UC San Diego
UC San Diego Previously Published Works

Title

An Excess of Risk-Increasing Low-Frequency Variants Can Be a Signal of Polygenic Inheritance in Complex Diseases

Permalink

Journal

ISSN

Authors

Chan, Yingleong
Lim, Elaine T
Sandholm, Niina
et al.

Publication Date

DOI

Peer reviewed

An Excess of Risk-Increasing Low-Frequency Variants Can Be a Signal of Polygenic Inheritance in Complex Diseases

Yingleong Chan,[1,2,3] Elaine T. Lim,[1,2,4] Niina Sandholm,[5,6,7] Sophie R. Wang,[1,2,3] Amy Jayne McKnight,[8] Stephan Ripke,[2,4] DIAGRAM Consortium, GENIE Consortium, GIANT Consortium, IIBDGC Consortium, PGC Consortium, Mark J. Daly,[1,2,4] Benjamin M. Neale,[2,4] Rany M. Salem,[1,2,3] and Joel N. Hirschhorn[1,2,3,]*

In most complex diseases, much of the heritability remains unaccounted for by common variants. It has been postulated that lower-frequency variants contribute to the remaining heritability. Here, we describe a method to test for polygenic inheritance from lower-frequency variants by using GWAS summary association statistics. We explored scenarios with many causal low-frequency variants and showed that there is more power to detect risk variants than to detect protective variants, resulting in an increase in the ratio of detected risk to protective variants (R/P ratio). Such an excess can also occur if risk variants are present and kept at lower frequencies because of negative selection. The R/P ratio can be falsely elevated because of reasons unrelated to polygenic inheritance, such as uneven sample sizes or asymmetric population stratification, so precautions to correct for these confounders are essential. We tested our method on published GWAS results and observed a strong signal in some diseases (schizophrenia and type 2 diabetes) but not others. We also explored the shared genetic component in overlapping phenotypes related to inflammatory bowel disease (Crohn disease [CD] and ulcerative colitis [UC]) and diabetic nephropathy (macroalbuminuria and end-stage renal disease [ESRD]). Although the signal was still present when both CD and UC were jointly analyzed, the signal was lost when macroalbuminuria and ESRD were jointly analyzed, suggesting that these phenotypes should best be studied separately. Thus, our method may also help guide the design of future genetic studies of various traits and diseases.

Introduction

Most common diseases involve a mix of both genetic and environmental factors and do not follow simple patterns of Mendelian inheritance. In such diseases, the genetic component is usually polygenic: genetic variation in many genes individually contribute a small or a moderate component of disease risk.[1] Genome-wide association studies (GWASs) have identified numerous genomic loci in which common variants (≥5% frequency) are associated with complex diseases.[2] Even in some of the largest and most successful GWASs to date, much of the genetic contribution to phenotype remains unexplained (sometimes called "missing heritability"),[3,4] suggesting that lower-frequency variants, not well surveyed by GWASs, may also contribute to the missing heritability. Indeed, in some diseases such as autism spectrum disorders (ASD [MIM 209850]), inherited rare (<1% frequency) and low-frequency (<5% frequency) variants have been recently shown to play an important role in the genetic architecture of the disorder,[5,6] suggesting that more loci with low-frequency variants could be identified if appropriate additional studies were performed. In other diseases, there is as yet little evidence of a substantial role for low-frequency

variation, leaving open the question of whether studies of low-frequency variation will be fruitful for those diseases.

The relative success of different approaches in identifying more contributing loci will depend on what type of variation accounts for the missing heritability. Low-frequency variants might remain undetected because they might not be well represented or well tagged by markers on genotyping arrays and therefore would not be well imputed.[7] Along these lines, the statistical power to detect low-frequency variants in GWASs is much lower than that of common variants if their underlying effect sizes are similar.[8] Knowing whether low-frequency variants contribute to the missing heritability of a disease is important because approaches better suited to identify additional common variants differ from those aimed at identifying rarer variants (genotyping arrays with common variants compared to arrays with lower-frequency variants or sequencing).

Methods for detecting a contribution from common variants to the missing heritability have been described previously. In a GWAS of schizophrenia (SCZ [MIM 181500]),[9] Purcell and colleagues developed the concept of a polygenic score by combining the effects of multiple common variants that are modestly associated with

schizophrenia. They showed that the score is predictive of schizophrenia in an independent cohort, thus indicating that there is a polygenic signal from many yet-to-be-detected common variants in schizophrenia. Yang and colleagues adopted a different approach by assessing the narrow-sense heritability of human height with a linear-model analysis by using hundreds of thousands of common variants.[10] They found that at least 45% of the variance of height can be accounted for by common variants, indicating that there are many common variants associated with height that have yet to be discovered. Although both methods can be used to detect a signal of polygenic inheritance from common variants in complex diseases, these tests were not designed to specifically test for low-frequency variants and also require individual-level genotype data.

In this manuscript, we describe an approach that can be applied directly to GWAS summary statistics to ascertain the presence of polygenic inheritance from low-frequency variants. We observed that, if low-frequency variants contribute to disease susceptibility, there can be an excess of associated risk variants compared to protective variants at a given significance level. Here, risk variants are defined as variants for which the minor allele is associated with increased risk of disease and protective variants are defined as those for which the minor allele is associated with decreased risk of disease. Under the null model, there should be no excess of associated risk variants compared to protective variants. We calculated the risk to protective ratio (R/P ratio) (the ratio of the number of detected risk variants over the number of detected protective variants) to test for such an excess of risk variants. We explored various scenarios that could give rise to an increase in the R/P ratio. First, we showed empirically and analytically that when low-allele-frequency variants contribute to polygenic inheritance of a disease with low prevalence, there is an elevated R/P ratio because of greater power to detect risk variants than protective variants. Next, we showed through simulations that under a scenario of polygenic inheritance that includes negative selection, risk variants can have lower average frequencies than protective variants, leading to an elevated R/P ratio within the lower-frequency range. However, we also showed that such an elevated R/P ratio can occur because of reasons unrelated to polygenic inheritance. First, we showed that an uneven sample size (a substantially larger control group than case group) can produce an apparent increase in the R/P ratio and therefore, where the sample size is not balanced between the case and control groups, one should compare the observed R/P ratio against that obtained through simulations with the same sized groups of cases and controls. Next, we showed that particular scenarios of asymmetric population stratification can produce a similar excess of low-frequency risk variants and recommend that precautions for detecting and correcting for such stratification should be performed before one can

confidently interpret an excess of risk variants as being a signal of polygenic inheritance.

We then applied our method to results from published GWASs for several diseases, including schizophrenia,[11] bipolar disorder (BIP [MIM 125480]),[12] major depressive disorder (MDD [MIM 608516]),[13] type 2 diabetes (T2D [MIM 125853]),[14] and various classes of obesity (OB [MIM 601665]).[15] We observed strong signals of increased risk variants in several of the diseases but little or no signal in others, suggesting that efforts to discover low-frequency and rare variants will be more fruitful for the diseases with such a signal. We further used our method to test whether apparently related phenotypes share low-frequency or rare genetic contributors and hence should be analyzed together or separately. By applying the method to phenotypes related to diabetic nephropathy (DN [MIM 603933])[16] and inflammatory bowel disease (IBD [MIM 266600]),[17] we found that the polygenic signal was eliminated when individuals with macroalbuminuria and individuals with end-stage renal disease were analyzed together, whereas we still observed a significant signal when individuals with Crohn disease and ulcerative colitis were analyzed together. Thus, our method has the potential to guide the strategy in searching for additional genetic loci as well as in prioritizing the choice of phenotype for future studies of rare genetic variation in polygenic traits and diseases.

Material and Methods

Testing for an Excess of Risk Variants from GWAS Summary Statistics

Calculating the R/P Ratio Statistic from Observed GWAS Summary Statistics

The four input fields we used for R/P ratio calculations for each SNP are as follows: an identifier (rsID), the minor allele frequency, the association p value, and a field to determine the direction of effect, i.e., either an odds-ratio (OR) or an effect size (β). The ORs or βs were adjusted to reflect the effect of the minor allele by inverting the ORs or changing the sign of the βs if they were reported for the major allele. Each variant was assigned as risk if the OR > 1 or β > 0 and protective if the OR < 1 or β < 0. Neutral variants, i.e. OR = 1 or β = 0 were discarded from the analysis. We removed SNPs not present in the HapMap CEU population (phase 2, release 28),[18,19] not in the 1000 Genomes EUR population,[20] or with minor allele frequency less than 1%. We sorted the remaining variants in order from most significant to least and performed LD-pruning by systematically going through the variants and removing variants that have an $r^2 > 0.1$ with any of the more significantly associated variants. We used PLINK[21] to calculate r^2 correlations of variant pairs within a 1 megabase window from 379 EUR individuals of the 1000 Genomes. To measure the excess of risk variants in the lower-frequency range, we separated the low-frequency variants into three distinct bins, i.e., 1%–5%, 5%–10%, and 10%–15%. We also included the 30%–50% bin as a negative control, where we should not observe any excess of risk variants. For each bin, we counted the number of detected risk variants and the number of detected protective variants that meet

significance cutoffs of p < 0.001 and p < 0.01. We calculated the R/P ratio as

$$R/P\ ratio = \frac{No.\ of\ detected\ risk\ variants}{No.\ of\ detected\ protective\ variants}.$$

Assessing the Significance of the Observed R/P Ratio
To assess the significance of an elevation in R/P ratio, we simulated individuals with HAPGEN[22] by using parameters from the HapMap CEU population (phase 3, release 2) to obtain the null distribution of the $\log_2$ R/P ratio statistic. We first simulated 100,000 individuals to form a pool of individuals that we could subsequently sample from. Next, we randomly sampled the same number of individuals in the case and control groups as were used in the actual GWAS, performed the association test with PLINK, with LD-pruning and R/P ratio calculations identical to the procedure described above. We repeated this process 1,000 times to obtain accurate estimates of the sample mean (μ) and standard deviation (σ) of the $\log_2$ R/P ratio under the null for each of our frequency bins and p value cutoffs. We calculated the significance of the observed $\log_2$ R/P ratio by performing a one-tailed Z test to obtain the Z score and p value (p), i.e.,

$$Z\ score = \frac{observed\ \log_2 R/P\ ratio - \mu}{\sigma}$$

$$p = \int_{Zscore}^{\infty} N(x, 0, 1)dx.$$

We defined p < 0.01 as our significance threshold for calling a significant excess of risk variants. We used the $\log_2$ R/P ratio as our test statistic because the $\log_2$ R/P ratio is normally distributed for all the frequency bins and p value cutoffs used (Figure S1 available online).

Calculating Noncentrality Parameter for Comparing Power between Risk and Protective Variants

Power Calculation
The power of a variant is expressed by calculating the expected noncentrality parameter (NCP) of the χ^2 distribution for the alternative distribution. The greater the NCP, the more power there is to detect the effective variant. The algorithm for calculating NCP is identical to the genetic power calculator[8] for case-control threshold-selected quantitative traits, assuming an additive model of the QTL effect, i.e., the dominance to additive QTL effect parameter is set to 0. The variance explained for a SNP with allele frequency as p and effect size as β is $\beta^2 2p(1 - p)$. For risk variants, we calculated the NCP (NCP_{risk}) for multiple values of effect sizes (β), ranging from 0 to 0.5 with intervals of 0.01. Similarly, for protective variants, we calculated the NCP ($NCP_{protective}$) for multiple values of β, ranging from 0 to −0.5 with intervals of 0.01. The relative difference in power between risk and protective variants is measured by the NCP ratio. The NCP ratio is calculated as

$$NCP\ ratio = \frac{NCP_{risk}}{NCP_{protective}}.$$

Base Model
We define the base model as a set of parameters used for calculating NCP: 10,000 case subjects, 10,000 control subjects, and effective and marker variant frequency set to 1%. The prevalence is set as 1%, i.e., the trait threshold's lower and upper limit is 2.33 and 9, respectively, for case subjects and −9 and 2.33 for control subjects. We have used 9 and −9 as surrogates for infinity (+∞ and -∞, respectively), but any sufficiently large number will not change the conclusions of the downstream analyses. Complete linkage disequilibrium (LD) between the causal variant and marker variant is assumed, i.e., $D' = 1$.

Simulating R/P Ratios for Negative Selection

Obtaining Frequencies and Effect Sizes
If the variants that have an effect on the phenotype are under negative selection, it can lead to scenarios where there are more risk variants than protective variants to begin with, especially for low-frequency variants. To illustrate this, we simulated neutral variants and causal variants under negative selection by using previously published models and parameters that result in an allele spectrum similar to that observed in the European population.[23,24] We used the forward simulation package ForSim[25] to simulate coding sequence variation in the European population in 1,000 genes. The average gene coding length was set as 1,500 bp. We used a mutation rate per site of 2×10^{-8} and a uniform locus-wide recombination rate of 2 Mb/cM. We modeled the distribution of selection coefficients (s) for de novo missense mutations by a gamma distribution.[26] We used the conventional 4-parameter model of the history of the European population with long-term constant size (N = 8,100 for 45,000 generations) followed by a bottleneck (N = 2,000) and then by exponential growth (1.5% increase per generation for 370 generations) to achieve a final population size of approximately 500,000 individuals.[23,24] We obtained 823 nonneutral variants that have minor allele frequencies ≥1% and assigned them as effective variants and assuming that the allele under negative selection confers risk, i.e., positive effect (Figure S2). By considering only additive genetic effects, we assigned effect sizes as $\beta = s^{\tau}(1 + \varepsilon)$ as suggested in Eyre-Walker.[27] Here, β is the variant's additive effect on the quantitative trait, s is the absolute value of the variant's selection coefficient, ε is a normally distributed random noise parameter that was set to having mean 0 and standard deviation 0.05, and τ is the degree of coupling between β and s and was set at 0.5 for our analyses. The effect sizes are scaled so that these 823 variants explain 60% of the phenotypic variance.

Obtaining Phenotypes and Calculating R/P Ratio for the Selection Model
We use the 100,000 HAPGEN-simulated individuals and selected 823 matched SNPs such that the frequency matches the variants generated by ForSim. We then assigned these matched SNPs with effect sizes determined earlier. We calculated the phenotypic Z score for each of our 100,000 individuals in the same way that we did previously,[28] i.e., by calculating the weighted allele score (*WAS*) and adding it to a randomly generated variable sampled from a normal distribution of mean 0 and variance 0.4 such that the total variance explained is 1. We then sampled 2,000 individuals with phenotypic Z scores > 1.645 (5% prevalence) as case subjects and another 2,000 individuals with phenotypic Z scores ≤ 1.645 as control subjects. We used PLINK to perform the association test on all the variants and calculated the R/P ratio within the same frequency bins as well as p value cutoffs as described above. This process was repeated 1,000 times to obtain the distribution of the R/P ratio. For the control model, we randomly sampled 2,000 individuals as case subjects and 2,000 individuals as control subjects and calculated the R/P ratio as described above.

Simulating R/P Ratios for Population Stratification

We use HAPGEN to simulate 4,000 distinct individuals from the HapMap CEU population (phase 3, release 2) as well as another 4,000 distinct individuals from the HapMap TSI population (phase 3, release 2). For complete stratification, we randomly sampled 1,000 individuals from the CEU pool as control subjects and 1,000 individuals from the TSI pool as case subjects. We simulated asymmetric mixtures of 1%, 5%, and 10% by randomly sampling 1,000 individuals from the CEU pool as control subjects and sampling 10, 50, and 100 individuals from the TSI pool as case subjects, respectively, and made up the remainder of the case group from the CEU pool. We used PLINK to perform the association test on all the variants and calculated the R/P ratio within the same frequency bins as well as p value cutoffs as described above. Each process was repeated 1,000 times to obtain the distribution of the R/P ratio. All PCA analysis was performed with smartpca from the EIGENSOFT 3.0 package.[29] All meta-analysis of GWAS summary statistics were performed with METAL.[30] Inflation of the GWAS test statistic because of population stratification was assessed by genomic control inflation factor (λ_{GC}).[31]

Calculating R/P Ratio from Published GWAS Summary Statistics

Schizophrenia, Major Depressive Disorder, and Bipolar Disorder

GWAS summary statistics were provided from published results of schizophrenia,[11] bipolar disorder,[12] and major depressive disorder.[13] SNPs that failed imputation (INFO < 0.6) were discarded. The sizes of the case and control groups used for simulating the null distribution are as follows: schizophrenia (SCZ), 9,394 case subjects and 12,462 control subjects; major depressive disorder (MDD), 9,240 case subjects and 9,519 control subjects; and bipolar disorder (BIP), 7,481 case subjects and 9,250 control subjects.

Type 2 Diabetes

GWAS summary statistics were provided from published results of type 2 diabetes.[14] SNPs that passed imputation for fewer than 15,000 individuals ($N_{cases} < 15,000$) were discarded. A total of 15,000 case subjects and 50,337 control subjects were used for simulating the null distribution.

Obesity

GWAS summary statistics were provided from published results of various classes of obesity.[15] SNPs that passed imputation for fewer than 50,000 individuals ($N_{cases} < 50,000$), 10,000 individuals ($N_{cases} < 10,000$), 2,000 individuals ($N_{cases} < 2,000$), and 1,000 individuals ($N_{cases} < 1,000$) were discarded for the overweight (BMI > 25), class 1 (BMI > 30), class 2 (BMI > 35), and class 3 (BMI > 40) data sets, respectively. The sizes of the case and control groups used for simulating the null distribution are as follows: overweight, 50,000 case subjects and 35,715 control subjects; class 1, 10,000 case subjects and 20,325 control subjects; class 2, 2,000 case subjects and 12,466 control subjects; and class 3, 1,000 case subjects and 18,346 control subjects.

Inflammatory Bowel Disease

GWAS summary statistics were provided from published results of Crohn disease (CD),[32] ulcerative colitis (UC),[33] and the combined case cohort of both Crohn disease and ulcerative colitis (CD+UC).[17] SNPs that failed imputation (INFO < 0.6) were discarded. The sizes of the case and control groups used for simulating the null distribution are as follows: CD, 5,956 case subjects and 14,927 control subjects; UC, 6,968 case subjects and 20,464 control subjects; and CD+UC, 12,882 case subjects and 21,770 control subjects.

Diabetic Nephropathy

GWAS summary statistics were provided from published results of phenotypes related to diabetic nephropathy[16] that are macroalbuminuria (MACRO) and end-stage renal disease (ESRD). SNPs that failed imputation in at least 1 cohort were discarded. The sizes of the case and control groups used for simulating the null distribution are as follows: macroalbuminuria versus control ($MACRO_{ctrl}$), 1,478 case subjects and 3,315 control subjects; end-stage renal disease versus control ($ESRD_{ctrl}$), 1,399 case subjects and 3,315 control subjects; ESRD versus controls that include MACRO ($ESRD_{ctrl+macro}$), 1,399 case subjects and 5,253 control subjects; and combined MACRO and ESRD versus control ($[MACRO + ESRD]_{ctrl}$), 2,916 case subjects and 3,315 control subjects.

Results

We developed a method to detect and assess the significance of an excess of risk variants, measured by the ratio of risk variants to protective variants (R/P ratio) within a series of frequency bins and p value cutoffs (see Material and Methods). We proceeded to show that under an assumption of polygenic inheritance from low-frequency variants, there is more statistical power to detect risk variants than to detect protective variants, which can result in an increased R/P ratio. We also showed that such an excess can also occur if risk variants are kept at lower frequencies because of negative selection. However, such an excess can also occur because of reasons unrelated to a contribution of rare variants to disease risk: uneven sample sizes or asymmetric population stratification. Therefore, steps have to be taken to account for these latter possibilities before one can confidently interpret the excess of risk variants as a true signal of polygenic inheritance. Finally, we applied the method to GWAS summary statistics from several published studies.

Significantly Higher Power to Detect Low-Frequency Risk Variants of Moderate to Large Effect

The liability threshold model for disease[34] has been shown to be consistent with results from GWASs for multiple diseases.[35] This model assumes that there is an underlying unmeasured trait related to disease risk and that individuals are affected with disease only when the value of the trait exceeds a particular threshold. Under such a model, we discovered that the statistical power to detect risk variants is higher than the power to detect protective variants, even when they have the same effect size with respect to the underlying unmeasured trait. For example, we calculated power by using a predefined set of parameters defined as the "base model" (see Material and Methods). From our calculations, we observed that as effect size increases, there is significantly more power to detect risk than protective variants as indicated by the increase in the NCP ratio (Figure 1). This result shows that for this scenario, where the number of risk and protective variants are equal and have similar absolute effect sizes, the difference in power can create an excess of
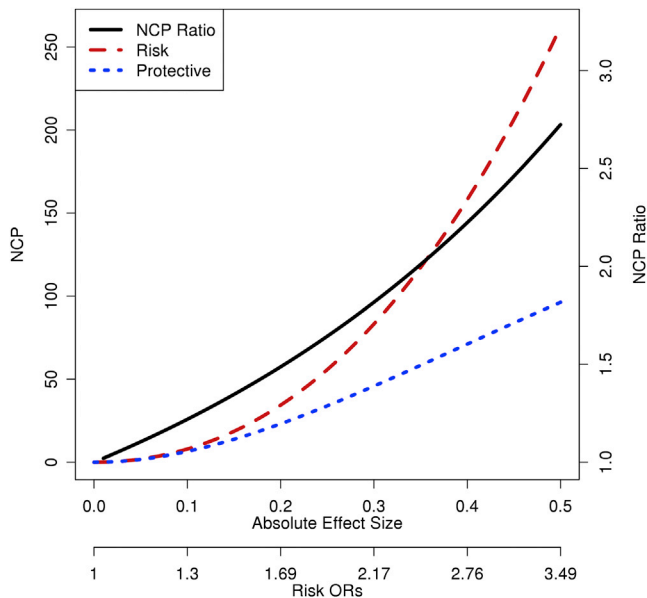
Figure 1. Comparing the Power to Detect Risk and Protective Variants with the Same Underlying Effect Size
The plot shows the power as the noncentrality parameter (NCP) for detecting minor alleles that confer risk (risk variants) and minor alleles that confer protection (protective variants) with varying absolute effect sizes ($0 < \beta < 0.5$ in standard deviation units) via parameters from the base model (see Material and Methods). It also shows the NCP ratio, which is the NCP of risk variants divided by the NCP of protective variants with the same absolute effect size (right vertical axis). The equivalent odds ratio (OR) for the risk variants is also shown on the horizontal axis.

detected risk variants over protective variants, which can result in an increased R/P ratio.

The Difference in Power Is Larger under Certain Scenarios

We explored how the difference in power to detect risk and protective variants would be affected when we varied the parameters in the model under which we calculated power. First, we calculated power via the base model but varied the minor allele frequency from 1% to 15%. The difference in power for risk and protective variants decreases as the variant frequency increases (Figure 2A). Second, we varied the disease prevalence from 1% (trait Z score > 2.33) to 15% (trait Z score > 1.03). Here, the difference in power decreases with increasing disease prevalence (Figure 2B), and there is no difference in power at any effect size when the disease prevalence is exactly 50%. Third, we varied the linkage disequilibrium (LD) between the associated variant and the causal variant from moderate LD ($D' = 0.5$) to strong LD ($D' = 0.8$). Although there is a general loss of power with decreasing LD, the difference in power between risk and protective variants increases with decreasing LD (Figure 2C). Along similar lines, when we assumed that low-frequency causal variants are being tagged by variants of higher frequencies (fixing the frequency of the tagged variant at 5% and varying the frequency of the causal variant from 4% to 1%), we also

observed a greater difference in power as the causal variant frequency decreased (Figure 2D). These results show that the difference in power between risk and protective variants should be more obvious when testing variants within the low-frequency range (<5% frequency), in polygenic diseases with lower prevalence, and when the markers being tested are proxies for lower-frequency causal variants. The driving force behind this result is that the case group is ascertained from individuals with an extreme distribution of liability scores whereas the control group has a much broader distribution of liability scores. Consequently, given equal sizes of the case and control groups, the increase in minor allele count of a risk variant in the case group is greater than the increase in minor allele count of an equally strong protective variant in the control group, leading to higher power for detecting the risk variant (see Appendix A for derived formulae that confirm the increase in power). Thus, if rare or low-frequency variants play a substantial role in certain diseases with polygenic architecture, these results predict that we could observe an increased R/P ratio for low-frequency variants in the GWAS summary statistics for these diseases.

Excess of Risk Variants Can Be Caused by Negative Selection

Beyond the differences in power, an excess of risk compared to protective variants can also occur if there is negative selection against the disease, leading risk variants to be kept at lower frequencies than protective variants. To illustrate this scenario, we simulated negative selection by coupling effects on evolutionary fitness and on a quantitative trait for a set of variants (frequency $\geq 1\%$) and then assigning case-control status based on the trait values (see Material and Methods). We observed an increase in the R/P ratio for the frequency bins within 1% to 15% but not for the 30% to 50% frequency bin (Figures 3A and S3). These results show that under a model where rare variants contribute to disease and are under negative selection, we could observe an increase in the R/P ratio for low-frequency variants in the GWAS summary statistics for these diseases.

Excess of Risk Variants Arise from Having More Control than Case Subjects

The previous results show that polygenic inheritance from lower-frequency variants can lead to an increase in the R/P ratio but that such an increase can occur in other settings as well. Under the null hypothesis, one would expect that on average, the number of detected risk variants would be equal to the number of detected protective variants, resulting in an expected R/P ratio of 1. However, in our simulations, we observed that the expected R/P ratio can deviate from 1 because of an imbalance between the sizes of the case and control groups. Specifically, if there are substantially more control than case subjects, a feature present in some GWASs of dichotomous traits, it would result in the increase of the expected R/P ratio (R/P
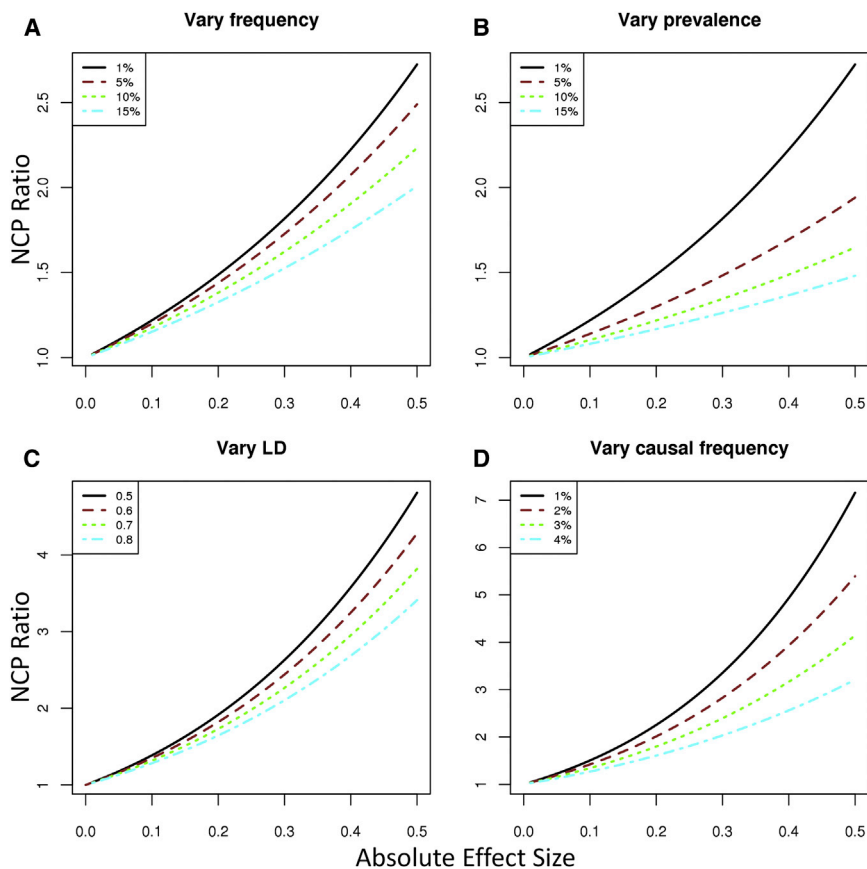
Figure 2. Effects of Varying Various Parameters on the NCP Ratio

The plots show the difference in power for detecting risk versus protective variants through the NCP ratio under varying parameters. Unless otherwise specified, the parameters used for calculating NCP are from the base model (see Material and Methods).

(A) Minor allele frequency of the associated variant varying from 1% to 15%.

(B) Disease prevalence (threshold of liability) varying from 1% to 15%.

(C) Linkage disequilibrium (LD) between the causal variant and the marker variant as a function of D′ (varying from 0.5 to 0.8).

(D) The marker variant frequency is set at 5% with the causal variant frequency ranging from 1% to 4%.

ratio > 1). To illustrate this, we randomly simulated 1,000 case subjects and 3,000 control subjects (1k/3k) and measured the distribution of the R/P ratio under a null model of no association (see Material and Methods). We observed that there is an increase in the R/P ratio distribution for 1k/3k for the low-frequency bins (Figures 3B and S4). This increase is not seen with common variants (30%–50% frequency bin), nor if the numbers of case and control subjects are equal (Figures 3B and S4). Of note, with larger sample sizes (10,000 case subjects and 30,000 control subjects; 10k/30k), we observed that the increase in R/P ratio is substantially attenuated (Figures 3B and S4). These results show that an excess of control subjects can increase the expected R/P ratio and should be accounted for by comparing the observed R/P ratio against those obtained through simulations under a null model. These results also show that with sufficiently large number of case subjects (e.g., >10,000), the increase in the expected R/P ratio resulting from this imbalance will be minimal.

Excess of Risk Variants Can Result from Asymmetric Population Stratification

We also considered whether an excess of risk variants could be seen in GWASs that are confounded by population stratification. As a first test, we randomly simulated 1,000 individuals of either northern European ancestry (CEU, based on allele frequencies in the CEU HapMap sample) or southern European ancestry (TSI, based on allele frequencies in the TSI HapMap sample). In one experiment, we simulated 1,000 CEU individuals as control subjects and 1,000 TSI individuals as case subjects (see Material and Methods), and as a stratification-free experiment, we simulated 1,000 CEU control subjects and 1,000 CEU case subjects. The simulated TSI and CEU populations show the expected differences in principal component analysis (Figure S5). We found that although there was a large excess of apparent associations for both risk and protective variants, leading to enormous inflation of the genomic control test statistic ($\lambda_{GC} \sim 22.9$), the resulting R/P ratio did not deviate substantially from expectations under the null (Figures 3C and S6). Therefore, even extreme scenarios with the usual forms of population stratification should not cause substantial deviations of the R/P ratio.

However, we reasoned that a special case of asymmetric population stratification could potentially cause the R/P ratio to depart from expectations under the null. Specifically, if there were a mixture of different populations in only the case group and not in the control group, or vice versa, it could lead to an increase or decrease of the R/P ratio. To test this, we randomly simulated a series of models where the control group is homogenous (CEU) but the case group is a mixture of CEU and TSI individuals (see Material and Methods). At a 1% mixture in the case group ($\lambda_{GC} \sim 1.01$), we did not observe any significant excess of risk variants, but at 5% mixture ($\lambda_{GC} \sim 1.06$), we observed an excess of risk variants within the low-frequency ranges (Figures 3D and S7). This excess is even larger with a 10% mixture ($\lambda_{GC} \sim 1.24$) (Figures 3D and S7). Variants within the common frequency range do not show an excess of risk variants (Figures 3D and S7). These results show that
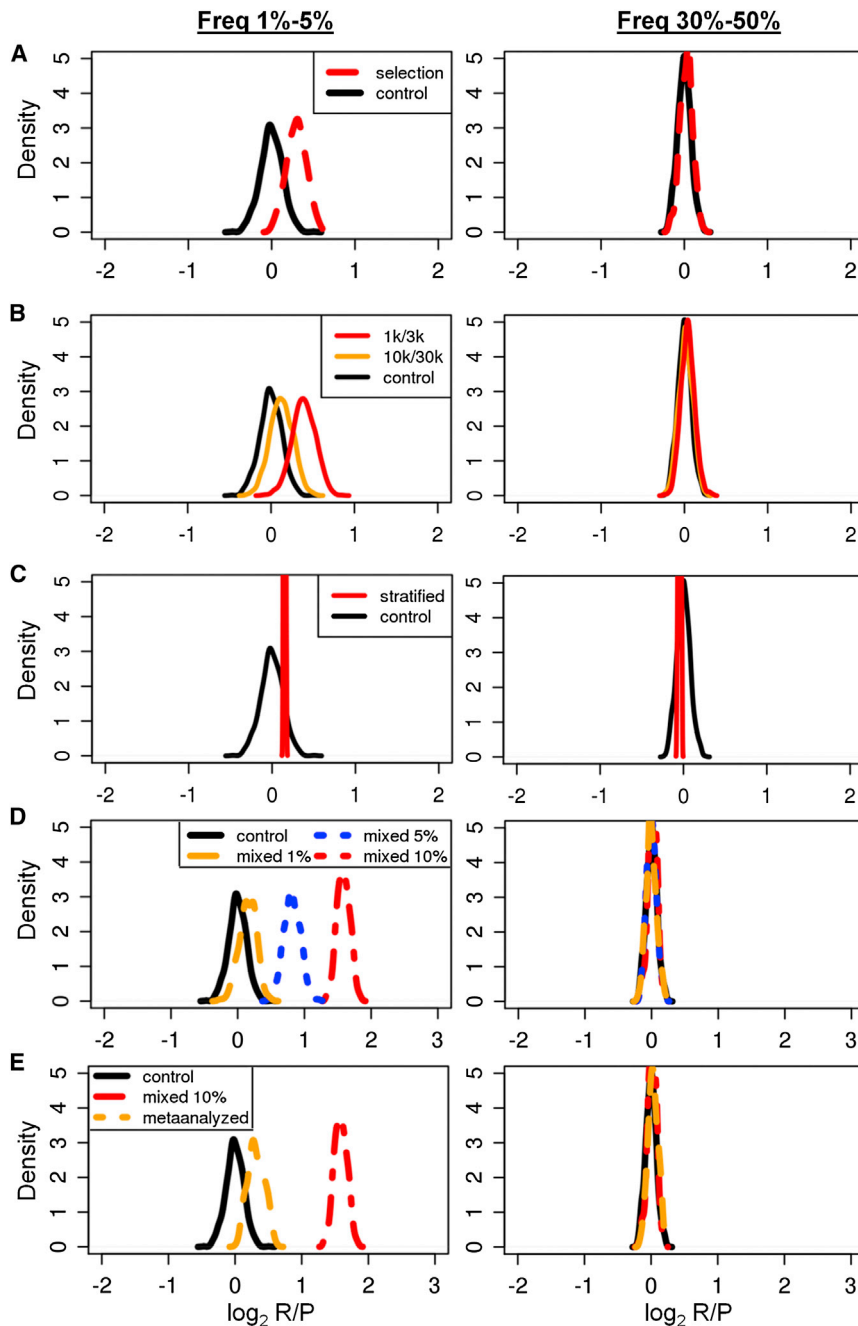
Figure 3. The Distribution of the R/P Ratio from Simulating Variants under Various Scenarios

The figure shows the distribution of the $\log_2$ R/P ratio for the 1%–5% and 30%–50% frequency bins from simulating variants under various scenarios. The p value cutoff for each of the bins is 0.01.

(A) Simulating variants under negative selection. The selection model (red) uses the 823 effective variants whereas the control (black) model assumes that no variants affect the phenotype.

(B) Simulating larger size of control than case group. The 1k/3k (red) model simulates the null distribution of the $\log_2$ R/P ratio for 1,000 case subjects and 3,000 control subjects. The 10k/30k (orange) model simulates the null distribution of the $\log_2$ R/P ratio for 10,000 case subjects and 30,000 control subjects. The control (black) model simulates the null distribution of the $\log_2$ R/P ratio for 1,000 case subjects and 1,000 control subjects.

(C) Simulating population stratification. The stratification model (red): case group simulated from TSI population and control group simulated from the CEU population. The control model (black): both case and control groups simulated from the CEU population.

(D) Simulating asymmetric population stratification. The models for asymmetric population stratification are as follows. Mixed 10%, 5%, and 1% indicate that 10%, 5%, and 1% of the case group is simulated from TSI individuals, respectively, and the rest of the individuals used are simulated from CEU individuals. The control model is comprised of case subjects simulated only from CEU individuals, i.e., without any population stratification.

(E) Simulating asymmetric population stratification after meta-analysis with nonstratified data. The model "mixed 10%" and "meta analyzed" refers to asymmetric population stratification of 10% mixture of TSI individuals of the case subjects before and after being meta-analyzed with four other data sets without such stratification, respectively. The control model indicates no asymmetric population stratification.

such asymmetric population stratification can increase the R/P ratio, with only moderate increases in the genomic control statistics. As a corollary, if the mixture were to exist in the control group but not in the case group, we would expect the R/P ratio to decrease.

Finally, we meta-analyzed the results from the asymmetrically stratified GWASs with results from nonstratified GWASs (see Material and Methods) to determine the effect on the R/P ratio if only a subset of the studies had asymmetric population stratification. We found that the increase in the R/P ratio is attenuated after meta-analysis (Figures 3E and S8). These results indicate that whereas asymmetric population stratification can give rise to an

excess of risk variants, combining such results with nonstratified results can reduce the magnitude of the signal. Because this particular type of stratification is unlikely to be present in most of the cohorts prior to meta-analysis, it may be useful to examine the summary statistics of each study individually to determine whether the increased R/P ratio is derived from a subset of studies in the GWAS meta-analysis. Ideally, if an increased R/P ratio is observed, principal component analysis or other methods should also be applied to the primary data to search for outliers present exclusively in the case group to further rule out asymmetric population stratification as a cause of an increased R/P ratio.

Table 1. Schizophrenia, Major Depressive Disorder, and Bipolar Disorder

Freq (%)	p Value Cutoff	SCZ			MDD			BIP		
		O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p
1–5	0.001	1.864	1.127	0.0298	1.210	1.058	0.269	0.884	1.110	0.748
	0.01	1.623	1.032	2.42×10^{-7}*	1.169	1.006	0.048	0.953	1.028	0.778
5–10	0.001	1.348	1.057	0.1279	0.933	1.039	0.623	1.038	1.077	0.509
	0.01	1.230	1.019	0.0111	0.914	1.005	0.865	0.973	1.013	0.678
10–15	0.001	1.050	1.082	0.4926	1.348	1.035	0.126	1.038	1.055	0.473
	0.01	1.054	1.019	0.3335	1.193	1.005	0.027	1.046	1.015	0.349
30–50	0.001	1.063	1.022	0.3736	1.098	1.003	0.264	1.122	1.039	0.291
	0.01	1.001	1.003	0.5010	0.944	1.001	0.836	1.070	1.009	0.165

The observed and expected R/P ratios and p values obtained from analyzing GWAS summary statistics of psychiatric disorders: schizophrenia (SCZ), major depressive disorder (MDD), and bipolar disorder (BIP). O(R/P) refers to the observed R/P ratio and E(R/P) refers to the expected R/P ratio obtained through simulations. p refers to the p value obtained from a one-tailed Z test (*p < 0.01).

Using the R/P Ratio in Actual GWAS Results to Search for Signals of Low-Frequency Variants Contributing to Disease Risk

Schizophrenia, Major Depressive Disorder, and Bipolar Disorder

We applied our method to data from several psychiatric disorders: schizophrenia,[11] bipolar disorder,[12] and major depressive disorder.[13] We observed a significant increase in the R/P ratio only for schizophrenia in the 1%–5% frequency bin, at a cutoff of p < 0.01 (p = 2.42×10^{-7}) (Table 1). We did not observe any significant differences in the other frequency bins nor for any of the other psychiatric disorders (Table 1). These results are indicative of polygenic inheritance from low-frequency variants in schizophrenia but do not provide similar support for a role of low-frequency variants in major depressive disorder or bipolar disorder.

Type 2 Diabetes

Next, we applied our method to GWAS results of type 2 diabetes.[14] The R/P ratio for type 2 diabetes was signifi-

cantly increased in the low-frequency bins (Table 2). The most significant difference was observed in the 1%–5% bin with cutoff of p < 0.01 (p = 3.08×10^{-15}). We also observed a significant excess of risk variants in the 10%–15% bin (p < 0.01, p = 2.28×10^{-5}). Because the difference in power between risk and protective variants becomes minimal as the variant frequency increases, this observed excess of risk variants is more probably due to negative selection on diabetes risk alleles, tagging of low-frequency variants by the more common SNPs in this frequency range, and/or possibly asymmetric population stratification. Nonetheless, these results are indicative of polygenic inheritance from low-frequency variants in type 2 diabetes.

Obesity

We also applied our method to GWAS results for various classes of obesity:[15] overweight (BMI > 25), class 1 (BMI > 30), class 2 (BMI > 35), and class 3 (BMI > 40). The control group used for each class of obesity were individuals with BMI < 25. We observed a significant increase in the 1%–5% frequency bin with a cutoff of p < 0.01 for only the class 1 data set (p = 8.8×10^{-6}) (Table 3). Also, although we generally observed a gradual increase in the R/P ratio with increasing BMI definitions of obesity, which could be consistent with a role of lower-frequency variants, the increase in R/P ratio could also be explained by having a larger control than case group. We did not observe any significant excess of risk variants for the low-frequency bins in the class 2 or class 3 data sets, probably because of the severely reduced sample sizes for the more extreme BMI definitions of obesity.

Testing whether Related Phenotypes Are Likely to Share Low-Frequency Causal Variants

To increase the power of GWASs, some studies have pooled apparently related phenotypes into a single case group.[16,17] We applied our method to measure the R/P ratio on published GWAS results of these related

Table 2. Type 2 Diabetes

Freq (%)	p Value Cutoff	T2D		
		O(R/P)	E(R/P)	p
1–5	0.001	3.833	1.205	5.89×10^{-6}*
	0.01	2.009	1.069	3.08×10^{-15}*
5–10	0.001	1.636	1.131	0.043
	0.01	1.439	1.051	2.28×10^{-5}*
10–15	0.001	1.660	1.081	0.031
	0.01	1.400	1.033	8.36×10^{-4}*
30–50	0.001	1.041	1.038	0.459
	0.01	1.035	1.008	0.308

The observed and expected R/P ratios and p values obtained from analyzing GWAS summary statistics of type 2 diabetes (T2D). O(R/P) refers to the observed R/P ratio and E(R/P) refers to the expected R/P ratio obtained through simulations. p refers to the p value obtained from a one-tailed Z test (*p < 0.01).

Table 3. Obesity

Freq (%)	p Value Cutoff	Overweight			Class 1			Class 2			Class 3		
		O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p
1–5	0.001	1.188	0.997	0.228	0.917	1.164	0.758	2.462	2.410	0.410	3.700	3.454	0.354
	0.01	1.120	0.986	0.078	1.536	1.050	8.8×10^{-6}*	1.533	1.376	0.114	1.814	1.617	0.111
5–10	0.001	1.026	0.998	0.408	1.139	1.098	0.393	0.697	1.640	0.999	1.857	2.067	0.607
	0.01	1.023	0.991	0.328	0.937	1.023	0.838	1.108	1.222	0.871	1.227	1.346	0.845
10–15	0.001	0.784	0.999	0.826	0.971	1.087	0.610	1.276	1.567	0.713	1.385	1.766	0.779
	0.01	1.109	1.003	0.113	1.013	1.028	0.544	1.066	1.208	0.883	1.269	1.267	0.479
30–50	0.001	1.121	0.991	0.194	1.059	1.020	0.380	0.949	1.094	0.763	1.019	1.112	0.696
	0.01	1.022	0.999	0.340	1.045	1.004	0.225	0.985	1.035	0.816	0.955	1.044	0.946

The observed and expected R/P ratios and p values obtained from analyzing GWAS summary statistics of clinical classes of obesity: overweight (BMI > 25), class 1 (BMI > 30), class 2 (BMI > 35), and class 3 (BMI > 40). O(R/P) refers to the observed R/P ratio and E(R/P) refers to the expected R/P ratio obtained through simulations. p refers to the p value obtained from a one-tailed Z test (*p < 0.01).

phenotypes. We reasoned that our method could also be used to test whether pooling related phenotypes would increase power to detect low-frequency variants, using only the GWAS summary statistics. We applied our method to GWAS results from two different pairs of related phenotypes, one pair for inflammatory bowel disease and one pair for diabetic nephropathy.

Inflammatory Bowel Disease

The two major types of inflammatory bowel disease are Crohn disease (CD) and ulcerative colitis (UC).[36] We examined the R/P ratio in GWAS results for Crohn disease,[32] ulcerative colitis,[33] and the combined case cohort of both Crohn disease and ulcerative colitis.[17] We observed significant increases in the R/P ratio for both Crohn disease and ulcerative colitis within the low-frequency bins (Table 4). The most significant increases were found in the 1%–5% bin with cutoff of p < 0.01 (CD, p = 1.55 × 10^{-10}; UC, p = 2.25 × 10^{-9}), consistent with a polygenic role of low-frequency variants in both diseases. However, when Crohn disease and ulcerative colitis were combined as a single case group (CD + UC), the increase in R/P ratio is less significant than in the individual GWAS results (Table 4). These results suggest that there are some low-frequency genetic contributors to Crohn disease and ulcerative colitis that are not shared by both diseases. However, because the signal is still present (albeit attenuated) when both diseases were studied together, it also suggests that the two diseases do share some overlapping low-frequency genetic contributors, although the attenuated signal could reflect persistence of two separate individual signals that are diluted after combination of the two sets of cases.

Diabetic Nephropathy

We performed a similar analysis on two phenotypes used to characterize diabetic nephropathy:[18] macroalbuminuria (MACRO) and end-stage renal disease (ESRD). Unlike inflammatory bowel disease, MACRO and ESRD are not necessarily distinct; MACRO is a milder form of diabetic nephropathy and some of the individuals thus affected

progress to develop ESRD. The control group used for that study were diabetic individuals that did not develop nephropathy. We analyzed the GWAS results performed for individuals with macroalbuminuria versus control subjects (MACRO$_{ctrl}$), individuals with end-stage renal disease versus control subjects (ESRD$_{ctrl}$), individuals with end-stage renal disease versus control subjects that also include individuals with macroalbuminuria (ESRD$_{ctrl+macro}$), and a combined case cohort that includes both individuals with macroalbuminuria and end-stage renal disease versus control subjects ([MACRO + ESRD]$_{ctrl}$). For the analyses of MACRO$_{ctrl}$ and of ESRD$_{ctrl}$, we observed significant increases to the R/P ratio in the 1%–5% bin with cutoff of p < 0.01 (MACRO$_{ctrl}$, p = 0.001; ESRD$_{ctrl}$, p = 6.4 × 10^{-5}) (Table 5). For the ESRD$_{ctrl+macro}$ analysis, where individuals with macroalbuminuria are included within the control group, there is an even larger increase of the R/P ratio (ESRD$_{ctrl+macro}$, p = 9 × 10^{-11}) (Table 5). However, when MACRO$_{ctrl}$ and ESRD$_{ctrl}$ were combined into a single case group ([MACRO + ESRD]$_{ctrl}$), none of the frequency bins showed significant increases in the R/P ratio (Table 5). These results suggest that although there are low-frequency contributors to both macroalbuminuria and end-stage renal disease, these contributors do not substantially overlap. There is no detectable increase in the R/P ratio when both phenotypes are combined, unlike our observations for inflammatory bowel disease. Thus, these results indicate that studies of low-frequency variation for diabetic nephropathy would be more fruitful if MACRO and ESRD are tested separately.

Discussion

We have shown that our method for measuring the R/P ratio can be used as a test for the presence of multiple low-frequency or rare genetic contributors to disease risk. This method can be applied to GWAS summary statistics,

Table 4. Inflammatory Bowel Disease: Crohn Disease and Ulcerative Colitis

Freq (%)	p Value Cutoff	CD			UC			CD+UC		
		O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p
1–5	0.001	2.545	1.347	0.017	1.958	1.358	0.075	1.385	1.159	0.222
	0.01	1.994	1.111	1.55×10^{-10}*	1.866	1.106	2.25×10^{-9}*	1.457	1.048	1.6×10^{-4}*
5–10	0.001	1.148	1.162	0.477	1.490	1.192	0.153	1.099	1.107	0.463
	0.01	1.314	1.069	1.4×10^{-3}*	1.460	1.066	8.59×10^{-5}*	1.239	1.027	0.012
10–15	0.001	1.200	1.181	0.424	1.279	1.186	0.337	1.583	1.076	0.059
	0.01	1.043	1.059	0.551	1.213	1.066	0.075	1.104	1.026	0.205
30–50	0.001	0.925	1.035	0.743	1.163	1.037	0.217	1.036	1.026	0.445
	0.01	1.052	1.018	0.266	1.004	1.009	0.524	1.043	1.005	0.251

The observed and expected R/P ratios and p values obtained from analyzing GWAS summary statistics of inflammatory bowel diseases: Crohn disease (CD), ulcerative colitis (UC), and the combined CD and UC as a single case group (CD+UC). O(R/P) refers to the observed R/P ratio and E(R/P) refers to the expected R/P ratio obtained through simulations. p refers to the p value obtained from a one-tailed Z test (*p < 0.01).

even if there are few or no genome-wide significant associations. We analyzed results from multiple published GWASs and found significant signals in some but not all diseases. These results support the hypothesis that the diseases where the R/P ratio is increased have a polygenic contribution from as-yet-undetected low-frequency or rare variants.

Some existing methods for detecting polygenic inheritance[9,10,37] use variants that achieve nominal significance in GWASs to determine whether they are informative as predictors of phenotype. Because our method assesses the direction of effect of these variants against the null model, our method represents a rather different, independent approach for assessing polygenic inheritance of low-frequency variants. Furthermore, our method does not require having identified associated loci or the availability of individual level data. For example, in schizophrenia, it has been shown that a substantial proportion of schizophrenia disease risk is the result of variants with frequency >1%.[38] Our finding suggests that some disease risk is accounted for by variants within the low-frequency range (frequency < 5%). In a recent exome-sequencing study of 2,536 schizophrenia cases and 2,543 controls,[39] Purcell and colleagues showed a polygenic burden of rare disruptive mutations, which is consistent with our observation. Similarly, for type 2 diabetes, our results suggest the presence of low-frequency or rare variants contributing to disease risk, even though most of the variants known to be associated with disease risk are common (frequency $\geq$ 5%).[14]

We also showed that negative selection under polygenic inheritance can increase the R/P ratio for low-frequency variants, because risk variants would be kept at lower frequencies while the protective variants could drift to higher frequencies. Indeed, in a previous study,[40] Park and colleagues showed that across most qualitative traits, minor alleles conferred risk more often than protection, which they concluded to be evidence for purifying

selection. Although this can be the case for some diseases, we showed that this increase in the R/P ratio can also arise because there is more power to detect risk variants than to detect protective variants. Furthermore, we have established that if there are substantially more control than case subjects, a feature present in many GWASs, this imbalance can distort the null distribution such that there would appear to be more risk than protective variants. However, this imbalance can be accounted for through simulations, as we have demonstrated.

Our method also provides a simple and early way of assessing the utility of different phenotype definitions for genetic studies of low-frequency variation simply from GWAS summary statistics. Our results for inflammatory bowel disease are consistent with the idea that Crohn disease and ulcerative colitis have some overlapping genetic contributors. Indeed, a previous study exploring the effect of common Crohn disease variants on ulcerative colitis identified significant overlaps between the two diseases, but also loci specific to Crohn disease.[41] For diabetic nephropathy, where there are few established loci from which to draw conclusions, we observed signals for both macroalbuminuria and particularly for end-stage renal disease when analyzed separately, but no significant signal when both diseases were combined as a single case group. This suggests that macroalbuminuria and end-stage renal disease are distinct in their genetic architecture and would be more productive if they were to be studied separately. Interestingly, the same GWAS on diabetic nephropathy discovered a single genome-wide significant locus only when end-stage renal disease was treated separately from macroalbuminuria,[16] consistent with our observation.

Finally, asymmetric population stratification between the case and control groups can lead to both false-positive associations (as evidenced by an increased genomic control inflation factor)[42] and also an increase in the R/P ratio. Thus, although our observations of higher-than-expected R/P ratios in some of the published GWAS data

Table 5. Diabetic Nephropathy: Macroalbuminuria and End-Stage Renal Disease

Freq (%)	p Value Cutoff	MACRO$_{ctrl}$			ESRD$_{ctrl}$			ESRD$_{ctrl+macro}$			[MACRO + ESRD]$_{ctrl}$		
		O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p	O(R/P)	E(R/P)	p
1–5	0.001	2.000	1.655	0.205	1.944	1.706	0.283	2.667	2.008	0.146	1.087	1.133	0.504
	0.01	1.560	1.198	1.4×10^{-3}*	1.705	1.207	6.4×10^{-5}*	2.270	1.285	9×10^{-11}*	1.026	1.042	0.550
5–10	0.001	1.563	1.359	0.253	1.278	1.404	0.585	1.533	1.584	0.496	0.875	1.071	0.754
	0.01	1.200	1.116	0.175	1.240	1.143	0.147	1.552	1.187	2.9×10^{-4}*	1.045	1.017	0.352
10–15	0.001	0.893	1.275	0.892	1.343	1.304	0.403	1.462	1.397	0.380	0.912	1.038	0.640
	0.01	1.208	1.104	0.150	1.190	1.128	0.258	1.310	1.160	0.078	1.053	1.009	0.290
30–50	0.001	1.122	1.066	0.343	1.198	1.051	0.197	0.968	1.076	0.719	1.037	1.001	0.382
	0.01	0.990	1.023	0.690	1.152	1.014	0.017	1.038	1.032	0.449	0.981	1.003	0.652

The observed and expected R/P ratios and p values obtained from analyzing GWAS summary statistics of diabetic nephropathy: macroalbuminuria (MACRO$_{ctrls}$), end-stage renal disease (ESRD$_{ctrls}$), ESRD versus controls that include MACRO (ESRD$_{ctrls+macro}$), and the combined MACRO and ESRD as a single case group ([MACRO + ESRD]$_{ctrls}$). O(R/P) refers to the observed R/P ratio and E(R/P) refers to the expected R/P ratio obtained through simulations. p refers to the p value obtained from a one-tailed Z-test (*p < 0.01).

sets are suggestive of a role of low-frequency variants, we cannot completely rule out that some of these signals could be in part explained by asymmetric population stratification. Of note, none of the R/P ratios showed a deficit of risk variants (which would be expected under some models of asymmetric population stratification), suggesting that asymmetric population stratification is not widespread. Furthermore, these GWASs have used methods to detect and correct for population stratification.

In conclusion, our method can be used to screen for polygenic inheritance from low-frequency or rare variants in diseases where GWASs have been performed. Our method can also be extended to other summary statistics, e.g., studies from sequencing or exome-chip genotyping, to assess low-frequency variants that were directly genotyped rather than imputed. This method can serve as a simple approach to guide researchers in prioritizing strategies in searching for as-yet-unexplained heritability for specific diseases. For example, in a study of epilepsy,[43] Heinzen and colleagues failed to identify any rare variants of large effect through exome sequencing; analysis of GWAS data for epilepsy can in theory help guide decisions about embarking on additional studies of low-frequency or rare variants with larger sample sizes. Although a lack of a signal from our method does not rule out a role for low-frequency variants and may reflect a combination of small sample sizes and a set of effect sizes and frequencies that do not significantly alter the R/P ratio, a positive signal can provide greater confidence about the likelihood that low-frequency or rare variants contribute to disease risk.

Appendix A

Calculating NCP from Various Given Parameters

We define the following parameters required to calculate the noncentrality parameter (NCP) as a function of effect size of minor allele (β), minor allele frequency (p), liability threshold (t), number of case individuals (N_d), and number of control individuals (N_c). We denote the minor allele (effect allele) as a_1 and the major allele (noneffect allele) as a_2. As such, the liability distribution of a_1 is $N(x, \mu_1, \sigma^2)$ and the liability distribution of a_2 is $N(x, \mu_2, \sigma^2)$ such that $N(x, \mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2.

The mean liabilities for a_1 and a_2 are as follows:

$$\text{Mean liability for } a_1 = \mu_1 = \beta - \beta p = \beta q$$

$$\text{Mean liability for } a_2 = \mu_2 = -\beta p$$

where q is the major allele frequency such that $p + q = 1$. The variance remaining σ^2 is:

$$\text{Variance remaining} = \sigma^2 = 1 - \beta^2 pq.$$

Next, we calculate a series of conditional probabilities as follows:

$$P(case \mid a_1) = \int_t^\infty N(x, \mu_1, \sigma^2) \mathrm{d}x$$

$$P(case \mid a_2) = \int_t^\infty N(x, \mu_2, \sigma^2) \mathrm{d}x$$

$$P(control \mid a_1) = \int_{-\infty}^t N(x, \mu_1, \sigma^2) \mathrm{d}x$$

$$P(control \mid a_2) = \int_{-\infty}^t N(x, \mu_2, \sigma^2) \mathrm{d}x.$$

With these conditional probabilities, we proceed to calculate the expected allele frequencies of both the

minor allele and major allele in both case subjects and control subjects by using Bayes' theorem. These are calculated as:

$$P_{d1} = P(a_1 \mid case) = \frac{P(case \mid a_1)p}{\int_t^{\infty} N(x, 0, 1)dx}$$

$$P_{d2} = P(a_2 \mid case) = 1 - P_{d1}$$

$$P_{c1} = P(a_1 \mid control) = \frac{P(control \mid a_1)p}{\int_{-\infty}^{t} N(x, 0, 1)dx}$$

$$P_{c2} = P(a_2 \mid control) = 1 - P_{c1}.$$

We then calculate the NCP by the χ^2 statistic from a 2 by 2 contingency table for the expectation of the observed number of a_1 and a_2 in both the case and control groups.

	Case	Control	Total
a_1	$2\,N_d\,P_{d1}$	$2\,N_c\,P_{c1}$	$2\,A$
a_2	$2\,N_d\,(1 - P_{d1})$	$2\,N_c\,(1 - P_{c1})$	$2\,B$
Total	$2\,N_d$	$2\,N_c$	$2\,T$

where

$$A = N_d\,P_{d1} + N_c\,P_{c1}$$

$$B = N_d(1 - P_{d1}) + N_c(1 - P_{c1})$$

$$T = A + B = N_d + N_c.$$

The expected number for each cell is the row total times the column total divided by the grand total.
Thus, the NCP is calculated as:

$$NCP = \sum_{Each\ cell} \frac{(Observed - Expected)^2}{Expected}$$

$$NCP = \frac{\left(2N_d\,P_{d1} - \frac{4AN_d}{2T}\right)^2}{\frac{4AN_d}{2T}} + \frac{\left(2N_c\,P_{c1} - \frac{4AN_c}{2T}\right)^2}{\frac{4AN_c}{2T}}$$
$$+ \frac{\left(2N_d\,P_{d2} - \frac{4BN_d}{2T}\right)^2}{\frac{4BN_d}{2T}} + \frac{\left(2N_c\,P_{c2} - \frac{4BN_c}{2T}\right)^2}{\frac{4BN_c}{2T}}$$

$$NCP = \frac{2TN_d P_{d1}^2}{A} + \frac{2AN_d}{T} - 4N_d P_{d1} + \frac{2TN_c P_{c1}^2}{A} + \frac{2AN_c}{T}$$
$$- 4N_c P_{c1} + \frac{2TN_d(1 - P_{d1})^2}{B} + \frac{2BN_d}{T} - 4N_d(1 - P_{d1})$$
$$+ \frac{2TN_c(1 - P_{c1})^2}{B} + \frac{2BN_c}{T} - 4N_c(1 - P_{c1})$$

$$NCP = \frac{2TN_d P_{d1}^2}{A} + \frac{2TN_c P_{c1}^2}{A} + \frac{2TN_d(1 - P_{d1})^2}{B}$$
$$+ \frac{2TN_c(1 - P_{c1})^2}{B} - 2T.$$

After some algebra and simplification,

$$NCP = \frac{2T}{AB} N_d N_c (P_{d1} - P_{c1})^2.$$

Therefore,

$$NCP = 2N_d N_c (P_{d1} - P_{c1})^2 \left(\frac{N_d + N_c}{(N_d P_{d1} + N_c P_{c1})(N_d P_{d2} + N_c P_{c2})} \right).$$

We verified that these formulae were correct by comparing to simulated results.

Determining NCP Ratio between Risk and Protective Variants with the Same Magnitude of Effect
We formulated the various probabilities between risk and protective variants. Assuming β to be positive, the risk variant would have the following probabilities:

$$P_{d1} = \frac{p \int_t^{\infty} N(x, \beta q, \sigma^2)dx}{\int_t^{\infty} N(x, 0, 1)dx}$$

$$P_{c1} = \frac{p \int_{-\infty}^{t} N(x, \beta q, \sigma^2)dx}{\int_{-\infty}^{t} N(x, 0, 1)dx}$$

and the protective variant with the same magnitude of effect would have the following probabilities:

$$P_{d1} = \frac{p \int_t^{\infty} N(x, -\beta q, \sigma^2)dx}{\int_t^{\infty} N(x, 0, 1)dx}$$

$$P_{c1} = \frac{p \int_{-\infty}^{t} N(x, -\beta q, \sigma^2)dx}{\int_{-\infty}^{t} N(x, 0, 1)dx}.$$

Assuming that there are equal number of case and control subjects ($N_1 = N_2$), then

$$NCP \ \alpha \ \left(\frac{\int_t^{\infty} N(x, \beta q, \sigma^2)dx}{\int_t^{\infty} N(x, 0, 1)dx} - \frac{\int_{-\infty}^{t} N(x, \beta q, \sigma^2)dx}{\int_{-\infty}^{t} N(x, 0, 1)dx} \right)^2.$$

The ratio between risk and protective variants with the similar magnitude of β is therefore

$$NCP\ ratio = \frac{\left(\dfrac{\int_t^\infty N(x, \beta q, \sigma^2)\mathrm{dx}}{\int_t^\infty N(x,0,1)\mathrm{dx}} - \dfrac{\int_{-\infty}^t N(x, \beta q, \sigma^2)\mathrm{dx}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2}{\left(\dfrac{\int_t^\infty N(x, -\beta q, \sigma^2)\mathrm{dx}}{\int_t^\infty N(x,0,1)\mathrm{dx}} - \dfrac{\int_{-\infty}^t N(x, -\beta q, \sigma^2)\mathrm{dx}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2},$$

We can transform the distributions such that

$$NCP\ ratio = \frac{\sigma\left(\dfrac{\int_{\frac{t-\beta q}{\sigma}}^\infty N(z,0,1)\mathrm{dz}}{\int_t^\infty N(x,0,1)\mathrm{dx}} - \dfrac{\int_{-\infty}^{\frac{t-\beta q}{\sigma}} N(z,0,1)\mathrm{dz}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2}{\sigma\left(\dfrac{\int_{\frac{t+\beta q}{\sigma}}^\infty N(y,0,1)\mathrm{dy}}{\int_t^\infty N(x,0,1)\mathrm{dx}} - \dfrac{\int_{-\infty}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2},$$

where $z = (x - \beta q)/\sigma$, $y = (x + \beta q)/\sigma$, and $dx = \sigma dz = \sigma dy$.

Then,

$$NCP\ ratio = \frac{\left(\dfrac{\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz}}{\int_t^\infty N(z,0,1)\mathrm{dz}}\right)^2}{\left(\int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}\right)^2}$$

When prevalence is 50% ($t=0$),

$$\int_{\frac{-\beta q}{\sigma}}^0 N(z,0,1)\mathrm{dz} = \int_0^{\frac{+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}.$$

and therefore

$$NCP\ ratio = 1.$$

This shows that when prevalence is 50% ($t = 0$) and there are equal sample numbers in the case and control groups ($N_1 = N_2$), the NCP between risk and protective variants with identical magnitudes of effect (β) would be the same regardless of any other parameters.

For the case where $t > 0$, if

$$\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz} - \int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy} > 0,$$

$$NCP\ ratio = \frac{\left(\dfrac{\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz} + \int_t^\infty N(z,0,1)\mathrm{dz}}{\int_t^\infty N(x,0,1)\mathrm{dx}} - \dfrac{\int_{-\infty}^t N(z,0,1)\mathrm{dz} - \int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2}{\left(\dfrac{\int_t^\infty N(y,0,1)\mathrm{dy} - \int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_t^\infty N(x,0,1)\mathrm{dx}} - \dfrac{\int_{-\infty}^t N(y,0,1)\mathrm{dy} + \int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2}$$

$$NCP\ ratio = \frac{\left(\left(1 + \dfrac{\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz}}{\int_t^\infty N(x,0,1)\mathrm{dx}}\right) - \left(1 - \dfrac{\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)\right)^2}{\left(\left(1 - \dfrac{\int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_t^\infty N(x,0,1)\mathrm{dx}}\right) - \left(1 + \dfrac{\int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)\right)^2}$$

$$NCP\ ratio = \frac{\left(\dfrac{\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz}}{\int_t^\infty N(x,0,1)\mathrm{dx}} + \dfrac{\int_{\frac{t-\beta q}{\sigma}}^t N(z,0,1)\mathrm{dz}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2}{(-1^2)\left(\dfrac{\int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_t^\infty N(x,0,1)\mathrm{dx}} + \dfrac{\int_t^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^t N(x,0,1)\mathrm{dx}}\right)^2}$$

then the NCP for risk variants will be greater than the NCP for protective variants and the NCP ratio will be greater than 1. When $t > \beta q$, this will be true because the normal distribution is monotonic decreasing above $z = 0$ ($y = 0$).

To extend this to the more general case of $t > 0$, we first examine the individual components,

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)dz = \int_{-\infty}^{t} N(z,0,1)dz - \int_{-\infty}^{\frac{t-\beta q}{\sigma}} N(z,0,1)dz$$

$$= \frac{1}{2}\left[1 + \text{eft}\left(\frac{t}{\sqrt{2}}\right)\right] - \frac{1}{2}\left[1 + \text{eft}\left(\frac{t-\beta q}{\sigma\sqrt{2}}\right)\right]$$

$$= \frac{1}{2}\left[\text{eft}\left(\frac{t}{\sqrt{2}}\right) - \text{eft}\left(\frac{t-\beta q}{\sigma\sqrt{2}}\right)\right]$$

where eft is the error function. Similarly,

$$\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)dy = \int_{-\infty}^{\frac{t+\beta q}{\sigma}} N(y,0,1)dy - \int_{-\infty}^{t} N(y,0,1)dy$$

$$= \frac{1}{2}\left[1 + \text{eft}\left(\frac{t+\beta q}{\sigma\sqrt{2}}\right)\right] - \frac{1}{2}\left[1 + \text{eft}\left(\frac{t}{\sqrt{2}}\right)\right]$$

$$= \frac{1}{2}\left[\text{eft}\left(\frac{t+\beta q}{\sigma\sqrt{2}}\right) - \text{eft}\left(\frac{t}{\sqrt{2}}\right)\right].$$

Therefore,

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)dz - \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)dy$$

$$= \frac{1}{2}\left[\text{eft}\left(\frac{t}{\sqrt{2}}\right) - \text{eft}\left(\frac{t-\beta q}{\sigma\sqrt{2}}\right)\right] - \frac{1}{2}\left[\text{eft}\left(\frac{t+\beta q}{\sigma\sqrt{2}}\right) - \text{eft}\left(\frac{t}{\sqrt{2}}\right)\right]$$

$$= \text{eft}\left(\frac{t}{\sqrt{2}}\right) - \frac{1}{2}\text{eft}\left(\frac{t-\beta q}{\sigma\sqrt{2}}\right) - \frac{1}{2}\text{eft}\left(\frac{t+\beta q}{\sigma\sqrt{2}}\right).$$

Taking the first two terms of the Taylor-series expansion of the error function and approximating σ to 1 ($\sigma \approx 1$),

$$\text{eft}\left(\frac{t}{\sqrt{2}}\right) - \frac{1}{2}\text{eft}\left(\frac{t-\beta q}{\sigma\sqrt{2}}\right) - \frac{1}{2}\text{eft}\left(\frac{t+\beta q}{\sigma\sqrt{2}}\right)$$

$$\cong \frac{2}{\sqrt{\pi}}\left(\frac{t}{\sqrt{2}} - \frac{t^3}{6\sqrt{2}}\right) - \frac{1}{2}\left(\frac{2}{\sqrt{\pi}}\right)\left(\frac{t-\beta q}{\sqrt{2}} - \frac{(t-\beta q)^3}{6\sqrt{2}}\right)$$

$$- \frac{1}{2}\left(\frac{2}{\sqrt{\pi}}\right)\left(\frac{t+\beta q}{\sqrt{2}} - \frac{(t+\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{12t}{6\sqrt{2}} - \frac{2t^3}{6\sqrt{2}} - \frac{6t - 6\beta q}{6\sqrt{2}} + \frac{(t-\beta q)^3}{6\sqrt{2}}\right.$$

$$\left. - \frac{6t + 6\beta q}{6\sqrt{2}} + \frac{(t+\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{-2t^3 + (t-\beta q)^3 + (t+\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{-2t^3 + t^3 - 3t^2\beta q + 3t(\beta q)^2 - (\beta q)^3 + t^3 + 3t^2\beta q + 3t(\beta q)^2 + (\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{t(\beta q)^2}{\sqrt{2}}\right).$$

As such, if $t > 0$,

$$\frac{1}{\sqrt{\pi}}\left(\frac{t(\beta q)^2}{\sqrt{2}}\right) > 0.$$

Therefore, if $t > 0$,

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)dz > \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)dy$$

$$NCP\ ratio > 1.$$

Therefore, for diseases with low prevalence ($t > 0$), there is more power to detect risk variants compared with the protective variant.

Supplemental Data

Supplemental Data include consortia information and eight figures and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgements

Web Resources

The URL for data presented herein is as follows:

PISA, http://www.broadinstitute.org/~rigel/pisa/

References

1. Hirschhorn, J.N., and Gajdos, Z.K.Z. (2011). Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annu. Rev. Med. *62*, 11–24.
2. Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am. J. Hum. Genet. *86*, 6–22.
3. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

4. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.

5. Lim, E.T., Raychaudhuri, S., Sanders, S.J., Stevens, C., Sabo, A., MacArthur, D.G., Neale, B.M., Kirby, A., Ruderfer, D.M., Fromer, M., et al.; NHLBI Exome Sequencing Project (2013). Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron *77*, 235–242.

6. Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N., et al. (2013). Using whole-exome sequencing to identify inherited causes of autism. Neuron *77*, 259–273.

7. Iyengar, S.K., and Elston, R.C. (2007). The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol. Biol. *376*, 71–84.

8. Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.

9. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

10. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

11. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

12. Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. *43*, 977–983.

13. Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H., Boomsma, D.I., Cichon, S., et al.; Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry *18*, 497–511.

14. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.

15. Berndt, S.I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. *45*, 501–512.

16. Sandholm, N., Salem, R.M., McKnight, A.J., Brennan, E.P., Forsblom, C., Isakova, T., McKay, G.J., Williams, W.W., Sadlier, D.M., Mäkinen, V.-P., et al.; DCCT/EDIC Research Group (2012). New susceptibility loci associated with kidney disease in type 1 diabetes. PLoS Genet. *8*, e1002921.

17. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

18. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

19. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

20. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

21. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

22. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.

23. Adams, A.M., and Hudson, R.R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics *168*, 1699–1712.

24. Agarwala, V., Flannick, J., Sunyaev, S., and Altshuler, D.; GoT2D Consortium (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat. Genet. *45*, 1418–1427.

25. Lambert, B.W., Terwilliger, J.D., and Weiss, K.M. (2008). ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. Bioinformatics *24*, 1821–1822.

26. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871–3876.

27. Eyre-Walker, A. (2010). Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. USA *107* (*Suppl 1*), 1752–1756.

28. Chan, Y., Holmen, O.L., Dauber, A., Vatten, L., Havulinna, A.S., Skorpen, F., Kvaløy, K., Silander, K., Nguyen, T.T., Willer, C., et al. (2011). Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. PLoS Genet. *7*, e1002439.

29. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

30. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191.

31. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics 55, 997–1004.

32. Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42, 1118–1125.

33. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A., et al. (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat. Genet. 43, 246–252.

34. Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. 29, 51–76.

35. Slatkin, M. (2008). Exchangeable models of complex inherited diseases. Genetics 179, 2253–2261.

36. Baumgart, D.C., and Carding, S.R. (2007). Inflammatory bowel disease: cause and immunobiology. Lancet 369, 1627–1640.

37. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2013). Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol. Biol. 1019, 215–236.

38. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R.; Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS) (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. 44, 247–250.

39. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185–190.

40. Park, J.-H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., Jr., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. USA 108, 18026–18031.

41. Anderson, C.A., Massey, D.C.O., Barrett, J.C., Prescott, N.J., Tremelling, M., Fisher, S.A., Gwilliam, R., Jacob, J., Nimmo, E.R., Drummond, H., et al.; Wellcome Trust Case Control Consortium (2009). Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. Gastroenterology 136, 523–529.e3.

42. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11, 459–463.

43. Heinzen, E.L., Depondt, C., Cavalleri, G.L., Ruzzo, E.K., Walley, N.M., Need, A.C., Ge, D., He, M., Cirulli, E.T., Zhao, Q., et al. (2012). Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. Am. J. Hum. Genet. 91, 293–302.