# Chapter 2

# Case-control association testing using sequencing data

## 2.1 Introduction

The emergence of 'next-generation' technology has caused the cost of DNA sequencing to plummet over the last ten years. This has already led to a number of very successful large-scale sequencing studies using healthy human populations, such as the 1000 Genomes, UK10K, and Exome Aggregation Consortium projects. However, researchers are now looking to extend this success to the identification of disease risk variants using case-control cohorts. Through the direct capture of millions of rare and low frequency variants, such studies offer an unprecedented opportunity to better understand the genetic architecture of complex disease, uncover novel associations underlying disease risk, and further resolve signals down to causal variants of potential therapeutic relevance.

Despite the promise offered by such studies, in practice they are hampered by the high costs associated with sequencing at scale, and the complexity of analysing such data. One cost-saving approach that has been used very successfully in array-based genome wide association studies is to borrow control samples from publicly-available datasets, allowing a maximal number of disease cases to be assayed. However,

attempts to use the same study design in a sequencing setting are faced with a number of difficulties associated with combining multi-source sequencing data at scale. In particular, systematic biases in exome capture technology and sequencing depth lead to crucial sensitivity and specificity differences when performing variant calling; for case-control studies, the effects of these systematic biases can be observed as a slew of false associations.

### 2.1.1 Chapter overview

In this chapter, I shall describe methods that can be used for the case-control analysis of sequencing data in the presence of a known bias in sensitivity and specificity between the cohorts, as may arise through systematic differences in, amongst other things, sequencing depth. Existing methods to approach this problem include the incorporation of population-level information, through the use of joint calling, genotype refinement, and imputation into GWAS datasets, in order to improve the ability to test for association at sites of low frequency variation.

For rare variation, where the minor allele is observed too infrequently for population-based methods to be effective, I implement a new statistic proposed by Derkach et al. (2014) that is able to account for systematic biases between cases and controls directly in the association test. In order to obtain a well-behaved test statistic on real data, I develop a number of additional filtering recommendations that can be used to identify both errors and variants that are likely to be true sites of variation but have been poorly captured in one of the groups due to systematically lower sequencing depth.

Together, these methods demonstrate that it is possible, albeit difficult, to perform large-scale association testing in sequencing data that suffers from widespread systematic biases between cases and controls. This opens up the opportunity for researchers to perform case-control analyses on datasets that have been obtained from multiple sources, such as can often occur when merging datasets in large-scale efforts by disease consortia, or when looking to maximise sample sizes in a fixed-cost study through the use of publicly available control datasets.

### 2.1.2 Contributions

In order to test the methods described here, I used a low coverage sequencing study of inflammatory bowel disease performed by the UK IBD Genetics Consortium. Variant calling, genotype refinement, and many of the quality control analyses on this dataset were performed by Yang Luo. Further details on this dataset, and those who contributed to preparing it, will be provided in Chapter 3. Of particular relevance to the work in this chapter, the analysis of low quality sites using support vector machines was performed by Yang Luo. Unless stated, I carried out all other analyses.

## 2.2 Next-generation sequencing studies

### 2.2.1 Study design considerations

Next-generation sequencing offers an exciting opportunity to improve our understanding of the genetics underlying complex traits. However, in reality this excitement is tempered by the high costs still associated with sequencing. Because expenditure increases approximately linearly with the number of short sequencing reads produced, a crucial design decision in a fixed cost study revolves around how best to distribute these reads to maximise information: towards increased sample size, increased individual coverage, or an increased number of interrogated sites.

To date, the majority of sequencing studies have focused on the exome. This cost-effective approach to sequencing captures just the protein-coding portion of the genome to high coverage, which makes it well suited for use in clinical diagnostics and the discovery of rare, coding disease variants. Initial studies were therefore focused on individuals or small family groups with unexplained Mendelian disorders. However, exome sequencing has seen an explosion in popularity over the past decade, culminating in the recent release of over 60,000 exomes by the Exome Aggregation Consortium (Lek et al., 2016). During this time, exome studies have offered important insights into a number of aspects of human health and disease, ranging from the identification of causal mutations in rare disorders (Choi et al., 2009; Ng et al., 2010; Wright et al., 2015) and driver mutations in cancers (Barbieri et al., 2012; Stephens et al., 2012), through to more general characterisations of rare coding variation across large cohorts (Walter et al., 2015; Lek et al., 2016).

An alternative study design involves redistributing the sequencing reads to capture the whole genome, but to much lower coverage. This allows for large sample sizes, and the detection of potentially interesting non-coding variation, but comes at the cost of data quality at the individual sample level. This type of study has proven to be a valuable way of obtaining comprehensive genome-wide catalogues of variation across human populations, via studies such as the 1000 Genomes and UK10K projects (1000 Genomes Project Consortium et al., 2015; Walter et al., 2015). Furthermore, through the cost-effective collection of large whole genome

cohorts, such studies have led to the development of a haplotype reference panel containing over 32,000 individuals, providing a very important public resource that can be used for the accurate imputation of low frequency variants from existing genotyping arrays (McCarthy et al., 2016).

### 2.2.2 Challenges of performing case-control analyses

These large-scale exome and low coverage whole genome efforts have highlighted not only the importance of generating very large sequencing cohorts, to reveal patterns of human population biology and provide vital resources for interpreting the clinical relevance of variation, but also the practical difficulties in managing multi-source data at this scale. The lack of a standardised approach for the generation of sequencing data has resulted in a number of slight variations on the basic study design, whether it be high-coverage exomes or low-coverage whole genomes, as investigators try to fine-tune their designs to answer a variety of scientific questions. As a result, when combining data from 14 different studies, the Exome Aggregation Consortium pointed out that variations in exome capture technology and sequencing depth across their 60,706 exomes required a joint analysis of such computational intensity and analytical complexity that it would be impossible using the limited resources available to most research centres (Lek et al., 2016).

I will note here that the systematic differences between cohorts being referred to here are not the same as the batch effects that can arise through the course of an experimental study. Just as is often seen with genotyping data, sequencing studies are still plagued by such issues: the specific reagents and machines used, slight variations in experimental conditions, or even the day on which a sample was processed can all lead to differences in the quality of the data produced (Figure 2.1). Naturally, these problems are important to consider, and indeed if samples are processed at multiple sequencing facilities then these effects can become even more pronounced. However, generally, these sorts of batch effects can be accounted for using careful quality control.

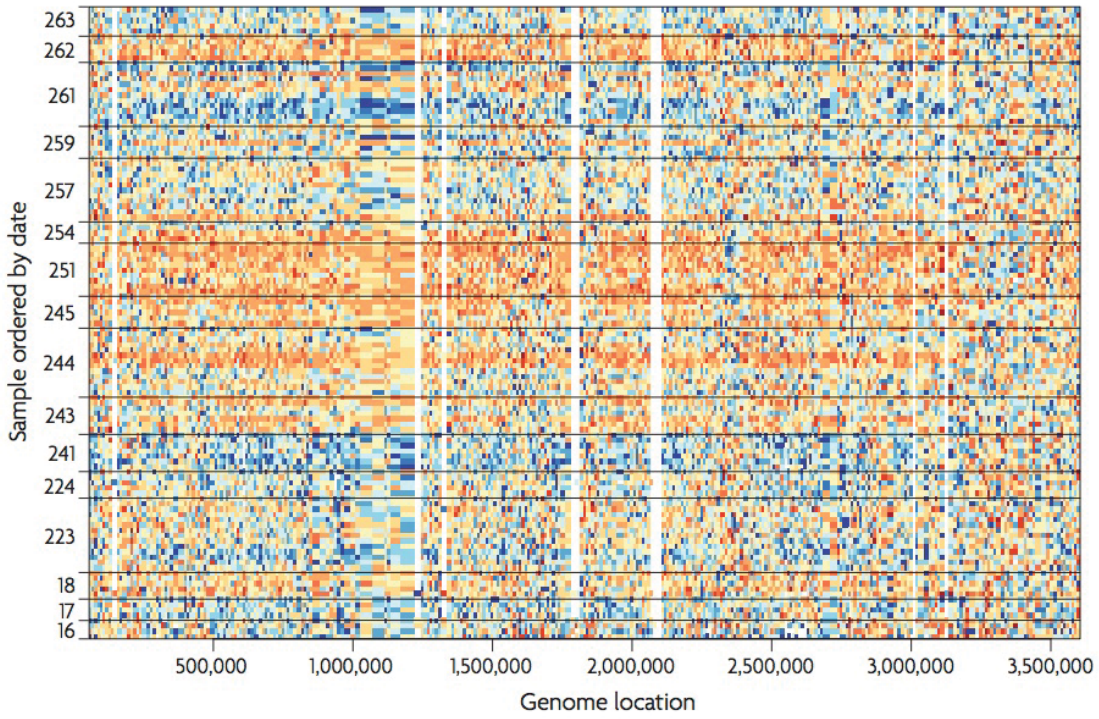


Figure 2.1: Batch effects observed in the 1000 Genomes project sequence data. Each row repesents a different HapMap sample, sorted by processing date. Every sample was processed in the same sequencing facility, using the same platform. Colours represent the standardised coverage data for each sample: blue indicates three standard deviations below average, and orange indicates three standard deviation above average. A large batch effect is observed between days 243 and 251. Figure sourced from Leek et al. (2010).

Of greater concern when combining sequence data from multiple sources are more widespread systematic differences that have arisen due to variation in the study designs. One example of this is the exome capture kit used, which defines the regions of the genome that will be sequenced and (through variable probe efficacies) the relative read depth that is likely to be obtained for certain regions. Systematic differences in read depth can also be observed on a more global scale, when data has simply been collected to different average coverages.

Because variants are detected in sequence data using the distribution of alleles across all the reads that overlap at a given position, sequencing depth has a direct impact on the sensitivity and specificity of variant calling. In particular, increased read depth leads to both improved sensitivity (the detection of true variant sites)

and improved specificity (the ability to distinguish true variants from sequencing errors). As a result, a cohort sequenced to higher depth (whether that be globally or locally) can be expected to contain more sites of true variation, and fewer errors, than a cohort sequenced to lower average depth across the same regions.

This observation is likely to be a serious problem as we extend the success of sequencing-based studies in healthy human populations to explore disease associations in case-control cohorts. I shall describe one such effort in Chapter 3, where we use low coverage sequencing to search for rare and low frequency variation associated with IBD. In that example, the cases were sequenced to a lower average depth than the controls (which were sourced from the UK10K project), in order to maximise sample size and therefore power to detect associations. Although this study may represent a particularly extreme example of differing read depths between cases (2-4x) and controls (7x), we envision that similar issues are likely to arise in other studies that use publicly available controls to save on costs. In this sort of case-control setting, any systematic differences between sequencing data from different sources is likely to heavily bias attempts to perform association testing.

In the following sections, I shall describe a range of methods that can be used to overcome systematic differences in sequencing depth between cases and controls. These consist of two broad approaches, depending on the prevalence of the variant of interest in the population. Firstly, for more common variants, population level information can be used to improve the overall sensitivity of both datasets and reduce differences between cohorts, thereby allowing standard association testing methods to be used. For rare variants, where this information is not available, I instead describe the development of a new approach to perform association testing in the presence of coverage bias between cases and controls.

## 2.3 Low frequency and common variants

### 2.3.1 Joint calling across samples

A powerful means of overcoming systematic sequencing differences between cases and controls at sites of low frequency and common variation is to perform joint variant calling (Figure 2.2). This method uses population-level detail about a given site to improve sensitivity to detect variation in carriers that have only intermediate levels of sequence support. It also allows for better specificity in variant detection: essentially, when more information is incorporated, it becomes easier to model errors and detect false positives. This is particularly important for sequencing data where, unlike the extensively curated variant lists that are included on genotyping arrays, there has been no pre-selection for true sites of variation.

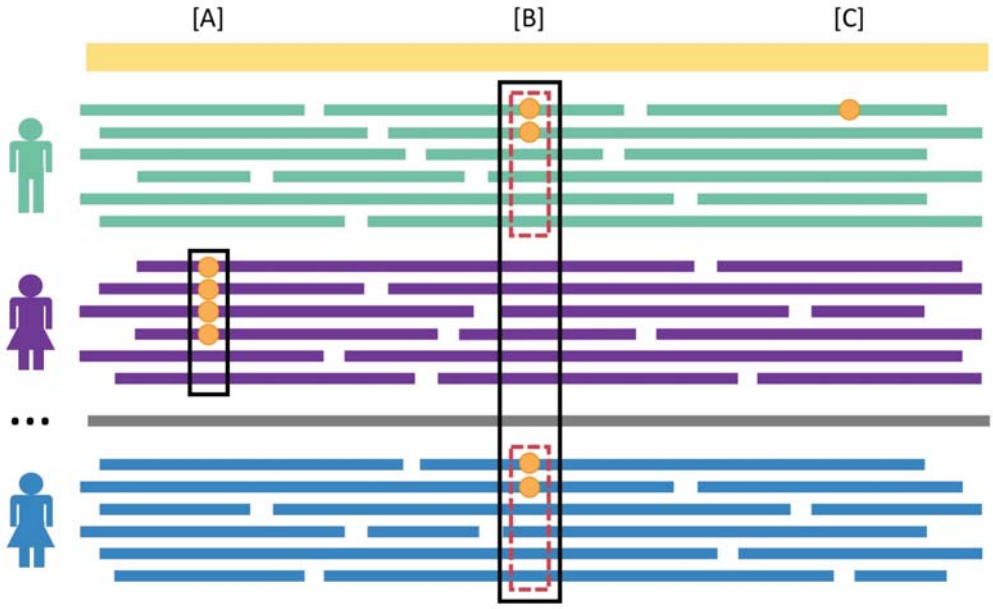


Figure 2.2: Calling variants jointly across a number of individuals can increase both the sensitivity and specificity of variant detection. While some variants may be accurately captured through variant calling on a single sample (A), for some individuals a lack of sequence support can cause the variant to be missed (B). However, if the variant is jointly called across reads pooled from a number of samples, these variants can be more accurately detected (B). Joint calling also helps to improve the detection of errors (C).

By performing variant calling jointly across the entire case-control cohort, the genotype calls for all samples will utilise information from reads accumulated over both cases and controls. This can greatly improve the sensitivity and specificity of variant calling for both groups, and reduce calling differences that may have arisen due to variations in average sequencing depths.

### 2.3.2 Genotype refinement

After joint calling, some variants that have been poorly captured for a given individual can be improved using genotype refinement (Figure 2.3), which infers specific genotypes by imputing from other individuals and neighbouring variation. As Li (2011) explains, this method improves the genotype call for an individual, $I$, who happens to have poor sequence coverage at the site of interest, $S_0$. If there are other samples that have high coverage at $S_0$ then, if there exists a second site $S_1$ which is in high linkage disequilibrium with $S_0$, and for which both $I$ and the other samples have sufficient sequence support, the likely genotype for individual $I$ at position $S_0$ can be inferred.

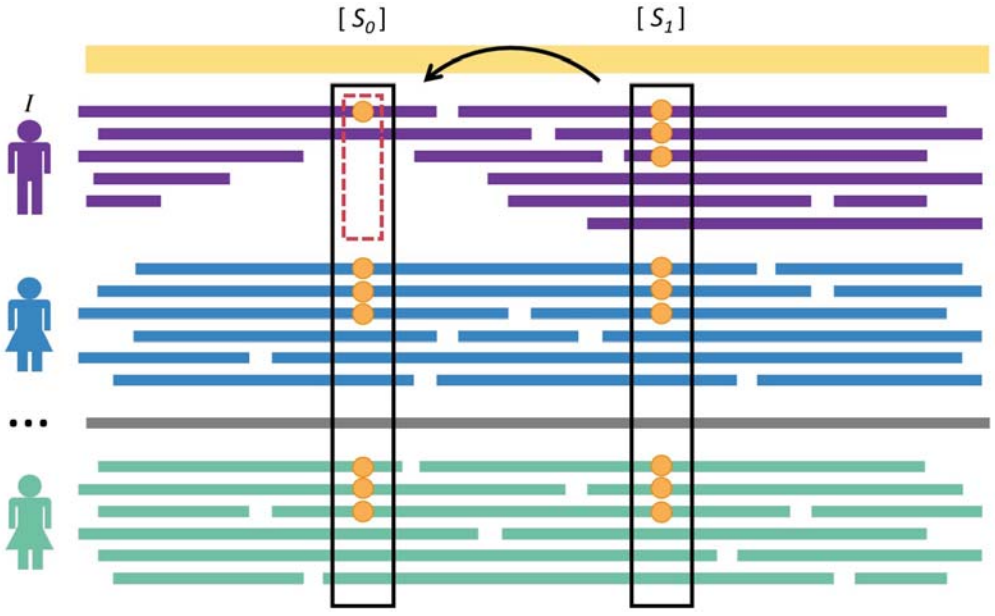


Figure 2.3: Genotype refinement through imputation, where the poor quality genotype at position $S_0$ for individual $I$ is improved by imputing from position $S_1$.

### 2.3.3 Imputation of GWAS cohorts

A combination of joint variant calling and genotype refinement is an effective way of improving variant calls in sequencing data, particularly when the average read depth is low. Both methods were used successfully in the 1000 Genomes and UK10K projects to generate high-quality variant call sets, and when applied simultaneously to both case and control cohorts they are also able to help alleviate the variable sensitivity and specificity that can arise from systematic differences in sequencing coverage (Figure 2.4).

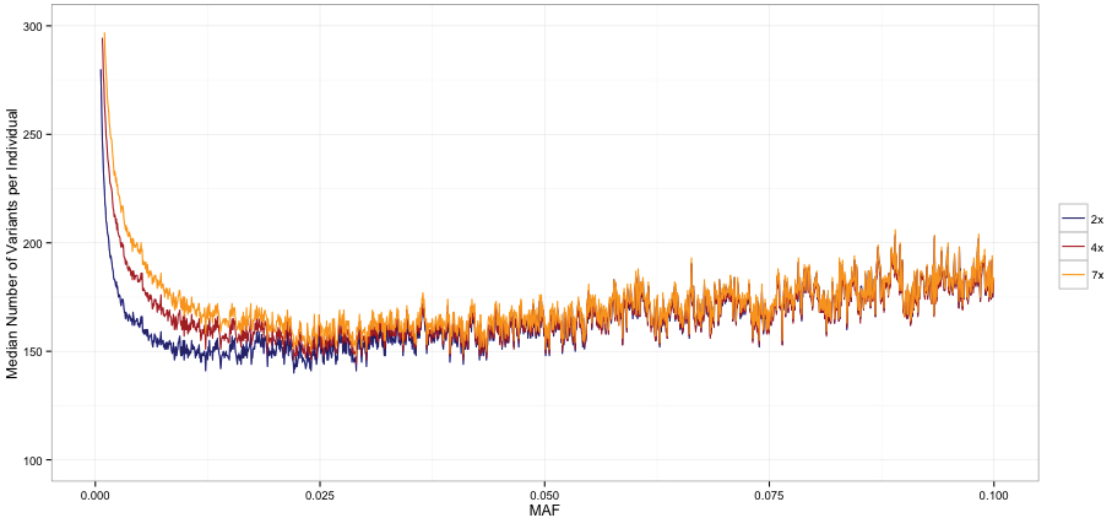


Figure 2.4: I investigate the effect of read depth on sensitivity and specificity across the allele frequency spectrum, for a range of average sequencing depths as shown with blue (2x), red (4x) and yellow (7x) lines. Variants have been jointly called across three cohorts (1,767 2x, 2,513 4x, and 3,652 7x samples), followed by genotype refinement. Sensitivity is then approximated as the median number of variants called per individual. Compared to rare variant calls, which do not have sufficient population-level information to be improved through joint calling and genotype refinement, the differences in sensitivity between each cohort have been notably improved for low frequency and common variation.

However, association testing using low frequency and common variation (MAF $\geq$ 0.1%) is still susceptible to residual bias due to sequencing depth. As will be discussed in more detail in section 3.5, despite using joint calling, genotype refinement, and very stringent quality control on our low coverage IBD sequences, there was still an excess of extremely significant sites ($P < 1 \times 10^{-15}$) falling outside

of known IBD-associated loci, the majority of which had a MAF $< 5\%$. Most (if not all) of these are likely to be false associations that simply reflect the greater number of observations in the higher coverage group due to better sensitivity, rather than any true effect on disease risk.

Although residual bias from sequencing depth differences can prevent case-control association testing of low frequency variation in differentially sequenced cohorts alone, these datasets still provide valuable imputation reference panels. With quality variant call sets produced using joint calling and genotype refinement, a set of haplotypes from across both cases and controls can be used to impute these variants into large panels of genotyped individuals. This approach not only increases sample size, and therefore power to detect associations, but will also produce case-control datasets that are not affected by the original coverage bias present in the sequenced reference panels. For example, imputation into GWAS was used successfully by a recent case-control association study of Type 2 diabetes to increase the utility of their low-coverage whole genome sequences (Fuchsberger et al., 2016).

## 2.4 Rare variant association testing

Because the minor allele of a given rare variant is observed so infrequently, methods that rely on the incorporation of population-level information, such as joint calling, genotype refinement, and imputation, cannot be usefully applied (Figure 2.4). This leads to two major issues when performing rare variant association studies in case-control cohorts. Firstly, testing can only be performed in directly sequenced individuals, limiting sample sizes. Given the scarcity of these variants in the population, obtaining a significantly large difference in minor allele frequency between cases and controls is simply not possible with achievable sample sizes. Secondly, any systematic bias in read depth between the cohorts cannot be overcome by processing the data prior to association testing, requiring new association test statistics that are tailored to this specific situation. I shall discuss the development

of an approach that can be used to address each of these problems in the following sections.

### 2.4.1 Increasing power using burden testing

Single-variant association tests can only be successfully applied to rare variants if the sample sizes are sufficiently large, or the variant effects are particularly strong. Because of this, rare variant association testing generally relies on the aggregation of signals from across multiple variants in order to increase power. The most common methods by which variants are aggregated and their cumulative effects are tested can be broadly broken into three categories: burden tests, variance-component tests, and combined tests (Lee et al., 2014b; Moutsianas and Morris, 2014). Depending on the underlying genetic architecture of the disease being tested, different methods will be better powered to detect an association (Table 2.1).

The simplest approach is to perform a burden test, which combines information across a number of variants in a target region (e.g. by counting the number of occurrences of each minor allele) and then tests the resulting summary score. However, such methods only work well if the majority of variants included are causal, and all have the same direction of association with the trait. One way to overcome these limitations is to use a variance component test, which compares the observed variance with the expected variance of the distribution of allele frequencies in a target region. If the variance is over-dispersed, meaning an increase from the expected binomial variance, this can indicate a subset of variants that are preferentially observed in either cases or controls (Figure 2.5). In this way, it is possible to efficiently test for a combination of effect directions (risk, neutral or protective), although this does come at the cost of reduced power if all variants do in fact act in the same direction (Neale et al., 2011).

Table 2.1: A comparison of current rare variant association testing methods, adapted from Lee et al. (2014b).

| Method | Advantages | Disadvantages | Examples |
|---|---|---|---|
| Burden tests | | | |
| Collapse multiple variants into a single summary score | Well-powered when a large proportion of the variants included are causal, and all have the same direction of association with the trait | Performs poorly in the presence of both risk, protective and null variants | CAST (Morgenthaler and Thilly, 2007)<br>CMC (Li and Leal, 2008)<br>ARIEL(Asimit et al., 2012)<br>MZ test (Morris and Zeggini, 2010)<br>WSS (Madsen and Browning, 2009) |
| Variance component tests | | | |
| Test for over-dispertion in the variance of genetic effects | More robust to the presence of both risk and protective variants, or if only a small proportion of included variants are causal | Less powerful than burden tests if most variants are in fact causal and operating in the same direction of effect | C-alpha (Neale et al., 2011)<br>SKAT(Wu et al., 2011)<br>SSU test(Pan, 2009) |
| Combined tests | | | |
| Linear combinations of burden and variance component tests | Useful for datasets when the genetic architecture is unknown, reduces the power loss associated with applying the incorrect model | Is less powerful than applying either a burden test or a variance component test to data where their respective assumptions about the genetic architecture hold | SKAT-O (Lee et al., 2012b)<br>Fisher method (Derkach et al., 2013)<br>MiST (Sun et al., 2013) |

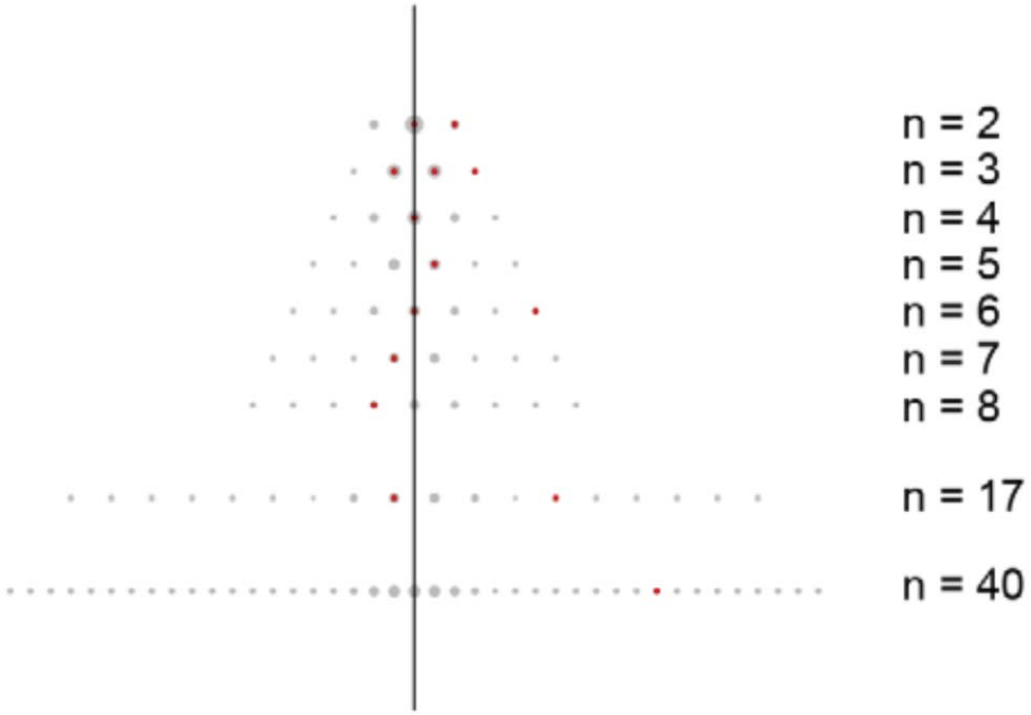


Figure 2.5: An example of the distribution of recurrent, low frequency non-synoymous variants in *NOD2*, comparing 350 CD cases to 350 controls. Each row defines variants observed n times in the dataset, with the observations split between controls (left of the vertical line) and cases (right of the vertical line). As an example, the n = 3 row describes three observed variants in red, one seen in 3 cases and 0 controls, one seen in 2 cases and 1 control, and one seen in 1 case and 2 controls. The variance component test determines if there is a difference in the variance of the observed data (red) and the binomial probability distribution (grey). Figure sourced from Neale et al. (2011).

While variance component tests are generally the preferred approach when faced with the aggregation of variable effect sizes and directions, their loss of power compared to simple burden tests when effect direction is consistent means that many people who are testing data of unknown genetic architecture will turn to tests that combine both burden and variance component approaches. Rather than simply applying each test separately and taking the minimum $p$-value, which can lead to an inflated type I error rate, these combined tests attempt to find the optimal linear combination of both the burden and variance-component tests (Lee et al., 2012b).

### 2.4.2 Accounting for differences in sensitivity and specificity between cases and control

In general, the rare variant association tests discussed above assume the case and control datasets have been well matched. In particular, the minor allele frequencies to be tested are derived directly from genotype calls, thereby assuming that these calls are equivalent for the two datasets. Unfortunately, when there are systematic biases in coverage between the cohorts this assumption does not hold. In practice, there is increased sensitivity to detect variation in the higher coverage group, and decreased specificity to avoid errors in the lower coverage group (Figure 2.6). This can lead to two types of false association signals: an excess of erroneous variants that have been called in the lower coverage group, and an excess of true variant calls in the higher coverage group that failed to be detected in the lower coverage cohort. Depending on how different subsets of these variants (which have opposing false signals) are selected for aggregation into a burden test, it is possible that significant false associations may be observed.

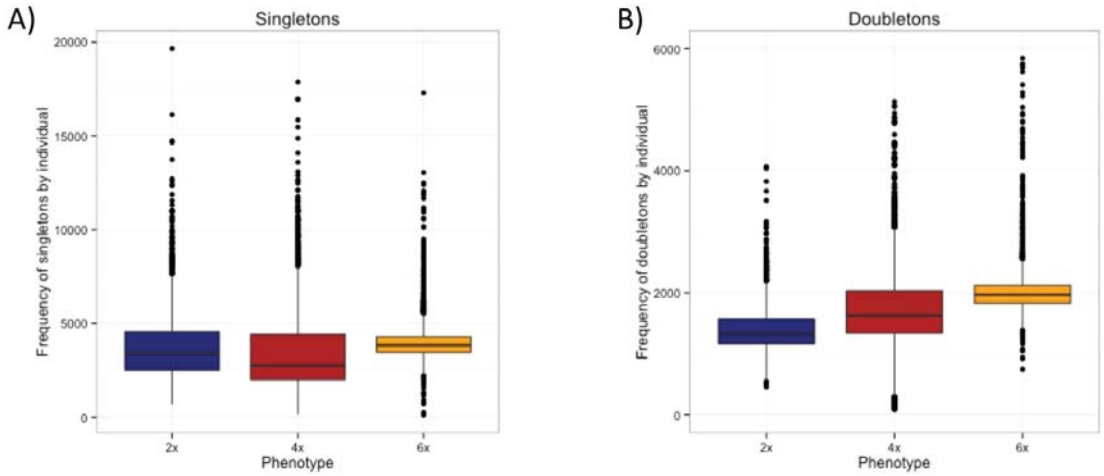


Figure 2.6: The effect of read depth on the sensitivity and specificity of calling genotypes for rare variants. Variants were jointly called across three cohorts (1,767 2x (blue), 2,513 4x (red), and 3,652 7x (yellow) samples), followed by genotype refinement. For singletons, which are observed just once in the population, there is an excess of variants called in the 2x cohort, indicating a loss of specificity at low coverage (panel A). For doubletons, which are observed twice in the population, we see a more general pattern of increasing sensitivity as read depth is increased (panel B).

One way to overcome this issue would be to down-sample the higher coverage group so that the average read depth is consistent across both cases and controls prior to variant calling, and then perform association testing using one of the standard methods from Table 2.1. However, this requires the removal of potentially useful sequence information. To avoid the loss of valuable data, another commonly proposed solution is to test for association using a logistic regression analysis that includes the read depth as a covariate (Garner, 2011), or weights variants based on quality scores (Daye et al., 2012). However, if the cases and controls can be perfectly separated by read depth then it cannot be used as a covariate, as it will cause the parameters of the logistic regression to no longer be estimable (Derkach et al., 2014).

Instead, the solution I use here is to account for known differences in the sensitivity and specificity of variant calling by replacing the hard genotype calls with genotype dosages. Rather than discrete counts of the minor allele, such that a genotype call for individual $i$ at position $j$ can be defined as $G_{ij} \in \{0, 1, 2\}$, the dosage is calculated as the expected genotype given the sequencing data $D$, such that $E\left(G_{ij}|D_{ij}\right) = \sum_{g=0}^{2} gP(G_{ij} = g|D_{ij})$. Here, $P(G_{ij} = g|D_{ij})$ is the probability of each genotype given the sequencing data. The resulting dosage estimate better reflects the confidence of a variant call, allowing for the effects of read depth to be incorporated into the test.

### Association testing using genotype dosage

Skotte et al. (2012) developed a score statistic that performs association testing using this genotype dosage data. Their statistic is derived from the joint likelihood of phenotype and sequencing data across all individuals at a given locus (Equation 2.1). This assumes that, across $n$ samples, for any one individual $i$ their phenotype $Y$ depends on the observed sequencing data $D$ through the unobserved genotype $G$ at locus $j$.

$$P\left(\mathbf{Y} = (Y_1, ..., Y_n), \mathbf{D} = (D_{1j}, ..., D_{nj})\right) = \prod_{i=1}^{n} \left( \sum_{g=0}^{2} P\left(Y_i|G_{ij} = g\right) P\left(G_{ij} = g, D_{ij}\right) \right) \quad (2.1)$$

The main component of interest in this likelihood is the relationship between the phenotype and the genotype, $P(Y_i|G_{ij} = g)$: if we were to consider $logit\left(P\left(Y_i|G_{ij} = g\right)\right) = B_0 + B_1 g$ then a test to determine if the slope is null ($H_0 : B_1 = 0$) can be used to indicate if there is any association between the two. $S_j$, the score statistic for $B_1$, has been derived in Equation 2.2, and has the variance as shown in Equation 2.3. The corresponding test statistic $T_j = \frac{{S_j}^2}{Var(S_j)}$ is chi-squared, with one degree of freedom. Under the null hypothesis, $S_j = 0$.

$$S_j = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) E\left(G_{ij}|D_{ij}\right) \tag{2.2}$$

$$Var\left(S_j\right) = \sum_{cases} \left(1 - \bar{Y}\right)^2 Var\left(E\left(G_{ij}|D_{ij}\right)\right) + \sum_{controls} \left(\bar{Y}\right)^2 Var\left(E\left(G_{ij}|D_{ij}\right)\right) \tag{2.3}$$

Importantly, the variance of $E(G_{ij}|D_{ij})$ is read depth dependent. Intuitively, as read depth increases the data will better reflect the true genotype, so that $E(G_{ij}|D_{ij})$ will approach the true $G_{ij}$ while $Var(E(G_{ij}|D_{ij}))$ approaches the true $Var(G_{ij})$. This is because we obtain less information about the true genotype at lower coverages, and thus the expected variance of the genotype given the data, $E(Var(G_{ij}|D_{ij}))$, is greater. At sufficiently high coverage, when we can consider the data to perfectly reflect the true genotype, this value should converge to 0. Therefore, by the law of total variances (Equation 2.4), estimating the variance of the true genotypes using $Var(E(G_{ij}|D_{ij}))$ will lead to an underestimate of this value at low depths.

$$Var(G_{ij}) = Var(E(G_{ij}|D_{ij})) + E(Var(G_{ij}|D_{ij})) \tag{2.4}$$

How this corresponds to the variance component of the test statistic depends on the relative depths and sample sizes of the two groups, as the group with the smallest sample size will contribute the most to the variance calculation, due to the inclusion of the average phenotype $\bar{Y}$ in the weights (see Equation 2.3). For

example, if we assume that the high coverage group has sufficient information to obtain reasonable variance estimates, while the lower coverage group does not, then when $N_{Low} >> N_{High}$ the variance component will be underestimated, while if $N_{High} >> N_{Low}$ the variance component may actually be overestimated. Underestimation of the variance component will lead to an overinflated test statistic, and vice versa.

Derkach et al. (2014) therefore proposed that, in the presence of systematic read depth differences between cases and controls, a more accurate test statistic could be obtained by calculating the variance components for the two groups separately (Equation 2.5).

$$\begin{aligned}\hat{V}ar\left(S_j\right) = N_{case} \left(\frac{N_{control}}{N}\right)^2 \hat{V}ar_{case} \left(E(G_{ij}|D_{ij})\right) \\ + N_{control}\left(\frac{N_{case}}{N}\right)^2 \hat{V}ar_{control}\big(E(G_{ij}|D_{ij})\big)\end{aligned} \tag{2.5}$$

This 'Robust Variance Score' (RVS) statistic can be extended to perform a burden test for multiple rare variants, using a similar approach as standard burden tests like CAST and CMC. The individual variant score statistics are simply summed together to given an overall score, while the variance component is calculated by combining the covariance matrices of the cases and controls, after estimating them separately. Unfortunately, however, the distribution of the resulting test statistic for the joint variant analysis is unknown. Instead, a permutation-style procedure needs to be used, whereby a $p$-value is generated by creating $X$ bootstrap samples and counting up the number of times they generate a test statistic that is more significant than the original sample. Usually, evaluating significance using permutation would involved randomly permuting case and control status, but the different read depths between the groups precludes this. Instead, both the case and control groups are separately centred around their respective means, and then (still separately) sampled with replacement from these centred values, maintaining the same numbers of cases and controls as the original sample. In this way, the difference between the groups is reduced to one dimension (variance only), forming

an empirical null set from which bootstrap samples can be generated without swapping case and control status (Derkach et al., 2014).

### 2.4.3 Testing in a dataset with systematic read depth bias between cases and controls

In order to test the performance of the RVS in the presence of a known systematic bias in read depth between cases and controls, I considered a low coverage whole genome sequencing study of inflammatory bowel disease. The sample collection, sequencing and quality control procedures used to generate this dataset will be described in more detail in Chapter 3. However, briefly, it consists of 1,767 patients with ulcerative colitis (median coverage of 2x), 2,513 patients with Crohn's disease (4x), and 3,652 population controls (7x).

#### Implementing the RVS statistic in C++

Testing a dataset of this size using the original R implementation of the RVS statistic as provided by Derkach et al. (2014) would lead to extensive computer memory demands and excessive run times, such that it was not possible even given the sizeable computational resources available at the Wellcome Trust Sanger Institute. I therefore had to first implement the RVS statistic as an extension to the software ANGSD (Korneliussen et al., 2014), which makes use of the compiled language C++ and multi-threading to generate much more efficient run times. My implementation can be found at https://github.com/katiedelange/angsd.

I developed the algorithm described in Box 2.1 to perform the RVS association test within the framework defined by ANGSD. I optimised this solution to minimise memory requirements (currently the most limiting resource within the cluster computer framework to be used for association testing) and made use of multi-threading in order to parallelise steps wherever possible.

Box 2.1: Algorithm used to implement the RVS statistic within the ANGSD framework.

```
// Request the following inputs from the user
    - The number of burn-in bootstrap resampling permutations to
      perform before significance is evaluated
    - The number of bootstrap resampling permutations to perform
      (-1 specifies that adaptive permutation should be used)

// Extract the relevant summary data from the genotype probabilities
For each site j
    For each individual i
        Compute and store the expected genotype
          E(G_ij|D_ij) = sum_g P(G_ij = g|D_ij), for g=0,1,2
        Compute and store the expected variance
          Var(G_ij = g|D_ij) = E(G_i^2|D_ij) - E(G_ij|D_ij)^2
    Determine the population allele frequency estimate
      (G_ij|D_ij)/2N across both samples at this site.

// Compute the score statistic components for the unpermuted sample
Append the burden score S to the list of scores
    S = sum_{j=0}^{N}(S_j), where S_j = sum(Y_i - Ybar)E(G_ij|D_ij)
Append the burden variance Var(S) to the list of variances
    Var(S) = sum_i sum_j sum_k cov(E(G_ij|D_ij), E(G_ik|D_ik)

// Centre the stored genotype dosages around their respective means
Separately for cases and controls
    For each site j
        Compute the mean expected genotype
        Subtract this from each individual using a matrix transform

// Run permutation testing to evaluate the significance of the test
For the requested number of permutations
    Separately for N_0 controls and N_1 cases
        Randomly sample N_(0,1) times (with replacement)
    Append the permuted sample score S to the list of scores
    Append the permuted sample variance Var(S) to the list of variances

// Return the fraction of times that a permuted sample is more
// significant than the original sample
```

### Performance of the RVS in systematically biased data

I tested the performance of the RVS burden test on rare (0.0001<MAF<0.01) functional coding variation within genes. I define functional coding variants to be those with one of the following Variant Effect Predictor (McLaren et al., 2010) annotations: frameshift_variant, stop_gained, initiator_codon_variant, splice_donor_variant, splice_acceptor_variant, missense_variant, stop_lost, inframe_deletion, or inframe_insertion. The MAF range used is also defined so as to exclude singletons, due to the lack of specificity at this frequency for very low coverage data (Figure 2.6). Despite these restrictions, I observe a very large excess of apparently significant associations after $10^6$ permutations (Figure 2.7), and systematic over-inflation of the test statistic ($\lambda = 1.34$).

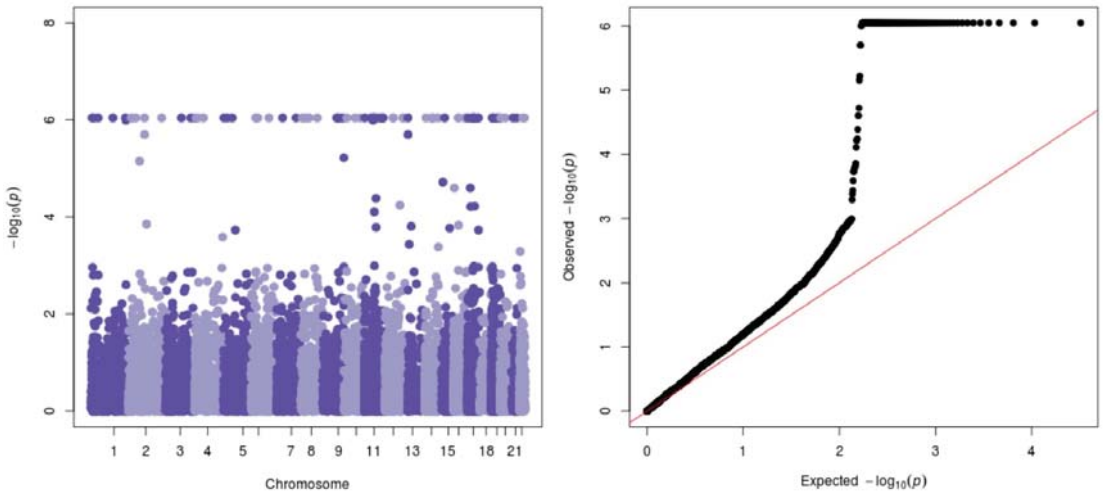


Figure 2.7: Burden testing using the RVS statistic (up to 1,000,000 permutations) on rare (0.0001<MAF<0.01) functional coding variation within genes.

When trying to determine why this statistic does not appear to be adjusting for the systematic coverage bias in this dataset as well as the authors suggest it will (Derkach et al., 2014), I note that there are a few crucial assumptions that must be met. In particular, the method assumes that 1) the variants being tested are true sites of variation; and 2) a variant has been successfully detected if it is present. However, particularly when dealing with rare variants in very low coverage datasets, it is likely that these assumptions will be violated at a number of tested sites. This

includes both errors that have been mistakenly included in the lower coverage group due to reduced specificity, and rare variants that have failed to be detected in the lower coverage group due to reduced sensitivity. I therefore looked to modify the standard sequencing quality control procedure that was applied to this test dataset (see Chapter 3 for details) to include additional filters tailored to rare variants, in order to both better remove potential errors and try to identify sites that, whilst true sites of variation, failed to be identified in one group due to low coverage (rather than disease association).

### 2.4.4 Adjusting the quality control procedures

#### Identifying variants sites that were missed at lower coverage

I first focus on trying to deal with rare variant calls that are likely to be true sites of variation, but were missed in the lower coverage group due to a lack of sensitivity. Hu et al. (2016) show that this particular problem can sometimes be overcome by modelling the error rate and using it to predict loci that are likely to be true variants. In particular, they aim to include the maximal set of possible variants in the test, applying only minimal filtering to try and remove sites that are predicted to be truly monomorphic in both datasets. This is done by screening out sites that are predicted to be uninformative, in that they have a score $S = 0$ and therefore do not contribute to the burden test. However, because this minimal screening step is unlikely to capture all problematic sites, they then adjust the permutation procedure to try and generate bootstrap datasets that have identical allele frequencies between cases and controls, but match the read depths, error rates, and the number of true variants and monomorphic loci that are seen in the original dataset. Unfortunately, this method relies on a sufficiently strong signal-to-noise ratio at very rare sites in at least one of the groups being tested, in order to properly model errors for the initial screening step. For situations where both cases and controls are of low coverage, this method is not expected to offer any significant advantages over Derkach et al's RVS model.

I therefore looked to capture these sites as part of the filtering process instead, by trying to measure how accurately a given site is likely to have been captured across all the individuals in each cohort. To do this, I calculate the INFO score $\alpha$, which can be interpreted as describing the amount of 'missing' information such that the observed data at a site is equivalent to a set of perfectly observed genotypes in a sample of size $\alpha N$ (Marchini and Howie, 2010), separately for each cohort. It is computed using the likelihood of the true population allele frequency $\theta_j$ at a given site $j$ if we had observed genotypes $G_{ij}$, as shown in Equation 2.6.

$$\mathcal{L}(\theta_j) = \prod_{i=1}^{N} \theta_j^{G_{ij}} (1-\theta_j)^{2-G_{ij}} \tag{2.6}$$

The score (first derivative) and information (second derivative) for this likelihood are shown in Equations 2.7 and 2.8, where N is the sample size, and $X = \sum_{i=1}^{N} G_{ij}$. The score reflects how sensitively $\mathcal{L}(\theta_j)$ depends on $\theta_j$, while the information describes how much information the observable variable $G_{ij}$ carries about $\theta_j$.

$$U(\theta_j) = \frac{d \log \mathcal{L}(\theta_j)}{d\theta_j} = \frac{X - 2N\theta_j}{\theta_j(1-\theta_j)} \tag{2.7}$$

$$I(\theta_j) = \frac{d^2 \log \mathcal{L}(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N - X}{(1-\theta_j)^2} \tag{2.8}$$

If we then consider that the genotypes $G_{ij}$ are not perfectly observable, but are instead approximated through the data $D_{ij}$, we can compute a similar likelihood for the allele frequency parameter $\theta_j$ that is integrated over the missing data that comes from estimating $G_{ij}$ using $D_{ij}$ (Equation 2.9). In order to do this, the data is partitioned into the observed data $Y_O$ and the missing data $Y_M$.

$$\mathcal{L}^*(\theta_j, Y_O) = \log(P(Y_O|\theta)) = \log \int P(Y_O, Y_M|\theta)\, \mathrm{d}Y_M. \tag{2.9}$$

The score and information of this observed data likelihood is heavily related to that of the full likelihood, as shown in Equations 2.10 and 2.11 (Louis, 1982).

$$U^*(\theta) = \frac{d\mathcal{L}^*(\theta_j)}{d\theta_j} = E_{Y_M|Y_O.G_{ij}}[U(\theta_j)] \tag{2.10}$$

$$I^*(\theta_j) = \frac{d^2\mathcal{L}^*(\theta_j)}{d\theta_j^2} = E_{Y_M|Y_O.G_{ij}}[I(\theta_j)] - V_{Y_M|Y_O.G_{ij}}[U(\theta_j)] \tag{2.11}$$

Of particular interest here is the information statistic, which we can use to describe the amount of missing information about the true allele frequency due to estimation using observed data as opposed to true genotypes. If we consider $I^*(\theta_j)$ to represent the observed information, and $E_{Y_M|Y_O.G_{ij}}[I(\theta_j)]$ the complete information, it follows that $V_{Y_M|Y_O.G_{ij}}[U(\theta_j)]$ is the missing information. These components can be calculated using Equations 2.12 and 2.13. Importantly, we can see that the top line of Equation 2.13 is actually calculating $Var(G_{ij}|D_{ij})$: as mentioned earlier, this converges to 0 as the read depth improves. Therefore, we expect more missing data in lower coverage samples.

$$E_{Y_M|Y_O.G_{ij}}[I(\theta_j)] = \frac{2N}{\hat{\theta}(1-\hat{\theta})} \tag{2.12}$$

$$V_{Y_M|Y_O.G_{ij}}[U(\theta_j)] = \frac{\sum_{i=1}^{N} E(G_{ij}|D_{ij}) - E(G_{ij}^2|D_{ij})}{\hat{\theta}^2(1-\hat{\theta})^2} \tag{2.13}$$

Using these two terms, we can compute the ratio of observed data to complete data (Equation 2.14), giving the INFO score $\alpha$ that can then be used to generate an effective sample size $\alpha N$ for the amount of informative data in the sample set at site $j$.

$$\alpha = \frac{E_{Y_M|Y_O.G_{ij}}[I(\theta_j)] - V_{Y_M|Y_O.G_{ij}}[U(\theta_j)]}{E_{Y_M|Y_O.G_{ij}}[I(\theta_j)]} \tag{2.14}$$

This INFO score provides an estimate of how well a variant has been captured across all the individuals in each cohort, and (as can be seen in Equations 2.12 and 2.13) is also closely related to the terms being tested by the RVS statistic. I

therefore computed this statistic for each site separately in each of the test cohorts, and plotted the distributions as shown in Figure 2.8.

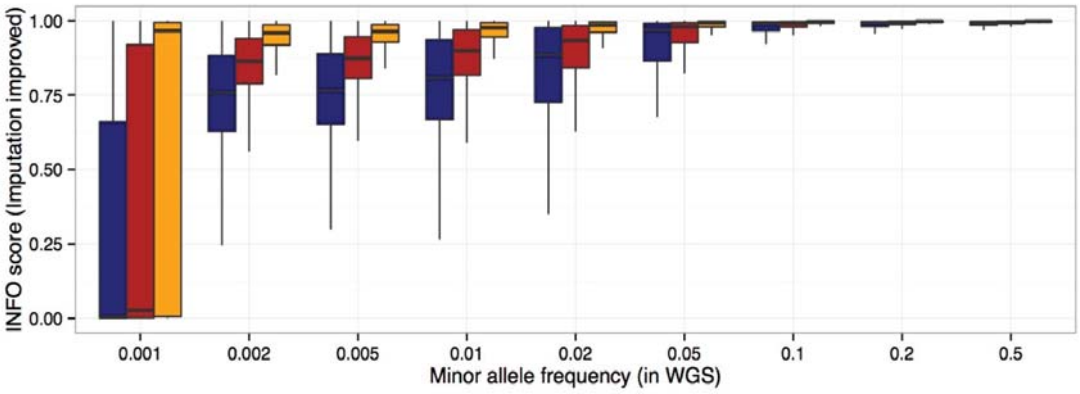


Figure 2.8: The distribution of the INFO score for 2x (blue), 4x (red) and 7x (yellow) data across a range of minor allele frequencies.

Immediately apparent are the large differences in median INFO scores between each of the cohorts below a minor allele frequency of $\sim 2\%$. This is particularly pronounced for very rare variants, where the datasets sequenced to 2-4x average coverage retain almost no information about sites with a MAF $< 0.2\%$. Given these observations, it is unsurprising that a score statistic calculated using datasets that are so distinct in their ability to capture the true genotypes resulted in such an excess of false positive associations. However, the extent to which each cohort differed on their median INFO measure, and how this changed between rare and common sites, was more unexpected.

One possibility is that this effect may be related to the use of genotype refinement via imputation, which is the major MAF-dependent factor affecting the genotype probabilities from which both the INFO score and RVS statistic are calculated. This process aims to remove noise and improve confidence in genotype calls made: in essence, producing a set of 'smoothed' genotype probabilities through the incorporation of population-level information. However when the true signal is low, such as for sites of rare variation, it may be that this refinement step is overzealous. To evaluate if this is the case, I investigated the use of genotype probabilities generated directly from the samtools Genotype Quality (GQ) field, without any genotype refinement.

The GQ value represents the phred-scaled genotype probability of the most likely genotype, as calculated by GQ$= -10 \log_{10} \max \left( P(G_{ij} = g | D_{ij}) \right.$ , for $g \in \{0, 1, 2\}\big)$. Unfortunately, this does not provide enough information to resolve all three possible genotype probabilities (homozygous reference, RR; heterozygous, RA; and homozygous alternate, AA). Therefore, in order to produce a set of genotype probabilities I assign the probability reflected in the GQ score to the genotype called in the VCF file, and all the remaining probability to the most likely alternate call:

$$P(\text{Call}) = 1 - 10^{\frac{-\text{GQ}}{10}}$$

$$P(\text{Alt}) = 1 - P(\text{Call})$$

$$P(\text{Remainder}) = 0$$

When the called genotype is homozygous, the next most likely genotype is assumed to be the heterozygous genotype (i.e. if Call=RR or AA, then Alt=RA). If the genotype call was heterozygous I assume, given the low MAF ($\leq 0.01$) of the variants being considered for burden testing, that the rare homozygote is not likely to be observed and thus I define the next most likely genotype as being homozygous reference (i.e. Call=RA, Alt=RR).

I compute these unrefined genotype probabilities across the complete dataset, and recalculate the INFO score separately for each of the three cohorts, across all sites. As can be seen in Figure 2.9, using unrefined genotype data leads to a dramatic improvement in the amount of information obtained at sites of rarer variation (MAF $\leq 2\%$). The utility of performing genotype refinement at common sites is also apparent, with improved INFO score distributions for higher MAFs (particularly MAF $\geq 10\%$, Figure 2.8).

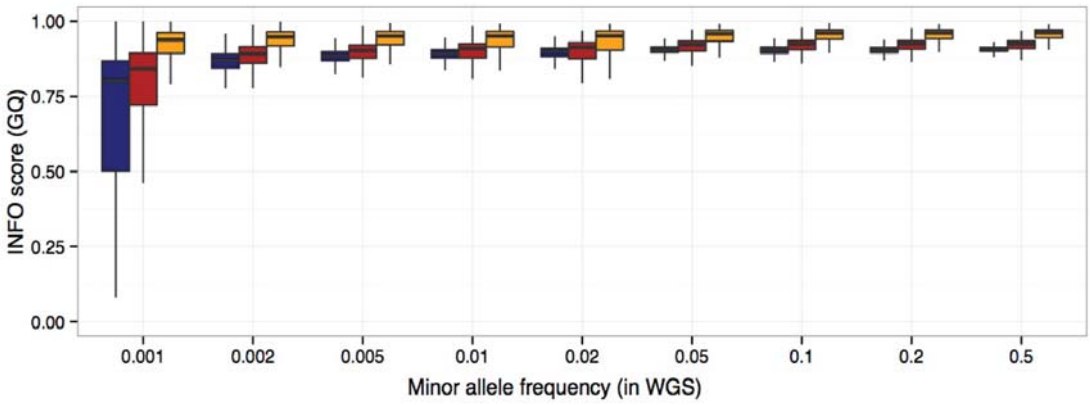


Figure 2.9: The distribution of the INFO score for 2x (blue), 4x (red) and 7x (yellow) data across a range of minor allele frequencies, using raw genotype probabilities estimated directly from the samtools genotype quality score.

In order to minimise the possible differences in INFO score between the case and control cohorts during association testing, and thus attempt to reduce the inclusion of rare variants that have been detected in the high coverage group but missed in the low coverage group due to reduced sensitivity, I filter out any sites with INFO $< 0.6$ in either of the relevant cohorts for each test. In general, this allows more sites to be retained when comparing the 4x cases (as opposed to the 2x cases) to the 7x controls.

### Additional error filtering

I then applied the following additional quality control filters, to try and reduce the number of erroneous sites included (particularly from the lower coverage group, which has poorer specificity during variant calling):

– Sites with a missingness rate $> 0.9$. When using unrefined genotype probabilities, the missingness rate across all sites is greatly increased, compared to the refined set that has attempted to infer a number of missing genotypes. I remove any sites with a high number of samples where a genotype could not be called.

– Sites with low confidence observations comprising $\geq 1\%$ of non-missing data. I define a low confidence observation as one with a maximum genotype probability $\leq 0.9$. This filter helps to capture sites where it is particularly difficult to confidently call variants, or where a large number of samples happen to have particularly low coverage.

– 'Uncertain' sites. These are sites that I first identified by analysing some of the most significant associations originally produced by the RVS, that did not lie in known IBD loci. In general, I noted a number of sites with low quality scores and a high proportion of individuals with a maximum genotype probability less than one (although not sufficiently low so as to be captured by the low-confidence filter described above). As can be seen in Figure 2.10, these sites have quite different distributions of genotype probabilities compared to high-quality sites. In order to systematically detect such variants, I used the output of five independent Support Vector Machines (SVMs) that were trained on 1,000 high-quality sites that overlapped with the HapMap3 dataset (Altshuler et al., 2010), and 1,000 poor-quality sites with a quality score $< 10$ in the raw VCF files. Any site with an SVM score $< 0.1$ in any of the five runs was removed.

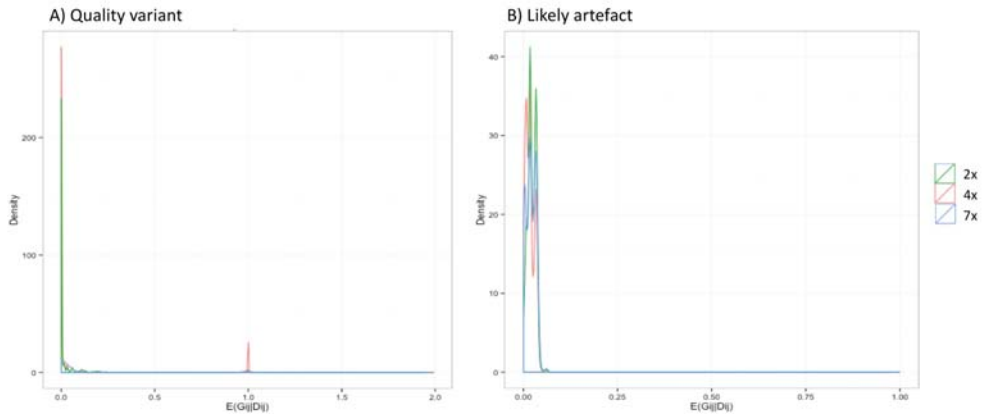


Figure 2.10: An example of a site captured to high quality (panel A), compared to a site with mostly low confidence genotype probabilities (panel B).

Using these additional quality control filters, and unrefined genotype probabilities, I repeated the RVS burden test on rare (0.0001<MAF<0.01) functional coding variation within genes. As can be seen in Figure 2.11, the Type I (false positive) error rate is now properly controlled and no systematic over-inflation of the test statistic is observed ($\lambda$=1.06).

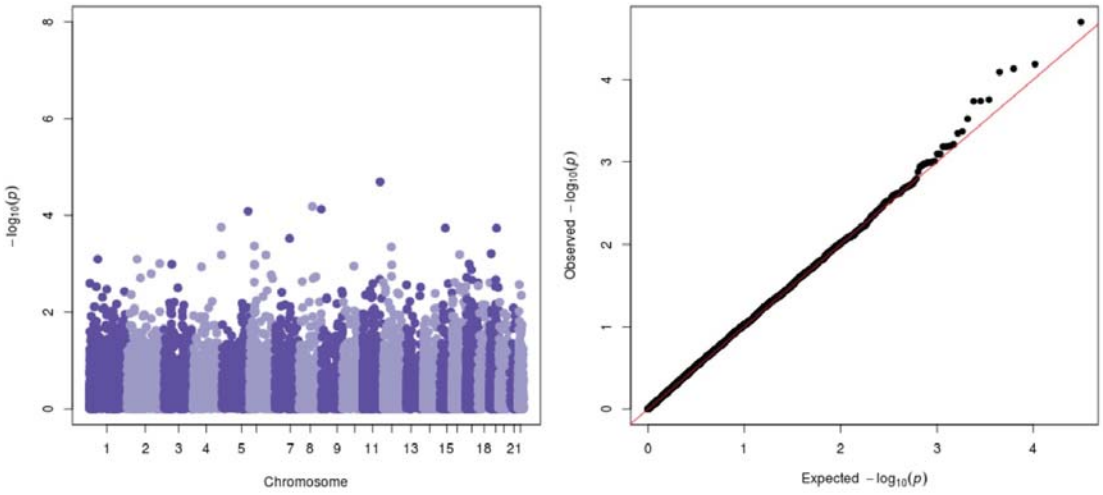


Figure 2.11: The performance of the RVS statistic in a dataset with systematic read depth bias between the cases (4,280 samples at 2-4x coverage) and controls (3,652 samples at 7x).

### 2.4.5 Increasing the size of the burden test

The logical extension of these gene-based rare variant burden tests is to combine individual tests together into larger, more powerful, gene set tests. However, the RVS statistic is a simple burden test, and does not account for potential differences in the direction of effect of its constituent variants. Within individual genes, one possible way to try and overcome this is to select for variation that is predicted to have a damaging effect on the protein, in the hope that all variation affecting a given gene will therefore act in the same direction. However, for larger gene set tests this is unlikely to help, particularly as previous research has already shown that loss of some genes will lead to an increase in risk, while loss of others will be protective. For example, if we consider just the two most strongly associated genes in IBD, variation in *NOD2* is risk-increasing, while variation in *IL23R* is risk-decreasing.

I therefore extended the RVS statistic to perform larger burden set tests using an enrichment procedure that allows for opposing directions of effect. For each gene (or other form of primary aggregation set, such as enhancers or promoters), the absolute scores are summed together to form an overall score statistic that is independent of effect direction. Overall variances are also summed together, meaning that whilst covariance is included when computing the variance component for an individual gene, the inter-gene covariance is not accounted for. This decision was made in order to greatly reduce the number of between-variant comparisons that were required, which generated massive improvements in the computational efficiency of this method. However, overall I expect the loss of inter-gene covariances to be of minimal consequence. In general, covariance is used to capture the effects of linkage disequilibrium between variants in the test, increasing the overall variance component of the test statistic when highly-correlated variants are present, in order to avoid over-estimating the significance of an association. It is therefore retained for individual gene tests, where all included variation is in very close proximity, but overall it is expected to be relatively small given the rarity of the variants being tested (and therefore their low correlation with other variation in the region). For gene set tests in particular, where many of the contributing genes are not even on the same chromosome, linkage disequilibrium between variants from different genes should be very low.

The resulting set statistic is then divided by the equivalent statistic produced using the set consisting of all genes, in an approach based on the SMP method devised by Purcell et al. (2014). Accounting for the exome-wide statistic in this way helps to remove any residual case-control coverage bias that may accumulate over the large numbers of variants contributing to these gene set tests. Significance is evaluated using permutation testing, where individual gene statistics are re-computed in bootstrapped samples (with the exact same samples drawn for every gene during each permutation round) and summed to produce both set and exome-wide permutation statistics.

## 2.5 Discussion

Large-scale sequencing studies such as the Exome Aggregation Consortium (Lek et al., 2016), the 1000 Genomes project (1000 Genomes Project Consortium et al., 2015), and the UK10K project (Walter et al., 2015) have revealed important insights into human population biology, and provided vital resources for interpreting the clinical relevance of variation. However, they have also highlighted the practical difficulties associated with combining multi-source sequencing data at scale, as systematic biases in exome capture technology and sequencing depth lead to crucial sensitivity and specificity differences when performing variant calling. As researchers now look to extend the success of these cohort studies to investigate genetic disease risk using large case-control comparisons, the effects of these systematic biases can be observed as a slew of false associations.

In this chapter, I have described various methods that can be used to overcome systematic biases in read depth in a case-control setting, in order to prevent over-inflation of the test statistic and tightly control the Type I error rate. While the effects of sequencing coverage can be largely overcome at sites of low frequency variation, through joint calling of variants followed by genotype refinement, ultimately disease associations for such variants are best tested by imputing them into the wealth of existing GWAS cohorts currently available. Not only does this increase sample size, and therefore power to detect association, but the resulting imputed sequences will not be affected by any of the systematic sequencing biases present in the original cohorts.

For rare variation, which is poorly correlated with nearby variation and therefore cannot be accurately imputed, studies must be performed in the directly sequenced data. As the rare allele for these sites is observed so infrequently in the population, joint calling and genotype refinement offer little power to alleviate the effects of sequencing depth on the sensitivity and specificity of variant calling. Rare variant association testing in the presence of systematic read depth bias between cases and controls therefore required the development of a novel approach that accounts for this bias directly in the association test.

To this end, I implemented the RVS statistic described by Derkach et al. (2014), which adjusts for read depth bias by using genotype dosages (as opposed to hard genotype calls) and calculating the variance component of the test statistic (which is read depth dependent) separately for cases and controls. I then test the performance of this statistic in real data, using cases that had been sequenced at 2-4x average coverage, and controls that were sequenced to 7x. Unfortunately, when using a standard sequencing processing and quality control pipeline, this statistic failed to control the Type I error rate. However, I overcame this problem by reverting to the use of unrefined genotype probabilities, as the genotype refinement process is overzealous when acting upon sites of rare variation, and applying additional quality control filters. Using these adjustments, the number of false positive associations when performing rare variant burden testing across genes can be well controlled, and no systematic over-inflation of the test statistic is observed.

This process has emphasised the difficulties associated with performing large-scale sequencing studies, particularly in a case-control setting. However, I have also shown that, through the use of carefully chosen methods and very stringent quality control, it is possible to perform association testing on this scale even in the presence of systematic read depth bias between cases and controls. This analysis proves that it is feasible for researchers to cost-effectively investigate the role of low frequency and rare variation in genetic disease risk by combining their own sequenced cases with large, publicly-available control datasets.