Stony Brook University

Distribution of Number of Rare Variants Appearing in Cases but Not Controls in Genome-wide Studies

A Dissertation Presented

by

Wenjie Xu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2010

Stony Brook University

The Graduate School

Wenjie Xu

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Stephen J. Finch – Advisor

Professor, Applied Mathematics and Statistics

Nancy R. Mendell – Chairperson of Defense

Professor, Applied Mathematics and Statistics

Haipeng Xing

Assistant Professor, Applied Mathematics and Statistics

Eli Hatchwell

Adjunct Professor, Department of Pathology

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of the Graduate School

Abstract of the Dissertation

Distribution of Number of Rare Variants Appearing in Cases but Not Controls in Genome-wide Studies

by

Wenjie Xu

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2010

Whole genome sequencing and whole exome sequencing are developing techniques to explore the associations between rare variants and complex diseases. The number of variants that are expected to appear in a randomly selected group that do not appear in a different group randomly selected from the same population has unknown mean and variance. Expressions for these quantities are derived here. Numerical values are calculated assuming that the frequency of a rare variant has a beta distribution using parameters estimated for four populations. Extensions to the number of variants that appear in r ($r \geq 2$) members of a randomly selected group with none in the comparison group are given. These calculations suggest that a genome wide study of rare variants would generate an extremely large number of false positives. Similarly, an exome wide search would also generate a smaller but still overwhelming number of false positives. A search restricted to variants in a specified gene would not generate excessive numbers of false positives. The expectations using the beta model fit a SNP database well when the underlying beta distribution was restricted to variant frequencies greater than 0.001.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Literature Review

A human genetic variant can be classified as common or rare based on the frequency of the minority allele (MAF) in the human population. Frazer et al. (2009) defined a common variant as one with $MAF \geq 0.01$ and a rare variant as one with $MAF < 0.01$. Other authors have differing definitions. For example, Bodmer and Bonilla (2008) define a common variant as one with $MAF \geq 0.05$.

Human genetic variants can also be classified as single nucleotide polymorphisms (SNPs) or structural variants by their nucleotides' compositions (Feuk et al. 2006). A SNP is a variant restricted to a single nucleotide (A, T, C, or G). Structural variants are DNA sequence variations occurring when more than one connected base pairs differ between individuals.

A third way to classify human genetic variants is to divide them into neutral, near-neutral and non-neutral variants (Frazer et al. 2009). Neutral variants are defined as genetic variants that do not contribute to phenotypic variation. Non-neutral variants are those contributing to phenotypic variation. Near-neutral variants are intermediate in effect. It is hypothesized that most genetic variants are neutral. They may have achieved significant frequencies in the population (Kimura 1968).

Hundreds of complex phenotypic traits determine our physical characteristics and our probability of developing certain diseases (Frazer et al. 2009). These traits are

thought to be influenced by genetic variants, environmental factors, or both. Human genetics research has been trying to identify which variant inheritably determines the components of phenotypes. There are two main hypotheses: the common disease – common variant (CDCV) hypothesis (Lander 1996) and the common disease – rare variant (CDRV) hypothesis (Pritchard 2001). The CDCV hypothesis, as the name implies, states that most of the complex polygenic diseases are largely governed by common variants. The CDRV hypothesis states that complex polygenic diseases are largely governed by rare variants. There is now interest in finding structural variants that are associated with specific complex traits (Conrad and Hurles 2007).

Genome-wide association studies (GWAS) have been widely used to identify the common variants and the statistical associations between SNPs and common variants (Donnelly et al. 2008; McCarthy et al. 2008). Their foundation is the CDCV hypothesis. Many findings report low odds ratios associating a gene with a disease (Iles 2008). This has created doubts about the value of the CDCV hypothesis (Bodmer and Bonilla 2008). Most common structural variants are in linkage disequilibrium (LD) with SNPs and thus should have been assayed by proxy in GWAS (Frazer et al. 2009; Redon et al. 2008; Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006; McCarroll et al. 2008). Since rare variants are not in LD with common variants in general, they are not likely to be detected in a GWAS.

Large-scale sequencing can be used to find causal rare variants. Specifically, whole genome sequencing technologies have steadily decreased in cost. Exome sequencing (Ng et al. 2008), which refers to sequencing the coding regions of the

genome (roughly about 1% of the total (Choi et al. 2009; Ng et al. 2009)), is a less intensive approach. There have been some successful applications of whole genome sequencing and exome sequencing to the analysis of rare, recessive disorders, including Charcot-Marie-Tooth (Lupski et al. 2010), Joubert Syndrome (Edvardson et al. 2010) and Miller Syndrome (Ng et al. 2010). These are examples in which inference is relatively straightforward, in that causal inference of variants was based on the identification of homozygous (or doubly heterozygous) mutations in affected individuals supplemented with family data.

1.2 Determining sample sizes to detect a rare variant in case-control studies

Both single-marker and multiple-marker tests can be used to analyze sequencing data. The single-marker tests include the chi-square test of homogeneity, Fisher's exact test, and the linear trend test (Cochran-Armitage). Li and Leal (2008) proposed several multiple-marker tests including the Hotelling T-squared test, the Combined Multivariate and Collapsing(CMC) method, and the collapsing method, while Madsen and Browning (2009) proposed the weighted sum method. Additional tests are proposed regularly in the genetic epidemiology literature.

Case-control studies are often used to detect the association between a rare variant and a phenotype. Candidate genes are sequenced for each member of the case

and control groups. If there is a significant difference in frequency between the two groups for a variant, then it is considered as a candidate variant which may have contributed to inherited susceptibility.

Here I report the power of Fisher's exact test to detect the association between a rare variant and a disease for a range of sample sizes for the control and case groups. I assume: The prevalence of disease *D* is 1%. There are a number of rare variants that cause disease *D*. Of those affected, 1% possess a specific variant *V*. Every individual with variant *V* has the disease *D*; that is, the relative risk of having disease *D* for those with variant *V* compared to those without *V* is extremely large. An association study with N_A cases and N_U controls will be run. The null hypothesis is that there is no association between those who have disease *D* and those who have variant *V*. The alternative hypothesis is there is an association. The question I answer is how many N_A cases and N_U controls are needed to have power=0.9 when α=0.05? The definitions that I use are shown in Table 1.

Table 1
Definition of Terms for Power Analysis

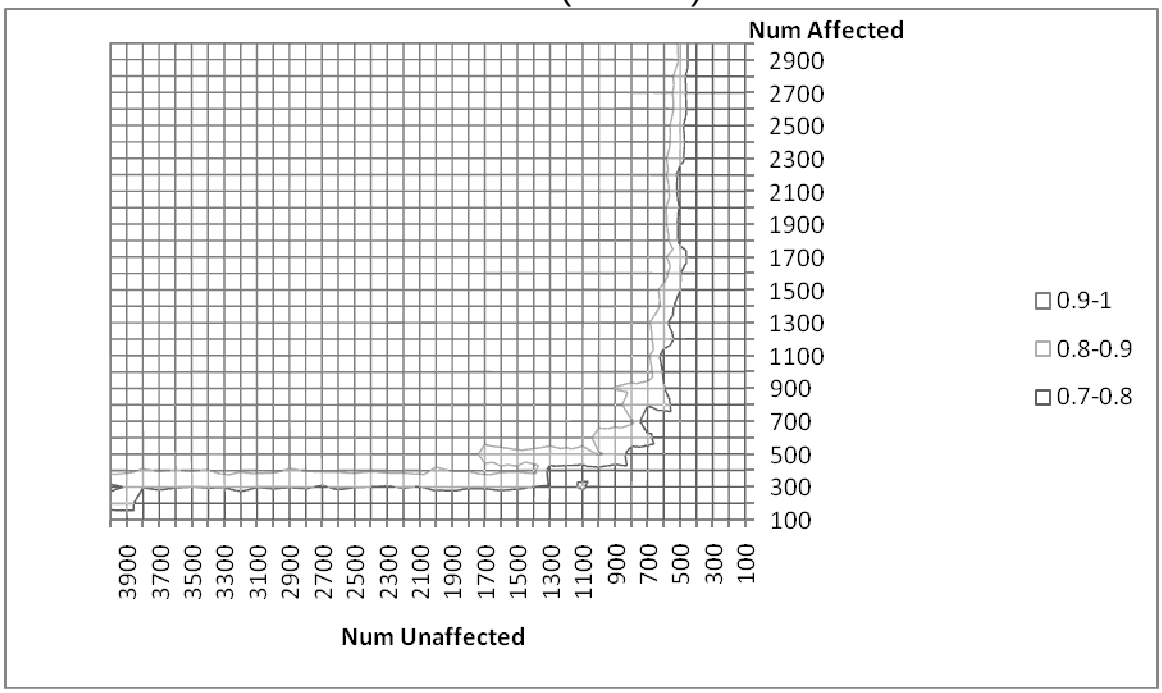
Variant\Affection Status	Affected	Unaffected
With V	*X*	*Y*
Without V	$N_A - X$	$N_U - Y$
Total	N_A	N_U

That is, the assumptions are that $E(X) = 0.01N_A$ and $E(Y) = 0$. Figure 1 contains the simulated power calculated using R for Fisher's exact test (Smyth 2007).

The horizontal axis is the number of unaffected; the vertical axis is the number of affected; and the contour represent combinations of sample size with power 0.80 or 0.90 for Fisher's exact statistic. The powers are simulated values based on 1000 replications at each combination of N_A cases and N_U controls.

The figure 1 shows that with 100 affected people in the case group the power of the Fisher's exact test did not have power above 0.8, even with 4,000 in the control group. With 300 affected people, 1300 people are needed in the control group to reach power 0.80. Further increase of the number of unaffected does not substantially increase the power. In terms of the summation of the number of affected and the number of unaffected, a design with 600 affected and 700 unaffected has the smallest number of subjects and power above 0.80 for the values studied. A design with 700 affected and 800 unaffected has the smallest total number and power above 0.90.

Figure 1
Contour Plot of Number of Cases, Number of Controls,
and Simulated Power of Fisher's Exact Test
($\alpha = 0.05$)



1.3 Estimates of total number of variants in genome and in coding regions

Ionita-Laza et al. (2009) estimated that the total number of variants in a 5 Mb region of the genome was 13,270 in the CEPH European population, using the ENCODE database (ENCODE Project Consortium 2004; Birney et al. 2007). Linear extrapolation using 3,000 Mb as the length of the whole genome yields an estimate of approximately 8 million variants in the European population. Similarly, the extrapolation of the report of Ionita-Laza et al. (2009) using the NIEHS SNP database was

approximately 12 million variants for the European population. That is, the number of variants in human populations, *I*, is of the order of 10,000,000. Table 2 gives estimates of the total number of variants for other populations using these two databases.

Table 2
Estimates of the total number of variants in selected human populations

	African	European	Chinese	Japanese
ENCODE	10,541,000	7,978,000	6,508,000	5,818,000
NIEHS SNPs	20,733,000	12,065,000	11,636,000	11,100,000

With regard to exome sequencing, Li et al. (2010) identified 53,081 coding SNPs with MAF greater than 0.02 by sequencing 200 Danish exomes. Extrapolation using estimates of parameters reported by Ionita-Laza et al. (2009) yields an estimate of approximately 115,000 exonic variants in the ENCODE CEPH population. Similarly the extrapolation is approximately 185,000 exonic variants in the NIEHS SNPs European population.

1.4 Objective of the dissertation

Current sequencing techniques are still limited by the following problems:

1. Complex disorders may have multiple (perhaps hundreds) genetic components (Ropers 2006; Collins 1999; Frayling 2007; Frazer et al. 2009);

2. Most causal variants are found in coding regions, but there are also examples showing the non-coding regions play important roles for certain disorders (Maller et al. 2006; Li et al. 2009; Hammoud et al. 2007);
3. The number of variants between the whole genome sequences of individuals is very large. The average difference between two individuals is on the order of 3-4,000,000 variants (including SNPs, indels, CNVs, and more complex alterations) (Lewontin 2005; Kruglyak and Nickerson 2001);
4. The significance of variants even within a single gene of interest may be problematic. Variants previously believed to be of no significance have been found to be relevant (Ramamurthi, 2005; Duan 2003; Greally 2007). Variants previously believed to be causal to some disorders have also been found not to be associated (Mutsuddi et al. 2006).

In order for inferences to be made for variants found in disease cohorts, I aim to:

1. Provide estimates for the total number of whole genome sequences that must be obtained from normal (control) populations;
2. Where exonic variants are sought, provide estimates for the total number of exome sequences that must be obtained from normal (control) populations;
3. Where causal variants in a specific gene are sought, provide estimates for the total number of gene sequences that must be obtained from normal (control) populations. For example, FBN1 is related with Marfan syndrome and related

disorders (Dietz et al. 1993), and NF1 is related with neurofibromatosis and related diseases (Johnson et al. 1993).

My approach, under the assumption of variable MAF for each variant, following Ionita-Laza et al. (2009), has been as follows:

1. theoretical assessment assuming 10,000,000 variants in the genome;
2. theoretical assessment assuming 150,000 variants in the exome;
3. theoretical assessment considering the number of variants in specific genes;
4. empirical assessment of Chr1-22/500k SNP data using permutation modeling.

The approach to assessment using the (*r, h*) specification can be formulated as follows. The SNP $i, i = 1,\mathrm{K}\,, I$, where I is the number of SNPs genotyped on the platform, is said to appear in the smaller group when at least *r* individuals in the group has the minority allele. It is said to appear in the larger group when it appears in at least *h* members. In the comparison of two groups, the number of SNPs in both groups is denoted $N_{++}^{(r,h)}$; the number in the smaller group but not in the larger is denoted $N_{+-}^{(r,h)}$; the number in the larger group but not in the smaller is denoted $N_{-+}^{(r,h)}$; and the number not in either group is denoted $N_{--}^{(r,h)}$. The questions are: How many variants $N_{++}^{(r,h)}$ are expected to appear in both samples? How many variants $N_{+-}^{(r,h)}$ are expected to appear in the smaller sample but not the larger sample? How many variants $N_{-+}^{(r,h)}$ expected to appear in the larger sample but not the smaller? How many variants $N_{--}^{(r,h)}$ will not

appear in either sample? For design purposes, since there are usually fewer affected subjects than unaffected subjects in a study of rare variants, $N_{+-}^{(r,h)}$ is the most important of these variables. I seek to identify specifications that have small $E(N_{+-}^{(r,h)})$.

2. Methods

2.1 Empirical assessment

In my empirical assessment of the modeling of the minority frequency of the variants, I use the Framingham Heart Study (FHS) data as a proxy for sequencing data as provided in the Genetics Analysis Workshop 16 (Cupples et al. 2009). The FHS data has 1,599 unrelated participants who have been genotyped on a 500K platform. I compare 10 sets of 100 participants selected randomly without replacement to those not selected. I also compare 10 sets of 400 participants selected randomly without replacement, and 10 sets of 400 randomly selected without replacement to those not selected (about 1200). Finally I compare 10 sets of 800 randomly selected participants to those not selected.

2.2 Theoretical assessment – expectations when the variants appear at least once in each group

I assume that the total number of variants is I and that the occurrence of variant i is independent of the occurrence of variant j, $i \neq j, 1 \leq i, j \leq I$. Ionita-Laza et al. (2009) modeled the frequency of a variant with a beta distribution. That is, let f_i be the frequency of variant i. Then, the Ionita-Laza model is that $f_i \sim Beta(a,b)$. Ionita-Laza et al. (2009) report estimates of a and b for populations in the ENCODE project (http:www.hapmap.org/downloads/encode1.html.en) and NIEHS study (http://egp.gs.washington.edu/) in their supplemental material. Table 3 presents their estimates.

Table 3
Estimates of parameters for populations from ENCODE and NIEHS SNPs Databases

		African	European	Chinese	Japanese
ENCODE	$\hat{a}$	0.07	0.14	0.22	0.35
	$\hat{b}$	0.97	0.73	0.77	0.86
NIEHS SNPs	$\hat{a}$	0.036	0.076	0.058	0.064
	$\hat{b}$	1.06	0.72	0.63	0.6

(Source: Ionita-Laza et al. 2009 Supplemental Material)

In this dissertation, I use the beta distribution model with left and right truncation. In my empirical analysis of SNP data to justify the beta distribution assumption, I use left truncation at *c*, a value chosen to make the theoretical results as close as possible to the empirical results. The right truncation *d* is allowed to keep the analysis as general as possible. That is, the truncation condition is $c \leq f_i \leq d$.

For sample size n in group 1 and m in group 2 and specific variant *i*,

$\Pr(i \text{ not in group } 1) = g_i^{\,n}$, where $g_i = 1 - f_i$ with truncation condition that $1 - d \le g_i \le 1 - c$;

$\Pr(i \text{ not in group } 2) = g_i^{\,m}$.

$\Pr(i \text{ in group } 1) = 1 - g_i^{\,n}$.

$\Pr(i \text{ in group } 2) = 1 - g_i^{\,m}$.

$\Pr(i \text{ in both groups}) = (1 - g_i^{\,n})(1 - g_i^{\,m}) = p_{i1}$.

$\Pr(i \text{ in group 1 but not in group 2}) = g_i^{\,m}(1 - g_i^{\,n}) = p_{i2}$.

$\Pr(i \text{ in group 2 but not in group 1}) = g_i^{\,n}(1 - g_i^{\,m}) = p_{i3}$.

$\Pr(i \text{ not in either group}) = g_i^{\,n+m} = p_{i4}$.

Then, N_{++}, the number of variants in both groups, is given by

$N_{++} = \sum_{i=1}^{I} X_i$, where $X_i \mid g_i \sim \text{Bernoulli}(p_{i1})$ so that

$$EN_{++} = \sum_{i=1}^{I} E(E(X_i \mid g_i)) = \sum_{i=1}^{I} E(p_{i1}) = IE(1 - g_i^{\,n})(1 - g_i^{\,m}) = I(1 - Eg_i^{\,n} - Eg_i^{\,m} + Eg_i^{\,n+m}).$$

Since

$$Eg_i^{\,n} = \frac{\int_{1-d}^{1-c} t^n \frac{1}{B(b,a)} t^{b-1}(1-t)^{a-1} dt}{\int_{1-d}^{1-c} \frac{1}{B(b,a)} t^{b-1}(1-t)^{a-1} dt} = \frac{\int_{1-d}^{1-c} t^{n+b-1}(1-t)^{a-1} dt}{\int_{1-d}^{1-c} t^{b-1}(1-t)^{a-1} dt} = \frac{B(1-c;b+n,a) - B(1-d;b+n,a)}{B(1-c;b,a) - B(1-d;b,a)},$$

where $B(b,a)$ is the beta function and $B(c;b,a)$ is the incomplete beta function, that is,

$B(b,a) = \int_0^1 t^{b-1}(1-t)^{a-1} dt$ and $B(c;b,a) = \int_0^c t^{b-1}(1-t)^{a-1} dt$.

Let $M(a,b,c,d,n) = \frac{B(1-c;b+n,a) - B(1-d;b+n,a)}{B(1-c;b,a) - B(1-d;b,a)}$. Then

$EN_{++} = I[1 - M(a,b,c,d,n) - M(a,b,c,d,m) + M(a,b,c,d,n+m)]$, and

$$VAR(N_{++}) = \sum_{i=1}^{I} Var(X_i) = \sum_{i=1}^{I} [Var(E(X_i \mid g_i)) + E(Var(X_i \mid g_i))]$$

$$= \sum_{i=1}^{I} [Var(p_{i1}) + E(p_{i1}(1-p_{i1}))] = \sum_{i=1}^{I} [Ep_{i1}^2 - (Ep_{i1})^2 + Ep_{i1} - Ep_{i1}^2] = \sum_{i=1}^{I} [Ep_{i1} - (Ep_{i1})^2]$$

$$= I[1 - M(a,b,c,d,n) - M(a,b,c,d,m) + M(a,b,c,d,n+m)]$$

$$- I[1 - M(a,b,c,d,n) - M(a,b,c,d,m) + M(a,b,c,d,n+m)]^2$$

Similarly, the expected number of variants in group 1 but not in group 2 and its variance are

$$EN_{+-} = IEg_i^{\ m}(1 - g_i^{\ n}) = I(Eg_i^{\ m} - Eg_i^{\ n+m}) = I[M(a,b,c,d,m) - M(a,b,c,d,n+m)].$$

$$VAR(N_{+-})$$

$$= I[M(a,b,c,d,m) - M(a,b,c,d,n+m)] - I[M(a,b,c,d,m) - M(a,b,c,d,n+m)]^2$$

The expected number of variants in group 2 but not in group 1 and its variance are given by:

$$EN_{-+} = IEg_i^{\ n}(1 - g_i^{\ m}) = I(Eg_i^{\ n} - Eg_i^{\ n+m}) = I[M(a,b,c,d,n) - M(a,b,c,d,n+m)]$$

$$VAR(N_{-+})$$

$$= I[M(a,b,c,d,n) - M(a,b,c,d,n+m)] - I[M(a,b,c,d,n) - M(a,b,c,d,n+m)]^2.$$

The number of variants not appearing in both group and its variance are:

$$EN_{--} = IEg_i^{\ n+m} = IM(a,b,c,d,n+m).$$

$$VAR(N_{--}) = I[M(a,b,c,d,n+m) - M^2(a,b,c,d,n+m)].$$

2.3 Theoretical assessment – expectations when the variants appear at least twice in case group and at least once in control group

I next consider an extension in which the SNP $i, i = 1,\text{K}\ , I$, is said to appear in the smaller group when at least two individuals in that group have the minority allele. It is said to appear in the larger group when it appears at least once. I call this the specification (2,1). Group 1 is a random sample of *n* individuals and is used as a model of a case group, and group 2 is a random sample of *m* individuals and is used as a model of a control group; that is $n \le m$. Then,

$$\Pr(i \text{ not in group } 1) = g_i^{\ n} + n(1-g_i)g_i^{n-1} = ng_i^{n-1} - (n-1)g_i^{\ n} .$$

$$\Pr(i \text{ in group } 1) = 1 - ng_i^{n-1} + (n-1)g_i^{\ n} .$$

$$\Pr(i \text{ not in group } 2) = g_i^{\ m} .$$

$$\Pr(i \text{ in group } 2) = 1 - g_i^{\ m} .$$

$$\Pr(i \text{ in both groups}) = (1 - ng_i^{n-1} + (n-1)g_i^{\ n})(1 - g_i^{\ m})$$
$$= 1 - ng_i^{n-1} + (n-1)g_i^{\ n} - g_i^{\ m} + ng_i^{n+m-1} - (n-1)g_i^{n+m} = p_{i1}^{(2,1)}$$

$$\Pr(i \text{ in group 1 but not in group 2}) = (1 - ng_i^{n-1} + (n-1)g_i^{\ n})g_i^{\ m}$$
$$= g_i^{\ m} - ng_i^{n+m-1} + (n-1)g_i^{n+m} = p_{i2}^{(2,1)} .$$

$$\Pr(i \text{ in group 2 but not in group 1}) = (1 - g_i^{\ m})(ng_i^{n-1} - (n-1)g_i^{\ n})$$
$$= ng_i^{n-1} - (n-1)g_i^{\ n} - ng_i^{n+m-1} + (n-1)g_i^{n+m} = p_{i3}^{(2,1)} .$$

$$\Pr(i \text{ not in either group}) = (ng_i^{n-1} - (n-1)g_i^{\ n})g_i^{\ m}$$
$$= ng_i^{n+m-1} - (n-1)g_i^{n+m} = p_{i4}^{(2,1)} .$$

Then the expectation and variance of the number of variants appearing in both groups are

$$EN_{++}^{(2,1)}$$
$$= I[1 - nEg_i^{n-1} + (n-1)Eg_i^n - Eg_i^m + nEg_i^{n+m-1} - (n-1)Eg_i^{n+m}]$$
$$= I[1 - nM(a,b,c,d,n-1) + (n-1)M(a,b,c,d,n) - M(a,b,c,d,m) + nM(a,b,c,d,n+m-1)$$
$$- (n-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{++}^{(2,1)})$$
$$= I[1 - nM(a,b,c,d,n-1) + (n-1)M(a,b,c,d,n) - M(a,b,c,d,m) + nM(a,b,c,d,n+m-1)$$
$$- (n-1)M(a,b,c,d,n+m)]$$
$$- I[1 - nM(a,b,c,d,n-1) + (n-1)M(a,b,c,d,n) - M(a,b,c,d,m) + nM(a,b,c,d,n+m-1)$$
$$- (n-1)M(a,b,c,d,n+m)]^2$$

The expectation and variance of the number of variants appearing in group 1 but not in group 2 are

$$EN_{+-}^{(2,1)} = I[M(a,b,c,d,m) - nM(a,b,c,d,n+m-1) + (n-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{+-}^{(2,1)}) = I[M(a,b,c,d,m) - nM(a,b,c,d,n+m-1) + (n-1)M(a,b,c,d,n+m)]$$
$$- I[M(a,b,c,d,m) - nM(a,b,c,d,n+m-1) + (n-1)M(a,b,c,d,n+m)]^2$$

The expectation and variance of the number of variants appearing in group 2 but not in group 1 are

$$EN_{-+}^{(2,1)} = I[nM(a,b,c,d,n-1) - (n-1)M(a,b,c,d,n) - nM(a,b,c,d,n+m-1)$$
$$+ (n-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{-+}^{(2,1)}) = I[nM(a,b,c,d,n-1) - (n-1)M(a,b,c,d,n) - nM(a,b,c,d,n+m-1)$$
$$+ (n-1)M(a,b,c,d,n+m)] - I[nM(a,b,c,d,n-1) - (n-1)M(a,b,c,d,n)$$
$$- nM(a,b,c,d,n+m-1) + (n-1)M(a,b,c,d,n+m)]^2$$

The expectation and variance of the number of variants not appearing in both groups are

$$EN_{--}^{(2,1)} = IE(ng_i^{n+m-1} - (n-1)g_i^{n+m})$$
$$= I[nM(a,b,c,d,n+m-1) - (n-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{--}^{(2,1)}) = I[nM(a,b,c,d,n+m-1) - (n-1)M(a,b,c,d,n+m)]$$
$$- I[nM(a,b,c,d,n+m-1) - (n-1)M(a,b,c,d,n+m)]^2$$

2.4 Theoretical assessment – expectations when the variants appear at least *r* times in case group and at least once in control group

I next consider an extension in which the SNP $i, i = 1,\text{K}\ , I$, is said to appear in the smaller group when at least *r* individuals in that group have the minority allele. I call this the specification (*r*,1). As before, group 1 is a random sample of *n* individuals: group 2 is a random sample of *m* individuals; and $n \le m$. Then,

$$\Pr(i \text{ not in group } 1) = g_i{}^n + \binom{n}{1}(1-g_i)g_i^{n-1} + \text{K} + \binom{n}{r-1}(1-g_i)^{r-1}g_i^{n-r+1}$$
$$= \left[1 - \binom{n}{1} + \binom{n}{2} - \text{K} + (-1)^{r-1}\binom{n}{r-1}\right]g_i{}^n + \left[\binom{n}{1} - 2\binom{n}{2} + 3\binom{n}{3} + ... + (-1)^r r\binom{n}{r-1}\right]g_i^{n-1}$$
$$+ \left[\binom{n}{2} - 3\binom{n}{3} + 6\binom{n}{3} + ... + (-1)^{r+1}u_{3,r}\binom{n}{r-1}\right]g_i^{n-2} + \text{K}$$
$$+ \left[\binom{n}{j} - u_{j+1,j+2}\binom{n}{3} + u_{j+1,j+3}\binom{n}{3} + ... + (-1)^{j+r-1}u_{j,r}\binom{n}{r-1}\right]g_i^{n-j}$$
$$+ ... + \binom{n}{r-1}g_i^{n-r+1}$$
$$= (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \text{K}\ , (-1)^{r-1}g_i^{n-r+1})U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T$$

where $U = (u_{i,j})$ is an *r* by *r* Pascal upper triangular matrix. A Pascal upper triangular matrix is of the form:

$$U = \begin{pmatrix} 1 & 1 & 1 & 1 & \Lambda & 1 \\ & 1 & 2 & 3 & \mathrm{K} & r-1 \\ & & 1 & 3 & \mathrm{K} & \binom{r-1}{2} \\ & & & \mathrm{O} & \mathrm{O} & \mathrm{M} \\ & & & & 1 & r-1 \\ & & & & & 1 \end{pmatrix};$$

the entries below the diagonal are 0.

$$\Pr(i \text{ in group } 1) = 1 - \Pr(i \text{ not in group } 1)$$

$$= 1 - (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T.$$

$$\Pr(i \text{ not in group } 2) = g_i^m.$$

$$\Pr(i \text{ in group } 2) = 1 - g_i^m.$$

$$\Pr(i \text{ in both groups})$$

$$= (1 - g_i^m) \left(1 - (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T \right)$$

$$= 1 - g_i^m - (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T$$

$$+ (g_i{}^{n+m}, -g_i^{n+m-1}, g_i^{n+m-2}, \mathrm{K}, (-1)^{r-1} g_i^{n+m-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T = p_{i1}^{(r,1)}$$

Pr(i in group 1 but not in group 2)

$$= \left(1-(g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T \right) g_i^m$$

$$= g_i^m - (g_i{}^{n+m}, -g_i^{n+m-1}, g_i^{n+m-2}, \mathrm{K}, (-1)^{r-1} g_i^{n+m-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T$$

$$= p_{i2}^{(r,1)} .$$

Pr(i in group 2 but not in group 1)

$$= (1-g_i^m)(g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T$$

$$= (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T$$

$$- (g_i{}^{n+m}, -g_i^{n+m-1}, g_i^{n+m-2}, \mathrm{K}, (-1)^{r-1} g_i^{n+m-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T = p_{i3}^{(r,1)} .$$

Pr(i not in either group)

$$= g_i^m (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T .$$

$$= (g_i{}^{n+m}, -g_i^{n+m-1}, g_i^{n+m-2}, \mathrm{K}, (-1)^{r-1} g_i^{n+m-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T = p_{i4}^{(r,1)}$$

Since the concern is N_{+-}, I focus on its mean and variance under selected specifications. The expectation and variance of the number of variants appearing in group 1 but not in group 2 are

$$EN_{+-}^{(r,1)}$$
$$= IEg_i{}^m$$
$$- I(Eg_i{}^{n+m}, -Eg_i^{n+m-1}, Eg_i^{n+m-2}, \mathrm{K}\ , (-1)^{r-1} Eg_i^{n+m-r+1}) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T$$
$$= I[M(a,b,c,d,m) - (M(a,b,c,d,n+m), -M(a,b,c,d,n+m-1), M(a,b,c,d,n+m-2), \mathrm{K}\ ,$$
$$(-1)^{r-1} M(a,b,c,d,n+m-r+1)) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T\Bigg]$$

$$VAR(N_{+-}^{(r,1)})$$
$$= I[M(a,b,c,d,m) - (M(a,b,c,d,n+m), -M(a,b,c,d,n+m-1), M(a,b,c,d,n+m-2), \mathrm{K}\ ,$$
$$(-1)^{r-1} M(a,b,c,d,n+m-r+1)) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T\Bigg]$$
$$- I[M(a,b,c,d,m) - (M(a,b,c,d,n+m), -M(a,b,c,d,n+m-1), M(a,b,c,d,n+m-2), \mathrm{K}\ ,$$
$$(-1)^{r-1} M(a,b,c,d,n+m-r+1)) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T\Bigg]^2$$

2.5 Theoretical assessment – expectations when the variants appear at least twice in each group

I next consider an extension in which the SNP $i, i = 1, \mathrm{K}\ , I$, is said to appear in a group when at least two individuals in that group have the minority allele. That is, the specification is (2,2). As before, group 1 is a random sample of *n* individuals, and group 2 is a random sample of *m* individuals. Then,

$\Pr(i \text{ not in group } 1) = g_i^{\,n} + n(1-g_i)g_i^{n-1} = ng_i^{n-1} - (n-1)g_i^{\,n}$.

$\Pr(i \text{ not in group } 2) = mg_i^{m-1} - (m-1)g_i^{\,m}$.

$\Pr(i \text{ in group } 1) = 1 - ng_i^{n-1} + (n-1)g_i^{\,n}$.

$\Pr(i \text{ in group } 2) = 1 - mg_i^{m-1} + (m-1)g_i^{\,m}$.

$\Pr(i \text{ in both groups}) = (1 - ng_i^{n-1} + (n-1)g_i^{\,n})(1 - mg_i^{m-1} + (m-1)g_i^{\,m})$
$= 1 - ng_i^{n-1} + (n-1)g_i^{\,n} - mg_i^{m-1} + (m-1)g_i^{\,m}$
$+ nmg_i^{n+m-2} + (n+m-2nm)g_i^{n+m-1} + (n-1)(m-1)g_i^{n+m} = p_{i1}^{(2,2)}$

$\Pr(i \text{ in group 1 but not in group } 2) = (1 - ng_i^{n-1} + (n-1)g_i^{\,n})(mg_i^{m-1} - (m-1)g_i^{\,m})$
$= mg_i^{m-1} - (m-1)g_i^{\,m} - nmg_i^{n+m-2} + (2nm-n-m)g_i^{n+m-1} - (n-1)(m-1)g_i^{n+m} = p_{i2}^{(2,2)}$.

$\Pr(i \text{ in group 2 but not in group } 1) = (1 - mg_i^{m-1} + (m-1)g_i^{\,m})(nf_i^{\,n-1} - (n-1)g_i^{\,n})$
$= ng_i^{n-1} - (n-1)g_i^{\,n} - nmg_i^{n+m-2} + (2nm-n-m)g_i^{n+m-1} - (n-1)(m-1)g_i^{n+m} = p_{i3}^{(2,2)}$.

$\Pr(i \text{ not in either group}) = (ng_i^{n-1} - (n-1)g_i^{\,n})(mg_i^{m-1} - (m-1)g_i^{\,m})$
$= nmg_i^{n+m-2} + (n+m-2nm)g_i^{n+m-1} + (n-1)(m-1)g_i^{n+m} = p_{i4}^{(2,2)}$.

Then the expectation and variance of the number of variants appearing in both groups are

$$EN_{++}^{(2,2)} = \sum_{i=1}^{I} E(E(X_i \mid g_i)) = \sum_{i=1}^{I} E(p_{i1}^{(2,2)})$$
$$= I[1 - nEg_i^{n-1} + (n-1)Eg_i^{\,n} - mEg_i^{m-1} + (m-1)Eg_i^{\,m}]$$
$$+ I[nmEg_i^{n+m-2} + (n+m-2nm)Eg_i^{n+m-1} + (n-1)(m-1)Eg_i^{n+m}]$$
$$= I[1 - nM(a,b,c,d,n-1) + (n-1)M(a,b,c,d,n) - mM(a,b,c,d,m-1)$$
$$+ (m-1)M(a,b,c,d,m) + nmM(a,b,c,d,n+m-2) + (n+m-2nm)M(a,b,c,d,n+m-1)$$
$$+ (n-1)(m-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{++}^{(2,2)}) = \sum_{i=1}^{I} Var(X_i) = \sum_{i=1}^{I} [Var(E(X_i \mid g_i)) + E(Var(X_i \mid g_i))]$$
$$= \sum_{i=1}^{I} [Var(p_{i1}^{(2,2)}) + E(p_{i1}^{(2,2)}(1 - p_{i1}^{(2,2)}))]$$
$$= \sum_{i=1}^{I} [E(p_{i1}^{(2,2)})^2 - (Ep_{i1}^{(2,2)})^2 + Ep_{i1}^{(2,2)} - E(p_{i1}^{(2,2)})^2] = \sum_{i=1}^{I} [Ep_{i1}^{(2,2)} - (Ep_{i1}^{(2,2)})^2]$$
$$= I[1 - nM(a,b,c,d,n-1) + (n-1)M(a,b,c,d,n) - mM(a,b,c,d,m-1)$$
$$+ (m-1)M(a,b,c,d,m) + nmM(a,b,c,d,n+m-2) + (n+m-2nm)M(a,b,c,d,n+m-1)$$
$$+ (n-1)(m-1)M(a,b,c,d,n+m)] - I[1 - nM(a,b,c,d,n-1) + (n-1)M(a,b,c,d,n)$$
$$- mM(a,b,c,d,m-1) + (m-1)M(a,b,c,d,m) + nmM(a,b,c,d,n+m-2)$$
$$+ (n+m-2nm)M(a,b,c,d,n+m-1) + (n-1)(m-1)M(a,b,c,d,n+m)]^2$$

The expectation and variance of the number of variants appearing in group 1 but not in group 2 are

$$EN_{+-}^{(2,2)} = IE(mg_i^{m-1} - (m-1)g_i^m - nmg_i^{n+m-2} + (2nm-n-m)g_i^{n+m-1} - (n-1)(m-1)g_i^{n+m})$$
$$= I[mM(a,b,c,d,m-1) - (m-1)M(a,b,c,d,m) - nmM(a,b,c,d,n+m-2)$$
$$+ (2nm-n-m)M(a,b,c,d,n+m-1) - (n-1)(m-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{+-}^{(2,2)}) = I[mM(a,b,c,d,m-1) - (m-1)M(a,b,c,d,m) - nmM(a,b,c,d,n+m-2)$$
$$+ (2nm-n-m)M(a,b,c,d,n+m-1) - (n-1)(m-1)M(a,b,c,d,n+m)]$$
$$- I[mM(a,b,c,d,m-1) - (m-1)M(a,b,c,d,m) - nmM(a,b,c,d,n+m-2)$$
$$+ (2nm-n-m)M(a,b,c,d,n+m-1) - (n-1)(m-1)M(a,b,c,d,n+m)]^2$$

The expectation and variance of the number of variants appearing in group 2 but not in group 1 are

$$EN_{-+}^{(2,2)} = I[nM(a,b,c,d,n-1) - (n-1)M(a,b,c,d,n) - nmM(a,b,c,d,n+m-2)$$
$$+ (2nm-m-n)M(a,b,c,d,n+m-1) - (n-1)(m-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{-+}^{(2,2)}) = I[nM(a,b,c,d,n-1) - (n-1)M(a,b,c,d,n) - nmM(a,b,c,d,n+m-2)$$
$$+ (2nm-m-n)M(a,b,c,d,n+m-1) - (n-1)(m-1)M(a,b,c,d,n+m)]$$
$$- I[nM(a,b,c,d,n-1) - (n-1)M(a,b,c,d,n) - nmM(a,b,c,d,n+m-2)$$
$$+ (2nm-m-n)M(a,b,c,d,n+m-1) - (n-1)(m-1)M(a,b,c,d,n+m)]^2$$

The expectation and variance of the number of variants not appearing in either group are

$$EN_{--}^{(2,2)} = IE(nmg_i^{n+m-2} + (n+m-2nm)g_i^{n+m-1} + (n-1)(m-1)g_i^{n+m})$$
$$= I[nmM(a,b,c,d,n+m-2) - (2nm-n-m)M(a,b,c,d,n+m-1).$$
$$+ (n-1)(m-1)M(a,b,c,d,n+m)]$$

$$VAR(N_{--}^{(2,2)})$$
$$= I[nmM(a,b,c,d,n+m-2) - (2nm-n-m)M(a,b,c,d,n+m-1)$$
$$+ (n-1)(m-1)M(a,b,c,d,n+m)]$$
$$- I[nmM(a,b,c,d,n+m-2) - (2nm-n-m)M(a,b,c,d,n+m-1)$$
$$+ (n-1)(m-1)M(a,b,c,d,n+m)]^2$$

2.6 Theoretical assessment – expectations when the variants appear at least *r* times in case group and at least twice in control group

I next consider an extension in which the SNP $i, i = 1,\mathrm{K}, I$, is said to appear in the smaller group when at least *r* individuals in that group have the minority allele, and is said to appear in the larger group when at least two individuals in that group have the minority allele. That is, the specification is (*r*, 2). Group 1 is a random sample of *n* individuals; group 2 is a random sample of *m* individuals; and $n \le m$. Then,

$$\Pr(i \text{ not in group } 1)$$
$$= (g_i^{\,n}, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1} g_i^{n-r+1}) U \left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1} \binom{n}{r-1} \right)^T.$$

$$\Pr(i \text{ in group } 1) = 1 - (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \text{K}, (-1)^{r-1} g_i^{n-r+1}) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}, (-1)^{r-1}\binom{n}{r-1}\right)^T .$$

$$\Pr(i \text{ not in group } 2) = m g_i^{m-1} - (m-1) g_i^m .$$

$$\Pr(i \text{ in group } 2) = 1 - m g_i^{m-1} + (m-1) g_i^m .$$

I continue to focus on the mean and variance of N_{+-} for selected specifications.

$$\begin{aligned}
&\Pr(i \text{ in group } 1 \text{ but not in group } 2) \\
&= (m g_i^{m-1} - (m-1) g_i^m) \\
&\cdot \left[1 - (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \text{K}, (-1)^{r-1} g_i^{n-r+1}) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}, (-1)^{r-1}\binom{n}{r-1}\right)^T \right] \\
&= m g_i^{m-1} - (m-1) g_i^m \\
&- m(g_i{}^{n+m-1}, -g_i^{n+m-2}, g_i^{n+m-3}, \text{K}, (-1)^{r-1} g_i^{n+m-r}) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}, (-1)^{r-1}\binom{n}{r-1}\right)^T \\
&+ (m-1) g_i{}^{n+m}, -g_i^{n+m-1}, g_i^{n+m-2}, \text{K}, (-1)^{r-1} g_i^{n+m-r+1} U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}, (-1)^{r-1}\binom{n}{r-1}\right)^T
\end{aligned}$$

The expectation and variance of the number of variants appearing in group 1 but not in group 2 are

$$\begin{aligned}
&EN_{+-}^{(r,2)} = \\
&= I[mM(a,b,c,d,m-1) - (m-1)M(a,b,c,d,m) - m(M(a,b,c,d,n+m-1), \\
&- M(a,b,c,d,n+m-2), M(a,b,c,d,n+m-3), \text{K}, (-1)^{r-1} M(a,b,c,d,n+m-r)) \\
&* U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}, (-1)^{r-1}\binom{n}{r-1}\right)^T \\
&+ (m-1)(M(a,b,c,d,n+m), -M(a,b,c,d,n+m-1), M(a,b,c,d,n+m-2), \text{K}, \\
&(-1)^{r-1} M(a,b,c,d,n+m-r+1)) U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \text{K}, (-1)^{r-1}\binom{n}{r-1}\right)^T
\end{aligned}$$

$$
\begin{aligned}
&VAR(N_{+-}^{(r,2)}) \\
&= I[mM(a,b,c,d,m-1) - (m-1)M(a,b,c,d,m) - m(M(a,b,c,d,n+m-1), \\
&- M(a,b,c,d,n+m-2), M(a,b,c,d,n+m-3), \mathrm{K}\ , (-1)^{r-1} M(a,b,c,d,n+m-r)) \\
&* U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T \\
&+ (m-1)(M(a,b,c,d,n+m), - M(a,b,c,d,n+m-1), M(a,b,c,d,n+m-2), \mathrm{K}\ , \\
&(-1)^{r-1} M(a,b,c,d,n+m-r+1))U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T \Bigg] \\
&- I[mM(a,b,c,d,m-1) - (m-1)M(a,b,c,d,m) - m(M(a,b,c,d,n+m-1), \\
&- M(a,b,c,d,n+m-2), M(a,b,c,d,n+m-3), \mathrm{K}\ , (-1)^{r-1} M(a,b,c,d,n+m-r)) \\
&* U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T \\
&+ (m-1)(M(a,b,c,d,n+m), - M(a,b,c,d,n+m-1), M(a,b,c,d,n+m-2), \mathrm{K}\ , \\
&(-1)^{r-1} M(a,b,c,d,n+m-r+1))U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}\ , (-1)^{r-1}\binom{n}{r-1}\right)^T \Bigg]^2
\end{aligned}
$$

2.7 Theoretical assessment – expectations when the variants appear at least *r* times in case group and at least *h* times in control group

My definition of the specification (*r*,*h*) is that SNP *i* is present in the smaller group of *n* subjects when at least *r* have the minority allele of SNP *i* and that SNP *i* is present in the larger group (group 2) of *m* subjects (that is, $n \le m$) when at least *h* have the minority allele of SNP *i.* Then,

$$\Pr(i \text{ not in group } 1) = g_i{}^n + \binom{n}{1}(1-g_i)g_i^{n-1} + \mathrm{K} + \binom{n}{r}(1-g_i)^{r-1}g_i^{n-r+1}$$

$$= (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1}g_i^{n-r+1})U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1}\binom{n}{r-1}\right)$$

$$\Pr(i \text{ not in group } 2)$$

$$= (g_i{}^m, -g_i^{m-1}, g_i^{m-2}, \mathrm{K}, (-1)^{h-1}g_i^{m-h+1})U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{h-1}\binom{n}{h-1}\right).$$

$$\Pr(i \text{ in group } 1)$$

$$= 1 - (g_i{}^n, -g_i^{n-1}, g_i^{n-2}, \mathrm{K}, (-1)^{r-1}g_i^{n-r+1})U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{r-1}\binom{n}{r-1}\right).$$

$$\Pr(i \text{ in group } 2)$$

$$= 1 - (g_i{}^m, -g_i^{m-1}, g_i^{m-2}, \mathrm{K}, (-1)^{h-1}g_i^{m-h+1})U\left(\binom{n}{0}, -\binom{n}{1}, \binom{n}{2}, \mathrm{K}, (-1)^{h-1}\binom{n}{h-1}\right).$$

Then $\Pr(i \text{ in both groups})$, $\Pr(i \text{ in group 2 but not in group } 1)$, $\Pr(i \text{ in group } 1 \text{ but not in group 2})$, $\Pr(i \text{ not in either group})$ are each polynomial functions of g_i. I can then use $Eg_i^n = M(a,b,c,d,n)$ to get the expectations and variances of the number of variants appearing in both groups, in group 1 but not group 2, in group 2 but not in group 1, or in neither group. The analysis is similar to that in previous Sections.

I report the mean and variance of N_{+-} for selected specifications. When $I = 10{,}000{,}000$, $n = 100$, $m = 1500$, $a = 0.14$, $b = 0.73$, $c = 0$, $d = 1$ (that is, the European population with 100 cases and 1500 controls assuming a genome wide search), Table 4 gives the expectation and standard deviation of $N_{+-}^{(r,h)}$. The table shows that the specification (4,1), which means that a variant is considered present in the case group of 100 when it appears at least four times and present in the control group of 1500 when

it appears at least once, does not generate an overwhelming number of false positives. Specifically, the expected value of the number of false positives $N_{+-}^{(4,1)}$ is 4 with standard deviation 2.

Table 4

The expectation and standard deviation of $N_{+-}^{(r,h)}$

*r**h*	1	2
1	28,206 (168)	59,362 (243)
2	1,004 (32)	3,064 (55)
3	44 (7)	287 (17)
4	4 (2)	10 (3)

Table 5 supplements the evaluation of the various specifications. It contains the one-sided p-values of Fisher's exact test for selected specifications. A study with 100 cases and 1500 controls has Fisher's exact test p-value less than 0.05 for the specification (*r*,1), $r \geq 2$, for any variant appearing *r* times in the cases and not appearing in the controls. Analogously, a study with 400 cases and 1200 controls has Fisher's exact test p-value less than 0.05 for (*r*,1), $r \geq 3$, A study with 400 cases and 400 controls and the study with 800 cases and 800 controls has Fisher's exact test p-value less then 0.05 for (*r*,1), $r \geq 5$.

Table 5

P-values for one sided Fisher's exact test under selected specifications (*r, h*)

Fisher's exact test p-value	400 vs 400	400 vs 1,200	800 vs 800	100 vs 1,500
Once in cases and never in controls; specification (1,1)	0.5000	0.2500	0.5000	0.0625
Twice in cases and never in controls; specification (2,1)	0.2497	0.0623	0.2498	0.0039
3 times in cases and never in controls; specification (3,1)	0.1245	0.0155	0.1248	0.0002
4 times in cases and never in controls; specification (4,1)	0.0620	0.0039	0.06227	0.0001
5 times in cases and never in controls; specification (5,1)	0.0309	0.0010	0.0311	0.0000

.

2.8 Truncation

When using the FHS data, the proportion of variants not appearing in either group is much larger than expected. Since most SNPs in FHS data are common, rare variants are precluded from appearing in the data. Consequently, I consider a model in which the allele frequency followed a beta distribution restricted to be greater than 0.001 in the

empirical assessment using the formulas in section 2.2 - 2.7 above. The following table compares the actual and expected numbers of variants in neither group using 488,146 SNPs on Chromosome 1-22 from FHS data. A left truncation point equal to 0.001 gives expected results that are in the range of the observed results. This left truncation point is used for the empirical comparison of expectations from the beta distribution to FHS data. The choice of truncation point while using this approach for other datasets should be such that the correspondence between calculated expected value and observed value be as close as possible.

Table 6

Comparison of actual and expected numbers of variants in not appearing in either group (N_{--}) under specification (1,1)

Total Number of cases and controls	800	1,600
Average	11,857.3	5,671
Low	9,775	5,671
High	14,439	5,671
Expectation(σ) when min frequency>0.005 (ENCODE)	208 (14)	2 (1)
Expectation(σ) when min frequency>0.005 (NIEHS SNPs)	251 (16)	2 (2)
Expectation(σ) when min frequency>0.001(ENCODE)	12,515 (110)	3,393 (58)
Expectation(σ) when min frequency>0.001 (NIEHS SNPs)	15,689 (123)	4,302 (65)
Expectation(σ) when min frequency>0.0005 (ENCODE)	25,205 (155)	10,815 (103)
Expectation(σ) when min frequency>0.0005 (NIEHS SNPs)	32,065 (173)	13,954 (116)
Expectation(σ) (ENCODE)	167,186 (332)	151,728(323)
Expectation(σ) (NIEHS SNPs)	270,825(347)	256,930(349)

3. Results

3.1 Results when the variants appear at least once in each group (under the specification (1,1))

3.1.1 Theoretical expected values assuming 10,000,000 variants in the population

3.1.1.1 In the European population

Table 7 contains the expected values of N_{++}, N_{+-}, N_{-+}, and N_{--} and their standard deviations using the beta distribution(0.14, 0.73) and the beta distribution (0.076, 0.72). These are the estimates in Ionnita-Laza et al. (2009) for the European population using the Encode database and the NIEHS SNP database respectively. I focus on E(N_{+-}), the number of variants observed in the smaller group but not in the larger group. These values are very large, suggesting that a genome wide study using rare variants would report an extremely large number of variants in the smaller group that did not appear in the larger (control) group when both samples were from exactly the same population. For example, the expected number in a random group of 100 that did not appear in a random group of 1500 is on the order of 28,000. In a comparison of 400 with 1200, the expected value increases to about 120,000. Finally, in a comparison of 800 vs. 800, the expected number is over 300,000.

Table 7

Number of variants in the genome expected to appear in two samples among European population under specification (1,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	5,877,447 (1,557)	6,098,551 (1,543)	5,391,187 (1,576)	6,258,426 (1,530)
	Expectation(σ) NIEHS SNPs	3,852,054 (1,539)	4,035,688 (1,551)	3,477,329 (1,506)	4,167,333 (1,559)
N_{+-}	Expectation(σ) ENCODE	348,838 (580)	127,714 (355)	28,206 (168)	316,658 (554)
	Expectation(σ) NIEHS SNPs	299,958 (539)	116,323 (339)	25,876 (161)	284,637 (526)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	665,476 (788)	1,472,348 (1,121)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		584,594 (742)	1,233,401 (1,040)	
N_{--}	Expectation(σ) ENCODE	3,424,917 (1,501)	3,108,259 (1,464)	3,108,259 (1,464)	3,108,259 (1,464)
	Expectation(σ) NIEHS SNPs	5,548,031 (1,572)	5,263,394 (1,579)	5,263,394 (1,579)	5,263,396 (1,579)

Table 7

(Continued)

	European	100 vs 1,500	100 vs 15,000	100 vs 150,000	100 vs 1,500,000
N_{++}	Expectation(σ) ENCODE	5,391,187 (1,576)	5,417,280 (1,576)	5,419,239 (1,576)	5,419,382 (1,576)
	Expectation(σ) NIEHS SNPs	3,477,329 (1,506)	3,500,963 (1,508)	3,503,016 (1,509)	3.503,189 (1,509)
N_{+-}	Expectation(σ) ENCODE	28,206 (168)	2,113 (46)	153 (12)	11 (1)
	Expectation(σ) NIEHS SNPs	25,876 (161)	2,242 (47)	189 (14)	16 (4)
N_{-+}	Expectation(σ) ENCODE	1,472,348 (1,121)	2,310,494 (1,333)	2,934,675 (1,440)	3,388,134 (1,497)
	Expectation(σ) NIEHS SNPs	1,233,401 (1,040)	2,058,854 (1,279)	2,769,624 (1,415)	3,367,841 (1,495)
N_{--}	Expectation(σ) ENCODE	3,108,259 (1,464)	2,270,113 (1,325)	1,645,932 (1,173)	1,192,473 (1,025)
	Expectation(σ) NIEHS SNPs	5,263,394 (1,579)	4,437,941 (1,571)	3,727,171 (1,529)	3,128,954 (1,466)

Table 7

(Continued)

	European	100 vs 1,500	1,000 vs 15,000	10,000 vs 150,000	100 ,000vs 1,500,000

	Expectation(σ) ENCODE	5,391,187 (1,576)	6,659,952 (1,491)	7,580,258 (1,354)	8,247,045 (1,202)
N_{++}	Expectation(σ) NIEHS SNPs	3,477,329 (1,506)	4,523,512 (1,574)	5,402,627 (1,576)	6,140,682 (1,539)
	Expectation(σ) ENCODE	28,206 (168)	20,438 (143)	14,806 (122)	10,726 (104)
N_{+-}	Expectation(σ) NIEHS SNPs	25,876 (161)	21,725 (147)	18,237 (135)	15,310 (124)
	Expectation(σ) ENCODE	1,472,348 (1,121)	1,067,823 (977)	773,656 (845)	560,471 (727)
N_{-+}	Expectation(σ) NIEHS SNPs	1,233,401 (1,040)	1,036,305 (964)	870,013 (891)	730,348 (823)
	Expectation(σ) ENCODE	3,108,259 (1,464)	2,251,788 (1,321)	1,631,280 (1,168)	1,181,758 (1,021)
N_{--}	Expectation(σ) NIEHS SNPs	5,263,394 (1,579)	4,418,458 (1,570)	3,709,123 (1,528)	3,113,660 (1,464)

3.1.1.2 Other populations

Table 8 presents the expected numbers for the African, Chinese and Japanese populations. The expected number of variants that appear in the smaller group compared to the larger group are of the same order of magnitude as in the European population.

Table 8

Number of variants in the genome expected to appear in two samples among African, Chinese, and Japanese population under specification (1,1)

	African	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	3,384,183 (1,496)	3,566,520 (1,515)	3,105,264 (1,451)	3,697,116 (1,527)
N_{++}	Expectation(σ) NIEHS SNPs	1,876,584 (1,235)	1,993,283 (1,263)	1,649,819 (1,174)	2,076,478 (1,283)
N_{+-}	Expectation(σ) ENCODE	298,915 (538)	116,579 (339)	25,950 (161)	284,897 (526)

	Expectation(σ) NIEHS SNPs	195,187 (437)	78,488 (279)	17,538 (132)	190,479 (432)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	583,813 (741)	1,225,696 (1,037)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		385,665 (609)	790,079 (853)	
N_{--}	Expectation(σ) ENCODE	6,017,986 (1,548)	5,733,089 (1,564)	5,733,089 (1,564)	5,733,089 (1,564)
	Expectation(σ) NIEHS SNPs	7,733,043 (1,324)	7,542,563 (1,361)	7,542,563 (1,361)	7,542,563 (1,361)

Table 8
(Continued)

	Chinese	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	7,439,376 (1,380)	7,648,449 (1,341)	6,934,968 (1,458)	7,801,221 (1,310)
	Expectation(σ) NIEHS SNPs	3,207,628 (1,476)	3,363,617 (1,494)	2,895,801 ((1,434)	3,475,173 (1,506)
N_{+-}	Expectation(σ) ENCODE	317,109 (554)	108,037 (327)	23,646 (154)	272,374 (515)
	Expectation(σ) NIEHS SNPs	257,438 (501)	101,449 (317)	22,612 (150)	247,330 (491)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	589,483 (745)	1,387,624 (1,093)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		504,768 (692)	1,051,421 (970)	
N_{--}	Expectation(σ) ENCODE	1,926,405 (1,247)	1,654,031 (1,175)	1,654,031 (1,795)	1,654,031 (1,175)
	Expectation(σ) NIEHS SNPs	6,277,497 (1,529)	6,030,166 (1,547)	6,030,166 (1,547)	6,030,166 (1,547)

Table 8
(Continued)

	Japanese	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	8,761,759 (1,042)	8,914,626 (984)	8,332,725 (1,179)	9,028,195 (937)
	Expectation(σ) NIEHS SNPs	3,511,574 (1,509)	3,675,613 (1,525)	3,181,277 (1,473)	3,793,023 (1,534)
N_{+-}	Expectation(σ) ENCODE	219,308 (463)	66,440 (257)	14,329 (120)	172,179 (411)
	Expectation(σ) NIEHS SNPs	269,776 (512)	105,737 (323)	23,552 (153)	258,103 (501)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	391,486 (613)	1,025,498 (959)	Same as N_{+-}
	Expectation(σ)		527,878	1,104,399	

	NIEHS SNPs		(707)	(991)	
N_{--}	Expectation(σ) ENCODE	799,626 (858)	627,447 (767)	627,447 (767)	627,447 (767)
	Expectation(σ) NIEHS SNPs	5,948,874 (1,552)	5,690,772 (1,566)	5,690,772 (1,566)	5,690,772 (1,566)

3.1.2 Theoretical expected values assuming 150,000 variants in the exome

I now report the expected numbers for a study considering only exonic regions, assuming 150,000 variants there. As shown in Table 9, In a comparison of a random sample of 100 to a random sample of 1500 from the same population, the number of variants appearing in the smaller group but not the larger is around 400 (with a standard deviation of about 20). The values increase substantially with larger numbers in each group. These expected values are also so large as to suggest that this specification is not practical in the sense that there would be a large number of traits identified simply by random chance.

Table 9

Number of variants in the exome expected to appear in two samples among European population under specification (1,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	88,162 (191)	91,478 (189)	80,868 (193)	93,876 (187)
	Expectation(σ) NIEHS SNPs	57,781 (188)	60,535 (190)	52,160 (184)	62,510 (191)
N_{+-}	Expectation(σ) ENCODE	5,232 (71)	1,916 (43)	423 (21)	4,750 (68)
	Expectation(σ) NIEHS SNPs	4,499 (66)	1,745 (42)	388 (20)	4,270 (64)
N_{-+}	Expectation(σ)	Same as	9,982 (97)	22,085 (137)	Same as

	ENCODE	N_{+-}			N_{+-}
	Expectation(σ) NIEHS SNPs		8,769 (91)	18,501 (127)	
	Expectation(σ) ENCODE	51,374 (184)	46,624 (179)	46,624 (179)	46,634 (179)
N_{--}	Expectation(σ) NIEHS SNPs	83,220 (192)	78,951 (193)	78,951 (193)	78,951 (193)

Table 9
(Continued)

	African	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	50,763 (183)	53,498 (186)	45,229 (178)	55,467 (187)
N_{++}	Expectation(σ) NIEHS SNPs	28,149 (151)	29,899 (155)	24,747 (144)	31,147 (157)
	Expectation(σ) ENCODE	4,484 (66)	1,749 (42)	389 (20)	4,273 (64)
N_{+-}	Expectation(σ) NIEHS SNPs	2928 (54)	1,177 (34)	263 (16)	2,857 (53)
	Expectation(σ) ENCODE		8,757 (91)	18,385 (127)	
N_{-+}	Expectation(σ) NIEHS SNPs	Same as N_{+-}	5,785 (75)	11,851 (104)	Same as N_{+-}
	Expectation(σ) ENCODE	90,270 (190)	85,996 (192)	85,996 (192)	85,996 (192)
N_{--}	Expectation(σ) NIEHS SNPs	115,996 (162)	113,138 (167)	113,138 (167)	113,138 (167)

Table 9
(Continued)

	Chinese	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	111,591 (169)	114,727 (164)	104,021 (179)	117,018 (160)
N_{++}	Expectation(σ) NIEHS SNPs	48,114 (181)	50,454 (183)	43,437 (176)	52,128 (184)
	Expectation(σ) ENCODE	4,757 (68)	1,621 (40)	355 (19)	4,085 (63)
N_{+-}	Expectation(σ) NIEHS SNPs	3,862 (61)	1,522 (39)	339 (18)	3,710 (60)
	Expectation(σ) ENCODE		8,842 (91)	20,814 (134)	
N_{-+}	Expectation(σ) NIEHS SNPs	Same as N_{+-}	7,572 (85)	15,771 (119)	Same as N_{+-}
N_{--}	Expectation(σ) ENCODE	28,896 (153)	24,810 (144)	24,810 (144)	24,810 (144)

	Expectation(σ) NIEHS SNPs	94,162 (187)	90,452 (189)	90,452 (189)	90,452 (189)

Table 9
(Continued)

	Japanese	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	131,426 (128)	133,719 (120)	124,991 (144)	135,422 (115)
N_{++}	Expectation(σ) NIEHS SNPs	52,674 (185)	55,134 (187)	47,719	56,895 (188)
	Expectation(σ) ENCODE	3,290 (57)	996 (31)	215 (15)	2,583 (50)
N_{+-}	Expectation(σ) NIEHS SNPs	4,047 (63)	1,586 (87)	353 (19)	3,872 (61)
	Expectation(σ) ENCODE	Same as N_{+-}	5,872 (75)	15,382 (117)	Same as N_{+-}
N_{-+}	Expectation(σ) NIEHS SNPs		7,918 (187)	16,566 (121)	
	Expectation(σ) ENCODE	11,994 (105)	9,411 (94)	9,411 (94)	9,411 (94)
N_{--}	Expectation(σ) NIEHS SNPs	89,233 (190	85,362 (192)	85,362 (192)	85,362 (192)

3.1.3 Theoretical expected values in specific genes

I next consider whether this strategy could be effective at the gene level. According to dbSNP Build 130 and CCDS (Consensus CoDing Sequence) database, gene FBN1 contains 1,301 SNPs in total and 65 SNPs in exonic regions. Table 10 contains the expected numbers for this gene for the European population. The expected number appearing in a randomly selected group of 100 that did not appear in a randomly selected group of 1500 is about 4 with a standard deviation of 2. The expected number increases to about 40 when comparing two randomly selected groups of 800. These

numbers are relatively practical. Table 11 contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts range from 0.2 to 2.

Table 10

Number of variants in gene FBN1 expected to appear in two samples among European population under specification (1,1)

	FBN1: 1,301 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	765 (18)	793 (18)	701 (18)	814 (17)
	Expectation(σ) NIEHS SNPs	501 (18)	525 (18)	452 (17)	542 (18)
N_{+-}	Expectation(σ) ENCODE	45 (7)	17 (4)	4 (2)	41 (6)
	Expectation(σ) NIEHS SNPs	39 (6)	15 (4)	3 (2)	37 (6)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	87 (9)	192 (13)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		76 (8)	160 (12)	
N_{--}	Expectation(σ) ENCODE	446 (17)	404 (17)	404 (17)	404 (17)
	Expectation(σ) NIEHS SNPs	722 (18)	685 (18)	685 (18)	685 (18)

Table 11

Number of variants in gene FBN1 in the exome expected to appear in two samples among European population under specification (1,1)

	FBN1 (exon): 65 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	38 (4)	40 (4)	35 (4)	41 (4)
	Expectation(σ) NIEHS SNPs	25 (4)	26 (4)	23 (4)	27 (4)
N_{+-}	Expectation(σ) ENCODE	2 (1)	0.8 (0.9)	0.2 (0.4)	2 (1)
	Expectation(σ) NIEHS SNPs	2 (1)	0.8 (0.8)	0.2 (0.4)	2 (1)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	4 (2)	10 (3)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		4 (2)	8 (3)	

N_{--}	Expectation(σ) ENCODE	22 (4)	20 (4)	20 (4)	20 (4)
	Expectation(σ) NIEHS SNPs	36 (4)	34 (4)	34 (4)	34 (4)

According to dbSNP Build 130 and CCDS database, gene NF1 contains 1,659 SNPs in total and 57 SNPs in exonic regions. Table 12 contains the expected numbers for this gene for the European population. The expected number appearing in a randomly selected group of 100 that did not appear in a randomly selected group of 1500 is about 5 with a standard deviation of 2. The expected number increases to about 50 when comparing two randomly selected groups of 800. Table 13 contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts range from 0.1 to 2.

Table 12

Number of variants in gene NF1 expected to appear in two samples among European population under specification (1,1)

	NF1: 1,659 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	975 (20)	1,012 (20)	894 (20)	1,038 (20)
	Expectation(σ) NIEHS SNPs	639 (20)	670 (20)	577 (19)	691 (20)
N_{+-}	Expectation(σ) ENCODE	58 (7)	21 (5)	5 (2)	53 (7)
	Expectation(σ) NIEHS SNPs	50 (7)	19 (4)	4 (2)	47 (7)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	110 (10)	244 (14)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		97 (10)	205 (13)	
N_{--}	Expectation(σ) ENCODE	568 (19)	516 (19)	516 (19)	516 (19)
	Expectation(σ) NIEHS SNPs	920 (20)	873 (20)	873 (20)	873 (20)

Table 13

Number of variants in gene NF1 in the exome expected to appear in two samples among European population under specification (1,1)

	NF1 (exome): 57 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	34 (4)	35 (4)	31 (4)	36 (4)
	Expectation(σ) NIEHS SNPs	22 (4)	23 (4)	20 (4)	24 (4)
N_{+-}	Expectation(σ) ENCODE	2 (1)	0.7 (0.8)	0.2 (0.4)	2 (1)
	Expectation(σ) NIEHS SNPs	2 (1)	0.7 (0.8)	0.1 (0.4)	2 (1)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	4 (2)	8 (3)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		3 (4)	7 (2)	
N_{--}	Expectation(σ) ENCODE	20 (4)	18 (3)	18 (3)	18 (3)
	Expectation(σ) NIEHS SNPs	32 (4)	30 (4)	30 (4)	30 (4)

Appendix 1 contains the results for numbers of variants in gene SYNE1, HMCN1, UBR4, RYR1. The results are similar to the tables above.

3.1.4 Empirical assessment of Chr1-22/500k SNP data using permutation modeling

3.1.4.1 Empirical assessment in the genome

I calculate expectations using the parameters for a European population (Ionnita-Laza et al. 2009) assuming a minimum variant frequency of 0.001 and without truncation. I examine 488,146 SNPs on Chromosome 1-22 with results as shown in Table 14. The observed average number of SNPs that appeared in a random sample of

100 but not in a random sample of 1500 is 559. This is relatively close to expected count of about 600 assuming truncation and 1300 assuming no truncation. The standard deviation of N_{+-} is modeled to be about 25, which is much smaller than the sample standard deviation of 336 observed in 10 random comparisons.

The counts increase as the minimum sample size increases. For example, the number in one group but not the other ranges from 4000 to 9800 in two random samples of 800.

Table 14

Categorization of SNPs in the genome by appearance in two randomly selected groups under specification (1,1)

		400 vs 400[1]	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Average($\hat{\sigma}$)	457,996 (1,847)	464,739.5 (1,424)	445,674.5 (4,143)	469,263.3 (306)
	Low	455,448	462,157	440,859	468,722
	High	461,037	466,915	450,837	469,542
	Expectation(σ) ENCODE* [2]	442,328 (204)	456,085 (173)	406,093 (261)	466,509 (144)
	Expectation(σ) SIEHS SNPs*	431,724 (223)	448,456 (191)	390,162 (280)	461,069 (160)
	Expectation(σ) ENCODE	286,905 (344)	297,698(341)	263,169(348)	305,503 (338)
	Expectation(σ) SIEHS SNPs	188,036 (340)	197,001(343)	169,744(333)	203,427 (344)
N_{+-}	Average($\hat{\sigma}$)	9,146.4 (1,469)	2,394.3 (515)	558.6 (336)	6605.9 (1,596)
	Low	6,710	1,602	211	3,996
	High	12,386	3,131	999	9,757
	Expectation(σ) ENCODE*	16,652 (126)	2,894 (53)	547 (23)	9,122 (95)
	Expectation(σ) SIEHS SNPs*	20,366 (140)	3,635 (60)	689 (26)	11,387 (105)
	Expectation(σ) ENCODE	17,027 (128)	6,234(78)	1,377(37)	15,458 (122)

	Expectation(σ) SIEHS SNPs	14,642 (119)	5,678(75)	1,263(35)	13,894 (116)
N_{-+}	Average($\hat{\sigma}$)	Same as N_{+-}	15,341.2 (1,928)	36,241.9 (4,475)	Same as N_{+-}
	Low		12,429	30,677	
	High		18,716	41,387	
	Expectation(σ) ENCODE*		25,773 (156)	78,112 (256)	
	Expectation(σ) SIEHS SNPs*		31,754 (172)	92,992 (274)	
	Expectation(σ) ENCODE		32,485 (174)	71,872(248)	
	Expectation(σ) SIEHS SNPs		28,536(164)	60,208(230)	
N_{--}	Average($\hat{\sigma}$)	11,857.3 (1,508)	5,671[3] (0)	5,671 (0)	5,671 (0)
	Low	9,775	5,671	5,671	5,671
	High	14,439	5,671	5,671	5,671
	Expectation(σ) ENCODE*	12,515 (110)	3,393 (58)	3,393 (58)	3,393 (58)
	Expectation(σ) SIEHS SNPs*	15,689 (123)	4,302 (65)	4,302 (65)	4,302 (65)
	Expectation(σ) ENCODE	167,186 (332)	151,728 (323)	151,728 (323)	151,728 (323)
	Expectation(σ) SIEHS SNPs	270,825(347)	256,930 (349)	256,930 (349)	256,930 (349)

[1] Some groups have one fewer individual than reported. For example, in the 400 vs. 400 comparisons, one group has 399 randomly selected individuals.

[2] An expectation marked with "*" denotes a result assuming a truncated beta distribution with minimum variant frequency=0.001. Otherwise, the expectation calculation assumes a beta distribution (that is, one that ranges from 0 to 1).

[3]The observed numbers of N_{--} is constant for three comparisons (400 vs 1200, 100 vs 1500, and 800 vs 800) since the dataset contains only about 1,600 individuals, that is, the number of SNPs that do not appear in the dataset is constant.

3.1.4.2 Empirical assessment in the exome

I calculate expectations using the parameters for an European population (Ionnita-Laza et al. 2009) assuming a minimum frequency of 0.001 and a frequency without truncation. I examine 3,342 SNPs in exonic regions on Chromosome 1-22 with results shown in Table 15. The observed average number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 5. This compares to an expected count of about 4 assuming truncation and 9 assuming no truncation. The modeled standard deviation is 2, while the sample standard deviation is 3 and the range is 10. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 27 to 82 in two random samples of 800.

Table 15

Categorization of SNPs in the exome by appearance in two randomly selected groups under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Average($\hat{\sigma}$)	3,097.6 (19)	3,151.7 (13)	2,986.7 (34)	3,195 (6)
	Low	3,075	3,129	2,944	3,187
	High	3,125	3,169	3,041	3,206
	Expectation(σ) ENCODE*	3,028 (17)	3,123 (14)	2,780 (22)	3,194 (12)
	Expectation(σ) SIEHS SNPs*	2956 (18)	3,070 (16)	2,671 (23)	3,157 (13)
	Expectation(σ) ENCODE	1,964(28)	2,038(28)	1,802(29)	2,092(28)
	Expectation(σ) SIEHS SNPs	1,287(28)	1,349 (28)	1,162(28)	1,393(29)
N_{+-}	Average($\hat{\sigma}$)	74.7 (14)	17.2 (7)	4.6 (3)	51.5 (16)
	Low	51	6	0	27
	High	111	26	10	82
	Expectation(σ) ENCODE*	114 (10)	20 (4)	4 (2)	62 (8)
	Expectation(σ) SIEHS SNPs*	139 (12)	25 (5)	5 (2)	78 (9)
	Expectation(σ)	117(11)	43(6)	9(3)	106(10)

	ENCODE				
	Expectation(σ) SIEHS SNPs	100(10)	39(6)	9(3)	95(10)
N_{-+}	Average($\hat{\sigma}$)	Same as N_{+-}	129.1 (19)	306.7 (37)	Same as N_{+-}
	Low		103	247	
	High		161	352	
	Expectation(σ) ENCODE*		176 (13)	535 (21)	
	Expectation(σ) SIEHS SNPs*		217 (14)	637 (23)	
	Expectation(σ) ENCODE		222(14)	492(20)	
	Expectation(σ) SIEHS SNPs		195(14)	412(19)	
N_{--}	Average($\hat{\sigma}$)	95 (17)	44 (0)	44 (0)	44 (0)
	Low	75	44	44	44
	High	129	44	44	44
	Expectation(σ) ENCODE*	86 (9)	23 (5)	23 (5)	23(5)
	Expectation(σ) SIEHS SNPs*	107 (10)	29 (5)	29 (5)	29 (5)
	Expectation(σ) ENCODE	1,145(27)	1,039(27)	1,039(27)	1,039(27)
	Expectation(σ) SIEHS SNPs	1,854(29)	1,759(29)	1,759(29)	1,759(29)

3.1.4.3 Empirical assessment in specific genes

I examine 73 SNPs in gene FBN1 on Chromosome 15 with results shown in Table 16. The observed average number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0.6. This compares to an expected count of about 0.1 assuming truncation and 0.2 assuming no truncation. The counts increase as the

minimum sample size increases. For example, the average number in one group but not the other is 3 in two random samples of 800.

Table 16
Categorization of SNPs in gene FBN1with 73 SNPs by appearance in two randomly selected groups under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Average($\hat{\sigma}$)	56.1 (3)	59.5 (4)	51.8 (2)	63.5 (3)
	Low	52	53	49	58
	High	64	64	54	67
	Expectation(σ) ENCODE*	66 (2)	68 (2)	61 (3)	70 (2)
	Expectation(σ) SIEHS SNPs*	65 (3)	67 (2)	58 (3)	69 (2)
	Expectation(σ) ENCODE	43 (4)	45 (4)	39 (4)	46 (4)
	Expectation(σ) SIEHS SNPs	28 (4)	29 (4)	25 (4)	30 (4)
N_{+-}	Average($\hat{\sigma}$)	4.9 (3)	1.2 (2)	0.6 (1)	3.3 (3)
	Low	0	0	0	0
	High	10	4	3	12
	Expectation(σ) ENCODE*	2 (2)	0.4 (0.7)	0.1 (0.3)	1 (1)
	Expectation(σ) SIEHS SNPs*	3 (2)	0.5 (0.7)	0.1 (0.3)	2 (1)
	Expectation(σ) ENCODE	3 (2)	0.9 (1)	0.2 (0.5)	2 (1)
	Expectation(σ) SIEHS SNPs	2 (1)	0.8 (0.9)	0.2 (0.4)	2 (1)
N_{-+}	Average($\hat{\sigma}$)	Same as N_{+-}	9.3 (4)	17.6 (3)	Same as N_{+-}
	Low		3	13	
	High		17	21	
	Expectation(σ) ENCODE*		4 (2)	12 (3)	
	Expectation(σ) SIEHS SNPs*		5 (2)	14 (3)	
	Expectation(σ) ENCODE		5 (2)	11 (3)	
	Expectation(σ) SIEHS SNPs		4 (2)	9 (3)	
N_{--}	Average($\hat{\sigma}$)	7.2 (5)	3 (0)	3 (0)	3 (0)
	Low	3	3	3	3

	High	15	3	3	3
	Expectation(σ) ENCODE*	2 (1)	0.5 (0.7)	0.5 (0.7)	0.5 (0.7)
	Expectation(σ) SIEHS SNPs*	2 (2)	0.6 (0.8)	0.6 (0.8)	0.6 (0.8)
	Expectation(σ) ENCODE	25 (4)	23 (4)	23 (4)	23 (4)
	Expectation(σ) SIEHS SNPs	41 (4)	38 (4)	38 (4)	38 (4)

I examine 16 SNPs in gene NF1 (which is on Chromosome 17) with results as shown in Table 17. The observed average number of SNPs that appeared in a random sample of 100 but not in a random sample of 1500 is 0.1. This compares to an expected count of about 0.02 assuming truncation and 0.04 assuming no truncation. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 0 to 1 in two random samples of 800.

Table 17

Categorization of SNPs in gene NF1 with 16 SNPs by appearance in two randomly selected groups under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Average($\hat{\sigma}$)	13.9 (0.3)	14 (0)	13.4 (0.5)	14 (0)
	Low	13	14	13	14
	High	14	14	14	14
	Expectation(σ) ENCODE*	15 (1)	15 (1)	13 (1)	15 (0.8)
	Expectation(σ) SIEHS SNPs*	14 (1)	15 (1)	13 (2)	15 (0.9)
	Expectation(σ) ENCODE	9 (2)	10 (2)	9 (2)	10 (2)
	Expectation(σ) SIEHS SNPs	6 (2)	6 (2)	6 (2)	7 (2)
N_{+-}	Average($\hat{\sigma}$)	0.3 (0.5)	0.2 (0.4)	0.1 (0.3)	0.5 (0.5)
	Low	0	0	0	0
	High	1	1	1	1
	Expectation(σ) ENCODE*	0.5 (0.7)	0.1 (0.3)	0.02 (0.1)	0.3 (0.5)

	Expectation(σ) SIEHS SNPs*	0.7 (0.8)	0.1 (0.3)	0.02 (0.2)	0.4 (0.6)
	Expectation(σ) ENCODE	0.6 (0.7)	0.2 (0.4)	0.05 (0.2)	0.5 (0.7)
	Expectation(σ) SIEHS SNPs	0.5 (0.7)	0.2 (0.4)	0.04 (0.2)	0.5 (0.7)
N_{-+}	Average($\hat{\sigma}$)	Same as N_{+-}	0.8 (0.4)	1.5 (0.5)	Same as N_{+-}
	Low		0	1	
	High		1	2	
	Expectation(σ) ENCODE*		0.8 (0.9)	3 (1)	
	Expectation(σ) SIEHS SNPs*		1 (1)	3 (2)	
	Expectation(σ) ENCODE		1 (1)	2 (1)	
	Expectation(σ) SIEHS SNPs		0.9 (0.9)	2 (1)	
N_{--}	Average($\hat{\sigma}$)	1.5 (0.5)	1	1	1
	Low	1	1	1	1
	High	2	1	1	1
	Expectation(σ) ENCODE*	0.4 (0.6)	0.1 (0.3)	0.1 (0.3)	0.1 (0.3)
	Expectation(σ) SIEHS SNPs*	0.5 (0.7)	0.1 (0.4)	0.1 (0.4)	0.1 (0.4)
	Expectation(σ) ENCODE	5 (2)	5 (2)	5 (2)	5 (2)
	Expectation(σ) SIEHS SNPs	9 (2)	8 (2)	8 (2)	8 (2)

Appendix 2 contains the results for numbers of variants in gene SYNE1, HMCN1, UBR4, RYR1. The results are comparable.

3.2 Results when the variants appear at least twice in case group and at least once in control group (under the specification (2,1))

3.2.1 Theoretical expected values assuming 10,000,000 variants in the population under specification (2,1)

Under specification (2,1), the SNP $i, i = 1,\mathrm{K}, I$, is said to appear in case group when at least two individuals in that group have the minority allele, and is said to appear in control group when at least one individual has the minority allele. I calculate the expected number of variants observed in the smaller group but not in the larger group, which I denote with $E(N_{+-}^{(2,1)})$, and I report these values in Table 18. These values are comparatively smaller than those under specification (1, 1). For example, the expected number in a random group of 100 that do not appear in a random group of 1500 decreased from the order of 28,000 to about 1,000. The expected numbers are still very large, suggesting that a genome wide study using rare variants would report an extremely large number of variants in the smaller group that do not appear in the larger (control) group. The expected number in a random group of 400 that do not appear in a random group of 1200 is on the order of 19,000. In a comparison of 800 with 800, the expected value increases to about 100,000.

Table 18

Number of variants in the genome expected to appear in two samples among European population under specification (2,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Expectation(σ) ENCODE	5,588,592 (1,570)	5,678,679 (1,567)	4,775,368 (1,580)	5,996,390 (1,549)
	Expectation(σ) SIEHS SNPs	3,618,195 (1,520)	3,690,952 (1,526)	3,007,192 (1,450)	3,945,579 (1,546)
$N_{+-}^{(2,1)}$	Expectation(σ) ENCODE	108,993 (328)	18,906 (137)	1,004 (32)	99,043 (313)
	Expectation(σ) SIEHS SNPs	89,059 (297)	16,301 (128)	870 (29)	84,593 (290)
$N_{-+}^{(2,1)}$	Expectation(σ)	637,673	1,085,348	2,088,167	578,693

	ENCODE	(773)	(984)	(1,285)	(738)
	Expectation(σ) SIEHS SNPs	533,817 (711)	929,331 (918)	1,703,538 (1,189)	506,391 (693)
$N_{--}^{(2,1)}$	Expectation(σ) ENCODE	3,664,742 (1,524)	3,217,066 (1,477)	3,135,461 (1,467)	3,325,874 (1,490)
	Expectation(σ) SIEHS SNPs	5,758,930 (1,563)	5,363,416 (1,577)	5,288,400 (1,578)	5,463,438 (1,574)

3.2.2 Theoretical expected values assuming 150,000 variants in the exome under specification (2,1)

In Table 19, I report the expected numbers for a study considering only the exome under the specification (2, 1), assuming 150,000 variants in the exome. In a comparison of a random sample of 100 to a random sample of 1500 from the same population, the number of variants appearing in the smaller group but not the larger is around 15 (with a standard deviation of about 4), decreasing from the 400 expected under the specification (1,1). The expected value of 15 in an exome wide study is relatively practical. The expected number increases to about 1,500 when comparing two randomly selected groups of 800. This is smaller than the expectation of around 4,800 under the specification (1,1). These expected counts for comparisons 400 vs 400, 400 vs 1200, 800 vs 800 are still too large to be practical.

Table 19

Number of variants in the exome expected to appear in two samples among European population under specification (2,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Expectation(σ)	83,829 (192)	85,180 (192)	71,631 (193)	89,946 (190)

	ENCODE				
	Expectation(σ) SIEHS SNPs	54,273 (186)	55,364 (187)	45,108 (178)	59,184 (189)
$N_{+-}^{(2,1)}$	Expectation(σ) ENCODE	1,635 (40)	284 (17)	15 (4)	1,486 (38)
	Expectation(σ) SIEHS SNPs	1,336 (36)	245 (16)	13 (4)	1,269 (35)
$N_{-+}^{(2,1)}$	Expectation(σ) ENCODE	9,565 (95)	16,280 (120)	31,323 (157)	8,680 (90)
	Expectation(σ) SIEHS SNPs	8,007 (87)	13,940 (112)	25,553 (146)	7,596 (85)
$N_{--}^{(2,1)}$	Expectation(σ) ENCODE	54,971 (187)	48,256 (181)	47,032 (180)	49,888 (182)
	Expectation(σ) SIEHS SNPs	86,384 (191)	80,451 (193)	79,326 (193)	81,952 (193)

3.2.3 Theoretical expected values in specific genes under specification (2,1)

I then consider the specification applied at the gene level. Table 20 contains the expected numbers in gene FBN1 for the European population. The expected number appearing in a randomly selected group of 100 that do not appear in a randomly selected group of 1500 is about 0.1 with a standard deviation of 0.4. The expected number increases to about 12 when comparing two randomly selected groups of 800. The expected number of false positives is small enough to be practical. Table 20 also contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts are close to 0. These expectations suggest that this specification would be practical at the gene level.

Table 20

Number of variants in gene FBN1 expected to appear in two samples among European population under specification (2,1)

	FBN1: 1,301 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Expectation(σ) ENCODE	727 (18)	739 (18)	621 (18)	780 (18)
	Expectation(σ) SIEHS SNPs	471 (17)	480 (17)	391 (17)	513 (18)
$N_{+-}^{(2,1)}$	Expectation(σ) ENCODE	14 (4)	2 (2)	0.1 (0.4)	13 (4)
	Expectation(σ) SIEHS SNPs	12 (3)	2 (1)	0.1 (0.3)	11 (3)
$N_{-+}^{(2,1)}$	Expectation(σ) ENCODE	83 (9)	141 (11)	272 (15)	75 (8)
	Expectation(σ) SIEHS SNPs	69 (8)	121 (10)	222 (14)	66 (8)
$N_{--}^{(2,1)}$	Expectation(σ) ENCODE	477 (17)	419 (17)	408 (17)	433 (17)
	Expectation(σ) SIEHS SNPs	749 (18)	698 (18)	688 (18)	712 (18)

Table 20
(Continued)

	FBN1 (exon): 65 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Expectation(σ) ENCODE	36 (4)	37 (4)	31 (4)	39 (4)
	Expectation(σ) SIEHS SNPs	24 (4)	24 (4)	20 (4)	26 (4)
$N_{+-}^{(2,1)}$	Expectation(σ) ENCODE	0.7 (0.8)	0.1 (0.4)	0 (0.1)	0.6 (0.8)
	Expectation(σ) SIEHS SNPs	0.6 (0.8)	0.1 (0.3)	0 (0.1)	0.5 (0.7)
$N_{-+}^{(2,1)}$	Expectation(σ) ENCODE	4 (2)	7 (3)	14 (3)	4 (2)
	Expectation(σ) SIEHS SNPs	3 (2)	6 (2)	11 (3)	3 (2)
$N_{--}^{(2,1)}$	Expectation(σ) ENCODE	24 (4)	21 (4)	20 (4)	22 (4)
	Expectation(σ) SIEHS SNPs	37 (4)	35 (4)	34 (4)	36 (4)

Table 21 contains the expected numbers for another gene, NF1, for the European population. The expected number appearing in a randomly selected group of 100 that

do not appear in a randomly selected group of 1500 is about 0.2 with a standard deviation of 0.4. The expected number increases to about 15 when comparing two randomly selected groups of 800. Table 21 also contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts are also close to 0.

Table 21

Number of variants in gene NF1 expected to appear in two samples among European population under specification (2,1)

	NF1: 1,659 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	927 (20)	942 (20)	792 (20)	995 (20)
$N_{++}^{(2,1)}$	Expectation(σ) SIEHS SNPs	600 (20)	612 (20)	499 (19)	655 (20)
	Expectation(σ) ENCODE	18 (4)	3 (2)	0.2 (0.4)	16 (4)
$N_{+-}^{(2,1)}$	Expectation(σ) SIEHS SNPs	15 (4)	3 (2)	0.1 (0.4)	14 (4)
	Expectation(σ) ENCODE	106 (10)	180 (13)	346 (17)	96 (10)
$N_{-+}^{(2,1)}$	Expectation(σ) SIEHS SNPs	89 (9)	154 (12)	283 (15)	84 (9)
	Expectation(σ) ENCODE	608 (20)	534 (19)	520 (19)	552 (19)
$N_{--}^{(2,1)}$	Expectation(σ) SIEHS SNPs	955 (20)	890 (20)	877 (20)	906 (20)

Table 21

(Continued)

	NF1 (exome): 57 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	32 (4)	32 (4)	27 (4)	34 (4)
$N_{++}^{(2,1)}$	Expectation(σ) SIEHS SNPs	21 (4)	21 (4)	17 (3)	22 (4)
	Expectation(σ) ENCODE	0.6 (0.8)	0.1 (0.3)	0 (0.1)	0.6 (0.7)
$N_{+-}^{(2,1)}$	Expectation(σ) SIEHS SNPs	0.5 (0.7)	0.1 (0.3)	0 (0.1)	0.5 (0.7)
$N_{-+}^{(2,1)}$	Expectation(σ)	4 (2)	6 (2)	12 (3)	3 (2)

	ENCODE				
	Expectation(σ) SIEHS SNPs	3 (2)	5 (2)	10 (3)	3 (2)
$N_{--}^{(2,1)}$	Expectation(σ) ENCODE	21 (4)	18 (4)	18 (4)	19 (4)
	Expectation(σ) SIEHS SNPs	33 (4)	31 (4)	30 (4)	31 (4)

3.2.4 Empirical assessment of Chr1-22/500k SNP data using permutation modeling under specification (2,1)

3.2.4.1 Empirical assessment in the genome under specification (2,1)

I calculate expectations using the parameters for a European population (Ionnita-Laza et al.) assuming a minimum variant frequency of 0.001 and without truncation. I examine 488,146 SNPs on Chromosome 1-22 with results as shown in Table 22. The observed average number of SNPs that appeared in a random sample of 100 but not in a random sample of 1500 is 30. This compares to an expected count of about 50 assuming truncation and 45 assuming no truncation. The modeled standard deviation is 7, which is smaller than the sample standard deviation of 20. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 1,466 to 4,531 in two random samples of 800.

Table 22
Categorization of SNPs in the genome by appearance in two randomly selected groups under specification (2,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Average($\hat{\sigma}$)	448,857.4	451,576.9	429,278.8	461,094.9

		(1,363)	(1,373)	(1,858)	(1,057)
	Low	447,231	449,393	426,757	459,012
	High	451,517	454,147	431,499	462,106
	Expectation(σ) ENCODE*	422,373 (239)	428,832 (228)	361,092 (307)	451,913 (183)
	Expectation(σ) SIEHS SNPs*	407,875 (259)	415,614 (249)	339,228 (322)	443,085 (202)
	Expectation(σ) ENCODE	272,805 (347)	277,202 (346)	233,108 (349)	292,711 (342)
	Expectation(σ) SIEHS SNPs	176,621 (336)	180,172 (337)	146,795 (320)	192,602 (341)
$N_{+-}^{(2,1)}$	Average($\hat{\sigma}$)	3,249.9 (589)	530.4 (173)	30.3 (20)	3,056.1 (1,009)
	Low	2,192	275	8	1,466
	High	4,150	787	53	4,531
	Expectation(σ) ENCODE*	7,431 (85)	881 (30)	44 (7)	5,096 (71)
	Expectation(σ) SIEHS SNPs*	8,837 (93)	1,098 (33)	55 (7)	6,314 (79)
	Expectation(σ) ENCODE	5,320 (73)	923 (30)	49 (7)	4,835 (69)
	Expectation(σ) SIEHS SNPs	4,347 (66)	596 (28)	42 (7)	4,129 (64)
$N_{-+}^{(2,1)}$	Average($\hat{\sigma}$)	18,293.5 (2,058)	28,503.8 (1,876)	52,637.6 (2,174)	14,647.5 (2,655)
	Low	15,176	25,197	50,015	11,258
	High	21,959	31,480	55,489	19,467
	Expectation(σ) ENCODE*	36,606 (184)	53,026 (217)	123,114 (303)	23,718 (150)
	Expectation(σ) SIEHS SNPs*	44,216 (201)	64,595 (237)	143,927 (319)	29,372 (166)
	Expectation(σ) ENCODE	31,128 (171)	52,981 (217)	101,933 (284)	28,249 (163)
	Expectation(σ) SIEHS SNPs	26,058 (157)	45,365 (203)	83,158 (263)	24,719 (153)
$N_{--}^{(2,1)}$	Average($\hat{\sigma}$)	17,745.2 (2,113)	7,534.9 (344)	6,199.2 (317)	9,347.5 (630)
	Low	15,377	6,998	5,874	8,201
	High	20,826	8,025	6,617	10,251
	Expectation(σ) ENCODE*	21,826 (144)	5,406 (73)	3,866 (62)	7,419 (85)
	Expectation(σ) SIEHS SNPs*	27,218 (168)	6,839 (82)	4,937 (70)	9,375 (96)

	Expectation(σ) ENCODE	178,893 (337)	157,040 (326)	153,056 (324)	162,351 (329)
	Expectation(σ) SIEHS SNPs	281,120 (345)	261,813 (348)	258,151 (349)	266,696 (348)

3.2.4.2 Empirical assessment in the exome under specification (2,1)

I next examine 3,342 SNPs in exonic regions on Chromosome 1-22 with results as shown in Table 23. The observed average number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0.4. This compares to an expected count of about 0.3 with or without truncation. The modeled standard deviation of 0.7 is close to the sample standard deviation of 0.6. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 10 to 31 in two random samples of 800.

Table 23
Categorization of SNPs in the exome by appearance in two randomly selected groups under specification (2,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Average($\hat{\sigma}$)	3,024.4 (13)	3,047.4 (11)	2,844.7 (20)	3,128.8 (8)
	Low	3,011	3,033	2,816	3,114
	High	3,048	3,066	2,873	3,138
	Expectation(σ) ENCODE*	2,892 (20)	2,936 (19)	2,472 (25)	3,094 (15)
	Expectation(σ) SIEHS SNPs*	2,792 (21)	2,845 (21)	2,322 (27)	3,033 (17)
	Expectation(σ) ENCODE	1,868 (29)	1,898 (29)	1,596 (29)	2,004 (28)
	Expectation(σ) SIEHS SNPs	1,209 (28)	1,234 (28)	1,005 (27)	1,319 (28)
	Average($\hat{\sigma}$)	26.3 (7)	3.3 (2)	0.4 (0.7)	21.2 (8)

$N_{+-}^{(2,1)}$	Low	19	0	0	10
	High	38	9	2	31
	Expectation(σ) ENCODE*	50 (7)	6 (2)	0.3 (0.6)	35 (6)
	Expectation(σ) SIEHS SNPs*	61 (8)	8 (3)	0.4 (0.6)	43 (7)
	Expectation(σ) ENCODE	36 (6)	6 (3)	0.3 (0.6)	33 (6)
	Expectation(σ) SIEHS SNPs	30 (5)	5 (2)	0.3 (0.5)	28 (5)
$N_{-+}^{(2,1)}$	Average($\hat{\sigma}$)	151.3 (21)	233.4 (17)	448.7 (22)	119.7 (20)
	Low	129	206	415	95
	High	182	257	480	155
	Expectation(σ) ENCODE*	251 (15)	363 (18)	843 (25)	162 (12)
	Expectation(σ) SIEHS SNPs*	303 (17)	442 (20)	985 (26)	201 (14)
	Expectation(σ) ENCODE	213 (14)	363 (18)	698 (23)	193 (13)
	Expectation(σ) SIEHS SNPs	178 (13)	311 (17)	569 (22)	169 (13)
$N_{--}^{(2,1)}$	Average($\hat{\sigma}$)	140 (22)	57.9 (5)	48.2 (3)	72.3 (7)
	Low	111	50	44	60
	High	170	67	53	82
	Expectation(σ) ENCODE*	149 (12)	37 (6)	27 (5)	51 (7)
	Expectation(σ) SIEHS SNPs*	186 (13)	47 (7)	34 (6)	64 (8)
	Expectation(σ) ENCODE	1,225 (28)	1,075 (27)	1,048 (27)	1,112 (27)
	Expectation(σ) SIEHS SNPs	1,925 (29)	1,792 (29)	1,767 (29)	1,826 (29)

3.2.4.3 Empirical assessment in specific genes under specification (2,1)

I examine 73 SNPs in gene FBN1 on Chromosome 15 with results as shown in Table 24. The observed number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0. This compares to an expected count of about 0 with or without truncation. The average counts increase to about 2 in two random samples of 800.

Table 24
Categorization of SNPs in gene FBN1with 73 SNPs by appearance in two randomly selected groups under specification (2,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Average($\hat{\sigma}$)	52.6 (0.8)	53 (0.9)	47.5 (2)	55.7 (0.9)
	Low	51	51	46	54
	High	54	54	51	57
	Expectation(σ) ENCODE*	63 (3)	64 (3)	54 (4)	68 (2)
	Expectation(σ) SIEHS SNPs*	61 (3)	62 (3)	51 (4)	66 (2)
	Expectation(σ) ENCODE	41 (4)	41 (4)	35 (4)	44 (4)
	Expectation(σ) SIEHS SNPs	26 (4)	27 (4)	22 (4)	29 (4)
$N_{+-}^{(2,1)}$	Average($\hat{\sigma}$)	0.7 (1)	0.3 (0.7)	0 (0)	2.3 (3)
	Low	0	0	0	0
	High	3	2	0	9
	Expectation(σ) ENCODE*	1 (1)	0.1 (0.4)	0 (0.1)	0.8 (0.9)
	Expectation(σ) SIEHS SNPs*	1 (1)	0.2 (0.4)	0 (0.1)	0.9 (1)
	Expectation(σ) ENCODE	0.8 (0.9)	0.1 (0.4)	0 (0.1)	0.7 (0.8)
	Expectation(σ) SIEHS SNPs	0.7 (0.8)	0.1 (0.3)	0 (0.1)	0.6 (0.8)
$N_{-+}^{(2,1)}$	Average($\hat{\sigma}$)	8.6 (4)	15,8 (2)	21.9 (2)	9.9 (3)
	Low	3	12	19	4
	High	17	19	24	14
	Expectation(σ) ENCODE*	5 (2)	8 (3)	18 (4)	4 (2)
	Expectation(σ) SIEHS SNPs*	7 (2)	10 (3)	22 (4)	4 (2)

	Expectation(σ) ENCODE	5 (2)	8 (3)	15 (3)	4 (2)
	Expectation(σ) SIEHS SNPs	4 (2)	7 (2)	12 (3)	4 (2)
$N_{--}^{(2,1)}$	Average($\hat{\sigma}$)	11.1 (5)	3.9 (1)	3.6 (1)	5.1 (1)
	Low	3	3	3	3
	High	17	6	6	6
	Expectation(σ) ENCODE*	3 (2)	0.8 (0.9)	0.6 (0.8)	1 (1)
	Expectation(σ) SIEHS SNPs*	4 (2)	1 (1)	0.7 (0.9)	1 (1)
	Expectation(σ) ENCODE	27 (4)	23 (4)	23 (4)	24 (4)
	Expectation(σ) SIEHS SNPs	42 (4)	39 (4)	39 (4)	40 (4)

I examine 16 SNPs in gene NF1 on Chromosome 17 with results as shown in Table 25. The observed number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0. This compares to an expected count of about 0 with or without truncation. The counts are close to 0 for other comparisons.

Table 25

Categorization of SNPs in gene NF1with 16 SNPs by appearance in two randomly selected groups under specification (2,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(2,1)}$	Average($\hat{\sigma}$)	13.6 (0.5)	13.7 (0.5)	12.9 (0.6)	14 (0)
	Low	13	13	12	14
	High	14	14	14	14
	Expectation(σ) ENCODE*	14 (1)	14 (1)	12 (2)	15 (1)
	Expectation(σ) SIEHS SNPs*	13 (1)	14 (1)	11 (2)	15 (1)
	Expectation(σ) ENCODE	9 (2)	9 (2)	8 (2)	10 (2)
	Expectation(σ) SIEHS SNPs	6 (2)	6 (2)	5 (2)	6 (2)
$N_{+-}^{(2,1)}$	Average($\hat{\sigma}$)	0.1 (0.3)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0
	High	1	0	0	0

	Expectation(σ) ENCODE*	0.2 (0.5)	0 (0.2)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs*	0.3 (0.5)	0 (0.2)	0 (0)	0.2 (0.5)
	Expectation(σ) ENCODE	0.2 (0.4)	0 (0.2)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs	0.1 (0.4)	0 (0.2)	0 (0)	0.1 (0.4)
$N_{-+}^{(2,1)}$	Average($\hat{\sigma}$)	0.6 (0.7)	1.1 (0.6)	2 (0.7)	0.6 (0.5)
	Low	0	0	1	0
	High	2	2	3	1
	Expectation(σ) ENCODE*	1 (1)	2 (1)	4 (2)	0.8 (0.9)
	Expectation(σ) SIEHS SNPs*	1 (1)	2 (1)	5 (2)	1 (1)
	Expectation(σ) ENCODE	1 (1)	2 (1)	3 (2)	0.9 (0.9)
	Expectation(σ) SIEHS SNPs	0.9 (0.9)	1 (1)	3 (2)	0.8 (0.9)
$N_{--}^{(2,1)}$	Average($\hat{\sigma}$)	1.7 (0.5)	1.2 (0.4)	1.1 (0.3)	1.4 (0.5)
	Low	1	1	1	1
	High	2	2	2	2
	Expectation(σ) ENCODE*	0.7 (0.8)	0.2 (0.4)	0.1 (0.4)	0.2 (0.5)
	Expectation(σ) SIEHS SNPs*	0.9 (0.9)	0.2 (0.5)	0.2 (0.4)	0.3 (0.5)
	Expectation(σ) ENCODE	6 (2)	5 (2)	5 (2)	5 (2)
	Expectation(σ) SIEHS SNPs	9 (2)	9 (2)	8 (2)	9 (2)

3.3 Results when the variants appear at least 3 times in case group and at least once in control group (under the specification (3,1))

3.3.1 Theoretical expected values assuming 10,000,000 variants in the population under specification (3,1)

Under the specification (3, 1), the SNP $i, i = 1,\mathrm{K}\ , I$, is said to appear in case group when at least three individuals in that group have the minority allele. It appears in the larger group when it appears in at least one member. I calculate the expected number of variants observed in the smaller group but not in the larger group, which I denote with $E(N_{+-}^{(3,1)})$, and I report these values in Table 26. These values are comparatively smaller than those under specification (1,1) and (2,1). For example, the expected number in a random group of 100 that do not appear in a random group of 1500 decreases from the order of 1,000 under the specification (2,1) to about 44. The expected numbers are still large, suggesting that a genome wide study using rare variants would report a large number of variants in the smaller group that do not appear in the larger (control) group. The expected number in a random group of 400 that do not appear in a random group of 1200 is on the order of 3,400. In a comparison of 800 with 800, the expected value increases to about 37,000.

Table 26

Number of variants in the genome expected to appear in two samples among European population under specification (3,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Expectation(σ) ENCODE	5,355,328 (1,577)	5,392,607 (1,576)	4,408,803 (1,570)	5,784,894 (1,562)
	Expectation(σ) SIEHS SNPs	3,435,428 (1,502)	3,464,934 (1,505)	2,740,883 (1,411)	3,772,357 (1,533)
$N_{+-}^{(3,1)}$	Expectation(σ) ENCODE	40,706 (201)	3,427 (59)	44 (7)	37,047 (192)
	Expectation(σ) SIEHS SNPs	32,377 (180)	2,871 (54)	37 (6)	30,805 (175)
$N_{-+}^{(3,1)}$	Expectation(σ) ENCODE	870,937 (892)	1,371,420 (1,088)	2,454,731 (1,361)	790,189 (853)
	Expectation(σ)	716,584	1,155,349	1,969,847	679,612

	SIEHS SNPs	(816)	(1,011)	(1,258)	(796)
$N_{--}^{(3,1)}$	Expectation(σ) ENCODE	3,733,029 (1,530)	3,232,546 (1,479)	3,136,421 (1,467)	3,387,869 (1,497)
	Expectation(σ) SIEHS SNPs	5,815,611 (1,560)	5,376,846 (1,577)	5,289,233 (1,578)	5,517,226 (1,573)

3.3.2 Theoretical expected values assuming 150,000 variants in the exome under specification (3,1)

In Table 27, I report the expected numbers for a study considering only the exome, assuming 150,000 variants in the exome. In a comparison of a random sample of 100 to a random sample of 1500 from the same population, the number of variants appearing in the smaller group but not the larger is around 0.7 (with a standard deviation of about 0.8), decreasing from the 15 expected under the specification (2,1). The expected number increases to about 560 when comparing two randomly selected groups of 800. This is smaller than the expectation of around 1,500 under the specification (2, 1), but is still not practical in an exome wide study.

Table 27
Number of variants in the exome expected to appear in two samples among European population under specification (3,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Expectation(σ) ENCODE	80,329 (193)	80,889 (193)	66,132 (192)	86,773 (191)
	Expectation(σ) SIEHS SNPs	51,531 (184)	51,974 (184)	41,113 (173)	56,585 (188)
$N_{+-}^{(3,1)}$	Expectation(σ) ENCODE	611 (25)	51 (7)	0.7 (0.8)	556 (24)
	Expectation(σ)	486 (22)	43 (7)	0.6 (0.7)	462 (21)

	SIEHS SNPs				
$N_{-+}^{(3,1)}$	Expectation(σ) ENCODE	13,064 (109)	20,571 (133)	36,821 (167)	11,853 (104)
	Expectation(σ) SIEHS SNPs	10,749 (100)	17,330 (124)	29,548 (154)	10,194 (97)
$N_{--}^{(3,1)}$	Expectation(σ) ENCODE	55,995 (187)	48,488 (181)	47,046 (180)	50,818 (183)
	Expectation(σ) SIEHS SNPs	87,234 (191)	80,653 (193)	79,338 (193)	82,758 (193)

3.3.3 Theoretical expected values in specific genes under specification (3,1)

I then consider the specification (3,1) applied at the gene level. Table 28 contains the expected numbers in gene FBN1 for the European population. The expected number appearing in a randomly selected group of 100 that do not appear in a randomly selected group of 1500 is about 0 with a standard deviation of 0.1. The expected number increases to about 5 when comparing two randomly selected groups of 800. These numbers are much practical. Table 28 also contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts are close to 0. These expectations suggest that the specification (3, 1) would be practical at the gene level.

Table 28

Number of variants in gene FBN1 expected to appear in two samples among European population under specification (3,1)

	FBN1: 1,301 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Expectation(σ) ENCODE	697 (18)	702 (18)	574 (18)	753 (18)

	Expectation(σ) SIEHS SNPs	447 (17)	451 (17)	357 (16)	491 (17)
$N_{+-}^{(3,1)}$	Expectation(σ) ENCODE	5 (2)	0.4 (0.7)	0 (0.1)	5 (2)
	Expectation(σ) SIEHS SNPs	4 (2)	0.4 (0.6)	0 (0.1)	4 (2)
$N_{-+}^{(3,1)}$	Expectation(σ) ENCODE	113 (10)	178 (12)	319 (16)	103 (10)
	Expectation(σ) SIEHS SNPs	93 (9)	150 (12)	256 (14)	88 (9)
$N_{--}^{(3,1)}$	Expectation(σ) ENCODE	486 (17)	420 (17)	408 (17)	441 (17)
	Expectation(σ) SIEHS SNPs	757 (18)	700 (18)	688 (18)	718 (18)

Table 28
(Continued)

	FBN1 (exon): 65 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Expectation(σ) ENCODE	35 (4)	35 (4)	29 (4)	38 (4)
	Expectation(σ) SIEHS SNPs	22 (4)	23 (4)	18 (4)	25 (4)
$N_{+-}^{(3,1)}$	Expectation(σ) ENCODE	0.3 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)
	Expectation(σ) SIEHS SNPs	0.2 (0.5)	0 (0.1)	0 (0)	0.2 (0.4)
$N_{-+}^{(3,1)}$	Expectation(σ) ENCODE	6 (2)	9 (3)	16 (3)	5 (2)
	Expectation(σ) SIEHS SNPs	5 (2)	8 (4)	13 (3)	4 (2)
$N_{--}^{(3,1)}$	Expectation(σ) ENCODE	24 (4)	21 (4)	20 (4)	22 (4)
	Expectation(σ) SIEHS SNPs	38 (4)	35 (4)	34 (4)	36 (4)

Table 29 contains the expected numbers for another gene, NF1, for the European population. The expected number appearing in a randomly selected group of 100 that do not appear in a randomly selected group of 1500 is about 0 with a standard deviation of 0.1. The expected number increases to about 6 when comparing two randomly

selected groups of 800. Table 29 also contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts are also close to 0.

Table 29

Number of variants in gene NF1 expected to appear in two samples among European population under specification (3,1)

	NF1: 1,659 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	888 (20)	895 (20)	731 (20)	960 (20)
$N_{++}^{(3,1)}$	Expectation(σ) SIEHS SNPs	570 (19)	575 (19)	455 (18)	626 (20)
	Expectation(σ) ENCODE	7 (3)	0.6 (0.8)	0 (0.1)	6 (2)
$N_{+-}^{(3,1)}$	Expectation(σ) SIEHS SNPs	5 (2)	0.5 (0.7)	0 (0.1)	5 (2)
	Expectation(σ) ENCODE	144 (11)	228 (14)	407 (18)	131 (11)
$N_{-+}^{(3,1)}$	Expectation(σ) SIEHS SNPs	119 (11)	192 (13)	327 (16)	113 (10)
	Expectation(σ) ENCODE	619 (20)	536 (19)	520 (19)	562 (19)
$N_{--}^{(3,1)}$	Expectation(σ) SIEHS SNPs	965 (20)	892 (20)	877 (20)	915 (20)

Table 29

(Continued)

	NF1 (exome): 57 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	31 (4)	31 (4)	25 (4)	33 (4)
$N_{++}^{(3,1)}$	Expectation(σ) SIEHS SNPs	20 (4)	20 (4)	16 (3)	22 (4)
	Expectation(σ) ENCODE	0.2 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)
$N_{+-}^{(3,1)}$	Expectation(σ) SIEHS SNPs	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) ENCODE	5 (2)	8 (3)	14 (3)	5 (2)
$N_{-+}^{(3,1)}$	Expectation(σ) SIEHS SNPs	4 (2)	7 (2)	11 (3)	4 (2)
$N_{--}^{(3,1)}$	Expectation(σ)	21 (4)	18 (4)	18 (4)	19 (4)

	ENCODE				
	Expectation(σ) SIEHS SNPs	33 (4)	31 (4)	30 (4)	31 (4)

3.3.4 Empirical assessment of Chr1-22/500k SNP data using permutation modeling under specification (3,1)

3.3.4.1 Empirical assessment in the genome under specification (3,1)

I calculate expectations using the parameters for a European population (Ionnita-Laza et al. 2009) assuming a minimum variant frequency of 0.001 and without truncation. I examine 488,146 SNPs on Chromosome 1-22 with results as shown in Table 30. The observed average number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 1.4. This compares to an expected count of about 3 assuming truncation and 2 assuming no truncation. The modeled and sample standard deviations are both 2. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 517 to 2,035 in two random samples of 800.

Table 30

Categorization of SNPs in the genome by appearance in two randomly selected groups under specification (3,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Average($\hat{\sigma}$)	442,084.7 (942)	443,106.2 (973)	420,469.2 (1,074)	454,029.5 (1,263)
	Low	440,931	441,619	419,066	451,724
	High	444,123	445,186	421,831	455,538
	Expectation(σ)	404,977	407,764	333,420	437,203

	ENCODE*	(263)	(259)	(325)	(214)
	Expectation(σ) SIEHS SNPs*	387,579 (283)	390,864 (279)	309,247 (337)	425,249 (234)
	Expectation(σ) ENCODE	261,418 (348)	263,238 (348)	215,214 (347)	282,387 (345)
	Expectation(σ) SIEHS SNPs	167,699 (332)	169,139 (332)	133,795 (312)	184,146 (339)
$N_{+-}^{(3,1)}$	Average($\hat{\sigma}$)	1,125.8 (206)	104.3 (43)	1.4 (2)	1,279.8 (509)
	Low	730	42	0	517
	High	1,374	179	7	2,035
	Expectation(σ) ENCODE*	3,003 (55)	217 (15)	3 (2)	2,434 (49)
	Expectation(σ) SIEHS SNPs*	3,553 (59)	267 (16)	3 (2)	2,986 (54)
	Expectation(σ) ENCODE	1,987 (44)	167 (13)	2 (1)	1,808 (42)
	Expectation(σ) SIEHS SNPs	1,580 (40)	140 (12)	2 (1)	1,504 (39)
$N_{-+}^{(3,1)}$	Average($\hat{\sigma}$)	25,066.2 (2,255)	36,974.5 (1,457)	61,447.2 (1,371)	21,712.9 (2,879)
	Low	21,467	34,158	59,790	17,826
	High	28,504	39,254	63,152	26,755
	Expectation(σ) ENCODE*	54,002 (219)	74,095 (251)	150,786 (323)	38,428 (188)
	Expectation(σ) SIEHS SNPs*	64,512 (237)	89,345 (270)	173,908 (335)	47,207 (207)
	Expectation(σ) ENCODE	42,514 (197)	66,945 (240)	119,827 (301)	38,573 (188)
	Expectation(σ) SIEHS SNPs	34,980 (180)	56,398 (223)	96,157 (278)	33,175 (176)
$N_{--}^{(3,1)}$	Average($\hat{\sigma}$)	19,869.3 (2,356)	7,061 (474)	6,228.2 (335)	11,123.8 (1,130)
	Low	17,460	7,231	5,881	9,150
	High	23,602	8,635	6,663	12,747
	Expectation(σ) ENCODE*	26,164 (157)	6,071 (77)	3,938 (62)	10,081 (99)
	Expectation(σ) SIEHS SNPs*	32,503 (174)	7,670 (87)	4,988 (70)	12,703 (111)
	Expectation(σ) ENCODE	182,226 (338)	157,795 (327)	153,103 (324)	165,378 (331)
	Expectation(σ) SIEHS SNPs	283,887 (345)	262,469 (348)	258,192 (349)	269,321 (347)

3.3.4.2 Empirical assessment in the exome under specification (3,1)

I next examine 3,342 SNPs in exonic regions on Chromosome 1-22 with results as shown in Table 31. The observed average number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0.1. This compares to an expected count of about 0 with or without truncation. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 4 to 14 in two random samples of 800.

Table 31
Categorization of SNPs in the exome by appearance in two randomly selected groups under specification (3,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Average($\hat{\sigma}$)	2,962.3 (9)	2,972 (9)	2,770.9 (13)	3,069.7 (12)
	Low	2,947	2,959	2,759	3,048
	High	2,979	2,985	2,788	3,087
	Expectation(σ) ENCODE*	2,773 (22)	2,792 (21)	2,283 (27)	2,993 (18)
	Expectation(σ) SIEHS SNPs*	2,653 (23)	2,676 (23)	2,117 (28)	2,911 (19)
	Expectation(σ) ENCODE	1,790 (29)	1,802 (29)	1,473 (29)	1,933 (29)
	Expectation(σ) SIEHS SNPs	1,148 (27)	1,158 (28)	916 (26)	1,261 (28)
$N_{+-}^{(3,1)}$	Average($\hat{\sigma}$)	10.5 (3)	0.8 (0.6)	0.1 (0.3)	8.5 (4)
	Low	7	0	0	4
	High	15	2	1	14
	Expectation(σ) ENCODE*	21 (5)	1 (1)	0 (0.1)	17 (4)
	Expectation(σ)	24 (5)	2 (1)	0 (0.2)	20 (5)

	SIEHS SNPs*				
	Expectation(σ) ENCODE	14 (4)	1 (1)	0 (0.1)	12 (4)
	Expectation(σ) SIEHS SNPs	11 (3)	1 (1)	0 (0.1)	10 (3)
$N_{-+}^{(3,1)}$	Average($\hat{\sigma}$)	213.4 (23)	308.8 (12)	522.5 (15)	178.8 (24)
	Low	185	287	500	142
	High	249	324	539	219
	Expectation(σ) ENCODE*	370 (18)	507 (21)	1,032 (27)	263 (16)
	Expectation(σ) SIEHS SNPs*	442 (20)	612 (22)	1,191 (28)	323 (17)
	Expectation(σ) ENCODE	291 (16)	458 (20)	820 (25)	264 (16)
	Expectation(σ) SIEHS SNPs	239 (15)	386 (18)	658 (23)	227 (15)
$N_{--}^{(3,1)}$	Average($\hat{\sigma}$)	155.8 (22)	60.4 (6)	48.5 (3)	85 (11)
	Low	125	50	44	68
	High	181	69	54	101
	Expectation(σ) ENCODE*	179 (13)	42 (6)	27 (5)	69 (8)
	Expectation(σ) SIEHS SNPs*	223 (14)	53 (7)	34 (6)	87 (9)
	Expectation(σ) ENCODE	1,248 (28)	1,080 (27)	1,048 (27)	1,132 (27)
	Expectation(σ) SIEHS SNPs	1,944 (29)	1,797 (29)	1,768 (29)	1,844 (29)

3.3.4.3 Empirical assessment in specific genes under specification (3,1)

I examine 73 SNPs in gene FBN1 on Chromosome 15 with results as shown in Table 32. The observed number of SNPs that appear in a random sample of 100 but not

in a random sample of 1500 is 0. This compares to an expected count of about 0 with or without truncation. The average counts increase to 0.6 in two random samples of 800.

Table 32
Categorization of SNPs in gene FBN1with 73 SNPs by appearance in two randomly selected groups under specification (3,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Average($\hat{\sigma}$)	50.9 (1)	51.1 (1)	45.3 (2)	53 (0.8)
	Low	48	49	41	52
	High	52	52	49	54
	Expectation(σ) ENCODE*	61 (3)	61 (3)	50 (4)	65 (3)
	Expectation(σ) SIEHS SNPs*	58 (3)	58 (3)	46 (4)	64 (3)
	Expectation(σ) ENCODE	39 (4)	39 (4)	32 (4)	42 (4)
	Expectation(σ) SIEHS SNPs	25 (4)	25 (4)	20 (4)	28 (4)
$N_{+-}^{(3,1)}$	Average($\hat{\sigma}$)	0.2 (0.4)	0 (0)	0 (0)	0.6 (1)
	Low	0	0	0	0
	High	1	0	0	0
	Expectation(σ) ENCODE*	0.4 (0.7)	0 (0.2)	0 (0)	0.4 (0.6)
	Expectation(σ) SIEHS SNPs*	0.5 (0.7)	0 (0.2)	0 (0)	0.4 (0.7)
	Expectation(σ) ENCODE	0.3 (0.5)	0 (0.2)	0 (0)	0.3 (0.5)
	Expectation(σ) SIEHS SNPs	0.2 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)
$N_{-+}^{(3,1)}$	Average($\hat{\sigma}$)	10.3 (5)	17.7 (2)	24.1 (2)	12.6 (4)
	Low	5	14	21	6
	High	19	21	29	17
	Expectation(σ) ENCODE*	8 (3)	11 (3)	23 (4)	6 (2)
	Expectation(σ) SIEHS SNPs*	10 (3)	13 (3)	26 (4)	7 (3)
	Expectation(σ) ENCODE	6 (2)	10 (3)	18 (4)	6 (2)
	Expectation(σ) SIEHS SNPs	5 (2)	8 (3)	14 (3)	5 (2)
$N_{--}^{(3,1)}$	Average($\hat{\sigma}$)	11.6 (4)	4.2 (2)	3.6 (1)	6.8 (3)
	Low	3	3	3	3

	High	17	7	6	12
	Expectation(σ) ENCODE*	4 (2)	0.9 (0.9)	0.6 (0.8)	2 (1)
	Expectation(σ) SIEHS SNPs*	5 (2)	1 (1)	0.7 (0.9)	2 (1)
	Expectation(σ) ENCODE	27 (4)	24 (4)	23 (4)	25 (4)
	Expectation(σ) SIEHS SNPs	42 (4)	39 (4)	39 (4)	40 (4)

I examine 16 SNPs in gene NF1 on Chromosome 17 with results as shown in Table 33. The observed number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0. This compares to an expected count of about 0 with or without truncation. The counts for other comparisons are also close to 0.

Table 33

Categorization of SNPs in gene NF1with 16 SNPs by appearance in two randomly selected groups under specification (3,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(3,1)}$	Average($\hat{\sigma}$)	13.5 (0.5)	13.6 (0.5)	12.5 (0.5)	14 (0)
	Low	13	13	12	14
	High	14	14	13	14
	Expectation(σ) ENCODE*	13 (2)	13 (1)	11 (2)	14 (1)
	Expectation(σ) SIEHS SNPs*	13 (2)	13 (2)	10 (2)	14 (1)
	Expectation(σ) ENCODE	9 (2)	9 (2)	7 (2)	9 (2)
	Expectation(σ) SIEHS SNPs	5 (2)	6 (2)	4 (2)	6 (2)
$N_{+-}^{(3,1)}$	Average($\hat{\sigma}$)	0.1 (0.3)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0
	High	1	0	0	0
	Expectation(σ) ENCODE*	0.1 (0.3)	0 (0.1)	0 (0)	0 (0.3)
	Expectation(σ) SIEHS SNPs*	0.1 (0.3)	0 (0.1)	0 (0)	0 (0.3)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.2)

	Expectation(σ) SIEHS SNPs	0.1 (0.2)	0 (0.1)	0 (0)	0 (0.2)
$N_{-+}^{(3,1)}$	Average($\hat{\sigma}$)	0.7 (0.7)	1.2 (0.6)	2.4 (0.5)	0.6 (0.5)
	Low	0	0	2	0
	High	2	2	3	1
	Expectation(σ) ENCODE*	2 (1)	2 (1)	5 (2)	1 (1)
	Expectation(σ) SIEHS SNPs*	2 (1)	3 (2)	6 (2)	2 (1)
	Expectation(σ) ENCODE	1 (1)	2 (1)	4 (2)	1 (1)
	Expectation(σ) SIEHS SNPs	1 (1)	2 (1)	3 (2)	1 (1)
$N_{--}^{(3,1)}$	Average($\hat{\sigma}$)	1.7 (0.5)	1.2 (0.4)	1.1 (0.3)	1.4 (0.5)
	Low	1	1	1	1
	High	2	2	2	2
	Expectation(σ) ENCODE*	0.9 (0.9)	0.2 (0.4)	0.1 (0.4)	0.3 (0.6)
	Expectation(σ) SIEHS SNPs*	1 (1)	0.3 (0.5)	0.2 (0.4)	0.4 (0.6)
	Expectation(σ) ENCODE	6 (2)	5 (2)	5 (2)	5 (2)
	Expectation(σ) SIEHS SNPs	9 (2)	9 (2)	8 (2)	9 (2)

3.4 Results when the variants appear at least 4 times in case group and at least once in control group (under the specification (4,1))

3.4.1 Theoretical expected values assuming 10,000,000 variants in the population under specification (4,1)

The specification (4,1) states the SNP $i, i = 1,\mathrm{K}\ , I$, is said to appear in a group when at least four individuals in that group have the minority allele. I calculate the

expected number of variants observed in the smaller group but not in the larger group, which I denoted with $E(N_{+-}^{(4,1)})$, and I report these values in Table 34. These values are comparatively smaller than those under the specification (3,1). For example, the expected number in a random group of 100 that do not appear in a random group of 1500 decreased from the order of 44 to about 4. The expected numbers are practical in a genome wide study. The expected number in a random group of 400 that did not appear in a random group of 1200 is on the order of 800. In a comparison of 800 with 800, the expected value increases to about 16,000.

Table 34

Number of variants in the genome expected to appear in two samples among European population under specification (4,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Expectation(σ) ENCODE	5,164,466 (1,580)	5,179,934 (1,580)	4,145,951 (1,558)	5,609,750 (1,569)
	Expectation(σ) SIEHS SNPs	3,289,072 (1,486)	3,301,435 (1,487)	2,555,525 (1,379)	3,634,701 (1,521)
$N_{+-}^{(4,1)}$	Expectation(σ) ENCODE	16,304 (128)	836 (29)	4 (2)	16,242 (127)
	Expectation(σ) SIEHS SNPs	12,965 (114)	602 (25)	2 (1)	12,580 (112)
$N_{-+}^{(4,1)}$	Expectation(σ) ENCODE	1,061,799 (974)	1,584,093 (1,155)	2,717,584 (1,407)	965,333 (934)
	Expectation(σ) SIEHS SNPs	862,940 (888)	1,318,847 (1,070)	2,155,205 (1,300)	817,268 (866)
$N_{--}^{(4,1)}$	Expectation(σ) ENCODE	3,757,431 (1,532)	3,235,137 (1,479)	3,136,461 (1,467)	3,408,674 (1,499)
	Expectation(σ) SIEHS SNPs	5,835,023 (1,559)	5,379,116 (1,577)	5,289,268 (1,578)	5,535,451 (1,572)

3.4.2 Theoretical expected values assuming 150,000 variants in the exome under specification (4,1)

In Table 35, I report the expected numbers for a study considering only the exome, assuming 150,000 variants in the exome, when a variant is declared present in a group when at least two individuals have the variant. In a comparison of a random sample of 100 to a random sample of 1500 from the same population, the number of variants appearing in the smaller group but not the larger is around 0.1 (with a standard deviation of about 0.2), decreasing from the 0.7 expected under the specification (3, 1). The expected number increases to about 200 when comparing two randomly selected groups of 800. This is smaller than the expectation of around 560 under the specification (3, 1). These expected counts under this specification are much more practical than the counts under previous specifications in an exome wide study.

Table 35
Number of variants in the exome expected to appear in two samples among European population under specification (4,1)

	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Expectation(σ) ENCODE	77,467 (194)	77,699 (194)	62,189 (191)	84,146 (192)
	Expectation(σ) SIEHS SNPs	49,336 (182)	49,522 (182)	38,333 (169)	54,521 (186)
$N_{+-}^{(4,1)}$	Expectation(σ) ENCODE	245 (16)	13 (4)	0.1 (0.2)	244 (16)
	Expectation(σ) SIEHS SNPs	194 (14)	9 (3)	0 (0.2)	189 (14)
$N_{-+}^{(4,1)}$	Expectation(σ) ENCODE	15,927 (119)	23,761 (141)	40,764 (172)	14,480 (114)
	Expectation(σ) SIEHS SNPs	12,944 (109)	19,783 (131)	32,328 (159)	12,259 (106)
$N_{--}^{(4,1)}$	Expectation(σ) ENCODE	56,361 (188)	48,527 (181)	47,047 (180)	51,130 (184)
	Expectation(σ) SIEHS SNPs	87,525 (191)	80,687 (193)	79,839 (193)	83,032 (193)

3.4.3 Theoretical expected values in specific genes under specification (4,1)

I then consider the specification (4, 1) applied at the gene level. Table 36 contains the expected numbers in gene FBN1 for the European population. The expected number appearing in a randomly selected group of 100 that do not appear in a randomly selected group of 1500 is about 0. The expected number increases to 2 when comparing two randomly selected groups of 800. These numbers are much practical. Table 36 also contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts are close to 0. These expectations suggest that this specification would be practical at the gene level.

Table 36

Number of variants in gene FBN1 expected to appear in two samples among European population under specification (4,1)

	FBN1: 1,301 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	672 (18)	674 (18)	539 (18)	730 (18)
$N_{++}^{(4,1)}$	Expectation(σ) SIEHS SNPs	428 (17)	430 (17)	332 (16)	473 (17)
	Expectation(σ) ENCODE	2 (1)	0.1 (0.3)	0 (0)	2 (1)
$N_{+-}^{(4,1)}$	Expectation(σ) SIEHS SNPs	2 (1)	0.1 (0.3)	0 (0)	2 (1)
	Expectation(σ) ENCODE	138 (11)	206 (13)	354 (16)	126 (11)
$N_{-+}^{(4,1)}$	Expectation(σ) SIEHS SNPs	112 (10)	172 (12)	280 (15)	106 (10)
$N_{--}^{(4,1)}$	Expectation(σ)	489 (17)	421 (17)	408 (17)	443 (17)

	ENCODE				
	Expectation(σ) SIEHS SNPs	759 (18)	700 (18)	688 (18)	720 (18)

Table 36
(Continued)

	FBN1 (exon): 65 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	34 (4)	34 (4)	27 (4)	36 (4)
$N_{++}^{(4,1)}$	Expectation(σ) SIEHS SNPs	21 (4)	21 (4)	17 (4)	24 (4)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
$N_{+-}^{(4,1)}$	Expectation(σ) SIEHS SNPs	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
	Expectation(σ) ENCODE	7 (2)	10 (3)	18 (4)	6 (2)
$N_{-+}^{(4,1)}$	Expectation(σ) SIEHS SNPs	6 (2)	9 (3)	14 (3)	5 (2)
	Expectation(σ) ENCODE	24 (4)	21 (4)	20 (4)	22 (4)
$N_{--}^{(4,1)}$	Expectation(σ) SIEHS SNPs	38 (4)	35 (4)	34 (4)	36 (4)

Table 37 contains the expected numbers for another gene, NF1, for the European population. The expected number appearing in a randomly selected group of 100 that do not appear in a randomly selected group of 1500 is 0. The expected number increases to about 3 when comparing two randomly selected groups of 800. Table 37 also contains the expected numbers for this gene when attention is restricted to exonic SNPs in the European population. The expected counts are also close to 0.

Table 37
Number of variants in gene NF1 expected to appear in two samples among European population under specification (4,1)

	NF1: 1,659 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Expectation(σ) ENCODE	857 (20)	859 (20)	688 (20)	931 (20)

	Expectation(σ) SIEHS SNPs	546 (19)	548 (19)	424 (18)	603 (20)
$N_{+-}^{(4,1)}$	Expectation(σ) ENCODE	3 (2)	0.1 (0.4)	0 (0)	3 (2)
	Expectation(σ) SIEHS SNPs	2 (1)	0.1 (0.3)	0 (0)	2 (1)
$N_{-+}^{(4,1)}$	Expectation(σ) ENCODE	176 (13)	263 (15)	451 (18)	160 (12)
	Expectation(σ) SIEHS SNPs	143 (11)	219 (14)	358 (17)	136 (11)
$N_{--}^{(4,1)}$	Expectation(σ) ENCODE	623 (20)	537 (19)	520 (19)	565 (19)
	Expectation(σ) SIEHS SNPs	968 (20)	892 (20)	877 (20)	918 (20)

Table 37
(Continued)

	NF1 (exome): 57 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Expectation(σ) ENCODE	29 (4)	30 (4)	24 (4)	32 (4)
	Expectation(σ) SIEHS SNPs	19 (4)	19 (4)	15 (3)	21 (4)
$N_{+-}^{(4,1)}$	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
	Expectation(σ) SIEHS SNPs	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
$N_{-+}^{(4,1)}$	Expectation(σ) ENCODE	6 (2)	9 (3)	15 (3)	6 (2)
	Expectation(σ) SIEHS SNPs	5 (2)	8 (3)	12 (3)	5 (2)
$N_{--}^{(4,1)}$	Expectation(σ) ENCODE	21 (4)	18 (4)	18 (4)	19 (4)
	Expectation(σ) SIEHS SNPs	33 (4)	30 (4)	30 (4)	32 (4)

3.4.4 Empirical assessment of Chr1-22/500k SNP data using permutation modeling under specification (4,1)

3.4.4.1 Empirical assessment assuming 10,000,000 variants in the genome under specification (4,1)

I calculate expectations using the parameters for a European population (Ionnita-Laza et al. 2009) assuming a minimum variant frequency of 0.001 and without truncation. I examine 488,146 SNPs on Chromosome 1-22 with results as shown in Table 38. The observed average number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0.1. This compares to an expected count of about 0.1 with or without truncation. The modeled standard deviation of 0.3 is close to the sample standard deviation of 0.4. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 186 to 863 in two random samples of 800.

Table 38

Categorization of SNPs in the genome by appearance in two randomly selected groups under specification (4,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Average($\hat{\sigma}$)	437,277.5 (639)	437,658.1 (654)	413,590.3 (881)	448,535.2 (1,175)
	Low	436,506	436,769	412,279	446,563
	High	438,802	439,208	414,663	450,015
	Expectation(σ) ENCODE*	390,567 (279)	391,747 (278)	313,543 (335)	424,298 (236)
	Expectation(σ) SIEHS SNPs*	371,097 (298)	372,494 (297)	288,335 (344)	410,081 (256)
	Expectation(σ) ENCODE	252,101 (349)	252,856 (349)	202,383 (344)	273,838 (348)
	Expectation(σ) SIEHS SNPs	160,555 (328)	161,158 (329)	124,747 (305)	177,427 (336)
$N_{+-}^{(4,1)}$	Average($\hat{\sigma}$)	402.5 (70)	21.9 (10)	0.1 (0.3)	513.3 (225)
	Low	269	6	0	186
	High	481	39	1	863
	Expectation(σ) ENCODE*	1,228 (35)	48 (7)	0.1 (0.4)	1,082 (33)
	Expectation(σ) SIEHS SNPs*	1,456 (38)	59 (8)	0.2 (0.4)	1,312 (36)
	Expectation(σ) ENCODE	796 (28)	41 (6)	0.2 (0.5)	793 (28)
	Expectation(σ) SIEHS SNPs	633 (25)	29 (5)	0.1 (0.3)	614 (25)
$N_{-+}^{(4,1)}$	Average($\hat{\sigma}$)	29,873.4 (2,361)	42,422.6 (1,109)	68,326.1 (1,173)	27,207.2 (2,791)
	Low	26,126	40,136	66,851	23,349
	High	32,929	44,104	69,967	31,836
	Expectation(σ) ENCODE*	68,412 (243)	90,112 (271)	170,663 (333)	51,333 (214)
	Expectation(σ) SIEHS SNPs*	80,993 (260)	107,715 (290)	194,820 (342)	62,376 (233)
	Expectation(σ) ENCODE	51,831 (215)	77,327 (255)	132,658 (311)	47,122 (206)
	Expectation(σ) SIEHS SNPs	42,124 (196)	64,379 (236)	105,206 (287)	39,895 (191)
$N_{--}^{(4,1)}$	Average($\hat{\sigma}$)	20,592.6 (2,450)	8,043.4 (505)	6,229.5 (336)	11,890.3 (1,412)

	Low	18,122	8,043.4	5,882	9,481
	High	24,510	7,267	6,669	13,964
	Expectation(σ) ENCODE*	27,939 (162)	6,239 (78)	3,940 (63)	11,433 (106)
	Expectation(σ) SIEHS SNPs*	34,600 (179)	7,878 (88)	4,991 (70)	14,377 (118)
	Expectation(σ) ENCODE	183,418 (338)	157,922 (327)	153,105 (324)	166,393 (331)
	Expectation(σ) SIEHS SNPs	284,834 (344)	262,579 (348)	258,194 (349)	270,211 (347)

3.4.4.2 Empirical assessment assuming 150,000 variants in the exome under specification (4,1)

I next examine 3,342 SNPs in exonic regions on Chromosome 1-22 with results as shown in Table 39. The observed number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 is 0. This compares to an expected count of about 0 with or without truncation. The counts increase as the minimum sample size increases. For example, the numbers in one group but not the other range from 1 to 8 in two random samples of 800.

Table 39
Categorization of SNPs in the exome by appearance in two randomly selected groups under specification (4,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Average($\hat{\sigma}$)	2,920.1 (10)	2,923.6 (10)	2,712.4 (10)	3,020.4 (14)
	Low	2,907	2,909	2,701	2,997
	High	2,943	2,944	2,724	3,049
	Expectation(σ) ENCODE*	2,674 (23)	2,682 (23)	2,147 (28)	2,905 (19)

	Expectation(σ) SIEHS SNPs*	2,541 (25)	2,550 (25)	1,974 (28)	2,808 (21)
	Expectation(σ) ENCODE	1,726 (29)	1,731 (29)	1,386 (28)	1,875 (29)
	Expectation(σ) SIEHS SNPs	1,099 (27)	1,103 (27)	854 (25)	1,215 (28)
$N_{+-}^{(4,1)}$	Average($\hat{\sigma}$)	3.6 (2)	0.1 (0.3)	0 (0)	3.6 (2)
	Low	1	0	0	1
	High	6	1	0	8
	Expectation(σ) ENCODE*	8 (3)	0.3 (0.6)	0 (0)	7 (3)
	Expectation(σ) SIEHS SNPs*	10 (3)	0.4 (0.6)	0 (0)	9 (3)
	Expectation(σ) ENCODE	5 (2)	0.3 (0.5)	0 (0)	5 (2)
	Expectation(σ) SIEHS SNPs	4 (2)	0.2 (0.4)	0 (0)	4 (2)
$N_{-+}^{(4,1)}$	Average($\hat{\sigma}$)	255.6 (24)	357.2 (13)	581 (12)	228.1 (27)
	Low	224	328	565	180
	High	301	369	595	270
	Expectation(σ) ENCODE*	468 (20)	617 (22)	1m168 (28)	351 (18)
	Expectation(σ) SIEHS SNPs*	555 (22)	737 (24)	1,334 (28)	427 (19)
	Expectation(σ) ENCODE	355 (18)	529 (21)	908 (26)	323 (17)
	Expectation(σ) SIEHS SNPs	288 (16)	441 (20)	720 (24)	273 (16)
$N_{--}^{(4,1)}$	Average($\hat{\sigma}$)	162.7 (23)	61.1 (7)	48.6 (3)	89.9 (12)
	Low	132	50	44	72
	High	189	70	54	107
	Expectation(σ) ENCODE*	191 (13)	43 (6)	27 (5)	78 (9)
	Expectation(σ) SIEHS SNPs*	237 (15)	54 (7)	34 (6)	98 (10)
	Expectation(σ) ENCODE	1,256 (28)	1,081 (27)	1,048 (27)	1,139 (27)
	Expectation(σ) SIEHS SNPs	1,950 (28)	1,798 (29)	1,768 (29)	2,850 (29)

3.4.4.3 Empirical assessment in specific genes under specification (4,1)

I examine 73 SNPs in gene FBN1 on Chromosome 15 with results as shown in Table 40. The observed numbers of SNPs is 0. And the expected count is also 0 with or without truncation.

Table 40

Categorization of SNPs in gene FBN1with 73 SNPs by appearance in two randomly selected groups under specification (4,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Average($\hat{\sigma}$)	50.2 (1)	50.2 (1)	42.8 (2)	52 (0)
	Low	48	48	40	52
	High	51	51	45	52
	Expectation(σ) ENCODE*	58 (3)	59 (3)	47 (4)	63 (3)
	Expectation(σ) SIEHS SNPs*	55 (4)	56 (4)	43 (4)	61 (3)
	Expectation(σ) ENCODE	38 (4)	38 (4)	30 (4)	41 (4)
	Expectation(σ) SIEHS SNPs	24 (4)	24 (4)	19 (4)	27 (4)
$N_{+-}^{(4,1)}$	Average($\hat{\sigma}$)	0 (0)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0
	High	0	0	0	0
	Expectation(σ) ENCODE*	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs*	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.2 (0.3)
	Expectation(σ) SIEHS SNPs	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
$N_{-+}^{(4,1)}$	Average($\hat{\sigma}$)	11 (5)	18.6 (2)	26.6 (2)	13.6 (4)
	Low	6	16	25	6
	High	21	22	30	18
	Expectation(σ) ENCODE*	10 (3)	13 (3)	26 (4)	8 (3)

	Expectation(σ) SIEHS SNPs*	12 (3)	16 (4)	29 (4)	9(3)
	Expectation(σ) ENCODE	8 (3)	12 (3)	20 (4)	7 (3)
	Expectation(σ) SIEHS SNPs	6 (2)	10 (3)	16 (4)	6 (2)
$N_{--}^{(4,1)}$	Average($\hat{\sigma}$)	11.8 (4)	4.2 (2)	3.6 (1)	7.4 (4)
	Low	3	3	3	3
	High	17	7	6	15
	Expectation(σ) ENCODE*	4 (2)	0.9 (1)	0.6 (0.8)	2 (1)
	Expectation(σ) SIEHS SNPs*	5 (2)	1 (1)	0.7 (0.9)	2 (1)
	Expectation(σ) ENCODE	27 (4)	24 (4)	23 (4)	25 (4)
	Expectation(σ) SIEHS SNPs	43 (4)	39 (4)	39 (4)	40 (4)

I examine 16 SNPs in gene NF1 on Chromosome 17 with results as shown in Table 41. The observed numbers of SNPs that appear in case group but not in another are all 0's. This compares to an expected count of about 0 with or without truncation.

Table 41

Categorization of SNPs in gene NF1with 16 SNPs by appearance in two randomly selected groups under specification (4,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
$N_{++}^{(4,1)}$	Average($\hat{\sigma}$)	13.2 (0.4)	13.2 (0.4)	12.2 (0.4)	13.7 (0.5)
	Low	13	13	12	13
	High	14	14	13	14
	Expectation(σ) ENCODE*	13 (2)	13 (2)	10 (2)	14 (1)
	Expectation(σ) SIEHS SNPs*	12 (2)	12 (2)	9 (2)	13 (1)
	Expectation(σ) ENCODE	8 (2)	8 (2)	7 (2)	9 (2)
	Expectation(σ) SIEHS SNPs	5 (2)	5 (2)	4 (2)	6 (2)
$N_{+-}^{(4,1)}$	Average($\hat{\sigma}$)	0 (0)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0
	High	0	0	0	0

	Expectation(σ) ENCODE*	0 (0.2)	0 (0)	0 (0)	0 (0.2)
	Expectation(σ) SIEHS SNPs*	0 (0.2)	0 (0)	0 (0)	0 (0.2)
	Expectation(σ) ENCODE	0 (0.2)	0 (0)	0 (0)	0 (0.2)
	Expectation(σ) SIEHS SNPs	0 (0.1)	0 (0)	0 (0)	0 (0.1)
$N_{-+}^{(4,1)}$	Average($\hat{\sigma}$)	1 (0.8)	1.6 (0.7)	2.7 (0.5)	0.9 (0.7)
	Low	0	0	2	0
	High	2	2	3	2
	Expectation(σ) ENCODE*	2 (1)	3 (2)	6 (2)	2 (1)
	Expectation(σ) SIEHS SNPs*	3 (1)	4 (2)	6 (2)	2 (1)
	Expectation(σ) ENCODE	2 (1)	3 (1)	4 (2)	2 (1)
	Expectation(σ) SIEHS SNPs	1 (1)	2 (1)	3 (2)	1 (1)
$N_{--}^{(4,1)}$	Average($\hat{\sigma}$)	1.8 (0.6)	1.2 (0.4)	1.1 (0.3)	1.4 (0.5)
	Low	1	1	1	1
	High	3	2	2	2
	Expectation(σ) ENCODE*	0.9 (0.9)	0.2 (0.4)	0.1 (0.4)	0.4 (0.6)
	Expectation(σ) SIEHS SNPs*	1 (1)	0.3 (0.5)	0.2 (0.4)	0.5 (0.7)
	Expectation(σ) ENCODE	6 (2)	5 (2)	5 (2)	5 (2)
	Expectation(σ) SIEHS SNPs	9 (2)	9 (2)	8 (2)	9 (2)

3.5 Results under selected specifications

3.5.1 Theoretical expected values assuming 10,000,000 variants in the population under selected specifications

Under the specification (*r, h*) , the SNP $i, i = 1, \mathrm{K}, I$, is said to appear in the case group when at least *r* individuals in that group have the minority allele, and is said to appear in the control group when at least *h* individuals have the minority allele. I calculate the expected number of variants observed in the smaller group but not in the larger group under selected specifications, and I report these values in Table 42. As *r* goes up, the expected values decrease. As *h* goes up, the expected values increase. For example, the expected number in a random group of 100 that do not appear in a random group of 1500 decreases from the order of 28,000 under the specification (1, 1) to a range between 2 and 4 under the specification (4, 1). The expected number increases from 1,000 under the specification (2, 1) to about 2,500 under the specification (2, 2). The expected numbers suggest that the specification (4, 1) and (4, 2) are practical in a genome wide study using 100 affected subjects and 1500 controls. Other specifications would report an extremely large number of variants in the smaller group that did not appear in the larger (control) group.

Table 42

Number of variants in the genome expected to appear in $N_{+-}^{(r,h)}$ under selected specifications

(*r,h*)	European	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Expectation(σ) ENCODE	348,838 (580)	127,714 (355)	28,206 (168)	316,658 (554)
	Expectation(σ) SIEHS SNPs	299,958 (539)	116,323 (339)	25,876 (161)	284,637 (526)
(2,1)	Expectation(σ) ENCODE	108,993 (328)	18,906 (137)	1,004 (32)	99,043 (313)
	Expectation(σ) SIEHS SNPs	89,059 (297)	16,301 (128)	870 (29)	84,593 (290)
(3,1)	Expectation(σ) ENCODE	40,706 (201)	3,427 (59)	44 (7)	37,047 (192)
	Expectation(σ)	32,377 (180)	2,871 (54)	37 (6)	30,805 (175)

	SIEHS SNPs				
	Expectation(σ) ENCODE	16,304 (128)	836 (29)	4 (2)	16,242 (127)
(4,1)	Expectation(σ) SIEHS SNPs	12,965 (114)	602 (25)	2 (1)	12,580 (112)
	Expectation(σ) ENCODE	260,931 (504)	52,513 (229)	3,064 (55)	236,932 (481)
(2,2)	Expectation(σ) SIEHS SNPs	209,271 (453)	44,407 (210)	2,063 (51)	198,637 (441)
	Expectation(σ) ENCODE	119,070 (343)	13,598 (117)	287 (17)	112,365 (333)
(3,2)	Expectation(σ) SIEHS SNPs	94,062 (305)	10,448 (102)	178 (13)	90,036 (299)
	Expectation(σ) ENCODE	56,742 (238)	2,904 (54)	10 (3)	51,472 (226)
(4,2)	Expectation(σ) SIEHS SNPs	43,817 (209)	2,383 (49)	9 (3)	41,921 (204)

3.5.2 Theoretical expected values assuming 150,000 variants in the exome under selected specifications

In Table 43, I report the expected numbers for a study considering only the exome under selected specifications, assuming 150,000 variants in the exome. In a comparison of a random sample of 100 to a random sample of 1500 from the same population, the expected number of variants appearing in the smaller group but not the larger decreases from about 400 expected under the specification (1,1) to about 0 under the specification (4,1). In the comparison of two groups of 800, the expected value of N_{+-} decreases from 4,500 under the specification (1,1) to about 200 under the specification (4,1).

Table 43

Number of variants in the exome expected to appear in $N_{+-}^{(r,h)}$ under selected specifications

(*r,h*)		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Expectation(σ) ENCODE	5,232 (71)	1,916 (43)	423 (21)	4,750 (68)
	Expectation(σ) SIEHS SNPs	4,499 (66)	1,745 (42)	388 (20)	4,270 (64)
(2,1)	Expectation(σ) ENCODE	1,635 (40)	284 (17)	15 (4)	1,486 (38)
	Expectation(σ) SIEHS SNPs	1,336 (36)	245 (16)	13 (4)	1,269 (35)
(3,1)	Expectation(σ) ENCODE	611 (25)	51 (7)	0.7 (0.8)	556 (24)
	Expectation(σ) SIEHS SNPs	486 (22)	43 (7)	0.6 (0.7)	462 (21)
(4,1)	Expectation(σ) ENCODE	245 (16)	13 (4)	0.1 (0.2)	244 (16)
	Expectation(σ) SIEHS SNPs	194 (14)	9 (3)	0 (0.2)	189 (14)
(2,2)	Expectation(σ) ENCODE	3,914 (62)	788 (28)	46 (7)	3,554 (59)
	Expectation(σ) SIEHS SNPs	3,139 (55)	666 (26)	39 (6)	2,980 (54)
(3,2)	Expectation(σ) ENCODE	1,786 (42)	204 (14)	4 (2)	1,685 (41)
	Expectation(σ) SIEHS SNPs	1,411 (37)	157 (13)	3 (2)	1,351 (37)
(4,2)	Expectation(σ) ENCODE	851 (29)	44 (7)	0.2 (0.4)	772 (28)
	Expectation(σ) SIEHS SNPs	657 (26)	36 (6)	0.1 (0.4)	629 (25)

3.5.3 Theoretical expected values in specific genes under selected specifications

I then consider the specifications applied at the gene level. Table 44 contains the expected numbers in gene FBN1 for the European population. The expected number decreases from 4 under the specification (1, 1) to about 0 under the specification (4, 1). These expectations suggest that the specification (2, 1) would be practical at the gene level.

Table 44

Number of variants in gene FBN1 expected to appear in $N_{+-}^{(r,h)}$ under selected specifications

(*r,h*)	FBN1: 1,301 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Expectation(σ) ENCODE	45 (7)	17 (4)	4 (2)	41 (6)
	Expectation(σ) SIEHS SNPs	39 (6)	15 (4)	3 (2)	37 (6)
(2,1)	Expectation(σ) ENCODE	14 (4)	2 (2)	0.1 (0.4)	13 (4)
	Expectation(σ) SIEHS SNPs	12 (3)	2 (1)	0.1 (0.3)	11 (3)
(3,1)	Expectation(σ) ENCODE	5 (2)	0.4 (0.7)	0 (0.1)	5 (2)
	Expectation(σ) SIEHS SNPs	4 (2)	0.4 (0.6)	0 (0.1)	4 (2)
(4,1)	Expectation(σ) ENCODE	2 (1)	0.1 (0.3)	0 (0)	2 (1)
	Expectation(σ) SIEHS SNPs	2 (1)	0.1 (0.3)	0 (0)	2 (1)
(2,2)	Expectation(σ) ENCODE	34 (6)	7 (3)	0.4 (0.6)	31 (5)
	Expectation(σ) SIEHS SNPs	27 (5)	6 (2)	0.3 (0.6)	26 (5)
(3,2)	Expectation(σ) ENCODE	15 (4)	2 (1)	0 (0.2)	15 (4)
	Expectation(σ) SIEHS SNPs	12 (3)	1 (1)	0 (0.2)	12 (3)
(4,2)	Expectation(σ) ENCODE	7 (3)	0.4 (0.6)	0 (0)	7 (3)
	Expectation(σ) SIEHS SNPs	6 (2)	0.3 (0.6)	0 (0)	5 (2)

Table 44
(Continued)

(*r*,*h*)	FBN1 (exon): 65 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	2 (1)	0.8 (0.9)	0.2 (0.4)	2 (1)
(1,1)	Expectation(σ) SIEHS SNPs	2 (1)	0.8 (0.8)	0.2 (0.4)	2 (1)
	Expectation(σ) ENCODE	0.7 (0.8)	0.1 (0.4)	0 (0.1)	0.6 (0.8)
(2,1)	Expectation(σ) SIEHS SNPs	0.6 (0.8)	0.1 (0.3)	0 (0.1)	0.5 (0.7)
	Expectation(σ) ENCODE	0.3 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)
(3,1)	Expectation(σ) SIEHS SNPs	0.2 (0.5)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
(4,1)	Expectation(σ) SIEHS SNPs	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
	Expectation(σ) ENCODE	2 (1)	0.3 (0.6)	0 (0.1)	2 (1)
(2,2)	Expectation(σ) SIEHS SNPs	1 (1)	0.3 (0.5)	0 (0.1)	1 (1)
	Expectation(σ) ENCODE	0.8 (0.9)	0.1 (0.3)	0 (0)	0.7 (0.8)
(3,2)	Expectation(σ) SIEHS SNPs	0.6 (0.8)	0.1 (0.3)	0 (0)	0.6 (0.8)
	Expectation(σ) ENCODE	0.4 (0.6)	0 (0.1)	0 (0)	0.3 (0.6)
(4,2)	Expectation(σ) SIEHS SNPs	0.3 (0.5)	0 (0.1)	0 (0)	0.3 (0.5)

Table 45 contains the expected numbers of N_{+-} for another gene, NF1, for the European population. The expected number decreases from 5 under the specification (1, 1) to about 0 under the specification (4, 1). These expectations suggest that the specification (2, 1) would be practical at the gene level for all comparisons.

Table 45

Number of variants in gene NF1 expected to appear in $N_{+-}^{(r,h)}$ under selected specifications

(*r,h*)	NF1: 1,659 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	58 (7)	21 (5)	5 (2)	53 (7)
(1,1)	Expectation(σ) SIEHS SNPs	50 (7)	19 (4)	4 (2)	47 (7)
	Expectation(σ) ENCODE	18 (4)	3 (2)	0.2 (0.4)	16 (4)
(2,1)	Expectation(σ) SIEHS SNPs	15 (4)	3 (2)	0.1 (0.4)	14 (4)
	Expectation(σ) ENCODE	7 (3)	0.6 (0.8)	0 (0.1)	6 (2)
(3,1)	Expectation(σ) SIEHS SNPs	5 (2)	0.5 (0.7)	0 (0.1)	5 (2)
	Expectation(σ) ENCODE	3 (2)	0.1 (0.4)	0 (0)	3 (2)
(4,1)	Expectation(σ) SIEHS SNPs	2 (1)	0.1 (0.3)	0 (0)	2 (1)
	Expectation(σ) ENCODE	43 (6)	9 (3)	0.5 (0.7)	39 (6)
(2,2)	Expectation(σ) SIEHS SNPs	35 (6)	7 (3)	0.4 (0.7)	33 (6)
	Expectation(σ) ENCODE	20 (4)	2 (2)	0 (0.2)	19 (4)
(3,2)	Expectation(σ) SIEHS SNPs	16 (4)	2 (1)	0 (0.2)	15 (4)
	Expectation(σ) ENCODE	9 (3)	0.5 (0.7)	0 (0)	9 (3)
(4,2)	Expectation(σ) SIEHS SNPs	7 (3)	0.4 (0.6)	0 (0)	7 (3)

Table 45

(Continued)

(*r,h*)	NF1 (exome): 57 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	2 (1)	0.7 (0.8)	0.2 (0.4)	2 (1)
(1,1)	Expectation(σ) SIEHS SNPs	2 (1)	0.7 (0.8)	0.1 (0.4)	2 (1)
(2,1)	Expectation(σ) ENCODE	0.6 (0.8)	0.1 (0.3)	0 (0.1)	0.6 (0.7)

	Expectation(σ) SIEHS SNPs	0.5 (0.7)	0.1 (0.3)	0 (0.1)	0.5 (0.7)
	Expectation(σ) ENCODE	0.2 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)
(3,1)	Expectation(σ) SIEHS SNPs	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
(4,1)	Expectation(σ) SIEHS SNPs	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
	Expectation(σ) ENCODE	1 (1)	0.3 (0.5)	0 (0.1)	1 (1)
(2,2)	Expectation(σ) SIEHS SNPs	1 (1)	0.3 (0.5)	0 (0.1)	1 (1)
	Expectation(σ) ENCODE	0.7 (0.8)	0.1 (0.3)	0 (0)	0.6 (0.8)
(3,2)	Expectation(σ) SIEHS SNPs	0.5 (0.7)	0.1 (0.2)	0 (0)	0.5 (0.7)
	Expectation(σ) ENCODE	0.3 (0.6)	0 (0.1)	0 (0)	0.3 (0.5)
(4,2)	Expectation(σ) SIEHS SNPs	0.3 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)

3.5.4 Empirical assessment of Chr1-22/500k SNP data using permutation modeling under selected specifications

3.5.4.1 Empirical assessment in the population under selected specifications

I calculate expectations of N_{+-} using the parameters for a European population (Ionnita-Laza et al. 2009) assuming a minimum variant frequency of 0.001 and without truncation. I examine 488,146 SNPs on Chromosome 1-22 with results for N_{+-} as shown in Table 46. The observed number of SNPs that appeared in a random sample

of 100 but not in a random sample of 1500 decreases from 559 under the specification (1, 1) to 0.1 under the specification (4,1). The counts are within the range of the expected counts.

Table 46

Number of variants in the genome appearing in $N^{(r,h)}_{+-}$ under selected specifications

(*r,h*)		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Average($\hat{\sigma}$)	9,146.4 (1,469)	2,394.3 (515)	558.6 (336)	6605.9 (1,596)
	Low	6,710	1,602	211	3,996
	High	12,386	3,131	999	9,757
	Expectation(σ) ENCODE*	16,652 (126)	2,894 (53)	547 (23)	9,122 (95)
	Expectation(σ) SIEHS SNPs*	20,366 (140)	3,635 (60)	689 (26)	11,387 (105)
	Expectation(σ) ENCODE	17,027 (128)	6,234(78)	1,377(37)	15,458 (122)
	Expectation(σ) SIEHS SNPs	14,642 (119)	5,678(75)	1,263(35)	13,894 (116)
(2,1)	Average($\hat{\sigma}$)	3,249.9 (589)	530.4 (173)	30.3 (20)	3,056.1 (1,009)
	Low	2,192	275	8	1,466
	High	4,150	787	53	4,531
	Expectation(σ) ENCODE*	7,431 (85)	881 (30)	44 (7)	5,096 (71)
	Expectation(σ) SIEHS SNPs*	8,837 (93)	1,098 (33)	55 (7)	6,314 (79)
	Expectation(σ) ENCODE	5,320 (73)	923 (30)	49 (7)	4,835 (69)
	Expectation(σ) SIEHS SNPs	4,347 (66)	596 (28)	42 (7)	4,129 (64)
(3,1)	Average($\hat{\sigma}$)	1,125.8 (206)	104.3 (43)	1.4 (2)	1,279.8 (509)
	Low	730	42	0	517
	High	1,374	179	7	2,035
	Expectation(σ) ENCODE*	3,003 (55)	217 (15)	3 (2)	2,434 (49)
	Expectation(σ)	3,553 (59)	267 (16)	3 (2)	2,986 (54)

	SIEHS SNPs*				
	Expectation(σ) ENCODE	1,987 (44)	167 (13)	2 (1)	1,808 (42)
	Expectation(σ) SIEHS SNPs	1,580 (40)	140 (12)	2 (1)	1,504 (39)
(4,1)	Average($\hat{\sigma}$)	402.5 (70)	21.9 (10)	0.1 (0.3)	513.3 (225)
	Low	269	6	0	186
	High	481	39	1	863
	Expectation(σ) ENCODE*	1,228 (35)	48 (7)	0.1 (0.4)	1,082 (33)
	Expectation(σ) SIEHS SNPs*	1,456 (38)	59 (8)	0.2 (0.4)	1,312 (36)
	Expectation(σ) ENCODE	796 (28)	41 (6)	0.2 (0.5)	793 (28)
	Expectation(σ) SIEHS SNPs	633 (25)	29 (5)	0.1 (0.3)	614 (25)
(2,2)	Average($\hat{\sigma}$)	7,082.7 (1,179)	1,700.1 (464)	103.3 (72)	7,616.1 (1,959)
	Low	5,792	1,062	28	4,466
	High	10,480	2,433	196	11,335
	Expectation(σ) ENCODE*	18,598 (134)	3,030 (55)	171 (13)	14,362 (118)
	Expectation(σ) SIEHS SNPs*	22,092 (145)	3,745 (61)	213 (15)	17,634 (130)
	Expectation(σ) ENCODE	12,737 (111)	2,563 (50)	150 (12)	11,566 (106)
	Expectation(σ) SIEHS SNPs	10,216 (100)	2,168 (46)	127 (11)	9,696 (97)
(3,2)	Average($\hat{\sigma}$)	3,209 (426)	403.5 (137)	5.8 (4)	3,700.8 (1,192)
	Low	2,365	232	1	1,804
	High	3,779	639	12	5,442
	Expectation(σ) ENCODE*	8,907 (94)	841 (29)	12 (3)	7,632 (87)
	Expectation(σ) SIEHS SNPs*	10,484 (101)	1,029 (32)	15 (4)	9,271 (95)
	Expectation(σ) ENCODE	5,812 (76)	664 (26)	14 (4)	5,485 (74)
	Expectation(σ) SIEHS SNPs	4,592 (67)	510 (26)	9 (3)	4,395 (66)

3.5.4.2 Empirical assessment in the exome under selected specifications

I next examine 3,342 SNPs in exonic regions on Chromosome 1-22 with results for N_{+-} as shown in Table 47. The observed number of SNPs that appear in a random sample of 100 but not in a random sample of 1500 decreases from 5 under the specification (1, 1) to 0 under the specification (4, 1), which are close to the expected counts. The number in one group but not the other decreases from 50 under the specification (1,1) to 4 under the specification (4, 1) in two random samples of 800.

Table 47

Number of variants in the exome appearing in $N_{+-}^{(r,h)}$ under selected specifications

(*r*,*h*)		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Average($\hat{\sigma}$)	74.7 (14)	17.2 (7)	4.6 (3)	51.5 (15)
	Low	51	6	0	27
	High	111	26	10	82
	Expectation(σ) ENCODE*	114 (10)	20 (4)	4 (2)	62 (8)
	Expectation(σ) SIEHS SNPs*	139 (12)	25 (5)	5 (2)	78 (9)
	Expectation(σ) ENCODE	117(11)	43(6)	9(3)	106(10)
	Expectation(σ) SIEHS SNPs	100(10)	39(6)	9(3)	95(10)
(2,1)	Average($\hat{\sigma}$)	26.3 (7)	3.3 (2)	0.4 (0.7)	21.2 (8)
	Low	19	0	0	10
	High	38	9	2	31
	Expectation(σ) ENCODE*	50 (7)	6 (2)	0.3 (0.6)	35 (6)
	Expectation(σ) SIEHS SNPs*	61 (8)	8 (3)	0.4 (0.6)	43 (7)
	Expectation(σ) ENCODE	36 (6)	6 (3)	0.3 (0.6)	33 (6)
	Expectation(σ)	30 (5)	5 (2)	0.3 (0.5)	28 (5)

	SIEHS SNPs				
(3,1)	Average($\hat{\sigma}$)	10.5 (3)	0.8 (0.6)	0.1 (0.3)	8.5 (4)
	Low	7	0	0	4
	High	15	2	1	14
	Expectation(σ) ENCODE*	21 (5)	1 (1)	0 (0.1)	17 (4)
	Expectation(σ) SIEHS SNPs*	24 (5)	2 (1)	0 (0.2)	20 (5)
	Expectation(σ) ENCODE	14 (4)	1 (1)	0 (0.1)	12 (4)
	Expectation(σ) SIEHS SNPs	11 (3)	1 (1)	0 (0.1)	10 (3)
(4,1)	Average($\hat{\sigma}$)	3.6 (2)	0.1 (0.3)	0 (0)	3.6 (2)
	Low	1	0	0	1
	High	6	1	0	8
	Expectation(σ) ENCODE*	8 (3)	0.3 (0.6)	0 (0)	7 (3)
	Expectation(σ) SIEHS SNPs*	10 (3)	0.4 (0.6)	0 (0)	9 (3)
	Expectation(σ) ENCODE	5 (2)	0.3 (0.5)	0 (0)	5 (2)
	Expectation(σ) SIEHS SNPs	4 (2)	0.2 (0.4)	0 (0)	4 (2)
(2,2)	Average($\hat{\sigma}$)	64.5 (12)	11.2 (4)	0.9 (1)	60.2 (15)
	Low	42	5	0	36
	High	87	19	4	88
	Expectation(σ) ENCODE*	127 (11)	21 (5)	1 (1)	98 (10)
	Expectation(σ) SIEHS SNPs*	151 (12)	26 (5)	1 (1)	121 (11)
	Expectation(σ) ENCODE	87 (9)	18 (4)	1 (1)	79 (9)
	Expectation(σ) SIEHS SNPs	70 (8)	15 (4)	0.9 (0.9)	66 (8)
(3,2)	Average($\hat{\sigma}$)	28 (7)	3.2 (1)	0.2 (0.4)	28.2 (10)
	Low	20	2	0	13
	High	38	6	1	45
	Expectation(σ) ENCODE*	61 (8)	6 (2)	0.1 (0.3)	52 (7)
	Expectation(σ) SIEHS SNPs*	72 (8)	7 (3)	0.1 (0.3)	63 (8)
	Expectation(σ)	40 (6)	5 (2)	0.1 (0.3)	38 (6)

	ENCODE				
	Expectation(σ) SIEHS SNPs	31 (6)	3 (2)	0.1 (0.2)	30 (5)

3.5.4.3 Empirical assessment in specific genes under selected specifications

I examine 73 SNPs in gene FBN1 on Chromosome 15 with results for N_{+-} as shown in Table 48. The observed numbers of SNPs that appear in a random sample of 100 but not in a random sample of 1500 are all close 0. The expected counts are also close to 0. The number in one group but not the other decreased from 3 under the specification (1, 1) to 0 under the specification (4, 1) in two random samples of 800.

Table 48

Number of variants in gene FBN1with 73 SNPs appearing in $N_{+-}^{(r,h)}$ under selected specifications

(*r,h*)		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Average($\hat{\sigma}$)	4.9 (3)	1.2 (2)	0.6 (1)	3.3 (3)
	Low	0	0	0	0
	High	10	4	3	12
	Expectation(σ) ENCODE*	2 (2)	0.4 (0.7)	0.1 (0.3)	1 (1)
	Expectation(σ) SIEHS SNPs*	3 (2)	0.5 (0.7)	0.1 (0.3)	2 (1)
	Expectation(σ) ENCODE	3 (2)	0.9 (1)	0.2 (0.5)	2 (1)
	Expectation(σ) SIEHS SNPs	2 (1)	0.8 (0.9)	0.2 (0.4)	2 (1)
(2,1)	Average($\hat{\sigma}$)	0.7 (1)	0.3 (0.7)	0 (0)	2.3 (3)
	Low	0	0	0	0
	High	3	2	0	9
	Expectation(σ) ENCODE*	1 (1)	0.1 (0.4)	0 (0.1)	0.8 (0.9)

	Expectation(σ) SIEHS SNPs*	1 (1)	0.2 (0.4)	0 (0.1)	0.9 (1)
	Expectation(σ) ENCODE	0.8 (0.9)	0.1 (0.4)	0 (0.1)	0.7 (0.8)
	Expectation(σ) SIEHS SNPs	0.7 (0.8)	0.1 (0.3)	0 (0.1)	0.6 (0.8)
(3,1)	Average($\hat{\sigma}$)	0.2 (0.4)	0 (0)	0 (0)	0.6 (1)
	Low	0	0	0	0
	High	1	0	0	0
	Expectation(σ) ENCODE*	0.4 (0.7)	0 (0.2)	0 (0)	0.4 (0.6)
	Expectation(σ) SIEHS SNPs*	0.5 (0.7)	0 (0.2)	0 (0)	0.4 (0.7)
	Expectation(σ) ENCODE	0.3 (0.5)	0 (0.2)	0 (0)	0.3 (0.5)
	Expectation(σ) SIEHS SNPs	0.2 (0.5)	0 (0.1)	0 (0)	0.2 (0.5)
(4,1)	Average($\hat{\sigma}$)	0 (0)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0
	High	0	0	0	0
	Expectation(σ) ENCODE*	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs*	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.2 (0.3)
	Expectation(σ) SIEHS SNPs	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.3)
(2,2)	Average($\hat{\sigma}$)	1.6 (1)	0.9 (1)	0 (0)	4 (2)
	Low	0	0	0	1
	High	4	3	0	9
	Expectation(σ) ENCODE*	3 (2)	0.5 (0.7)	0 (0.2)	2 (1)
	Expectation(σ) SIEHS SNPs*	3 (2)	0.6 (0.7)	0 (0.2)	3 (2)
	Expectation(σ) ENCODE	2 (1)	0.4 (0.6)	0 (0.1)	2 (1)
	Expectation(σ) SIEHS SNPs	2 (1)	0.3 (0.6)	0 (0.1)	1 (1)
(3,2)	Average($\hat{\sigma}$)	0.4 (0.5)	0.4 (0.5)	0 (0)	1.6 (1)
	Low	0	0	0	0
	High	1	1	0	3

	Expectation(σ) ENCODE*	1 (1)	0.1 (0.4)	0 (0)	1 (1)
	Expectation(σ) SIEHS SNPs*	2 (1)	0.2 (0.4)	0 (0)	1 (1)
	Expectation(σ) ENCODE	0.9 (0.9)	0.1 (0.3)	0 (0)	0.8 (0.9)
	Expectation(σ) SIEHS SNPs	0.7 (0.8)	0.1 (0.3)	0 (0)	0.7 (0.8)

I examine 16 SNPs in gene NF1 on Chromosome 17 with results as shown in Table 49. The observed numbers of SNPs that appear in one group but not in another are all close 0. The expected counts are also close to 0.

Table 49

Number of variants in gene NF1with 16 SNPs appearing in $N_{+-}^{(r,h)}$ under selected specifications

(*r*,*h*)		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
(1,1)	Average($\hat{\sigma}$)	0.3 (0.5)	0.2 (0.4)	0.1 (0.3)	0.5 (0.5)
	Low	0	0	0	0
	High	1	1	1	1
	Expectation(σ) ENCODE*	0.5 (0.7)	0.1 (0.3)	0.02 (0.1)	0.3 (0.5)
	Expectation(σ) SIEHS SNPs*	0.7 (0.8)	0.1 (0.3)	0.02 (0.2)	0.4 (0.6)
	Expectation(σ) ENCODE	0.6 (0.7)	0.2 (0.4)	0.05 (0.2)	0.5 (0.7)
	Expectation(σ) SIEHS SNPs	0.5 (0.7)	0.2 (0.4)	0.04 (0.2)	0.5 (0.7)
(2,1)	Average($\hat{\sigma}$)	0.1 (0.3)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0
	High	1	0	0	0
	Expectation(σ) ENCODE*	0.2 (0.5)	0 (0.2)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs*	0.3 (0.5)	0 (0.2)	0 (0)	0.2 (0.5)
	Expectation(σ) ENCODE	0.2 (0.4)	0 (0.2)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs	0.1 (0.4)	0 (0.2)	0 (0)	0.1 (0.4)
(3,1)	Average($\hat{\sigma}$)	0.1 (0.3)	0 (0)	0 (0)	0 (0)
	Low	0	0	0	0

	High	1	0	0	0
	Expectation(σ) ENCODE*	0.1 (0.3)	0 (0.1)	0 (0)	0 (0.3)
	Expectation(σ) SIEHS SNPs*	0.2 (0.3)	0 (0.1)	0 (0)	0 (0.3)
	Expectation(σ) ENCODE	0.1 (0.3)	0 (0.1)	0 (0)	0.1 (0.2)
	Expectation(σ) SIEHS SNPs	0.1 (0.2)	0 (0.1)	0 (0)	0 (0.2)
(4,1)	Average($\hat{\sigma}$)	0	0	0	0
	Low	0	0	0	0
	High	0	0	0	0
	Expectation(σ) ENCODE*	0 (0.2)	0 (0)	0 (0)	0 (0.2)
	Expectation(σ) SIEHS SNPs*	0 (0.2)	0 (0)	0 (0)	0 (0.2)
	Expectation(σ) ENCODE	0 (0.2)	0 (0)	0 (0)	0 (0.2)
	Expectation(σ) SIEHS SNPs	0 (0.1)	0 (0)	0 (0)	0 (0.1)
(2,2)	Average($\hat{\sigma}$)	0.25 (0.4)	0 (0)	0 (0)	0.1 (0.2)
	Low	0	0	0	0
	High	1	0	0	1
	Expectation(σ) ENCODE*	0.6 (0.8)	0.1 (0.3)	0 (0.1)	0.5 (0.7)
	Expectation(σ) SIEHS SNPs*	0.7 (0.8)	0.1 (0.3)	0 (0.1)	0.6 (0.7)
	Expectation(σ) ENCODE	0.4 (0.6)	0.1 (0.3)	0 (0.1)	0.4 (0.6)
	Expectation(σ) SIEHS SNPs	0.3 (0.6)	0.1 (0.3)	0 (0.1)	0.3 (0.6)
(3,2)	Average($\hat{\sigma}$)	0.2 (0.4)	0 (0)	0 (0)	0.1 (0.3)
	Low	0	0	0	0
	High	1	0	0	1
	Expectation(σ) ENCODE*	0.3 (0.5)	0 (0.2)	0 (0)	0.3 (0.5)
	Expectation(σ) SIEHS SNPs*	0.3 (0.6)	0 (0.2)	0 (0)	0.3 (0.5)
	Expectation(σ) ENCODE	0.2 (0.4)	0 (0.1)	0 (0)	0.2 (0.4)
	Expectation(σ) SIEHS SNPs	0.2 (0.4)	0 (0.1)	0 (0)	0.1 (0.4)

4. Discussion

My research objectives were:

1. Provide estimates for the total number of whole genome sequences that must be obtained from normal (control) populations, in order for inferences to be made about variants found in disease cohorts;
2. Where exonic variants are sought, provide estimates for the total number of exome sequences that must be obtained from normal (control) populations;
3. Where causal variants in a specific gene are sought, provide estimates for the total number of gene sequences that must be obtained from normal (control) populations.

For two randomly selected groups from the same population, I calculated the expected number of variants appearing in both groups (EN_{++}), the expected number appearing in the smaller but not the larger group (EN_{+-}), the expected number appearing in the larger but not the smaller group (EN_{-+}), and the expected number not appearing in either group (EN_{--}). I then calculated these expected values assuming that the distribution of variant frequency followed a beta distribution. I used the parameters estimated by Ionnita-Laza et al. (2009) and calculated expectations for four populations. I also calculated expectations for a specification that a variant appeared in a group when at least two members of the group had the variant and confirmed these

expectations with the SNP data from the FHS. I also gave the general process to calculate the expectations for the specification (r,h). This specification meant that a variant appeared in the smaller group when it appeared in at least *r* members and that a variant appeared in the larger group when it appeared in at least *h* members.

I confirmed empirically that the assumption of a beta distribution for the frequency of a variant was consistent with SNP data from the FHS when the frequency was truncated on the left at 0.001 with regard to the average number of SNPs appearing in the smaller group but not the larger. The variability of this number was larger than the estimates from the beta distribution model.

With regard to objective 1, which dealt with genome wide studies, EN_{+-} was extremely large. This suggested that this approach is not practical for a genome wide study. Increasing group sizes increased EN_{+-}. I examined the expectations for the specification that a variant is present in a group if it appears at least twice. These expected values are smaller but still extremely large. I considered in detail the subset of specifications in which a variant appeared in the smaller group when it appeared in at least *r* subjects and it appeared in the larger group when it appeared at least once. In a comparison of a group of 100 to a group of 1500, EN_{+-} was about 28,000: $EN_{+-}^{(2,1)}$ was about1,000; $EN_{+-}^{(3,1)}$ was about 45; $EN_{+-}^{(4,1)}$ was about 5; and $EN_{+-}^{(5,1)}$ was about 0.1. For this choice of group sizes, the expected number of false positives with the specification (4, 1) was manageable.

With regard to objective 2 (a study of the whole exome), in the model describing NGS (that is, assuming a beta distribution without truncation), EN_{--} was extremely large

when the sum of the group sizes was 1600. Specifically it was about 40% of the total variants. The variants not appearing in either group constituted a pool of variants that would appear in the groups if the sum of the sizes of the groups were increased.

In empirical assessment using SNPs with no limitation on frequency, the observed values of N_{++}, the number appearing in both random groups, was much greater than the estimates based on the beta distribution. Conversely, the observed counts of N_{+-}, N_{-+}, and N_{--}, the numbers appearing in one group but not the other and the number not appearing in either group, were much smaller than the estimates using the beta distribution assumptions. The estimated standard deviations for all the estimates were so small that the observed differences could not be explained by statistical variation within the assumed model. Some possible explanations for these differences included: (1) most of the variants from FHS appeared more frequently than average; (2) there existed correlations among the SNPs; (3) the estimated parameters were misleading; (4) the assumption of a beta distribution was incorrect. Under the assumption that the minimum frequency>0.001, among European population or in the exome among the European population, the observed counts of N_{++} and N_{--}, were usually slightly greater than the expected estimates. Conversely, the observed counts of N_{-+}, were smaller than the expected estimates. The observed counts of N_{+-} were within the range of the expected estimates. On specific genes such as FBN1 and NF1, the observed counts of N_{++} were smaller than expected, while the observed counts of N_{--} were greater than the expected, and the observed counts of N_{+-} and N_{-+} were closer to their expected values.

In conclusion, genome and exome wide studies to identify rare variants associated with a disease require careful choice of group sizes and of specifications. Poor choices lead to unmanageably large numbers of false positives. There are specifications, however, that have manageable expected numbers of false positives. In studies of specified genes, the expected number of false positives is manageable.

REFERENCES

1. Biesecker LG (JAN 2010). Exome sequencing makes medical genomics a reality. *NATURE GENETICS* , Volume: 42 Issue: 1 Pages: 13-14.

2. Bodmer W & Bonilla C (JUN 2008). Common and rare variants in multifactorial susceptibility to common diseases. NATURE *GENETICS* , Volume: 40 Issue: 6 Pages: 695-701.

3. Choi M, Scholl UI, Ji WZ, et al. (NOV 10 2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* , Volume: 106 Issue: 45 Pages: 19096-19101.

4. Collins F (SEP 2000). Medical and societal consequences of the human genome project. *JOURNAL OF MEDICAL GENETICS*, Volume: 37 Pages: S32-S32 Supplement: Suppl. 1 Meeting Abstract: SP62.

5. Conrad DF, Andrews TD, Carter NP, et al. (JAN 2006). A high-resolution survey of deletion polymorphism in the human genome. *NATURE GENETICS* , Volume: 38 Issue: 1 Pages: 75-81.

6. Conrad DF, Hurles ME (JUL 2007). The population genetics of structural variation. *NATURE GENETICS* , Volume: 39 Pages: S30-S36 Supplement: Suppl. 7.

7. Cupples LA, Heard-Costa N, Lee M, Atwood LD, and the Framingham Heart Study Investigators (DEC 2009). Genetics Analysis Workshop 16 Problem 2: the Framingham Heart Study Data. *BMC Proceedings* , 3 (Suppl 7): S3.

8. Dietz, HC; McIntosh, I; Sakai, LY, et al. (AUG 1993). 4 novel FBN1 mutations - significance for mutant transcript level and EGF-like domain calcium-binding in the pathogenesis of MARFAN-syndrome. *GENOMICS* , Volume: 17 Issue: 2 Pages: 468-475.

9. Donnelly P (DEC 11 2008). Progress and challenges in genome-wide association studies in humans. *NATURE* , Volume: 456 Issue: 7223 Pages: 728-731.

10. Duan JB, Wainwright MS, Comeron JM, et al. (FEB 1 2003). Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and

synthesis of the receptor. *HUMAN MOLECULAR GENETICS* , Volume: 12 Issue: 3 Pages: 205-216.

11. Edvardson S, Shaag A, Zenvirt S, et al. (FEB 12 2010). Joubert Syndrome 2 (JBTS2) in Ashkenazi Jews Is Associated with a TMEM216 Mutation. *AMERICAN JOURNAL OF HUMAN GENETICS* , Volume: 86 Issue: 2 Pages: 294-294.

12. ENCODE Project Consortium (Oct 2 2004). The ENCODE (ENCyclopedia Of DNA Elements) project. *SCIENCE* , Volume: 306 Issue: 5696 Pages: 636-40.

13. Birney, E; Stamatoyannopoulos, JA; Dutta, A, et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *NATURE* , Volume: 447 Issue: 7146 Pages: 799-816.

14. Feuk L, Carson AR, Scherer SW (FEB 2006). Structural variation in the human genome. *NATURE REVIEWS GENETICS* , Volume: 7 Issue: 2 Pages: 85-97.

15. Frayling TM (SEP 2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *NATURE REVIEWS GENETICS* , Volume: 8 Issue: 9 Pages: 657-662.

16. Frazer KA, Murray SS, Schork NJ, et al. (APR 2009). Human genetic variation and its contribution to complex traits. *NATURE REVIEWS GENETICS* , Volume: 10 Issue: 4 Pages: 241-251.

17. Greally JM (JUN 14 2007). Genomics - Encyclopaedia of humble DNA. *NATURE* , Volume: 447 Issue: 7146 Pages: 782-783.

18. Hammoud S, Emery BR, Aoki VW, et al. (SEP-OCT 2007). Identification of genetic variation in the 5 ' and 3 ' non-coding regions of the protamine genes in patients with protamine deregulation*. ARCHIVES OF ANDROLOGY* , Volume: 53 Issue: 5 Pages: 267-274.

19. Hinds DA, Kloek AP, Jen M, et al. (JAN 2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *NATURE GENETICS* , Volume: 38 Issue: 1 Pages: 82-85.

20. Iles MM (FEB 2008). What can genome-wide association studies tell us about the genetics of common disease? *PLOS GENETICS* , Volume: 4 Issue: 2 Article Number: e33.

21. Ionita-Laza I, Lange C, Laird NM (MAR 31 2009). Estimating the number of unseen variants in the human genome. *PROCEEDINGS OF THE NATIONAL*

ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA , Volume: 106 Issue: 13 Pages: 5008-5013.

22. JOHNSON MR, LOOK AT, DECLUE JE, et al. (JUN 15 1993). Inactivation of the NF1 gene in human-melanoma and neuroblastoma cell-lines without impaired regulation of GTP.RAS. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* , Volume: 90 Issue: 12 Pages: 5539-5543.

23. Kruglyak L, Nickerson DA (MAR 2001). Variation is the spice of life. *NATURE GENETICS* , Volume: 27 Issue: 3 Pages: 234-236.

24. Lander ES (OCT 25 1996). The new genomics: Global views of biology. *SCIENCE* , Volume: 274 Issue: 5287 Pages: 536-539.

25. Lewontin, Richard C (Jul-Aug 2005). The fallacy of racial medicine: confusions about human races. *GENEWATCH* , Volume: 18 Issue: 4 Pages: 5-7, 17.

26. Li BS, Leal SM (SEP 12 2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *AMERICAN JOURNAL OF HUMAN GENETICS* , Volume: 83 Issue: 3 Pages: 311-321.

27. Li, M; Cheng, TS; Ho, PWL, et al. (2009). -459C > T point mutation in 5' non-coding region of human GJB1 gene is linked to X-linked Charcot-Marie-Tooth neuropathy. *JOURNAL OF THE PERIPHERAL NERVOUS SYSTEM* , Volume: 14 Issue: 1 Pages: 14-21.

28. Li, YR; Vinckenbosch, N; Tian, G, et al. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *NATURE GENETICS* , Volume: 42 Issue: 11 Pages: 969-NIL_82.

29. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. (APR 1 2010). Whole-Genome Sequencing in a Patient with Charcot-Marie-Tooth Neuropathy. *NEW ENGLAND JOURNAL OF MEDICINE* , Volume: 362 Issue: 13 Pages: 1181-1191.

30. Madsen BE, Browning SR (FEB 2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLOS GENETICS* , Volume: 5 Issue: 2 Article Number: e1000384.

31. Maller J, George S, Purcell S, et al. (SEP 2006). Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *NATURE GENETICS* , Volume: 38 Issue: 9 Pages: 1055-1059.

32. McCarroll SA, Hadnott TN, Perry GH, et al. (JAN 2006). Common deletion polymorphisms in the human genome. *NATURE GENETICS* , Volume: 38 Issue: 1 Pages: 86-92.

33. McCarroll SA, Kuruvilla FG, Korn JM, et al. (OCT 2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *NATURE GENETICS* , Volume: 40 Issue: 10 Pages: 1166-1174.

34. McCarthy MI, Abecasis GR, Cardon LR, et al. (MAY 2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *NATURE REVIEWS GENETICS* , Volume: 9 Issue: 5 Pages: 356-369.

35. Mutsuddi M, Morris DW, Waggoner SG, et al. (NOV 2006). Analysis of high-resolution HapMap of DTNBP1 (dysbindin) suggests no consistency between reported common variant associations and schizophrenia. *AMERICAN JOURNAL OF HUMAN GENETICS* , Volume: 79 Issue: 5 Pages: 903-909.

36. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC (AUG 2008). Genetic variation in an individual human exome. *PLOS Genetics* , Volume: 4 Issue: 8 Article Number: e1000160.

37. Ng SB, Turner EH, Robertson PD, et al. (SEP 10 2009). Targeted capture and massively parallel sequencing of 12 human exomes. *NATURE* , Volume: 461 Issue: 7261 Pages: 272-U153.

38. Ng SB, Buckingham KJ, Lee C, et al. (JAN 2010). Exome sequencing identifies the cause of a mendelian disorder. *NATURE GENETICS* , Volume: 42 Issue: 1 Pages: 30-U41.

39. Pritchard JK (JUL 2001). Are rare variants responsible for susceptibility to complex diseases? *AMERICAN JOURNAL OF HUMAN GENETICS* , Volume: 69 Issue: 1 Pages: 124-137.

40. Ramamurthi KS, Schneewind O (JAN 2005). A synonymous mutation in Yersinia enterocolitica yopE affects the function of the YopE type III secretion signal. *JOURNAL OF BACTERIOLOGY* , Volume: 187 Issue: 2 Pages: 707-715.

41. Redon, R; Ishikawa, S; Fitch, KR, et al. (2006). Global variation in copy number in the human genome. *NATURE* , Volume: 444 Pages: 444-454.

42. Ropers, HH (JUN 20). X-linked mental retardation: many genes for a complex disorder. *CURRENT OPINION IN GENETICS & DEVELOPMENT* , Volume: 16 Issue: 3 Pages: 260-269.

43. Smyth G (SEPT 2007). The statmod package. *http://www.statsci.org/r* , License: LGPL version 2 or newer.

Appendix 1

Table A1
Number of variants in gene SYNE1 expected to appear in two samples among European population under specification (1,1)

	SYNE1: 3,318 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	1,950 (28)	2,023 (28)	1,789 (29)	2,077 (28)
	Expectation(σ) NIEHS SNPs	1,278 (28)	1,339 (28)	1,154 (27)	1,383 (28)
N_{+-}	Expectation(σ) ENCODE	116 (11)	42 (28)	9 (3)	105 (10)
	Expectation(σ) NIEHS SNPs	100 (10)	39 (6)	9 (3)	94 (10)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	221 (14)	486 (20)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		194 (14)	409 (19)	
N_{--}	Expectation(σ) ENCODE	1,136 (27)	1,031 (27)	1,031 (27)	1,031 (27)
	Expectation(σ) NIEHS SNPs	1,841 (29)	1,746 (29)	1,746 (29)	1,746 (29)

Table A1
(Continued)

	SYNE1 (exome): 145 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	85 (6)	88 (6)	78 (6)	91 (6)
	Expectation(σ) NIEHS SNPs	56 (6)	59 (6)	50 (6)	60 (6)
N_{+-}	Expectation(σ) ENCODE	5 (2)	2 (1)	0.4 (0.6)	5 (2)
	Expectation(σ) NIEHS SNPs	4 (2)	2 (1)	0.4 (0.6)	4 (2)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	10 (3)	21 (4)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		8 (6)	18 (4)	

	Expectation(σ) ENCODE	50 (6)	45 (6)	45 (6)	45 (6)
N_{--}	Expectation(σ) NIEHS SNPs	80 (6)	76 (6)	76 (6)	76 (6)

Table A2

Number of variants in gene HMCN1 expected to appear in two samples among European population under specification (1,1)

	HMCN1: 2,301 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	1,352 (24)	1,403 (23)	1,241 (24)	1,440 (23)
N_{++}	Expectation(σ) NIEHS SNPs	886 (23)	929 (24)	929 (24)	959 (24)
	Expectation(σ) ENCODE	80 (9)	29 (5)	6 (3)	73 (8)
N_{+-}	Expectation(σ) NIEHS SNPs	69 (8)	27 (5)	27 (5)	65 (8)
	Expectation(σ) ENCODE	Same as N_{+-}	153 (12)	339 (17)	Same as N_{+-}
N_{-+}	Expectation(σ) NIEHS SNPs		135 (11)	135 (11)	
	Expectation(σ) ENCODE	788 (23)	715 (22)	715 (22)	715 (22)
N_{--}	Expectation(σ) NIEHS SNPs	1,277 (24)	1,211 (24)	1,211 (24)	1,211 (24)

Table A2

(Continued)

	HMCN1 (exome): 107 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	63 (5)	65 (5)	58 (5)	67 (5)
N_{++}	Expectation(σ) NIEHS SNPs	41 (5)	43 (5)	37 (5)	45 (5)
	Expectation(σ) ENCODE	4 (2)	1 (1)	0.3 (0.5)	3 (2)
N_{+-}	Expectation(σ) NIEHS SNPs	3 (2)	1 (1)	0.3 (0.5)	3 (2)
	Expectation(σ) ENCODE	Same as N_{+-}	7 (3)	16 (4)	Same as N_{+-}
N_{-+}	Expectation(σ) NIEHS SNPs		6 (2)	13 (3)	
N_{--}	Expectation(σ) ENCODE	37 (5)	33 (5)	33 (5)	33 (5)

	Expectation(σ) NIEHS SNPs	59 (5)	56 (5)	56 (5)	56 (5)

Table A3

Number of variants in gene UBR4 expected to appear in two samples among European population under specification (1,1)

	UBR4: 628 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	369 (12)	383 (12)	339 (12)	393 (12)
N_{++}	Expectation(σ) NIEHS SNPs	242 (12)	253 (12)	218 (12)	262 (12)
	Expectation(σ) ENCODE	22 (5)	8 (3)	2 (1)	20 (4)
N_{+-}	Expectation(σ) NIEHS SNPs	19 (4)	7 (3)	2 (1)	18 (4)
	Expectation(σ) ENCODE	Same as N_{+-}	42 (6)	92 (9)	Same as N_{+-}
N_{-+}	Expectation(σ) NIEHS SNPs		37 (6)	77 (8)	
	Expectation(σ) ENCODE	215 (12)	195 (12)	195 (12)	195 (12)
N_{--}	Expectation(σ) NIEHS SNPs	348 (12)	331 (13)	331 (13)	331 (13)

Table A3

(Continued)

	UBR4 (exome) : 106 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Expectation(σ) ENCODE	62 (5)	65 (5)	57 (5)	66 (5)
N_{++}	Expectation(σ) NIEHS SNPs	41 (5)	43 (5)	37 (5)	44 (5)
	Expectation(σ) ENCODE	4(2)	1 (1)	0.3 (0.5)	3 (2)
N_{+-}	Expectation(σ) NIEHS SNPs	3 (2)	1 (1)	0.3 (0.5)	3 (2)
	Expectation(σ) ENCODE	Same as N_{+-}	7 (33)	16 (4)	Same as N_{+-}
N_{-+}	Expectation(σ) NIEHS SNPs		6 (2)	13 (3)	
	Expectation(σ) ENCODE	36 (5)	33 (5)	33 (5)	33 (5)
N_{--}	Expectation(σ) NIEHS SNPs	59 (5)	56 (5)	56 (5)	56 (5)

Table A4

Number of variants in gene RYR1 expected to appear in two samples among European population under specification (1,1)

	RYR1: 901 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	530 (15)	549 (15)	486 (15)	564 (15)
	Expectation(σ) NIEHS SNPs	347 (15)	364 (15)	313 (14)	375 (15)
N_{+-}	Expectation(σ) ENCODE	31 (6)	12 (3)	3 (2)	29 (5)
	Expectation(σ) NIEHS SNPs	27 (5)	10 (3)	2 (2)	26 (5)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	60 (14)	133 (10)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		53 (7)	111 (10)	
N_{--}	Expectation(σ) ENCODE	309 (14)	280 (14)	280 (14)	280 (14)
	Expectation(σ) NIEHS SNPs	500 (15)	474 (15)	474 (15)	474 (15)

Table A4

(Continued)

	RYR1 (exome): 106 SNPs	400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Expectation(σ) ENCODE	62 (5)	65 (5)	57 (5)	66 (5)
	Expectation(σ) NIEHS SNPs	41 (5)	43 (5)	37 (5)	44 (5)
N_{+-}	Expectation(σ) ENCODE	4(2)	1 (1)	0.3 (0.5)	3 (2)
	Expectation(σ) NIEHS SNPs	3 (2)	1 (1)	0.3 (0.5)	3 (2)
N_{-+}	Expectation(σ) ENCODE	Same as N_{+-}	7 (33)	16 (4)	Same as N_{+-}
	Expectation(σ) NIEHS SNPs		6 (2)	13 (3)	
N_{--}	Expectation(σ) ENCODE	36 (5)	33 (5)	33 (5)	33 (5)
	Expectation(σ) NIEHS SNPs	59 (5)	56 (5)	56 (5)	56 5)

Appendix 2

Table A5

Categorization of SNPs in gene SYNE1with 128 SNPs by appearance in two randomly selected groups under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Average	121	124	118	127
	Low	118	121	NA	NA
	High	123	127	NA	NA
	Expectation(σ) ENCODE*	116 (3)	120 (3)	106 (4)	122 (2)
	Expectation(σ) SIEHS SNPs*	113 (4)	118 (1)	102 (5)	121 (3)
	Expectation(σ) ENCODE	75 (6)	78 (6)	69 (6)	80 (5)
	Expectation(σ) SIEHS SNPs	49 (6)	52 (6)	45 (5)	53 (6)
N_{+-}	Average	3	0	0	0
	Low	0	0	NA	NA
	High	6	0	NA	NA
	Expectation(σ) ENCODE*	4 (2)	0.8 (0.9)	0.1 (0.4)	2 (2)
	Expectation(σ) SIEHS SNPs*	5 (2)	1 (1)	0.2 (0.4)	3 (2)
	Expectation(σ) ENCODE	4 (2)	2 (1)	0.4 (0.6)	4 (2)
	Expectation(σ) SIEHS SNPs	4 (2)	1 (1)	0.3 (0.6)	4 (2)
N_{-+}	Average	Same as N_{+-}	4	9	Same as N_{+-}
	Low		3	NA	
	High		6	NA	
	Expectation(σ) ENCODE*		7 (3)	20 (4)	
	Expectation(σ) SIEHS SNPs*		8 (3)	24 (4)	
	Expectation(σ) ENCODE		9 (3)	19 (4)	
	Expectation(σ) SIEHS SNPs		7 (3)	16 (4)	

	Average	2	1	1	1
	Low	1	NA	NA	NA
	High	5	NA	NA	NA
	Expectation(σ) ENCODE*	3 (2)	0.9 (0.9)	0.9 (0.9)	0.9 (0.9)
	Expectation(σ) SIEHS SNPs*	4 (2)	1(1)	1(1)	1(1)
	Expectation(σ) ENCODE	44 (5)	40 (5)	40 (5)	40 (5)
N_{--}	Expectation(σ) SIEHS SNPs	71 (6)	67 (6)	67 (6)	67 (6)

Table A6

Categorization of SNPs in gene HMCN1with 93 SNPs by appearance in two randomly selected groups under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Average	79	83	81	87
	Low	74	75	NA	NA
	High	87	88	NA	NA
	Expectation(σ) ENCODE*	84 (3)	87 (2)	77 (4)	89 (2)
	Expectation(σ) SIEHS SNPs*	82 (3)	85 (3)	74 (4)	88 (2)
	Expectation(σ) ENCODE	58 (5)	57 (5)	50 (5)	58 (5)
N_{++}	Expectation(σ) SIEHS SNPs	36 (5)	38 (5)	32 (5)	39 (5)
	Average	5	0.8	0	2
	Low	0	0	NA	2
	High	14	1	NA	2
	Expectation(σ) ENCODE*	3 (2)	0.6 (0.7)	0.1 (0.3)	2 (1)
	Expectation(σ) SIEHS SNPs*	4 (2)	0.7 (0.8)	0.1 (0.4)	2 (1)
	Expectation(σ) ENCODE	3 (2)	2 (1)	0.3 (0.5)	3 (2)
N_{-+}	Expectation(σ) SIEHS SNPs	3 (2)	1 (1)	0.2 (0.5)	3 (2)
	Average		7	10	
	Low		2	NA	
N_{-+}	High	Same as	15	NA	

	Expectation(σ) ENCODE*	N_{+-}	5 (2)	15 (4)	Same as N_{+-}
	Expectation(σ) SIEHS SNPs*		6 (2)	18 (4)	
	Expectation(σ) ENCODE		6 (2)	14 (3)	
	Expectation(σ) SIEHS SNPs		5 (2)	11 (3)	
	Average	5	2	2	2
	Low	3	NA	NA	NA
	High	9	NA	NA	NA
	Expectation(σ) ENCODE*	2 (2)	0.6 (0.8)	0.6 (0.8)	0.6 (0.8)
	Expectation(σ) SIEHS SNPs*	3 (2)	0.8 (0.9)	0.8 (0.9)	0.8 (0.9)
	Expectation(σ) ENCODE	29 (4)	29 (4)	29 (4)	29 (4)
N_{--}	Expectation(σ) SIEHS SNPs	52 (5)	49 (5)	49 (5)	49 (5)

Table A7

Categorization of SNPs by appearance in two randomly selected groups in gene UBR4 with 211SNPs under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
	Average	205	205	202	206
	Low	203	204	NA	NA
	High	206	206	NA	NA
	Expectation(σ) ENCODE*	191 (4)	197 (4)	176 (5)	202 (3)
	Expectation(σ) SIEHS SNPs*	187 (5)	194 (4)	169 (6)	199 (3)
	Expectation(σ) ENCODE	124 (7)	129 (7)	114 (7)	132 (7)
N_{++}	Expectation(σ) SIEHS SNPs	81 (7)	85 (7)	73 (7)	88 (7)
	Average	2	1	0	2
	Low	0	0	NA	0
	High	6	4	NA	4
	Expectation(σ) ENCODE*	7 (3)	1 (1)	0.2 (0.5)	4 (2)
	Expectation(σ) SIEHS SNPs*	9 (3)	2 (1)	0.3 (0.5)	5 (2)
N_{+-}	Expectation(σ) ENCODE	7 (3)	3 (2)	0.6 (0.8)	7 (3)

	Expectation(σ) SIEHS SNPs	6 (2)	2 (2)	0.5 (0.7)	6 (2)
N_{-+}	Average	Same as N_{+-}	4	8	Same as N_{+-}
	Low		0	NA	
	High		6	NA	
	Expectation(σ) ENCODE*		11 (3)	34 (5)	
	Expectation(σ) SIEHS SNPs*		14 (4)	40 (6)	
	Expectation(σ) ENCODE		14 (4)	31 (5)	
	Expectation(σ) SIEHS SNPs		12 (3)	26 (5)	
N_{--}	Average	3	1	1	1
	Low	1	NA	NA	NA
	High	5	NA	NA	NA
	Expectation(σ) ENCODE*	5 (2)	1 (1)	1 (1)	1 (1)
	Expectation(σ) SIEHS SNPs*	7 (3)	2 (1)	2 (1)	2 (1)
	Expectation(σ) ENCODE	72 (7)	66 (7)	66 (7)	66 (7)
	Expectation(σ) SIEHS SNPs	117 (7)	111 (7)	111 (7)	111 (7)

Table A8

Categorization of SNPs by appearance in two randomly selected groups in gene RYR1with 18 SNPs under specification (1,1)

		400 vs 400	400 vs 1200	100 vs 1500	800 vs 800
N_{++}	Average	18	18	18	18
	Low	17	17	NA	NA
	High	18	18	NA	NA
	Expectation(σ) ENCODE*	16 (1)	17 (1)	15 (2)	17 (0.9)
	Expectation(σ) SIEHS SNPs*	16 (1)	17 (1)	14 (2)	17 (1)
	Expectation(σ) ENCODE	11 (2)	11 (2)	8 (2)	11 (2)
	Expectation(σ) SIEHS SNPs	7 (2)	7 (2)	6 (2)	8 (2)
N_{+-}	Average	0.3	0	0	0
	Low	0	0	NA	NA

	High	1	0	NA	NA
	Expectation(σ) ENCODE*	0.6 (0.8)	0.1 (0.3)	0.02 (0.1)	0.3 (0.6)
	Expectation(σ) SIEHS SNPs*	0.8(0.8)	0.1 (0.4)	0.03 (0.2)	0.4 (0.6)
	Expectation(σ) ENCODE	0.6 (0.8)	0.2 (0.5)	0.05 (0.2)	0.6 (0.7)
	Expectation(σ) SIEHS SNPs	0.5 (0.7)	0.2 (0.5)	0.05 (0.2)	0.5 (0.7)
N_{-+}	Average	Same as N_{+-}	0.3	0	Same as N_{+-}
	Low		0	NA	
	High		1	NA	
	Expectation(σ) ENCODE*		1 (0.9)	3 (2)	
	Expectation(σ) SIEHS SNPs*		1 (1)	3 (2)	
	Expectation(σ) ENCODE		1 (1)	3 (2)	
	Expectation(σ) SIEHS SNPs		1 (1)	2 (1)	
N_{--}	Average	0	0	0	0
	Low	0	0	NA	NA
	High	0	0	NA	NA
	Expectation(σ) ENCODE*	0.5 (0.7)	0.1 (0.4)	0.1 (0.4)	0.1 (0.4)
	Expectation(σ) SIEHS SNPs*	0.6 (0.7)	0.2 (0.4)	0.2 (0.4)	0.2 (0.4)
	Expectation(σ) ENCODE	6 (2)	6 (2)	6 (2)	6 (2)
	Expectation(σ) SIEHS SNPs	10 (2)	9 (2)	9 (2)	9 (2)