Chapter 5

High-throughput genomic studies of multiplex families

5.1 Introduction

The previous two chapters have discussed methods for mapping and interpreting disease associations in unrelated case/control cohorts. This has proven extremely successful at discovering common risk loci, including a large number of risk alleles for inflammatory bowel disease (IBD). However, case-control studies, using genotyping chips, are far from the only method of studying genetic risk.

As I discussed in the introduction, there are many types of risk variant that case-control GWAS studies are not well suited to study. In particular, the tag SNP approach is poorly powered to detect very low frequency

variants, even if they have large effect sizes. The rise of next-generation sequencing, however, gives us the opportunity to directly assay such variants via whole genome or whole-exome sequencing. The question is how to distinguish the (very small) number of causal risk variants from the (very large) number of low-frequency variants that have no effect on disease.

One potentially powerful tool is the study of multiplex (or multiply affected) families. Multiplex families have long been a staple of human disease genetics, and are the starting point for both the heritability and linkage studies that underlie much of our knowledge of complex disease. In recent years family studies have fallen out of favour in complex disease genetics as a result of the relatively poor performance of linkage studies and the success of GWAS. However, as we will see, multiplex families are more likely to harbour rare, high penetrance causal variants than unrelated cases. Furthermore, the fact that these variants are shared across multiple affected individuals gives us information that can allow us to whittle down the list of candidate variants by focusing only on those that are shared by many affecteds within the family.

I will start this chapter with a brief discussion of the history of multiplex family studies in complex disease (section 5.2). This section will also outline the approach to studying multiplex families that I describe in this chapter, in the context of the studies that have come before. I will then introduce some statistical models for analysing multiplex families in terms of high penetrance and polygenic risk factors (section 5.3). This will lead to the introduction of a new method for prioritising multiplex families that are most likely to carry a high penetrance mutation, using GWAS risk variants.

Section 5.4 will discuss a large multiplex family with over 40 family members suffering from IBD, collected with the aim of identifying rare causal mu-

tations. We have performed a detailed genetic investigation into this family, using targeting and whole-genome genotyping, as well as whole-exome and whole-genome sequencing. I will discuss the known risk in this family, and explore the linkage and haplotype evidence for association. I will then describe the analysis of the sequencing data, calling SNPs, indels and structural variants, and combining them with the linkage information. Finally, I will describe a filtering procedure designed to identify candidate causal variants on the basis of their frequency, function and segregation within the family. This identifies a total of 120 candidate variants, including coding and regulatory SNPs and indels, and structural variants.

In the final section (section 5.5) I will describe a validation and replication experiment designed to discover which of these candidate variants may be causal. I will describe the error modes that can create false candidates, and how they can be counteracted. Finally I will describe three methods for genetically replicating these associations, including using case-control cohorts, unaffected siblings and other multiplex families, and explore the power of these approaches.

5.2 A history of multiplex family studies in complex disease

It was the existence of multiply affected families that first led scholars to begin investigating what we now call disease genetics. At the turn of the 19th century John C. Otto published an extensive pedigree analysis of a haemophilic family in New Hampshire, tracing it back for three generations (Raabe, 2008). He also hypothesised that haemophilia may be traceable to only a few pilgrim families, the first description of what would now be called a founder effect. It is this very concept of family and population specific causal mutations that underlies the research in this chapter.

Studies of disease families were a focus of many Victorian scientists. Both French physician Paul Broca, and the English surgeon James Paget documented many multiplex cancer families, leading to the first studies into familial aggregation in what is now called complex disease (Schneider et al., 1986). While we may see the roots of the modern concept of family history in these developments, in other fields the recognition of familial clustering took a darker and more ideologically driven form. For instance, the hereditary degeneracy theories of psychiatric disease in late 19th century France fed rapidly into contemporary prejudices about the mentally ill that lay far from modern concepts of medical care (Dowbiggin, 1991).

Paget specifically argued that these cancer families were the result of a hereditary factor, but both he and Broca noted that the high (and not precisely known) prevalence of cancer made it difficult to rule out these families as merely chance occurrences. The reality of familial clustering was only established with the rise of systematic epidemiological studies, and the statistical frameworks required to analyse them, at the start of the 20th century.

As in the 19th century, studies of cancer lead the way (Schneider et al., 1986), and these studies came of age when the pioneering epidemiological studies of Janet Lane-Claypon (Lane-Claypon et al., 1926) conclusively demonstrated an enrichment of familial clustering in cancer. Even at this stage the genetic studies were informing biological knowledge: familial clustering was shown to occur strongly in cancer at a single location (in particular breast cancer), but only weakly in cancers from distinct locations, highlighting the importance of considering cancers of different tissues as distinct diseases.

Despite having (as we know now) a higher heritability, the study of multiplex families in inflammatory bowel disease developed later. This is partly because the current diagnostic landscape of IBD solidified later: while diagnoses of IBD stretch back to the 19th century, the distinct diagnoses of ulcerative and Crohn's colitis emerged only at the beginning of the 20th century (Kirsner, 1995). The existence of families with multiple affected individuals was noted from 1906, and nuclear families with three or more affecteds were documented from the 1930s (Kirsner, 1995). However, it was not until the 1960s and the advent of twin studies (Kirsner, 1973) that a hereditary role for IBD was widely accepted. Around this time the existence of very large IBD families in the Jewish population began to be noted, with a particularly striking family with seven affected members being reported in 1963 (Sherlock et al., 1963).

In the latter half of the 20th century, family history was recognised as the single strongest known predictor of IBD (Satsangi et al., 1997). Many collections of multiplex families were made during this time: in 2004 Russell and Satsangi (2004) reviewed studies of 19 distinct multiplex IBD family collections. These studies were important in establishing the broad strokes of IBD genetics. They gave the first indication that CD and UC were genet-

ically distinct, yet related, diseases. Furthermore, they hinted at significant substructure within IBD aetiology, by demonstrating a genetic effect on disease location, and suggesting a genetic role in disease progression. Overall, family studies established IBD as a complex genetic disease, comparable in heritability to other immune-mediated diseases such as type 1 diabetes and multiple sclerosis.

Around the turn of the 21st century, linkage studies of multiplex IBD families led to the identification of the major IBD susceptibility loci *NOD2* (Hampe et al., 1999; Hugot et al., 2001) (in CD) and HLA (Williams et al., 2002) (in UC). However, large meta-analyses of linkage studies, including nearly 2000 families, failed to identify further genome-wide significant loci (van Heel et al., 2004), and even had difficulty consistently replicating the (by then fine-mapped) *NOD2* locus. This ultimately led to the replacement of family-based methods with genome-wide association studies (a phenomenon reviewed in Chapter 1).

The failure of linkage meta-analysis in IBD showed that IBD is not caused solely by high penetrance alleles at a small number of loci. However, it does not imply that high penetrance alleles do not exist; only that, if they do exist, they are individually at low frequency and are located in a number of different loci (so-called locus heterogeneity). Indeed, many of these multiplex families are likely to harbour high penetrance mutations, which can potentially be detected via their co-segregation with disease status within that family. It was this approach that identified mutations in the IL10 receptor subunits as an important contributor to early onset IBD (Glocker et al., 2009a).

Recent developments in whole-genome and whole-exome sequencing have opened up new avenues for the discovery of high penetrance causal variants. The power of this approach was demonstrated with the discovery of the gene

underlying the previously unsolved Mendelian disease Miller syndrome (Ng et al., 2010). This study used whole-exome sequencing of four patients, combined with filtering based on databases of common variation and software for predicting the severity of coding mutations, and identified candidate causal mutations in the gene *DHODH*. Over recent years, this approach has become the dominant means of solving Mendelian diseases (Bamshad et al., 2011), and has even been used to identify mutations that underlie syndromic forms of IBD (Worthey et al., 2011a; Fiskerstrand et al., 2012).

Given the success of this sequencing approach, we would like to also use it to identify penetrant mutations in multiplex families with complex IBD. However, there are a number of challenges in generalising this approach. Firstly, there is no guarantee that any given affected individual, and even any given multiplex family, will carry a penetrant mutation. Ideally we would like to sequence families that are likely to carry such mutations, and thus we require methods to decide which families to select for study. Secondly, even if a causal mutation is present in a family it is unlikely to be fully penetrant. Likewise, because the disease is relatively common compared to Mendelian diseases some family members may have the disease despite not carrying the mutation (so-called "phenocopies"). We thus need methods that can discover such mutations in families that may include both affected non-carriers and unaffected carriers. Finally, as we saw in Chapter 4 many common IBD risk variants lie in regulatory rather than coding regions, and it is possible that this will also be true for rare risk variants. We would thus like to generalise the variant prioritisation procedure to include potential non-coding candidate risk variants.

5.3 Modelling and controlling polygenic risk in multiplex families

There are many potential factors that can lead to familial aggregation in a disease without leading to families suitable for locus mapping. An obvious reason (and one that has been discussed since the 19th century) is chance co-occurrence: the large number of families in the world makes it likely that there exist families that have a large number of affecteds despite the absence of an underlying genetic risk factor. This effect can be additionally confounded by uncertainty in the prevalence, or population stratification, both of which could inflate the chance of seeing multiplex families by chance. For instance, the higher prevalence of IBD in individuals of Ashkenazi Jewish individuals will lead to a larger number of multiplex families in the Jewish population, even if the increased risk in this group was entirely due to environment.

Additionally, a shared exposure to an environmental risk factor can lead a family to develop a higher incidence than would be expected by chance. Diagnostic bias can also lead to familial clustering, as a strong family history may lead to more vigilant screening or overdiagnosis (this is particularly likely to occur for diseases with a high rate of undiagnosed cases, such as prostate cancer (Fleshner, 1995)). These non-genetic causes all highlight the importance of careful screening of multiplex families to establish a genetic cause.

Furthermore, for the purposes of mapping loci an excess of familial aggregation as a result of genetics may not be enough to make a family useful for study. It is now becoming clear that a substantial portion of the heritability of complex traits is due to highly polygenic risk. Williams et al. (2002)

estimated the contribution of polygenic risk in three complex diseases in the Wellcome Trust Case-Control Consortium data, by applying a linear mixed-model method. This gave lower bounds on the liability-scale variance due to polygenic risk from common loci of 22% for Crohn's disease, 31% for Type I Diabetes and 38% for Bipolar Disorder. In many cases a significant minority of this polygenic risk has already been characterised, for example via the 193 independent IBD risk factors identified via the IIBDGC Immunochip study (see Chapter 4), but much still remains undiscovered.

The risk variants that make up this polygenic risk each have a small effect size, and thus are unlikely to individually co-segregate with affection status in multiplex families. They are therefore outside of the scope of what can be studied by sequencing families. However, it will contribute to familial aggregation of cases within multiplex families, creating another class of families that need to be excluded from family sequencing studies.

A good first stage in understanding the impact of polygenic and penetrant risk on multiplex families is to construct and examine theoretical models of risk in families. Recent theoretical studies have investigated models of high penetrance mutations (Al-Chalabi and Lewis, 2011), as well as models of continuous polygenic risk (Yang et al., 2010) in multiplex families. However, to answer questions about the relative contribution of penetrant and polygenic risk, we need to construct a model that contained both elements.

In this section, I will develop a model of genetic risk that combines a polygenic risk with the presence of dominant, high penetrance alleles, and study how different parameterisations of this model (corresponding to different heritabilities, prevalence and balances of polygenic/penetrant risk) alter the distribution of affecteds in multiplex families. I will also develop and test a method for performing genetic risk prediction in a partially genotyped

pedigree, and using such risk prediction to prioritise multiplex families that are likely to carry high penetrance mutations over those that are likely to carry only polygenic risk.

5.3.1 A combined polygenic/penetrant model of multiplex families

To describe the combined polygenic/penetrant model of genetic risk, I will first lay out the two components: a liability threshold model for polygenic risk due to common variants of low effect, and a dominant Mendelian model for higher penetrance variants. I will then combine these two models together to produce a general model of which both component models are special cases.

Throughout this section I will consider a nuclear family, with two parents denoted by subscripts m and f (for mother and father, treated as interchangeable), and O offspring denoted by subscripts $c_i : i = 1, ..., O$. I will use indicator variables d_i to denote the affection status of individuals. I use a parameter K to denote the disease prevalence in the population.

The polygenic model

We model the polygenic component of the disease using a liability threshold model (as described in Chapter 2). To recap, each individual in the family is given a liability $L_i = A_i + E_i$, where the genetic liability $A_i \sim N(0, h^2)$ is an additive polygenic component of risk, and the environmental liability $E_i \sim N(0, 1 - h^2)$ is an (individual-specific) environmental component. h^2 is called the heritability of liability, and measures the proportion of liability that is shared by identical twins: as this model assumes additive polygenic risk, h^2 is also the narrow-sense heritability. An individual is affected (i.e.

$d_i = 1$) if $L_i > T$, where T is the liability threshold $T = \Phi^{-1}(1-K)$ and Φ is the cumulative distribution function of the standard normal distribution.

The liabilities for each family member are

$$L_m = A_m + E_m \tag{5.1}$$

$$L_f = A_f + E_f \tag{5.2}$$

$$L_{c_i} = A_{c_i} + E_{c_i} = \frac{1}{2}(A_m + A_f) + M_{c_i} + E_{c_i}, \tag{5.3}$$

where $M_{c_i} \sim N(0, h^2/2)$ is a Mendelian segregation term. We can reformulate these equations in terms of $4 + O$ standard normal variables Z_i,

$$L_m = hZ_1 + \sqrt{1-h^2}Z_3 \tag{5.4}$$

$$L_f = hZ_2 + \sqrt{1-h^2}Z_4 \tag{5.5}$$

$$L_{c_i} = \frac{h}{2}(Z_1 + Z_2) + \sqrt{1 - \frac{h^2}{2}}Z_{c_i}, \tag{5.6}$$

The probability of an individual having disease state d_i given a genetic liability a_i is given by

$$P(d_i|A_i) = \begin{cases} \Phi(\frac{T-A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 1; \\ 1 - \Phi(\frac{T-A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 0 \end{cases}. \tag{5.7}$$

We can write down a similar expression conditional on parental genetic liabilities

$$P(d_c|A_m, A_f) = \begin{cases} \Phi(\frac{T-(A_m+A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } d_i = 1; \\ 1 - \Phi(\frac{T-(A_m+A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } d_i = 0 \end{cases}. \tag{5.8}$$

The probably mass function for a set of affection statuses $\vec{d} = (d_m, d_f, d_{c_1}, ..., d_{c_O})$ is thus given by

$$\begin{aligned} P(\vec{d}) &= \iint_{-\infty}^{\infty} P(d_m|hz_1)P(d_f|hz_2)\phi(z_1)\phi(z_2) \\ &\quad \times \prod_{i=1}^{O} P(d_{c_i}|hz_1, hz_2)dz_1dz_2. \end{aligned} \tag{5.9}$$

Because siblings are interchangeable and independent conditional on parental genetic liabilities, we can model the number of affected offspring using a binomial distribution. The joint probability of observing parent genotypes (d_m, d_f), and also observing a total of y_c affected offspring is thus

$$\begin{aligned} &P(d_m, d_f, \textstyle\sum d_{c_i} = y_c) = \\ &\textstyle\iint_{-\infty}^{\infty} P(d_m|hz_1)P(d_f|hz_2)\phi(z_1)\phi(z_2)\binom{O}{y_c} \\ &\times P(d=1|hz_1, hz_2)^{y_c} P(d=0|hz_1, hz_2)^{O-y_c} dz_1dz_2. \end{aligned} \tag{5.10}$$

Finally, because parents are interchangeable, we can write down the probability of observing y total affecteds in the family (including parents and children) as

$$\begin{aligned} P(\textstyle\sum d = y) = \quad & P(d_m = 1, d_f = 1, \textstyle\sum d_{c_i} = y - 2) \\ & + 2P(d_m = 1, d_f = 0, \textstyle\sum d_{c_i} = y - 1) \\ & + P(d_m = 0, d_f = 0, \textstyle\sum d_{c_i} = y). \end{aligned} \tag{5.11}$$

The dominant penetrant model

The dominant penetrant model assumes that a large number of individually rare variants exist in the population, each of which has a dominant effect with intermediate penetrance. Certain diseases are known to show such a heterogeneity of genetic architecture, for instance in diabetes (Molven and Njølstad, 2011) and breast cancer (Chen and Parmigiani, 2007), and it is possible this is true for other diseases.

This model assumes that a proportion R of cases have a dominant mutation with a penetrance of $\pi > K$. The total combined frequency of these mutations is thus KR/π (and therefore $\pi/R > K$): note that this is the proportion of people who carry at least one mutation, not the allele frequency. We will use the indicator variable $r_i = 1$ to denote that individual i carries a mutation, and assume that each individual carries at most one mutation.

The disease probabilities, conditional on genotype, are given by

$$P(d_i = 1|r_i = 1) = \pi \tag{5.12}$$

$$P(d_I = 1|r_i = 0) = \frac{K(1-R)}{1-KR/\pi}, \tag{5.13}$$

and transmission probabilities from parents to child are given by

$$P(r_{c_i} = 1|r_m = 0, r_f = 0) = 0 \tag{5.14}$$

$$P(r_{c_i} = 1|r_m = 1, r_f = 0) = \frac{1}{2} \tag{5.15}$$

$$P(r_{c_i} = 1|r_m = 1, r_f = 1) = \frac{3}{4}. \tag{5.16}$$

We can combine these two together to give disease probabilities condi-

tional on parental genotype

$$P(d_{c_i} = 1|r_m = 0, r_f = 0) = \frac{K - KR}{1 - KR/\pi} \quad (5.17)$$

$$P(d_{c_i} = 1|r_m = 1, r_f = 0) = \frac{K + \pi - 2KR}{2(1 - KR/\pi)} \quad (5.18)$$

$$P(d_{c_i} = 1|r_m = 1, r_f = 1) = \frac{K + 3\pi - 4KR}{4(1 - KR/\pi)}. \quad (5.19)$$

As with the polygenic model, offspring are interchangeable and independent conditional on parental genotype, so again we model the number of affected offspring binomially:

$$P(\sum d_{c_i} = y_c|r_m = 1, r_f = 0) =$$
$$\binom{O}{y_c} [P(d_{c_i} = 1|r_m, r_f)]^{y_c} [1 - P(d_{c_i} = 1|r_m, r_f)]^{O-y_c}. \quad (5.20)$$

We can then incorporate parental affection status, conditional on genotype, into the total count of affecteds y

$$P(\sum d_i = y|r_m, r_f) =$$
$$P(d_m = 1, d_m = 1|r_m, r_f)P(\sum d_{c_i} = y - 2|r_m, r_f)$$
$$+P(d_m = 1 \text{ or } d_f = 1|r_m, r_f)P(\sum d_{c_i} = y - 1|r_m, r_f)$$
$$+(1 - P(d_m = 1|r_m))(1 - P(d_f = 1|r_m))P(\sum d_{c_i} = y|r_m, r_f), \quad (5.21)$$

where

$$P(d_m = 1 \text{ or } d_f = 1 | r_m, r_f) =$$
$$P(d_m = 1|r_m) + P(d_f = 1|r_f) - 2P(d_m = 1|r_m)P(d_f = 1|r_f). \quad (5.22)$$

Finally we marginalize out parental genotypes using the population frequency

$$\begin{aligned} P(\sum d_i = y) = & \left(\tfrac{KR}{\pi}\right)^2 P(\textstyle\sum d_i = y | r_m = 1, r_f = 1) \\ & + 2\tfrac{KR}{\pi}(1 - \tfrac{KR}{\pi}) P(\textstyle\sum d_i = y | r_m = 1, r_f = 0) \\ & + (1 - \tfrac{KR}{\pi})^2 P(\textstyle\sum d_i = y | r_m = 0, r_f = 0). \end{aligned} \quad (5.23)$$

The combined polygenic/dominant penetrant model

The combined model takes into account both polygenic risk and the presence of penetrant dominant risk alleles. To do this we set two thresholds, one for non-carriers for the dominant risk alleles $T_{wt} = \Phi^{-1}\left(1 - \frac{K-KR}{1-KR/\pi}\right)$, and one for carriers $T_{dom} = \Phi^{-1}(1-\pi)$. We then model transmission of both the penetrant risk alleles and a continuous liability.

The continuous liability is again given as $L_i = A_i + E_i$, where the genetic liability $A_i \sim N(0, h_d^2)$ only includes heritability due to common variants, excluding the rare penetrant mutations. This polygenetic heritability is given by $h_p^2 = h^2 - h_d^2$, where $h_d^2 = \frac{\sigma_d^2}{1+\sigma_d^2}$ is the variance explained on the liability scale by the penetrant risk alleles, where

$$\sigma_d^2 = \frac{KR}{\pi}\left[T_{dom} - T\right]^2 + (1 - \frac{KR}{\pi})\left[T_{wt} - T\right]^2. \quad (5.24)$$

Note that $h_d^2 \to 1$ as $\pi \to 1$ and as $R \to 1$.

We now specify the disease probability conditional on both the polygenetic liability (A_i) and the presence of absence of a penetrant mutation (r_i)

$$P(d_i = 1|A_i, r_i) = \begin{cases} \Phi(\frac{T_{dom} - A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 1 \text{ and } r = 1; \\ \Phi(\frac{T_{wt} - A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 1 \text{ and } r = 0 \end{cases} . \quad (5.25)$$

Again we can give a child's disease probability conditional on the genetic liability and presence of penetrant mutations in the parents, by taking into account the multiple thresholds with different transmission probabilities

$$P(d_{c_i} = 1|A_m, A_f, r_m, r_f) =$$
$$\begin{cases} \Phi(\frac{T_{wt} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } r_m = 0 \text{ and } r_f = 0; \\ \frac{1}{2}\Phi(\frac{T_{dom} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) + \frac{1}{2}\Phi(\frac{T_{wt} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } r_m = 1 \text{ xor } r_f = 1; \\ \frac{3}{4}\Phi(\frac{T_{dom} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) + \frac{1}{4}\Phi(\frac{T_{wt} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } r_m = 1 \text{ and } r_f = 1 \end{cases} \quad (5.26)$$

As before, we can write down the probability of observing y_c affected offspring given parental genotypes by modelling the number of affecteds as a binomial

$$P(d_m, d_f, \sum d_{c_i} = y_c|r_m, r_f) =$$
$$\iint_{-\infty}^{\infty} P(d_m|h_p z_1, r_m) P(d_f|h_p z_2, r_f) \phi(z_1)\phi(z_2)$$
$$\times \binom{O}{y_c} P(d = 1|h_p z_1, h_p z_2, r_m, r_f)^{y_c}$$
$$\times (1 - P(d = 1|h_p z_1, h_p z_2, r_m, r_f)^{y_c})^{O - y_c} dz_1 dz_2. \quad (5.27)$$

We then include parental affection status to give the probability mass

function for the total number of affecteds y given parental genotypes

$$\begin{aligned} P(\sum d = y|r_m, r_f) = \quad & P(d_m = 1, d_f = 1, \sum d_{c_i} = y - 2|r_m, r_f) \\ & +2P(d_m = 1, d_f = 0, \sum d_{c_i} = y - 1|r_m, r_f) \\ & +P(d_m = 0, d_f = 0, \sum d_{c_i} = y|r_m, r_f), \end{aligned} \quad (5.28)$$

and finally we marginalize out parental genotypes using the population frequency to give the final probability mass function

$$\begin{aligned} P(\sum d_i = y) = \quad & \left(\tfrac{KR}{\pi}\right)^2 P(\sum d_i = y|r_m = 1, r_f = 1) \\ & +2\tfrac{KR}{\pi}(1 - \tfrac{KR}{\pi})P(\sum d_i = y|r_m = 1, r_f = 0) \\ & +(1 - \tfrac{KR}{\pi})^2 P(\sum d_i = y|r_m = 0, r_f = 0). \end{aligned} \quad (5.29)$$

Results

I have implemented the above combined model using R, and used it to explore how the expected number of affecteds in multiplex families for a relatively uncommon disease ($K = 0.01$) varies depending on model and model parameters.

Figures 5.1a and 5.1b show the results of this multiplex model to families of 8 ($O = 6$), with dominant penetrance of $\pi = 0.5$. The solid lines give the purely polygenic model $R = 0$, the black lines give the purely penetrant model $h^2 = 0$, and other lines give various parameterisations of the combined model.

The first thing to note is that multiplex nuclear families can be very common given only a moderate degree of polygenic risk. Families with 5 or

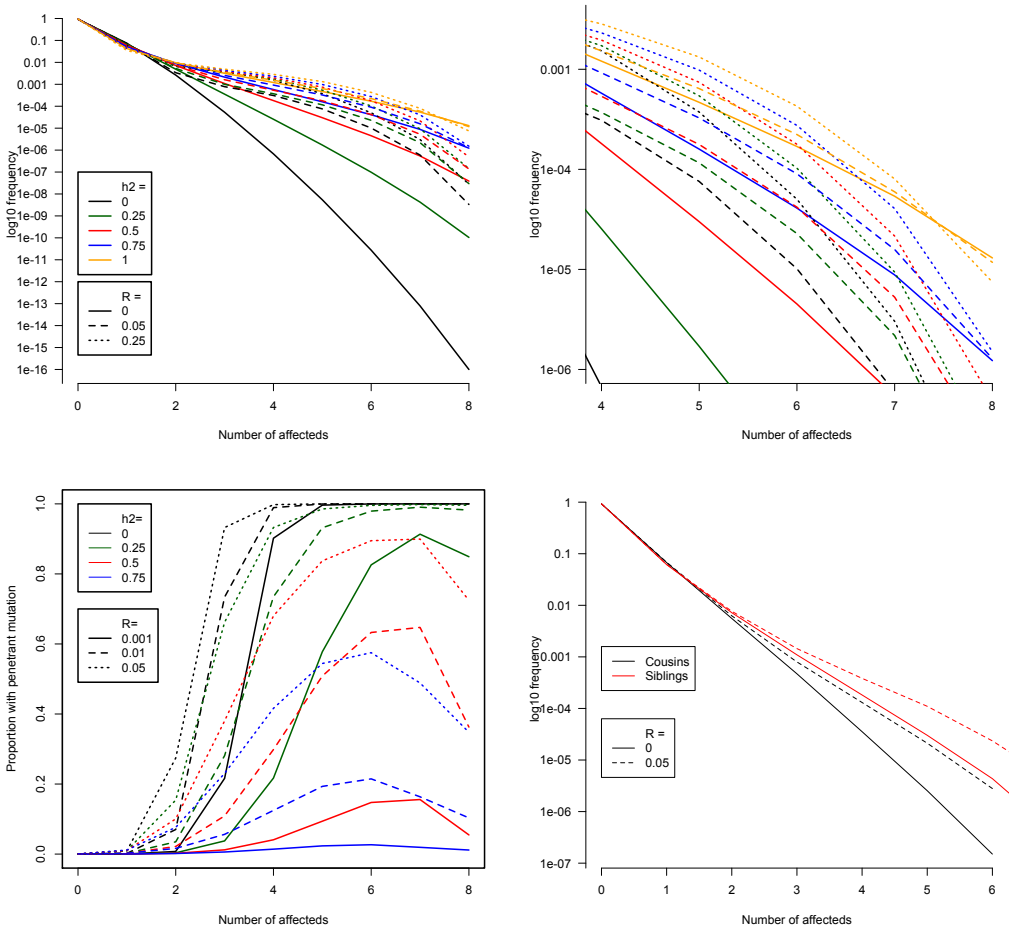


Figure 5.1: The results of the combined multiplex family model. a) Distribution of number of affecteds per nuclear family with 6 children, for different values of h^2 and R. b) A zoomed in view of the same model c) The probability that a nuclear family harbours a penetrant mutation, for different values of h^2 and R d) Comparison of sibships ($O = 6$) and cousinships ($k = 2, O_1 = 2, O_2 = 2$), with $h^2 = 0.5$ (generated by simulation). In both cases I used $K = 0.01$ and $\pi = 0.5$.

more affecteds, an occurrence that is virtually impossible under the null (less than one family in 200 million) become possible (one family in 500 thousand) for a moderate polygenic heritability of 0.25 and positively common (one family in 30 thousand) for strong heritability of 0.5. Multiplex families are likely to be relatively common, even without high penetrance mutations.

However, the flip-side of this is that a high degree of familial aggregation can be seen even for not particularly heritable diseases given a small contribution of dominant alleles. A disease with no polygenic liability, but 5% of cases caused by penetrant mutations, will show as many families with 4 affecteds as a disease with 50% heritability (despite the former case having a heritability of less than 1%). This seems to lead to the somewhat counterintuitive conclusion that families multiply affected by a weakly heritable disease will be easier to map than equivalent families with a strongly heritable disease, though this may be confounded by correlations between polygenic and penetrant heritabilities.

We can turn these results around and instead ask what proportion of multiplex families of a certain size harbour a penetrant mutation (Figure 5.1c). In the absence of polygenic risk, the vast majority of nuclear families with more than 4 affecteds harbour a penetrant mutation, even if such mutations explain a very small proportion of the total disease burden ($R > 0.001$). However, this becomes progressively less true as the heritability rises, and for highly heritable diseases penetrant mutations only become common in multiplex nuclear families if they already explain a non-trivial amount of all cases to start with (greater than 1% for $h^2 = 0.5$, and greater than 5% for $h^2 = 0.75$).

Figure 5.1d compares the results of the combined model for nuclear and extended families of the same size. Specifically, I have compared a nuclear

family with eight individuals (two parents and six children, i.e. $O = 6$ offspring) to an extended family with eight individuals (two siblings, their partners, and their two children each, i.e. $k = 2$,$O_1 = O_2 = 2$, with grandparental state disregarded). I consider only a heritability of $h^2 = 0.5$.

As we have already seen, under the polygenic model with $h^2 = 0.5$, observing five or more affected nuclear family members is not unlikely (1 in 30 thousand). However, Figure 5.1d shows that for an extended family of the same size this is a relatively rare even (1 in 400 thousand). This gap between nuclear and extended families is reduced if the presence of high penetrance mutations is considered. Introducing a small number of penetrant mutations ($R = 0.05$, $\pi = 0.5$) increases the number of families with at least 5 affecteds 9-fold for the cousinship (to 1 in 42 thousand), but only 4-fold for the nuclear family (to 1 in 7200). This corresponds to a 93% of cousinships with 5 affecteds carrying a penetrant mutation, compared to 83% for nuclear families.

From these analyses we can draw a number of lessons for studying multiplex families

- Even if only a minority of variance is explained by rare variants, these rare variants can still result in the occurrence of a relatively large number of multiplex families.

- However, relatively large numbers of multiplex families are also expected given the levels of polygenic risk ($h^2 = 0.2 - 0.5$) that have been shown to exist for many complex diseases. Thus the presence of strong familial clustering is not alone evidence of a penetrant mutation.

- Extended multiplex families, with aggregation occurring across cousins as well as siblings, is stronger evidence of a penetrant mutation.

- Using additional methods to decrease or factor out the contribution of polygenic risk will be valuable in identifying families that are likely to harbour penetrant risk variants

5.3.2 Risk prediction in multiplex families

An outline of risk prediction in families

As we have seen in the above section, the presence of polygenic risk can lead to a high frequency of multiplex families even in the absence of penetrant mutations. However, for many diseases we already have a grasp on this polygenic variation via the results of GWAS. For instance, the 193 independent associations to Crohn's Disease explain 12.7% of variance in disease liability (see Chapter 4). Using the upper bound of 84% (calculated in Chapter 1) and a lower bound of 22% (from Williams et al. (2002)), we know that we have discovered somewhere between 15% and 58% of the polygenic risk for Crohn's disease.

We can use these GWAS loci to produce estimates of polygenic risk, and use this polygenic risk to prioritise those families that are more likely to harbour penetrant mutations. Assume that a given family has N members, of whom y are affected. We wish to select families for which y is significantly larger than what would be expected given the observed genotypes, G, i.e. those that minimize:

$$P(\hat{y} > y|N, G) \tag{5.30}$$

If G is known for all family members then disease probabilities for each individual can be calculated directly from the odds ratios as described in Chapter 2, and then used to calculate equation (5.30) by sampling. How-

ever, most family based experiments will not generate genotype data across all members of the pedigree for a variety of reasons, including cost, DNA availability, consent, or death. A solution is to sample disease status as in the complete information case, conditional on a set of unobserved genotypes G_{unobs} that are themselves sampled from the conditional distribution

$$P(G_{unobs}|\mathbf{f}, T, G_{obs}), \tag{5.31}$$

where $\mathbf{f}$ is the population allele frequency, T is the family structure, and G_{obs} are the known genotypes. Sampling from this distribution is not trivial, but is possible via a modified Inside-Outside algorithm (Baker, 1979) (itself a generalisation of the forward-backwards algorithm used in Hidden Markov Models). The Inside-Outside is used for inference on tree-like data structures, and has been applied to certain multiple sequence alignment problems (Durbin, 1998). Here, we instead use Inside-Outside to sample from the posterior distribution of genotypes across a family. Briefly, we decompose the marginal genotype posteriors into inside and outside probabilities, similar to the forward and backward probabilities from an HMM. The inside probability accounts for information from each individual and their descendants, whereas the outside probability accounts for the individual's other relatives (including ancestors, siblings and cousins).

These values can be computed recursively via the standard Inside-Outside approach (Section 5.3.2), which enables the sampling of one individual's genotypes. When sampling an entire family, however, we must sample down the tree from the root, with each individual's genotypes conditioned on their parents' sampled genotypes (Section 5.3.2). We accomplish this by modifying the outside probability to include parental genotypes (Section 5.3.2).

Description of the Inside-Outside algorithm in trees

Definitions

The Inside-Outside algorithm is a generalisation of the Forward Backward algorithm, originally designed to extend parameter estimation from Hidden Markov Models to stochastic context-free grammars (Baker, 1979). Here we reformulate the Inside Outside algorithm as a method of performing parameter estimation and sampling on a directed tree.

A directed tree is a directed acyclic graph in which all nodes have a unique path originating from a single node. We will denote nodes by subscripts i, j, k. Each node i may have a parent p_i, offspring o_i and/or siblings s_i. A node without parents is called a "root node" or "root", and a node without children is called a "leaf node" or "leaf".

Each node i has an associated emission d_i (e.g, an observed genotype), as well as a hidden state x_i (e.g. an unobserved genotype) with statespace S_i. The values of hidden states will be denoted a, b, c etc, e.g. $(x_i = a)$ denotes that node i has hidden state value a.

The tree defines a graphical model that specifies the probability density functions for all the variables (hidden states and emissions) as conditional probabilities. Specifically, the probability density function of emission d_i is specified conditional on hidden state x_i taking value a by the likelihood

$$L_i(a) = P(d_i|x_i = a). \tag{5.32}$$

The probability density function for a non-root hidden state variable x_i taking on value b is specified conditional on the parent's hidden state x_{p_i} taking on value a by the transition probability

$$T_i(b|a) = P(x_i = b|x_{p_i} = a). \tag{5.33}$$

The probability distribution of the hidden state associated with the root x_{root} is given by the root prior

$$\pi(a) = P(x_{root} = a). \tag{5.34}$$

We will refer to all emissions associated with node i and nodes descended from node i as D_i, and all emissions not associated with node i or its descendants as $D_{!i}$. Note that these can both be expressed recursively

$$D_i = \{d_i, D_{o_i}\} \tag{5.35}$$

for non-leaves and $D_i = d_i$ for leaves, and

$$D_{!i} = \{D_{s_i}, D_{!p_i}, d_{p_i}\} \tag{5.36}$$

for non-roots and $D_{!i} = \emptyset$ for the root. All emissions associated with all nodes can be expressed as D, and $D = \{D_i, D_{!i}\}$ for any i.

We will use the Inside-Outside algorithm to deduce the probability density functions of hidden states x_i conditional on observed emissions associated with all nodes D.

The Inside Probability

The inside probability $\alpha_i(a)$ is defined as the probability of observing emission associated with node i and all its descendants, given that the hidden state x_i takes on value a

$$\alpha_i(a) = P(D_i|x_i = a). \tag{5.37}$$

For leaves, $D_i = d_i$, and hence $\alpha_i(a) = L_i(a)$. For non-leaves we have

$$\begin{aligned}
\alpha_i(a) &= P(D_i|x_i = a) \\
&= P(d_i|x_i = a) \prod_{j \in o_i} P(D_j|x_i = a) \\
&= P(d_i|x_i = a) \prod_{j \in o_i} \sum_{b \in S_j} P(D_j|x_i = b) P(x_j = b|x_i = a) \\
&= L_i(a) \prod_{j \in o_i} \sum_{b \in S_j} \alpha_j(b) T_j(b|a).
\end{aligned} \tag{5.38}$$

Because we require the inside probabilities of all offspring of a node to calculate its own inside probability we calculate the inside probabilities first for the leaves, and then propagate them recursively up the tree. The overall likelihood of all emissions D is

$$P(D) = \sum_{a \in S_{root}} \alpha_{root}(a)\pi(a). \tag{5.39}$$

The Outside Probability

The outside probability $\beta_i(a)$ is defined as the joint probability of observing emissions not associated with node i and its descendants, and the node i being in hidden state $x_i = a$ is

$$\beta_i(a) = P(D_{\setminus i}, x_i = a). \tag{5.40}$$

For the root node, $D_{\setminus i} = \emptyset$, so $\beta_{root}(a) = P(x_{root} = a) = \pi(a)$. For non-root nodes, we can calculate the outside probability recursively as

$$
\begin{aligned}
\beta_i(a) &= P(D_{\setminus i}, x_i = a) \\
&= \sum_{c \in S_{p_i}} P(x_{p_i} = c, x_i = a, D_{\setminus i}) \\
&= \sum_{c \in S_{p_i}} P(x_{p_i} = c, x_i = a, D_{\setminus p_i}) P(d_i | x_{p_i} = c) \prod_{j \in s_i} P(D_j | x_{p_i} = c) \\
&= \sum_{c \in S_{p_i}} P(x_{p_i} = c, D_{\setminus p_i}) P(x_i = a | x_{p_i} = c) P(d_i | x_{p_i} = c) \\
&\quad \times \prod_{j \in s_i} \sum_{b \in S_j} P(D_j | x_j = b) P(x_j = b | x_{p_i} = c) \\
&= \sum_{c \in S_{p_i}} \beta_{p_i}(c) T_i(a|c) L_{p_i}(c) \prod_{j \in s_i} \sum_{b \in S_j} \alpha_j(b) T_j(b|c). \qquad (5.41)
\end{aligned}
$$

The outside probability for each node requires the outside probability of the node's parent. We thus calculate it first for the root, and then propagate recursively down the tree. The outside probabilities are also dependent on the inside probabilities, which are therefore calculated first.

Conditional sampling across the tree

We can calculate the posterior distribution of hidden state x_i conditional on all emissions D in terms of the inside and outside probabilities as

$$
P(x_i = a | D) = \frac{\alpha_i(a)\beta_i(a)}{P(D)}. \qquad (5.42)
$$

We can sample from this posterior distribution for each node. However, this approach cannot jointly sample hidden states across the entire tree. To do this we need to propagate sampled states down the tree, starting with the root. The hidden state for the root can be sampled from the posterior distribution

$$P(x_{root} = a|D) = \frac{\alpha_{root}(a)\pi(a)}{P(D)}. \tag{5.43}$$

To sample non-roots, we must first calculate the partial outside variable, which includes the hidden state c of the parent, and can be calculated as

$$\begin{aligned} \beta_i^p(a,c) &= P(x_i = a, x_{p_i} = c, D_{\setminus i}) \\ &= \beta_{p_i}(c)T_i(a|c)L_{p_i}(c)\prod_{j\in s_i}\sum_{b\in S_j}\alpha_j(b)T_{jp_i}(b|c). \end{aligned} \tag{5.44}$$

The hidden state of node i can then be sampled from the posterior conditional on the sampled state of the parent c

$$P(x_i = a|D, x_{p_i} = c) = \frac{\beta_i^p(a,c)\alpha_i(a)}{\sum_{a\in S_i}\beta_i^p(a,c)\alpha_i(a)}. \tag{5.45}$$

Like the calculation of the outside probabilities, the samples are propagated down the tree.

Application of the Inside-Outside algorithm to family trees

A family is not strictly a directed tree, due to the addition of new founders (via marriage) in each generation. However, we can make a family into a directed tree by treating parent couples as a single node, consisting of a founder and a non-founder individual. The root node of this directed family tree consists of the top pair of founders. While I have currently only used this method for family trees with only one founder-founder couple, in fact any family relationships that do not include inbreeding (i.e. any that take the form of a polytree) can be modelled if the polytree is transformed to a directed tree by reversing the transition matrix (using $T_{p_i}(a|b) = T_i(b|a)P(x_{p_i})/P(x_i)$).

We use the Inside-Outside algorithm to sample unobserved genotypes conditional on all other genotypes for a single biallelic polymorphism with allele frequency f (although this is readily generalised to an arbitrary number of independent polymorphisms). We model individuals as nodes, and genotypes as hidden states for each node. For non-parent couples the state-space is

$$x_i \in S_i = \{AA, AB, BB\}, \tag{5.46}$$

and for parent couples it is

$$x_i = (x_i^f, x_i^{nf}) \in \{AA, AB, BB\}^2, \tag{5.47}$$

where x_i^f is the founder's genotype state and x_i^{nf} is the non-founder's genotype.

Genotype calls for each individual are modelled as emissions, and we assume that these genotypes are certain and thus for genotyped individuals x_i and d_i are identical (though genotype error can be included by modifying the likelihoods below). Genotypes can also be missing (N). Thus the emissions for a non-parent couple node is

$$d_i = g_i, \tag{5.48}$$

and for parent couples is

$$d_i = \{g_i^f, g_i^{nf}\}. \tag{5.49}$$

Likelihoods for non-parent couples are

$$L_i(a) = \begin{cases} 1 & \text{if } a = g_i \text{ or } g_i = N; \\ 0 & \text{otherwise.} \end{cases}, \tag{5.50}$$

and for parent couples are

$$L_i(a) = \begin{cases} 1 & \text{if } a_i^f = g_i^f \text{ and } a_i^{nf} = g_i^{nf}; \\ 1 & \text{if } a_i^f = g_i^f \text{ and } g_i^{nf} = N \text{ or } a_i^{nf} = g_i^{nf} \text{ and } g_i^f = N; \\ 1 & \text{if } g_i^f = g_i^{nf} = N; \\ 0 & \text{otherwise.} \end{cases}. \tag{5.51}$$

Transitions can only occur from a parent couple to a non-parent couple, or from a parent couple to a parent couple. For a parent couple to a non-parent couple, transmission is simple Mendelian inheritance

$$T_{ij}(a|b) = P(C = a|P1 = b^f, P2 = b^{nf}), \tag{5.52}$$

where C is the child's genotype, and $P1$ and $P2$ are parental genotypes. For parent couple to parent couple transmission, we need to include the probability density on the founder genotype

$$T_i(a|b) = P(C = a^{nf}|P1 = b^f, P2 = b^{nf})P(a^f|f), \tag{5.53}$$

where $P(a^f|f)$ is the population frequency of the founder's genotype, assuming Hardy-Weinberg equilibrium. Finally, the prior on the root node is given by the population frequency

$$\pi(a) = P(a^f|f)P(a^{nf}|f). \tag{5.54}$$

Using this formulation, marginal posteriors can be calculated for each unob-

served genotype, and joint genotypes for the entire family can be sampled from the joint posterior distribution.

Mangrove: An R package for risk prediction in families

To summarise the above approach, we can calculate the probability of seeing at least y affected families members in a family given known GWAS risk loci $P(y|G_{obs}, \beta, f)$ using the following process:

1. Convert the family tree with genotype data into a true directed tree with emissions as described in section 5.3.2

2. Calculate α_i, β_i and β_i^p statistics using the Inside-Outside algorithm as described in section 5.3.2

3. Sample N sets of genotypes for ungenotyped family members using the method in 5.3.2

4. Sample affection status for each individual conditional on samples genotyped, using standard risk prediction (Chapter 2)

5. Count the number of families with more than y affected family members

These stages have all been implemented in the R package Mangrove, which is available from the Comprehensive R Archive Network (CRAN). Mangrove is specifically designed to use genetic risk prediction to prioritise individuals or families for sequencing. As well as risk prediction in families, Mangrove can also perform both risk prediction and quantitative trait prediction in unrelated individuals. I have provided detailed documentation, and a vignette containing usage examples for both families and unrelated individuals, with the package.

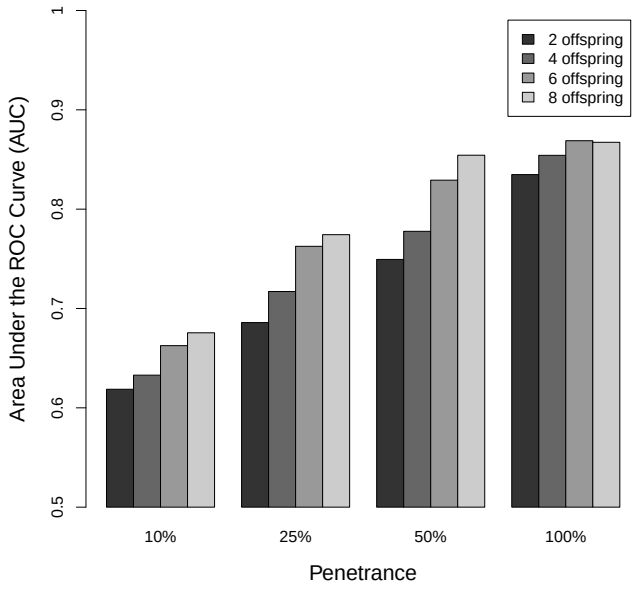


Figure 5.2: Ability to predict the presence of a high penetrance mutation (measured by AUC) in multiplex families using a polygenic risk score. We assume a disease with a prevalence of 1%, a heritability of 50%, and a genetic risk score that captures 12.5% of variance. All families have three affected individuals, and the AUC is shown for families of different total size and dominant mutations of varying penetrance.

Assessing the efficacy of risk prediction in families in prioritising penetrant mutations

The aim of the risk prediction prioritisation described above is to increase the chance that a family selected for sequencing carries a high penetrance mutation. To investigate how powerful this approach is I performed simulations of families with and without a high penetrance mutation.

Consider two families both subject to polygenic risk for a disease and one additionally containing a high penetrance dominant mutation. We would like to be able to identify the latter family for the type of family sequencing experiment described above. To evaluate the ability of the above method

to identify families containing such high penetrance mutations I simulated nuclear families with between 2 and 8 offspring, where three total family members were affected by a disease having 1% prevalence and heritability of 50% (these values correspond approximately to immune mediated diseases such as Crohn's disease). Half the families contained a dominant mutation with a penetrance from 10–100%, and the other half arose simply from polygenic risk and chance.

For each family, we computed the value of equation (5.30) based on a GWAS risk predictor explaining 25% of heritability (again by analogy to Crohn's disease). Figure 5.2 shows the area under the ROC curve (AUC), which in this instance can be interpreted as the probability of correctly distinguishing between one family with a penetrant mutation and one without. For a low-penetrance mutation in a small family AUC is only $\sim$0.6, but for a medium-penetrance mutation in a large family, AUC is $\sim$0.85, which would provide a substantial advantage over simply selecting the family with the largest number of affected individuals.

5.4 Linkage and sequence analysis of a multiplex IBD family

We have seen how multiplex families are likely to show an enrichment for rare, high penetrance risk variants. This is particularly true for multiplex families that span extended pedigrees, and in pedigrees with a low predicted risk given common variants. Via linkage and haplotyping methods, these families can also be analysed for candidate regions that may harbour such mutations. The falling cost of sequencing means that whole-exome or whole-genome sequencing can then be used to attempt to identify causal candidates in the family using linkage data and functional information.

To attempt to discovery such high penetrance mutations, we collected samples from extended families with multiple members affected by inflammatory bowel disease (IBD). Here I discuss the analysis of one such family.

Note that some non-important details of the family have been altered in this chapter to ensure anonymity. These include the gender of subjects, the number of offspring and the details of family relationships. In no case does this affect the conclusions drawn, though it may lead to small inconsistencies in the precise details of results.

5.4.1 Description of the family

The family comprises over 800 individuals of Ashkenazi Jewish descent, spanning four generations connected via a founding couple born at the turn of the 20th century (Figure 5.3). The family is characterised by its large number of offspring per parental couple, with an average of 9. The founding couple had seven offspring (including two identical twins), six of these have at least two descendants with IBD.

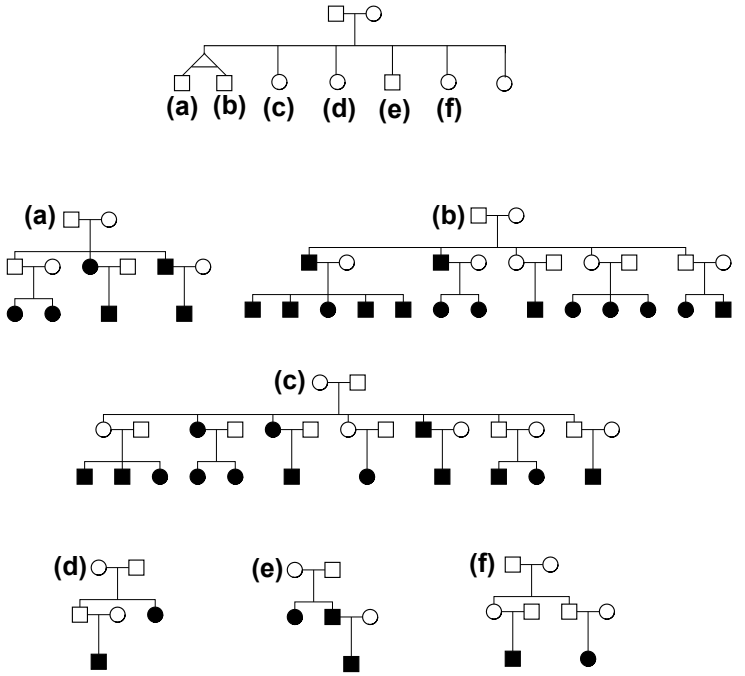


Figure 5.3: A pedigree for the family under study, showing affecteds and parents of affecteds. The top figure shows how the founders of the six subpedigrees (a-f) are related. The founders of subpedigrees a) and b) are identical twins.

A total of 41 individuals have been diagnosed with IBD, including 35 with a diagnosis of Crohn's disease and 7 with a diagnosis of ulcerative colitis. We were able to independently confirm the diagnosis via medical records in all but five cases. The location of disease in the bowel was variable. The average age of onset was 18.8 years (95% CI: 16-22, n=30) and at the time of sample collection, one-quarter of the patients had undergone surgical resections.

This family is a good candidate for discovering a high penetrance mutation. They have a wide geographic distribution, with affected individuals present in seven cities around the world, making an environmental cause of the disease less likely. Additionally, because the affecteds are spread across first and second cousins, polygenic risk is far less likely to explain the large number of affecteds.

5.4.2 Segregation analysis

Before looking at any genetic data, we can use the structure of the family to make a plausible guess at what sort of genetic risk factors we may be looking for. We will look specifically at subfamilies (a) and (b) as the identical twin founders make the analysis significantly easier.

Suppose we take the most optimistic view of the genetics of this family, i.e. that all cases are explained by a single dominant mutation. Together, the two identical twins have 18 offspring, of which 10 are either affected, or have affected children (or both). The most favourable model would be to suppose that these 10 individuals all inherited a causal mutation from these identical twins, and the rest did not. Furthermore, we will assume that all affected family members carry this mutation.

Under this favourable model, 9 parents and 18 affected children, as well as approximately half of their 66 unaffected siblings, will carry the mutation, of which 21 have the disease. This gives a penetrance of 35% (21 out of 60). In fact, as discussed in section 5.5.2, unaffected siblings are less likely to inherit a causal mutation. If we correct for this, the estimated penetrance in the highly favourable model is 41%, with a 95% confidence interval of 24-48%.

This model is almost certainly overly optimistic, as in a family of this size many of the cases are likely to be phenocopies, and likewise causal mutations may be segregating in parts of the family with no affecteds. It is also possible that the mutation is recessive, interacts with another risk factor (either genetic or environmental), or is only one of many undiscovered risk factors in the family. However, the model does illustrate how, even in the best-case scenario, we are looking for a mutation with incomplete penetrance (<50%).

Family	N	y	$E(y\|K)$	$E(y\|G)$	$\frac{y}{E(y\|G)}$	$P(y\|G)$
Whole family	806	41	6.04 (1 - 11)	10.24	4.00	$< 10^{-4}$
Subfamily (a)	112	6	0.84 (0-3)	1.02 (0-4)	5.90	0.0012
Subfamily (b)	112	15	0.84 (0 - 3)	0.97 (0-4)	15.42	$<10^{-4}$
Subfamily (c)	140	14	1.05 (0 - 3)	1.56 (0 - 5)	8.97	$<10^{-4}$
Subfamily (d)	147	2	1.10 (0-3)	1.24 (0-4)	1.62	0.352
Subfamily (e)	81	3	0.61 (0 - 2)	1.63 (0 - 5)	1.84	0.243
Subfamily (f)	138	2	1.04 (0 - 3)	3.11 (0 - 7)	0.74	0.706

Table 5.1: A Mangrove analysis of the IBD family, including analyses of the six subfamilies. N is the total number of individuals in this subfamily, y is the number of affected individuals, $E(y|K)$ is the expected number of affected given the prevalence alone, $E(y|G)$ is the expected number given genotyped common variants. $\frac{y}{E(y|G)}$ is the enrichment of cases over that predicted by common variants, and $P(y|G)$ is the probability of observing y or more affected in this pedigree given common variation. Numbers in brackets are 95% confidence intervals.

5.4.3 Known IBD risk variants in the family

We successfully genotyped 38 CD and UC risk variants in 152 family members across the entire family in order to assess the extent to which the increased incidence may be explained by known genetic risk factors. I used odds ratios and frequencies taken from the IIBDGC GWAS meta-analysis data (using only Jewish samples), except for the 3 *NOD2* variants for which I used the Immunochip data (described in Chapter 4). Together, these variants explain 7.8% of variance in CD liability and 2.0% in UC liability.

I used the R package Mangrove (described in Section 5.3.2) to assess the number of cases we would expect in the family given these common variants. I used population prevalence of CD and UC of 0.6% and 0.15%, collected by Adam Levine from Jewish patients in GP surgeries in North London (personal communication).

Compared to the baseline prevalence, the family shows a 6.8-fold enrichment in IBD. While the family does show a marked increase in risk (1.7-fold)

SNP	OR_{het}	OR_{hom}	P-value
rs2066844	1.83	8.65	1.6 x 10^{-12}
rs2066845	1.90	11.61	2.1 x 10^{-4}
rs2066847	2.56	29.8	1.8 x 10^{-16}
Compound heterozygous (Excess odds ratio over additivity)	x3.46	-	2.0 x 10^{-55}

Table 5.2: Odds ratios for *NOD2* mutations under a non-additive model, fitted from the IIBDGC Immunochip data described in Chapter 4. The p-values give the significance of the full model compared to a model with this term replicated with a purely additive term.

due to common risk variants, there is still a 4-fold enrichment in IBD even given these common variants (Table 5.1).

We can further break this down by subfamily (Table 5.1). Subfamilies (d)-(f) show a particularly marked enrichment in common risk variants, which would predict a 2.2-fold increase in prevalence. The expected number of affected given common risk variants (5.98) is remarkable close to the observed number (7), suggesting that there is unlikely to be any high penetrance mutations in this area of the family. By contrast, subfamilies (a)-(c) show a very large gap between the predicted and actual number of affecteds (9.9 times that predicted by common variants), suggesting that these subfamilies are good candidates for harbouring high penetrance mutations.

Modelling non-additivity in NOD2 risk variants

One complication is that the above analysis assumes an additivity genetic architecture. While this model fits most of the IBD risk variants well, it does not accurately model the *NOD2* risk variants, which show significant evidence of both recessive effects at single coding variants and epistatic interaction between coding variants (Table 5.2).

In subfamilies (a) and (b) *NOD2* mutations are relatively uncommon,

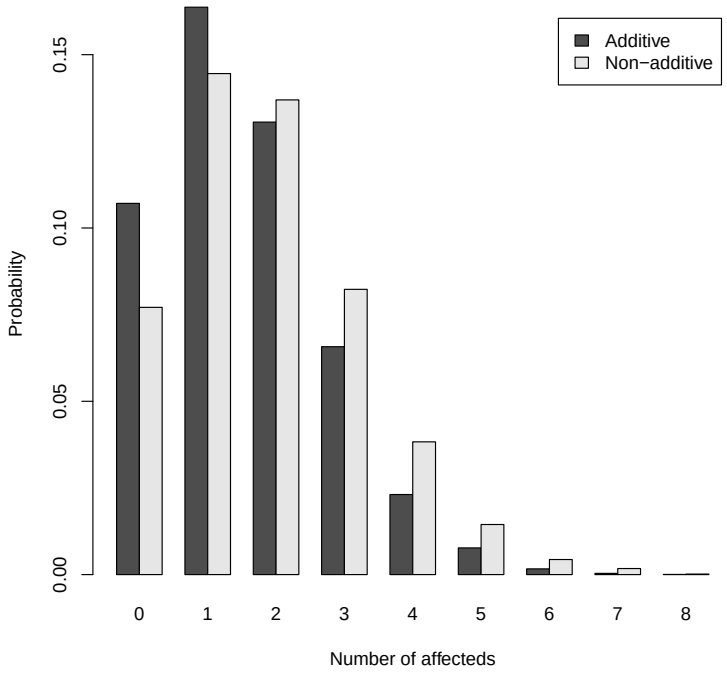


Figure 5.4: The distributed of cases expected in subfamily (c) under an additive and a non-additive model of NOD2 risk.

and no individuals were homozygous or compound heterozygous for *NOD2*, suggesting that the non-additive model will only decrease the total number of predicted affecteds. However, in subfamily (c) seven individuals are either homozygous or compound heterozygous for one of the three classical *NOD2* mutations, suggesting that the contribution of known genetics in this family could be larger than an additive analysis suggests.

I used data from the IIBDGC Immunochip dataset (described in Chapter 4) to fit a non-additive *NOD2* model by logistic regression (Table 5.2), and used the Mangrove method to perform risk prediction in subfamily (c) using this model. Non-additivity increases the expected number of affecteds slightly, from 1.56 to 1.88 ($p = 5.5 \times 10^{-11}$). However, the real increase is on the extremes (Figure 5.4), where the probability of seeing 6 or more affecteds increases by a factor of three (from 0.4% to 1.3%). Despite this increase, the

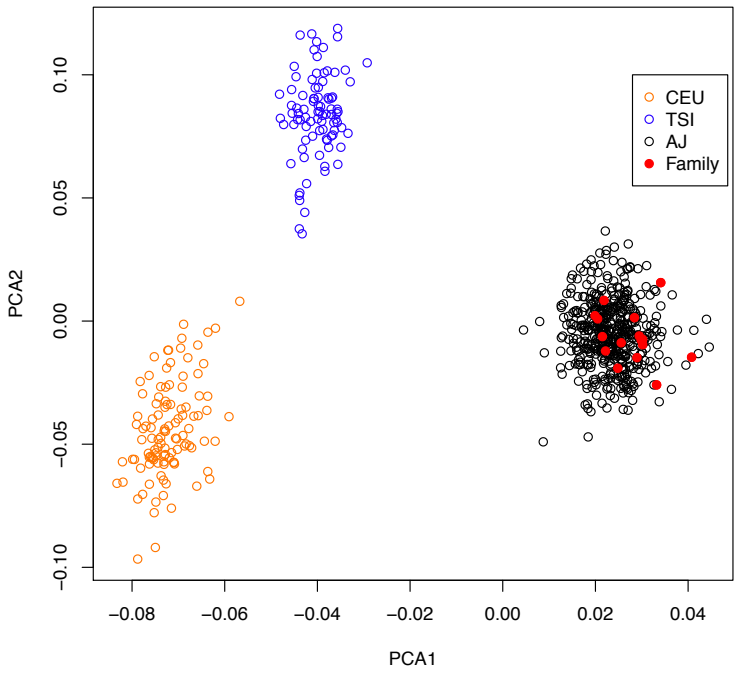


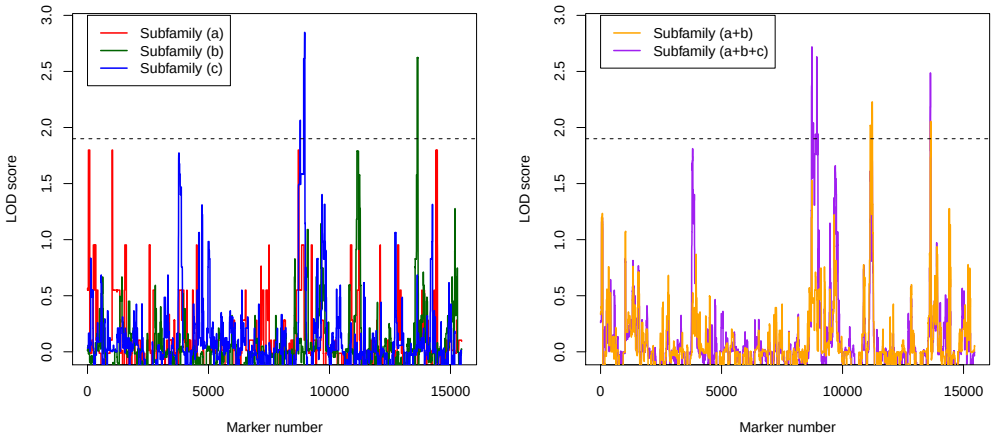
Figure 5.5: A principal component analysis of the family, using HapMap populations (TSI=Italian, CEU=Northern European) and Ashkenazi Jewish (AJ) reference populations.

probability of seeing 14 affecteds in subfamily (c) given common variation remains very small ($<< 10^{-4}$).

5.4.4 Linkage and haplotype analysis of the family

Genotyping data

A total of 60 individuals (30 affected and 30 unaffected) from subfamilies (a)-(c) were genotyped on an Illumina CytoSNP 12 BeadChip array. Genotypes were called using BeadStudio. Genotypes inconsistent with Mendelian segregation were set to missing, and SNPs with greater than 1% missingness, minor allele frequency less than 1% in founders or Hardy-Weinberg Equilibrium p-value less than 10^{-5} in founders were removed.



(a) Within subfamilies **(b)** Across subfamilies

Figure 5.6: Non-parametric linkage results for the family.

As a reference population, we used genetic data from a study of 471 Ashkenazi Jewish individuals genotyped on the Affymetrix Human SNP Array 6.0 (Bray et al., 2010), obtained via the NCBI's Gene Expression Omnibus (GEO) database (Barrett et al., 2011). Principal component analysis confirmed that the family members clustered with the Ashkenazi reference population (Figure 5.5).

We created a 1cM maximally informative genetic map by taking all SNPs present in both the reference set and the family, and for which there was no missing data in the family. We performed LD thinning in the reference dataset (such that $r^2 < 0.2$ for all SNPs). We then selected the SNP with the highest heterozygosity in the family founders in every 1cM block. Allele frequencies for these SNPs were calculated from the reference set.

Linkage analysis

We performed non-parametric linkage using Merlin (Abecasis et al., 2002) (v1.1.2). As we expect large increases in allele sharing due to high pene-

trance mutations, the standard linear approximation used by Merlin is too conservative, so we used the more accurate Kong and Cox exponential model (Kong and Cox, 1997). We used the maximally informative map and allele frequencies described above.

We ran linkage separately on the three subfamilies (a)-(c). We also used Fisher's method to combine the results for subfamilies (a)-(b) (i.e. the offspring of the identical twins), and for all subfamilies (a)-(c). The results are shown in Figure 5.6. None of the results meet the criteria for genome-wide significance (a LOD score of 3.3 (Lander and Kruglyak, 1995)). A number of linkage peaks reached the level of significance that Lander and Kruglyak (1995) suggest can be interpreted as "suggestive evidence" (a LOD score of 1.9). These are shown in Table 5.3.

The linkage peaks inferred are broad, and contain many genes. Even if we reduce this down to genes that are expressed in the immune or digestive systems, there are still between 7 and 89 genes in each linkage peak (Table 5.3). Low-throughput sequencing of exons in some of these candidates did not produce any likely candidate causal variants.

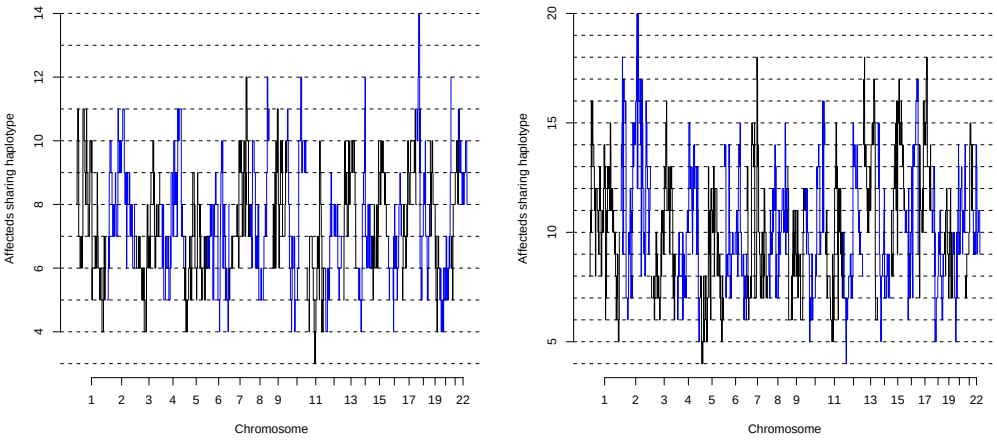
Haplotype analysis

As well as using the genotype data to find evidence of significant linkage, we can also use it for the related purpose of inferring the flow of haplotypes within the family. This can allow us to identify regions of the genome that are widely shared across subfamilies, and identify which family members do and do not share a candidate mutation on a particular haplotype. It can be used to inform the analysis of sequence data.

The computing resources required to carry out a full haplotype analysis grows exponentially with the number of samples. As a result, directly infer-

Chr	Pos in Mb	LOD score (subfamilies)	P-value	Genes (expressed)
Subfamily (b)				
18	6.98-9.71	2.62	2.54×10^{-4}	10 (7)
Subfamily (c)				
10	72.59-82.39	2.81	1.62×10^{-4}	81 (23)
Subfamily (a)+(b)				
13	89.61-96.75	2.23 (0.95, 1.58)	6.78×10^{-4}	24 (8)
18	6.98-9.71	2.05 (0.01, 2.62)	1.05×10^{-3}	10 (7)
Subfamily (a)-(c)				
10	19.17-81.96	2.72 (1.80, 0.12, 1.49)	2.01×10^{-4}	256 (89)
18	6.99-9.71	2.49 (0.01, 2.62, 0.69)	3.57×10^{-4}	10 (7)

Table 5.3: Suggestive linkage peaks (LOD > 1.9) in the family. Positions are given as the region in which markers have LOD > MAXLOD - 1. Numbers in brackets are LOD scores of the individual subfamilies that went into the analysis. The number of genes expressed in either the immune or digestive systems in the linkage peak is calculated from the expression datasets described in section 5.4.7



(a) Subfamilies (a)+(b) **(b)** Subfamilies (a)-(c)

Figure 5.7: Haplotype sharing in affecteds across the genome for subfamilies (a)+(b) (of 18 total) and subfamilies (a)-(c) (of 31 total).

ring haplotypes across subfamilies using Merlin was not possible. Instead, we developed a method for parallelising the calculating of haplotypes across subfamilies, involving the following steps:

1. Perform haplotype analysis in two subfamilies separately
2. For each pair of individuals across the two subfamilies, produce a small pedigree consisting of siblings of these individuals, and ancestors that connect them together. Use this to perform genome-wide identity-by-descent estimation in these two individuals.
3. For every possible set of haplotype assignments at every point in the genome, calculate the difference between the calculated identity-by-descent value and the value predicted by the haplotypes generated in step 1, summed across all pairs of individuals.
4. At each position in the genome, pick the haplotype assignment that minimises this value

We carried out this analysis on subfamilies (a)+(b) using this method, and on subfamilies (a)-(c) by then matching up haplotypes between subfamilies (a)+(b) and (c).

Haplotype sharing in subfamilies (a) and (b)

The maximum number of affected family members sharing the same haplotype across the genome for subfamilies (a) and (b) is shown in Figure 5.7a. The most widely shared haplotype is on chromosome 18 (corresponding to the suggestive linkage peak in Table 5.3), and is shared by 14 of the 18 genotyped affecteds. This haplotype is present in all five affected nuclear families in subfamily (b), and two of the four in subfamily (a).

Using the same approach as described in section 5.4.2, we can use this haplotype information to estimate the potential penetrance of a dominant mutation that lies on this haplotype. This model produces an estimate of the penetrance of 39% (95% CI 27-56%). It also implies between 4 and 7 phenocopies, corresponding to a phenocopy rate of 2.6% (95% CI 1.0-6.3%). While this is elevated compared to the population prevalence, this may be partly explained by ascertainment bias: this family, and in particular this subfamily, was selected for investigation due to the large number of affecteds, and this is likely to slightly inflate the number of affecteds due to winner's curse.

Haplotype sharing in subfamilies (a)-(c)

The maximum degree of haplotype sharing in subfamilies (a)-(c) is found on chromosome 2 (between 13.3Mb and 14.3Mb). This does not correspond to any of the suggestive peaks in the linkage analysis. This haplotype is shared across 10 of the 16 affected nuclear families, and affects 20 of the 31 genotyped affecteds in this part of the family.

A dominant causal mutation on this haplotype could have a relatively high penetrance (48%, 95% CI 36-64%). However, it would also imply between 11 and 14 phenocopies, corresponding to a phenocopy rate of 4.2% (95% CI: 2.4%-7.1%). This is more than 5-fold higher than the population prevalence, and 4-fold higher than the rate predicted from common risk variants in this part of the family, suggesting that a dominant mutation on this haplotype alone would be insufficient to explain the incidence of IBD in this family.

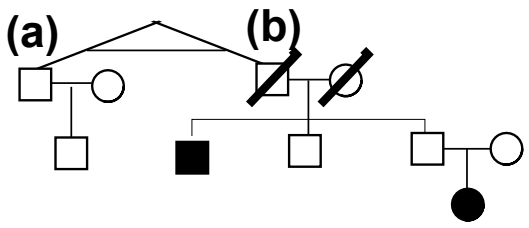


Figure 5.8: The founder subpedigree used for whole-genome sequencing.

5.4.5 Whole-genome sequencing in the family

Samples chosen for sequencing

Whole-genome sequencing allows a complete survey of variation within a family. It allows us to characterise structural variation, as well as SNPs and indels in non-coding DNA that may have a regulatory function. However, the cost is substantially higher, and thus we can only perform sequencing on a limited number of individuals. We decided to concentrate on subfamilies (a) and (b), as they are descended from two identical twins. This both increases the chance that a shared mutation is acting in both families, and reduces the cost of sequencing (because two founders can be sequenced for the price of one).

Figure 5.8 shows the 8 samples that we decided to sequence. These samples have been picked to capture the shared haplotypes introduced by the identical twins who founded subfamilies (a) and (b). Additionally, we included enough offspring to allow us to assign mutations to haplotypes, and thus allow us to impute variants on shared haplotypes into all affected members of subfamilies (a) and (b).

Generating and quality controlling raw sequence

We performed whole-genome sequencing using the Illumina HiSeq 2000, generating 2x100bp reads. A total of 407.5Gb of sequence was generated, and

Call set	SNPs	% dbSNP	Ts/Tv
Union	7.46M	79.5%	1.76
Intersection	6.09M	90.2%	2.07
VQSR (99%)	5.86M	92.2%	2.04
VQSR (90%)	5.16M	94.7%	2.12

Table 5.4: Summary statistics for various whole-genome sequencing call sets

aligned to build 37 of the human genome using BWA (Li and Durbin, 2009) v0.5.9. The mapping rate was 95.49% (range 94.07-96.35%), and the average coverage across the eight individuals was 16.1X (range 12.3 - 23.6X).

QC of the sequence data was performed using the BAMCheck pipeline developed by Petr Danecek, and all sequencing lanes passed. Samples were checked against their CytoSNP12 genotyping data (described above) to assure that samples swaps had not occurred. GATK (McKenna et al., 2010) v1.2 was used to perform local realignment around known indels, and to recalibrate base pair quality scores.

Calling SNPs and indels

Raw lists of SNPs and indels were generated using the GATK UnifiedGenotyper and samtools mpileup (Li et al., 2009) (v0.1.17). A total of 7.46M SNPs and 1.50M indels were called, of which 82% and 53% respectively were called by both approaches. This union SNP set is relatively poor: over 20% of SNPs are not seen in dbSNP, and the transition to transversion ratio (which should be above 2) is only 1.76 (Table 5.4). To improve the dataset, we carried out Variant Quality Score Recalibration (VQSR) using GATK. This technique fits a mixture model of true and false positive variants using QC metrics and a truth set of known polymorphic variants, and uses this to produce a calibrated quality score (the VQSLOD) for each variant.

Statistic	Call sets	Description
QD	SNP/Indel	Variant quality divided by depth
HaplotypeScore	SNP/Indel	Data consistency with exactly two haplotypes per individual
MQ	SNP	RMS mapping quality of reads mapping to site
MQRankSum	SNP	Test statistic for bias in MQ
DP	SNP	Total depth of reads at site
FS	SNP/Indel	Test statistic for bias in strand
ReadPosRankSum	SNP/Indel	Test statistic for bias in position in read

Table 5.5: QC statistics used for VQSR. In all cases "bias" refers to a difference in reference and non-reference reads. RMS stands for "root-mean-square", i.e. $\sqrt{\frac{1}{N}\sum_i^N x_i^2}$.

We used a variety of QC statistics as input for VQSR (Table 5.5). For SNP truth datasets, we used HapMap3 and 1000 Genomes Omni2.5 polymorphic sites, and for an indel truth dataset we used indels observed twice in the Mills and Devine (Mills et al., 2011a) dataset. A total of 5.86M SNPs and 1.22M indels passed the basic VQSR filter (VSQR99, equivalent to VQSLOD $>$ 2.52 for SNPs and $>$ 0.13 for indels), and these call sets had very favourable statistics (Table 5.4). A more stringent level of filtering (VQSR90, equivalent to VQSLOD $>$ 5.18 for SNPs and VQSLOD $>$ 3.20 for indels) provides a very high quality dataset at the expense of calling fewer variants.

We can use the CytoSNP 12 genotype data to test the sensitivity of the SNP call sets. Figure 5.9 shows this sensitivity as a function of non-reference allele count. As well as showing good quality statistics, the VQSR datasets have a very high sensitivity: the basic VQSR99 set has a 99.7% sensitivity for variants present in at least two individuals, and the stringent VQSR90 set, while less sensitive, still has a very high sensitivity (99.0%). A caveat to this analysis is that the CytoSNP 12 was designed late in the Illumina BeadChip

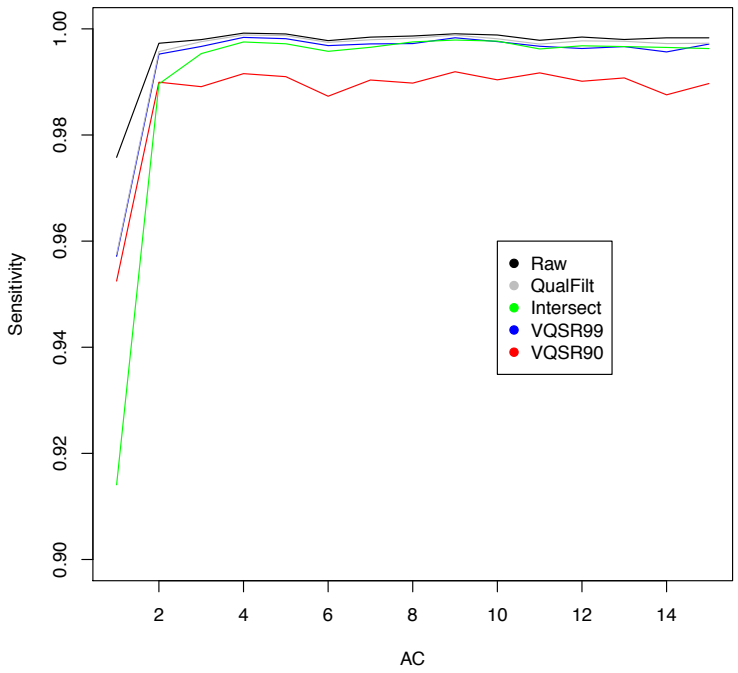


Figure 5.9: The sensitivity of the various WGS call-sets compared to array genotyping, as a function of non-reference allele count (AC).

line (in 2008) in order to genotype low concentrations of DNA, and as such is strongly biased towards "genotypeable" (i.e. complex, well-behaved) SNPs. The sensitivity values should thus be considered the sensitivity to detect "easy" SNPs.

Calling structural variants

Unlike for SNP and indel calling, there is no single well-established method for calling structural variants (SVs) from sequence data. Instead, most SV calling efforts combine information from a range of different complementary calling methods (Mills et al., 2011b).

To call SVs from the whole-genome sequencing data I used six different

Method	Insertions	Deletions	Inversions	Complex
BreakDancer	0	4630	517	0
CNVnator	2816	17371	0	0
Pindel	2573	2574	165433	0
RDXplorer	491	335	0	0
SECluster	1347	0	0	0
Genome STRiP	0	1377	0	0
SVMerge confirmed	814	3519	19184	8355

Table 5.6: Summary statistics for the different whole-genome sequencing structural variant callsets, along with the combined SVMerge set

calling methods to generate candidates. These included two methods that call SVs based on read-depth (RDXplorer (Yoon et al., 2009) and CNVnator (Abyzov et al., 2011)), two that call based on paired end reads (BreakDancer (Chen et al., 2009) and SECluster (Wong et al., 2010)), one that uses a combined read-length and paired-end method (Genome STRiP (Handsaker et al., 2011)) and one that calls based on split reads (Pindel (Wong et al., 2010)). We used the program SVMerge to combine these candidates together into a single set. We used the recommended SVMerge settings for filtering candidate sets, and removed calls that overlapped centromeres, teleomeres or gaps in the reference. The merged list of variants was then checked by local assembly (using the assembly program Velvet (Zerbino and Birney, 2008)) to confirm breakpoints. A breakdown of the number of variants called is shown in Table 5.6. Note that a very large number of inversions and complex events are called, coming almost exclusively from Pindel. As Pindel already uses local realignment, the 19,184 inversions could not actually be confirmed by an independent method, and should thus be considered suspect.

A total of 1210 SVs had at least a 50% reciprocal overlap with known structural variants (taken from Zhang et al. (2006), Conrad et al. (2010) and Mills et al. (2011b)). Of these, 179 of the 814 insertions had been previously

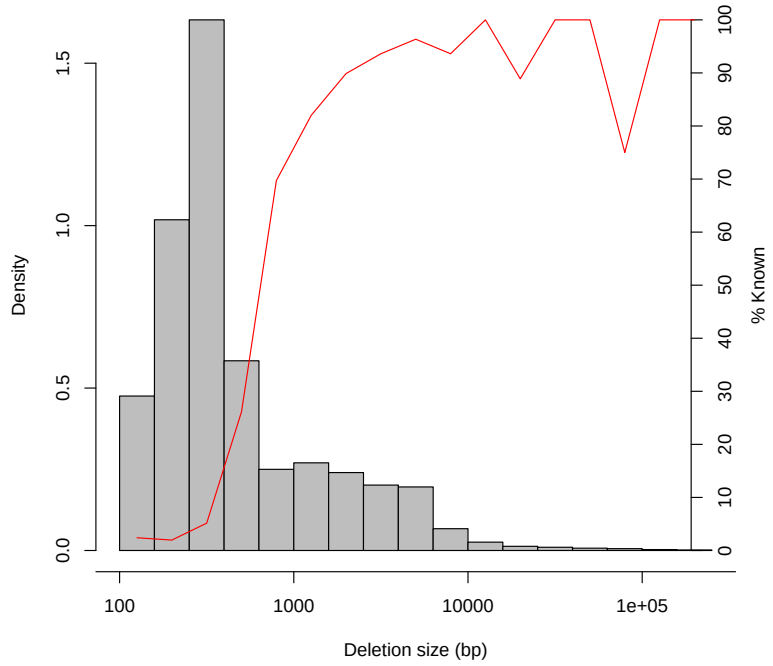


Figure 5.10: The distribution of deletion size in our call set, combined with the proportion observed (with >50% reciprocal overlap) in at least one of the three external datasets.

discovered, and 968 of the 3519 deletions. However, only 23 of the inversions and 37 of the complex events were previously known, suggesting again that these classes of variants are unreliably called. We decided that the likely very high false positive rate in inversions and complex events made them unreliable, and discarded them.

Looking in detail at the deletions, the number of called mutations also seen in the databases varies widely with the size of deletion (Figure 5.10) . 88.8% of deletions sized between 100 and 1000bp are novel, compared to only 9.6% of deletions greater than 1000bp. This likely represents a combination of false negatives in the database (for instance, Conrad et al. (2010) only

examined SVs larger than 443bp), and false positives in our call set.

5.4.6 Whole-exome sequencing in the family

Samples chosen for sequencing

Whole-exome sequencing is a more limited approach than whole-genome sequencing, and only allows the assessment of small-scale variation in coding regions. However, the substantially lower cost means that many more samples can be sequenced, potentially allowing a far more extensive study of coding variation than can be afforded by whole-genome sequencing.

All affected individuals from the family with DNA available (a total of 40) were sequenced, along with 13 unaffected family members to allow phasing. Additionally, we sequenced 26 control exomes, taken from unaffected members from the same ethnic group and geographic region as the family, to allow us to identify population-specific variation that may otherwise be mistaken for risk variants.

Processing of whole-exome sequencing data

We performed whole-exome sequencing, using a SureSelect Human All Exon 50 Mb kit for target enrichment and the Illumina HiSeq 2000 for sequencing. We used the same pipeline for quality control, mapping, realignment, recalibration and variant calling that was developed for the whole-genome sequencing (sections 5.4.5 and 5.4.5). The samples had a mean coverage of 154.0X in the target region (range 131.2X - 186.4X).

The VQSR99 set contained 128410 SNPs (87% known, Ts/Ts = 2.84), of which 105243 were also in the VQSR90 set (89% known, Ts/Tv = 2.96). The indel dataset contained too few indels to apply VQSR, so instead we

used the default GATK hard-filters to create a high-quality indel set. This included 9906 indels (52% known).

5.4.7 Identifying candidate variants in the family

Identifying candidate mutations

Between the whole-genome and whole-exome sequencing we have called over 7.5 million SNPs, indels and structural variants. Given the analyses reported above, we can be nearly certain that there exists, somewhere in this list, at least one mutation that causes a substantial increase in risk of inflammatory bowel disease. To identify such mutations, we need to filter out the vast majority of variants that do not contribute to IBD risk.

We have developed separate filtering procedures for the three different classes of variants: coding SNPs and indels, non-coding SNPs and indels and structural variants (laid out in detail in Table 5.7). Each filtering procedure begins with a platform-specific quality filter to remove poorly performing variants, followed by the removal of high-frequency variants using various databases of common variation.

Our next stage is to filter out any variants that are not present in at least half of the family members being considered. In the case of the data deriving from whole-genome sequencing we infer this from the haplotype flow information discussed in Section 5.4.4. For the SNPs and indels, we examine the consistency of the genotypes with what would be expected if the variant lay on the maximimally shared haplotype at that point in the genome. If the genotypes are consistent with this haplotype (given at most one genotyping error), and the haplotype is shared by at least half of affecteds in subfamilies (a)+(b), we include the variant. This approach has the notable

Filter	Description
Filters for coding variants:	
High quality	Genotype quality > 10 in at least 60% of samples
Uncommon coding variant	Frequency <2.5% in ESP[a], and less than <5% in our 26 AJ controls (annotated using ANNOVAR[b])
Affected sharing	Is shared by at least 50% of sequenced affecteds in either subfamilies (a)+(b), subfamily (c), or the entire family
Coding consequence	Is a missense, nonsense, essential splice, stop or frameshift mutation (annotated using Ensembl VEP[c])
Deleteriousness	Predicted to be deleterious to protein function (measured using Condel[d]).
Filters for non-coding variants:	
Haplotype consistency	Genotypes are consistent with maximally shared haplotype in linkage data (given at most one genotyping error).
Uncommon variant	Has an non-reference allele frequency <2.5% in 1000 Genomes Phase 1 Europeans[e]
Haplotype sharing	Variant is predicted to lie on a haplotype shared by at least 9 affected members of subfamilies (a) and (b)
Conserved	GERP[f] score > 2 or phastCons[g] score > 0.5, using UCSC vertebrate alignments[h]
Regulatory function	Within an Ensembl regulatory region (via VEP[c]) or within both a transcription factor binding site (TFBS) and a region of open chromatic (DNase1) in at least one ENCODE cell line[i] (via UCSC[j])
Filters for structural variants:	
Novel	Does not have >50% reciprocal overlap with a variant in Conrad *et al*[k], 1000 Genomes[l] or HGV[m].
Not a CNV region	Overlaps no more than 5 variants in HGV[m]
Haplotype sharing	Variant overlaps a haplotype shared by at least 9 affected members of subfamilies (a) and (b)
Potential function	Overlaps at least one coding base
Filters for all variants:	
Genic variant	Overlaps a gene region in GenCode release 7[n]
Expressed gene	Gene is expressed in at least one immune or gut tissue type, either in the Gene Expression Barcode[o] or our gene expression datasets.

Table 5.7: Filters used to identify candidate causal variants. [a]NHLBI GO Exome Sequencing Project (ESP) (2012). [b]Wang et al. (2010) [c]McLaren et al. (2010) [d]Gonzalez-Perez and Lopez-Bigas (2011) [e]Project (2012) [f]Davydov et al. (2010) [g]Siepel et al. (2005) [h]Dreszer et al. (2012) [i]The ENCODE Project Consortium (2012) [j]Rosenbloom et al. (2012) [k]Conrad et al. (2010) [l]Mills et al. (2011b) [m]Zhang et al. (2006) [n]Harrow et al. (2006) [o]McCall et al. (2011)

advantage of allowing us to assess the variant in more affecteds than were sequenced. However, if a causal variant has been introduced to the family multiple times on separate haplotypes, this variant will be missed (in the family this is true for the *NOD2* mutations, for example). Thus for the exome sequencing, where data is available for nearly all affecteds, we did not use the haplotype information, instead directly counting the number of affected individuals carrying each haplotype.

The next stage involves removing variants that are unlikely to have a functional impact. Coding SNPs and indels are filtered based on their predicted impact on protein function. Non-coding SNPs and indels from the whole-genome sequence are filtered based on their level of evolutionary conservation and their presence in putative regulatory features. Structural variants are filtered based on whether they delete coding sequence.

The final stage is to remove variants that, while possibly functional, are unlikely to be functionally relevant to IBD risk. We use two sets of gene expression data (one public reference set, one dataset generated by us) to identify genes that are expressed in tissues relevant to IBD (tissues of the immune or digestive systems). All mutations are filtered out if they do not overlap a gene identified as expressed in a relevant tissue.

In the next three sections I will describe the results of this filtering on the three different classes of variant, and discuss some of the candidate variants that this analysis uncovers.

Coding SNPs and indels

Across the entire family there were 7,626 protein-changing mutations that are at low frequency in the general population. Of these, 223 were shared by at least 50% of affecteds in at least one subfamily, and 36 were implicated as

Filter	SNPs	Indels
Low frequency protein mutations	7462	164
Shared by 50% of a subfamily	220	3
Deleterious	72	3
Expressed	35	1

Table 5.8: Summary of the filtering procedure for exome variants

functional in a relevant tissue (Table 5.8).

Ordering by the maximum frequency in affecteds in either subfamily, or across the entire family, the *NOD2* frameshift mutation ranks second in the list of candidates (Table 5.9). This mutation is a the strongest known risk factor for Crohn's disease, and acts as a reassuring positive control, demonstrating that this method can prioritise mutations with relatively low penetrance. This is particularly reassuring as the *NOD2* region was not identified as a suggestive linkage peak or widely shared haplotype, due to it being introduced by multiple founders: this shows that the sequencing and prioritisation approach can identify true associations that the linkage approaches cannot.

The most widely shared novel candidate mutation across the family was a missense mutation in the gene *PDE4FIP*, encoding the protein Myomegalin. This gene has not previously been implicated as having a role in immunity. Next down, a mutation in the gene *PIK3C2A* was found to be widely shared in subfamily (c): this gene is relatively poorly understood, but may play a role in autophagy (Vanhaesebroeck et al., 2010). Towards the top of the list we also find a missense variant in *NLRP2* (a protein known to regulate inflammation in macrophages (Fontalba et al., 2007)) that is shared across subfamilies.

Chr:Pos	Alleles	Affected carriers			Gene	Mutation
		(a+b)	(c)	All		
1:144871738	C/A	16	11	27	*PDE4DIP*	Aka1742Ser (0.73)
16:50763778	G/GC	0	10	15	*NOD2*	Leu1007Fs
11:17191207	T/C	0	10	10	*PIK3C2A*	Lys28Glu (0.55)
11:64527189	C/T	14	0	14	*PYGM*	Arg61His (0.82)
19:55481394	C/T	4	9	13	*NLRP2*	Ser4Leu (0.74)
3:148601439	G/C	1	9	11	*CPA3*	Arg273Pro (0.70)
11:5536759	G/A	0	9	10	*UBQLNL*	Gln305X
3:136664737	C/T	13	1	15	*NCK1*	Ala180Val (0.50)
11:5424701	T/C	5	8	15	*OR51B5*	Ile292Thr (0.86)
11:64854223	C/A	0	8	8	*ZFPL1*	Pro147His (0.55)

Table 5.9: Top 10 SNP protein coding candidate mutations. The number after the amino acid change is the Condel score on the canonical transcript.

Filter	SNPs	Indels
Low frequency mutations on maximal haplotype	125189	38290
Shared by at least 9 affecteds	26993	8501
Conserved base	3143	584
Regulatory function	110	12
Expressed in relevant tissue	74	7

Table 5.10: Summary of the filter procedure for non-coding variants

Non-coding SNPs and indels

A total of 35,494 SNPs and indels were at low frequency in the population, and were shared by at least 9 affecteds in subfamilies (a)+(b) (Table 5.10). Further filtering produced 81 candidate variants, which were both conserved and lay in putative regulatory regions (Table 5.11).

Chrom:Pos	Alleles	Affected	Conservation	Gene	Regulatory features
18:9380839	G/A	14	1.04	*TWSG1*	Ensembl
14:61835929	C/T	11	2.98	*PRKCH*	Ensembl, TFBS(EBF), DNase1
4:169330682	T/C	11	2.11	*DDX60L*	Ensembl, TFBS(c-Fos, junD), DNAse1
1:7018684	G/A	10	4.30	*CAMTA1*	Ensembl, TFBS(KAP1,p300,CEBPB), DNase1
9:112012029	G/C	10	4.17	*EPB41L4B*	TFBS(CTCF,Rad21), DNase1
10:247426	G/A	10	3.58	*ZMYND11*	Ensembl, TFBS(c-Fos,STAT3,c-Jun), DNase1
1:7717739	G/C	10	3.47	*CAMTA1*	Ensembl
1:108439711	A/C	10	3.12	*VAV3*	TFBS(EBF,EBF1), DNase1
22:19347974	C/T	10	2.27	*HIRA*	Ensembl, TFBS(CTCF,Pol2)
2:103039025	T/A	10	2.05	*IL18RAP*	TFBS (GATA2, STAT2, others) DNase1
17:62181435	C/CA	10	0.98	*ENR1*	TFBS(c-Fos,STAT3,others), DNase1
13:98919267	C/CAA	10	0.91	*FARP1*	Ensembl
10:99441651	G/GA	10	0.78	*AVPI1*	Ensembl
1:52150585	C/CT	10	0.61	*OSBPL9*	Ensembl, TFBS(p300), DNase1

Table 5.11: Top 10 SNP candidates, and all 4 indel candidates, shared by at least 10 individuals. SNP conservation is given as the GERP score, and indel conservation is given by the phastCons score.

Filter	SNPs	Indels
Novel insertions or deletions >100bp	2332	262
Shared by at least 9 affecteds	645	80
Delete coding sequence	5	0
Expressed	3	0

Table 5.12: Summary of the filtering procedure for structural variants

No single candidate stood out as both clearly functional and widely shared. Only one potential regulatory mutation was on the maximally shared haplotype (i.e. shared by 14 individuals). This was a novel mutation in a putative regulatory region of *TWSG1* (a gene implicated in BMP signalling and B cell differentiation). However, of the 4 cell lines the regulatory feature was detected as active in, none was related to the immune or digestive system, and there was no clear evidence of transcription factor binding at this position.

There were some promising candidate mutations that were shared by a reduced number of affecteds. A mutation in a B- and T-cell active regulatory region near *PRKCH* (involved in T-cell activation (Fu et al., 2011)) is shared by eleven affecteds. This gene has previously been implicated in susceptibility to atrophic gastritis by a candidate gene study (Goto et al., 2010). Another strong candidate is *IL18RAP*, a receptor for interleukin-18 (known to be important in Crohn's disease (Maerten et al., 2004)), and a candidate causal gene in the IIBDGC Immunochip analysis (Chapter 4). The mutation itself is in a binding site for STAT2, a transcription factor known to be downregulated in IBD (Mudter et al., 2005), though only 10 individuals share this mutation.

Chrom:Pos	Alleles	Affected	Gene (bases deleted)
7:142494034-142495142	1108bp deletion	12	*TRBJ2* (50bp) to *TRBJ6* (48bp)
13:95363645-95363829	184bp deletion	10	*SOX21* (184bp)
4:84221936-84222193	257bp deletion	9	*HSPE* (77bp)

Table 5.13: The three candidate structural variants

Structural variants

725 novel structural variants lay within regions of the genome with haplotypes shared by at least 9 individuals (Table 5.12). Because of the difficulty in genotyping structural variants we were not able to test whether these variants fell on the maximally shared haplotype. Of the 725 mutations, 5 deleted coding sequence, and 3 of these lay within genes expressed in the digestive or immune system (Table 5.13).

The functional structural variant that is most widely shared lies in the T-cell receptor β (TCRB) locus, and appears to delete seven TCRBJ genes (including all the most commonly used ones (Freeman et al., 2009)). At first glance, this makes it an excellent candidate. However, the TCRB region undergoes VDJ recombination during T cell development, and the deletion may well have occurred during normal somatic development. Furthermore, parts of the TCRB region are known to be copy number variable in healthy individuals (Mackelprang et al., 2002), meaning that even a germ-line mutation may be benign. The other two candidate SVs are not particularly widely shared, and do not lie in any obvious candidate genes.

New *NOD2* variants

I mentioned above that the well-established *NOD2* frameshift mutation ranked second in the list of coding candidate variants in the family. The importance of this mutation in the genetics of Crohn's disease led us to specifically investigate *NOD2* variants that our above prioritisation analysis may have missed. Doing so uncovered two new *NOD2* mutations that are likely increasing the risk of IBD in this family.

One mutation was carried in a heterozygous state by one of the spouses that underwent whole-genome sequencing. This mutation (Arg791Gln) is present in dbSNP (rs104895464), but is at very low frequency in the general population (0.1%). It has a high Condel score (0.997) and lies in the middle of the LRR domain: this places it in the "CD sensitive region" described in Chapter 4, section 6.1, and thus is very likely to increase the risk of Crohn's disease. However, the mutation is not very common in the family: It was observed once in the sequencing, and from the haplotype flow we can infer that it was only passed on to one affected offspring.

A second novel *NOD2* mutation is found in the exome sequencing, and occurs at the same base pair as one of the traditional *NOD2* mutations (Gly908Arg). This mutation, Gly908Cys, is not present in 1000 Genomes or ESP datasets, though it has been observed twice (in 662 individuals) in the NIH ClinSeq project (Biesecker et al., 2009; Biesecker, 2012). This allele has an even higher Condel score than the established variant (0.999 vs 0.997), suggesting that it too will increase the risk of Crohn's disease. This mutation was introduced by a spouse, and was passed on to two affected children (both of whom are thus discovered to be compound heterozygous for this and a second *NOD2* mutation). Because of the striking nature of this mutation and the fact that it had (at the time) never been reported before,

we performed capillary sequence validation to confirm its existence.

While both of these mutations likely increase the risk of IBD, both were introduced by spouses and thus are not carried on a haplotype shared across nuclear families. Additionally, as they are together only carried by three affected individuals, they can only explain only a small fraction of the affecteds in the family.

Sample set	*N*	Reason
Affecteds and parents	74	Validation of sites and genotypes
Jewish controls	∼100	Validation of allele frequency
Case/control cohort	∼600	Replication via association
Unaffected siblings	∼250	Replication via transmission
Other multiplex families	∼200	Replication via additional families
Total	∼1200	

Table 5.14: Summary of the samples used for in the replication and validation effort. The columns give the name of the sample set, the number of samples included, the reason for their inclusion.

5.5 Follow-up of candidate causal variants

In the previous section I described a number of candidate variants (120) that could be driving the prevalence of IBD in a multiplex family. The vast majority of these are not associated with IBD: instead, they are likely to be a combination of technical errors and variants that have risen to high frequency in the family by chance.

To reduce the number of candidate causal variants, we have designed a validation and replication exercise to identify erroneous and non-associated candidates. This involves genotyping approximately 1200 samples using 8 Sequenom plexes (around 220 variants). These will consist of candidates from the IBD family discussed above, as well as candidate variants from other families and other important known risk variants (such as the *NOD2* mutations). The samples to be used, as well as the reasons that they are included, are shown in Table 5.14. In this section I will describe the intended validation and replication tests, and discuss their power to confirm or falsify candidate causal variants.

Once these tests have been carried out, and if candidate variants still remain, these variants will be carried forward into functional studies to iden-

tify likely causal mechanisms. I will not discuss these functional experiments here.

5.5.1 Technical validation of causal variants

During the 1000 Genomes loss-of-function project described in chapter 3, we learned that LoF variants are greatly enriched for technical errors compared to other classes of variations (MacArthur et al., 2012). This was not due to any particular property of the variants themselves, but instead due to the fact that loss-of-function variants are extremely rare. In essence, because the number of true loss-of-function variants is depleted relative to other categories, while the number of technical errors is approximately constant regardless of functional category, the proportion of errors is much higher.

The list of candidate variants from the family suffers from a similar effect. We have picked these candidates based on a number of criteria that will diminish the pool of true variants and increase the relative number of errors. The classes of functional variants that we have selected for are known to be under negative selection: coding SNPs predicted to be damaging to protein structure are under strong negative selection (Barreiro et al., 2008), and mutations inside non-coding regulatory regions are also known to be rarer than in the genome as a whole (The ENCODE Project Consortium, 2012). We have also selected variants that are common within the family, but rare in the general population, which itself will inflate the error rate.

In this section I will discuss some sources of technical error in the candidates, and discuss validation strategies that can overcome these problems.

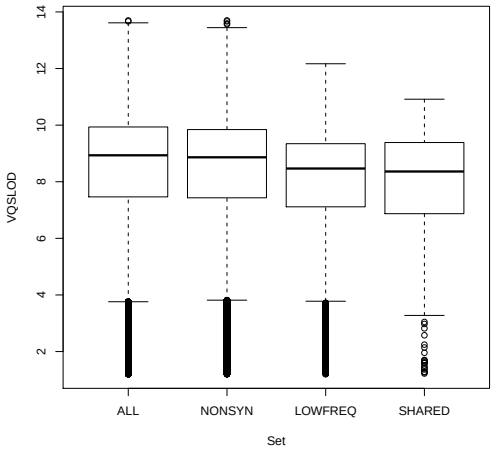


Figure 5.11: The VQSLOD scores for the exome call sets after various sequential filtering steps.

False positive variants

Some of the candidate variants will be false, the result of systematic errors in sequencing. The VQSR calibration will have given us a degree of robustness to such errors, but it is likely that at least some will remain. Figure 5.11 show the VQSLOD score for the exome variants after various stages of filtering. There is a difference in score of approximately 0.8 between the entire exome dataset and the shared, low-frequency coding variants. This shows that systematic errors of the type measured by VQSR are more common in our datasets. More specifically, it corresponds to an estimated 2.2-fold increase in false positive rate in filtered variants (95% CI: 1.6-3.1).

Ideally, all candidate variants should be validated using an independent technology. Capillary resequencing is perhaps the most accurate form of validation (for example, we use this technology to validate the novel *NOD2* variant discussed above), but it is low throughput. PCR amplification is another low-throughput method that can be used to validate structural variants. A

more high-throughput validation technology is the Sequenom (Bradic et al., 2011) mass spectrometry method (the main method used for validation of LoF variants in the 1000 Genomes project). This requires processing a large number of samples to accurately validate sites, but can be combined with the various genotyping efforts described below.

Poorly genotyped variants

Another potential source of false candidates is genotype error. A variant may be real, and present in the family, but some samples have been assigned the wrong genotypes. This can lead a variant that is present only in a small number of individuals to seem to be present in a larger number. This is particularly likely to be a problem in the whole-genome sequencing data, where the coverage is much lower, and incorrect genotypes in a small number of individuals can lead to a variant being incorrectly inferred to lie on a shared haplotype. Again, the most reliable method of detecting these problems is to perform genotyping on the same samples using an independent technology. This can be combined with the site validation described in the previous section.

Common variants

Some of the candidate variants may in fact be at high frequency in the general population. While we have filtered these datasets based on population frequency, there are two factors that may lead a high-frequency variant to remain in the list. Firstly, the variant may be absent from the reference set used, either because it was not detected in the original call list, or was filtered out as poorly performing. Secondly, the variant may be at high frequency exclusively in the Ashkenazi Jewish population. For instance, of the

35,191 exome variants that were below 2.5% in Americans of both European and African descent, 222 were detected at above 10% in the Ashkenazi Jewish control exomes. For the whole-genome sequencing no Jewish controls were available, meaning many of our non-coding candidates may be at high frequency in the Jewish population.

The solution to this problem is to genotype all candidate variants in a control population taken from the same ethnic group and geographic region as the family.

5.5.2 Independent replication of causal variants

Even if the variant is real, is truly low frequency and has been correctly genotyped, it still may be present in a large number of affected family members merely by chance. This is especially true in our case, where we know that this family does not show a genome-wide significance linkage peak, and many of our candidate variants do not lie within even suggestive linkage peaks. To demonstrate that a variant is causal, we need to provide independent replication of the association. In this section I will discuss three different methods of replicating a candidate mutation by genotyping in further samples.

Validation in a case-control cohort

While we filtered out variants with an allele frequency of above 2.5%, many of our candidate variants are still polymorphic in the general population. Such variants may well not be well tagged in GWAS, but case-control cohorts well powered to detect them if genotyped directly. For variants at intermediate frequency (between 0.1% and 1%) we can attempt to replicate these variants in standard case-control cohorts of IBD.

Assuming a risk allele frequency of f, a prevalence of K and a dominant

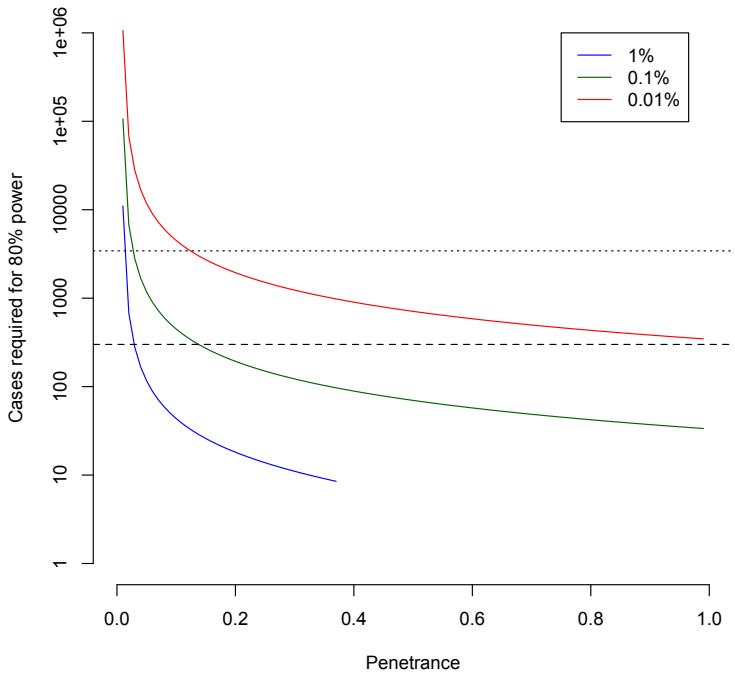


Figure 5.12: The same size required to have 80% power to replicate a mutation with a given penetrance with $p < 0.01$, assuming a prevalence of $K = 0.0075$. The colours of the lines represent the allele frequency in the general population. The dashed line represents a small replication effort (300 cases and 300 controls), and the dotted line represents a large effort (3000 cases and 3000 controls).

penetrance of π (such that $\pi < \frac{K}{f(2-f)}$), the proportion of affecteds in the general population who carry this mutation is

$$P(r=1|d=1) = \frac{P(d=1|r=1)P(r=1)}{P(d=1)} \quad (5.55)$$

$$= f(2-f)\frac{\pi}{K}. \quad (5.56)$$

Similarly, the proportion of unaffected carriers is

$$P(r = 1|d = 0) = f(2 - f)\frac{1 - \pi}{1 - K}. \quad (5.57)$$

The sample size required to detect a difference in the number of carriers between cases and controls, for a given penetrance and allele frequency, is shown in Figure 5.12. A small genotyping effort (300 cases and 300 controls) is well powered to detect (and therefore also to rule out) medium penetrance mutations (>10%) with an allele frequency of greater than 0.1%. A large genotyping effort (such as the whole-genome sequencing experiment described in Chapter 6) would have a power to detect and rule out medium penetrance mutations with a population frequency of greater than 0.01%.

Replicating truly rare mutations is extremely difficult using case-control cohorts, though datasets on the scale of the International IBD Genetics Consortium's replication cohort (discussed in chapter 4) would be well powered to replicate intermediate penetrance mutations with allele frequencies as low as 1 in 200,000.

Validating using unaffected siblings

A standard way to validate a potential causal variant is to track its co-segregation with affection status within the family it was discovered in. In the approach described above, we have prioritised variants for follow-up based on their presence in a large number of affecteds. However, the unaffected siblings of these affected individuals have not been tested, and these unaffected individuals can provide an additional validation set. Where a parent is heterozygous for the candidate mutation, we can test for evidence of causality by testing whether it is transmitted to less than half of unaffected children. Here I will consider what allele frequencies we expect in unaffected siblings as a function of penetrance, and what power these unaffected siblings can

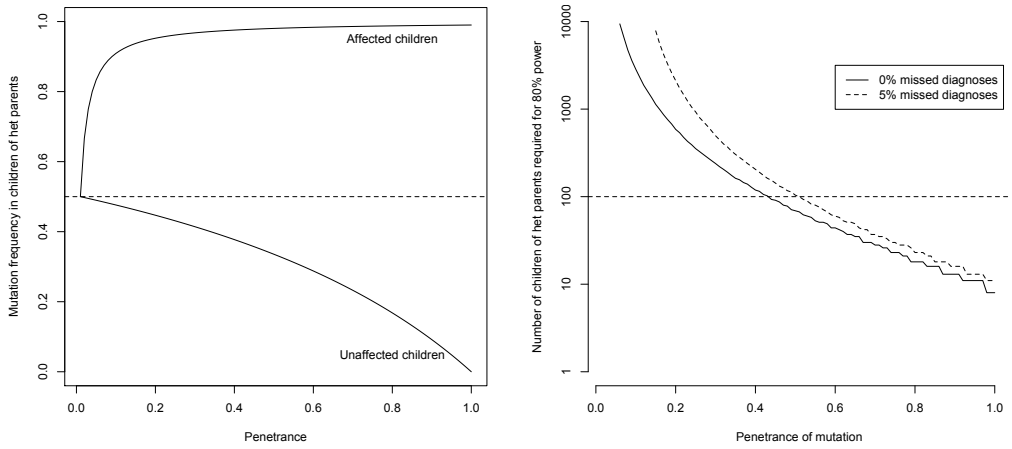


Figure 5.13: a) The frequency of a dominant mutation in affected and unaffected children of an individual heterozygous for this mutation. b) The number of unaffected children of parents heterozygous for the mutation required to validate causality with $p < 0.01$ by a binomial hypothesis test, as a function of the penetrance of the mutation. The solid line represents the case where all unaffected individuals are correctly diagnosed, whereas the dashed line represents a scenario in which 5% of unaffected siblings in fact are (or will become) affected. In both cases I assume $K = 0.0075$.

provide to validate causality.

We will assume that one parent caries the mutation, and there is therefore an even chance that a child will inherit it, i.e. $P(r = 1) = P(r = 0) = \frac{1}{2}$. The overall disease prevalence in the children is thus

$$\begin{aligned} P(d = 1) &= P(d = 1|r = 0)P(r = 0) + P(d = 0|r = 0)P(r = 1) \\ &= \frac{K + \pi}{2}. \end{aligned} \tag{5.58}$$

The proportion of unaffected children who are wild-type is thus

$$P(r=0|d=0) = \frac{P(d=0|r=0)P(r=0)}{P(d=1)} \quad (5.59)$$

$$= \frac{1-K}{(1-K)+(1-\pi)}. \quad (5.60)$$

We can calculate the same value for affected children

$$P(r=0|d=1) = \frac{P(d=1|r=0)P(r=0)}{P(d)} \quad (5.61)$$

$$= \frac{K}{\pi+K}. \quad (5.62)$$

These two equations are plotted as a function of π (for a fixed $K = 0.0075$) in Figure 5.13a. While the mutation frequency in affected children rises very rapidly with the penetrance, the corresponding frequency in unaffecteds falls much more slowly. Figure 5.13b shows the number of unaffected children of heterozygous parents required to validate a candidate mutation at $p < 0.05$. For high penetrance mutations ($\pi > 0.7$) validation can be performed in a modest number of unaffected siblings ($N < 30$), though for intermediate penetrance mutations ($\pi > 0.4$) larger number of unaffecteds are required ($N \sim 100$).

This analysis assumes that all individuals who are currently believed to be unaffected are truly unaffected. However, a proportion of these individuals are likely to have the disease but not yet have been diagnosed, or will go on to develop the disease later in life. This could seriously increase the frequency of the mutation in unaffecteds.

To model this, we assume that a proportion α of the unaffected siblings are in fact cases. We will denote the true affection status with d^T, such that

$P(d^T = 1|d = 0) = \alpha$. The proportion of individuals classified as unaffected who are wild-type is given by

$$\begin{aligned} P(r = 0|d = 0) &= P(r = 0|d^T = 0)P(d^T = 0|d = 0) \\ &\quad + P(r = 0|d^T = 1)P(d^T = 1|d = 0) \qquad (5.63) \\ &= (1 - \alpha)\frac{1 - K}{(1 - K) + (1 - \pi)} + \alpha\frac{K}{K + \pi}. \qquad (5.64) \end{aligned}$$

This diagnostic uncertainty can seriously reduce the power of validation using unaffected siblings. The dashed line in Figure 5.13b shows how many more siblings are needed to account for this diagnostic uncertainty. For instance, to validate a mutation with a penetrance of $\pi = 0.4$ requires $N = 115$ siblings under perfect diagnostic conditions, but $N = 190$ when there is a 5% underdiagnosis rate.

For the candidate variants in the family we are studying, the number of unaffected offspring of carrier parents varies from 50 to 250, depending on the number of subfamilies the mutation is segregating in. We thus will have power to replicate mutations with a high penetrance (>60%) for most mutations, down to about 30% for more widely shared mutations.

Replication in other multiplex families

Perhaps the gold standard for replicating a causal mutation found in a family is to show that it segregates with disease status it in a second family. As we saw in section 5.3.1, multiplex families are more likely to carry more penetrant mutations, and thus screening a large enough number of multiplex families is likely to turn up other instances of the mutation even if the allele frequency in the population is low. For instance, a 0.1% variant with a

penetrance of 50% will be present in 1.3% of cases, but will be present in approximately 10% of patients with at least 2 affected first degree relatives (calculated using the model described in section 5.3). Once such families are identified, affected children of mutation carriers can be tested for an over-inheritance of the mutation.

As we saw in Figure 5.13a, providing that penetrance is above around 20%, affected children of heterozygous parents should be carriers at least 95% of the time. If 8 such affected children can be collected from additional families, and the mutation is causal, more often than not (>65% of the time) all will carry the mutation. However, if the mutation is not causal, there is only a 0.4% chance of all children carrying this mutation. Even for a disease with a 10% penetrance, only 12 children are required to produce the same effect. Thus, identifying less than a dozen affected children in families carrying the mutation is often sufficient to demonstrate causality.

5.6 Conclusions

Discovering high penetrance mutations in multiplex families is, unsurprisingly, a more complex endeavour for complex diseases than for Mendelian disease. We have seen how a large number of multiplex families can arise as a result of polygenic risk alone, and great care must be taken to select families that are likely to carry penetrant mutations. Even if an affected family is detected, a combination of phenocopies, incomplete penetrance and less obviously severe mutations can make correct identification of the causal variants difficult.

Given this, it is not surprising that the above approach did not produce the single, clearly highly damaging mutation shared by all affecteds that would be expected from a Mendelian disease family. Instead, a detailed genotyping, sequencing and filtering experiment produced a series of over a hundred plausible candidates. One of the most valuable resources in the identification of these variants has been tools for inferring both coding and non-coding function, including variant effect prediction, information on regulatory regions, and tissue specific gene expression data. This has allowed us to drastically reduce the list of candidates on the basis of putative function.

A list of multiple candidate variants is likely to be the standard output for family sequencing studies in complex disease. As has been the case with common associations, the key to turning these candidate variants into established associations will be independent replication. I have shown, for certain variants there is potential for replication with unaffected siblings, and within case-control cohorts. However, the most valuable form of replication is likely to be the detection of evidence of co-segregation with affection status in other multiplex families. This highlights the value of collecting samples from many multiplex families, and of collaboration between different research

groups studying multiplex families.

From these observations, I believe that we can identify the two most important developments that will drive forward the study of multiplex families in coming years. The first will be the integration of increasingly detailed functional datasets, and in particular datasets that can assess regulatory function. The second will be collaboration, and in particular reciprocal replication, between research groups in order to establish causal variants in multiple families.