OPEN ACCESS Freely available online

PLoS GENETICS

A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease

Iuliana Ionita-Laza[1]*, Joseph D. Buxbaum[2], Nan M. Laird[3☯], Christoph Lange[3,4,5☯]

1 Department of Biostatistics, Columbia University, New York, New York, United States of America, **2** Department of Psychiatry, Mount Sinai School of Medicine, New York, New York, United States of America, **3** Department of Biostatistics, Harvard University, Boston, Massachusetts, United States of America, **4** Institute for Genomic Mathematics, University of Bonn, Bonn, Germany, **5** German Center for Neurodegenerative Diseases, Bonn, Germany

Abstract

Rapid advances in sequencing technologies set the stage for the large-scale medical sequencing efforts to be performed in the near future, with the goal of assessing the importance of rare variants in complex diseases. The discovery of new disease susceptibility genes requires powerful statistical methods for rare variant analysis. The low frequency and the expected large number of such variants pose great difficulties for the analysis of these data. We propose here a robust and powerful testing strategy to study the role rare variants may play in affecting susceptibility to complex traits. The strategy is based on assessing whether rare variants in a genetic region collectively occur at significantly higher frequencies in cases compared with controls (or vice versa). A main feature of the proposed methodology is that, although it is an overall test assessing a possibly large number of rare variants simultaneously, the disease variants can be both protective and risk variants, with moderate decreases in statistical power when both types of variants are present. Using simulations, we show that this approach can be powerful under complex and general disease models, as well as in larger genetic regions where the proportion of disease susceptibility variants may be small. Comparisons with previously published tests on simulated data show that the proposed approach can have better power than the existing methods. An application to a recently published study on Type-1 Diabetes finds rare variants in gene *IFIH1* to be protective against Type-1 Diabetes.

Citation: Ionita-Laza I, Buxbaum JD, Laird NM, Lange C (2011) A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease. PLoS Genet 7(2): e1001289. doi:10.1371/journal.pgen.1001289

Editor: Suzanne M. Leal, Baylor College of Medicine, United States of America

Received May 26, 2010; **Accepted** December 31, 2010; **Published** February 3, 2011

Funding: This work was supported by NIH grants 1R03HG005908, R01MH087590, and R01MH081862. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ii2135@columbia.edu

☯ These authors contributed equally to this work.

Introduction

Common diseases such as diabetes, heart disease, schizophrenia, etc., are likely caused by a complex interplay among many genes and environmental factors. At any single disease locus allelic heterogeneity is expected, i.e., there may be multiple, different susceptibility mutations at the locus conferring risk in different individuals [1].

Common and rare variants could both be important contributors to disease risk. Thus far, in a first attempt to find disease susceptibility loci, most research has focused on the discovery of common susceptibility variants. This effort has been helped by the widespread availability of genome-wide arrays providing almost complete genomic coverage for common variants. The genome-wide association studies performed so far have led to the discovery of many common variants reproducibly associated with various complex traits, showing that common variants can indeed affect risk to common diseases [2,3]. However, the estimated effect sizes for these variants are small (most odds ratios are below 1.5), with only a small fraction of trait heritability explained by these variants [4]. For example, at least 40 loci have been identified for height, but these loci together explain only 5% of the 80% estimated heritability for this trait [5]. One possible explanation for this missing heritability is that, in addition to common variants, rare variants are also important.

Evidence to support a potential role for rare variants in complex traits comes from both empirical and theoretical studies. There is an increasing number of recent studies on obesity, autism, schizophrenia, epilepsy, hypertension, HDL cholesterol, some cancers, Type-1 diabetes etc. [6–15] that implicate rare variants (both single position variants and structural variants) in these traits. From a theoretical point of view, population genetics theory predicts that most disease loci do not have susceptibility alleles at intermediate frequencies [16,17].

With rapid advances in next-generation sequencing technologies it is becoming increasingly feasible to efficiently sequence large number of individuals genome-wide, allowing for the first time a systematic assessment of the role rare variants may play in influencing risk to complex diseases [18–21]. The analysis of the resulting rare genetic variation poses many statistical challenges. Due to the low frequencies of rare disease variants (as low as 0.001, and maybe lower) and the large number of rare variants in the genome, studies with realistic sample sizes will have low power to detect such loci one at a time, the way we have done in order to find common susceptibility variants [5,22]. It is then necessary to perform an overall test for all rare variants in a gene or, more generally a candidate region, under the expectation that cases with disease are different with respect to rare variants compared with control individuals. Several methods along these lines have already been proposed. One of the first statistical methods proposed for

Author Summary

Risk to common diseases, such as diabetes, heart disease, etc., is influenced by a complex interaction among genetic and environmental factors. Most of the disease-association studies conducted so far have focused on common variants, widely available on genotyping platforms. However, recent advances in sequencing technologies pave the way for large-scale medical sequencing studies with the goal of elucidating the role rare variants may play in affecting susceptibility to complex traits. The large number of rare variants and their low frequencies pose great challenges for the analysis of these data. We present here a novel testing strategy, based on a weighted-sum statistic, that is less sensitive than existing methods to the presence of both risk and protective variants in the genetic region under investigation. We show applications to simulated data and to a real dataset on Type-1 Diabetes.

the analysis of rare variants [23] is based on testing whether the proportion of carriers of rare variants is significantly different between cases and controls. A subsequent paper by Madsen and Browning [24] introduced the concept of weighting variants according to their estimated frequencies in controls, so that less frequent variants are given higher weight compared with more common variants. Price et al. [25] extended the weighted-sum approach in [24] to weight variants according to externally-defined weights, such as the probability of a variant to be functional. One potential drawback for these methods is that they are very sensitive to the presence of protective and risk variants.

We introduce here a new testing strategy, which we call replication-based strategy, and which is based on a weighted-sum statistic, but that has the advantage of being less sensitive to the presence of both risk and protective variants in a genetic region of interest. We illustrate the proposed approach on simulated data, and a real sequence dataset on Type-1 diabetes.

Methods

We assume for ease of exposition, and without loss of generality, that an equal number of cases and controls have been sequenced in a genetic region of interest. In what follows, for the sake of fixation, we will be concerned with the situation where rare variants in the region increase susceptibility to disease. We discuss first a one-sided testing strategy to test for the presence of variants conferring risk to disease.

We partition the variants observed in cases and controls into distinct groups, according to the observed frequencies of the minor allele in cases and controls. More precisely, group (k,k') contains all variants that have exactly k copies of the minor allele in controls, and exactly k' copies of the minor allele in cases. Let $n_k^{k'}$ be the size of group (k,k'). Note that the set of $n_k^{k'}$ represents a summary of the original data, that in some sense contains all the information the data can tell us about the presence of disease variants in the region under investigation. For the purpose of testing for the presence of risk variants, we choose to focus only on variants that are likely to be risk variants, i.e., those variants with $k'>k$. A summary of the data is shown in Table 1.

We define the following weighted-sum statistic, where each variant in group (k,k') is assigned a weight $w_k^{k'}$, and hence:

$$S=\sum_{k=0}^{N_r}\sum_{k'>k} n_k^{k'} w_k^{k'} \qquad (1)$$

Table 1. Data summary.

k/k′	1	2	3	4	5	...
0	n_0^1	n_0^2	n_0^3	n_0^4	n_0^5	...
1		n_1^2	n_1^3	n_1^4	n_1^5	...
2			n_2^3	n_2^4	n_2^5	...
3				n_3^4	n_3^5	...
4					n_4^5	...
...						

Variants are classified according to the number of times they appear in controls (k) and cases (k'). Only variants with higher observed count in cases compared with controls (i.e., more likely to be risk variants) are considered.
doi:10.1371/journal.pgen.1001289.t001

where N_r is an upper threshold on the number of occurrences of a variant in controls.

The choice of a good weighting scheme is very important for the performance of the approach. There are several possible ways to define the weights, including several already in the literature. Madsen and Browning [24] use data-dependent weights, with

$$w_k^{k'}=\frac{k'}{\sqrt{p_U(1-p_U)}},$$

where $p_U=\dfrac{k+1}{2(n_U+1)}$ is the estimated frequency based on controls only, and n_U is the number of controls. Price et al. [25] discuss the possibility of incorporating external weights, based on predictions about variants being functional.

For our approach we define a set of data-dependent weights, as follows. For a variant that occurs k times in controls and k' times in cases with $k'>k$, a natural weight is the negative log of the probability of a variant occurring at most k times in controls and at least k' times in cases, under the null hypothesis of the variant not being associated with the disease:

$$w_k^{k'}=-\log[p(k,k')] \text{ for } k'>k.$$

The statistic S above can then be written as:

$$S=\sum_{k=0}^{N_r}\sum_{k'>k} -n_k^{k'}\log[p(k,k')].$$

Since the number of mutations at a rare variant position follows approximately a Poisson distribution, the probability $p(k,k')$ of observing at a variant position at most k mutations in controls, and at least k' mutations in cases is calculated as

$$p(k,k')=\text{ppois}(k,\hat{f})\cdot(1-\text{ppois}(k'-1,\hat{f})),$$

where $\hat{f}=(k+k')/2$ is the estimated variant frequency based on the observed number of occurrences in both cases and controls, and ppois is the Poisson distribution function. Note that the higher the observed frequency in cases compared with controls (i.e., the higher $k'-k$), the higher the weight will be, and hence S tends to be larger when more variants are seen at higher frequencies in cases versus controls. We employ a standard permutation

procedure to evaluate the significance of S by randomly permuting the case/control label, and repeating the procedure described above for each permuted dataset, thus quantifying the extent to which the observed value of S is significantly higher compared to the null expectation.

The strategy described above is inherently one-sided, because we focus on variants that have higher observed frequency in cases compared with controls, i.e., more likely to be risk variants. This test can be used symmetrically to test for the presence of protective variants. Without any prior knowledge on the direction of the association, two one-sided statistics need to be computed. If S_+ and S_- are the two one-sided statistics as defined in eq. (1), then a max-statistic can be used that calculates the maximum of the two, i.e., $\max(S_+, S_-)$, and the statistical significance can be assessed by permutation.

Incorporation of External Biological Information

If external information is available on the plausibility of a rare variant to be related to disease, it is of interest to be able to incorporate such information into our testing strategy. Such information has proved essential in the mapping of the disease genes for two monogenic disorders [26], and may well prove important for mapping disease genes in more complex diseases. It is straightforward to extend the proposed approach to take into account such information. If we denote by $\varphi(v)$ the probability that a variant v is functional, then we can rewrite the statistic S above as:

$$S = \sum_{k=0}^{N_r} \sum_{k'>k} \sum_{v \in (k,k')} -\varphi(v) \log[p(k,k')],$$

where $v \in (k,k')$ signifies that variant v occurs k times in controls, and k' times in cases. In particular, if $\varphi(v)=1$ for all variants v then we recover the statistic S above, where functional information was not used. If on the other hand a variant is not functional, then $\varphi(v)=0$, and this variant is ignored.

Results

Simulated Data

Simulation model. We evaluated both the Type-1 error and the power for the proposed approach using data simulated under various genetic and disease models, and compared the results to those obtained using several existing approaches. Li and Leal [23] proposed one of the very first statistical methods for association testing with rare variants, based on collapsing rare variants in a genetic region together. In this approach, each individual is called a carrier if the individual contains at least one rare variant in the region. Then the strategy is to assess whether the proportion of carriers in affected individuals is significantly different from the proportion of carriers in unaffected individuals. A subsequent approach proposed by Madsen and Browning [24] is based on a weighted-sum statistic. A feature of this approach, especially relevant in large samples, is that variants are weighted according to their estimated frequencies from unaffected individuals, such that less frequent variants are assigned higher weights compared to more frequent variants.

The first set of simulations is based on a neutral Wright-Fisher model. Using the software package Genome [27] we generated 10,000 haplotypes according to a coalescent model, resulting in a total of 183 single nucleotide variants (SNVs) in the region (see Text S1 for more details). For the second set of simulations, we assume that the rare variants in the region are under weak purifying selection (as discussed in [16]), and use Wright's distribution [28] to sample the frequency at each variant:

$$f(p) = c p^{\beta_s - 1} (1-p)^{\beta_n - 1} e^{\sigma(1-p)},$$

where β_s and β_n are scaled mutation rates, and σ is the selection rate; c is a normalizing constant. As in [16] and [24], we take $\beta_s = 0.001$, $\beta_n = \beta_s/3$, and $\sigma = 12$. The main difference between the two simulation models is that the variant frequency spectra are different, with proportionally more rare variants under the second model compared with the first model (e.g., 141 out of the total of 183 variants have frequency below 0.01 under the second model, while only 83 have this low frequency under the first model).

With respect to the disease model, we assume varying number of disease susceptibility variants (DSVs) between 10 and 30, chosen at random from the generated polymorphisms that had low frequency (less than 0.01). We assume two possible values for the total population attributable risk (PAR): 0.03 and 0.05. The total PAR is distributed among all the disease variants. In one scenario, all disease variants have the same PAR, equal to the total PAR divided by the number of disease susceptibility variants. Perhaps a more realistic scenario is to assume unequal PAR for the different DSVs, and to this end we assume that individual variants' PARs are uniformly sampled from [0,1], and then renormalized to make them sum to the same total PAR of 0.03 or 0.05. In addition to the uniform distribution, we have also used an exponential model for the distribution of the individual PARs. In Supplemental Figure S-1 in Text S5 we show an example of the possible relationship between the odds ratio and the frequency at a disease variant, assuming 20 disease variants with frequencies between 0.0001 and 0.01, and a total PAR of 0.03 or 0.05.

Type-1 error. We evaluated the Type-1 error for both the proposed and the existing approaches using the two simulation models discussed above (neutral and weakly-purifying selection). We assume two possible sample sizes, 1000 and 2000, with equal number of cases and controls. As shown in Table 2, the Type-1

Table 2. Type-1 Error for the three approaches: collapsing (C), weighted-sum (WS_t), and replication-based (R).

Sim. Model	Sample Size	α	C	WS_t	R
1	1000	0.050	0.044	0.053	0.051
		0.025	0.022	0.027	0.025
		0.010	0.007	0.010	0.011
1	2000	0.050	0.055	0.043	0.052
		0.025	0.027	0.022	0.025
		0.010	0.012	0.010	0.008
2	1000	0.050	0.044	0.049	0.048
		0.025	0.023	0.027	0.022
		0.010	0.010	0.011	0.012
2	2000	0.050	0.047	0.051	0.050
		0.025	0.019	0.029	0.027
		0.010	0.008	0.011	0.012

Results for two genetic simulation models are shown: variants under a neutral evolution model (1), and variants under a weakly-purifying selection model (2). The sample size is the total number of individuals sequenced, with equal numbers of cases and controls. Nominal α levels: 0.05, 0.025, and 0.01.
doi:10.1371/journal.pgen.1001289.t002

error is well controlled at the nominal levels $\alpha = 0.05$, 0.025 and 0.01 for all three methods.

Power. We evaluated the power for all three methods assuming two models for generating the genetic data, and several complex disease models. For the genetic data, as explained above, two scenarios are illustrated: a first one where the variant frequencies are generated using a neutral coalescent model, and a second one where the variants are under weakly-purifying selection. For the underlying disease models, a varied number of disease susceptibility variants are assumed, that contribute equally, or unequally to the total PAR. For the latter scenario, the individual variant PAR are sampled from a uniform distribution (results for an exponential sampling distribution are shown in Supplemental Table S-1 in Text S3). To make the comparison fair among the different methods considered the same threshold of 0.01 was used on the frequency of the variants included in the three testing methods.

Power estimates for a series of simulation experiments are shown in Table 3. Note that the results are based on two-sided testing for all three methods and $\alpha = 0.05$. For the same total PAR, the power decreases with increasing number of disease variants, due to the correspondingly smaller contribution of each disease variant. Also, the power increases for all methods when the weakly-purifying selection simulation model is used as opposed to the neutral model, due to the lower number of rare variants that are actually *observed* under the former model compared with the latter model. However, given the same sampling distribution for the frequency of the variants, the power did not vary much between the different ways the individual PARs were sampled. The key factor is the total PAR for the region. Overall the proposed approach is consistently and substantially more powerful than both the collapsing and the weighted-sum approaches across the multiple scenarios that we have considered, and under both models to generate the variant frequencies.

Sensitivity to presence of both risk and protective variants. So far we have assumed scenarios where only variants that increase risk are present in a genetic region. However, sometimes it may be the case that both risk and protective variants are present in a genetic region of interest, for example when multiple genes in a set or pathway are tested together. This can also be true when individuals from the two extremes of a phenotype distribution are chosen to be studied. In such situations, the two existing methods discussed can suffer substantial loss of power, depending on the relative contributions of the two classes of variants. We show here that the proposed approach is less sensitive to such mixture, the principal reason being the inclusion of only those variants that may confer risk, and exclusion of the variants that are unlikely to be risk variants when we test for the presence of risk variants, and similarly when we test for the presence of protective variants.

Table 3. Power Estimates ($\alpha = 0.05$) for the three approaches: collapsing (C) versus weighted-sum (WS_t) versus replication-based (R).

				PAR = 0.03			PAR = 0.05		
Sim. Model	**Sample Size**	**Disease Model**	**#DSVs**	**C**	**WS_t**	**R**	**C**	**WS_t**	**R**
1	1000	Eq PAR	10	0.223	0.248	0.344	0.522	0.673	0.753
			20	0.210	0.242	0.331	0.474	0.514	0.690
			30	0.174	0.201	0.301	0.438	0.480	0.668
1	2000	Eq PAR	10	0.427	0.610	0.743	0.828	0.966	0.985
			20	0.384	0.521	0.701	0.783	0.917	0.975
			30	0.334	0.447	0.661	0.706	0.876	0.966
1	1000	Uneq PAR	10	0.214	0.276	0.367	0.514	0.670	0.768
			20	0.203	0.224	0.324	0.487	0.548	0.707
			30	0.163	0.210	0.298	0.442	0.469	0.652
1	2000	Uneq PAR	10	0.414	0.646	0.753	0.834	0.957	0.982
			20	0.394	0.551	0.712	0.769	0.920	0.972
			30	0.344	0.485	0.658	0.744	0.880	0.959
2	1000	Eq PAR	10	0.278	0.417	0.537	0.569	0.828	0.901
			20	0.262	0.310	0.481	0.538	0.724	0.870
			30	0.230	0.259	0.431	0.513	0.650	0.845
2	2000	Eq PAR	10	0.478	0.843	0.921	0.868	0.999	1.000
			20	0.455	0.762	0.911	0.859	0.996	1.000
			30	0.387	0.671	0.885	0.817	0.984	0.999
2	1000	Uneq PAR	10	0.265	0.385	0.494	0.586	0.854	0.893
			20	0.240	0.349	0.478	0.569	0.757	0.890
			30	0.235	0.276	0.455	0.494	0.680	0.846
2	2000	Uneq PAR	10	0.480	0.859	0.924	0.882	0.996	1.000
			20	0.431	0.761	0.899	0.860	0.993	0.999
			30	0.414	0.708	0.880	0.814	0.988	0.999

Two genetic simulation models are assumed: neutral variants (1), and mildly deleterious variants (2). Varying number of DSVs are assumed, that can contribute equally or unequally to the total PAR. The sample size is the total number of individuals sequenced, with equal numbers of cases and controls. All tests are two-sided, i.e., testing for the presence of risk or protective variants in the region of interest.

doi:10.1371/journal.pgen.1001289.t003

Table 4. Power Estimates ($\alpha = 0.05$) for two-sided tests, testing for the presence of risk or protective variants, when there is a mixture of risk and protective variants in the region of interest.

			PAR = 0.03			PAR = 0.05		
Sim. Model	**#Risk**	**#Protective**	**C**	**WS_t**	**R**	**C**	**WS_t**	**R**
1	20	0	0.210	0.242	0.331	0.474	0.514	0.690
		5	0.132	0.164	0.263	0.281	0.389	0.567
		10	0.081	0.126	0.202	0.145	0.300	0.476
		20	0.056	0.102	0.183	0.044	0.209	0.478
2	20	0	0.262	0.310	0.481	0.538	0.724	0.870
		5	0.155	0.212	0.383	0.336	0.559	0.788
		10	0.095	0.185	0.314	0.160	0.438	0.697
		20	0.054	0.133	0.286	0.056	0.320	0.711

In addition to 20 risk variants in the region, there are between $0-20$ protective variants as well. Simulation model corresponds to one of the two scenarios: neutral variants (1), and mildly deleterious variants (2). The total sample size is 1000 cases and controls. Collapsing (C) vs. weighted-Sum (WS_t) vs. replication-based (R).
doi:10.1371/journal.pgen.1001289.t004

In Table 4 we show power estimates when we test for the presence of risk or protective variants, given the existence of both risk and protective variants in the region. We assume that there are 20 risk variants in the region, and the number of protective variants is between 0 and 20. As in the previous simulations, the total PAR for the 20 risk variants can take two values, 0.03 and 0.05, while each protective variant has the same per-variant PAR, equal to the total PAR divided by 20. Therefore, when the number of protective variants is 20 the overall contribution to disease is the same for risk and protective variants. This is of course the worst case scenario, and the Collapsing and Weighted-Sum approaches suffer from substantial loss of power in such cases. Even the proposed approach is not insensitive to such scenarios; however the loss in power is considerably less than that for the other two methods.

Type-1 Diabetes Dataset

We also applied our approach to a dataset on Type 1 Diabetes (T1D), published by Nejentsev et al. [15]. In their paper, the authors resequenced exons and splice sites of ten candidate genes in 480 cases and 480 controls (more details on the dataset are in Text S2). In their study, rare variants were tested individually, and two SNVs in gene *IFIH1* and two other SNVs in gene *CLEC16A* were found to be protective against T1D.

Here we reanalyze the dataset using the proposed approach, and two of the existing approaches. For each gene and each method, we perform two-sided tests, testing for the presence of risk or protective variants. Results are in Table 5. As in [15] we found one gene, *IFIH1*, to be significant with all three methods (P-value < 0.001 for all three methods). For this gene, controls were enriched for rare mutations compared with cases. Some evidence of enrichment in protective variants was also observed in another gene, *CLEC16A*, although the P-values do not remain significant after multiple testing correction.

Table 5. Type-1 diabetes results.

Gene	**#SNVs**	**C**	**WS_t**	**R**
IFIH1	29	0.0005	0.0002	0.0001
CLEC16A	45	0.030	0.022	0.014

Two-sided P-values for the top two genes. An upper frequency threshold of 0.01 was used for the variants considered for testing.
doi:10.1371/journal.pgen.1001289.t005

Discussion

We have proposed here a new testing strategy to examine associations between rare variants and complex traits. The approach is based on a weighted-sum statistic that makes efficient use of the information the data provides on the presence of disease variants in the region being investigated. The proposed test is based on computing two one-sided statistics, designed to quantify enrichment in risk variants, and protective variants, respectively. This aspect allows the proposed approach to have substantially better power than existing approaches in the presence of both risk and protective variants in a region. Even when only one kind of variants is present, we have shown via simulations that the proposed approach has consistently better power than existing approaches. An application to a previously published dataset on Type-1 Diabetes [15] confirmed the original finding, namely that rare variants in *IFIH1* confer protection towards disease.

The weights underlying our weighted-sum statistic depend only on the data at hand. However, external information on the likelihood of a variant to be functional could prove very useful, and could be combined with the information present in the data to improve power to identify disease susceptibility variants. Such information has been successfully used to identify the genes for several monogenic disorders [26]. Price et al. [25] discuss a weighted-sum approach with externally-derived weights, and show that such information can be very useful using several empirical datasets. We have also described a natural way to take into account such external functional predictions within the proposed framework.

Since empirical data are only now becoming available, it is not known how often both risk and protective variants are present in a particular disease gene. When both types of variants are present, it seems appealing to be able to combine the two types of signals. It is possible to extend the proposed approach to take advantage of both kinds of disease variants, and we discuss such an extension in Text S4. We noticed in our simulation experiments that such a hybrid approach can have much improved power when both types of variants are present, but this comes at the price of reduced power when only one type of variants is present. Therefore, depending on the underlying disease model, both approaches could provide useful information.

The proposed approach is applicable to a case-control design and therefore is susceptible to spurious findings due to population stratification. Population stratification has been shown to be an important issue in the context of common variants. For rare variants, differences in rare variant frequencies between populations are likely to be even more pronounced. Development of new methods, and extension of existing methods are necessary to adequately address the issue. Alternatively, family-based designs offer the advantage of being robust to false positive findings due to population stratification.

Replication of association signals in independent datasets is an essential aspect of any disease association study, and has become standard practice for common variants. Rare variants, due to their low frequencies and potential modest effects, are normally tested together with other rare variants in the same unit, e.g., gene. Therefore a reasonable first replication strategy is at the level of the gene. Follow-up of individual variants in the gene can be performed to investigate whether any of the rare variants in the gene can be found to be significantly associated with disease.

Finding rare disease susceptibility variants is a challenging problem, especially due to their low frequencies and the probable moderate effects on disease. So far the methods proposed in the literature have focused on case-control designs. However, for rare variants, family-based designs may prove very useful. Not only are they robust against population stratification, but they may also offer increased power due to the increased likelihood of affected relatives to share the same rare disease variants. Continued development of novel statistical methods for identifying rare disease susceptibility variants is needed for population-based designs, and especially for family-based designs.

Software implementing these methods is available at: http://www.mailman.columbia.edu/our-faculty/profile?uni=ii2135.

Supporting Information

Text S1 Simulation model 1.
Found at: doi:10.1371/journal.pgen.1001289.s001 (0.04 MB PDF)

Text S2 Type-1 diabetes dataset.
Found at: doi:10.1371/journal.pgen.1001289.s002 (0.04 MB PDF)

Text S3 Power estimates when individual variants' PAR are sampled from an exponential distribution.
Found at: doi:10.1371/journal.pgen.1001289.s003 (0.05 MB PDF)

Text S4 Capitalizing on the presence of both risk and protective variants.
Found at: doi:10.1371/journal.pgen.1001289.s004 (0.11 MB PDF)

Text S5 Relationship between odds ratios and frequencies for the simulated scenarios.
Found at: doi:10.1371/journal.pgen.1001289.s005 (0.56 MB PDF)

Author Contributions

Conceived and designed the experiments: IIL NML CL. Performed the experiments: IIL. Analyzed the data: IIL. Contributed reagents/materials/analysis tools: IIL JDB. Wrote the paper: IIL NML CL.

References

1. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11: 241–247.
2. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356–369.
3. Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A catalog of published genome-wide association studies. Available at http://www.genome.gov/26525384.
4. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the Missing Heritability of Complex Diseases. Nature 461: 747–753.
6. Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, et al. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci USA 101: 15992–15997.
7. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869–872.
8. Fearnhead NS, Winney B, Bodmer WF (2005) Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. Cell Cycle 4: 521–525.
9. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, et al. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci USA 103: 1810–1815.
10. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40: 592–599.
11. Stefansson H, Rujescu D, Cichon S, PietilŠinen OP, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455: 232–236.
12. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539–543.
13. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358: 667–675.
14. Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, et al. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41: 160–162.
15. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387–389.
16. Pritchard JK (2001) Are rare variants responsible for susceptibility to common diseases? Am J Hum Genet 69: 124–137.
17. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common diseaseÐcommon variant… or not? Hum Mol Genet 11: 2417–2423.
18. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141.
19. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135–1145.
20. Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet 85: 142–154.
21. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46.
22. Morris AP, Zeggini E (2009) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.
23. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311–321.
24. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384. doi:10.1371/journal.pgen.1000384.
25. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
26. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2009) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30–35.
27. Liang L, Zoellner S, Abecasis GR (2007) GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics 23: 1565–1567.
28. Wright S (1949) Adaptation and selection. In: Genetics, Paleontology, and Evolution. Princeton Univ Press. pp 365–389.