Study Designs for Identification of Rare Disease Variants in Complex Diseases: The Utility of Family-Based Designs

Iuliana Ionita-Laza*[,1] and Ruth Ottman[†,‡]

*Department of Biostatistics and †G. H. Sergievsky Center and Departments of Epidemiology and Neurology, Columbia University, New York, New York 10032, and ‡Epidemiology Division, New York State Psychiatric Institute, New York, New York 10032

ABSTRACT The recent progress in sequencing technologies makes possible large-scale medical sequencing efforts to assess the importance of rare variants in complex diseases. The results of such efforts depend heavily on the use of efficient study designs and analytical methods. We introduce here a unified framework for association testing of rare variants in family-based designs or designs based on unselected affected individuals. This framework allows us to quantify the enrichment in rare disease variants in families containing multiple affected individuals and to investigate the optimal design of studies aiming to identify rare disease variants in complex traits. We show that for many complex diseases with small values for the overall sibling recurrence risk ratio, such as Alzheimer's disease and most cancers, sequencing affected individuals with a positive family history of the disease can be extremely advantageous for identifying rare disease variants. In contrast, for complex diseases with large values of the sibling recurrence risk ratio, sequencing unselected affected individuals may be preferable.

COMMON diseases such as diabetes, cancer, and autism are likely caused by a complex interaction among many genes and environmental factors. Both common and rare genetic variants are expected to play a role. Thus far the available technology has allowed for the identification of common disease susceptibility variants, mostly via genome-wide association studies. However, the common variants detected so far have small effect sizes and overall explain only a small fraction of the estimated trait heritability (Maher 2008; Manolio *et al.* 2009). The recent advances in next-generation sequencing technologies (Metzker *et al.* 2010; Tucker *et al.* 2009) allow for the first time an objective assessment of the importance of rare variants in complex diseases. An increasing number of recent studies on hypertension, schizophrenia, epilepsy, type-1 diabetes, autism, etc. (Ji *et al.* 2008; Stefansson *et al.* 2008; Helbig *et al.* 2009; Nejentsev *et al.* 2009; Pinto *et al.* 2010) implicate rare variants in these disorders.

Ongoing sequencing studies are already generating unprecedented amounts of genetic data. The large number of genetic variants in these data sets, most of them with low frequencies (<1%), creates great challenges for statistical analysis. Traditional association testing strategies that have worked well for common variants will have low power to identify rare disease susceptibility variants (Morris and Zeggini 2009; Bansal *et al.* 2010). To extract the rich information provided by large sequencing data sets, several novel statistical approaches have been proposed, especially designed to identify rare variants that influence disease risk (Li and Leal 2008; Madsen and Browning 2009; Bhatia *et al.* 2010; Han and Pan 2010; King *et al.* 2010; Liu and Leal 2010; Price *et al.* 2010; Ionita-Laza *et al.* 2011; Neale *et al.* 2011). The common idea underlying all these methods is to group variants in a region of interest, *e.g.*, a gene, and perform a gene-based test rather than individual tests for each of the variants in a gene.

An important question that has not yet been addressed is the relative power of designs based on affected relatives *vs.* designs using unselected affected individuals to identify rare disease variants. Since rare disease variants tend to be enriched in families containing multiple affected individuals, family-based designs can play an important role in the

doi: 10.1534/genetics.111.131813
Manuscript received June 19, 2011; accepted for publication August 3, 2011
Supporting information is available online at http://www.genetics.org/content/suppl/2011/08/12/genetics.111.131813.DC1.
[1] Corresponding author: 722 W. 168th St., 6th Floor, New York, NY 10025.
E-mail: ii2135@columbia.edu

identification of rare causal variants. The purpose of this article is to quantify this enrichment and to study its implications for the optimal design of studies that search for rare disease variants in complex traits.

Methods

Effective number of variants in related individuals

By analogy to the concepts of "effective population size" in population genetics (Wright 1931, 1938) and "effective number of markers" in a linkage disequilibrium block (Nyholt 2004), we introduce here a new concept for analysis of rare variants in related individuals: the *effective* number of variants at a position, *i.e.*, the number of "independent" variants at a specific position in a sample of related individuals. More precisely, the effective number of variants is the number of observed variants *corrected* for the known familial correlation among the individuals included in a sample. This concept is important as it allows a uniform comparison of designs based on various types of relatives, regardless of the relationship type.

For simplicity of presentation we define the effective number of variants at a position for a pair of individuals. If the individuals are unrelated, then the effective number of variants is equal to the observed number of variants since the two individuals are independent (hence no correction is necessary). However, the effective number of variants in *related* individuals is less than the total number of observed variants if some of the variants are *shared* among family members. For example, for a pair of siblings each of whom carries a rare variant in heterozygous state at a position, the effective number of variants will be less than the observed number of variants, *i.e.*, two, due to the high probability that these two variants are shared identical-by-descent (IBD) and hence are not independent. Similarly, for a pair of second cousins that each carry a rare variant in heterozygous state the effective number of variants is less than two, although as shown in the examples below, it is higher than for a pair of siblings due to the lower likelihood that a variant is shared IBD for second cousins compared with siblings. Below we define mathematically the concept of effective number of variants.

For a pair of relatives we define the effective number of variants, k_{eff}, as follows:

$$k_{\mathrm{eff}} = \begin{cases} k_{\mathrm{eff}|2} & \text{if both relatives carry a rare variant} \\ k_{\mathrm{eff}|1} & \text{if only one of the two relatives carries a rare variant} \\ k_{\mathrm{eff}|0} & \text{if neither of the two relatives carries a rare variant.} \end{cases}$$

By definition, $k_{\mathrm{eff}|1} = 1$ and $k_{\mathrm{eff}|0} = 0$. For $k_{\mathrm{eff}|2}$, we show in supporting information, File S1 that

$$k_{\mathrm{eff}|2} \cong \log_{2f}\left[4f\varphi + 4f^2\left(1 - 4\varphi + 4\delta\varphi^2\right)\right], \qquad (1)$$

where f is the frequency of the variant, and φ is the kinship coefficient; $\delta = 0$ if the two relatives can share a maximum of one allele IBD (*e.g.*, first cousins) and 1 if they can share two alleles IBD (*e.g.*, siblings). The approximation is based only on the assumption that the variant is rare (*i.e.*, $f \leq 0.01$) and is very accurate under this assumption.

When the two individuals are unrelated and each carries a rare variant, $\varphi = 0$ and we obtain the expected result that $k_{\mathrm{eff}|2} = 2$. For identical twins $\varphi = 0.5$, $\delta = 1$, and $k_{\mathrm{eff}|2} = 1$. For relatives in between, the effective number of variants is between 1 and 2. For two sibs, $\varphi = \frac{1}{4}$, $\delta = 1$, and hence for $f = 0.01$ we obtain $k_{\mathrm{eff}|2} = 1.17$. Similarly for two second cousins, $\varphi = \frac{1}{64}$, $\delta = 0$, and hence $k_{\mathrm{eff}|2} = 1.76$. These and other examples are summarized in Table 1. Note that the effective number of variants depends on the frequency f. Hence, as expected, the lower the frequency is, the lower the effective number of variants, reflecting the low probability that these shared variants are independent (and the greater chance they are identical-by-descent).

To summarize, for a pair of relatives k_{eff} is calculated as follows:

$$k_{\mathrm{eff}} = \begin{cases} \log_{2f}\left[4f\varphi + 4f^2\left(1 - 4\varphi + 4\delta\varphi^2\right)\right] \\ \quad \text{if both relatives carry a rare variant} \\ 1 \\ \quad \text{if only one of the two relatives carries a rare variant} \\ 0 \\ \quad \text{if neither of the two relatives carries a rare variant.} \end{cases}$$

Under the assumption that the variant is not associated with disease, we calculate the expected value of k_{eff} for a pair of relatives in File S1 (Equation S7) as

$$\begin{aligned} E\left[k_{\mathrm{eff}}\right] = {} & k_{\mathrm{eff}|2} \cdot 4f(1-f) \cdot \left[\varphi + f(1-f)\left(1 - 4\varphi + 4\delta\varphi^2\right)\right] \\ & + \; k_{\mathrm{eff}|1} \cdot 4f(1-f)^2 \cdot \left[(1-2\varphi) - f\left(1 - 4\varphi + 4\delta\varphi^2\right)\right]. \end{aligned} \qquad (2)$$

At a rare variant position, for a data set of relative pairs of the same type in N different families, we define $k_{\mathrm{eff}}^{\mathrm{Total}} = \sum_{i=1}^{N} k_{\mathrm{eff}}^{i}$, where k_{eff}^{i} is the effective number of variants in family i. Then

$$E\left[k_{\mathrm{eff}}^{\mathrm{Total}}\right] = NE\left[k_{\mathrm{eff}}\right]. \qquad (3)$$

If the variant is not associated with disease, then the distribution of $k_{\mathrm{eff}}^{\mathrm{Total}}$ can be approximated by a Poisson distribution with mean $= E[k_{\mathrm{eff}}^{\mathrm{Total}}]$ (as also shown empirically in File S1, Figure S1)

Effective number of variants at a disease locus

At a *disease* locus, the effective number of variants in two affected relatives is expected to be increased compared to a nondisease locus. We consider here a two-locus genetic heterogeneity model (Risch 1990a), where each locus is an independent cause for disease. We use the two-locus model for mathematical convenience; without loss of generality the second locus can be considered to encompass all the other disease loci that act additively to influence disease risk, in addition to the primary locus under investigation. The effective number of variants we expect to observe at the first disease locus in a pair of affected relatives is

Table 1 The effective number of variants, $k_{\text{eff}|2}$, at a rare variant position in two related individuals that each carry a variant, as defined in Equation 1

		$k_{\text{eff}\|2}$	
Relationship	φ	f = 0.001	f = 0.01
Identical twins	$\frac{1}{2}$	1.00	1.00
Parent–child	$\frac{1}{4}$	1.11	1.17
Sibs	$\frac{1}{4}$	1.11	1.17
Half-sibs	$\frac{1}{8}$	1.22	1.34
Uncle–nephew	$\frac{1}{8}$	1.22	1.34
First cousins	$\frac{1}{16}$	1.33	1.50
First cousins-1 (once removed)	$\frac{1}{32}$	1.44	1.64
Second cousins	$\frac{1}{64}$	1.55	1.76
Unrelateds	0	2.00	2.00

φ is the kinship coefficient. Results for two values of the frequency parameter f, 0.001 and 0.01, are shown.

$$
\begin{aligned}
E\left[k_{\text{eff}}^{\text{D}}\right] = {} & k_{\text{eff}|2}\cdot 4f(1-f) \\
& \cdot\ \left[\varphi + f(1-f)\left(1-4\varphi+4\delta\varphi^2\right)\right]\beta_{\text{R}} \\
& +\ k_{\text{eff}|1}\cdot 4f(1{-}f)^2 \\
& \cdot\ \left[(1-2\varphi) - f\left(1-4\varphi+4\delta\varphi^2\right)\right]\alpha_{\text{R}},
\end{aligned}
\tag{4}
$$

where β_{R}, $\alpha_{\text{R}} \geq 1$ are derived in File S1 and depend on the type of relationship R (*e.g.*, sibs, first cousins, etc.), the genotype relative risk (GRR) for the first locus, and the overall recurrence risk in siblings as measured by λ_{S} (Risch 1990a). The expression in Equation 4 is similar to the expression in Equation 2 for the effective number of variants expected at a nondisease locus. We performed simulations according to this model and show that the empirical estimates for $E[k_{\text{eff}}^{\text{D}}]$ agree very well with the analytical results in (4) (see File S1).

For a data set of N different families, each consisting of an affected relative pair of the same type, we have

$$
E\left[k_{\text{eff}}^{\text{D,Total}}\right] = NE\left[k_{\text{eff}}^{\text{D}}\right]. \tag{5}
$$

As above, the distribution of $k_{\text{eff}}^{\text{D,Total}}$ can be approximated by a Poisson distribution with mean $E[k_{\text{eff}}^{\text{D,Total}}]$ (as also shown empirically File S1).

Furthermore, using Equation 4 we can also calculate the effective number of variants we expect at the first disease locus in an affected individual *with* an affected relative. This is important to evaluate the importance of selecting affected individuals with a positive family history for a disease for inclusion in a sequencing study. We denote by k_1^{D} the number of variants at the first locus in an affected individual that is known to have an affected relative. Then using Equation 4,

$$
\begin{aligned}
E\left[k_1^{\text{D}}\right] = {} & 1\cdot 4f(1-f)\cdot\left[\varphi + f(1-f)\left(1-4\varphi+4\delta\varphi^2\right)\right]\beta_{\text{R}} \\
& +\ 0.5\cdot 4f(1{-}f)^2 \\
& \cdot\ \left[(1-2\varphi) - f\left(1-4\varphi+4\delta\varphi^2\right)\right]\alpha_{\text{R}},
\end{aligned}
$$

where β_{R}, $\alpha_{\text{R}} \geq 1$ are derived in File S1. As above, $E[k_1^{\text{D,Total}}] = NE[k_1^{\text{D}}]$ for a data set of N affected individuals, each with an affected relative. Again it is true that $k_1^{\text{D,Total}} \sim \text{Poisson}(E[k_1^{\text{D,Total}}])$.

Note that the description above pertains to a disease locus with only one disease variant position. However, a disease locus may have multiple disease variant positions (*i.e.*, allelic heterogeneity). The extension to this scenario is straightforward by summing the effective number of variants at each position and using the fact that a sum of independent Poisson random variables is also a Poisson random variable (we assume independence among the different variant positions within a locus, a reasonable approximation given the low frequency of the variants).

Expected P-value at a disease locus

To compare the power of designs on the basis of biologically related cases *vs.* unrelated cases, we calculate an expected *P*-value (EPV) (Dempster and Schatzoff 1965; Sackrowitz and Samuel-Cahn 1999). The expected *P*-value, or expected significance level as originally defined in the pioneering article of Dempster and Schatzoff (1965), has been proposed as a measure of the performance of a test. Unlike the power of a test, the EPV does not depend on any prespecified significance level and is in close connection with the common practice of reporting *P*-values in applied research. The EPV is a single number that can be used to judge the performance of a test; the smaller the EPV is, the better the test.

For our situation, the effective number of variants under the null hypothesis that neither of the two loci is associated with disease, $k_{\text{eff}}^{\text{Total}}$, follows approximately a Poisson distribution with mean $\lambda_1 = E[k_{\text{eff}}^{\text{Total}}]$ as derived in Equation 3. Similarly, the effective number of variants at the first disease locus in a two-locus disease model, $k_{\text{eff}}^{\text{D,Total}}$, follows approximately a Poisson distribution with $\lambda_2 = E[k_{\text{eff}}^{\text{D,Total}}]$ as derived in Equation 5. Then by definition the expected *P*-value for the first disease locus is

$$
\text{EPV} = P(T_1 \geq T_2) = P(T_1 - T_2 \geq 0),
$$

where $T_1 \sim \text{Poisson}(\lambda_1)$ and $T_2 \sim \text{Poisson}(\lambda_2)$. Since the difference between two independent Poisson random variables follows a Skellam distribution (Skellam 1946), we can calculate the EPV under a disease model only on the basis of the values of λ_1 and λ_2.

For comparison we also report results based on power, and, as shown below, they are highly correlated with those based on the EPV. Since we assume that the variant frequency in controls is known, the resulting power levels can be considered as achievable as the number of controls grows very large.

Results

Performance of affected relatives vs. unrelated affected individuals

For identifying rare disease variants, it is of great interest to study the circumstances where study designs involving

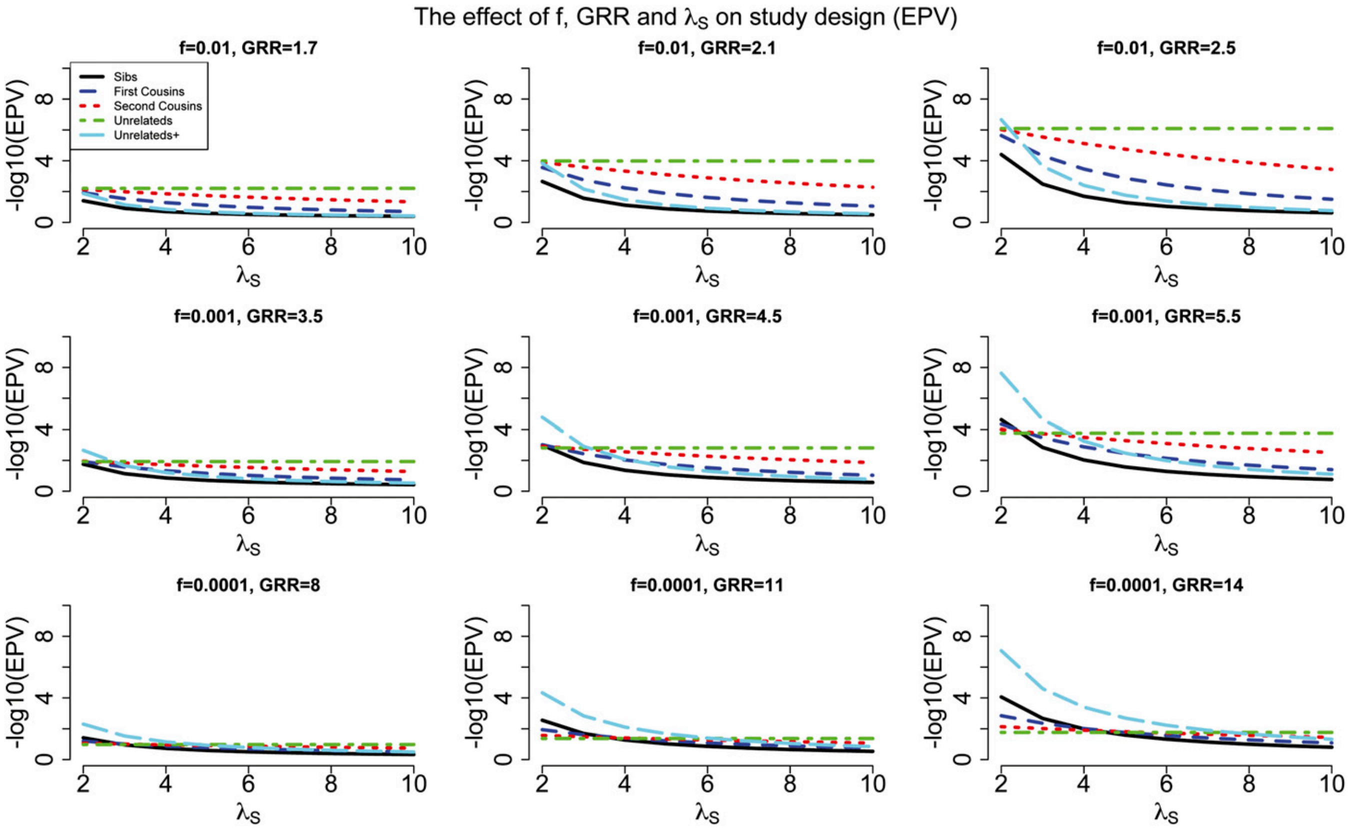

Figure 1 The effect of locus frequency, GRR, and overall λ_S on the relative performance (as measured by the expected *P*-value) of affected relatives *vs.* unrelated affected individuals. The three rows correspond to three different frequencies for the disease locus: 0.01, 0.001, and 0.0001. λ_S is between 2 and 10. The number of affected individuals is 2000: 1000 sib-pairs, 1000 first-cousin pairs, 1000 second-cousin pairs, 2000 unrelateds, and 2000 unrelated individuals known to have an affected sibling (*i.e.*, unrelateds+).

affected individuals from families containing multiple affected individuals are advantageous compared with those involving unselected affected individuals.

Under the assumption of a complex polygenic model, with many possible disease loci each with a small population attributable risk (Risch and Merikangas 1996), the two-locus disease model can be viewed as consisting of a locus of interest and a second locus accounting for the residual effect. We assume five study designs, each with the same number of affected individuals (*i.e.*, 2000): affected sib-pairs, affected first-cousin pairs, affected second-cousin pairs, unrelated affected individuals, and unrelated affected individuals known to have an affected sibling. We consider a complex trait with prevalence 0.03 and a sibling recurrence risk ratio (λ_S) between 2 and 10, as observed for many complex traits (Merikangas and Risch 2003). We assume that the variant frequency at the first disease locus is low, between 0.0001 and 0.01, and several possible values for the genotype relative risk with higher GRRs are assumed for lower frequency at the disease locus. In Figure 1 we show expected *P*-values associated with the first locus as a function of the sibling recurrence risk ratio, for each of the five study designs.

For complex diseases with low values for λ_S (*e.g.*, 2–4), affected individuals known to have an affected sibling are extremely advantageous to identify rare disease variants of moderate to large effect (*e.g.*, GRR $\sim$ 5–10, Figure 1). The difference between using affected individuals with an affected sibling *vs.* unselected affected individuals can be substantial in this case. For example, for a disease with $\lambda_S = 2$, for a locus with frequency 0.001 and a GRR of 5.5 the expected *P*-value is $10^{-7.65}$ for a study with 2000 unrelated affected individuals known to have an affected sibling and only $10^{-3.77}$ for a study with 2000 unselected affected individuals. Similarly, for a locus with frequency 0.0001 and a GRR of 14 the corresponding expected *P*-values are 10^{-7} and 10^{-18}. For small values of λ_S and large values of GRR (*e.g.*, $\lambda_S = 2$ and GRR = 14), even a design based on 1000 affected individuals known to have an affected sibling is expected to be more advantageous than a design based on 2000 unselected affected individuals (Figures 1 and 4). Similar results are obtained when the design comparison is based on power rather than expected *P*-value (Figure S2).

However, for complex diseases with larger values of λ_S such as autism ($\lambda_S \approx 22$), the advantage of using family-based designs (either affected relative pairs or affected individuals with a known affected sibling) diminishes, and unselected affected individuals can be more advantageous (Figure 2 and Figure S3). This trend of greater advantage with unselected affected individuals is more pronounced

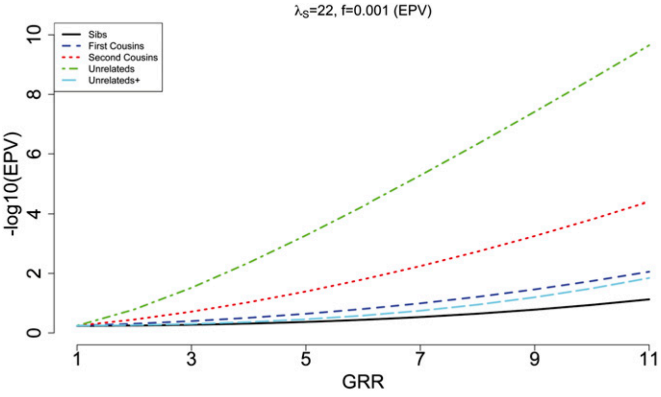

Figure 2 Usefulness of sequencing both affected individuals in a pair of affected relatives. The three rows correspond to three different frequencies for the disease locus: 0.01, 0.001, and 0.0001. The number of affected individuals is 2000 for the two affected relatives design, *i.e.*, 1000 sib-pairs and 1000 second-cousin pairs, and only 1000 for the design based on only one affected individual in a pair, *i.e.*, 1000 affected individuals known to have an affected sibling and 1000 affected individuals known to have an affected second cousin.

with increasing frequency of disease-causing variants (Figure 1). These results are in concordance with those in Risch (1990b, 1992) for linkage analysis, which state that for small values of overall λ_S and polymorphic information content (PIC) (Botstein *et al.* 1980) affected sib-pairs are optimal, while for large values of λ_S and PIC, more distantly related individuals are best.

Above we have assumed that the disease locus has only one disease variant position. Although this is clearly a simplified scenario, the results obtained are informative for the more realistic scenario when the locus of interest contains multiple disease variants. In Figure 3 we show the results for the case when 5 different disease variant positions are present at the locus. Two cases are illustrated: (1) the GRRs for the individual variants are a decreasing function of frequency, so that lower-frequency variants have substantially higher GRRs compared to more frequent variants (for example, $f = 0.0001$ and GRR $= 21$ *vs.* $f = 0.01$and GRR $= 1.2$), and (2) the GRR is the same for all variants (*i.e.*, 2). As can be seen, for the first case, due to the presence of high-risk low-frequency variants among the 5 disease variants, unrelated affected individuals that are known to have an affected sibling are best for small values of λ_S (as also shown previously in Figure 1), whereas for the second case where the GRR for low-frequency variants is modest (*i.e.*, 2), distantly related affected individuals, and in particular unselected affected individuals, perform better. Also shown are results when 10 additional random variants are added to the region. In this case, the relative performance of

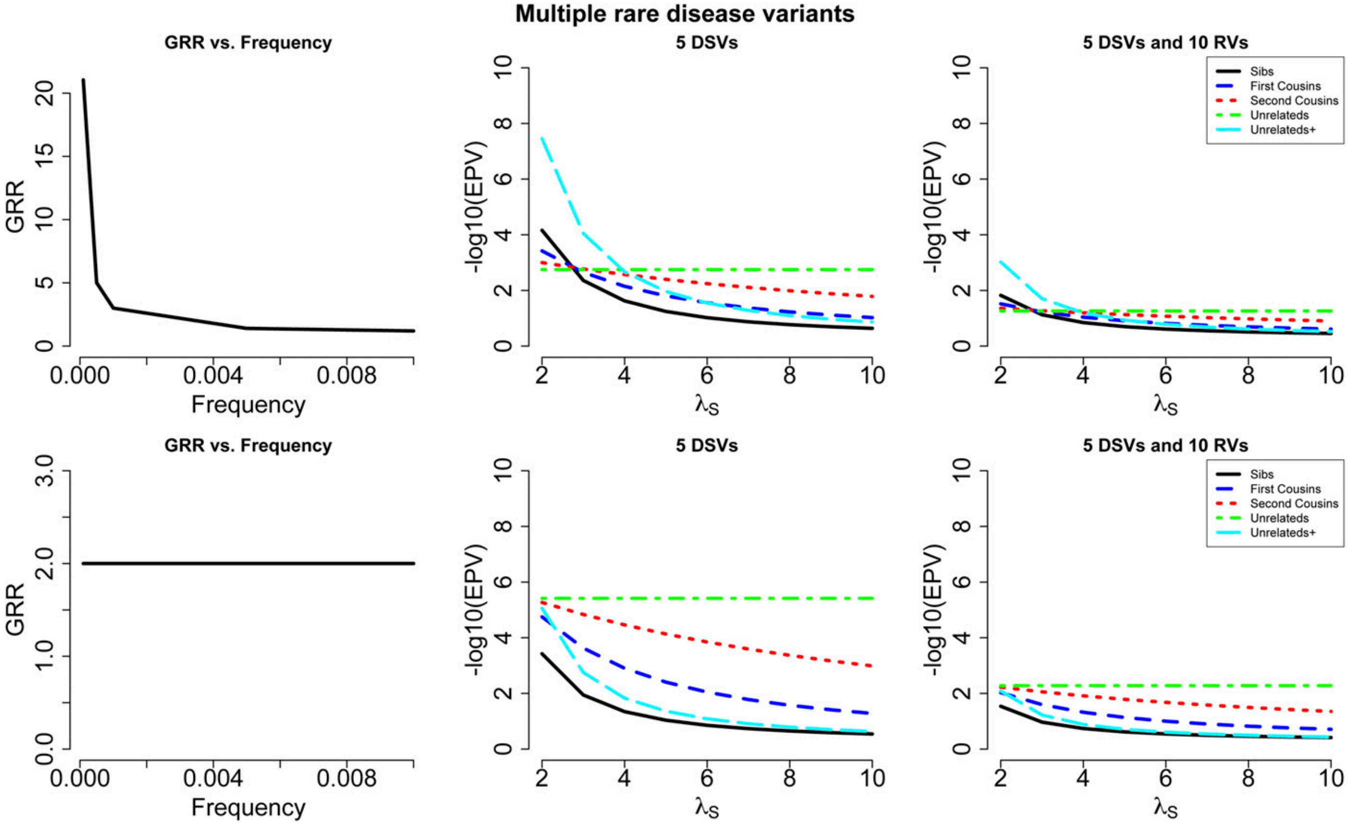

Figure 3 Results are shown for scenarios where multiple (*i.e.*, 5) disease susceptibility variants (DSVs) are present at the disease locus. The relationship between the GRR and disease variant frequency is shown in the first column. In the top panel, the GRR is a decreasing function of variant frequency, while in the bottom panel, the GRR is the same (*i.e.*, 2) regardless of variant frequency. The third column presents results when 10 additional random variants (*i.e.*, not disease variants) are also present.

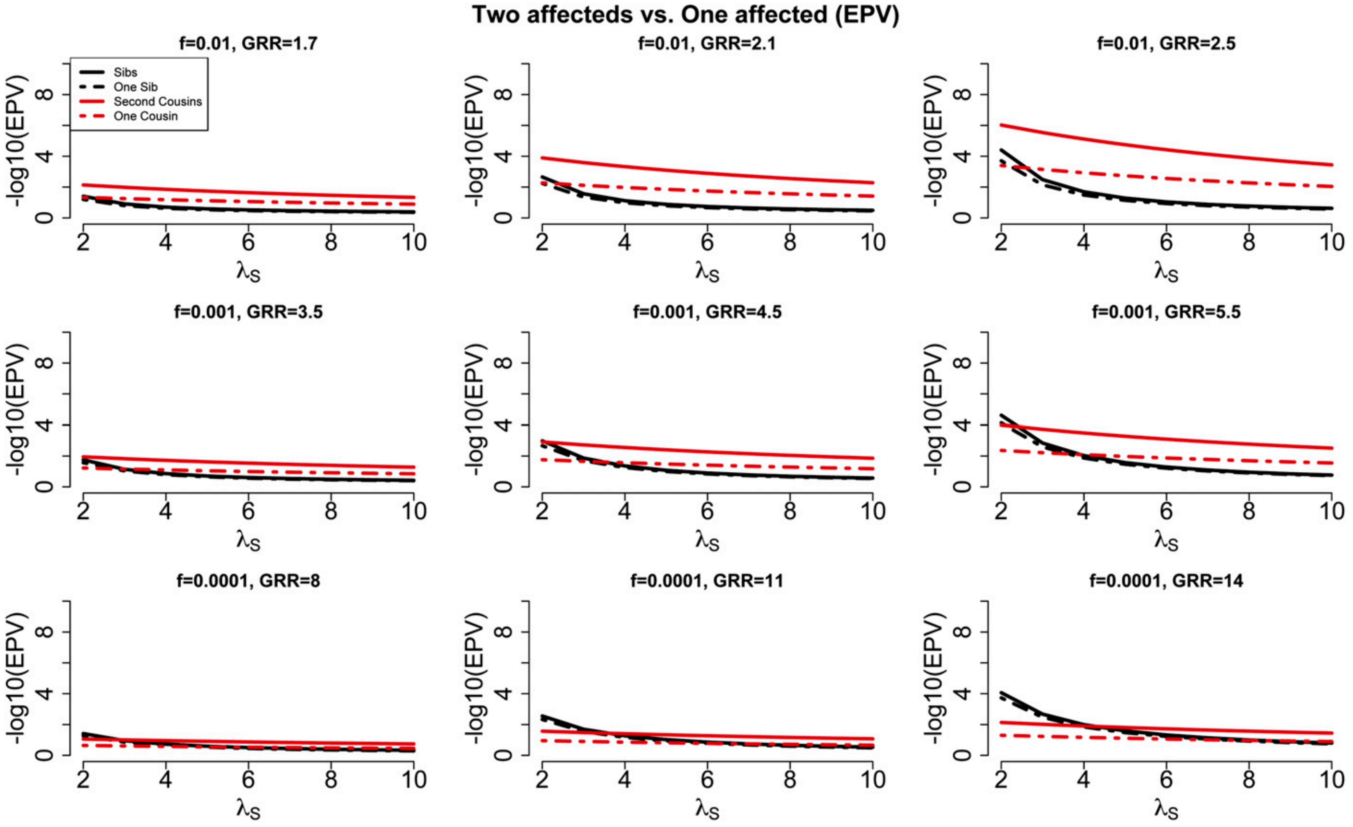

Figure 4 For complex traits with a large value for λ_S (*e.g.*, $\lambda_S = 22$ for autism) we show expected *P*-values for a disease locus with frequency *f* = 0.001 as a function of the GRR for five study designs, each with 2000 affected individuals: 1000 sib-pairs, 1000 first-cousin pairs, 1000 second-cousin pairs, 2000 unrelateds, and 2000 unrelated individuals known to have an affected sibling (*i.e.*, unrelateds+).

different designs is unchanged, but the performance is reduced for all designs.

The utility of sequencing both affected individuals in a pair of affected relatives

Sequencing both individuals from a pair of affected relatives doubles the sequencing costs, so it is important to consider the circumstances under which power is increased by this approach. We evaluated the utility of sequencing both affected individuals in an affected relative pair compared to sequencing only one individual from the pair (Figure 4 and Figure S4). We find that for affected sibling pairs, sequencing the second sibling contributes little additional information; hence sequencing only one of two affected siblings is expected to be on average almost as good as sequencing both of them, with the significant advantage of reducing the sequencing cost by half (Figure 4). However, for more distantly related individuals, such as second cousins, sequencing both individuals is expected to be more powerful than sequencing only one of them. The advantage can be significant. For example, when $\lambda_S = 2$, for a locus with frequency 0.001 and a GRR of 5.5 the expected *P*-value is 10^{-4} for a study with 1000 affected second-cousin pairs (power at $\alpha = 1.6 \times 10^{-6}$ is 87%) and only $10^{-2.1}$ for a study with 1000 affected individuals known to have a second cousin affected (power is 47%). If instead we compare the power of a study with 1000 affected second-cousin pairs with that of a study with 2000 affected individuals known to have a second cousin affected, hence the same sequencing cost for both studies, we find that the power levels are very similar (data not shown).

Discussion

In complex diseases characterized by extensive genetic heterogeneity, each disease locus is likely to be responsible for a very small fraction of affected individuals in a population. The choice of the affected individuals to be included in a study greatly influences the power to identify disease loci. The framework developed here, based on the effective number of variants in a pair of affected relatives, makes it feasible to quantify the enrichment in rare disease variants in family-based *vs.* population-based samples and to investigate the optimal study design for identifying rare disease variants in complex traits.

We have shown here that for diseases with small values for the sibling recurrence risk ratio, as observed with many complex traits, sequencing affected individuals with an affected close relative, such as a sibling, can be an extremely powerful strategy for identifying rare disease variants with moderate to large GRRs. This finding is in concordance with previous findings for breast cancer and common variants (Antoniou and Easton 2003). However, we find that the advantage of using affected individuals with a positive family history declines with increasing values of $\lambda_{S.}$ For complex diseases with large values of λ_S, such as autism ($\lambda_S \approx 22$), unselected affected individuals may be preferable. The main explanation for these results is that for a complex disease involving many contributing disease loci, one single locus is likely to explain less of the overall familial aggregation when λ_S is large than when λ_S is small. Hence when λ_S is large, focusing on affected relatives (or individuals with an affected relative) may enrich the sample in rare disease loci that are less likely to be shared among multiple families, and therefore unselected affected individuals may be preferable.

Designs based on affected relative pairs are commonly employed in linkage analysis (Weeks and Lange 1988). Although here we have not attempted to examine the performance of linkage analysis, we note that the results we obtained agree well with those for linkage analysis (Risch 1990b, 1992), which state that for small values of overall λ_S and disease locus frequency affected sib-pairs are optimal, while for large values of λ_S and disease locus frequency, more distantly related individuals are best.

The framework developed here is important for the statistical analysis of rare variant data. We have introduced here the concept of effective number of variants at a position in a set of relatives, which allows for a unified treatment of both biologically related and unrelated affected individuals and makes feasible natural extensions of statistical tests recently developed for population-based designs (such as the weighted-sum statistic in Madsen and Browning 2009) to general designs based on affected relative pairs and/or unrelated affected individuals. One simple statistic for affected relative pairs would be a weighted sum of the effective number of variants at different positions. Such extensions are currently under development.

The main goal of this article is to introduce a framework to assess the enrichment of rare disease mutations in affected relative pairs of different types. Such an enrichment analysis is of relevance for any association testing method, and therefore the results presented are expected to be valid in general for association testing. We have applied this framework to the problem of optimal study design for association studies with rare variants.

Software implementing the described methods is available in File S2, as well as on the first author's webpage at http://www.columbia.edu/~ii2135/, and should be helpful in the design of association studies with rare variants.

Acknowledgments

We are grateful to Mitchell Gail and Joseph Buxbaum for useful discussions. We also thank the associate editor and two reviewers for their comments that helped improve our article. This work was supported by grants R03HG005908-2 (to I.I.-L.) and RC2 NS070344 (to R.O.) from the National Institutes of Health and DMS-1100279 (to I.I.-L.) from the National Science Foundation.

Literature Cited

Antoniou, A. C., and D. F. Easton, 2003 Polygenic inheritance of breast cancer: implications for design of association studies. Genet. Epidemiol. 25: 190–202.

Bansal, V., O. Libiger, A. Torkamani, and N. J. Schork, 2010 Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. 11: 773–785.

Bhatia, G., V. Bansal, O. Harismendy, N. J. Schork, E. J. Topol *et al.*, 2010 A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput. Biol. 6: e1000954.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32: 314–331.

Dempster, A. P., and M. Schatzoff, 1965 Expected significance level as a sensitivity index for test statistics. J. Am. Stat. Assoc. 60: 420–436.

Han, F., and W. Pan, 2010 A data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. 70: 42–54.

Helbig, I., H. C. Mefford, A. J. Sharp, M. Guipponi, M. Fichera *et al.*, 2009 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat. Genet. 41: 160–162.

Ionita-Laza, I., J. Buxbaum, N. M. Laird, and C. Lange, 2011 A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 7: e1001289.

Ji, W., J. N. Foo, B. J. O'Roak, H. Zhao, M. G. Larson *et al.*, 2008 Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat. Genet. 40: 592–599.

King, C. R., P. J. Rathouz, and D. L. Nicolae, 2010 An evolutionary framework for association testing in resequencing studies. PLoS Genet. 6: e1001202.

Li, B., and S. M. Leal, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83: 311–321.

Liu, D. J., and S. M. Leal, 2010 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. 6(10): e1001156.

Madsen, B. E., and S. R. Browning, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5: e1000384.

Maher, B., 2008 Personal genomes: the case of the missing heritability. Nature 456: 18–21.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.

Merikangas, K. R., and N. Risch, 2003 Genomic priorities and public health. Science 302: 599–601.

Metzker, M. L., 2010 Sequencing technologies—the next generation. Nat. Rev. Genet. 11: 31–46.

Morris, A. P., and E. Zeggini, 2009 An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34: 188–193.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin *et al.*, 2011 Testing for an unusual distribution of rare variants. PLoS Genet. 7: e1001322.

Nejentsev, S., N. Walker, D. Riches, M. Egholm, and J. A. Todd, 2009 Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387–389.

Nyholt, D. R., 2004 A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am. J. Hum. Genet. 74: 765–769.

Pinto, D., A. T. Pagnamenta, L. Klei, R. Anney, D. Merico *et al.*, 2010 Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368–372.

Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 86: 832–838.

Risch, N., 1990a Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. 46: 222–228.

Risch, N., 1990b Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am. J. Hum. Genet. 46: 242–253.

Risch, N., 1992 Corrections to "Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs". Am. J. Hum. Genet. 51: 673–675.

Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science 273: 1516–1517.

Sackrowitz, H. B., and E. Samuel-Cahn, 1999 P-values as random variables: expected P-values. Am. Stat. 53: 326–331.

Skellam, J. G. 1946 The frequency distribution of the difference between two Poisson variates belonging to different populations. J. R. Stat. Soc. Ser. A, 109: 296.

Stefansson, H., D Rujescu, S Cichon, O. P Pietilinen, A Ingason *et al.*, 2008 Large recurrent microdeletions associated with schizophrenia. Nature 455: 232–236.

Tucker, T., M. Marra, and J. M. Friedman, 2009 Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85: 142–154.

Weeks, D. E., and K. Lange, 1988 The affected-pedigree-member method of linkage analysis. Am. J. Hum. Genet. 42: 315–326.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Wright, S., 1938 Size of population and breeding structure in relation to evolution. Science 87: 430–431.

Communicating editor: S. F. Chenoweth

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2011/08/12/genetics.111.131813.DC1

Study Designs for Identification of Rare Disease Variants in Complex Diseases: The Utility of Family-Based Designs

Iuliana Ionita-Laza and Ruth Ottman

DOI: 10.1534/genetics.111.131813

File S1

S1 Effective number of variants in related individuals

We derive here in detail the results in the text, in particular relations (1) and (2).

Assumption S1. *We assume for simplicity that all individuals who carry a rare variant are in fact heterozygous. This assumption is reasonable for our setting since homozygous genotypes would have very low probability for rare variants ($f^2 \leq 10^{-4}$ for $f \leq 0.01$).*

We are interested in computing the joint distribution of the genotypes at a position in two related individuals. Let us code the genotype of an individual as 0, 1 or 2, denoting the number of rare variants (or minor alleles) in the genotype. For example, a homozygous genotype for the rare variant is denoted by 2 (however, we ignore individuals with genotype 2 by Assumption S1).

Claim S2. *Let $p_{\{i,j\}}$ be the probability that the genotypes in two related individuals are i and j, with $i,\ j \in \{0,1\}$. Let φ be the kinship coefficient for the two relatives, f be the minor allele frequency at a variant position, and δ be 1 if the two relatives can share two alleles IBD, and 0 if the two relatives can share a maximum of one allele IBD. Then the following relationships are true:*

$p_{\{0,0\}}$	$(1-f)^2 \cdot [1 - 2f(1-2\varphi) + f^2(1 - 4\varphi + 4\delta\varphi^2)]$
$p_{\{0,1\}}$	$4f(1-f)^2 \cdot [(1-2\varphi) - f(1 - 4\varphi + 4\delta\varphi^2)]$
$p_{\{1,1\}}$	$4f(1-f) \cdot [\varphi + f(1-f)(1 - 4\varphi + 4\delta\varphi^2)]$

Table S-1: Joint probabilities for the genotypes in two related individuals.

Proof. To calculate $p_{\{i,j\}}$ we condition on the IBD sharing between the two individuals. There are at most two alleles that can be shared IBD between two individuals, depending on the precise relationship. Let $p_{ij|k}$ be the probability that in a set of relatives the first genotype is i and the second genotype is j conditional on the IBD sharing being k, where $i,\ j \in \{0,1\}$ and $0 \leq k \leq 2$. We now calculate the probabilities $p_{ij|k}$ in turn.

- IBD sharing is 0 with probability

$$P(IBD = 0) = 1 - 4\varphi + 4\delta\varphi^2.$$

Because the genotypes at the two relatives are independent in this case it is relatively easy to calculate $p_{ij|0}$ (Table S-2).

The results in the table are based on multiplying the probabilities of individual genotypes. For example,

$$p_{01|0} = P(\text{genotype 1 is 0 and genotype 2 is 1}|\ IBD = 0) = (1-f)^2 \cdot 2f(1-f) = 2f(1-f)^3.$$

	Genotype 2	
Genotype 1	0	1
0	$(1-f)^4$	$2f(1-f)^3$
1	$2f(1-f)^3$	$4f^2(1-f)^2$

Table S-2: Joint probabilities for the genotypes in two related individuals, conditional on the IBD sharing being 0, i.e., $p_{ij|0}$.

- IBD sharing is 1 with probability

$$P(IBD = 1) = 4\varphi - 8\delta\varphi^2.$$

Results for $p_{ij|1}$ are in Table S-3.

	Genotype 2	
Genotype 1	0	1
0	$(1-f)^3$	$f(1-f)^2$
1	$f(1-f)^2$	$f(1-f)$

Table S-3: Joint probabilities for the genotypes in two related individuals, conditional on the IBD sharing being 1, i.e., $p_{ij|1}$.

The results in the table are calculated as follows:

$$\begin{aligned} & p_{00|1} = P(\text{genotype 1 is 0 and genotype 2 is 0}|\ IBD = 1) = \\ = \ & P(\text{genotype 1 is 0}|\text{ genotype 2 is 0 and } IBD = 1)P(\text{ genotype 2 is 0}|IBD = 1) = \\ = \ & (1-f)\cdot(1-f)^2 = (1-f)^3. \end{aligned}$$

Also

$$\begin{aligned} & p_{10|1} = P(\text{genotype 1 is 1 and genotype 2 is 0}|\ IBD = 1) = \\ = \ & P(\text{genotype 1 is 1}|\text{ genotype 2 is 0 and } IBD = 1)P(\text{ genotype 2 is 0}|IBD = 1) = \\ = \ & f\cdot(1-f)^2. \end{aligned}$$

Similarly

$$\begin{aligned} & p_{11|1} = P(\text{genotype 1 is 1 and genotype 2 is 1}|\ IBD = 1) = \\ = \ & P(\text{genotype 1 is 1}|\text{ genotype 2 is 1 and } IBD = 1)P(\text{ genotype 2 is 1}|IBD = 1) = \\ = \ & \left(\frac{1}{2}f + \frac{1}{2}(1-f)\right)\cdot 2f(1-f) = f(1-f). \end{aligned}$$

	Genotype 2	
Genotype 1	0	1
0	$(1-f)^2$	0
1	0	$2f(1-f)$

Table S-4: Joint probabilities for the genotypes in two related individuals, conditional on the IBD sharing being 2, i.e., $p_{ij|2}$.

- IBD sharing is 2 with probability

$$P(IBD = 2) = 4\delta\varphi^2.$$

In this case, the two genotypes are identical. Results for $p_{ij|2}$ are in Table S-4.

It is relatively easy to derive the results in the table. For example:

$$\begin{aligned} & p_{11|2} = P(\text{genotype 1 is 1 and genotype 2 is 1}|\ IBD = 2) = \\ = \ & P(\text{genotype 1 is 1}|\text{ genotype 2 is 1 and } IBD = 2)P(\text{genotype 2 is 1}|IBD = 2) = \\ = \ & 1 \cdot 2f(1-f) = 2f(1-f). \end{aligned}$$

From Tables S-2, S-3 and S-4 we get:

$$\begin{aligned} & p_{\{0,0\}} = P(\text{genotype 1 is 0 and genotype 2 is 0}) = \\ = \ & p_{00|0}P(IBD = 0) + p_{00|1}P(IBD = 1) + p_{00|2}P(IBD = 2) \\ = \ & (1-f)^2 \cdot [(1 - 4\varphi + 4\delta\varphi^2)(1-f)^2 + (4\varphi - 8\delta\varphi^2)(1-f) + 4\delta\varphi^2]. \end{aligned} \quad \text{(S3)}$$

$$\begin{aligned} & p_{\{0,1\}} = P(\text{genotype 1 is 0 and genotype 2 is 1 } OR \text{ genotype 1 is 1 and genotype 2 is 0}) = \\ = \ & 2P(\text{genotype 1 is 0 and genotype 2 is 1}) = \\ = \ & 2(p_{01|0}P(IBD = 0) + p_{01|1}P(IBD = 1) + p_{01|2}P(IBD = 2)) \\ = \ & 2f(1-f)^2 \cdot [2(1 - 4\varphi + 4\delta\varphi^2)(1-f) + (4\varphi - 8\delta\varphi^2)]. \end{aligned} \quad \text{(S4)}$$

$$\begin{aligned} & p_{\{1,1\}} = P(\text{genotype 1 is 1 and genotype 2 is 1}) = \\ = \ & 2(p_{11|0}P(IBD = 0) + p_{11|1}P(IBD = 1) + p_{11|2}P(IBD = 2)) \\ = \ & f(1-f) \cdot [4(1 - 4\varphi + 4\delta\varphi^2)f(1-f) + (4\varphi - 8\delta\varphi^2) + 8\delta\varphi^2]. \end{aligned} \quad \text{(S5)}$$

Using simple algebraic manipulation it is possible to simplify the above expressions for $p_{\{0,0\}}$, $p_{\{0,1\}}$, and $p_{\{1,1\}}$ (Table S-5).

We can easily check the two extreme cases: $\varphi = 0$, $\delta = 0$ (unrelated individuals, Table S-6) and $\varphi = 0.5$, $\delta = 1$ (identical twins, Table S-7).

□

$p_{\{0,0\}}$	$(1-f)^2 \cdot [1-2f(1-2\varphi)+f^2(1-4\varphi+4\delta\varphi^2)]$
$p_{\{0,1\}}$	$4f(1-f)^2 \cdot [(1-2\varphi)-f(1-4\varphi+4\delta\varphi^2)]$
$p_{\{1,1\}}$	$4f(1-f) \cdot [\varphi+f(1-f)(1-4\varphi+4\delta\varphi^2)]$

Table S-5: Joint probabilities $p_{\{i,j\}}$ for the genotypes in two related individuals.

$p_{\{0,0\}}$	$(1-f)^2(1-f)^2$
$p_{\{0,1\}}$	$(1-f)^2 \cdot 2f(1-f) + 2f(1-f) \cdot (1-f)^2$
$p_{\{1,1\}}$	$[2f(1-f)] \cdot [2f(1-f)]$

Table S-6: Joint probabilities for the genotypes in two unrelated individuals ($\varphi = 0$, $\delta = 0$).

$p_{\{0,0\}}$	$(1-f)^2$
$p_{\{0,1\}}$	0
$p_{\{1,1\}}$	$2f(1-f)$

Table S-7: Joint probabilities for the genotypes in two identical twins ($\varphi = 0.5$, $\delta = 1$).

Claim S6. *For a pair of individuals that each carry a rare variant in heterozygous state we calculate the effective number of variants for the pair as:*

$$\boxed{k_{\mathit{eff}|2} \cong \log_{2f}[4f\varphi + 4f^2(1-4\varphi+4\delta\varphi^2)]}.$$

Proof. We define the *effective* number of variants in a pair of individuals that are both heterozygous, denoted by $k_{\text{eff}|2}$, to satisfy the equation:

$$\begin{aligned} & P(\text{genotype 1 is 1 and genotype 2 is 1}|\text{neither genotype is 2}) = \\ = \; & P(\text{genotype 1 is 1}|\text{genotype 1 is not 2})^{k_{\text{eff}|2}}. \end{aligned}$$

When the two individuals are unrelated we can see that $k_{\text{eff}|2}$ should be 2. Similarly, when the two individuals are identical twins $k_{\text{eff}|2}$ should be 1, as expected.

In order to define $k_{\text{eff}|2}$ we need to calculate first the probability that two relatives are both heterozygous, given that they are not homozygous (as by Assumption S1). More precisely, for a pair of relatives we have:

$$\begin{aligned} & P(\text{genotype 1 is 1 and genotype 2 is 1}|\text{neither genotype is 2}) = \\ = \; & \frac{P(\text{genotype 1 is 1 and genotype 2 is 1})}{P(\text{neither genotype is 2})} = \\ = \; & \frac{p_{\{1,1\}}}{p_{\{0,0\}} + p_{\{0,1\}} + p_{\{1,1\}}} = \\ = \; & \frac{4f \cdot [\varphi + f(1-f)(1-4\varphi+4\delta\varphi^2)]}{(1+f) - f^2(1-4\delta\varphi^2) + f^3(1-4\varphi+4\delta\varphi^2)} \\ \cong \; & \frac{4f \cdot [\varphi + f(1-4\varphi+4\delta\varphi^2)]}{(1+f)} \cong 4f\varphi + 4f^2(1-4\varphi+4\delta\varphi^2). \end{aligned}$$

In the approximation above, we approximate $f(1-f) \cong f$ and $(1+f) - f^2(1-4\delta\varphi^2) + f^3(1-4\varphi+4\delta\varphi^2) \cong 1$ for $f \leq 0.01$. These approximations are very good, and have only negligible effect.

Also the probability that an individual is heterozygous, conditional on not being homozygous is:

$$P(\text{ genotype is } 1 | \text{ genotype is not } 2) = \frac{2f(1-f)}{1-f^2} = \frac{2f}{1+f} \cong 2f$$

if we assume that $1+f \cong 1$ for $f \leq 0.01$. As can be seen, the conditioning on the genotype not being 2 has negligible effect, so we will ignore it in subsequent calculations.

By definition the *effective* number of variants in a pair of individuals that are both heterozygous, denoted by $k_{\text{eff}|2}$, satisfies the equation:

$$\begin{aligned} & P(\text{genotype 1 is 1 and genotype 2 is 1}|\text{neither genotype is 2}) = \\ = \; & P(\text{genotype 1 is 1}|\text{genotype 1 is not 2})^{k_{\text{eff}|2}}. \end{aligned}$$

Using the relationships above we get:

$$\boxed{k_{\text{eff}|2} \cong \log_{2f}[4f\varphi + 4f^2(1-4\varphi+4\delta\varphi^2)]}.$$

Note that the above approximation is extremely good. It is easy to see that for two unrelated individuals, since $\varphi = 0$ and $\delta = 0$, we get $k_{\text{eff}|2} = 2$; for two identical twins, since $\varphi = 0.5$ and $\delta = 1$ we get $k_{\text{eff}|2} = 1$. These values match our intuition, namely, that for unrelated individuals the two variants are indepedent, while for identical twins there is only one independent variant.

□

If only one of the two relatives carries a rare variant, then the *effective* number of variants in the pair, denoted by $k_{\text{eff}|1}$, is 1. Similarly, if neither of the two relatives is carrier of rare variant, then $k_{\text{eff}|0} = 0$. For a pair of relatives we calculate the *effective* number of variants, denoted by k_{eff}, as follows:

$$k_{\text{eff}} = \begin{cases} k_{\text{eff}|2} & \text{if both relatives carry a rare variant} \\ k_{\text{eff}|1} & \text{if only one of the two relatives carries a rare variant} \\ k_{\text{eff}|0} & \text{if neither of the two relatives carries a rare variant} \end{cases}$$

Using the derivations above, we can now calculate the *effective* number of variants expected at a variant position in a pair of individuals:

$$\begin{aligned} & E[k_{\text{eff}}] = k_{\text{eff}|2}p_{\{1,1\}} + k_{\text{eff}|1}p_{\{1,0\}} = \\ = \; & k_{\text{eff}|2} \cdot 4f(1-f) \cdot [\varphi + f(1-f)(1-4\varphi+4\delta\varphi^2)] + \\ + \; & k_{\text{eff}|1}4f(1-f)^2 \cdot [(1-2\varphi) - f(1-4\varphi+4\delta\varphi^2)]. \end{aligned} \quad \text{(S7)}$$

S2 Effective number of variants in affected relatives at a *disease* locus

In a similar fashion, we can also calculate the *effective* number of variants we expect to observe at a *disease* locus, and derive relation (4) in text. We assume a two-locus genetic heterogeneity model, with either locus being an independent disease cause. Similar to the derivations above we calculate $p^D_{\{i,j\}}$, i.e., the probability that two affected relatives have genotypes i and j at the first disease locus, with $i,\ j \in \{0,1\}$. Using Bayes rule we obtain:

$$p^D_{\{i,j\}} = \frac{p_{\{i,j\}} P(\text{two affected} \mid \text{the two genotypes at the first disease locus are } i \text{ and } j)}{P(\text{two affected relatives})}.$$

If we let K be the disease prevalence, and K_R be the risk to a type R relative of an affected individual, then it is true that:

$$P(\text{two affected relatives}) = KK_R.$$

In order to compute

$$P(\text{two affected relatives}| \text{ the two genotypes at the first disease locus are } i \text{ and } j),$$

we need to introduce some notation.

Notation. *We follow closely the notations in Risch (1990a). Let x_i and y_j be the marginal "penetrance summands" at the two disease loci in our two-locus model, defined such that the penetrance $w_{ij} = 1-(1-x_i)(1-y_j)$; i and j index the possible genotypes at the two loci with $0 \leq i \leq 2$, and $0 \leq j \leq 2$. Let p_i and q_j be the population frequencies of those genotypes, and τ_{jk} be the conditional probability that the relative has genotype k given that the proband has genotype j. K_1 and K_2 are the prevalences for the two disease loci; K_{1R} and K_{2R} are the recurrence risks in relatives corresponding to the two disease loci.*

Then the following relationships hold (see Risch 1990a):

$$\begin{aligned}
K &= 1-(1-K_1)(1-K_2). \\
K_1 &= \sum_{i=0}^{2} p_i x_i. \\
K_2 &= \sum_{j=0}^{2} q_j y_j. \\
K_2 K_{2R} &= \sum_{j=0}^{2}\sum_{k=0}^{2} q_j \tau_{jk} y_j y_k.
\end{aligned}$$

With these notations, we can now show how to derive

$$P(\text{two affected relatives}| \text{ the two genotypes at the first disease locus are 1 and 1}).$$

$$
\begin{aligned}
&P(\text{two affected relatives}|\text{ the two genotypes at the first disease locus are 1 and 1}) = \\
&= \sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}[1-(1-x_1)(1-y_j)][1-(1-x_1)(1-y_k)] = \sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk} - \\
&-(1-x_1)\sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}(1-y_k) - (1-x_1)\sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}(1-y_j) + \\
&+(1-x_1)^2\sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}(1-y_j)(1-y_k) = \\
&= 1-2(1-x_1)(1-K_2)+(1-x_1)^2(1-2K_2+K_2K_{2R}) = \\
&= x_1^2+2x_1(1-x_1)K_2+(1-x_1)^2K_2K_{2R}.
\end{aligned}
$$

If we let

$$\beta_R = \frac{x_1^2+2x_1(1-x_1)K_2+(1-x_1)^2K_2K_{2R}}{KK_R}, \tag{S8}$$

then

$$p^D_{\{1,1\}} = p_{\{1,1\}}\beta_R.$$

Note that $\beta_R = 1$ when neither of the two-loci is involved in disease.

Similarly, we can calculate

$$P(\text{two affected relatives}|\text{ the two genotypes at the first disease locus are 0 and 1}).$$

$$
\begin{aligned}
&P(\text{two affected relatives}|\text{ the two genotypes at the first disease locus are 0 and 1}) = \\
&= \sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}[1-(1-x_0)(1-y_j)][1-(1-x_1)(1-y_k)] = \sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk} - \\
&-(1-x_1)\sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}(1-y_k) - (1-x_0)\sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}(1-y_j) + \\
&+(1-x_0)(1-x_1)\sum_{j=0}^{2}\sum_{k=0}^{2} q_j\tau_{jk}(1-y_j)(1-y_k) = \\
&= 1-(1-x_0)(1-K_2)-(1-x_1)(1-K_2)+(1-x_0)(1-x_1)(1-2K_2+K_2K_{2R}).
\end{aligned}
$$

Let

$$\alpha_R = \frac{1-(1-x_0)(1-K_2)-(1-x_1)(1-K_2)+(1-x_0)(1-x_1)(1-2K_2+K_2K_{2R})}{KK_R}, \tag{S9}$$

then

$$p^D_{\{0,1\}} = p_{\{0,1\}}\alpha_R.$$

Note that $\alpha_R = 1$ when neither of the two-loci is involved in disease.

For one pair of affected relatives the *effective* number of rare variants we expect to observe at the first disease locus is:

$$\begin{aligned} & E[k^D_{\text{eff}}] = p^D_{\{1,1\}}k_{\text{eff}|2} + p^D_{\{0,1\}}k_{\text{eff}|1} = \\ = \; & k_{\text{eff}|2} \cdot 4f(1-f) \cdot [\varphi + f(1-f)(1-4\varphi+4\delta\varphi^2)]\beta_R + \\ + \; & k_{\text{eff}|1}4f(1-f)^2 \cdot [(1-2\varphi) - f(1-4\varphi+4\delta\varphi^2)]\alpha_R \end{aligned} \tag{S10}$$

i.e., relation (4) in text. Note the similarity between this expression and that of $E[k_{\text{eff}}]$ in eq. (S7).

Calculation of α_R and β_R for a disease model. When comparing different study designs for a two-locus disease model, we need to calculate α_R and β_R for a specific disease model, characterized by the disease prevalence K, the overall λ_S, the GRR and frequency f at the first locus.

To calculate x_0 and x_1 in the expressions of α_R (eq. S9) and β_R (eq. S8) above we use the following derivations. If we let GRR be the genotype relative risk, i.e., $GRR = \frac{P(\text{aff}|Aa)}{P(\text{aff}|AA)}$, then it is possible to show using derivations similar to those used in the calculation of α_R and β_R that:

$$GRR = \frac{x_1 + (1-x_1)K_2}{x_0 + (1-x_0)K_2}.$$

Also

$$K_1 = \sum_{i=0}^{2} p_i x_i \approx p_0x_0 + p_1x_1.$$

(By Assumption S1, we ignore the low probability case of an individual being a homozygous carrier for the rare allele.) From the two equations above we can now calculate x_0 and x_1, and obtain:

$$\begin{aligned} x_1 &= \frac{K_1 + p_0K_2(GRR-1)/(GRR(1-K_2))}{p_1 + p_0/GRR}. \\ x_0 &= \frac{K_1 - p_1x_1}{p_0}. \end{aligned}$$

If we let $\lambda_{1R} = \frac{K_{1R}}{K_1}$, and $\lambda_{2R} = \frac{K_{2R}}{K_2}$; R be any relationship such as S =sibling, C = first cousins, and SC = second cousins, then as in Risch (1990a) we have:

$$\begin{aligned}
\lambda_R - 1 &= \frac{K_R}{K} - 1 = \left(\frac{K_1}{K}\right)^2 (\lambda_{1R} - 1) + \left(\frac{K_2}{K}\right)^2 (\lambda_{2R} - 1)\,. \\
\lambda_S - 1 &= 4(\lambda_C - 1) = 16(\lambda_{SC} - 1). \\
\lambda_{1S} - 1 &= 4(\lambda_{1C} - 1) = 16(\lambda_{1SC} - 1). \\
\lambda_{2S} - 1 &= 4(\lambda_{2C} - 1) = 16(\lambda_{2SC} - 1).
\end{aligned}$$

To calculate λ_{1S} we make use of the relationship between the genotype relative risk (GRR) and the sibling risk ratio at a single locus (λ_{1S}), as derived in Rybicki and Elston (2000). Based on these relationships, we now have all the relations required to calculate α_R and β_R for a specified two-locus disease model.

S3 Empirical distribution of $k_{\text{eff}}^{\text{D, Total}}$

We have performed simulations in order to generate the empirical distribution of $k_{\text{eff}}^{\text{D,Total}}$, the *effective* number of variants at the first disease locus in a two-locus heterogeneity model, assuming a range of values for the relevant parameters. In particular, we assume the disease prevalence to be 0.03, the frequency at the first disease locus to be 0.001 or 0.01, and the GRR to be between $1 - 4$; λ_S is 2 or 4. We used Risch's two-locus heterogeneity model, and simulated datasets consisting of 1000 affected sib-pairs, or 2000 unrelated affected individuals. We generated 10000 such datasets, and for each dataset we calculated the *effective* number of variants observed at the locus, $k_{\text{eff}}^{\text{D,Total}}$. In Figure S-1 we show the empirical distribution of $k_{\text{eff}}^{\text{D,Total}}$ when the GRR= $\{1,\ 2.5\}$, with the corresponding theoretical Poisson density function overlaid. The mean values for $k_{\text{eff}}^{\text{D,Total}}$ over the simulated datasets are shown in Table S-8, together with the analytical results in eq. (4) in text or eq. (S10) in the Supplemental Material S2. As shown, the concordance is extremely good.

				GRR			
Relationship	λ_S	f	Type	1	2	3	4
Sib-Pairs	2	0.01	A	31.3	55.59	88.39	128.7
			S	31.19	54.97	86.36	123.3
		0.001	A	3.11	5.59	9.11	13.63
			S	3.13	5.55	9.07	13.62
	4	0.01	A	31.30	42.82	58.63	78.24
			S	31.02	42.29	57.77	76.28
		0.001	A	3.10	4.28	5.98	8.18
			S	2.99	4.29	5.78	7.99
Unrelateds	2	0.01	A	39.56	76.48	112.1	146.3
			S	40.2	78.3	114.8	150.3
		0.001	A	3.99	7.86	11.70	15.53
			S	4.06	7.89	11.96	15.92
	4	0.01	A	39.56	76.48	112.1	146.3
			S	39.78	77.64	115.32	149.41
		0.001	A	3.99	7.86	11.70	15.53
			S	3.96	7.99	12.07	15.73

Table S-8: *Effective* number of variants expected at a disease locus, $k_{\text{eff}}^{\text{D,Total}}$, with a GRR of $1 - 4$ and frequency of $0.001 - 0.01$ in a dataset of 1000 affected sib-pairs, or 2000 unrelated affected individuals, with $\lambda_S = \{2,\ 4\}$. A two-locus disease heterogeneity model is assumed. Results based on analytical calculations (A) – using eq. (4) in text or eq. (S10) in Supplemental Material S2 – and 10000 simulated datasets (S) are shown.

S4 Locus specific recurrence risk to siblings

In Supplemental Table S-9 for each of the disease variants in Figure 1 we report the locus specific recurrence risk to siblings (λ_{1S}).

f	GRR	λ_{1S}
0.01	1.7	1.004
	2.1	1.010
	2.5	1.020
0.001	3.5	1.006
	4.5	1.010
	5.5	1.020
0.0001	8	1.005
	11	1.010
	14	1.020

Table S-9: The locus specific λ_S for the disease variants in Figure 1.

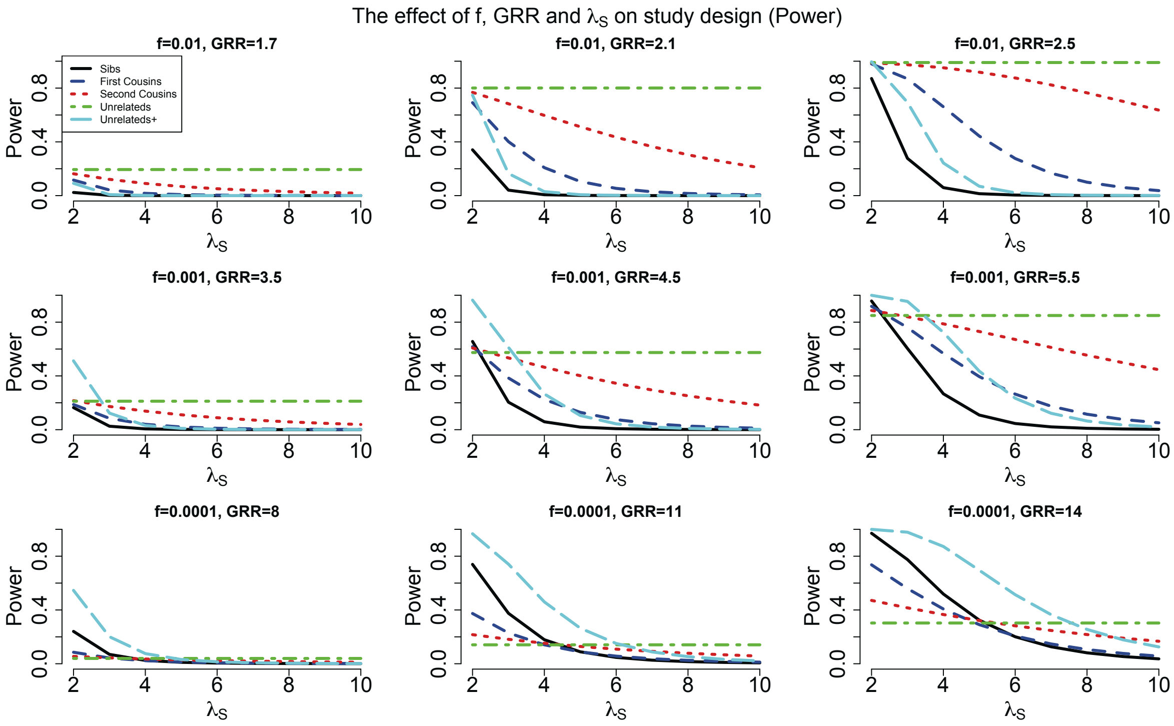

Figure S1 Empirical distribution of k_{eff} $^{D, Total}$ using 10000 simulated datasets of 1000 affected sib-pairs each, and a variant with frequency 0.01; GRR = 1 and GRR = 2.5; λ_S = 4. Overlaid are the Poisson density functions with their respective means.

In Supporting Figures S2-S4 we report results on the design performance based on Power rather than EPV (as done in the main text). The Power is for $\alpha = 1.6 \cdot 10^{-6}$ corresponding to the Bonferroni threshold when testing 30000 genes.

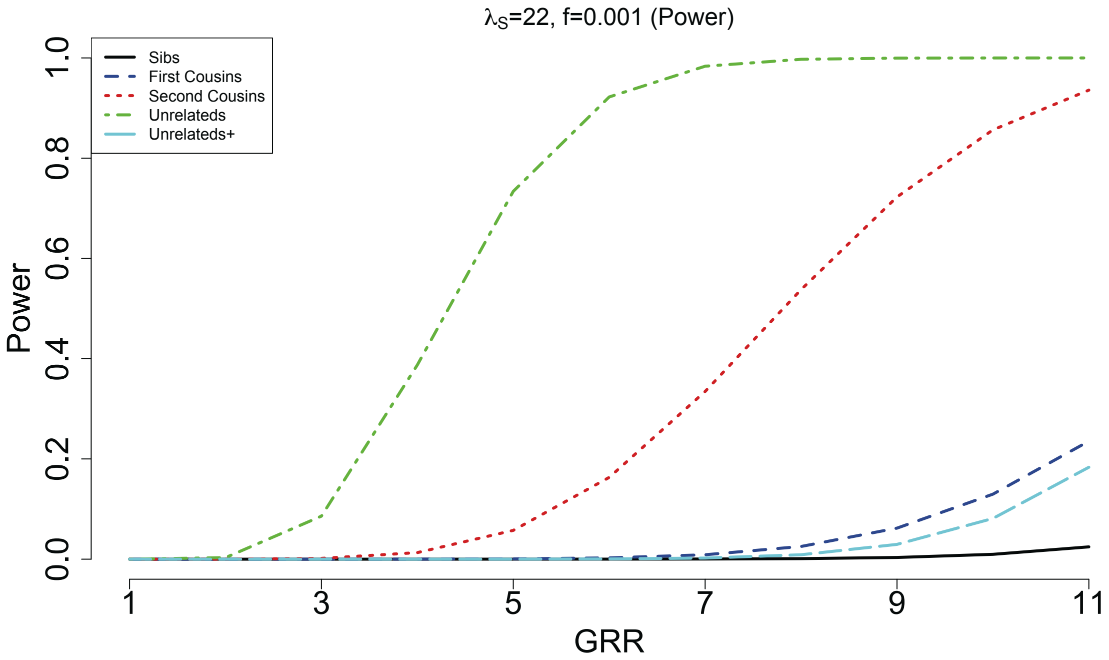

Figure S2 The effect of locus frequency, GRR, and overall λ_S on the relative performance (as measured by Power evaluated at $\alpha = 1.6 \cdot 10^{-6} = 0.05/30000$ if we assume 30000 genes) of affected relatives vs. unrelated affected individuals. The three rows correspond to three different frequencies for the disease locus: 0.01, 0.001 and 0.0001. λ_S is between 2 and 10. The number of affected individuals is 2000: 1000 sib-pairs, 1000 first-cousin pairs, 1000 second-cousin pairs, 2000 unrelateds, and 2000 unrelated individuals known to have an affected sibling (i.e., unrelateds +).

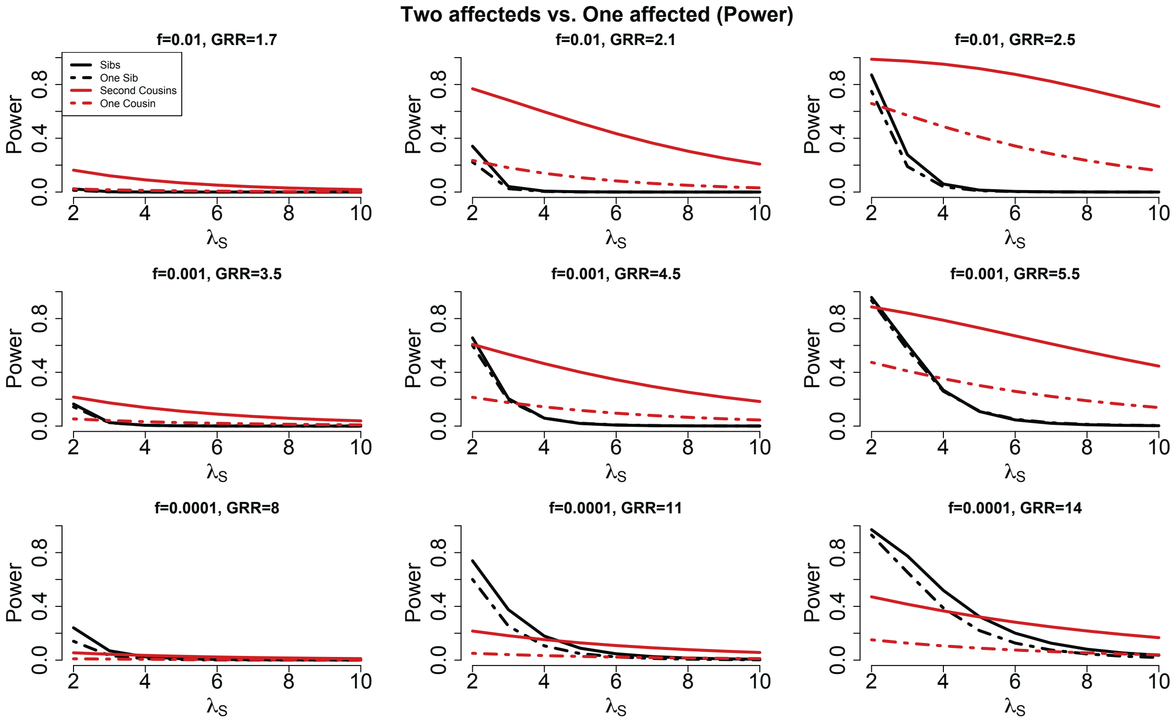

Figure S3 For complex traits with large values of λ_S (e.g., λ_S = 22 for autism) we show the Power (at $\alpha = 1.6 \cdot 10^{-6}$) at the first disease locus with frequency *f* = 0.001 as a function of the GRR for five study designs, each with 2000 affected individuals: 1000 sib-pairs, 1000 first-cousin pairs, 1000 second-cousin pairs, 2000 unrelateds, and 2000 unrelated individuals known to have an affected sibling (i.e., unrelateds +).

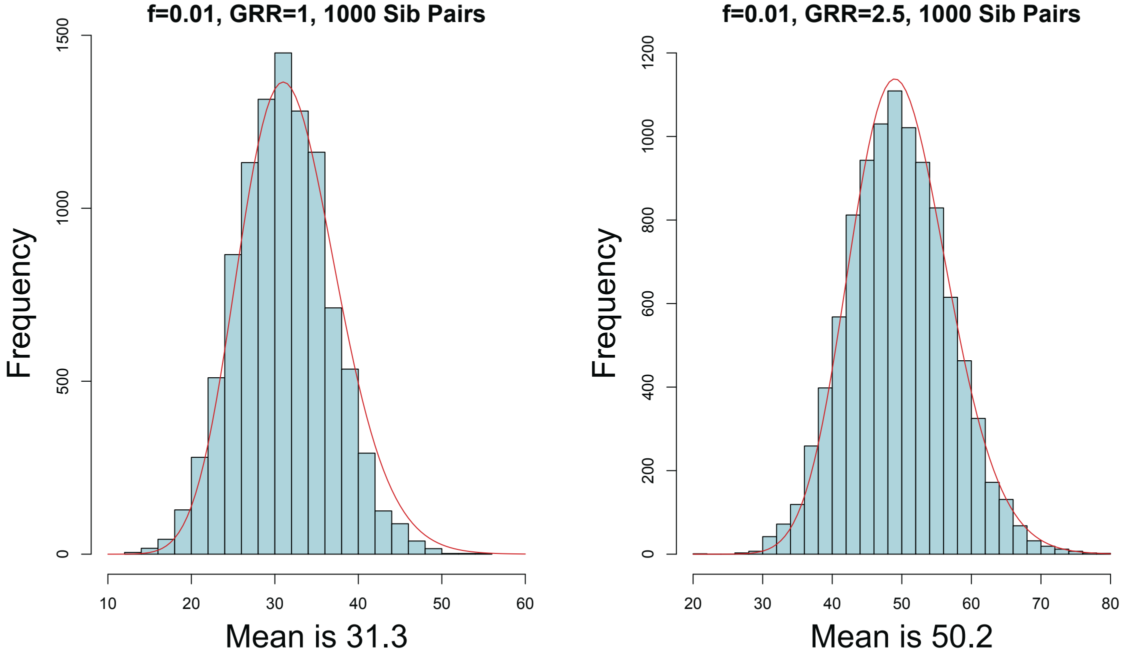

Figure S4 Usefulness of sequencing both affected individuals in a pair of affected relatives (comparisons based on Power at $\alpha = 1.6 \cdot 10^{-6}$). The three rows correspond to three different frequencies for the disease locus: 0.01, 0.001 and 0.0001. λ_S is between 2 and 10. The number of affected individuals is 2000 for the two affected relatives design, i.e., 1000 sib-pairs and 1000 second-cousin pairs, and only 1000 for the design based on only one affected individual in a pair, i.e., 1000 affected individuals known to have an affected sibling, and 1000 affected individuals known to have an affected second-cousin.

Rybicki, B. A., and R. C. Elston, 2000 The relationship between the sibling recurrence-risk ratio and genotype relative risk. Am. J. Hum. Genet. 66: 593–604.

File S2

Supporting software and documentation

File S2 is available for download at http://www.genetics.org/content/suppl/2011/08/12/genetics.111.131813.DC1 as a compressed folder.