Using Affected Sib-Pairs to Uncover Rare Disease Variants

Hervé Perdry^{a, b} Bertram Müller-Myhsok^c Françoise Clerget-Darpoux^d

^aUniversité Paris-Sud, and ^bInserm U669, Hôpital Paul Brousse, Villejuif, France; ^cStatistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany; ^dInserm U781, Hôpital Necker – Enfants Malades, Paris, France

Key Words

Methodology · Rare variants · Family data · Sib-pairs

Abstract

Objective: We propose a new test for rare variant mapping, based on an affected sib-pair sample and a control sample. In each sib-pair, only the index case needs to be sequenced, and the number of alleles shared identical-by-descent between the sibs is used as complementary information. The test makes use of both association and linkage information. We compare this test to the Armitage test on case-control data, with cases either from the general population of cases or from a sample of cases having an affected sib. **Methods:** A score test based on the likelihood in a multiplicative risk model is proposed. Its power is estimated by simulations and compared to Armitage test's power. **Results:** The affected sib-pairs design allows a tremendous gain of power over the case-control design (from 1 to 99% for a moderate sample size and relative risk values around 3, at an α level of 10^{-11}). When cases are ascertained in a sample of cases having an affected sib, the use of linkage information in our test allows a gain of power of more than 20% in certain situations. **Conclusion:** We demonstrate the interest in using familial data and both association and linkage information for rare variant mapping.

Copyright © 2013 S. Karger AG, Basel

Introduction

As the advances of technology made it possible, the genetic epidemiology community progressively moved from the study of monogenic diseases to complex diseases, for which the number and nature of genetic factors are a priori unknown. For several years, the strategic choice for the search of genetic risk factors involved in those diseases has been the genome-wide linkage study. Many susceptibility factors have been mapped and published, but most often replication has failed in further samples. For example, in 1998 Concannon et al. [1] used a sample of more than 600 affected sib-pairs (ASP) to test linkage of the 15 susceptibility loci that had been mapped so far for insulin-dependent diabetes mellitus (IDDM), which were named IDDM1 through IDDM15. Apart from the well-established HLA region (IDDM1), there was only one region (and not a previously reported one) for which the linkage test was significant.

This lack of power to detect linkage or to replicate it can be explained by the model underlying the factors involved in this disease. Indeed, the power to detect linkage is model dependent [2–4] and, in addition, the identity-by-descent (IBD) sharing statistics may be highly variable from one sample to another [5]. In a multiplicative model, the power of a linkage study depends both on the odds

Table 1. Notations for ASP and controls

	<i>IBD</i> = 2	<i>IBD</i> = 1	<i>IBD</i> = 0	
$G = 0$	n_{02}	n_{01}	n_{00}	m_0
$G = 1$	n_{12}	n_{11}	n_{10}	m_1
$G = 2$	n_{22}	n_{21}	n_{20}	m_2
<i>n</i> denotes ASP; <i>m</i> denotes controls.				

ratios (OR) and on the variant allele frequencies in the sample. This explains for example why the insulin-linked polymorphic region ILPR (also called IDDM2) can be detected by association study but not by linkage study: in the data by Bell et al. [6] on IDDM in the North American population, the class 1 allele (M1) of the variable number tandem repeats (VNTR) in ILPR has a frequency of 0.70 and, under a multiplicative model, an OR of 3.4 is estimated for this allele. A case-control association study has very good power to detect this effect, but there is virtually no power to detect it by linkage in ASP as the IBD distribution expected with these parameters hardly differs from the distribution under the hypothesis of no effect (OR = 1). On the other hand, had the allele frequency been 0.10, the same OR would have lead to an ASP linkage study with high power.

One of the few cases where the genome-wide linkage strategy was successful is the location of the first identified susceptibility gene (NOD2) on chromosome 16q in Crohn's disease [7]. The causal single nucleotide polymorphisms (SNPs) of NOD2 have a low allele frequency with high GRR for both homozygote and heterozygote composites, which gives enough statistical power for the linkage analysis to be successful.

Given the scarcity of positive results, the scientific community questioned the relevance of this approach. In particular, Risch and Merikangas [8] emphasized the poor power of linkage analysis for a multiplicative model with high allele frequency and low OR, and promoted an alternative approach – the genome-wide association analysis: ‘the future of the genetics of complex diseases is likely to require large-scale testing by association analysis’. The HapMap project [9–11] and the discovery of millions of SNPs throughout the human genome, together with the development of the bead and chip technology, have made it feasible to type hundreds of thousands of markers on huge samples of patients and controls. Thanks to the extensive linkage disequilibrium (LD) that exists, genome-wide association studies have been extremely

successful in discovering genetic variants associated with multifactorial diseases [12].

However, that approach designed for the detection of common variants is not appropriate for the identification of rare variants (those for which the risk allele has a frequency <5%) [13]. Consequently, with the advance of next-generation sequencing, whole-genome or whole-exome sequencing in samples of unrelated cases and controls is now advocated, and various testing strategies have been proposed [14–24].

Some authors have investigated the use of familial data and of linkage information for enrichment of rare variants [25–29]; some new methods using familial data have been proposed [30, 31]. In this paper, we propose a study design based on a sample of controls and a sample of ASP. Sequencing is performed in controls and in one of the affected sibs (denoted index case) of each sib-pair. In addition, the number of alleles shared IBD for each sib-pair will be determined either by sequencing or lower density genotyping of the other sib. This design allows us to formulate a test making use both of the association information, through contrasting the control genotypes with those of the index case, and of the linkage information through the IBD, conditional to the index genotype. This strategy will be compared to the classical case-control strategy.

Methods

We propose a test making use of both genotypic information and IBD information in ASP. We will compare this test with a classical case-control test (the Armitage trend test), with cases ascertained either from the general population of cases or from a sample of cases having an affected sib.

We consider a diallelic locus A/a. The 3 genotypes AA, Aa and aa are denoted in the sequel by 0, 1 and 2. We study a sample of n ASP together with a sample of m controls. In the sequel, m is fixed to $m = 1,000$, limiting the study to the case where a fixed control panel is used as a reference, and n varies from $n = 100$ to $n = 1,000$.

In each ASP, let us denote the genotype of the index case by G ($G = 0, 1$ or 2) and the number of alleles shared IBD by the 2 sibs at the locus under consideration by IBD ($IBD = 2, 1$ or 0). For $k = 0, 1, 2$ and $i = 2, 1, 0$, we denote the number of sib-pairs in which the index genotype is k and the number of IBD alleles is i by n_{ki} . The sample of controls is similarly composed of m_0, m_1, m_2 controls of genotypes $k = 0, 1, 2$. These notations are summarized in table 1.

In practice, the test will be applied on the whole genome, testing one rare variant at a time. In order to obtain the IBD status of each sib-pair, one will need a low-density genotyping of each member in the surrounding region. The IBD probabilities can be then estimated using for example the methods described in Lander and Green [32] or Han and Abney [33]. The IBD status can then

be chosen as the status with maximum probability. In this study, we will not address the impact of a possible misattribution of the IBD status.

Model and Likelihood

Let q denote the frequency of allele a in the general population; the frequency of allele A is $p = 1 - q$. We assume that the Hardy-Weinberg equilibrium holds in the control population: for a control of genotype G , the probabilities $Q_k = \mathbb{P}(G = k)$ are:

$$Q_1 = p^2, Q_2 = 2pq, Q_3 = q^2.$$

We now consider a sib-pair, the members of which are distinguished by indices 1 (for the index case) and 2 (for the other sib). Their genotypes are denoted by G_1 and G_2 , and Aff_1 and Aff_2 denote the event that they are affected. The event ASP is defined by $\text{ASP} = \text{Aff}_1$ and Aff_2 .

For an individual with genotype $k = 0, 1$ or 2 , the probability of being affected is denoted by $f_k = \mathbb{P}(\text{Aff}|G = k)$, and for an individual with genotype $k = 0, 1$ or 2 , who has an affected sib, the probability of being affected is denoted by $f'_k = \mathbb{P}(\text{Aff}_2|\text{Aff}_1, G_2 = k)$.

We assume a multiplicative model, in which the following relations hold:

$$\begin{aligned} f_1 &= rf_0 & f_2 &= r^2 f_0 \\ f'_1 &= rf'_0 & f'_2 &= r^2 f'_0 \end{aligned} \quad (1)$$

The number of alleles shared IBD by the 2 sibs is denoted by $\text{IBD} \in \{0, 1, 2\}$. Under this model, we can compute the probabilities

$$P_{ki} = \mathbb{P}(G_1 = k, \text{IBD} = i|\text{ASP}).$$

The details of this computation and the resulting formulas are given in the Appendix.

Given these observations, we can write the log-likelihood of the parameters (r, q) as:

$$L(r, q) = \sum_{k,i} n_{ki} \log P_{ki} + \sum_k m_k \log Q_k. \quad (2)$$

Score Test

Our test statistics for the null hypothesis $H_0: r = 1$ is based on the score

$$U = \left(\frac{\partial L}{\partial r} \right) (\hat{q}, 1)$$

where $\hat{q}$ is the maximum likelihood estimator of q under H_0 , which is trivially

$$\hat{q} = \frac{\sum_i n_{1i} + 2 \sum_i n_{2i} + m_1 + 2m_2}{2n + 2m}.$$

We show in the Appendix that U can be written as

$$U = U_1 \hat{q} + U_0 \quad (3)$$

where

$$\begin{aligned} U_1 &= -4n_{02} - 3n_{01} - 2n_{00} \\ &\quad - 4n_{12} - 3n_{11} - 2n_{10} \\ &\quad - 4n_{22} - 3n_{21} - 2n_{20} \end{aligned} \quad (4)$$

that is $U_1 = -\sum_{k,i} (2 + i)n_{ki}$ and

$$\begin{aligned} U_0 &= +2n_{12} + \frac{3}{2}n_{11} + n_{10} \\ &\quad + 4n_{22} + 3n_{21} + 2n_{20} \end{aligned} \quad (5)$$

that is $U_0 = 1/2 \sum_{k,i} (2 + i)kn_{ki}$.

Note that under H_0 , the expected value of U_1 is $-3n$ and the expected value of U_0 is $3nq$. Asymptotically, the variance of U is

$$\sigma_U^2 = \frac{1}{4} \frac{(1-q)qn(19m+n)}{n+m},$$

which we estimate by

$$\hat{\sigma}_U^2 = \frac{1}{4} \frac{(1-\hat{q})\hat{q}n(19m+n)}{n+m}.$$

We use as a test statistic

$$T = \frac{U}{\sqrt{\hat{\sigma}_U^2}}. \quad (6)$$

It is easy to show that for $r > 1$, the expected value of U is positive, which allows to produce a one-sided test for $r = 1$ versus $r > 1$ (note that this makes sense here because we look for rare causative variants; for frequent variants, a one-sided test would not be appropriate, unless previous information on the at-risk allele is available). It is known from the classical theory of likelihood that the asymptotic distribution of T under H_0 is a standard normal distribution $\mathcal{N}(0, 1)$.

Case-Control Design and the Armitage Test

In the case-control design, the probabilities $\mathbb{P}(G = k|\text{Aff})$ are readily obtained by an application of Bayes' theorem:

$$\mathbb{P}(G = 0|\text{Aff}) = \frac{p^2}{p^2 + 2pqr + q^2r^2},$$

$$\mathbb{P}(G = 1|\text{Aff}) = \frac{2pqr}{p^2 + 2pqr + q^2r^2},$$

$$\mathbb{P}(G = 2|\text{Aff}) = \frac{q^2r^2}{p^2 + 2pqr + q^2r^2}.$$

These probabilities will be used to perform simulations for the case-control design under the alternative hypothesis H_1 .

We use Armitage's test [34] for trend on case-control data, as defined in Morris and Zeggini [13] and advocated in Sasieni [35], with scores 0, 1, 2 as appropriate for the multiplicative model. We denote the test statistic by A , which can be computed as:

$$A = \frac{m(n_2 + 2n_3) - n(m_2 + 2m_3)}{\sqrt{\frac{nm}{n+m}((n_2 + m_2)(n_1 + m_1 + n_3 + m_3) + 4(n_1 + m_1)(n_3 + m_3))}}$$

where n_k and m_k are the number of cases and controls with genotype k (for $k = 0, 1$ or 2), respectively, and $n = \sum_k n_k$ and $m = \sum_k m_k$ are the total number of cases and controls, respectively.

Computation of p Values

Under H_0 , both our test statistic T and Armitage's test statistic A are asymptotically normally distributed. However, when q is small, for reasonable sample sizes ($n < 1,000$), we are not under

Table 2. Expected number of individuals for each genotype with a total sample size of $n = 500$

	$r = 1$	$r = 2.5$ (CC)	$r = 2.5$ (ASP)
$G = 0$	490.05	475.67	458.70
$G = 1$	9.90	24.02	40.41
$G = 2$	0.05	0.30	0.89

Column 1 = under H_0 ; CC (column 2) = cases from the general population; ASP (column 3) = index cases from a sample of ASP.

Table 3. 500 ASP, under H_0 ($r = 1$)

	IBD = 2	IBD = 1	IBD = 0	Total
$G_1 = 0$	122.5	245.0	122.5	490.05
$G_1 = 1$	2.5	5.0	2.5	9.90
$G_1 = 2$	0.0	0.0	0.0	0.05
Total	125.00	250.00	125.00	

Table 4. 500 ASP, $r = 2.5$

	IBD = 2	IBD = 1	IBD = 0	Total
$G_1 = 0$	113.0	229.3	116.4	458.70
$G_1 = 1$	14.3	20.3	5.9	40.41
$G_1 = 2$	0.5	0.4	0.1	0.89
Total	127.7	250.0	122.3	

asymptotic conditions, which can be seen on a qq-plot obtained from 10^6 p values simulated under H_0 (such a simulation is easily done by random sampling from multinomial distributions to generate contingency tables under H_0 , then compute the statistics and the associated p values).

We thus use importance sampling to obtain accurate empirical cumulative density functions $F_{n,m,q}^T$ and $F_{n,m,q}^A$ of T and A , respectively, under H_0 (more details are given in the Appendix). If q was known, the p value associated to a value of $T = t$ and $A = a$, respectively, would be $p = 1 - F_{n,m,q}^T(t)$ and $1 - F_{n,m,q}^A(a)$. If q is unknown, we simply replace it by its estimation $\hat{q}$, defining the p value to be $p = 1 - F_{n,m,\hat{q}}^T(t)$ and $1 - F_{n,m,\hat{q}}^A(a)$, respectively.

To check empirically that the p values computed by this procedure are uniformly distributed, when computed under H_0 , we again generate samples under H_0 , compute the p values and verify the uniformity by a qq-plot of the logarithms of the p values.

Power

Simulation studies were used to assess the power of the 3 tests and designs considered:

- (1) our test using controls and ASP, designed by T;
- (2) the Armitage test using controls and index cases from ASP, designed by A (ASP);
- (3) the Armitage test using controls and cases from the general cases population, designed by A (CC).

In a first set of simulations, the number of controls in the sample is fixed to $m = 1,000$ (reference panel), and the rare allele frequency is fixed to $q = 0.01$. For values of n (the number of ASP or cases) ranging from 100 to 1,000 in steps of 25, and values of r (the relative risk) ranging from 1 to 4 in steps of 0.1, we simulated 10^6 samples of ASP and controls or cases and controls, with random sampling from multinomial distributions with parameters according to the models described above. An additional set of simulations was generated for q varying from 0.001 to 0.1 (31 values uniformly spaced on a logarithmic scale), with $n = m = 1,000$, and $r = 2, 3$ and 4.

For each sample, the test statistics T and A are computed as well as the resulting p values. The power for a given α value is estimated by the proportion of p values below α . All computations were done in R.

Results

Expected Values under the Null and Alternative Hypotheses

Table 2 provides the expected number of patients with genotype $G = 0$, $G = 1$ and $G = 2$, respectively, in a sample of $n = 500$ when $q = 0.01$, $r = 1$ (H_0) and $r = 2.5$ (H_1). The first column ($r = 1$) corresponds to the genotypic distribution in the general population, the second column ($r = 2.5$ (CC)) gives the expectation for 500 patients randomly ascertained in the general population, while in the third column ($r = 2.5$ (ASP)), the patients have an affected sib. This table shows a higher proportion of variant carrier in the index cases from ASP (8%) than in the population of random patients (5%) and illustrates the advantage of ascertaining ASP well. It can be shown without difficulty that in our model this enrichment in causative variants in index cases from ASP will be observed as soon as $r > 1$, independently of the value of q .

To cast some light on the complementary information provided by the IBD, we display in tables 3 and 4 the expected numbers of ASP in each category $G_1 = k$, $IBD = i$, $q = 0.01$ and $r = 1$ or $r = 2.5$. Whereas when $r = 2.5$ the IBD sharing proportions (0.26, 0.50, 0.24) of the whole set of ASPs hardly differ from the proportions (0.25, 0.50, 0.25) expected under H_0 ($r = 1$), there is a big difference in this distribution in the sharing of pairs in which the index case is a variant carrier: for $G_1 = 1$ (0.35, 0.50, 0.15) and $G_1 = 2$ (0.51, 0.41, 0.08), respectively. This illustrates the important information provided by the IBD sharing conditionally on the index genotype. Table 4 shows that

Fig. 1. Empirical quantiles of p values for $n = 100$, $m = 1,000$, and $q = 0.01$, computed using the normal distribution (plot uses a simulation set of 10^6 samples under H_0 , **a**) and computed using our procedure (plot uses a simulation set of 10^6 samples with an importance sampling procedure, allowing to estimate small quantiles, **b**). Logarithmic scales are used.

Fig. 2. Power of the 3 tests and designs, with $\alpha = 10^{-7}$, $q = 0.01$, n ranging from 100 to 1,000, and $r = 2, 2.5$ and 3.5 for **a–c**, respectively.

the advantage of our statistics is the simultaneous use of association and linkage information.

Uniformity of p Values under the Null Hypothesis

Figure 1a displays a qq-plot of p values computed assuming the asymptotic normal distribution, using 10^6 statistics values simulated under H_0 for $n = 100$, $m = 1,000$ and $q = 0.01$; a logarithmic scale is used, so that it can be interpreted as a qq-plot of $-\log_{10}(p)$. An excess of low p

values appears for both our statistics T and Armitage's A : for example, p values lower than 10^{-4} are in proportion $1.4 \cdot 10^{-3}$, which is 14 times the expected value. Thus, using the asymptotic distribution to compute the p values would lead to an anti-conservative test.

As described in the Methods section, we compute the empirical cumulative distributions $F_{n,m,q}^T$ and $F_{n,m,q}^A$ of T and A using 10^6 weighted statistics values obtained by importance sampling, for $n = 100$, $m = 1,000$, and q ranging

Fig. 3. Power of the 3 tests and designs, with $\alpha = 10^{-7}$, $q = 0.01$, r ranging from 1 to 4, and $n = 200, 500$ and $1,000$ for **a–c**, respectively.

Fig. 4. Power of the 3 tests and designs, with $\alpha = 10^{-7}$, $n = 1,000$, q ranging from 0.001 to 0.1 (a logarithmic scale was used for q), and $r = 2, 3$ and 4 for **a–c**, respectively.

from 0.001 to 0.1. Figure 1b displays a qq-plot of p values computed using these distributions (assuming again $n = 100$, $m = 1,000$ and $q = 0.01$). Importance sampling was used here to calculate statistics values because we want to verify that the p values computed by our procedure do not lead to a biased test even for very low α thresholds. In figure 1b, we see that the qq-plot is very close to the diagonal, until the quantile level of 10^{-20} , from which we conclude that our procedure gives an unbiased test until α levels as low as 10^{-20} , which is much lower than needed for genome-wide multiple testing correction.

Power

We report here the power of the 3 tests and designs considered (as defined in the Methods section, designed by T, A (ASP) and A (CC)). The power is assessed for 2 α levels: $\alpha = 10^{-7}$ (fig. 2–4), which would be suitable for studying a large genomic region, and $\alpha = 10^{-11}$ (fig. 5–7), which corresponds to a Bonferroni correction for 5 million variants (for a family-wise error rate of 0.05), which is sufficient for a genome-wide screening. At both levels, there is a huge gain of power in the ASP designs over the case-control design. For example, for $q = 0.01$ at an α lev-

Fig. 5. Power of the 3 tests and designs, with $\alpha = 10^{-11}$, $q = 0.01$, n ranging from 100 to 1,000, and $r = 2, 2.5$ and 3.5 for **a–c**, respectively.

Fig. 6. Power of the 3 tests and designs, with $\alpha = 10^{-11}$, $q = 0.01$, r ranging from 1 to 4, and $n = 200, 500$ and $1,000$ for **a–c**, respectively.

el of 10^{-11} , the power of both ASP designs exceeds 99% for $n = 500$ and $r = 3.5$, while the power of the case-control design with the Armitage test is roughly 1%.

At first sight, the difference in power between the 2 ASP designs, i.e. our test T and the Armitage test A (ASP), can seem minor; however, the high/steep curves in figures 3 and 6 can be misleading: although the 2 curves are very close to one another, the difference in power for given values of r and n can exceed 20%. This can be seen in figure 5b: for $n = 1,000$ and $r = 2.5$, our test T has a power of 68%, whereas the Armitage test A (ASP) has a power

of 45%. Figures 4 and 7 show that the drop of power is severe for all methods when q decreases. Even with $r = 4$, the power of all methods is very low when $q = 0.001$.

Discussion

Genome-wide association studies efficiently allow targeting frequent variants with modest effects, associated with various multifactorial diseases. Data which prevailed for linkage tests, ASP or more extended pedigrees were ne-

Fig. 7. Power of the 3 tests and designs, with $\alpha = 10^{-11}$, $n = 1,000$, q ranging from 0.001 to 0.1 (a logarithmic scale was used for q), and $r = 2, 3$ and 4 for **a–c**, respectively.

glected for the benefit of massive case-control designs. Familial diseases and the joint use of linkage and association information can, however, lead to efficient designs for complex diseases [36]. The discussion about the relative merits of linkage and association analysis has led, mostly, to the concurring result that a combination of both types of information is the method of choice (see for example [37]). This becomes clear in the situation described in this paper.

Not only because the proportion of index cases bearing an at-risk rare variant is higher than in the general population, but mainly because the number of alleles shared IBD in sib-pairs, in which the index case bears a rare variant, provides an independent information to confirm or reject the involvement of this variant in a disease, can the situation presented here lead, when combined into a single test statistics, to a very powerful test. Besides the gain in power over more conventional methods, conditioning the linkage to the association observation [5] is also crucial when passing from the SNP association signal obtained in genome-wide association studies to the real measure of the gene effect (in terms of genotypic relative risks), as illustrated in many examples: HLA DQB and rheumatoid arthritis [38], PTPN22 and rheumatoid arthritis [39], IL2RA and multiple sclerosis [40]. More generally, the aim of a candidate gene strategy is not only to show its involvement but to model its effect. The more parameters there are in a model, the more information is needed. The MASC method [41] was developed in this regard.

In this paper, we propose a study design based on a sample of controls and a sample of ASP. The rare variants

are typed in the controls and the index case of each sib-pair; lower-density genotyping is required for the other sib, allowing to estimate the number of alleles shared IBD at the loci detected in the first step. Note that the costs of this low-density genotyping is negligible compared to the costs of sequencing the index case. This information could even be available from previous studies. We showed that our test with the proposed design is powerful enough to uncover rare variants in a genome-wide analysis. For some relative risk values associated with the rare variant considered, the use of IBD information increases the power by more than 20%.

Our study shows that using the familial history of patients can reduce the costs of rare variant association studies drastically. A recent paper [27] dealing with the same subject draws different conclusions: the familial design was found useful for diseases with a small relative recurrence risk in one sibling, but useless or even counterproductive when the risk is high. This discrepancy of results is due to the fact that we studied a different disease model. The model chosen in the study by Ionita-Laza and Ottman [27] is Risch's two-locus heterogeneity model [42], in which the disease has 2 different etiologies: it can be caused by the rare variant considered or by the remaining of the genome. In contrast, in our model, the disease is caused by the rare variant and by the remaining of the genome, and bearing the rare allele results in the same multiplicative risk increase regardless of familial history (equations 1 and 2), which is not true in the heterogeneity model. The large variety of models which may underly

multifactorial diseases must encourage the geneticists to not stick to a single strategy and, in any case, to enrich their information as much as possible. For the majority of multifactorial diseases, the model underlying the genetic susceptibility is likely to be too complex to reduce the information to the simple comparison of the genome sequences of patients and controls.

As already noted, if our model holds true, the enrichment of causative variants in the index case of ASP is systematic; thus, the ASP design will systematically be more powerful. As the distortion in IBD demonstrated in table 4 is also systematic, test T will be the more powerful as it makes use of this additional information. Table 4 also shows that the enrichment of rare variants is larger among index cases from sib-pairs with $IBD = 2$. Hence, under special circumstances, when only one candidate region is studied instead of the whole genome, and if the number of ASP is large compared to the number of individuals for which sequencing is possible, one can consider sequencing only index cases from ASP with $IBD = 2$ as an alternative strategy. Another power study would be needed to compare this strategy to the one we advocate here.

In some cases, the existence of rare protective variants can be conjectured, especially for frequent traits such as obesity. It is possible to use the statistics T for a left-sided test; however, another solution is worth consideration: comparing a set of non-obese sib-pairs to a set of obese controls. As a rare variant protective for obesity can be considered as causative for non-obesity, our test would apply. We conjecture that under certain circumstances, this design can be more powerful.

Our method can be extended to the situation of several rare variants within the same gene exerting an effect on the disease studied. A prime example of this situation is Crohn's disease, where more than 10 years ago the first gene (NOD2) identified as involved in disease susceptibility perfectly illustrated this situation with 3 main variants implicated. These 3 variants represented 81% of the causal mutations, whilst the remaining 19% were constituted from a total of 27 other mutations [43]. The concept of several functional/causative mutations within the same disease is the underlying paradigm for all the clustering techniques proposed by others (in particular, see [15, 16, 19, 21–23]). This has been shown to be a valid assumption for a variety of phenotypes, e.g. Crohn's disease (as above), essential tremor [44], holoprosencephaly [45], to name just 2 very recently published examples.

In order to extend the method to this situation, a set of variants can be considered as a multiallelic variant (an allele corresponding to each possible haplotype, with the

assumption that the probability of intragenic recombination is negligible). Considering all possible genotypes for this multiallelic variant, the 2 main principles used by the test proposed here would apply:

- (1) The risk genotypes are more frequent in patients with an affected sib than in random patients.
- (2) The IBD sharing distribution for affected sibs of index cases depends on the genotype of the index case. In theory, the number of conditional IBD distributions is equal to the number of possible genotypes. However, in practice, some rare genotypes would be neglected or pooled with other genotypes.

The most complicated situation occurs when there are both causative and protective rare variants within one gene. Of course, if there is no LD between these variants, and no interaction between the variants, so that our modeling assumptions still hold, our test remains usable, but this is unlikely to be the case: in this situation, the strategy of haplotype modeling sketched above would be inescapable. We are convinced that the more intricate the model is, the more useful can the additional information beared by the familial design be. Once a gene bearing causative or protective variants is identified, a precise estimation of the risks it confers requires adopting a candidate gene strategy, collecting as much information as possible.

We would also like to point out that the test presented here is not the only possible application of this design and this model. Here we focused on the increase in statistical power in terms of rare variant mapping, but an estimation of the r value can be obtained by maximization of the likelihood (2) in both q and r . A test for the direct role of the variant considered and the multiplicative model (1) can then be performed by comparing the maximum likelihood to the likelihood of the saturated model. These extensions will be the object of future work.

R scripts to compute our test T and the associated p values are available for download from <http://www.g2s.u-psud.fr/herve/asp>.

Appendix

The Model for ASP

We use Bayes' theorem to compute the probabilities $P_{ki} = \mathbb{P}(G_i = k, IBD = i | \text{ASP})$:

$$P_{ki} = \frac{\mathbb{P}(\text{ASP} | G_i = k, IBD = i) \mathbb{P}(G_i = k, IBD = i)}{\mathbb{P}(\text{ASP})} \quad (\text{A.1})$$

$$= \frac{\mathbb{P}(\text{ASP} | G_i = k, IBD = i) \mathbb{P}(G_i = k, IBD = i)}{\sum_{l,j} \mathbb{P}(\text{ASP} | G_i = l, IBD = j) \mathbb{P}(G_i = l, IBD = j)}. \quad (\text{A.2})$$

We thus have to compute the values of $\alpha_{ki} = \mathbb{P}(G_1 = k, IBD = i)$ and $\beta_{ki} = \mathbb{P}(\text{ASP}|G_1 = k, IBD = i)$. We have readily $\mathbb{P}(G_1 = k, IBD = i) = \mathbb{P}(G_1 = k)\mathbb{P}(IBD = i)$; the IBD distribution is well-known:

$$\mathbb{P}(IBD = i) = \begin{cases} \frac{1}{4} & \text{if } i = 0, 2, \\ \frac{1}{2} & \text{if } i = 1, \end{cases} \quad (\text{A.3})$$

and from the Hardy-Weinberg equilibrium we have:

$$\mathbb{P}(G_1 = k) = \begin{cases} p^2 & \text{if } k = 0, \\ 2pq & \text{if } k = 1, \\ q^2 & \text{if } k = 2. \end{cases} \quad (\text{A.4})$$

This allows to compute all α_{ki} . Now we turn to $\beta_{ki} = \mathbb{P}(\text{ASP}|G_1 = k, IBD = i)$. From the model assumptions (equation 1 in the main text), we deduce:

$$\begin{aligned} \beta_{ki} &= \mathbb{P}(\text{ASP}|G_1 = k, IBD = i) \\ &= \mathbb{P}(\text{Aff}_1|G_1 = k) \mathbb{P}(\text{Aff}_2|\text{Aff}_1, G_1 = k, IBD = i) \end{aligned} \quad (\text{A.5})$$

$$= r^k f_0 \mathbb{P}(\text{Aff}_2|\text{Aff}_1, G_1 = k, IBD = i) \quad (\text{A.6})$$

$$= r^k f_0 \gamma_{ki} \quad (\text{A.7})$$

and

$$\begin{aligned} \gamma_{ki} &= \mathbb{P}(\text{Aff}_2|\text{Aff}_1, G_1 = k, IBD = i) \\ &= \sum_l \mathbb{P}(\text{Aff}_2|\text{Aff}_1, G_2 = l) \mathbb{P}(G_2 = l|G_1 = k, IBD = i) \end{aligned} \quad (\text{A.8})$$

$$= \sum_l f_0' r^l \mathbb{P}(G_2 = l|G_1 = k, IBD = i). \quad (\text{A.9})$$

Finally, computing $\mathbb{P}(G_2 = l|G_1 = k, IBD = i)$ is straightforward. The case $IBD = 2$ is trivial:

$$\mathbb{P}(G_2 = l|IBD = 2, G_1 = k) = \begin{cases} 0 & \text{if } k \neq l \\ 1 & \text{if } k = l \end{cases} \quad (\text{A.10})$$

For $IBD = 1$, the following matrix recapitulates the values of $\mathbb{P}(G_2 = l|IBD = 1, G_1 = k)$:

$$\left[\mathbb{P}(G_2 = l|IBD = 1, G_1 = k) \right]_{kl} = \begin{bmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{bmatrix}, \quad (\text{A.11})$$

and if $IBD = 0$, clearly

$$\mathbb{P}(G_2 = l|IBD = 0, G_1 = k) = \begin{cases} p^2 & \text{if } l = 0 \\ 2pq & \text{if } l = 1 \\ q^2 & \text{if } l = 2 \end{cases} \quad (\text{A.12})$$

The following matrix recapitulates the values of $\gamma_{ki} = \mathbb{P}(\text{Aff}_2|\text{Aff}_1, G_1 = k, IBD = i)$, obtained from equations A.9–A.12:

$$[\gamma_{ki}]_{ki} = f_0' \times \begin{bmatrix} p^2 + 2pqr + q^2r^2 & p + qr & 1 \\ p^2 + 2pqr + q^2r^2 & \frac{1}{2}(p + r + qr^2) & r \\ p^2 + 2pqr + q^2r^2 & pr + qr^2 & r^2 \end{bmatrix} \quad (\text{A.13})$$

Putting all together, we readily obtain the values of β_{ki} and then of the products $\alpha_{ki} \beta_{ki}$:

$$\alpha_{02}\beta_{02} = f_0 f_0' \times \frac{1}{4} p^2 \times 1$$

$$\alpha_{01}\beta_{01} = f_0 f_0' \times \frac{1}{2} p^2 \times (p + qr)$$

$$\alpha_{00}\beta_{00} = f_0 f_0' \times \frac{1}{4} p^2 \times (p^2 + 2pqr + q^2r^2)$$

$$\alpha_{12}\beta_{12} = f_0 f_0' \times \frac{1}{4} 2pq \times r \times r$$

$$\alpha_{11}\beta_{11} = f_0 f_0' \times \frac{1}{2} 2pq \times r \times \frac{1}{2} (p + r + qr^2)$$

$$\alpha_{10}\beta_{10} = f_0 f_0' \times \frac{1}{4} 2pq \times r \times (p^2 + 2pqr + q^2r^2)$$

$$\alpha_{22}\beta_{22} = f_0 f_0' \times \frac{1}{4} q^2 \times r^2 \times r^2$$

$$\alpha_{21}\beta_{21} = f_0 f_0' \times \frac{1}{2} q^2 \times r^2 \times (pr + qr^2)$$

$$\alpha_{20}\beta_{20} = f_0 f_0' \times \frac{1}{4} q^2 \times r^2 \times (p^2 + 2pqr + q^2r^2)$$

The normalizing term $\mathbb{P}(\text{ASP}) = \sum_{ki} \alpha_{ki} \beta_{ki}$ can now be computed by summing all these expressions. It can be factorized as:

$$\mathbb{P}(\text{ASP}) = 1/4 f_0 f_0' (2 + 4q(r - 1) + q(1 + q)(r - 1)^2). \quad (\text{A.14})$$

The values of $\mathbb{P}_{ki}$ are now obtained simply from the above computed values as:

$$\mathbb{P}_{ki} = \frac{\alpha_{ki} \beta_{ki}}{\mathbb{P}(\text{ASP})}. \quad (\text{A.15})$$

Note that the term $f_0 f_0'$ becomes redundant. Note also that under H_0 , i.e. $r = 1$, all $\gamma_{ki} = f_0'$, thus $\beta_{ki} = f_0 f_0'$ and $\mathbb{P}(\text{ASP}) = f_0 f_0'$, and finally $\mathbb{P}_{ki}$ is simply equal to α_{ki} .

Computing the Score

The easiest way to compute the first order derivative

$$U(q, 1) = \left(\frac{\partial L}{\partial r} \right) (q, 1)$$

of the log-likelihood

$$L(r, q) = \sum_{k,i} n_{ki} \log P_{ki} + \sum_k m_k \log Q_k \quad (\text{A.16})$$

is to add a first-order Taylor expansion $(r - 1)$ to each term. For example, the term $\log \mathbb{P}(\text{ASP})$ has the simple Taylor expansion:

$$\log \mathbb{P}(\text{ASP}) = 4q(r - 1) + \dots \quad (\text{A.17})$$

putting aside the term $f_0 f_0'$, which ultimately drops out.

We now obtain the following first-order expansions easily:

$$\begin{aligned}
\log P_{02} &= 2\log(p) - 2\log 2 & -4q(r-1) & + \dots \\
\log P_{01} &= 2\log(p) - \log 2 & -3q(r-1) & + \dots \\
\log P_{00} &= 2\log(p) - 2\log 2 & -2q(r-1) & + \dots \\
\log P_{12} &= \log(p) + \log(q) - \log 2 & +(-4q+2)(r-1) & + \dots \\
\log P_{11} &= \log(p) + \log(q) & +\left(-3q+\frac{3}{2}\right)(r-1) & + \dots \\
\log P_{10} &= \log(p) + \log(q) - \log 2 & +(-2q+1)(r-1) & + \dots \\
\log P_{22} &= 2\log q - 2\log 2 & +(-4q+4)(r-1) & + \dots \\
\log P_{21} &= 2\log q - \log 2 & +(-3q+3)(r-1) & + \dots \\
\log P_{20} &= 2\log q - 2\log 2 & +(-2q+2)(r-1) & + \dots
\end{aligned}$$

The terms in $\log Q_k$ do not depend on r , so the value of $U(q, 1)$ can be easily deduced from the above equations alone. For their values, see equations 3–5 in the main text.

Variance of the Score

There can be various way to estimate the variance of the score. We simply want an estimate of n , m , and q under H_0 . We used the Delta-method, with the help of a computer algebra program (Maple 14), as follows: we fixed a value of q , n and m , and we denoted the gradient of U by $u = [u_1; u_2]$ seen as a function of $v = [n_{02}, n_{01}, \dots, n_{20}, m_0, m_1, m_2]$, evaluated for $\mathbb{E}(v)$ (with this expected value computed under H_0); u_1 is the vector obtained by taking only the first 9 components, which correspond to the variables $v_n = [n_{02}, n_{01}, \dots, n_{20}]$, and u_2 is obtained by taking the last 3 components, which correspond to the variables $v_m = [m_0, m_1, m_2]'$. We then estimated the variance of U (for a fixed value of q) by $u_1' \Sigma_1 u_1 + u_2' \Sigma_2 u_2$, where Σ_1 and Σ_2 are the variance-covariance matrices of the variables in v_n and v_m under H_0 , respectively (here the independence of the ASP sample and of the control sample is used to cut the variance into 2 terms). The matrices Σ_1 and Σ_2 are the usual variance-covariance matrices of the multinomial distributions with parameters given by $P_{ki} = \alpha_{ki}$ (as we assume $r = 1$) for Σ_1 , and by Q_k for Σ_2 .

Here are some intermediate steps of the computation. We have:

$$u_1 = \frac{1}{n+m} \begin{bmatrix} -q(n+4m) \\ -3mq \\ q(n-2m) \\ (n+4m)(1-2q) \\ \frac{3}{2}m(1-2q) \\ \frac{1}{2}(1-2q)(2m-n) \\ (n+4m)(1-q) \\ 3m(1-q) \\ (1-q)(2m-n) \end{bmatrix} \quad (\text{A.18})$$

$$u_2 = \frac{1}{n+m} \begin{bmatrix} 3nq \\ -\frac{3}{2}n(1-2q) \\ -3n(1-q) \end{bmatrix} \quad (\text{A.19})$$

and we obtain

$$u_1' \sum_i u_1 = \frac{1}{4} q(1-q) \frac{(n^2 + 19m^2 + 2nm)n}{(n+m)^2} \quad (\text{A.20})$$

$$u_2' \sum_i u_2 = \frac{9}{2} q(1-q) \frac{n^2 m}{(n+m)^2}, \quad (\text{A.21})$$

and summing up these 2 terms, after simplification one obtains:

$$\sigma_U^2 \approx \frac{1}{4} q(1-q) \frac{n(19m+n)}{n+m}.$$

Computation of p Values Using Importance Sampling

Importance sampling is a method for computing expectations using a random sample drawn from a distribution different from the target distribution. Our problem is to compute the empirical cumulative density function of the statistics T and A under H_0 (in the sequel we restrict the discussion to T); i.e. ideally any value of x can be estimated by:

$$F(x) = \mathbb{P}(T \leq x) = \mathbb{E}(\mathbb{1}_{T \leq x}). \quad (\text{A.22})$$

This can be done by repeatedly sampling data from the multinomial distribution associated with H_0 , which we denote P_0 , computing the associated statistics values $(t_i)_{i=1, \dots, N}$, and estimating $F(x)$ by:

$$F(x) \approx \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{t_i} \leq x. \quad (\text{A.23})$$

However, when x is located far to the right of the mode of distribution, which is the case when $F(x)$ is very close to 1, a huge number N of samples will be needed to get an accurate estimation for $1 - F(x)$ (which is the p value associated with x in our one-sided test).

The principle of random sampling is to sample from a distribution $\mathbb{P}_1$ instead of P_0 ; for each sample s_i , an importance weight

$$w_i = \frac{\mathbb{P}_0(s_i)}{\mathbb{P}_1(s_i)}$$

is recorded, as well as the statistics values $t_i = T(s_i)$, and we estimate $F(x)$ by:

$$F(x) \approx \frac{1}{\sum w_i} \sum_{i=1}^N w_i \mathbb{1}_{t_i} \leq x. \quad (\text{A.24})$$

For a theoretical justification of the method, see for example Gelman et al. [46]. This is interesting for our problem because an appropriate choice of $\mathbb{P}_1$ allows to obtain values of t_i that will be far enough from the mode of distribution under $\mathbb{P}_0$ to get good estimates for $1 - F(x)$ or $F(x)$ for the values of interest. In practice, we used a probability log proportional to the multinomial distribution, i.e. $\mathbb{P}_1(s) \propto \mathbb{P}_0(s)^\alpha$, where we found that $\alpha = 0.2$ produced satisfying results.

Acknowledgement

H.P. and B.M.-M. were partly funded by a grant from the ARSEP foundation (Fondation pour l'Aide à la Recherche sur la Sclérose en Plaques).

References

- Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, Mitra M, Farmer J, Williams SR, Cox NJ, Bell GI, Risch N, Spielman RS: A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat Genet* 1998;19:292–296.
- Louis EJ, Thomson G, Payami H: The affected sib method. II. The intermediate model. *Ann Hum Genet* 1983;47:225–243.
- Motro U, Thomson G: The affected sib method. I. Statistical features of the affected sib-pair method. *Genetics* 1985;110:525–538.
- Thomson G, Bodmer W: The genetic analysis of HLA and disease association; in Dausset J, Svejgaard A (eds): *HLA and Disease*. Baltimore, Williams and Wilkins, 1977, pp 84–93.
- Clerget-Darpoux G, Selinger-Leneman H, Babron MC: Why do complex traits resist DNA analysis? *Int J Hum Genet* 2001;1:55–63.
- Bell GI, Horita S, Karam JH: A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 1984;33:176–183.
- Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001;411:599–603.
- Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996;273:1516–1517.
- The International HapMap Consortium: The International HapMap Project. *Nature* 2003;426:789–796.
- The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005;437:1299–1320.
- The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarrroll SA, Visscher PM: Finding the missing heritability of complex diseases. *Nature* 2009;461:747–753.
- Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;34:188–193.
- Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V: A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 2010;6:e1000954.
- Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;70:42–54.
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C: A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 2011;7:e1001289.
- Ionita-Laza I, Laird NM: On the optimal design of genetic variant discovery studies. *Stat Appl Genet Mol Biol* 2010;9.
- King CR, Rathouz PJ, Nicolae DL: An evolutionary framework for association testing in resequencing studies. *PLoS Genet* 2010;6:e1001202.
- Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
- Liu DJ, Leal SM: A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 2010;6:e1001156.
- Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;5:e1000384.
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;7:e1001322.
- Price AL, Kryukov GV, De Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;86:832–838.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;89:82–93.
- Akula N, Detera-Wadleigh S, Shugart YY, Nalls M, Steele J, McMahon FJ: Identity-by-descent filtering as a tool for the identification of disease alleles in exome sequence data from distant relatives. *BMC Proc* 2011;5(suppl 9):S76.
- Choi SH, Liu C, Dupuis J, Logue MW, Jun G: Using linkage analysis of large pedigrees to guide association analyses. *BMC Proc* 2011;5(suppl 9):S79.
- Ionita-Laza I, Ottman R: Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 2011;189:1061–1068.
- Shi G, Rao DC: Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genet Epidemiol* 2011;35:572–579.
- Shi G, Simino J, Rao DC: Enriching rare variants using family-specific linkage information. *BMC Proc* 2011;5(suppl 9):S82.
- Feng T, Elston R, Zhu X: A novel method to detect rare variants using both family and unrelated case-control data. *BMC Proc* 2011;5(suppl 9):S80.
- Zhu X, Feng T, Li Y, Lu Q, Elston RC: Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 2010;34:171–187.
- Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 1987;84:2363–2367.
- Han L, Abney M: Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* 2011;35:557–567.
- Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11:375–386.
- Sasieni PD: From genotypes to genes: doubling the sample size. *Biometrics* 1997;53:1253–1261.
- Clerget-Darpoux F: Use of association and linkage information for the study of multifactorial disease; in Pawlowitzki IH, Edwards JH, Thompson EA (eds): *Genetic Mapping of Disease Genes*. San Diego, Academic Press, 1997, pp 179–187.
- Muller-Myhsok B: A note on the relative power of association studies and linkage analysis in the genetic analysis of disease; in Pawlowitzki IH, Edwards JH, Thompson EA (eds): *Genetic Mapping of Disease Genes*. San Diego, Academic Press, 1997, pp 173–178.
- Tezenas du Montcel S, Michou L, Petit-Teixeira E, Osorio J, Lemaire I, Lasbleiz S, Pierlot C, Quillet P, Bardin T, Prum B, Cornelis F, Clerget-Darpoux F: New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis Rheum* 2005;52:1063–1068.
- Bourgey M, Perdry H, Clerget-Darpoux F: Modeling the effect of PTPN22 in rheumatoid arthritis. *BMC Proc* 2007;1(suppl 1):S37.
- Babron MC, Perdry H, Handel AE, Ramagopalán SV, Damotte V, Fontaine B, Müller-Myhsok B, Ebers GC, Clerget-Darpoux F: Determination of the real effect of genes identified in GWAS: the example of IL2RA in multiple sclerosis. *Eur J Hum Genet* 2011;20:321–325.
- Clerget-Darpoux F, Babron MC, Prum B, Lathrop GM, Deschamps I, Hors J: A new method to test genetic models in HLA associated diseases: the MASC method. *Ann Hum Genet* 1988;52:247–258.
- Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 1990;46:222–228.

- 43 Lesage S, Zouali H, Cézard JP, Colombel JF, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V, Finkel Y, Modigliani R, Gower-Rousseau C, Macry J, Merlin F, Chamaillard M, Jannot AS, Thomas G, Hugot JP; EPWG-IBD Group, EPIMAD Group, and GETAID Group: *CARD15/NOD2* mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 2002;70:845–857.
- 44 Merner ND, Girard SL, Catoire H, Bourassa CV, Belzil VV, Rivière JB, Hince P, Levert A, Dionne-Laporte A, Spiegelman D, Noreau A, Diab S, Szuto A, Fournier H, Raelson J, Belouchi M, Panisset M, Cossette P, Dupré N, Bernard G, Chouinard S, Dion PA, Rouleau GA: Exome sequencing identifies *FUS* mutations as a cause of essential tremor. *Am J Hum Genet* 2012;91:313–319.
- 45 Ribeiro LA, Roessler E, Hu P, Pineda-Alvarez DE, Zhou N, Jones S, Abd Chandrasekharappa M, Richieri-Costa A, Muenke M: Comparison of mutation findings in *ZIC2* between microform and classical holoprosencephaly in a Brazilian cohort. *Birth Defects Res A Clin Mol Teratol* 2012;94:912–917.
- 46 Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis*. Boca Raton, CRC Press, 2004.