Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants

Ju-Hyun Park[a], Mitchell H. Gail[a], Clarice R. Weinberg[b], Raymond J. Carroll[c], Charles C. Chung[d], Zhaoming Wang[d], Stephen J. Chanock[a,d], Joseph F. Fraumeni, Jr.[a,1], and Nilanjan Chatterjee[a,1]

[a]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, MD 20852; [b]Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC 27709; [c]Department of Statistics, Texas A&M University, College Station, TX 77843-3143; and [d]Core Genotyping Facility, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Gaithersburg, MD 20877

Contributed by Joseph F. Fraumeni, Jr., September 12, 2011 (sent for review July 18, 2011)

Recent discoveries of hundreds of common susceptibility SNPs from genome-wide association studies provide a unique opportunity to examine population genetic models for complex traits. In this report, we investigate distributions of various population genetic parameters and their interrelationships using estimates of allele frequencies and effect-size parameters for about 400 susceptibility SNPs across a spectrum of qualitative and quantitative traits. We calibrate our analysis by statistical power for detection of SNPs to account for overrepresentation of variants with larger effect sizes in currently known SNPs that are expected due to statistical power for discovery. Across all qualitative disease traits, minor alleles conferred "risk" more often than "protection." Across all traits, an inverse relationship existed between "regression effects" and allele frequencies. Both of these trends were remarkably strong for type I diabetes, a trait that is most likely to be influenced by selection, but were modest for other traits such as human height or late-onset diseases such as type II diabetes and cancers. Across all traits, the estimated effect-size distribution suggested the existence of increasingly large numbers of susceptibility SNPs with decreasingly small effects. For most traits, the set of SNPs with intermediate minor allele frequencies (5–20%) contained an unusually small number of susceptibility loci and explained a relatively small fraction of heritability compared with what would be expected from the distribution of SNPs in the general population. These trends could have several implications for future studies of common and uncommon variants.

genetic prediction | missing heritability | population genetics

Large meta-analyses of genome-wide association studies (GWAS) have now identified more than 1,000 susceptibility loci for complex traits. Nevertheless, for most complex traits, the fraction of heritability explained by common variants remains below 10–15%, even for traits for which large numbers of loci have been detected [e.g., dozens to over 200 (1–3)]. We and others (4, 5) have projected that complex traits are likely to have an increasingly large number of susceptibility loci that have correspondingly smaller individual contributions to heritability. A fraction of these loci could be detected in future GWAS with large, but realistic, sample sizes, but they are still unlikely to fully explain a large fraction of missing heritability (4). The spectrum of genetic variation is greater than originally anticipated and suggests that the underlying genomic architecture of many diseases and traits could be more complex. Availability of newer-generation genotyping chips and sequencing technologies has raised the hope that future studies of uncommon and rare variants could increase the rate of discovery and explain an additional fraction of heritability, and eventually approach clinically useful discriminatory performance for genetic risk models.

The discoveries generated from GWAS provide important insights into the genetic architecture of complex traits while also providing new opportunities to understand the biology of complex diseases. Analyses of the distribution of susceptibility single-nucleotide polymorphism (SNPs) in relation to various genomic features and pathways (1, 3) have suggested clues for the biologic basis of genetic susceptibility for a number of traits. The distribution of various population genetic parameters across susceptibility loci may provide further insight into population genetic models with important implications for future studies. A major challenge, for example, for future studies of low-frequency susceptibility variants is that statistical power for their discovery may be low unless they have relatively larger effects. Although population genetic models suggest that such a trend is expected under purifying selection (6, 7), the implications are often unclear for traits, such as late-onset diseases, that are not directly related to fitness. Availability of a large number of susceptibility loci across many traits now provides the research community with the opportunity to test such hypotheses empirically.

A complication for investigating any population genetic hypothesis using only known susceptibility loci is that such a set may not be representative of the spectrum of underlying susceptibility loci for which the inference is desired. Based on statistical power considerations, for example, variants with lower allele frequencies and small effects on the trait are expected to be systematically underrepresented in the current set of known loci. Thus, an inverse relationship between allele frequencies and effects may be observed simply due to the nature of ascertainment of the current set of known loci.

In this report, we use data from about 400 susceptibility loci across 13 different traits to examine the distribution of allele frequencies, effect-size parameters, and heritability explained by common susceptibility loci. In these evaluations, we account for differential probabilities for ascertainment for different SNPs based on estimates of their statistical power for detection in the original discovery studies. We evaluate the relationship between allele frequencies and regression effects, such as log odds ratios and linear regression coefficients, that have been typically reported to summarize association strength in existing studies. In addition, we provide estimates for the number of underlying susceptibility loci and their contribution to genetic variance within categories of allele frequencies for different traits. These

Author contributions: J.-H.P., M.H.G., C.R.W., and N.C. designed research; J.-H.P. and N.C. performed research; J.-H.P., R.J.C., and N.C. contributed new reagents/analytic tools; J.-H.P., C.C., and Z.W. analyzed data; and J.-H.P., M.H.G., C.R.W., S.J.C., J.F.F., and N.C. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: fraumeni@nih.gov or chattern@mail.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1114759108/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1114759108

analyses provide empirical insights into the genetic architecture of select complex traits. Several implications for future discoveries are discussed.

Results

We selected a set of traits that include both quantitative [height (1) and lipid levels (2)] and qualitative phenotypes [type I (8) and II (9) diabetes, common cancers (10–15), and Crohn's disease (3)]; the traits also occur as early- (type I diabetes) and late-onset diseases (type II diabetes, certain forms of cancer, and Crohn's disease), and so far have demonstrated a diverse range of heritability (Table 1). For each trait, we only analyzed independent susceptibility SNPs, typically observed from distinct genomic regions, so that all results can be correspondingly interpreted in the context of the underlying set of independent susceptibility SNPs. We included traits that have a minimum of 20 susceptibility SNPs identified from recent GWAS in subjects of European background so that sufficient statistical power exists for investigating the underlying trends. For cancers, we considered a pooled analysis for six different sites (breast, prostate, colon or rectum, bladder, pancreas, and brain), as numbers of reported loci for some of the individual outcomes were small. We chose these specific sites for pooling because they are known to have a similar degree of familial aggregation (sibling recurrence risk = 2–3), and recent GWAS have indicated that these sites may have similar numbers of susceptibility loci in a similar range of effect sizes.

Distribution of Allele Frequency for Susceptibility SNPs. Investigation of the distribution of minor-allele frequencies (MAF) suggests (Fig. 1) that for all traits, except possibly for HDL level, the distribution of observed susceptibility SNPs is skewed toward higher minor-allele frequencies (MAF >20%) rather than intermediate frequencies (MAF 5–20%) in comparison with SNP allele-frequency distributions in general human populations or among tagging SNPs that have been included in common genotyping platforms. Overall, out of 387 SNPs included in the analysis for all traits combined, the fraction of SNPs with intermediate-frequency categories was only 23.0%, which was significantly lower than the corresponding fraction of 55.0% among independent representative SNPs (any pairwise $r^2 \leq 0.1$) from the HapMap (hapmap.ncbi.nlm.nih.gov) database ($P = 2.05 \times 10^{-30}$). The power-weighted analysis also estimated a relatively small fraction (26.4%) of susceptibility SNPs for the intermediate-frequency category, and thus indicated that the observed clustering of common susceptibility SNPs toward higher frequencies is unlikely to have resulted from the artifacts of study power.

Next, we investigated the frequency distribution of "risk" alleles (Fig. S1). For disease traits, we define risk alleles as variants that correspond to a disease odds ratio greater than one. For lipid level, we define risk alleles as variants that are positively associated with total cholesterol and LDL, the increased level of which is known to confer risk of heart disease, but negatively associated with HDL, the increased level for which is considered protective. For height, although such definitions are more ambiguous, we considered variants that are positively associated with height as a risk allele because increased mortality has been previously reported for taller subjects (16–19).

Table 1. Traits, estimates of total heritability, and number of known susceptibility SNPs used in reported analysis

Trait	Height	TC	HDL	LDL	Cancer*	CD	T1D	T2D
Estimate of heritability	0.8	0.275	0.275	0.275	∼2	27.5	15	3
No. of independent SNPs	114	51	46	37	59	68	31	25

Heritability for quantitative traits is reported as the fraction of total variance attributable to susceptibility, and that for qualitative traits is reported as sibling recurrence risks. TC, total cholesterol; CD, Crohn's disease; T1D, type I diabetes; T2D, type II diabetes.

*Includes cancers of breast, prostate, colon or rectum, pancreas, bladder, and brain.

For all disease traits, the risk variants tended to be minor alleles (frequency <50%) rather than major alleles in populations of European background (Fig. S1 and Table S1). The pattern is most profound and statistically significant for type I diabetes, for which 72.4% of the risk variants were minor alleles ($P = 0.004$ for testing $\pi \leq 0.5$ vs. $\pi > 0.5$). No such pattern was apparent for quantitative traits. We further investigated the distribution of minor versus major alleles among variants that conferred the highest risk for each trait. Among SNPs with risk coefficients (log odds ratio for disease trait and linear regression coefficient for continuous trait) in the highest quartile of coefficients, the likelihood for the risk variant being the less prevalent allele increased for all traits except for height, Crohn's disease, and type II diabetes (Table S1).

Distribution of Effect Sizes for Susceptibility SNPs. We define "effect size" for susceptibility SNPs using two alternative criteria. In one, we define it as the coefficient (β) for a SNP when its association with the outcome is modeled through a regression model, such as linear regression for a quantitative trait or logistic regression for a qualitative trait, assuming a linear trend per copy of an allele. In our analysis, the regression coefficients for quantitative traits are presented in units of standard deviation (SD) of the trait so that they are comparable across traits. In a second criterion, we define effect size as the contribution of the SNP to genetic variance of the trait, that is, $gv = 2\beta^2 f(1 - f)$, where f is the allele frequency for either of the two SNP alleles (4). It is noteworthy that the power for detection of a susceptibility SNP for most commonly used association tests that assume linear trend depends on β and f only through the quantity gv (4). In figures and tables, we present gv as a fraction of the total genetic variance σ_G^2 of a trait attributable to heritability. For qualitative traits, the variance due to heritability is computed from estimates of sibling recurrence risk (Table 1) using a log-normal model for risk (20).

Within trait (both quantitative and qualitative), the distributions of the absolute values of regression coefficients were comparable except for type II diabetes, in which the distribution of log-odds ratios was shifted toward lower absolute values compared with other diseases (Fig. S2). When effect sizes are expressed as a fraction of genetic variances explained by a susceptibility SNP, the loci for lipid levels and cancers appeared to have larger contributions than those for other traits of the same type (quantitative vs. qualitative) (Fig. S2). Smoothed nonparametric estimates of the effect-size distribution (Fig. 2 and Fig. S3) across susceptibility loci revealed a clear effect of adjustment for study power. Across all traits, the density of the effect sizes for the observed SNPs initially increased with decreasing effect sizes, reached a peak, and then decreased at the lowest range. The estimated power-adjusted density of effect sizes for all underlying susceptibility SNPs, however, continued to increase at an increasingly faster rate, as the effect size decreased. This analysis suggests there are increasingly large numbers of susceptibility loci with decreasing effect sizes for complex traits, regardless of trait.

We also explored alternative parametric models that could describe the effect-size distributions for the different traits. We observed that an exponential model, which has been argued based on population genetic theory (21–24), is often inadequate for describing the distribution of genetic variances for the common susceptibility SNPs in complex traits (Fig. S3). In particular, a single exponential distribution predicts more susceptibility SNPs with larger effect sizes than observed in current GWAS. Interestingly, models based on mixtures of exponential distributions

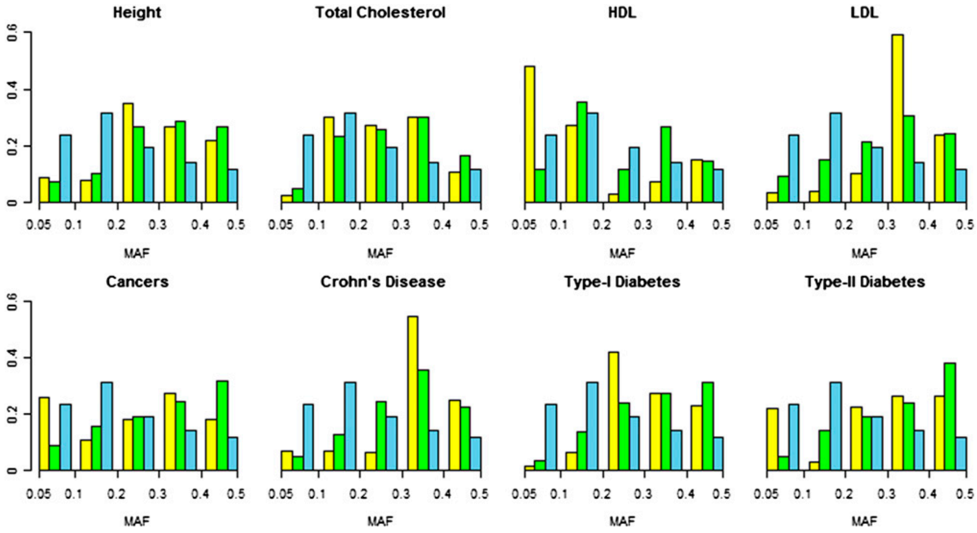

Fig. 1. Distribution of frequencies for minor alleles across an estimated number of susceptibility SNPs (yellow), observed susceptibility SNPs (green), and independent representative SNPs in the HapMap project (blue).

often provided a good description of underlying distribution of *gv* because they allowed for a very large number of small effect sizes and for a small number of relatively larger effect sizes. Analogously, we observed that scale mixtures of normal distributions, instead of a single normal distribution, can provide a stable description of distributions of regression coefficients (β) associated with common susceptibility SNPs.

Relationship Between Allele Frequency and Effect Size. We explored the relationship between allele frequency and effect size in different scales. An inverse relationship between the squared regression coefficient and $f(1 - f)$ was observed consistently across different traits (Fig. 3). For a number of these traits, however, the strengths of these relationships become less pronounced after adjustment for ascertainment due to study power. The strength of the trend, as captured by the slope of the fitted line (Table 2), markedly varies between traits, with an almost 10-fold change between the two extremes of distinct types of traits. After adjustment, the most pronounced trend was seen for type I diabetes and Crohn's disease among qualitative traits and LDL level among quantitative traits. In exploring the relationship between the frequency of the risk allele and the magnitude of the associated risk coefficient (Fig. S4), we observed a quadratic pattern that indicates increasing risk coefficients as the risk-allele frequency diverges away from 0.50 either toward 0 or toward 1. Thus, it appears that regression coefficients for common susceptibility SNPs increase in magnitude monotonically with decreasing minor-allele frequency, irrespective of whether the minor allele confers risk or protection. However, for some traits, such as type I diabetes, risk alleles were predominantly minor alleles, that is, they had frequencies of less than 0.50.

The genetic variance explained by individual SNPs generally remains constant over allele-frequency ranges (Fig. S5). Thus, the increasing trend for the regression effects for SNPs with decreasing $f(1 - f)$ (Fig. 3) compensates for diminishing contribution of a SNP to genetic variance due to its lower prevalence in the population. For some traits that demonstrated stronger inverse correlation between regression coefficients and allele frequencies, a slight increase was observed for genetic variance with decreasing allele frequency (Fig. S5 and Table S2).

Estimates for Number of Underlying Susceptibility SNPs and Their Contribution to Heritability. As a by-product of the weighted analysis for distribution of allele frequencies and effect sizes, we obtained estimates for the total number of underlying susceptibility loci for the different traits with effect sizes in the range that is observed in current studies (Table 3). These estimates show that for all traits there are a large number of additional loci that

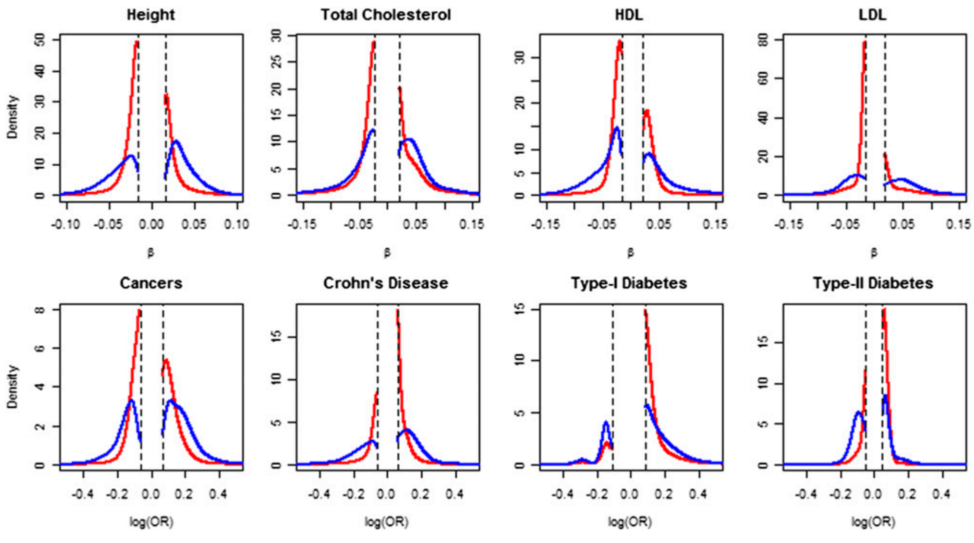

Fig. 2. Smoothed estimate of distribution of regression coefficients associated with minor alleles for susceptibility loci, shown with (red) and without power adjustment (blue). OR, odds ratio.

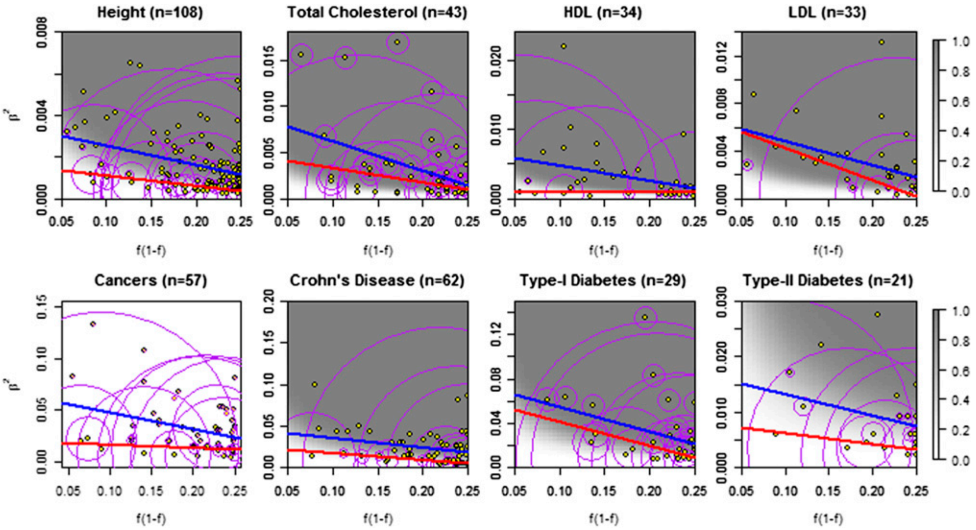

Fig. 3. Scatter plots for the 2D distribution of regression effects and minor-allele frequencies for observed susceptibility SNPs. The analysis is performed in the scale of squared regression coefficients β^2 and $f(1 - f)$, which are the components that define the contribution of a SNP to genetic variance. In the background of each plot, the power of the original discovery study is shown in gray scale over different regions of the parameter space. The weight for each SNP, which is the inverse of its power for detection, is proportional to the area of the purple circle surrounding it. Fitted lines with (red lines) and without weights (blue lines) are shown.

are currently undetected but might be discovered in larger GWAS. However, the total genetic variance explained by all of these SNPs, including those yet undetected, remains modest. Estimates for the total number of susceptibility SNPs for specific MAF categories correspond to the pattern we described earlier (Fig. 1), namely that susceptibility variants tend to cluster toward more common allele-frequency (MAF >20%) than intermediate-frequency (MAF =5–20%) categories. Consequently, our analysis indicates that although the contributions of individual SNPs to heritability are similar over different allele frequencies, collectively the contribution of SNPs with intermediate-frequency categories is substantially smaller than those in more common frequency categories.

Discussion

In this report, we used estimates of regression coefficients, minor-allele frequencies, and study power for a large number of recently reported susceptibility SNPs across a wide variety of traits to obtain empirical insight into the genetic architecture of SNPs related to complex traits. Our analysis is unique in that we draw inferences not only for the observed set of loci but also for an underlying "population" of susceptibility loci, many of which are not yet detected due to limited statistical power of current studies. In this regard, our results bear important implications discussed below.

Selection. Our analysis provides empirical support for a long-standing population genetic model that effect sizes for susceptibility variants should be inversely related to their allele frequencies. As the underlying theory is based on purifying selection, it is remarkable that among all of the different traits we studied it was type I diabetes, a disease of early onset that is most likely to be subject to selection pressure, that showed the strongest trend both in terms of correlation between allele frequencies and regression coefficients and the proportion of risk alleles that were less common (frequency <50%). For most other traits, including late-onset diseases such as cancers and type II diabetes, these trends, although consistently present, were modest. It is also noteworthy that for all traits, the absolute value of the regression coefficient showed an increasing trend with decreasing minor-allele frequency irrespective of whether the minor allele conferred risk or protection (Fig. S4). Such a pattern is consistent with a pleiotropic model (7) under which susceptibility variants could have effects in opposite directions for different traits, for example, across an underlying "fitness" trait that drives selection pressure and a specific trait we study.

Future Discovery and Prediction Based on Common Variants. We observed that commonly assumed models of effect sizes, that is, an exponential distribution for genetic variances or a normal distribution for regression coefficients, predict higher numbers of susceptibility SNPs with relatively larger effects than has been seen with current GWAS data. The distribution of effect sizes has important implications for future discoveries as well as for genetic prediction. For example, a fitted exponential distribution to genetic variances for susceptibility SNPs for height predicts that a total of 1,485 SNPs would be needed to explain 45% of variance of height, the fraction of heritability that was recently attributed to common susceptibility SNPs using a variance component analysis (5). In contrast, a mixture of two exponential distributions, which provided a much better fit to the data, estimated that at least 7,244 SNPs would be needed to explain the same fraction of heritability. The number of discoveries expected in the future and their contribution to heritability could be quite different under these two models. GWAS of 500,000 subjects, for example, could expect to discover a total of 1,216 and 2,333 loci explaining 44.1% or 32.9% of the heritability of height under the single-exponential versus mixture-exponential models, respectively.

The distribution of effect sizes also has implications for risk prediction using polygenic models that may include additional loci based on a more liberal significance threshold than commonly used for discovery in GWAS (25). Again assuming that 45% of the heritability of height can be explained by common variants, for example, we estimated that a polygenic model built on a training dataset of about 100,000 subjects could achieve a maximum out-of-sample predictive power (maximized over different significance levels) corresponding to an R^2 value about 27.8% under the single-exponential model, 23.7% under a two-component mixture model, and only 17% under a three-component mixture model. The reduction in predictive power under the later models occurs because more precise estimates of association coefficients are needed when the true coefficients are smaller for SNPs to contribute positively to the performance of risk models. Thus, even if in theory common susceptibility SNPs could explain a large fraction of heritability for complex traits, the distribution of effect sizes indicates that, in light of practical constraints for sample size in future GWAS, a large fraction of the remaining SNPs may have effects that are too small to be detected individually or to make a major contribution collectively in predictive models.

Implications for Power for Future Studies of Uncommon Variants. Assuming that the observed trend will continue for lower allele-

Table 2. Linear regression analysis between squared regression coefficients (β^2) and minor-allele frequencies for susceptibility SNPs

	Unweighted		Weighted	
Trait	Slope	*P* value	Slope	*P* value
Qualitative				
Type I diabetes	−2.20E-01	6.24E-02	−2.15E-01	1.03E-02
Crohn's disease	−1.14E-01	3.02E-02	−8.06E-02	8.21E-02
Cancers	−1.64E-01	8.85E-03	−3.05E-02	5.30E-01
Type II diabetes	−3.90E-02	1.83E-01	−2.00E-02	3.49E-01
Quantitative				
LDL	−2.05E-02	1.66E-02	−2.75E-02	9.45E-04
Total cholesterol	−3.20E-02	7.09E-03	−1.57E-02	2.54E-01
Height	−9.22E-03	1.78E-04	−4.80E-03	1.43E-02
HDL	−2.20E-02	5.99E-02	−1.80E-03	7.99E-01

The slopes are for the regression of squared regression coefficients against $f(1 - f)$, where f is allele frequency. These variables are the components that define the contribution of a SNP to genetic variance. The weighted and unweighted analyses are performed with and without adjustment for study powers, respectively. The traits are sorted by the strength of power-adjusted slope estimates, a measure of the strength of linear relationship.

frequency ranges, our analysis provides insight into what kind of effect sizes we might expect for future studies of uncommon variants, such as those with MAF in the range of 1–5%; what their statistical power for detection from association studies could be; and how much they may contribute to heritability. Although our analysis provides some support that less common alleles are likely to have larger effects, the trend appears to be quite modest after adjusting for lower power for discovery of loci with smaller effect sizes. Based on the fitted lines shown in Fig. 3, for example, we estimated that for a disease like Crohn's disease, the average regression effects correspond to an odds ratio of 1.08 for MAF = 0.45, 1.13 for MAF = 0.15, and 1.16 for MAF = 0.05. Such trends, although modest, suggest that the genetic variance due to individual susceptibility SNPs, on average, remains fairly constant over different ranges of allele frequency (Fig. S5 and Table S2). This indicates that the strength of an association test statistic, which is closely related to genetic variance, could also be expected to remain similar, on average, for common (including intermediate and perhaps uncommon) variants. Still, the power for detecting an individual uncommon susceptibility SNP, using a completely agnostic approach, may be reduced by a multiple-testing penalty as the number of markers increases in future studies for comprehensive coverage for the lower spectrum of allele frequencies.

Contribution to Heritability of Uncommon Variants. We observed that the contribution of individual SNPs to genetic variance is similar on average over different ranges of allele frequency. This raises the possibility that susceptibility SNPs in uncommon allele-frequency categories may explain a large fraction of heritability if such variants are present in larger proportions and follow the distribution of SNPs that are currently annotated in human populations. We observed, however, that for several different traits, both the number of susceptibility SNPs and their collective contribution to genetic variance were highest for more common MAF categories, 30–40% or 40–50%, and dropped substantially for lower allele-frequency categories, 5–10% or 10–20%. Our power-adjusted analysis confirmed that such observed patterns are not caused by lower statistical power for detection of association for SNPs with lower frequencies. Thus, trends in data from current GWAS do not suggest that susceptibility loci with intermediate and uncommon allele frequencies could explain a large fraction of missing heritability.

Certain population genetic models predict that in the future a large fraction of missing heritability for complex traits could be explained by loci that contain classes of rare (MAF <1%) susceptibility variants (7, 26). Our analysis of common susceptibility SNPs does not provide evidence for or against such a hypothesis, because we cannot extrapolate our results to loci that have complex allelic architecture and are not currently represented in our analysis.

Limitations. Some caveats of the current analysis merit discussion. It is noteworthy that our inference is based on common SNPs (MAF ≥5%) that are expected to have high coverage in current genotyping platforms used in existing GWAS. We cannot readily extrapolate the observed trends to uncommon and rare variants. It is possible, for example, that a stronger inverse correlation exists between regression effects and allele frequencies for the more rare variants than would be predicted based on common susceptibility SNPs included in our analyses. Nevertheless, we believe that the observed trends over a wide spectrum of allele frequencies provide clues to patterns that might emerge from future studies of less common variants.

It is also noteworthy that our analysis cannot be generalized to the population of all common susceptibility SNPs. It is likely that there are many common SNPs that have effect sizes so small that they virtually did not have any power to be detected and hence represented in current studies. Our power analysis marks the regions of parameter space for which the current studies have no representation of the underlying susceptibility SNPs (Fig. 2). Despite such truncation, the power-adjusted analyses provide useful population-based interpretation of the results for genetic architecture of complex traits. Moreover, given the large sample sizes for some of the existing studies, it seems that the effect sizes

Table 3. Estimates for the total number of underlying loci and their contribution to genetic variance for the underlying traits

	Height		TC		HDL		LDL		Cancer		CD		T1D		T2D	
MAF range	Est. no.	GV*	Est. no.	GV	Est. no.	GV	Est. no.	GV	Est. no.	GV	Est. no.	GV	Est. no.	GV	Est. no.	GV
0.05–0.1	55.3	1.3	2.0	1.2	113.4	5.2	4.0	0.9	17.4	3.2	29.4	1.2	1.0	0.2	52.3	2.5
0.1–0.2	50.1	1.9	24.9	4.3	63.8	8.4	5.0	2.0	7.1	2.3	29.1	1.8	4.7	0.9	6.7	1.1
0.2–0.3	224.2	5.7	22.6	7.8	6.8	1.5	12.7	5.0	12.1	3.2	27.5	2.6	31.1	3.5	54.4	3.9
0.3–0.4	172.6	4.7	24.9	5.0	16.9	3.8	74.7	6.4	18.2	9.4	233.8	8.7	20.4	2.4	63.6	4.2
0.4–0.5	140.5	4.9	8.8	2.2	35.8	2.4	30.1	3.1	12.0	4.3	107.1	5.4	17.1	2.5	62.7	5.1
Total	642.7	18.9	83.3	20.5	236.6	21.3	126.5	17.4	66.9	22.5	426.9	19.8	74.4	9.5	239.7	16.8

All projections are restricted to the effect-size ranges that are observed in current studies.

*All genetic variances (GVs) are shown as the percentage of the total variance of the trait attributable to heritability. For qualitative traits, the variance due to heritability is computed from estimates of sibling recurrence risk shown in Table 1 using a log-normal model for risk.

that remain completely undetectable would be so small that they are unlikely to be discovered in large proportion in future studies with realistic sample sizes. Thus, our analysis reveals those trends and patterns for susceptibility loci that are likely to be detectable in association studies.

Based on our estimates of statistical power for discovery of known susceptibility SNPs and the presentation of our statistical framework, our analysis of the distributions of allele frequencies and effect-size parameters for a large number of common susceptibility SNPs provides insight into the population genetic architecture of complex traits. Future studies with additional array content for lower allele-frequency SNPs together with sequencing will certainly discover new loci, and will provide an additional opportunity to investigate different population genetic models for further understanding of the differences and similarities we already observe in the genetic architecture across diverse complex traits.

Methods

SNP Selection. We attempted to follow a general algorithm for selecting the susceptibility SNPs to be included in our analysis across different traits. For each trait, we identified the largest GWAS reported to date. From published reports, we identified independent susceptibility SNPs that reached genome-wide significance (Table S3). In some multistage studies, susceptibility SNPs that had been reported in previous studies were not pursued beyond the first stage (or GWA meta-analysis). We only included previously reported SNPs if their association *P* values from the first stage reached the study's threshold for follow-up to subsequent stages. The underlying rationale here is that if the current study followed up all SNPs that met their first-stage selection criterion irrespective of results from previous studies, then our analysis would have included only those previously reported SNPs that reached the required significance at first stage.

Estimation of Effect Size and Powers. To avoid overestimation of effect size due to the problem of the winner's curse (27), we attempted to obtain estimates of regression coefficients and minor-allele frequencies from independent replication studies whenever such data were included as part of the original report. In the absence of such data, we obtained estimates of these parameters from the final stage of the studies if a multistage design was reported. For single-stage studies with no independent replication data, we used a statistical technique (27) to correct for the winner's curse and compared analyses with and without such correction. We evaluated the power of each of the original studies at the estimated values of effect-size parameters following the exact design of these studies (Table S3).

Power-Adjusted Analysis. In each power-adjusted analysis, a SNP is included with a weight as the inverse of its power of detection in the corresponding discovery study (see Fig. 3 and Figs. S4 and S5 for pictorial representations of weights). Intuitively, the set of observed susceptibility SNPs represents a random sample from the underlying population of susceptibility SNPs, where different SNPs are selected with different sampling probability due to their different effect sizes. By weighting each observed SNP by the inverse of its sampling probability, which in this case is its statistical power for detection, we allow it to represent the underlying population of SNPs that have similar effect sizes and hence have similar probabilities of sampling. For example, if a SNP has an effect size that corresponds to a statistical power of 25%, then the weight of the SNP is 1/0.25 = 4, implying that it is considered to represent four susceptibility SNPs with similar effect sizes from the underlying population. The use of inverse-probability weighting methods for unbiased estimation of population parameters is motivated by methods used in statistical sample surveys (28, 29), where unequal probability sampling is commonly used to increase study efficiency.

The weighted analysis of the SNPs allows generalization of inference only to the section of the population for which the sampling probability is nonzero. To reduce instability associated with SNPs with very low power and consequently large associated weights, we restricted our analysis to SNPs that had at least 1% power in the current studies. Thus, the conclusion we draw from our analysis should be taken as holding for the part of the parameter region where the current studies have ≥1% power. For ease of visualization of this region, we have shown how the power of the different studies varies over the parameter space in the background of Fig. 3 using a gray scale.

We assessed the statistical significance of linear relationship between allele frequencies and effect-size parameters by bootstrap resampling methods. In each bootstrap run, we randomly select a set of SNPs from the observed SNPs with replacement. For each such sample, we repeat the original analysis with the associated weights for the sampled SNPs. We evaluated the SE for the slope of the fitted line over 1,000 bootstrap samples to obtain an estimate of uncertainty of the underlying relationship that is due to randomness of the underlying sampling mechanism for observed SNPs.

ACKNOWLEDGMENTS. This research was supported by the intramural programs of the National Cancer Institute and National Institute of Environmental Health Sciences. The research of R.J.C. was supported by a grant from the National Cancer Institute (R27-CA057030).

1. Lango Allen H, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467:832–838.
2. Teslovich TM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466:707–713.
3. Franke A, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 42:1118–1125.
4. Park JH, et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 42:570–575.
5. Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42:565–569.
6. Wright S (1938) The distribution of gene frequencies in populations of polyploids. *Proc Natl Acad Sci USA* 24:372–377.
7. 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.
8. Barrett JC, et al.; Type 1 Diabetes Genetics Consortium (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 41:703–707.
9. Voight BF, et al.; MAGIC Investigators; GIANT Consortium (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579–589.
10. Turnbull C, et al.; Breast Cancer Susceptibility Collaboration (UK) (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* 42:504–507.
11. Eeles RA, et al.; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators; PRACTICAL Consortium (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* 41:1116–1121.
12. Houlston RS, et al.; Colorectal Cancer Association Study Consortium; CoRGI Consortium; International Colorectal Cancer Genetic Association Consortium (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 40:1426–1435.
13. Rothman N, et al. (2010) A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* 42:978–984.
14. Petersen GM, et al. (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* 42:224–228.
15. Shete S, et al. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* 41:899–904.
16. Benson VS, Pirie K, Green J, Casabonne D, Beral V; Million Women Study Collaborators (2008) Lifestyle factors and primary glioma and meningioma tumours in the Million Women Study cohort. *Br J Cancer* 99:185–190.
17. Dieckmann KP, et al. (2008) Tallness is associated with risk of testicular cancer: Evidence for the nutrition hypothesis. *Br J Cancer* 99:1517–1521.
18. Moore SC, et al. (2009) Height, body mass index, and physical activity in relation to glioma risk. *Cancer Res* 69:8349–8355.
19. Paajanen TA, Oksala NK, Kuukasjärvi P, Karhunen PJ (2010) Short stature is associated with coronary heart disease: A systematic review of the literature and a meta-analysis. *Eur Heart J* 31:1802–1809.
20. Pharoah PD, et al. (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 31:33–36.
21. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
22. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, New York).
23. Orr HA (1998) The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. *Evolution* 52:935–949.
24. Orr HA (2005) Theories of adaptation: What they do and don't say. *Genetica* 123:3–13.
25. Purcell SM, et al.; International Schizophrenia Consortium (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748–752.
26. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69:124–137.
27. Ghosh A, Zou F, Wright FA (2008) Estimating odds ratios in genome scans: An approximate conditional likelihood approach. *Am J Hum Genet* 82:1064–1074.
28. Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 47:663–685.
29. Korn EL, Graubard BI (1991) Epidemiologic studies utilizing surveys: Accounting for the sampling design. *Am J Public Health* 81:1166–1173.