Generalizing from sequencing studies

C Ryan King, Paul J Rathouz, Dan L Nicolae

April 7, 2014

1 Abstract

Several authors have suggested the use of case-control sequencing studies for rare diseases since they tend to enrich the sample for rare high-impact mutations. We demonstrate one price of this enrichment: case-control studies selectively ascertain risk-increasing SNPs, and when used for prediction tend underestimate the cumulative effect of SNPs which they observe and overestimate the effect of SNPs which are novel in other studies. This bias is functionally restricted to the set of SNPs with minor allele frequency (MAF) near 1 out of the sample size, and is exaggerated when SNP log-odds ratios come from a heavy-tailed distribution. Importantly, different MAF-dependent predictions are generated by different sample sizes. We illustrate how the effect of pooled rare SNPs varies with disease prevalence and the distribution of SNP effects even when the mean log odds ratio is zero, and how this effect interacts with the use of a case-control sample. We also show parallel effects on the power of allele-count tests.

2 Introduction

As the cost of sequencing declines there is increasing interest in its use for “personalized medicine” and prediction about disease risks. One might expect that prediction methods from GWAS-based research would be applicable; however, we will explore some shortcomings of those techniques in sequencing. In GWAS, data collected from either prospective (cohort) or retrospective (case-control) studies allows one to estimate the odds-ratio (OR) on the probability of having the trait. Those OR estimates can be taken from either a single study or pooled using meta-analysis and used for plug-in risk prediction. Because array genotyping is relatively inexpensive and GWAS requires large sample sizes to overcome strict multiple testing thresholds, effect estimates from GWAS tend to be fairly precise for common SNPs. In contrast, the unique advantage of sequencing association studies is the ability to detect previously unknown rare mutations for correlation to the trait. Gene-level association tests designed for sequence data make use of rare SNPs, but for their role in prediction and estimation has been less well defined in the literature. Sequencing studies will continue to discover new rare SNPs for the foreseeable future, and individuals sequenced for risk prediction will continue to carry previously unseen SNPs \[1\, 2\]; therefore, we will never be in the comfortable position of having precise plug-in estimates for each SNP. There best we
can hope for with seldomly observed SNPs is a pooled effect estimate - a “rare SNP effect” - which would could use to extrapolate to new samples.

In this paper we discuss two issues with estimated rare SNP effects. In Section 3.1 we calculate the relationship between the marginal effect of pooled rare SNPs and the underlying distribution of individual SNP effects. The marginal effect turns out to depend on the mean and variance of log-odds ratios as well as the disease prevalence rather than simply reflecting the mean. In addition to demonstrating the impact of these factors on the estimated rare-SNP effect, we show that the power of parallel allele-count tests have the same relationships. The interpretation of these tests therefore requires caution; they are not tests of the mean effect of SNPs.

Section 3.2 turns to the use of SNPs discovered in a case-control sequencing study for prediction. Case-control studies are somewhat more efficient than prospective designs for discovering rare risk-increasing SNPs since frequency among the cases is a function of the odds ratio. Because the sampling of SNPs is biased with respect to odds ratios, predictions based on the observed population of rare SNP effects are biased if that ascertainment is not accounted for. Explicitly, the observed effect tends to understate the extent to which those same SNPs increase risk in external samples. In contrast, previously unobserved SNPs (those novel in follow-up samples) tend to have a weaker or protective association with the trait. Conditioning on whether or not a SNP was previously detected may seem like an odd step; however, that is exactly the decision made when following up a case-control study with custom genotyping or further sequencing. Importantly, this effect is only of relevance for SNPs whose MAF is near one over the sample size. At higher frequencies, essentially all SNPs are ascertained regardless of their effect on trait; at lower frequencies very few SNPs are ascertained.

**Notaion.** We denote the disease status of the $i^{th}$ participant as $Y_i = 1$ for diseased and $= 0$ for undiseased for $N$ cases and $N$ controls. We take $G$ to be a $2N \times M$ matrix of participant genotypes, $G_{ij} = k$ for the $i^{th}$ participant having the minor allele at the $j^{th}$ locus $k = \{0, 1, 2\}$ times, and $G_i$ the row vector for the $i^{th}$ participant. To isolate the behavior of SNPs at different MAFs, our analysis will fix all SNPs to come from a common MAF $f$, which we will vary between simulations.

We describe the data model in two ways; the first of which we call the SNP-specific model. At the participant level,

$$\text{logit} \left( Pr\{Y_i = 1|G_i, \gamma\} \right) = G_i \gamma + \gamma_0,$$

where $\gamma_j$ denotes the log-odds ratio of the $j^{th}$ SNP and $\text{logit}()$ is the logistic link function. The SNP effects are described as coming from a population with $E[\gamma] = \mu$ and $V\text{ar}(\gamma) = \tau^2$.

We also define the population average or marginal model:

$$\text{logit} \left( Pr\{Y_i = 1|g_i, \beta\} \right) = g_i \beta + \beta_0,$$

where $g_i$ is a vector of counts of minor alleles in several buckets of frequency for the $i^{th}$ person. In our examples where all SNPs come from the same MAF, $g_i$ is just a scalar of the count of minor alleles in a gene for that person. We refer to a test of $\beta_g = 0$ as an allele-count test; for example a t-test of the number of minor alleles between cases and controls.
3 Results

3.1 Population average effects.

We can obtain a useful approximate formula for the $\beta$ parameters in equation (2) if we assume $\gamma \sim N(\mu, \tau^2)$. First, re-write the logistic model as the equivalent latent liability model,

$$Pr\{Y_i = 1|G_i, \mu, \tau\} = Pr\{X_i + G_i \gamma > c\} \quad (3)$$

where $X_i$ is a latent variable following a logistic distribution and $c$ is the threshold value, $c = -\log(p/1-p) \approx -\log(p)$ for $p$ the prevalence of the disease among those with no SNPs. Next approximate the logistic variable by a Gaussian scaled by 1.6, yielding

$$Pr\{Y_i = 1|G_i, \mu, \tau\} \approx Pr\{Z_i \sqrt{1.6^2 + \tau^2 g_i} + \mu g_i > c\},$$

where $Z$ is a standard normal, and re-apply the Normal-logistic approximation

$$Pr\{Y_i = 1|G_i, \mu, \tau\} \approx Pr\{X_i^* + \frac{\mu g_i - c}{\sqrt{1 + \tau^2 g_i/1.6^2}} > 0\},$$

where $X^*$ is again logistic distributed. Returning to a usual logistic regression form we have,

$$\text{logit} (Pr\{Y_i = 1|g_i, \mu, \tau\}) \approx -\frac{c + \mu g_i}{\sqrt{1 + \tau^2 g_i/1.6^2}}. \quad (4)$$

While approximation (4) is usually acceptable for prevalence greater than 5%, it fails for lower prevalences as the tails of the logistic and normal distributions diverge. The solid curve of the right panel of Figures 1, 2, 3 display the expected lOR varying the prevalence, mean, and standard deviation using equation 3. Figure S1 shows the divergence of formulas (3) and (4) at low prevalences. We validate the results of Figures 1, 2, 3 in simulation. We simulate large cohorts varying each of the parameters ($\mu, \tau, c$ and the mean $g$) and tabulate the mean disease status by $g$; Tables S1 - S4 display the results, which are in agreement with the Figures. The complete specification of the simulation details is found in the Supplement.

The important features of Equation (4) are that the pooled lOR increases with the mean SNP effect, the variance of SNP effects, and the negative log of the disease prevalence, down to about 5% as shown in Figure S1. One useful application of Equation (4) is the lOR contrasting those carrying one minor allele and those with zero minor alleles. When the mean IOR is zero and the standard deviation is less than 1, the IOR is approximately

$$\text{lOR}\{g = 1\} \approx \frac{c\tau^2/1.6^2}{2\sqrt{1 + \tau^2/1.6^2}}, \quad (5)$$

which increases sharply with the standard deviation of IORs and scales with $-\log(p)$. We illustrate the impact of these parameters on allele-count tests in Figures S3, S4, and S5 which plot the power of an allele-count test versus the standard deviation of log-odds ratios, disease-prevalence, and number of SNPs respectively when $\mu = 0$. As one expects, the effect on power of allele-count tests mirrors the effect on the IOR.
3.2 Sampling Bias by Odds Ratio

The distribution of the number of times a SNP is observed in cases and controls is binomial with odds among the controls and cases,

\[
\text{odds}_{\text{controls}} = \frac{f}{1 - f} \frac{1}{1 + p(OR - 1)},
\]

\[
\text{odds}_{\text{cases}} = \frac{OR}{1 - f} \frac{1}{1 + p(OR - 1)},
\]

where \( f \) is the MAF and \( p \) the disease prevalence among the unexposed. Using a Poisson approximation to the binomial, the probability that a SNP is observed in either the \( N \) cases or \( N \) controls can be derived:

\[
Pr\{\text{count observed} > 0\} \approx 1 - \exp(-f \ N(1 + OR)).
\]

There are three important things to note about formula (7). Risk-associated alleles are enriched into the case-sample, but the enrichment of disease-preventing alleles into the control-sample is negligible in a rare-disease setting. Figure 4 displays the difference and ratio of ascertainment probability between risk-increasing and risk-decreasing alleles with the same magnitude log-odds ratio; the risk-increasing allele is more likely to be sampled for all MAFs.
Figure 2: **Mean of SNP IORs versus predictive IOR**  
Left panel: mean of SNP IORs (x-axis) versus mean of novel (dotted) and previously sampled (dashed) SNP IORs. Right panel: mean of SNP IORs versus predictive IOR for novel (dotted), previously sampled (dashed) and all (solid) SNPs. IOR of 1 SNP versus 0 SNPs. Prevalence 5%, SD of SNP IORs 0.6, MAF 0.005, sample size 100 each group.

and magnitudes of IOR. The non-equality of expected log-odds ratios of observed SNPs and the generative population of SNPs is what we refer to as ascertainment bias in this context.

Figure 4 displayed the effect of retrospective ascertainment on fixed IORs, but it is more natural to consider IORs which come from a continuous distribution. Figure 5 shows the expected value of previously observed and novel IORs for several members of the truncated Gaussian and Student’s t distribution family. Distributions with heavier tails than a Gaussian exhibit a higher bias and an exaggerated effect at very low MAF. The cumulative distributions plots of the IORs under these scenarios are displayed in Figure S6 which shows that they are very similar in the region of zero and primarily differ in the likelihood of large uncommon IORs.

Second, due to the exponential form of formula (7), the above bias is meaningful only for a narrow range of MAF and peaks for SNPs with MAF = 1/2N. The ascertainment probability quickly saturates to nearly one as $N \cdot f$ becomes more than 5 regardless of the odds ratio ($1 - e^{-5} \approx 0.993$). The investigator is virtually assured of noticing the existence of common alleles, so differential ascertainment become of negligible magnitude. The difference in the right panel of Figure 4 also goes to zero for very low allele frequencies; even a quite deleterious OR does not include many very rare SNPs in the observed set. Although these outcomes become uncommon, this is not equivalent to the bias becoming negligible; as can be seen in Figure 5, the expected IOR of observed SNPs remains high for very small MAF, and the expected IOR of novel SNPs strongly negative for high MAF. Failing to discover a SNP with MAF > $1/N$ suggests that this SNP is protective and therefore has low frequency among
Figure 3: **SD of SNP IORs versus predictive IOR.** Left panel: SD of SNP IORs (x-axis) versus mean of novel (dotted) and previously sampled (dashed) SNP IORs. Right panel: SD of SNP IORs versus predictive IOR for novel (dotted), previously sampled (dashed) and all (solid) SNPs. IOR of 1 SNP versus 0 SNPs. Prevalence 5%, mean of IORs 0, MAF 0.005, sample size 100 each group.

While the variance of ascertained SNP effects varies with MAF, the change is small in the Gaussian case and only notable for the t distributions which allow large SNP effects and only for comparatively rare SNPs. Intuitively, case control sampling allows rare SNPs with quite large effects to be sampled into the cases, increasing the variability of observed IORs.

Third, the ascertainment probability in formula (7) depends on the sample size, rather than just the minor allele frequency. Above we argued that non-representative sampling can only be a problem for a particular range of low to rare frequencies; however, that range is a function of the sample size. This creates a differential bias between the SNP sets observed in studies of different sample sizes and ensures that as long as there are low MAF SNPs the above bias will exist regardless of the sample size. We recreate Figure 5 but with an increased sample size in Figure S2, which shifts the location of the curve but leaves the shape the same. Finally, a feature lost in approximation (7) is the dependence on baseline disease prevalence. Examining the probabilities in Equation 6, as the prevalence goes to zero the enrichment of risk increasing SNPs in cases is maximized. This is graphically illustrated in Figure 1.

The distribution of true effect sizes in the study is not necessarily the object of interest; we also illustrate how the sampling bias affects the observed data and the validity of predictions. Using equation (3) and (6), Figures 1, 2 and 3 display the IOR using previously observed and novel SNPs to predict the trait in an independent replication cohort. We also validate these calculations in simulation. In each simulation, we draw a balanced case-control study from a large population as well as an independent prospective replication cohort; details of the simulation setup are found in the Supplement and the results in Tables S5, S6, S7, S8.
Figure 4: **Sampling probabilities by OR and MAF** x-axis: \( N \cdot f \). Left panel y-axis: ratio of sampling probability for a SNP at the specified MAF with OR given by the label and the inverse of that OR. Right panel: difference in absolute sampling probabilities. 100 individuals each group, disease prevalence 1%.

The tables recapitulate the claims above; because of the shifted mean of novel versus previously observed SNPs (as in Figure 5), the correct prediction for previously observed SNPs is that they are much more risk increasing than novel SNPs, which tend to be protective. As the mean or variance of SNP effects increases, the difference becomes exaggerated (Tables S7 and S6). With MAFs which are small compared to the sample size, previously observed SNPs are more risk-increasing and at larger MAFs novel SNPs are very protective (Table S8 and S10). The sample size determines the MAF at which the effect is most prevalent, but not the magnitude of the effect (Table S10 and S9).
Figure 5: Expected log-odds ratios given previously sampled (left) or novel in follow-up (right) and MAF. x-axis: $N \cdot f$. y-axis: mean lOR. Source log-odds ratio distributed $N(0, 0.6^2)$ (black), Student’s t distributions with 1, 2, 3 degrees of freedom (green, red, blue) truncated at ±4 with same 20% and 80% quantile as $N(0, 0.6^2)$. 100 individuals each group; disease prevalence 1%.

Figure 6: Standard deviation of log-odds ratios given previously sampled (left) or novel in follow-up (right) and true minor allele frequency. Settings identical to Figure 5.
4 Discussion

We have two main points to discuss. First is the differing interpretation of population average models and tests based on those models (such as allele-count and weighted burden tests) which collapse genotype vectors into a summary statistic versus SNP specific models which attempt to explicitly model the effect of every SNP on phenotype. In Figures 1 and 3 we have provided an example in which an investigator could correctly claim than possessing more minor alleles increased risk of disease, but not that minor alleles on average increase risk. Even where the mean SNP IOR is zero, we show a non-trivial power for allele-count tests. Heuristically, there are more controls who are close to becoming cases than there are cases close to becoming controls after a small change in their risk factors. Therefore, a collection of SNPs will tend to materialize a net increase in risk. In the terms of the causal inference literature, SNPs are not “collapsible” over individual genotypes [8]. These two models ask different questions, and the primary confusion which we have encountered (and hope to have dispelled) is accidentally interpreting association tests based on the PA model as if they were based on the SNP specific model. It has often been observed that though a SNP specific model is richer and answers additional questions, the parameters have an awkward extrapolation to observed data. This distinction originated in biometrics modeling longitudinal data [9], but has been reiterated in many contexts [10, 11, 12, 13]. The failure of Equation (4) at very low prevalence displayed in Figure S1 is itself interesting, since it points to observable different behavior when the data come from a logistic versus probit (latent Gaussian) model; these two are usually difficult to distinguish in practice.

Second, among SNPs with MAF near 1/2N, case-control studies tend to selectively ascertain SNPs which are risk-increasing. Because risk-increasing minor alleles have an increased frequency among affecteds compared to their population MAF, they are more likely than SNPs which decrease risk to be observed more than zero times in the study. The relevance of this aspect of the study design is that in external data previously detected SNPs will increase risk more than average SNPs, and novel SNPs will increase risk less or even decrease it. We illustrated this as a problem of prediction, where log-odds ratios for novel SNPs are substantially lower than old SNPs. However, there are several other implications. As we showed, the affected set of SNPs changes depending on the sample size, and therefore predictions based on experiments of different size will appear inconsistent. If, as is common practice, small costly sequencing projects are replicated with larger custom genotyping projects, the replicated effect will tend to be larger than that initially observed and not representative of SNPs in that gene. Other study designs which depend on the effect sizes of the SNPs in question, such as family-based replication and studies of correlated phenotypes, may also be affected. We note that under the strict null hypothesis that no SNPs in a gene are causal, the bias is zero and therefore gene-based association tests maintain correct calibration.

An alternative way to understand the same phenomenon is to note that given its true MAF, observing a SNP zero times in cases and controls is weakly informative on its odds ratio. Consider a SNP known to exist at frequency 1% in the general population observed zero times in a study of 100 cases and 100 controls; this data excludes large odds ratios which would have made it likely to be sampled in cases. If a-priori these odds ratios were possible, then a-posteriori the expected odds ratio is necessarily shifted downward. Omitting all these weakly-shifted pieces of information will then create a bias when estimating the gene-level
average relative to having known and included the full set of zero-count SNPs.

We have not presented simulations or calculations where MAFs are drawn from a con-
tinuous distribution as they would be in real data. Because the ascertainment bias varies
with MAF, we felt this added an unnecessary level of complication to understanding the
phenomenon. Approximate results for a mixture of MAFs can be obtained by a weighted
average of the results for point distributions of MAF. Uncommon SNPs (especially single-
tons) observed in a real sequencing study are a mixture of those with MAF near 1/2N and
lucky representatives of the large pool of rarer SNPs. Fortunately, as seen in Figures 5 and
the relative bias stabilizes as the MAF becomes small. The impact of MAF is, however,
important to understand as comparisons across genes with different SNP frequency spectra
will experience differential bias. Similarly, there will be differential bias when comparing
human populations with different demographic history and therefore frequency distributions

A criticism that one might level at our results is that the calculations only produce
substantial bias in the presence of fairly strong SNP effects, which might be unrealistic.
We reply that the supposed existence of rare SNPs with large effects is a major rationale
for performing sequencing-based association studies [15], and that the failure to detect such
variants in GWAS or linkage studies does not exclude their existence. We suspect that a
distribution like the truncated Student’s t used in Figure 5 whose distribution is plotted in
Figure S6 represents a reasonable belief about log-odds ratios; most are very nearly zero with
a few outlying strong effects.

We are aware of a literature on ascertainment-adjustment to variance components analysis
of family based data, which stems from an identical problem [16, 17, 18, 19, 20, 21, 22,
23, 24]. In these cases the relatively simple structure of ascertainment makes correction
feasible. Similarly, SNP-panel and small sequencing study based inferences on site frequency
spectra and other population genetic quantities have been examined for ascertainment bias
[25, 26, 27]. Calculating the bias for known distributions of MAF and SNP effects is trivial;
however, applying this correction in practice is challenging for several reasons, and we have
therefore saved detailed discussion for future work. First, as we showed in Figure 5 the result
depends on the tails of the distribution of IORs, which can be difficult to robustly estimate.
Second, the observed data are minimally informative for estimation of the MAF spectrum
around 1/2N and especially for extrapolation to lower MAF. External information in the
form of demographic models and previously sequenced cohorts of the same ethnicity will be
crucial to developing accurate estimates and can not be spoken to in much generality. Third,
the joint distribution of MAF and SNP IORs matters. We have performed calculations only
where the two are independent, but in reality a more complex procedure is required which
estimates the distribution of SNP effects for low MAFs and even extrapolates to very low
MAFs.

Acknowledgments

We are grateful to Nancy Cox and Lin Chen for comments on a draft of the paper. We would
like to thank Jonathan Pritchard for feedback on the project.
References

1. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, et al. (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature Communications 1: 131.

2. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science (New York, NY) 336: 740–743.

3. Longmate JA, Larson GP, Krontiris TG, Sommer SS (2010) Three ways of combining genotyping and resequencing in case-control association studies. PLoS ONE 5: e14318.

4. Curtin K, Iles MM, Camp NJ (2009) Identifying rarer genetic variants for common complex diseases: diseased versus neutral discovery panels. Annals of human genetics 73: 54–60.

5. Edwards TL, Song Z, Li C (2011) Enriching targeted sequencing experiments for rare disease alleles. Bioinformatics 27: 2112–2118.

6. Li B, Leal SM (2009) Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. PLoS Genet 5: e1000481.

7. Yang F, Thomas DC (2011) Two-stage design of sequencing studies for testing association with rare variants. Human Heredity 71: 209–220.

8. Greenland S (1999) Confounding and collapsibility in causal inference. Statistical Science 14: 29–46.

9. Zeger SL, Liang K, Albert PS (1988) Models for longitudinal data: A generalized estimating equation approach. Biometrics 44: 1049–1060.

10. Crouchley R, Davies RB (1999) A comparison of population average and random-effect models for the analysis of longitudinal count data with base-line information. Journal of the Royal Statistical Society Series A (Statistics in Society) 162: 331–347.

11. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA (1998) Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. American Journal of Epidemiology 147: 694 –703.

12. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, et al. (2010) To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology (Cambridge, Mass) 21: 467–474.

13. Subramanian SV, O’Malley AJ (2010) Modeling neighborhood effects: the futility of comparing mixed and marginal approaches. Epidemiology (Cambridge, Mass) 21: 475–478; discussion 479-481.
14. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5: e1000695.

15. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.

16. Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, et al. (2000) Ascertainment adjustment: Where does it take us? American Journal of Human Genetics 67: 1505–1514.

17. Glidden DV, Liang K (2002) Ascertainment adjustment in complex diseases. Genetic Epidemiology 23: 201–208.

18. Bowden J, Thompson JR, Burton PR (2007) A two-stage approach to the correction of ascertainment bias in complex genetic studies involving variance components. Annals of Human Genetics 71: 220–229.

19. Ma J, Amos CI, Warwick Daw E (2007) Ascertainment correction for Markov chain Monte Carlo segregation and linkage analysis of a quantitative trait. Genetic Epidemiology 31: 594–604.

20. Noh M, Lee Y, Pawitan Y (2005) Robust ascertainment-adjusted parameter estimation. Genetic Epidemiology 29: 68–75.

21. Epstein M, Lin X, Boehnke M (2002) Ascertainment-adjusted parameter estimates revisited. The American Journal of Human Genetics 70: 886–895.

22. Sung YJ, Dawson G, Munson J, Estes A, Schellenberg GD, et al. (2005) Genetic investigation of quantitative traits related to autism: Use of multivariate polygenic models with ascertainment adjustment. American Journal of Human Genetics 76: 68–81.

23. Burton PR, Scurrah KJ, Tobin MD, Palmer LJ (2005) Covariance components models for longitudinal family data. International Journal of Epidemiology 34: 1063–1077.

24. Burton PR (2003) Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using Gibbs sampling. Genetic Epidemiology 24: 24–35.

25. Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168: 2373–2382.

26. Nielsen R, Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theoretical Population Biology 63: 245–255.

27. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Research 15: 1496–1502.
Supplement

Figure S1: **Comparison of formulas (4) and (3)**. Disease prevalence among those with no minor alleles (x-axis) versus predictive lOR from equation (3) (dotted) and formula (4) (solid). lOR of 1 SNP versus 0 SNPs. SD of SNP lORs 0.6, mean of lORs 0, MAF 0.005. Low limit of plot = 0.001.

4.1 Simulation settings

For Figure S3 we simulated 40 SNPs with minor allele frequencies drawn from a beta distribution mean .12 standard deviation .02 in a population of 1 million individuals with a baseline disease prevalence of 5%, drawing a sample of 500 cases 500 controls for each of the depicted values of the standard deviation of log-odds ratios. Log odds-ratios were always mean zero iid normal. We then conducted a t-test of the number of minor alleles in cases and controls; essentially identical results were obtained for rank-tests and generalized linear models. For Figure S4 we used the same parameters for SNPs, fixed the standard deviation of log-odds ratios at .6 and varied the baseline prevalence from .01 to .5. For Figure S5 we fixed the log-odds ratio standard deviation at 1, fixed the baseline prevalence at 5% and varied the number of SNPs from 10 to 100 keeping the distribution of minor allele frequencies the same. 2000 replicates for each point in each figure were produced, each repeating the entire procedure.

All tables use a Gaussian distribution for lORs and a large sample prospective \((10^6)\) generated by the logistic model (1) with all SNPs uncorrelated. 2000 replicates are drawn for each Table. Table S1 varies the disease prevalence in rows and holds the SD of SNP effects.
at 0.6, the mean SNP effect at 0, the MAF at 0.01, and the number of SNPs at 50. Table S2 varies the SD of SNP effects in rows and holds the the mean SNP effect at 0, the disease prevalence at 5%, the MAF at 0.01, and the number of SNPs at 50. Table S3 varies the mean of SNP effects in rows and holds the SD of SNP effects at 0.7, the disease prevalence at 5%, the MAF at 0.01, and the number of SNPs at 50. Table S4 varies the SNP MAF in rows and holds the SD of SNP effects at 0.7, the mean SNP effect at 0, the disease prevalence at 5%, and the number of SNPs at 50. Tables for retrospective simulation draw 100 cases and 100 controls from a source population of $10^6$ and compare the results in $10^6$ future independent individuals. We display only the contrast for having no minor alleles versus 1 minor allele for all alleles combined, novel in follow-up, and previously discovered SNPs; patterns in the impact of the number of alleles are similar to the prospective case (not shown). Table S5 varies the disease prevalence in rows and holds the SD of SNP effects at 0.6, the mean SNP effect at 0, the MAF at 0.01, and the number of SNPs at 50. Table S6 varies the SD of SNP effects in rows and holds the the mean SNP effect at 0, the disease prevalence at 5%, the MAF at 0.01, and the number of SNPs at 50. Table S7 varies the mean of SNP effects in rows and holds the SD of SNP effects at 0.7, the disease prevalence at 5%, the MAF at 0.01, and the number of SNPs at 50. Table S8 varies the SNP MAF in rows and holds the SD of SNP effects at 0.7, the mean SNP effect at 0, the disease prevalence at 5%, and the number of SNPs at 50. Table S9 jointly varies the number of SNPs and the MAF to keep the average number of minor alleles per person at 0.5; it holds the SD of SNP effects at 0.7, the mean SNP effect at 0, and the disease prevalence at 5%. Table S10 varies the case-control sample size in rows and holds the SD of SNP effects at 0.7, the mean SNP effect at 0, the disease prevalence at 5%, the MAF at 0.01, and the number of SNPs at 50. Table S11 jointly varies the MAF and the sample size in rows to keep the expected number of appearances for each SNP constant at 0.5; it holds the SD of SNP effects at 0.7, the mean SNP effect at 0, the disease prevalence at 5%, and the number of SNPs at 50.

|    | 1     | 2     | 3     |
|----|-------|-------|-------|
| 0.2| 0.09  | 0.09  | 0.16  | 0.17  | 0.22  | 0.23  |
| 0.1| 0.14  | 0.13  | 0.26  | 0.24  | 0.35  | 0.33  |
| 0.05| 0.19  | 0.15  | 0.34  | 0.28  | 0.48  | 0.40  |
| 0.01| 0.29  | 0.18  | 0.54  | 0.33  | 0.74  | 0.48  |

Table S1: IOR of outcome by number of derived alleles from formula (4) and simulation varying prevalence of disease using prospective sampling.
Table S2: lOR of outcome by number of derived alleles from formula (4) and simulation varying SD of SNP effects using prospective sampling

|      | 1    | 2    | 3    |
|------|------|------|------|
| 0.25 | 0.04 | 0.03 | 0.07 | 0.05 | 0.10 | 0.07 |
| 0.5  | 0.13 | 0.10 | 0.25 | 0.20 | 0.35 | 0.28 |
| 0.75 | 0.28 | 0.23 | 0.49 | 0.42 | 0.66 | 0.57 |
| 1    | 0.45 | 0.38 | 0.74 | 0.65 | 0.95 | 0.85 |
| 1.25 | 0.62 | 0.55 | 0.97 | 0.89 | 1.19 | 1.11 |

Table S3: lOR of outcome by number of derived alleles from formula (4) and simulation varying mean of SNP effects using prospective sampling

|      | 1    | 2    | 3    |
|------|------|------|------|
| -0.47| -0.18| -0.25| -0.36| -0.52| -0.53| -0.79|
| 0    | 0.25 | 0.21 | 0.44 | 0.38 | 0.60 | 0.52 |
| 0.26 | 0.49 | 0.46 | 0.89 | 0.86 | 1.22 | 1.21 |
| 0.47 | 0.68 | 0.66 | 1.24 | 1.24 | 1.72 | 1.74 |
| 0.69 | 0.88 | 0.87 | 1.62 | 1.62 | 2.26 | 2.25 |
| 0.92 | 1.09 | 1.08 | 2.00 | 2.01 | 2.79 | 2.80 |

Figure S2: **Effect of increasing sample size on ascertainment bias.** Identical to right panel of figure 4 with new lines the result with use 1000 cases and controls instead of 100.
Table S4: lOR of outcome by number of derived alleles from formula (4) and simulation varying SNP MAF using prospective sampling

|    | 1    | 2    | 3    |
|----|------|------|------|
| 0.005 | 0.25 | 0.20 | 0.44 | 0.37 | 0.60 | 0.51 |
| 0.01  | 0.25 | 0.19 | 0.44 | 0.36 | 0.60 | 0.49 |
| 0.02  | 0.25 | 0.21 | 0.44 | 0.38 | 0.60 | 0.52 |
| 0.03  | 0.25 | 0.20 | 0.44 | 0.37 | 0.60 | 0.51 |
| 0.05  | 0.25 | 0.20 | 0.44 | 0.37 | 0.60 | 0.51 |

Figure S3: Standard deviation of log-odds ratios versus simulated power of allele count test when the mean log-odds ratio is 0.

Table S5: lOR of outcome by SNP class for first allele from formula (4) and simulation varying prevalence of disease using case-control sampling

|    | all   | novel | old   |
|----|-------|-------|-------|
| 0.2 | 0.10  | 0.10  | -0.01 | -0.00 | 0.15 | 0.16 |
| 0.1 | 0.13  | 0.14  | -0.02 | -0.00 | 0.21 | 0.22 |
| 0.05 | 0.16 | 0.16  | -0.02 | -0.00 | 0.24 | 0.25 |
| 0.01 | 0.17  | 0.18  | -0.03 | -0.00 | 0.27 | 0.28 |
| 0.001 | 0.18 | 0.19  | -0.03 | -0.00 | 0.27 | 0.29 |
| 1e-04 | 0.18 | 0.19  | -0.03 | -0.00 | 0.27 | 0.29 |
Figure S4: Baseline disease prevalence versus simulated power of allele count test.

Figure S5: Number of SNPs in causal group versus simulated power of allele count test.
Figure S6: **Cumulative distribution plots for SNP log-odds ratios used in Figure 5.**

N(0,6^2) (black), Student’s t distributions with 1, 2, 3 degrees of freedom (green, red, blue) with same inner 80% quantile as N(0,6^2).

| all | novel | old |
|-----|-------|-----|
| 0.25 | 0.03 | 0.02 |
| 0.5  | 0.11 | -0.11 |
| 0.75 | 0.24 | -0.21 |
| 1    | 0.40 | -0.36 |
| 1.25 | 0.58 | -0.49 |

**Table S6:** IOR of outcome by SNP class for first allele from formula (4) and simulation varying SD of SNP effects using case-control sampling

| all | novel | old |
|-----|-------|-----|
| -0.47 | -0.25 | -0.56 |
| 0   | 0.21 | -0.20 |
| 0.26 | 0.46 | -0.01 |
| 0.47 | 0.66 | 0.13 |
| 0.69 | 0.87 | 0.28 |
| 0.92 | 1.09 | 0.43 |

**Table S7:** IOR of outcome by SNP class for first allele from formula (4) and simulation varying mean of SNP effects using case-control sampling
### Table S8: IOR of outcome by SNP class for first allele from formula \(^\text{[4]}\) and simulation varying SNP MAF using case-control sampling

|     | all     | novel   | old     |
|-----|---------|---------|---------|
| 0.005 | 0.21 0.18 | -0.03 -0.02 | 0.32 0.28 |
| 0.01 | 0.21 0.18 | -0.20 -0.16 | 0.26 0.23 |
| 0.02 | 0.21 0.18 | -0.44 -0.32 | 0.22 0.19 |
| 0.03 | 0.21 0.18 | -0.62 -0.49 | 0.21 0.18 |
| 0.05 | 0.21 0.18 | -0.88 -0.75 | 0.20 0.18 |

### Table S9: IOR of outcome by SNP class for first allele from formula \(^\text{[4]}\) and simulation varying MAF and number of SNPs with constant expected count per person using case-control sampling

|     | all     | novel   | old     |
|-----|---------|---------|---------|
| 0.005 | 0.21 0.19 | -0.03 -0.02 | 0.32 0.29 |
| 0.01 | 0.21 0.18 | -0.20 -0.16 | 0.26 0.23 |
| 0.02 | 0.21 0.16 | -0.44 -0.31 | 0.22 0.17 |
| 0.03 | 0.21 0.15 | -0.62 -0.46 | 0.21 0.15 |
| 0.05 | 0.21 0.12 | -0.88 -0.65 | 0.20 0.12 |

### Table S10: IOR of outcome by SNP class for first allele from formula \(^\text{[4]}\) and simulation varying sample size using case-control sampling

|     | all     | novel   | old     |
|-----|---------|---------|---------|
| 50  | 0.21 0.20 | 0.07 0.08 | 0.37 0.32 |
| 100 | 0.21 0.20 | -0.03 -0.02 | 0.32 0.29 |
| 200 | 0.21 0.20 | -0.20 -0.17 | 0.26 0.25 |
| 500 | 0.21 0.20 | -0.54 -0.43 | 0.21 0.20 |
| 1000| 0.21 0.20 | -0.87 -0.74 | 0.21 0.20 |

### Table S11: IOR of outcome by SNP class for first allele from formula \(^\text{[4]}\) and simulation varying MAF and sample size with constant expected count per SNP using case-control sampling

|     | all     | novel   | old     |
|-----|---------|---------|---------|
| 0.005 | 0.21 0.18 | -0.03 -0.02 | 0.32 0.28 |
| 0.01 | 0.21 0.18 | -0.03 -0.02 | 0.32 0.27 |
| 0.02 | 0.21 0.18 | -0.03 -0.02 | 0.31 0.27 |
| 0.03 | 0.21 0.18 | -0.03 -0.03 | 0.31 0.26 |
| 0.05 | 0.21 0.18 | -0.03 -0.02 | 0.31 0.25 |