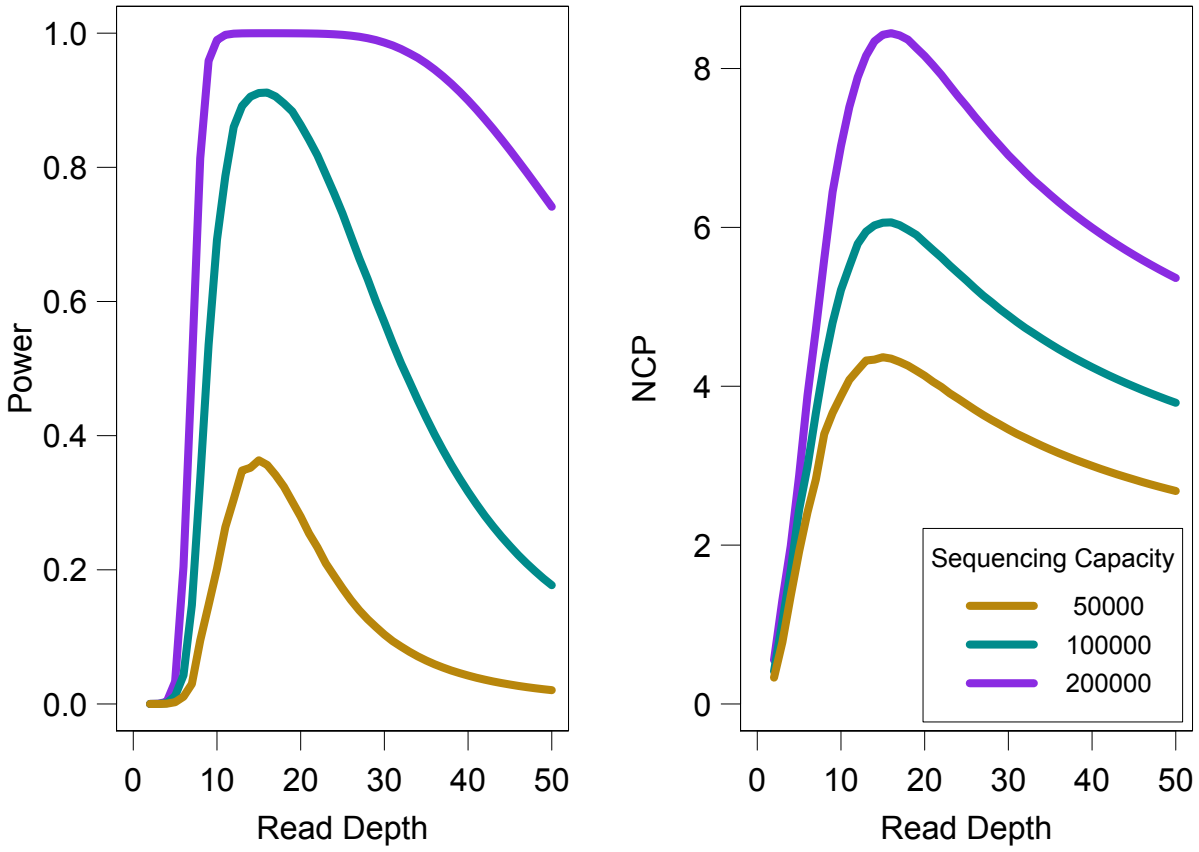


Figure II-6. Association study power and NCP by read depth for different sample preparation costs

Sequencing capacity 100,000x, relative risk 15, population frequency of singletons 0.008, prevalence 20%, case-control ratio 1:1

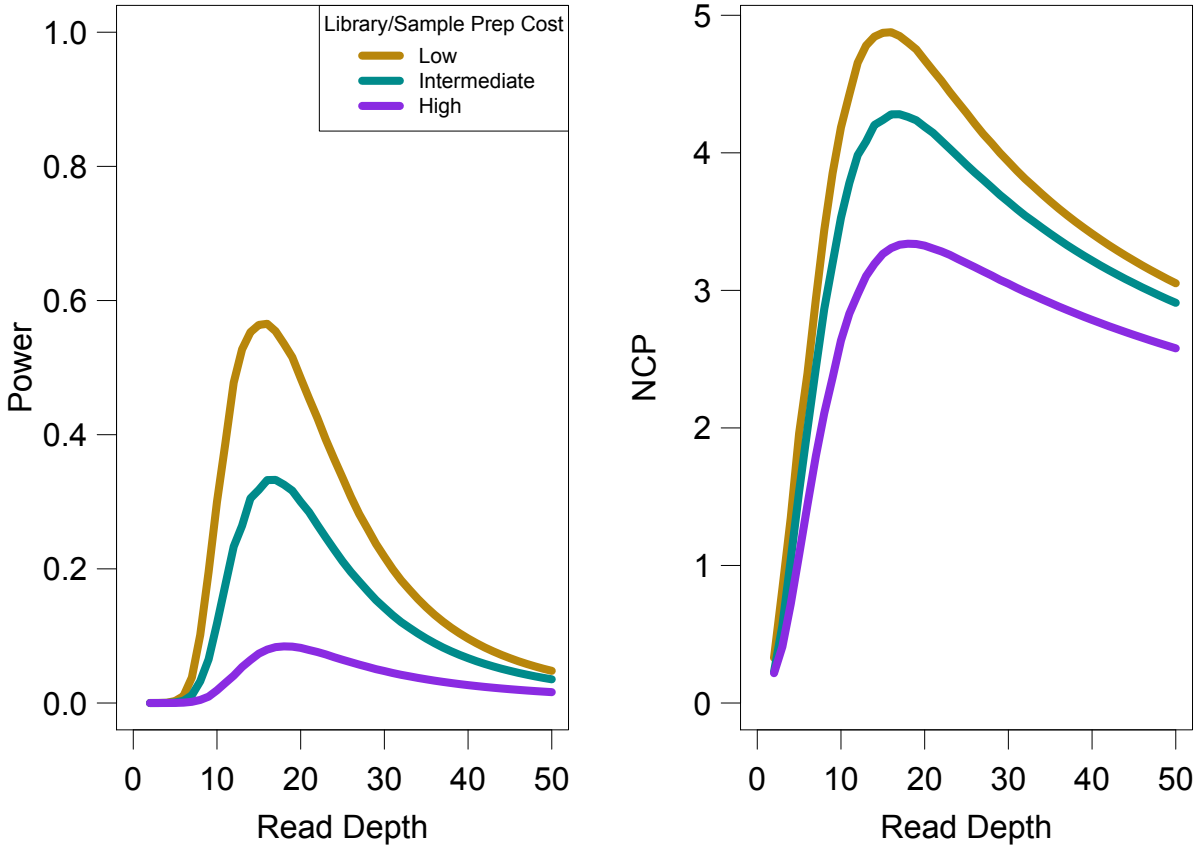
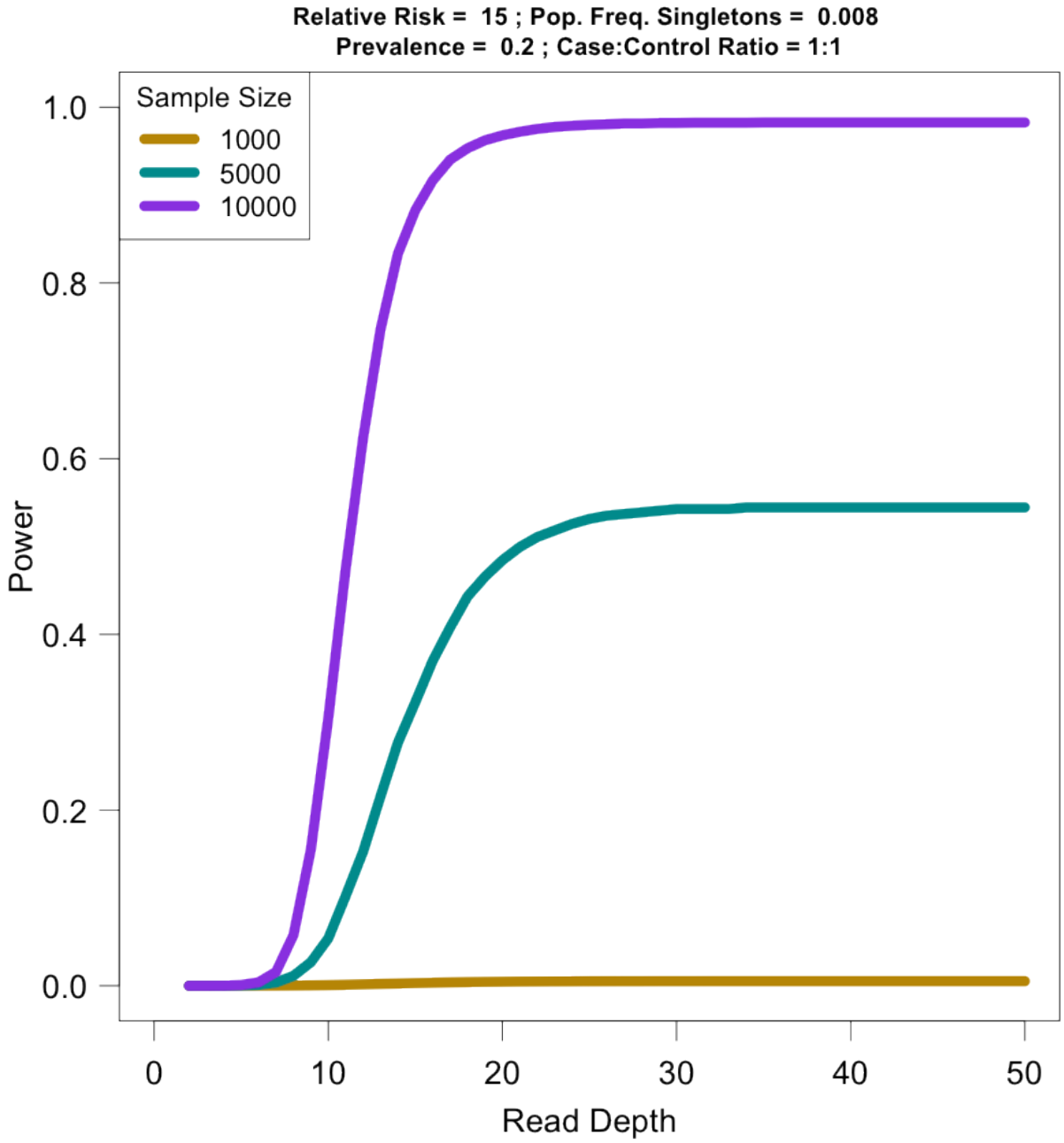


Figure II-7. Association study power by read depth for fixed sample size
Increasing coverage beyond a threshold does not increase power of an association study for constant sample size.



CHAPTER III. Optimal Sequencing Strategies for Identifying Singletons Associated with Quantitative Traits[2]

Abstract

A major focus of modern genetic association studies is the effect of rare variation on quantitative traits, so it is important to determine cost-effective sequencing strategies for these studies. Deep sequencing allows for accurate detection and genotyping of all variants per individual, but may restrict sample size. The reduced cost of low-pass sequencing allows for increased sample sizes (given a fixed total cost), which is beneficial in association studies, but may miss some variants. Here, we explore the effect of sample size and depth on studies of rare variants associated with quantitative traits, focusing on singleton variants that appear in a single individual in a sample and are the most difficult type of variant to detect. We examined the sensitivity to detect singletons in simulated data and down-sampled deep genome and exome data. We extended this to examine association study power comparing the burden of singletons in a gene to the trait value. We show that when sequencing capacity is fixed, power of association studies focused on singletons is typically maximized for 11-16x coverage, independent of frequency of singletons per gene per person and effect size. Our results suggest a sequencing depth of 11-16x as an appropriate compromise of sample size and sequencing depth for studies of rare variants and quantitative traits.

---

[2] Rashkin, S., Jun, G., and Abecasis, G. (In preparation). Optimal Sequencing Strategies for Identifying Singletons Associated with Quantitative Traits.

Introduction

Quantitative traits have become an area of intense focus in recent years.[16] Examples of quantitative traits include height, cholesterol levels, and blood glucose levels. Frequently, these can be used as markers for a disease, and they can also aid in understanding the basic mechanism of disease.[16] Knowing how variants affect a quantitative trait value can be informative about the biological mechanism by which that variant affects disease, rather than simply knowing that a variant increases or decreases disease susceptibility.

As with binary traits, there is still much we do not know about the biological basis of quantitative traits.[4; 16] Focusing entirely on common single nucleotide polymorphisms (SNPs) is insufficient for understanding the biology of quantitative trait variation.[2; 4; 8; 16] Rare variants are more likely to disrupt gene function and might, therefore, explain some of missing heritability of common, complex diseases and increase our mechanistic understanding of disease.[4-7] With the advent of better sequencing technologies, allowing for reduced costs of sequencing studies, it is possible to include rare variants in genetic association studies.

A typical study design for examining rare variants relies on deep sequencing to provide the highest power for variant discovery.[4; 10] However, deep sequencing increases the cost of sequencing a single sample, limiting the number of individuals for a sequencing study (given fixed resources), which typically reduces the power to detect association. Lower coverage can reduce sequencing costs but may result in missing many rare variants.[3; 39] Recent studies have shown that rare variants can be detected even with low coverage sequencing if the sample size is

large.[3; 4; 14] Examples of such studies have identified rare variants including an LDLR variant that is associated with low-density lipoprotein cholesterol levels[40] and several rare and low-frequency variants associated with hematological traits[41]. These results imply that a detailed comparison of sequencing study designs for the specific purpose of studying rare variants associated with quantitative traits is useful for designing modern genetic studies.

Here, we compare the power of sequencing based rare variant association studies using low, intermediate, and deep sequencing strategies. Since singleton variants that appear only once in a sample are the most difficult type of variant to detect, designs with power to detect associations between them and a trait value should have power to detect associations with a trait for more common variants. Any sequencing depth that works well for singletons should provide an upper bound of needed sequencing depth for more common variants, for which low pass data can be analyzed more effectively.[3; 15] Our goal was to determine optimal sequencing strategies that maximize power for a fixed sequencing effort for studies exploring the role of singletons in quantitative traits (we conducted a similar analysis for binary traits in a companion paper[42]).

We examined the balance between sensitivity to detect singletons (increases with coverage) and power to detect association (increases with sample size) with analyses using simulated data and actual sequencing data. We explored the power of association tests for study designs implementing low, intermediate, and deep sequencing strategies across a range of effect sizes and frequencies of singletons. Our results show that, for fixed sequencing capacity, power to detect association is maximized at 11-16x and decreases as coverage is increased beyond this threshold.

Methods

Overview

To determine the optimal study design for studying singletons associated with quantitative traits, we conducted several simulations under a variety of parameter values. We approximated sensitivity to detect singletons by computational simulations validated by simulations using real data. We then extended these results to examine how association study power varies under different study designs in terms of sample size and read depth.

Definitions

We consider sequencing studies that assess N individuals, sequencing each to an average depth d. We assume that the cost of the study is proportional to the cost of preparing samples for sequencing, Nc, and the total sequencing depth, Nd. (Here, c is a constant that places sequencing depth and per sample cost in the same scale).

Thus, we estimate the total cost of a sequencing study as:

$$\begin{aligned} Cost &= N \cdot \left(cost\ per\ sample\right) \\ &= N \cdot d \cdot \left(cost\ per\ depth\right) + N \cdot \left(cost\ of\ library\ and\ sample\ preparation\right) \\ &= N \cdot \left(cost\ per\ depth\right) \cdot \left(d + c\right) \\ &\propto N \cdot \left(c + d\right) \end{aligned}$$

where $N(c+d)$ is the total sequencing capacity. In this simple model, to keep total cost constant, sample size and depth must vary inversely of each other (*i.e.*, if sample size increases, coverage decreases). In our simulations, we first considered c=0. We then expanded our analyses to also consider c=5 and c=20 for total budgets of *sequencing capacity* = 50,000x, 100,000x, and

200,000x. With current genome sequencing costs of $1,000 - $3,000 per 30x genome, c=5 and c=20 correspond to costs of ~$250 to $600 and of ~$700 - $2,000 for sample collection and preparation, respectively.

Sensitivity to Detect Singleton Variants

We first implemented computational simulations to estimate the sensitivity of singleton discovery. For each combination of depth, sample size, error rate, and false positive rate, we used a simulated multi-sample caller that calculated the power of detecting a singleton at a site from the likelihood of calling the variant across 200,000 simulations. We validated these results using simulations from real-data, where we down-sampled from deeply sequenced genome and exome samples and, for each sample size and read depth combination, assessed our ability to detect previously called singletons. For a more detailed explanation of this part of our analysis, refer to our companion paper for binary traits.[42] We chose parameter settings for our computational simulations where results closely mimicked empirical estimates and used these results for our analysis of association study power across a broad range of sample sizes, depths, and cost models.

Power to Detect Association

We assumed a model where the trait value is dependent on burden of singletons in a region, at significance level alpha = 2.5×10^{-6}, corresponding to the analysis of ~20,000 independent gene regions. We assumed the data follow a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N\left(0, \sigma_e^2\right)$$

where, for the i^{th} individual, y_i is the phenotype value, x_i is the number of singletons in a gene, and ε_i is the error term. In this model, β_0 represents the average trait value for an individual with no singleton variants in the gene and β_1 represents the effect size, the change in the trait value for each additional singleton present in the gene for an individual. Since the power of association is dependent solely on the slope parameter (β_1), we fixed β_0 as 0 for our data generation. We allowed β_1 to vary from 0.5 to 3.

We modeled X following a *Poisson(p)* distribution, where the rate p is the population frequency of singletons per person per gene. In our companion paper for a similar analysis for binary traits, we estimated that singletons occur at a rate of 0.0079 per gene per person across the whole genome, assuming 20,000 genes.[42] However, this rate may vary when looking at specific genes – it is reasonable to assume that longer genes may have higher rates of occurrence. We, therefore, considered rates from 0.001 to 0.01 per gene per person to account for most situations.

The trait value given the number of singletons in a gene ($Y_i|X_i$) follows a $N(\beta_0+\beta_1 x_i, \sigma_e^2)$ distribution. For ease of simulations, we assumed that Y is standardized so the total phenotypic variance, σ_p^2, is 1. The variance of the error term can be calculated as:

$$\sigma_p^2 = Var(Y) = Var(\beta_0 + \beta_1 X + \varepsilon) = \beta_1^2 \cdot Var(X) + Var(\varepsilon)$$
$$\Rightarrow 1 = \beta_1^2 p + \sigma_e^2$$
$$\Rightarrow \sigma_e^2 = 1 - \beta_1^2 p$$

For each simulation, we set the sample size (N), read depth (d), population frequency of singletons (p), and effect size (β_1). For each individual, we sampled $X_i \sim Poisson(p)$ and $Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, 1-(\beta_1)^2 p)$. To account for uncertainty regarding the true number of singletons,

we adjusted X_i to determine the number of singletons detected for the i^{th} individual, Z_i. We first generated the number of singletons present for the i^{th} individual that are correctly detected $T_i \sim Bin(X_i, s)$, where s is the sensitivity for detecting singletons given the depth and sample size. We also generated the number of false positive singletons for the i^{th} individual, $F_i \sim Poisson(\gamma/N)$, where γ is the false positive rate for detecting singletons. We adjusted by N to approximate the number of false positives in a single individual rather than across the sample. Z_i is the sum of these quantities.

Using the generated data, we fit a simple linear regression on Y using Z and assessed β_1 for significance. We repeated this for 5000 replicates. Power was estimated as the fraction of simulations where β_1 was significant out of the total number of replicates.

Our goal was to identify the optimal study design in terms of sample size and depth of coverage, so we were interested in determining the depth at which power is maximized. Since power cannot increase beyond 100%, it is sometimes impossible to distinguish which of several depths results in the largest power if all have near 100% power. Therefore, we calculated the non-centrality parameter for a linear regression F-test[43]:

$$\lambda = \frac{\beta_1^2 \sum_{i=1}^{n} (z_i - \bar{z})^2}{\sigma_{Ze}^2}$$

where $\sigma^2_{Ze} = 1-(\beta_1)^2 Var(Z) = 1-(\beta_1)^2[Var(T)+Var(F)] = 1-(\beta_1)^2[sp+\gamma/N]$. As we are only interested in the slope term, we can use the non-centrality parameter for a t-test, which is the square root of the above term.

Results

Sensitivity to Detect Singletons

Our simulations show that sensitivity to detect singletons increases as coverage increases, until ~25x, beyond which increasing depth further has little effect on power. This pattern holds for both fixed sample size and fixed sequencing capacity (where sample size decreases as coverage increases). For a fixed sample size of 100 individuals, empirical estimates resemble simulation estimates for a sequencing error rate of e=0.01 with a false positive rate of γ=0.001, though computational sensitivity estimates are lower than estimates using real data for lower depths and higher than estimates using real data at higher depths. For a more detailed explanation of this part of our analysis, refer to our similar analysis for binary traits.[42]

Power to Detect Association

We first considered a scenario with fixed cost (sample size and depth vary inversely of each other) with no extra cost of library/sample preparation (c=0). As depth increases, power increases until a threshold is reached. Beyond this depth threshold, a further increase in coverage results in a decrease in power. For example, sequencing 20,000 samples at 5x provides 42.94% power, sequencing 8,333 samples at 12x provides 80.74% power, and sequencing 2,000 samples at 50x provides only 7.52% power for population frequency of singletons 0.008 and effect size 0.8. The magnitude of the power to detect association depends on the combination of the effect size (β_1) and the population frequency of singletons (p).

As the effect size increases, power increases, holding other parameters constant (see Figure III-1). For example, for a sample size of 8,333 at 12x coverage with a population frequency of

singletons of 0.008 and an effect size of 0.5, the power of an association study is 13.90%. When the effect size is increased to 0.7, the power increases to 59.14%, and, when the effect size is further increased to 1, the power increases to 98.02%.

Similarly, as the population frequency of singletons increases, power also increases, holding other parameters constant (see Figure III-2). For example, sequencing 8,333 samples at 12x coverage with an effect size of 1 and a population frequency of singletons of 0.001 per gene per person provides 2.00% power. When frequency of singletons increases to 0.005, power increases to 77.08%, and, when singleton frequency is increased further to 0.01, power increases to 99.84%.

Power calculated through the average non-centrality parameter across replicates follows a similar pattern as empirical power (see Figure III-3). For a sequencing capacity of 100,000x, the difference in power estimates has a mean of 0.48% (with a standard deviation of 2.76%). We determined the depth at which the non-centrality parameter is maximized for each combination of parameters. As sequencing capacity increases, maximum NCP increases (see Figure III-4). As relative cost of library and sample preparation increases, maximum NCP decreases (see Figure III-5). When either total sequencing capacity or c increases, the point at which NCP is maximized shifts to a higher depth. For c=0, NCP is maximized at 11-12x; for c=5, NCP is maximized at 12-14x; and for c=20, NCP is maximized at 15-16x. With increasing per sample preparation costs, it is advantageous to sequence fewer individuals at higher depth. For sequencing capacity=50,000x, NCP is maximized at 11-15x; for sequencing capacity=100,000x, NCP is maximized at 12-15x; and, for sequencing capacity=200,000x, NCP is maximized at 12-

16x. With increasing total sequencing capacity, the overall sample size increases and a slight increase in sequencing depth is needed to accommodate the greater stringency needed to maintain low false positive rates in variant calling. This point does not vary much with effect size or population frequency of singletons.

For fixed sample size, increasing depth does not always result in an increase in power (see Figure III-6). At low coverage, increasing depth increases power substantively. However, after a threshold is reached, power remains relatively constant. For instance, for an effect size of 0.7 and a population frequency of singletons of 0.008, sequencing 10,000 samples provides 57.62% power at 10x, 92.72% power at 25x, 93.36% power at 35x, and 93.36% power at 50x. Regardless of the parameter values examined, NCP is maximized by 35x, with 99% of maximal NCP occurring by 24x.

Discussion

We set out to compare low, intermediate, and deep sequencing strategies for sequencing studies exploring the association of singleton variants with quantitative traits. We found that power is never large unless effect size or frequency of singletons is large. If cost is fixed so sample size and depth vary inversely, power decreases as depth increases beyond 11-16x. Even for fixed sample size, increasing coverage beyond 24x has little effect on power. Therefore, we believe it will often be better to sequence larger samples at lower coverage rather than increasing coverage in smaller samples when searching for singletons associated with quantitative traits. We recommend increasing coverage beyond 16x only if sample size is restricted or reasons other

than genetic association studies can justify the advantages of sacrificing increased sample size in lieu of more complete sequencing of each individual.

While varying effect size and population frequency of singletons varies the power to detect an association, the read depth and sample size combination that maximizes power (for a fixed cost) remains constant. Consider two scenarios. In the first, we have an effect size of 0.7 and a population frequency of singletons of 0.007. In the second, we have an effect size of 1 and a population frequency of singletons of 0.009. In both situations, assume a total sequencing capacity of 100,000x with intermediate library/sample preparation costs (c=5). The first scenario attains a maximum power of 20.67% (NCP = 3.89); the second reaches a maximum power of 94.76% (NCP = 6.34). Both situations attain maximal power at a depth of 14x (with 5,263 samples). This suggests that deep sequencing is not justified for detecting rare variant association, irrespective of trait or gene under study. While power may be low, increasing coverage beyond a threshold at 11-16x will not increase power if it requires a decrease in sample size.

Of the parameters we considered, only the cost of library and sample preparation changed the depth required to maximize association study power, though this optimal depth remained between 11-16x. For larger library and sample preparation costs, the optimal depth increases slightly. For no extra cost of library/sample preparation (c=0), the ideal depth is 11-12x; for a moderate cost of library/sample preparation (c=5), the ideal depth is 12-14x; and when the cost of library/sample preparation is high (c=20), the ideal depth is 15-16x. When there is an increased penalty for increasing sample size rather than increasing coverage in the current

sample, a slightly deeper coverage is ideal to compensate for this penalty. For very large studies, the ideal depth shifts to slightly larger depths. For example, when the total sequencing capacity is 50,000x, the ideal depth is 11-15x per genome; when this increases to 100,000x, the ideal depth is 12-15x per genome; and, for sequencing capacity of 200,000x, the ideal depth is 12-16x per genome. This increase in per genome depth allows variant calling to become more stringent as sample size increases (there are more opportunities for false positive calls as more genomes are sequenced).

We explored effect sizes ranging from 0.5 to 3 in order to assess power in a variety of situations. Our analysis shows that, for really rare mutations, there is often insufficient power to be able to successfully detect an association between the number of singletons in a gene and a trait value for small to moderate effect sizes and a low frequency of singletons in a gene. It is only possible to detect an association if the effect size is large or the population frequency of singletons in a gene is large. We found that, while association study power increases as effect size increases, the depth at which power is maximized is unaffected by effect size modeled.

In summary, despite the benefits of deep sequencing in locating more rare variants per individual, a preferable design for identifying very rare variants associated with quantitative traits is to increase sample size at lower depths. We conduct a similar analysis for binary traits in a companion paper.[42] Our simulations focused on SNPs, although it is important to consider other types of variants such as insertions and deletions to further our understanding of the cause of disease. Since these non-SNP variants are, typically, more difficult to detect than SNPs, a

greater depth may be required to maximize the power of association.[38] A greater sequencing effort may be required to increase depth of coverage without decreasing sample size.

Figures and Tables

Figure III-1. Association study power by read depth for constant cost for different effect sizes

Fixed sequencing capacity of 100,000x with no extra cost of library/sample preparation (c=0) and population frequency of singletons 0.008

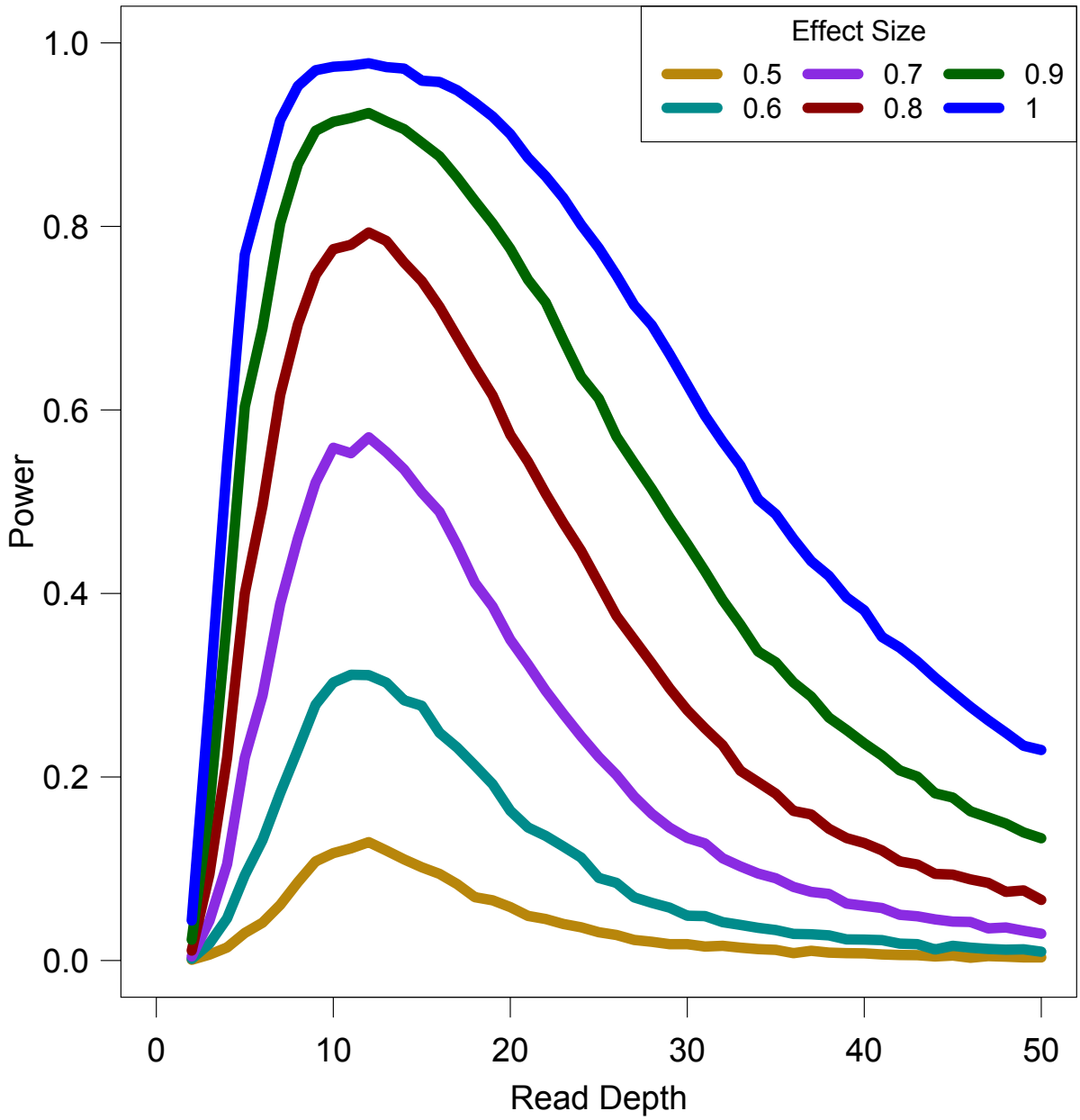


Figure III-2. Association study power by read depth for constant cost for varying population frequency of singletons
Fixed sequencing capacity of 100,000x with no extra cost of library/sample preparation (c=0) and effect size 1.

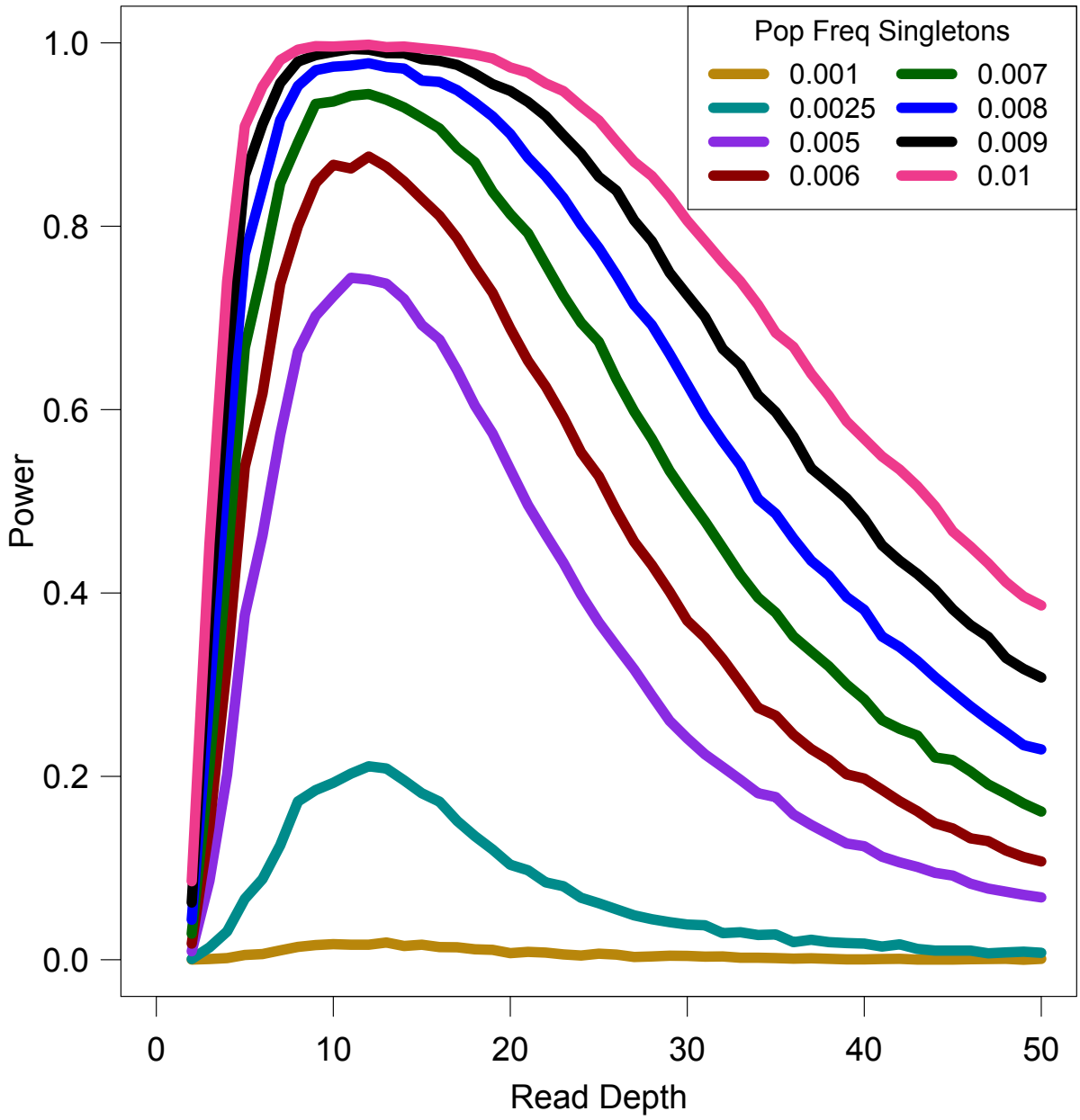


Figure III-3. Empirical power by read depth for fixed sequencing effort compared with power calculated via NCP

Fixed sequencing capacity of 100,000x with no extra cost of library/sample preparation (c=0), effect size 1, and population frequency of singletons 0.008.

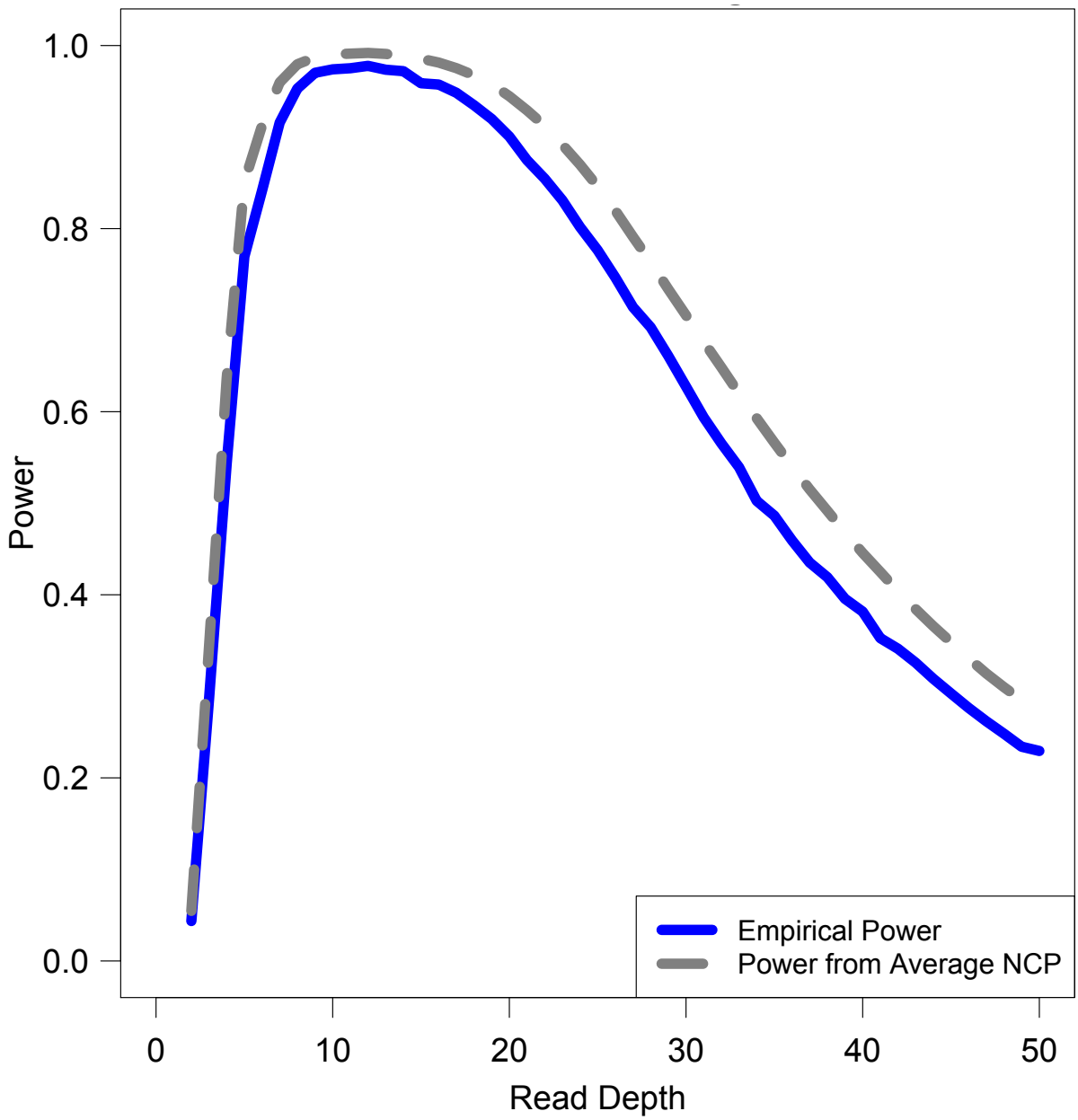


Figure III-4. Association study power and NCP by read depth for different total costs
Effect size =1; population frequency of singletons = 0.008; no library/sample preparation cost (c=0).

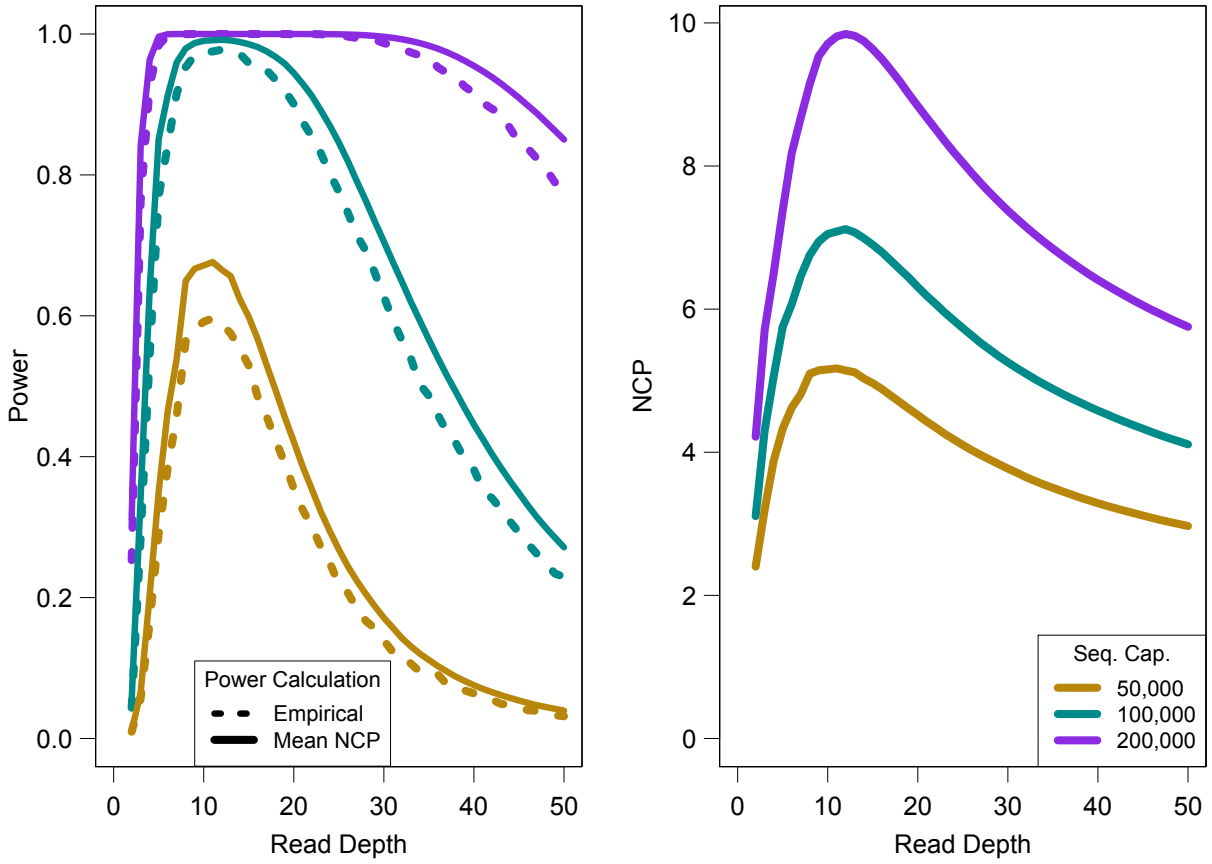


Figure III-5. Association study power and NCP by read depth for different library/sample preparation costs

Sequencing capacity = 100,000x; effect size = 1; population frequency of singletons = 0.008.

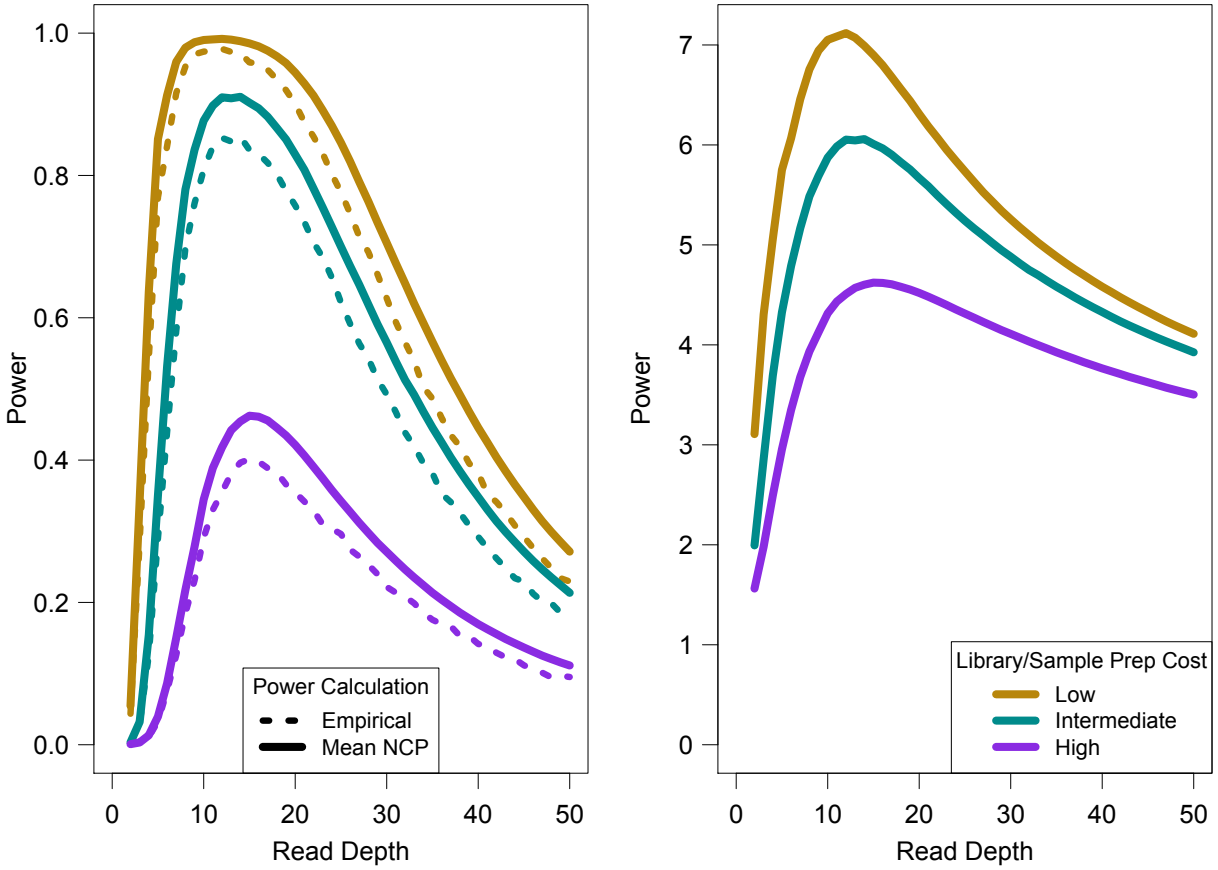
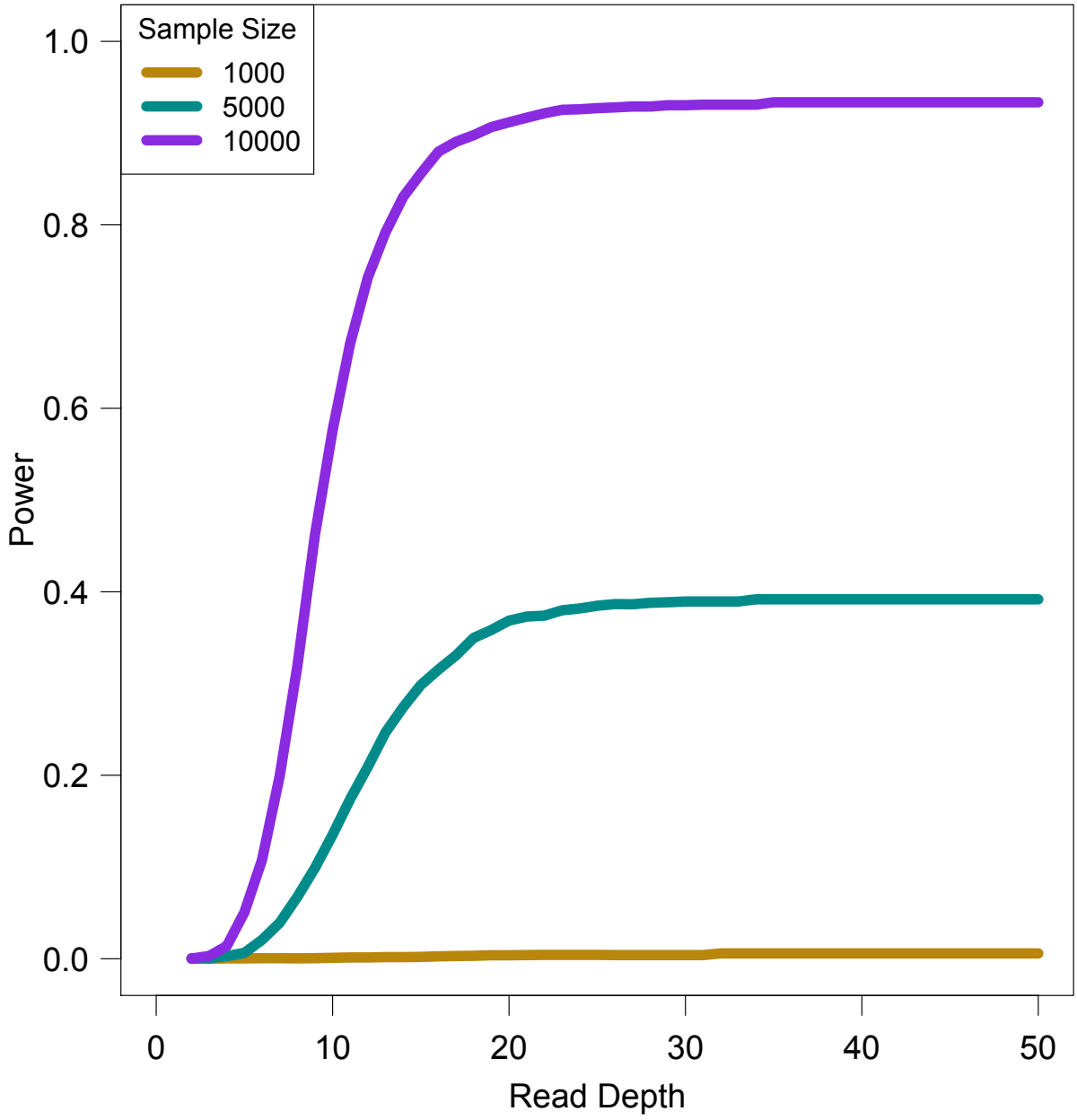


Figure III-6. Association study power by read depth for fixed sample size
Effect size = 0.7; population frequency of singletons = 0.008.



CHAPTER IV. Rapid Genotyping of Large Deletions from Next-Generation Sequencing Data[3]

Abstract

Many diseases, particularly psychiatric disorders such as schizophrenia and autism, have been associated with copy number variants. Detecting such variants is an important step in determining associations with diseases. Historically, this has been accomplished using array-based methods, but these do not typically allow for any other type of analysis to be performed from the resultant data. As sequencing costs decrease, it is increasingly desirable to use sequence data to assess copy number. There are several existing methods that implement approaches using sequence data, but most analyze samples individually. We developed a method that leverages population data across multiple individuals to determine copy number within a particular region, focusing specifically on number of reads mapped to the region. We tested our method using the 1000 Genomes Project Phase 1 samples and list of deletion regions. We were able to successfully identify 1,297 regions with easily distinguishable deletions, where each region has a mean accuracy of 95.49% (standard deviation of 8.16%) and a median of 97.67%. We also implemented our method in other samples, using this same list of regions. For AMD, GECCO, and Sardinia samples, we identified 935, 1,943, and 747 regions likely to encompass deletions, respectively. Our method is easy to use, fast, and accurate in calling deletions in regions

[3] Rashkin, S., Blackwell, T., and Abecasis, G. (In preparation). Rapid Genotyping of Large Deletions from Next-Generation Sequencing Data.

identified to have distinguishable deletions. Other regions likely require inclusion of information beyond mapped read counts or other, more complicated, methods.

Introduction

Copy number variants (CNVs) are segments of the genome where the number of copies varies between individuals. Typically, an individual has two copies of the entire genome: one maternally inherited copy and one paternally inherited copy. However, if there is a deletion, there can be fewer than two copies. Similarly, more copies can arise if there is a duplication or insertion. CNVs are longer than indels, ranging from approximately 100 bps to several thousand kbs. Copy number variation at some regions has been associated with susceptibility or resistance to a variety of diseases.[19] In particular, CNVs have been shown to be associated with psychiatric disorders, including schizophrenia and autism.[18-20] Detecting CNVs has, therefore, become an increasingly important area of research.

An important step in determining the association between CNVs and phenotypes of interest is to locate regions where copy number variation exists as well as the number of copies of each region that an individual possesses. Many studies have been conducted using microarray-based methods to locate CNVs.[19; 20] However, with the decreasing cost of genome sequencing, the amount of sequencing data available is increasing, and these data can be useful in determining CNVs, while allowing for other types of analyses simultaneously. Some recent studies have used sequencing information in order to locate regions of varying copy number.[19; 44] These methods consider at least one aspect of sequencing data in assessing copy number: (1) read depth analysis, which compares depth of coverage in a window to average depth to determine copy number; (2) read

pair analysis, which involves analysis of abnormally mapping pairs of reads; (3) split-read analysis, which assesses gapped sequence alignment; and (4) sequence assembly, which allows for structural variation detection.[19; 21; 38] While these methods all involve analysis of sequence data, most still analyze one genome at a time. One notable exception is Genome STRiP, which leverages population-level data into identification of deletions.[21; 44]

We developed a fast and simple method that works on sequence data within a specific region of interest. Focusing only on detecting deletions in key regions, we used sequencing data, specifically the number of reads mapped to a region, in order to predict CNVs via an expectation-maximization (EM) algorithm. We tested our method on the 1000 Genomes Project list of deletion regions using Phase 1 samples and successfully identified 1,297 regions (out of a total 13,973), with each region having a mean accuracy of 95.49% (standard deviation of 8.16%) and a median accuracy of 97.67%. Implementing our method on age-related macular degeneration (AMD)[45], Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)[31], and Sardinia[46] samples, we identified 935, 1,943, and 747 regions likely to have deletions, respectively. While our method might miss some deletion regions, these regions likely require more complicated detection methods, making them slower or more cumbersome to use.

Methods

It can be assumed that read depth in a region can be associated with the number of deletions at the region in question. As copy number decreases, read depth should also decrease. We measured an individual's read depth for each region as the number of reads mapped to the

region, adjusted by the total number of reads mapped for the individual. We further adjust read counts for a region by the mean read count across individuals for that region.

If there are two copies of the region, we can expect to have a high read depth, whereas we expect very low read depth if there are two deletions. Read depth can be modeled using a Gaussian mixture distribution:

$$d_i \sim p_0 \cdot N\left(\mu_0, \sigma_0^2\right) + p_1 \cdot N\left(\mu_1, \sigma_1^2\right) + p_2 \cdot N\left(\mu_2, \sigma_2^2\right)$$

where d_i is the adjusted number of reads mapped to a region for the i^{th} individual, μ_j represents the mean of the distribution component for j number of deletions (0, 1, 2), σ_j represents the standard deviation of the distribution component for j number of deletions, and p_j represents the probability of having j number of deletions.

To reduce the parameters being estimated, we relate the means of the three distributions to each other, based on the hypothesis: for an individual with two copies of the region, read depth is high; for an individual with no copies of the region, read depth is minimal; and, for an individual with one copy of the region, read depth falls in between these values. For this scenario, we set:

$$\mu_0 = \mu$$
$$\mu_1 = \mu + \Delta$$
$$\mu_2 = \mu + 2\Delta$$

where μ is the mean for the read count distribution for zero deletions and Δ is the difference in means between the groups. We further reduced the number of parameters by assuming each component of our mixture model has the same variance. Additionally, we assumed that the presence of a deletion is in Hardy-Weinberg Equilibrium (HWE), so the three probabilities can

be related to each other, so we only need to estimate one parameter, p_{del}, the probability of there being a deletion. Therefore, our model becomes:

$$d_i \sim (1-p_{del})^2 \cdot N\left(\mu,\sigma^2\right) + 2p_{del}(1-p_{del}) \cdot N\left(\mu+\Delta,\sigma^2\right) + p_{del}^2 \cdot N\left(\mu+2\Delta,\sigma^2\right)$$

We implemented an expectation-maximization (EM) algorithm to predict the maximum likelihood estimates for the mixture distribution and used these estimates to predict the number of deletions for each individual at each region. The EM-algorithm is a method of estimating parameters when there is missing data; here, the number of deletions is unknown. It is an iterative procedure resulting in the maximum likelihood estimates of all parameters. Since it is only guaranteed to find the local maximum and it can be sensitive to initial parameter values, we used multiple initialization values for each parameter (randomly selected within range of data) and kept the estimates resulting in the highest likelihood.

More specifically, for a set of observed data D that follows a statistical model (here, we assume d_i are independent, identically distributed and follow the mixture distribution described above), a set of unobserved data Z (here, z_i is the number of deletions for each individual), and a vector of parameters θ (here, $\theta=(p_{del}, \mu, \Delta, \sigma^2)$), the EM algorithm identifies the parameter estimates that maximize the likelihood function, $L(\theta;D,Z)$. The EM algorithm consists of two steps that are performed iteratively until parameter estimates converge.

In the expectation (E) step, we calculate $Q(\theta|\theta^{(r)})$, the expected value of the log likelihood with respect to Z given D and the current parameter estimates $\theta^{(r)}$, where r is the step number:

$$Q\left(\theta\middle|\theta^{(r)}\right) = E_{Z|d,\theta^{(r)}}\left[\log L(\theta;D,Z)\right]$$
$$\propto \sum_{j=0}^{2}\left\{\left(\log p_j^{(r)}\right)E\left[\sum_{i=1}^{n} I\left(Z_i = j\right)\middle|D,\theta^{(r)}\right] + \sum_{i=1}^{n}\left[\tfrac{-1}{2}\log\left(\sigma_j^{(r)}\right)^2 - \tfrac{1}{2\sigma_j^2}\left(d_i - \mu_j^{(r)}\right)^2\right]E\left[I\left(Z_i = j\right)\middle|D,\theta^{(r)}\right]\right\}$$

We can simplify this equation to calculating the following quantities:

$$E\left[I\left(Z_i = j\right)\middle|D,\theta^{(r)}\right] = P\left(Z_i = j\middle|D,\theta^{(r)}\right)$$
$$= \frac{p_j^{(r)} f\left(d_i;\mu_j^{(r)},\sigma_j^{(r)}\right)}{\sum_{j=0}^{2} p_j^{(r)} f\left(d_i;\mu_j^{(r)},\sigma_j^{(r)}\right)} = T_{j,i}^{(r+1)}$$
$$E\left[\sum_{i=1}^{n} I\left(Z_i = j\right)\middle|D,\theta^{(r)}\right] = \sum_{i=1}^{n} T_{j,i}^{(r+1)}$$

where the first is calculated for each individual *i* and group *j*, and the second is the summation of the first quantity across all individuals for each group *j*.

The maximization (M) step finds the parameter values that maximize $Q(\theta|\theta^{(r)})$:

$$\theta^{(r+1)} = \arg\max_{\theta} Q\left(\theta\middle|\theta^{(r)}\right)$$

We can simplify this to the following expressions for each parameter:

$$p_{del}^{(r+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(r+1)} + 2\sum_{i=1}^{n} T_{2,i}^{(r+1)}}{2n}$$
$$\mu^{(r+1)} = \bar{d} - \Delta^{(r+1)}\cdot\,\bar{y}^{(r+1)}, \quad where\ \bar{y}^{(r+1)} = \frac{\sum_{i=1}^{n}\left(T_{1,i}^{(r+1)} + 2\cdot\,T_{2,i}^{(r+1)}\right)}{n}$$
$$\Delta^{(r+1)} = \frac{\sum_{i=1}^{n}\sum_{j=0}^{2} T_{j,i}^{(r+1)}\left(j - \bar{y}^{(r+1)}\right)\left(d_i - \bar{d}\right)}{\sum_{i=1}^{n}\sum_{j=0}^{2} T_{j,i}^{(r+1)}\left(j - \bar{y}^{(r+1)}\right)}$$
$$\sigma^{(r+1)} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=0}^{2} T_{j,i}^{(r+1)}\left(d_i - \mu_j^{(r+1)}\right)^2}{n}}, \quad where\ \mu_j^{(r+1)} = \mu^{(r+1)} + j\cdot\,\Delta^{(r+1)}$$

We used the Kullback-Leibler (KL) distance between the null model of a single normal distribution and the estimated mixture model to assess whether our method is likely to be able to

accurately call deletions. The KL distance between distributions P and Q can be conceptualized as the information lost when using Q to approximate P and is calculated as:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx$$

where P is the null distribution and Q is the mixture distribution. Small values of KL distance indicate that Q, the estimated mixture distribution, is similar to P, the null distribution of a single normal distribution.

We first tested our method using 1,072 samples (average depth 11.36x) from the 1000 Genomes Project Phase 1 data and associated list of deleted regions.[47] We restricted our analysis to the 13,973 autosomal regions. We used these data to determine the value of KL distance that allows for detecting regions where deletions are present in some individuals and accurately genotyping each individual. We then applied our method to three other datasets: 1968 AMD samples (average depth ~5x)[45], 2975 samples (average depth ~13x) from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)[31], and 3441 Sardinian samples (average depth ~4x)[46].

Results

Our method works well to detect deletions in regions, especially for regions where the read count distributions for each number of deletions are well separated (see Figure IV-1). For a KL distance threshold of 0.3, our method identifies 1,853 regions likely to have deletions (see Table IV-A). Comparing deletion calls with those from the 1000 Genomes Project, these regions have a mean concordance of 75.12% (standard deviation of 39.12%) and a median of 95.62% (see Table IV-A).

Since we anticipate very few reads to be mapped to a region when there is a homozygous deletion, we expect the estimated mean for this group to be close to 0. Therefore, we further filter regions to exclude those where the mean for this group is further than two standard deviations away from zero. After applying this filter, we identify 1,297 regions that have a mean concordance of 95.49% (standard deviation of 8.16%) and a median of 97.67% (see Table IV-B). Furthermore, only 15 regions have a concordance below 80% (136 below 90%). For most of these regions, the read count distributions either are very overlapped (see Figure IV-2) or do not follow the distribution we expect (see Figure IV-3).

In regions where the read count distributions for each number of deletions are overlapped, some genotype calls are incorrect or ambiguous (see Figure IV-4). We attempted to resolve this uncertainty by performing genotype refinement via Beagle.[48] Use of this genotype refinement tends to improve genotype accuracy, especially in regions where the read count distributions are overlapping (see Figure IV-4). While some regions do not show an improvement after genotype refinement, most show moderate increases in concordance (see Figure IV-5).

We applied our method with the same filters for the same list of 13,973 deletion regions to three other datasets to identify regions likely to have deletions in these samples. For each of these datasets, our method identified an overlapping, but not identical, list of regions where deletions are likely (see Figure IV-6). Comparing the list of 1,297 regions identified by our method in the 1000 Genomes Project Phase 1 samples with the regions identified in other datasets, 454 regions are only identified in the 1000 Genomes samples. Our method identifies 935 regions in the AMD

samples, 77 of which are uniquely identified in these samples. Our method identifies 1,943 regions, where 1,023 regions are not identified in any other dataset. While having the largest sample size (3,441 individuals), our method identifies the fewest number of regions in the Sardinia samples: 747 regions, with only 36 uniquely identified in these samples. There are 557 regions our method identifies as likely to have deletions in all four datasets.

Discussion

We developed a novel method for detecting deletions from sequence data by leveraging information across multiple samples simultaneously. Our method is fast, easy to use, and accurate. Using the 1000 Genomes Project Phase 1 samples and list of deletion regions, our method identified 1,297 regions, with each region having a mean accuracy of 95.49% (standard deviation of 8.16%) and a median of 97.67%. Implementing our method on AMD, GECCO, and Sardinia samples, we identified 935, 1,943, and 747 regions likely to have deletions, respectively. While our method might miss some deletion regions, these regions likely require more complicated detection methods, making them slower or more cumbersome to use.

Our method runs quickly. Collecting metadata for each region for each individual is the rate-limiting step, compared to running our deletion caller. For 100 regions in a single sample processed on a single CPU, collecting the metadata takes approximately 5 minutes. Once the metadata are collected, implementing our deletion caller for 100 regions in a sample of 100 individuals takes approximately 10 seconds on a single CPU, running a single region at a time.

Since our method runs quickly, even though it may miss some deletion regions, it provides a useful step in an analysis where it is important to identify regions where deletions are likely. In particular, our method allows for a quick determination of deletion calls within specific regions of interest, without identifying particular breakpoints. In addition, its speed allows for a quick overview of specific regions and whether they are likely to contain deletions in a specific sample without committing to a more complicated, and time consuming, analysis.

Our method identifies the fewest number of deletion regions in the Sardinia data. While having the largest sample size (3441 individuals), these samples have the lowest average depth. Since our method compares read counts within a region, the deeper the coverage, the better it will be able to make comparisons.

Our method should work for all types of sequencing data, including exome data, but our analysis here focused on whole genome sequencing data. Further testing is necessary to assess method performance with exome data. Additionally, our current method only allows for detection of deletions within a specified region. Further work is necessary to include detection of insertions.

Figures and Tables

Figure IV-1. Example of region where distributions are well-separated
Region located at 1:152555542-152587742

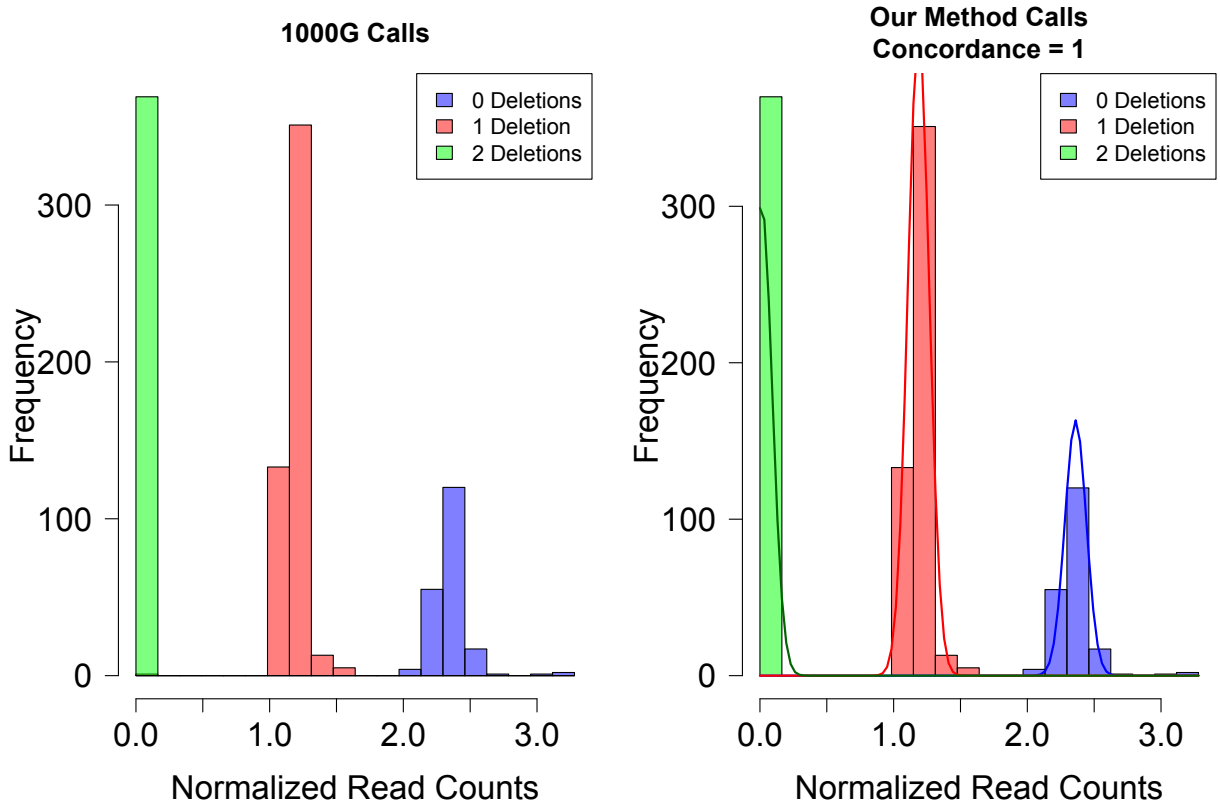


Table IV-A. Region concordance by KL threshold

KL Threshold	# Regions	Mean Concordance (SD) (%)	Median Concordance (%)
0.3	1853	75.12 (39.12)	95.62
0.35	1436	80.95 (34.94)	96.74
0.4	1145	86.34 (23.10)	97.29
0.45	918	90.71 (23.10)	97.67
0.5	761	92.87 (19.41)	97.95

Table IV-B. Region concordance by KL threshold for regions where μ_2-2σ<0

KL Threshold	Number of Regions	Mean Concordance (SD) (%)	Median Concordance (%)	# Regions where concordance <80%	# Regions where concordance <90%
0.3	1297	95.49 (8.16)	97.67	15	136
0.35	1108	95.85 (7.94)	97.85	11	86
0.4	954	95.97 (8.31)	97.85	9	62
0.45	801	96.21 (8.24)	97.95	8	43
0.5	680	96.53 (7.97)	98.04	7	27

Figure IV-2. Example of regions where distributions are very overlapped
Region located at 7:8698628-8698846

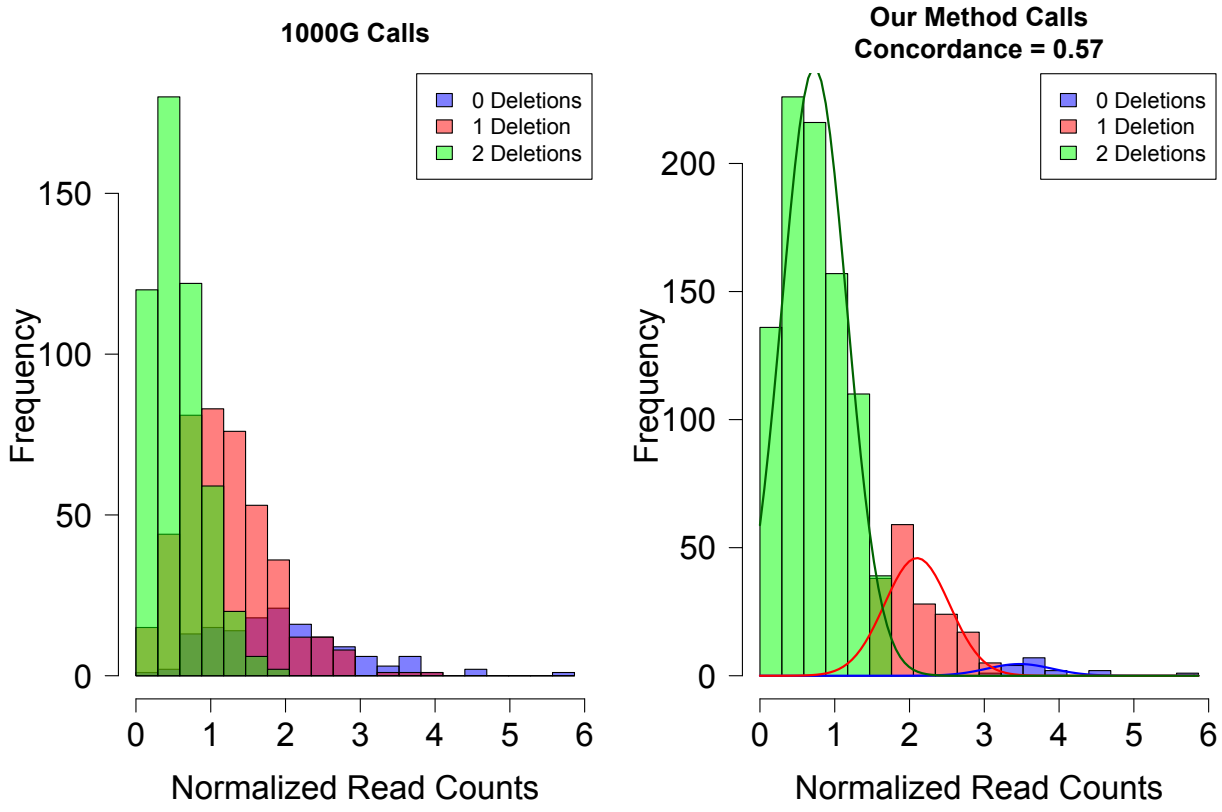


Figure IV-3. Example of a region where read counts do not follow expected distribution
Region located at 1:1564801-1568200

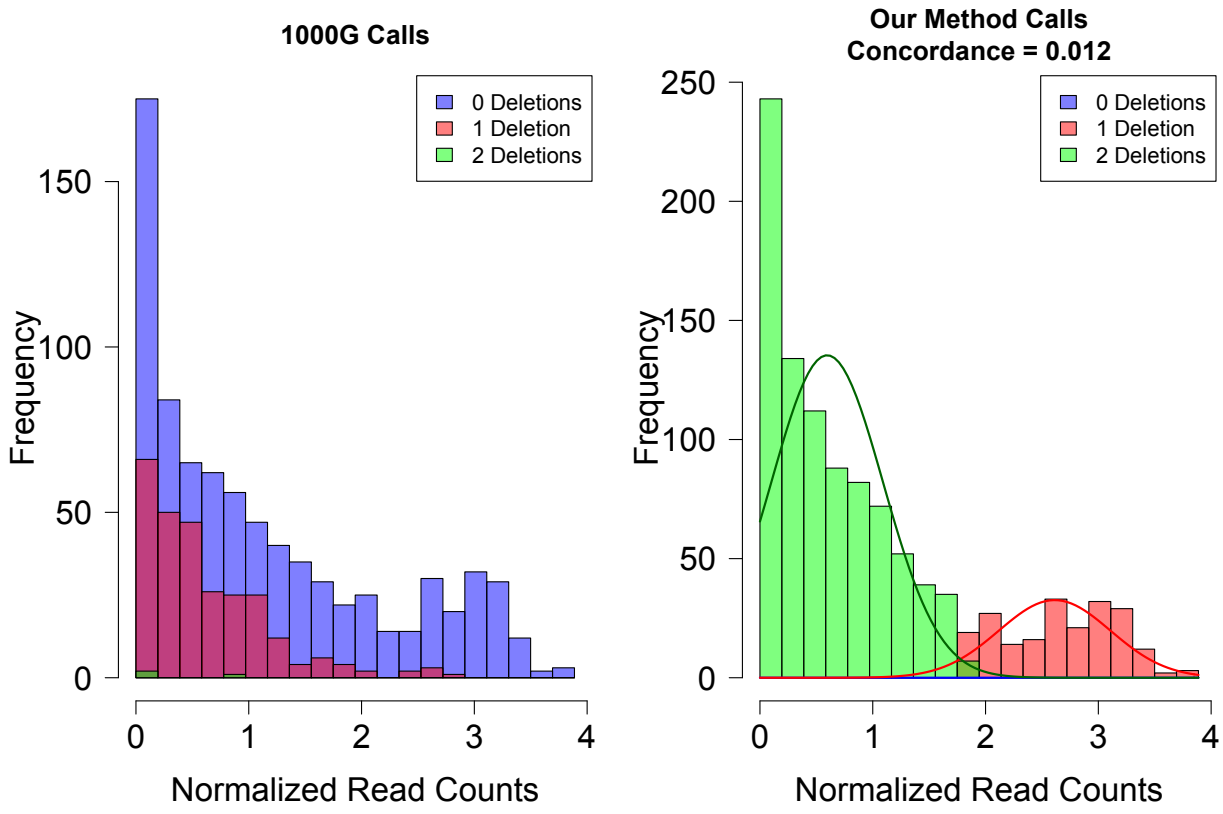


Figure IV-4. Example of region where distributions are overlapping
Region located at 21:15460780-15461864

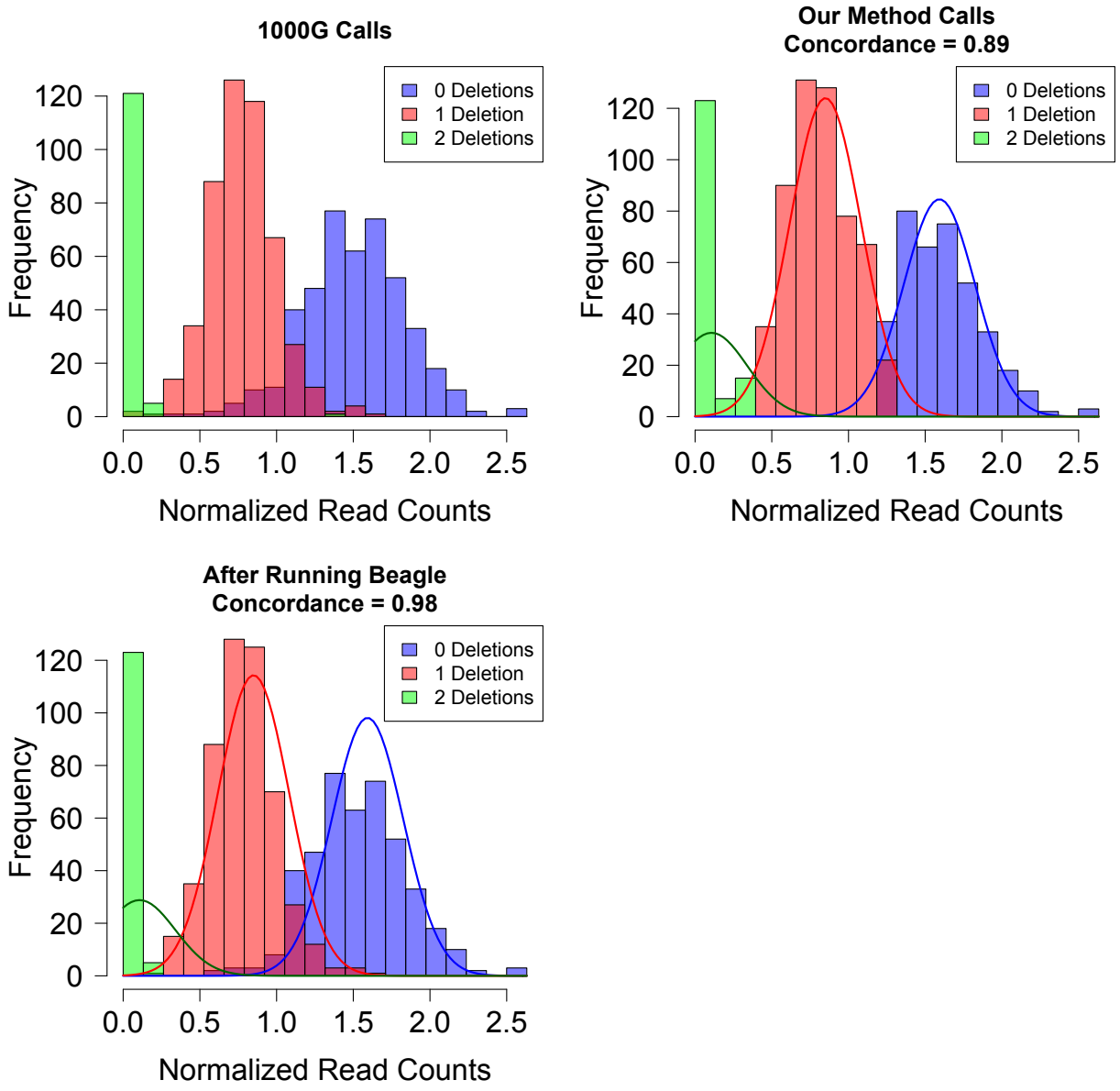


Figure IV-5. Concordance after Genotype Refinement vs. Original Concordance

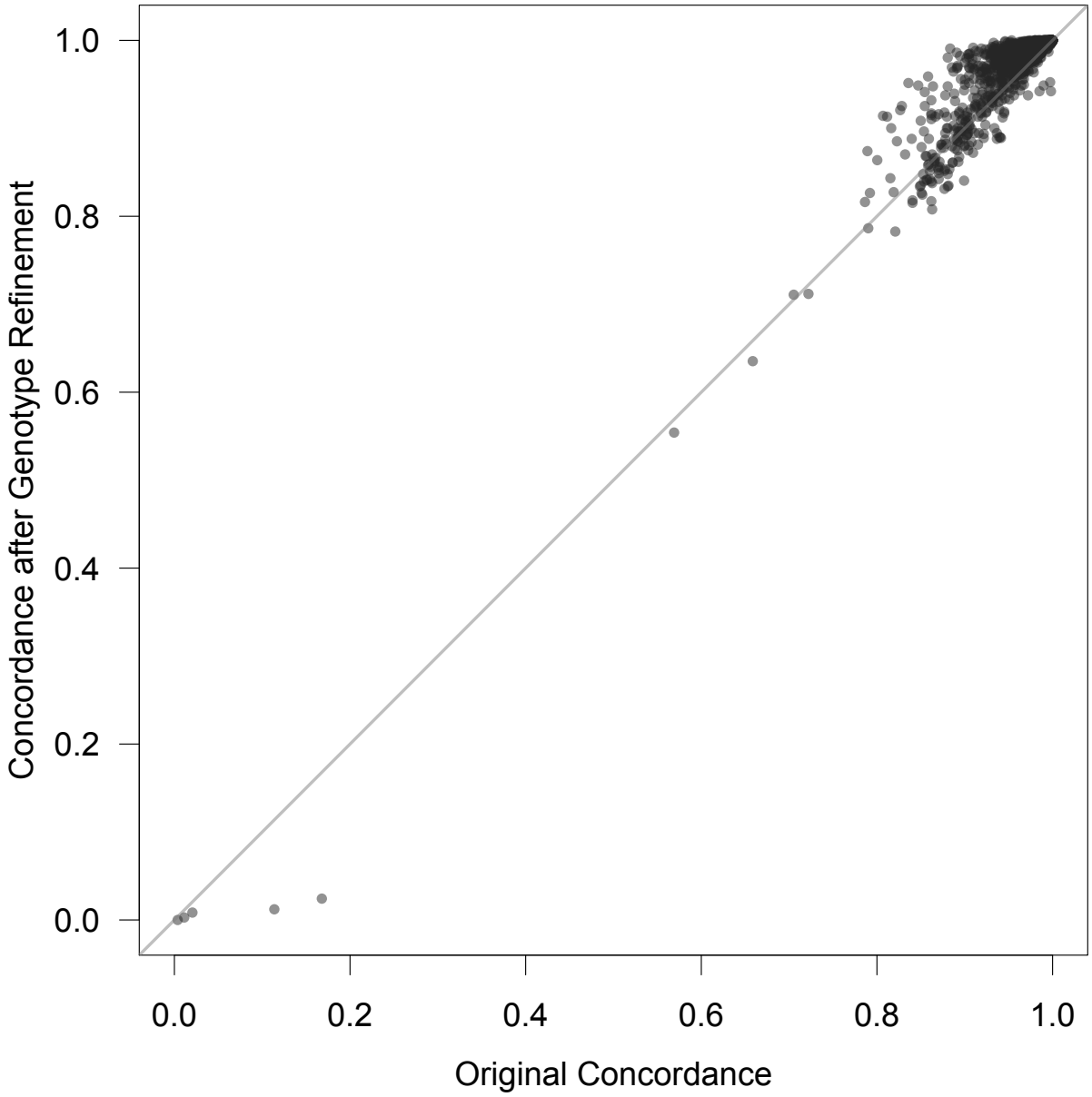
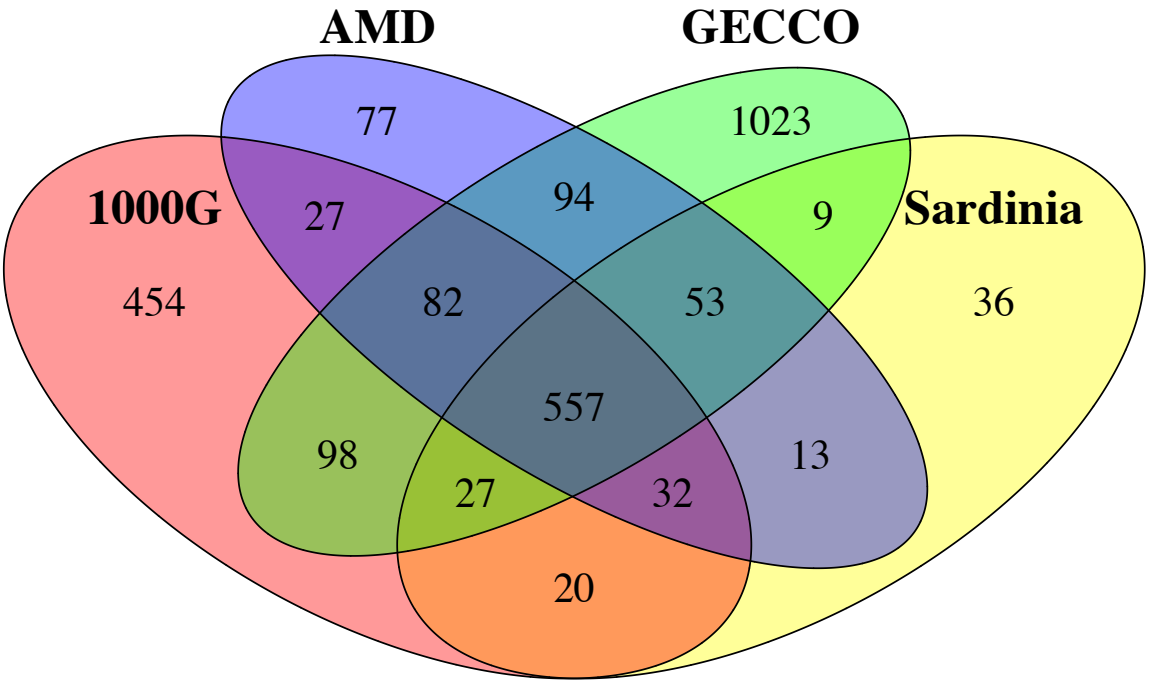


Figure IV-6. Venn diagram comparing lists of regions identified in each dataset



CHAPTER V. Discussion

While thousands of loci have been associated with complex diseases and traits, much is still unknown about the biology of disease. One possible reason is that much of genetic association studies has focused on common SNPs. Therefore, in this dissertation, we described methods that will aid in the design and analysis of complex traits using sequence data by allowing for improved assessments of other types of variants to disease, particularly rare variants and copy number variants.

In Chapters II and III, we assess different study designs to determine what analysis plan (deep, intermediate, or low coverage sequencing) is best for studying very rare variants. This will aid in future endeavors to design studies with goals of identifying rare variants associated with either discrete or quantitative traits. We showed that, when cost is fixed so read depth and sample size vary inversely, power to detect association between a singleton variant and a trait decreases as coverage is increased beyond a threshold (15-20x for discrete traits, 11-16x for quantitative traits). We recommend increasing coverage beyond this threshold only if sample size is restricted or reasons beyond genetic association studies justify decreasing sample size in favor of a more complete sequencing of each individual.

In Chapter IV, we describe a novel method for detecting deletion regions from sequence data by leveraging information across samples. This allows for determining deletion regions, which is

useful in downstream analysis, quickly, easily, and accurately, without the added expense of an extra assay. While our method might miss some deletion regions, these regions likely require more complicated detection methods. Our method's speed allows for a quick overview of specific regions to determine which ones are likely to contain deletions for a sample of individuals before committing to a more detailed and time consuming analysis.

REFERENCES

1. Peters, D.T., and Musunuru, K. (2012). Functional evaluation of genetic variation in complex human traits. Human molecular genetics 21, R18-23.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.
3. Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. Genome research 21, 940-951.
4. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. American journal of human genetics 95, 5-23.
5. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22, 231-238.
6. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science (New York, NY) 337, 64-69.
7. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. American journal of human genetics 82, 100-112.
8. Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature 456, 18-21.
9. Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40, 695-701.
10. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature reviews Genetics 11, 415-425.
11. Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 46, 294-298.
12. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N., et al. (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet 44, 1326-1329.
13. Kimura, H., Tsuboi, D., Wang, C., Kushima, I., Koide, T., Ikeda, M., Iwayama, Y., Toyota, T., Yamamoto, N., Kunimoto, S., et al. (2014). Identification of Rare, Single-Nucleotide Mutations in NDE1 and Their Contributions to Schizophrenia Susceptibility. Schizophrenia bulletin.

14. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
15. Le, S.Q., and Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome research 21, 952-960.
16. Mackay, T.F., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. Nature reviews Genetics 10, 565-577.
17. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics 11, 446-450.
18. Mehta, D., Iwamoto, K., Ueda, J., Bundo, M., Adati, N., Kojima, T., and Kato, T. (2014). Comprehensive survey of CNVs influencing gene expression in the human brain and its implications for pathophysiology. Neuroscience research 79, 22-33.
19. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al. (2011). Mapping copy number variation by population-scale genome sequencing. Nature 470, 59-65.
20. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. American journal of human genetics 91, 597-607.
21. Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S., and Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics (Oxford, England) 28, 2711-2718.
22. Morrison, A.C., Voorman, A., Johnson, A.D., Liu, X., Yu, J., Li, A., Muzny, D., Yu, F., Rice, K., Zhu, C., et al.; Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium. (2013). Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45, 899-901.
23. Helgason, H., Sulem, P., Duvvari, M.R., Luo, H., Thorleifsson, G., Stefansson, H., Jonsdottir, I., Masson, G., Gudbjartsson, D.F., Walters, G.B., et al. (2013). A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nat Genet 45, 1371-1374.
24. Raychaudhuri, S., Iartchouk, O., Chin, K., Tan, P.L., Tai, A.K., Ripke, S., Gowrisankar, S., Vemuri, S., Montgomery, K., Yu, Y., et al. (2011). A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nat Genet 43, 1232-1236.
25. Seddon, J.M., Yu, Y., Miller, E.C., Reynolds, R., Tan, P.L., Gowrisankar, S., Goldstein, J.I., Triebwasser, M., Anderson, H.E., Zerbib, J., et al. (2013). Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. Nat Genet 45, 1366-1370.
26. Zhan, X., Larson, D.E., Wang, C., Koboldt, D.C., Sergeev, Y.V., Fulton, R.S., Fulton, L.L., Fronick, C.C., Branham, K.E., Bragg-Gresham, J., et al. (2013). Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nat Genet 45, 1375-1379.
27. Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science (New York, NY) 324, 387-389.

28. (2014). Inactivating Mutations in NPC1L1 and Protection from Coronary Heart Disease. New England Journal of Medicine 371, 2072-2082.
29. Rashkin, S., Blackwell, T., and Abecasis, G. (In preparation). Rapid Genotyping of Large Deletions from Next-Generation Sequencing Data.
30. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.
31. Hsu, L., Jeon, J., Brenner, H., Gruber, S.B., Schoen, R.E., Berndt, S.I., Chan, A.T., Chang-Claude, J., Du, M., Gong, J., et al. (2015). A model to determine colorectal cancer risk using common genetic susceptibility loci. Gastroenterology 148, 1330-1339 e1314.
32. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25, 2078-2079.
33. Go, A.S., Mozaffarian, D., Roger, V.L., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S., et al. (2013). Heart disease and stroke statistics--2013 update: a report from the American Heart Association. Circulation 127, e6-e245.
34. Friedman, D.S., O'Colmain, B.J., Munoz, B., Tomany, S.C., McCarty, C., de Jong, P.T., Nemesure, B., Mitchell, P., and Kempen, J. (2004). Prevalence of age-related macular degeneration in the United States. Archives of ophthalmology 122, 564-572.
35. Maahs, D.M., West, N.A., Lawrence, J.M., and Mayer-Davis, E.J. (2010). Epidemiology of type 1 diabetes. Endocrinology and metabolism clinics of North America 39, 481-497.
36. NCBI Gene. AVP arginine vasopressin [Homo sapiens (human)]. In. ( ): National Center for Biotechnology Information [modified 2014 Sep 27; cited 2014 Oct 1]. Available from http://www.ncbi.nlm.nih.gov/gene/551.
37. NCBI Gene. (2014). TTN titin [Homo sapiens (human)]. In. ( ): National Center for Biotechnology Information [modified 2014 Sep 27; cited 2014 Oct 1]. Available from: http://www.ncbi.nlm.nih.gov/gene/7273.
38. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nature reviews Genetics 12, 363-376.
39. Panoutsopoulou, K., Tachmazidou, I., and Zeggini, E. (2013). In search of low-frequency and rare variants affecting complex traits. Human molecular genetics 22, R16-21.
40. van Zyl, T., Jerling, J.C., Conradie, K.R., and Feskens, E.J. (2013). Common and rare single nucleotide polymorphisms in the LDLR gene are present in a black South African population and associate with low-density lipoprotein cholesterol levels. J Hum Genet.
41. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dube, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat Genet 46, 629-634.
42. Rashkin, S., Jun, G., Chen, S., Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), and Abecasis, G. (In preparation). Optimal Sequencing Strategies for Identifying Disease-Associated Singletons.
43. Montgomery, D.C., Peck, E.A., and Vining, G.G. (2006). Introduction to linear regression analysis.(: Wiley-Interscience).
44. Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43, 269-276.

45. Kwong, A.M., Zhan, X., Fritsche, L.G., Bragg-Gresham, J., Branham, K.E., Othman, M., Gieser, L., Ratnapriya, R., Stambolian, D., Chew, E.Y., Swaroop, A., and Abecasis, G.R. Whole-Genome Sequencing of 4,806 Individuals for Age-Related Macular Degeneration (Abstract/Program#745). Presented at the 65th Annual Meeting of The American Society of Human Genetics, October 9, 2015 in Baltimore, MD.
46. Pilia, G., Chen, W.M., Scuteri, A., Orru, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS genetics 2, e132.
47. The 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
48. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics 81, 1084-1097.