European Journal of Human Genetics (2012) 20, 1085–1091

www.nature.com/ejhg

ARTICLE

Sampling strategies for rare variant tests in case–control studies

Sebastian Zöllner*[,1,2,3]

Advances in sequencing technology allow assessing the impact of rare variation on common disorders. For this purpose, methods combine rare variants across a gene and compare an aggregate statistic between cases and controls. However, sequencing many individuals is costly. Hence, it is necessary to identify case samples that are most likely to result in powerful tests under realistic model assumptions. Power can be increased by selecting cases that are highly likely to carry risk variants. As rare variants that contribute to the heritability of a disease co-segregate among affected family members, selecting cases that have affected family members may increase the power of rare variant tests considerably. Here I compare sequencing random cases to cases ascertained to have affected family members. I quantify the power of the different approaches and provide criteria for sample selection under different models of inheritance. Under a model of multiplicative gene–gene interaction, a sample of random cases has to be 2–16-fold larger to achieve the same power as a sample of cases ascertained to have affected family members. However, in traits with high heritability this power gain can be reduced or even reversed under models of additive gene–gene interaction. Hence study designs should depend on the studied disease's heritability and on the available sample size. I also show that selecting cases that share both chromosomes identical by descent with an affected sibling at candidate regions can result in a further power gain.
European Journal of Human Genetics **20,** 1085–1091; doi:10.1038/ejhg.2012.58; published online 18 April 2012

Keywords: rare variants; study design; family-based; burden test

INTRODUCTION

Genome-wide association studies have successfully identified many common variants that contribute to the risk of common disorders. However, identified variants have not explained the estimated heritability of most diseases and rare variants are now explored as likely contributors to disease risk.[1] Recently, functional rare variants have been identified in multiple genes that had previously been implicated by GWAS analysis[2] and implicated new genes as well.[3] Advances in sequencing technology now allow calling rare variation in large population samples of cases and controls on a genome-wide scale. Many studies use these data to assess the contribution of rare variation to the heritable risk of common disorders and to identify novel risk genes. However, single-marker tests of variants with low minor-allele count typically have insufficient power. To overcome this challenge, burden methods test genomic regions (typically genes) by combining putatively functional rare variants (eg missense variants) into aggregate statistics whose value is then compared between cases and controls.[4–7]

As sequencing studies are still costly, careful selection of sequenced samples is necessary to maximize power. Most genes only carry a small number of missense/nonsense alleles. Thus large sample sizes are required to achieve adequate power in case-control designs.[8] Studies can increase power by increasing the frequency difference of risk variants between cases and controls. This strategy has been successfully applied to quantitative traits such as plasma low-density lipoprotein levels[9,10] by selecting individuals from the extremes of the phenotypic distribution.

The equivalent strategy for binary traits such as disease affection status is selecting cases with multiple affected relatives.[11] Families with multiple affected relatives are more likely to segregate one or more risk variants and therefore cases sampled from such families are more likely to carry risk variants than random cases. This sampling method has been proposed in the past for common variants,[12–14] but the benefit for common variants with low effect size (odds ratio <1.2) is limited. As rare variants are expected to have higher effect sizes than common variants, gains from such strategies may be substantial.[15] However, such gains depend on underlying models of gene–gene interaction.[13,15]

Several features affect family-based designs for rare variants. First, little is known about the effect size distribution of rare variants. Presently, only a lack of linkage findings for most common complex diseases provides an upper bound on effect size. Second, each locus likely only contributes little to the overall heritability of a trait. Hence it is important to explore several models for interaction between the locus of interest and a large number of loci in the remaining genome. Third, when considering rare variants, it is necessary to model allelic heterogeneity, as each locus will carry multiple risk variants with differing effect sizes.

Here I explore a strategy of selecting cases conditional on having one affected relative. I develop closed-form equations that allow calculating the power of a burden test for a general model of rare risk variants where the effect sizes of variants at a locus are randomly distributed. On the basis of these equations, I examine the power of

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA; [2]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA; [3]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
*Correspondence: S Zöllner, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA. Tel: +734 647 9465; Fax: +734 763 2215; E-mail: szoellne@umich.edu
Received 31 October 2011; revised 9 February 2012; accepted 29 February 2012; published online 18 April 2012

burden test approaches under a wide range of scenarios consistent with an absence of linkage findings. I show that samples of cases collected conditional on having affected family members substantially outperform samples of random cases. This power gain depends on the distribution of effect size across risk variants. For realistic effect sizes the sample of random cases has to be 2–16-fold larger to achieve the same power as a sample of cases ascertained to have an affected family member. However, the gain in power is depended on the underlying model of gene–gene interaction. For models of additive interaction, the actual benefit of sampling conditional on affection status depends on the overall heritability of the trait.

I also consider re-sequencing studies that target candidate regions. For single regions, selecting cases that share the target segment with an affected family member further increases power. Selecting cases conditional on sharing two chromosomes with an affected family member can result in an increase in power equivalent to sequencing >10 times as many random cases.

MATERIALS AND METHODS

In the following, I calculate the summed frequency of rare risk variants at one locus of interest in cases sampled to have an affected relative, and from that frequency, the power of a burden test to identify this locus. To model linkage disequilibrium, I consider all haplotypes of risk variants at a locus rather than focusing on individual variants. By modeling the effect size of each haplotype as a random variable, haplotypes with multiple risk variants can be represented by having higher than average effect sizes. The overall heritability of the trait is affected by an unspecified number of unlinked loci. I consider two models for interaction between the genome and the locus of interest: a multiplicative interaction model under which each locus contributes independently to the heritability and an additive model.[16]

Genetic model

Assume a trait with prevalence K. For a pair of relatives with relationship status R, the probability of both relatives being affected is KK_R. I assume no inbreeding in either of the affected relatives. At the locus of interest, rare risk variants segregate in m distinct haplotypes $h_1, \ldots, h_m$; each haplotype carries an unspecified number of risk variants. Let $Pr(h_ih_j)$ indicate the probability of observing haplotypes h_i and h_j in an individual.

Sampling conditional on affected relatives

Let A indicate an affected individual and AA_R indicate a pair of affected individuals with relationship R. The probability of genotype h_ih_j in a case can be calculated by Bayes' Law:

$$\Pr(h_ih_j \mid A) \propto \Pr(A \mid h_ih_j)\Pr(h_ih_j). \quad (1)$$

When sampling cases conditional on having one affected relative of relationship status R, the probability of observing genotype h_ih_j in the index individual, is

$$\Pr(h_i \mid AA_R) \propto \Pr(AA_R \mid h_ih_j)\Pr(h_ih_j) \quad (2)$$

To calculate $Pr(AA_R|h_ih_j)$, I sum over all possible genotypes h_kh_l of the affected relative:

$$\Pr(AA_R \mid h_ih_j) = \sum_{k,l}^{m} \Pr(AA_R \mid h_ih_j, h_kh_l, R)\Pr(h_ih_j)\Pr(h_kh_l \mid h_ih_j, R). \quad (3)$$

$Pr(h_kh_l|h_ih_j,R)$ is calculated by conditioning on the number of chromosomes S shared identical by descent (IBD) by the relative pair. $Pr(AA_R|h_ih_j,h_kh_l)$ depends on the genetic model at the locus and the model of interaction with unlinked loci in the rest of the genome.

Gene–gene interaction

Consider an arbitrary number of risk loci that segregate independently of our locus of interest and result in a total of t multilocus genotypes. If multilocus genotype g_x, $x \in 1, \ldots, t$ has frequency $Pr(g_x)$, the probability of an individual being affected is

$$\Pr(A \mid h_ih_j) = \sum_{x=1}^{t} \Pr(A \mid h_ih_j, g_x)\Pr(g_x) \quad (4)$$

and the probability that a pair of relatives both are affected is:

$$\Pr(AA_R \mid h_ih_j, h_kh_l, R) = \sum_{x,y=1}^{t} \Pr(A \mid h_ih_j, g_x)\Pr(A \mid h_kh_l, g_y)\Pr(g_x)\Pr(g_y \mid g_x, R) \quad (5)$$

Assume we can separate the overall penetrance $Pr(A|h_i,h_j,g_x)$ into the penetrance component $\omega(h_ih_j)$ of genotype h_ih_j and the penetrance component $\Omega(g_x)$ of g_x. Interactions between loci in the remaining genome are then captured by Ω. The contribution to prevalence of this locus is then defined[16] as

$$K_L = \sum_{ij} \Pr(h_ih_j)\omega(h_ih_j)$$

and the locus' contribution to the recurrence risk (RR) among a pair of relatives is

$$K_LK_{LR} = \sum_{i}^{m}\sum_{j}^{m}\left[\Pr(h_ih_j)\omega(h_ih_j)\sum_{k}^{m}\sum_{l}^{m}\Pr(h_kh_l \mid h_ih_j)\omega(h_kh_l)\right]$$

The joint contribution of the rest of the genome to the prevalence can then be defined as $K_G = \sum_x \Pr(g_x)\Omega(g_x)$ and the contribution to the RR as

$$K_GK_{GR} = \sum_{x,y=1}^{t} p(g_x)p(g_y \mid g_x)\Omega(g_x)\Omega(g_y).$$

As shown below, $Pr(g_x)$, $\Omega(g_x)$ and $Pr(g_y|g_x, R)$ do not need to be specified beyond the overall prevalence and relative RR for the models under consideration.

Multiplicative interaction model. Under multiplicative interaction between the locus of interest and the remaining genome

$$\Pr(A \mid h_ih_j, g_x) = \omega(h_ih_j)\cdot\Omega(g_x). \quad (6)$$

The overall penetrance of the disease is then[16] $K = K_LK_G$ and $KK_R = K_GK_{GR}K_LK_{LR}$. By applying the definition (6) and factoring out $\omega(h_ih_j)$ in (4),

$$\Pr(A \mid h_ih_j) = K_G\cdot\omega(h_ih_j).$$

By solving (5) in a similar manner,

$$\Pr(AA_R \mid h_ih_j, h_kh_l) = \omega(h_ih_j)\omega(h_kh_l)\cdot K_GK_{GR}. \quad (7)$$

As K_GK_{GR} is present only as a multiplicative constant in this calculation, it will cancel when normalizing in (2) and thus it does not affect the probability of observing h_i.

Additive interaction model. Under additive interaction between the locus of interest and the remaining genome,

$$\Pr(A \mid h_ih_j, g_k) = \omega(h_ih_j) + \Omega(g_k). \quad (8)$$

The overall penetrance is $K = K_L + K_G$. The probability of observing an affected relative pair is[16] $KK_R = K_GK_{GR} + K_LK_{LR} + 2K_GK_L$. Thus, $K_G = K - K_L$ and $K_GK_{GR} = KK_R - K_LK_{LR} - 2K_L(K - K_L)$.

The probability of being affected conditional on carrying haplotypes h_i,h_j is then $Pr(A|h_ih_j) = K_G + \omega(h_ih_j)$ and the probability of an affected relative pair conditional on carrying haplotypes h_i,h_j,h_k,h_l is

$$\Pr(AA_R \mid h_ih_j, h_kh_l) = \omega(h_ih_j)\omega(h_kh_l) + \left(\omega(h_kh_l) + \omega(h_ih_j)\right)K_G + K_GK_{GR}$$

Effect size model

Let a proportion p of haplotypes carry one or more risk variants. Let $H \in \{0,1,2\}$ indicate the number of haplotypes with at least one risk variant in a sampled individual. The power of a burden test depends on the frequency of rare variant carrying haplotypes in cases, which is calculated from $Pr(H|A)$ in random cases and $Pr(H|AA_R)$ in selected cases. These probabilities can be

calculated by rewriting (2) and summing over the expected sharing S.

$$P(H \mid AA_R) \propto \sum_{S=0}^{2} P(S \mid R)P(H)P(AA_R \mid H,S). \quad (9)$$

The calculation depend on the penetrance model for $\omega(h_ih_j)$ are described below.

Multiplicative effect size model. The relative risk ω_i of haplotype i carrying a risk variant is sampled from a distribution f with expectation μ and variance σ^2. The relative risk of haplotypes not carrying a risk variants is 1. For all haplotypes i,j, I assume $\omega(h_ih_j)=\omega_i\omega_j$. As shown in Supplementary Information, all penetrances depend only on $\int \omega f(\omega)d\omega$ and $\int \omega^2 f(\omega)d\omega$, hence only the two moments of ω need to be specified. As the effect of all haplotypes is sampled from a distribution that is only specified by its first two moments, I assume without loss of generality, that each haplotype occurs only once in the population and modify the variance accordingly. Assuming Hardy–Weinberg Equilibrium (HWE) in the underlying population, the expected contribution to prevalence of the risk locus K_L is then

$$E(K_L) = (1+p(\mu-1))^2$$

The expected contribution to RR K_LK_{LR} among a pair with relationship R is

$$E(K_LK_{LR}) = \Pr(S=0 \mid R)E(K_L)^2 + \Pr(S=1 \mid R)E(K_L)\left(1+p\left(\mu^2+\sigma^2-1\right)\right) + \Pr(S=2 \mid R)\left(1+p\left(\mu^2+\sigma^2-1\right)\right)^2.$$

Details of calculating $Pr(AA_R|H,S)$ and proofs for the above equations are presented in Supplementary Text S1.

Additive effect size model. A proportion p of haplotypes carry risk variants and the risk contribution ω_i of each risk haplotype i is sampled from a distribution f with expectation μ and variance σ^2. The remaining haplotypes have risk contribution is 0. For all haplotypes i,j, let $\omega(h_ih_j)=\omega_i+\omega_j$. Again assuming each risk haplotypes occurs only once and that risk haplotypes are in HWE in the general population, $E(K_L)=2p\mu$ and

$$E(K_LK_{LR}) = \Pr(S=0 \mid R)\cdot E(K_L)^2 + \Pr(S=1 \mid R)\cdot\left(3p^2\mu^2+p\left(\mu^2+\sigma^2\right)\right) + (S=2 \mid R)\cdot 2p(\mu^2+\sigma^2+p\mu^2).$$

Details for this derivation and for calculations of $Pr(AA_R|H,S)$ are presented in Supplementary Text S2.

Other modeling concerns

Mis-specification. Markers that are included in a burden test without affecting the trait of interest can be modeled by adjusting the mean and variance of the effect size of functional variants. Assume a proportion $(1-q)$ of haplotypes without risk variants is falsely included in the test statistic. The remaining haplotypes have an effect size sampled from a distribution with mean μ_F and variance σ_F^2. Then, the mean and variance of included haplotypes is $\mu=q\mu_F$ in the additive model or $\mu=1+q(\mu_F-1)$ in the multiplicative model, and the variance is $\sigma^2=q\sigma_F^2+q(1-q)\mu_F^2$ in the additive model and $\sigma^2=q\sigma_F^2+q(1-q)(\mu_F-1)^2$ in the multiplicative model.

Power calculations. The modeled test uses a χ^2 test of independence in a sample of N_A affected individuals and N_U unaffected individuals to compare the number of haplotypes with at least one rare risk variant in affected individuals C_A and the number of haplotypes with at least one risk variant in random individuals C_U with $E(C_U)=2pN_U$. Under the null hypothesis of no effect, $E(C_A)=2pN_A$. Under the alternative, $E(C_A)=N_AE(H|A)$ if cases are sampled at random from the population, or $E(C_A)=N_AE(H|AA_R)$ if cases are sampled conditional on having an affected relative. The expectations for H can be calculated using equation (9) and from these expectations, the noncentrality parameter under the alternative is obtained. On the basis of the noncentrality parameter, I calculated the sample size required to achieve 80% power at a false positive rate of 10^{-6} for a range of parameters. This false positive rate maintains an experiment-wide type 1 error of 0.05 after Bonferroni correction for testing 50 000 regions in the genome and thus indicates genome-wide significance in a burden test.

Linkage test. To identify parameter settings that are consistent with an absence of linkage findings, I calculated the power of a genome-wide linkage scan using N affected sibpairs. On the basis of the model described above, the probability of sharing 0,1, or 2 alleles IBD in a pair of affected siblings conditional of the parameters p, μ, σ^2 can be calculated. Using those probabilities, I calculated the probability of the observed sharing being significantly higher than 1 at a genome-wide[17] significant $\alpha=10^{-5}$.

Sampling conditioning on sharing

If cases are selected from affects sib-pairs, it is possible to only select cases that share two chromosomes IBD with the other sibling. Then, the power of a burden test depends on $E(H|AA_R,S=2)$, which can be calculated with the equations given above. As for rare variants $Pr(H=i|AA_R,S=2)>Pr(H=i|AA_R)$ for $i=1,2$ and $p<\frac{1}{m+1}$ (multiplicative model; m is the mean relative risk) or $p<0.5$ (additive model). Thus $E(H|AA_R,S=2)>E(H|AA_R)$, therefore sampling conditional on sharing two haplotypes IBD has more power than sampling based on having an affected relative.

RESULTS

In the following, I compare using cases that are randomly selected (random cases) to cases that are selected based on having an affected sibling (selected cases) under a model where multiple risk variants with different effect sizes occur at a locus of interest. The distribution of effect sizes is specified only by its mean and variance. Moreover, I consider multiplicative and additive models of gene–gene interaction. These models are quite general; they are unaffected by the precise genetic architecture in the remaining genome. Finally I consider a study design that tests a region of interest by selecting cases from sibpairs that share two chromosomes IBD.

Multiplicative interaction

Assuming a model of multiplicative interaction (which could also be considered as a model of no interaction), I calculated the summed risk allele frequency in cases p_A for a random sample of cases and for a sample of cases selected to have an affected sibling assuming different summed population allele frequencies p and varying the mean relative risk m and the variance σ^2 of the effect size distribution (Figure 1). Under this model, the power in a design using selected cases is independent of the remaining genome and the population prevalence (see Materials and Methods). In samples taken from random cases, p_A increases almost linearly with f and m for small values of p (Figure 1a). In selected cases, p_A increases much faster. For example risk variants with $p=0.01$ and $m=3$ have a frequency of 0.029 in random cases and 0.057 in selected cases. The variance of risk between variants has no effect on p_A in random cases. In selected cases, p_A increases considerably with increasing variance (Figure 1b). This increase in frequency is observed for all values of p and m. Especially for low m, the frequency of haplotypes with risk variants can double in cases when comparing a model with variance 10 to a model with variance 0.

Using the p_A shown in Figure 1, I evaluated the performance of a simple burden test at a false positive rate of 10^{-6} (see Materials and Methods) by calculating the sample size required to achieve 80% power. I also calculated the power in a linkage study of 1000 affected sibpairs and indicated the range of parameters that is consistent with low power ($<10\%$) for positive findings using linkage. For samples of random individuals, the required sample size decreases with increasing relative risk m and with increasing summed minor allele frequency p (Figure 2a). However, high values of p and m are not consistent with the absence of strong linkage findings. In general, $m>4$ and $p>0.01$ result in linkage power $>10\%$. For parameter settings consistent with low linkage power, large sample sizes >400 random cases and controls are required for adequate power to achieve genomewide

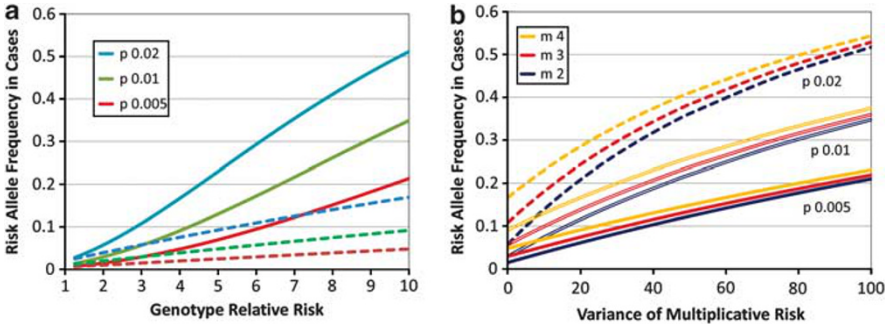

Figure 1 Summed risk allele frequency in cases. (**a**) Summed risk allele frequency in cases dependent on average effect sizes and summed population allele frequencies of risk variants. Samples drawn from random cases are shown as broken lines; samples drawn conditional on having an affected sibling are shown as solid lines. (**b**) Summed risk allele frequency in cases dependent on variance of effect sizes between risk variants. Broken lines represent a summed population frequency of risk variants $p=0.02$, double lines show results for $p=0.01$ and simple lines show results for $p=0.005$. Each color represents a mean multiplicative risk for each haplotype.

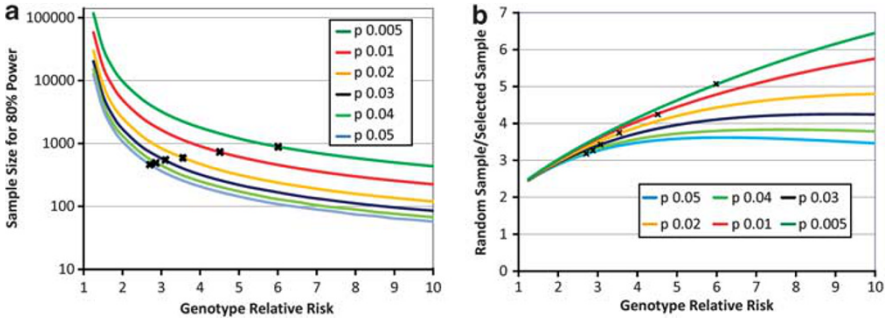

Figure 2 Sample size required for 80% power for genome-wide significant burden test assuming different mean genotype effect. Each line represents a summed population frequency; the X represents the genotype relative risk at which a region would obtain 10% linkage power ($\alpha=10^{-5}$). (**a**) The power generated by sampling random cases, (**b**) represents the ratio of sample sizes when sampling random cases to sampling cases conditional on having an affected sibling.

significance ($\alpha=10^{-6}$), regardless of p and m. For effect sizes more comparable with what is seen in common variants ($m=1.5$), sample sizes of 8300 random cases are required even for large values of $p=0.02$.

Sampling selected cases decreases the required sample sizes substantially (Figure 2b). For the lowest effect sizes considered ($m=1.25$), using selected cases reduces the required sample size to achieve the same power by a factor of 2.5 regardless of p. With increasing effect sizes, the benefit of using selected cases increases further. For $m=1.5$, $p=0.02$ the required sample size of selected cases is 3150; compared with 8100 random cases; for $m=2$, $p=0.02$ the required sample size is 840 selected cases compared with 2500 random cases. The relative benefit of using selected samples increases faster for small values of p. For the maximum effect size parameters consistent with the absence of linkage, the reduction ranges from 3.2-fold ($p=0.05$) to 5-fold ($p=0.005$).

As the variance of effect sizes across risk variants σ^2 does not affect p_A in random cases, the power of a burden test using random cases is independent from σ^2. On the other hand, when sampling selected cases, p_A increases as σ^2 increases and therefore the required sample size decreases (Figure 3). As σ^2 gets large this power is mostly determined by p and σ^2, and converges to the same value for all m. However, as σ^2 increases, so does linkage power, hence $\sigma^2>20$ is incompatible with an absence of linkage findings for all parameter setting considered here. But even for smaller σ^2 the reduction in required sample size in models with high variance can be considerable. For a model of moderate effect size and low cumulative frequency ($m=2$, $p=0.005$) the required sample size of random cases is 9750. If risk variants included in the test have the same effect

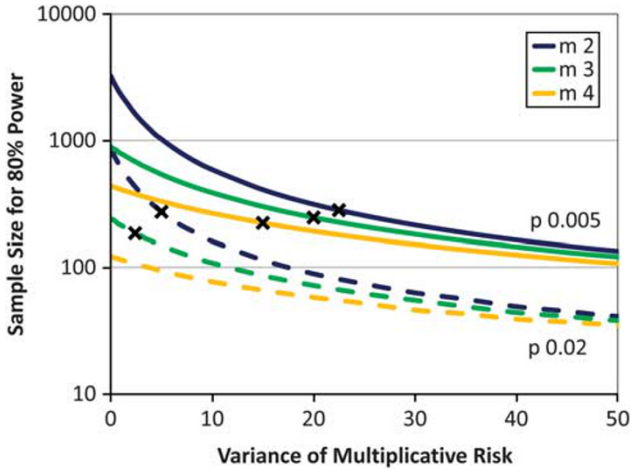

Figure 3 Sample size required to achieve 80% power in a genomewide significant burden test dependent on the variance of effect size among risk haplotypes. Broken lines represent results generated for a summed population allele frequency $p=0.02$; solid lines show results for $p=0.005$. The X represent the parameter setting for 10% linkage power.

size $\sigma^2 = 0$, the required selected sample size is 3230. For higher heterogeneity among effect sizes consistent with an absence of linkage ($\sigma^2 = 10$), the required sample size of a selected sample is 590, a 16-fold reduction in sample size. Such a σ^2 could for example be the result of most (94%) of variants having a relative risk of 1.2, whereas the remaining 6% of variants have a relative risk of 15.

An important contributor to σ^2 is the false inclusion of nonfunctional variants in the burden test. In practice, some variants included in a burden test at a true risk locus will not affect the trait of interest, thus decreasing the power of burden tests regardless of sampling strategy. However, including variants with no effect on disease risk also increases the variance of the effect size, thus increasing the power of a burden test in a sample of selected cases. Therefore, false inclusion of nonfunctional variants has a reduced power loss in designs using selected cases. This results in a higher benefit of using selected cases when a large proportion of variants are falsely included (Supplementary Figure 1), especially for variants with high effect size For $m = 5$ (blue line), the sample size of random cases is 4.4 times the sample size of selected cases for 0 false inclusion. This ratio increases to 7.2-fold for 80% false inclusion. For $m = 2$, this ratio increases from 3-fold to 3.7-fold over the same range. Note that the random/selected ratio increases, although m decreases when false inclusion increases (see Materials and Methods). With constant σ^2 decreasing m would result in a reduction in random/selected ratio (Figure 2b). However, as the misspecification increases, the variance of the relative risk also increases, resulting in the sample size ratio increasing instead. Hence, the benefit of using selected samples increases with increased number of falsely included variants.

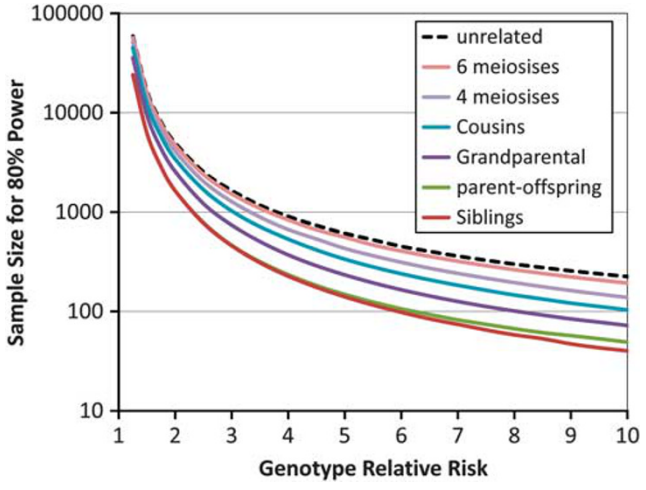

Figure 4 Benefit of conditioning on affected relatives with different relationships for a range of average genotype relative risk (horizontal axis).

Effect of relationship

So far I considered only the benefit of sampling affected individuals conditional on having an affected sibling. For comparison, I calculated the required sample size for sampling cases based on having an affected relative separated by up to six meiosises in a unilineal relationship. Figure 4 shows the sample size required to achieve 80% power in a genome-wide study for summed risk allele frequency $p = 0.01$, for other frequencies the results are similar. The benefit of conditioning on an affected relative is strongly dependent on the number of meiosises between the relatives. As sharing drops between distantly related relatives, the expected reduction in sample size from selected cases converges toward the sample size required from unconditional samples (black dotted line). For relationship-pairs split by >4 meioses, the benefit of conditioning on an affected relative is barely noticeable, resulting in <1.3-fold reduction in sample size for all m.

For cases sampled from relationships where both affected individuals share 50% IBD (siblings and parent-offspring pairs), the reduction in sample size is identical for $m<4$. Only for >4, the average IBD sharing of siblings exceeds the IBD sharing between parents and offspring. Therefore, linkage scans start having power for these values and conditioning on affected siblings performs better than conditioning on affected parents.

Additive model

In the additive model of interaction, the probability of observing a phenotype is the sum of the locus specific contribution and the contribution of the remaining genome. The additive risk contribution at the locus of interest is distributed with mean μ and variance σ^2. Under this model, p_A will depend on the RR between the relative pair (see Materials and Methods) in addition to p, μ and σ^2. For diseases with low RR = 2 the frequency of risk variants in selected cases is slightly lower than in random cases for $\mu<2$ but it increases much faster with μ than the frequency in random cases (Figure 5a). For diseases with higher RR, p_A in selected cases is lower than p_A in random cases for a wider range of average effect sizes. This effect of heritability is reflected in the sample size requirements of burden tests using selected cases and random cases (Figure 5b). For diseases with

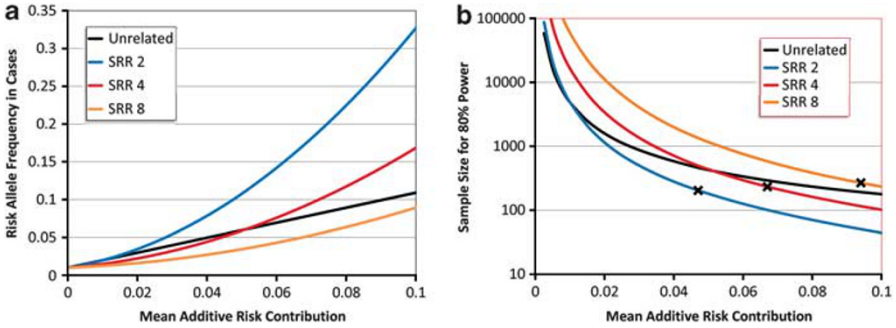

Figure 5 Family-based sampling on an additively interacting locus. I modeled a single locus with summed population frequency $p = 0.01$ and a prevalence $K = 0.01$. The X represent the parameter setting with 10% linkage power. Results are shown for three diseases with overall recurrence risks (RR) 2 (blue line), 4 (red line) and 8 (yellow line), as well as for random samples (black line). (**a**) Summed risk allele frequency in cases. (**b**) Sample size required for 80% power for genome-wide significant burden test.

Table 1 Benefit of sampling cases conditional on sharing two chromosomes identical by descent with an affected relative

	Summed risk allele frequency in cases						*Power in 1000 case/1000 controls*					
	Sharing 2			*Conditional sample*			*Sharing 2*			*Conditional sample*		
m\p	0.005	0.01	0.02	0.005	0.01	0.02	0.005	0.01	0.02	0.005	0.01	0.02
1.5	0.01	0.02	0.04	0.01	0.02	0.04	0.00	0.03	0.27	0.00	0.00	0.04
2	0.02	0.04	0.08	0.02	0.03	0.06	0.21	0.87	1.00	0.02	0.30	0.92
2.5	0.03	0.06	0.11	0.02	0.04	0.08	0.94	1.00	1.00	0.35	0.96	1.00
3	0.04	0.08	0.16	0.03	0.06	0.11	1.00	1.00	1.00	0.90	1.00	1.00

I modeled a disease with prevalence 0.01 and assumed a multiplicative model of interaction with 0 variance of effect size. The first three columns display the cumulative frequency of risk variants in cases sampled conditional on sharing two chromosomes with an affected sibling for average genotype relative risks varying from 1.5 to 3 and summed allele frequencies in random individuals of 0.005, 0.01 and 0.02. The second three columns show the same frequency in cases that are sampled conditional on having an affected sibling. The third three columns show the power in a population study of 1000 cases sharing two chromosomes with an affected sibling and 1000 random controls and the last three columns show the power in a population study of 1000 cases with an affected sibling and 1000 random controls.

RR = 2, the sample size requirements using random samples are similar or lower. For a disease with RR = 4, the power of using selected cases is smaller than the power of using a random cases for $\mu < 0.045$ and larger for bigger additive risk contributions. For highly heritable diseases (RR = 8), using selected cases is only advantageous for $\mu > 0.13$. Note that the linkage power at a locus also depends on the overall heritability, hence for large RR, larger values of μ are consistent with the absence of linkage.

Sampling conditional on sharing

When exploring specific regions in the genome, cases can be collected conditional on their degree of sharing with the affected family member. Table 1 compares the expected values of p_A in a sample of unrelated cases from affected sibpairs to the expected values of p_A in cases that share 2 chromosomes IBD at the locus of interest. On the basis of these values of p_A, I calculated the power in a sample of 1000 cases and 1000 controls for m between 1.5 and 3. In individuals sampled to be IBD 2, p_A is notably higher than p_A in cases sampled conditional on having an affected sibling only (Table 1). This p_A is in turn higher than p_A in random cases (Figure 1). Hence cases sampled conditional on IBD status have substantially higher power in an association test. For example, in a model with $m = 2.5$ and $p = 0.005$, the power of a study of 1000 random cases and 1000 controls is 0.002, although the power in a study collecting cases conditional on having affected relatives is 0.345 and the power of a study collecting cases that share two chromosomes IBD with an affected relative is 0.935.

DISCUSSION

Burden tests are expected to identify new genes for many common complex diseases. I have shown that for the rare variant allele frequencies observed in many genes[8] such tests likely require large sample sizes of at least several thousand cases and controls to achieve genome-wide significance. Further, I have discussed designing more powerful case–control studies of rare variation by sampling cases with a family history of being affected. In particular, I showed that samples with one affected close relative carry substantially more risk alleles than random samples. This increase in risk allele count increases the power of burden tests. This benefit is particularly pronounced under a model of multiple risk variants with varying effect sizes segregating at the same locus. For plausible models, the required sample of randomly selected cases is 16 times as large as the sample required for cases selected conditional on family history and more extreme models are conceivable. Even for effect size distributions with no power in large linkage studies, using selected samples results in a substantial gain in power. This benefit is maximal if cases are sampled conditional on having affected siblings and becomes progressively smaller if the second affected individual is more distantly related. I also considered a scenario where a specific region of interest, eg, a linkage peak was followed up by sequencing affected individuals conditional on sharing both chromosomes with an affected sibling. This strategy is considerably more powerful than just collecting cases conditional on affected relatives.

Beyond the upper bound on effect sizes provided by the absence of convincing linkage results, little data exists to support specific assumptions about the frequency distribution or the effect size distribution of rare variants affecting common diseases. Therefore, I developed equations for a general model of risk variants specified only by the average effect size of risk variants at one locus, the variance of the effect sizes across risk variants at one locus and the summed frequency of all risk variants. To calculate power under such a general model, I used a basic burden test comparable to methods proposed by Li and Leal[5] and Zawistowski *et al*.[4] For more complicated burden tests, the power gain depends on more specific aspects of the genetic architecture, such as the allele frequency of individual risk variants. However, the general conclusions of my results still apply as an increase of risk allele frequency in case samples will increase the power of any burden test.

In particular, our results can be extended to models that assume both protective and causal rare variants at the same locus. Under this scenario, tests that model both types of variants[6,7] may be more powerful than tests that assume that the effect of all rare variants has the same direction. Again, calculating the power of such tests requires a more specific model of rare variant architecture. However, regardless of the specific architecture, the variance of effect sizes is high if a locus has both protective and causal variants. Hence, cases sampled conditional on having affected relatives substantially increase the number of risk alleles in the case sample under this scenario.

Sampling cases conditional on having affected relatives has been originally proposed by Risch.[11] Li *et al*[13] have shown that this design can also increase power for single-marker tests of more common risk variants, however, substantial gains in power are only archived for relatively high effect sizes (relative risk ≥ 1.4). More recently, Peng *et al*[14] have shown that for variants with relative risks between 1.2 and 1.4 the benefit of sampling from affected sib-pairs increases with decreasing allele frequency. Finally Ionita-Laza and Ottman[15] studied the effect of family-based sampling on single-marker tests of rare variants, alhough focusing on a model of genetic heterogeneity that is similar to the model of additive penetrance presented here. For models where all risk variants have the same effect size their conclusions are similar to my results. Here, I illustrate that the gain

in power from using cases conditional on affected relatives is more pronounced for rare variants with high effect sizes that for common variants with low effect sizes. Moreover, I show that the benefit of family-based sampling is increased substantially under a model where effect sizes vary between risk variants. This scenario is likely for burden tests for two reasons: First, burden tests typically aim to combine the evidence across all missense and nonsense mutations. It is unlikely that all such variants have the same effect on disease risk. Second, it is not clear if all variants included in a burden test have any effect at all; in fact it seems likely that many included variants do not significantly affect the disease risk. To explain this power gain from family-based sampling in models with high variance of effect sizes, consider that in a scenario with intermediate mean effect size and high variance, some risk variants at the locus of interest will have low effect size and some variants at the same locus will have high effect size. A family with multiple affected individuals is substantially more likely to segregate the variants with high effect size without being less likely to segregate the variants with low effect size. Thus the overall number of risk variants observed in samples from high-risk families is increased.

A common concern when considering family-based sampling is the possibility of a segregating variant with high effect size that is oversampled in affected families. Under some models of interaction, this can in turn result in undersampling of other risk variants.[11] This will result in reduced power to identify other risk loci. As can be seen in my results, such 'crowding' out is not possible under a model of multiplicative interaction. Under a model of additive interaction such crowding out only depends on the overall RR among relatives, not on the effect size of specific variants. Thus, even if no single variant with high effect size is present, such crowding out is possible in diseases with high RR between the ascertained relatives. However, it is not clear how common additive interaction is in rare variants. Presently, all attempts at replicating findings of non-multiplicative gene–gene interaction have failed,[18] suggesting that between most common variants multiplicative interaction is often an appropriate model. Moreover, there are several strategies to avoid crowding out in diseases with high heritability. First cases can be chosen conditional on their genotype at known risk variants with high effect size. Second, selecting more distantly related relatives will reduce the RR.[15] Although selecting more distantly related relatives also reduces the benefit of conditional sampling, it can still result in an increase of power. A practical concern for family-based sampling designs may be the ease of ascertaining families. Especially for rare diseases, it may be very costly to collect families; in such cases the power gain of sampling families has to be evaluated together with the increased costs of generating such samples.

In summary, I have demonstrated that under a wide range of genetic models, sampling cases with affected relatives result in substantial power gains for rare variant sequencing studies over designs of sampling random cases and controls. Such power gains may be necessary to generate genome-wide significant results, especially if the summed frequency of rare variants is low in many genes.[8] However, in diseases with high sibling relative risk, family-based sampling may reduce power to detect genomic locations that interact additively with the remaining genome. Hence for such traits with high sibling relative risk (≥ 4), the optimal design depends on the available sample size. Small random samples (eg, <500 random cases) likely provide insufficient power to overcome Bonferroni correction for any locus, regardless of the underlying architecture. Hence using cases with affected relatives is advantageous as it increases the power of identifying those loci that interact multiplicatively. However, when larger case samples are sequenced for traits with high RR, random samples may be preferable, as power to map individual loci will be less dependent on the underlying model of gene–gene interaction. On the other hand, for diseases with low sibling relative risk (<4), sampling cases conditional on having affected relatives will almost always result in substantial gains in power and is thus advantageous over sampling random individuals.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

I thank Margit Burmeister, Michael Boehnke and Laura Scott for their helpful discussion. This work was supported by HG 005855.

1 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
2 Emison ES, Garcia-Barcelo M, Grice EA *et al*: Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *Am J Hum Genet* 2010; **87**: 60–74.
3 Holm H, Gudbjartsson DF, Sulem P *et al*: A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 2011; **43**: 316–320.
4 Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S: Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 2010; **87**: 604–617.
5 Li B, Leal SM: Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 2009; **5**: e1000481.
6 Neale B, Rivas M, Voight B *et al*: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; **7**: e1001322.
7 Wu MC, Kraft P, Epstein MP *et al*: Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 2010; **86**: 929–942.
8 Nelson MR, Wegmann D, Ehm MG *et al*: An abundance of rare functional variants in 202 drug target genes sequenced in 14002 people. Submitted.
9 Cohen J, Pertsemlidis A, Fahmi S *et al*: Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA* 2006; **103**: 1810–1815.
10 Cohen J, Pertsemlidis A, Kotowski I, Graham R, Garcia C: Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* 2005; **37**: 161–165.
11 Risch N: Implications of multilocus inheritance for gene-disease association studies. *Theor Popul Biol* 2001; **60**: 215–220.
12 Fingerlin T, Boehnke M, Abecasis G: Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. *Am J Hum Genet* 2004; **74**: 432–443.
13 Li M, Boehnke M, Abecasis G: Efficient Study designs for tests of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet* 2006; **78**: 778–792.
14 Peng B, Li B, Han Y, Amos C: Power analysis for case-control association studies of samples with known family histories. *Hum Genet* 2010; **127**: 699–704.
15 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.
16 Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 1990; **46**: 222–228.
17 Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241–247.
18 Reimherr M, Nicolae D: You've gotta be lucky: coverage and the elusive gene-gene interaction. *Ann Hum Genet* 2011; **75**: 105–111.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)