Statistics and Its Interface Volume 4 (2011) 285–294

Using ascertainment for targeted resequencing to increase power to identify causal variants

M. D. Swartz*, B. Peng, C. Reyes-Gibby and S. Shete

Researchers continue to use genome-wide association studies (GWAS) to find the genetic markers associated with disease. Recent studies have added to the typical two-stage analysis a third stage that uses targeted resequencing on a randomly selected subset of the cases to detect the causal single-nucleotide polymorphism (SNP). We propose a design for targeted resequencing that increases the power to detect the causal variant. The design features an ascertainment scheme wherein only those cases with the presence of a risk allele are selected for targeted resequencing. We simulated a disease with a single causal SNP to evaluate our method versus a targeted resequencing design using randomly selected individuals. The simulation studies showed that ascertaining individuals for the targeted resequencing can substantially increase the power to detect a causal SNP, without increasing the false-positive rate.

Keywords and phrases: Ascertainment, Genome-wide association study, Causal polymorphism, Targeted resequencing.

1. INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified polymorphisms associated with complex diseases such as lung cancer (Amos, et al. 2008, Hung, et al. 2008, Thorgeirsson, et al. 2008), prostate cancer (Eeles, et al. 2008, Thomas, et al. 2008), glioma (Shete, et al. 2009), and type II diabetes (Sladek, et al. 2007) and complex traits such as body mass index (BMI) (Frayling, et al. 2007). Often, a large number of subjects is required to achieve adequate power in a GWAS because the odds ratio of a given single-nucleotide polymorphism (SNP) associated with a complex disease can be small (Eberle, et al. 2007, Hirschhorn and Daly 2005, Marchini, Donnelly and Cardon 2005).

Typically, for a GWAS, researchers collect a large sample of individuals and divide the sample into two groups for a two-stage analysis. For stage 1, the investigator genotypes many markers spread across the genome using a contemporary commodity array of SNPs, such as Illumina's Human Hap 550 and Human 1M Duo chips or Affymetrix SNP array 6.0 (Satagopan and Elston 2003, Wellcome Trust Case Control Consortium 2007). Then, the investigator analyzes the data using a statistical test of choice and uses some screening criteria to select SNPs likely to be associated with the disease. For stage 2, the investigator selects the top n SNPs identified by the smallest p-values for association from stage 1 (Satagopan and Elston 2003). Originally, stage 1 and stage 2 were analyzed separately, with stage 2 treated as an independent replication set for the findings in stage 1, but Skol et al. showed that, under certain assumptions, it is more powerful to pool the information from stage 1 and stage 2 (Skol, et al. 2006, Skol, et al. 2007).

Recently, researchers have added a type of third stage to GWAS: for the set of SNPs that reach genome-wide significance, they sequence regions surrounding those SNPs and compare genotype distributions or allele frequencies in the cases with the distributions or frequencies computed from public sources, such as HapMap data or the 1000 Genome Project (Yamada, et al. 2009). This method, known as targeted resequencing, allows for the analysis of denser SNPs to better locate the causal variant.

Commonly, the true causal SNP can have lower minor allele frequency and, therefore, be somewhat sparsely represented in a randomly collected sample. Unless the SNP has complete penetrance for a disease, a SNP with a low minor allele frequency could have a lower occurrence among cases. Therefore, in a targeted sequencing analysis, even when sampling from cases only, the probability of sequencing the causal SNP is very low, and this probability can be increased through proper ascertainment.

Here, we propose a simple way to increase the probability of including the causal SNP in the sample selected for targeted resequencing and, as a result, improve the power of the analysis. The two-stage analysis remains the same. However, instead of randomly selecting a subset of the cases and then performing targeted resequencing analysis, as in (Yamada, et al. 2009), we randomly selected individuals from the cases carrying the minor alleles of SNPs achieving genome-wide significance from the GWAS. The rationale for this type of ascertainment is to increase the probability that the resequencing sample will include the causal allele. By definition of linkage disequilibrium (LD), the SNPs in strong LD would have similar frequencies, and thus the minor allele of the tagging SNP would be most likely in LD with the true rare causal allele. In Appendix A, we show that in the presence of LD between the tagging SNP and the causal SNP

*Corresponding author.

the probability of the causal sequence being contained in the sample increases if one uses ascertainment based on the minor allele of a SNP known to be tagging the causal allele. The SNPs identified by the two stage GWAS will most likely have the strongest association with the causal SNP. We call this method "ascertained targeted resequencing" because we ascertain samples based on the presence of the minor alleles at those SNPs detected by the two-stage GWAS. We analyzed simulated data and showed that ascertaining a sample based on the presence of an allele found to be significant in a two-stage GWAS does increase the power to detect the causal SNP using targeted resequencing.

2. METHODS

We investigated the usefulness of the ascertained targeted resequencing design by using simulation studies. We simulated 100 replicates of GWAS data for a disease generated by one disease locus. The data were simulated using a simuPOP (Peng and Kimmel 2005) script that extends the Hap-Sample method proposed in (Wright, et al. 2007). This method essentially resamples existing HapMap sequences using simulated recombination events. If a single-locus disease model is specified, it simulates genotypes at the disease susceptibility locus of cases and controls using Pr(genotype | affection status) before genotypes at other loci are simulated. The simulated datasets were validated according to their resemblance to the original HapMap dataset in terms of marker allele frequency, observed heterozygosity, Hardy-Weinberg deviation, and decay of linkage disequilibrium as a function of marker distance.

Using the simuPOP script, we simulated a total of 2000 cases and 2000 controls for each replicate. We used HapMap SNPs (Phase II data) from a 4.4 Mb region of chromosome 2. We simulated our genetic disease from a single SNP, as many GWAS have found only one SNP (Amos, et al. 2008, Hung, et al. 2008, Thorgeirsson, et al. 2008). To avoid overpowering the study, we simulated an odds ratio of 1.8 for the risk allele in the single-locus model. The SNP selected to be the causal SNP has a minor allele frequency on the order of 0.20. For each replicate, we used 1000 cases and 1000 controls for each stage. Since it is rare to have the causal SNP in stage 1 of a GWAS, the SNP we simulated as causal was not included in the Illumina Infinium Human Hap550 SNP chip set. However, when we simulated our disease, we generated two sets of replicates where the LD (measured by r^2) between the causal SNP and at least one marker on the HumanHap550 chip varied from 0.8 to .95, as described in more detail later.

For the stage 1 analysis, we used the SNPs from the HumanHap550 chip along chromosome 2. We conducted univariate logistic regression analyses to test for the association of each SNP with the disease of interest. For stage 2, we followed up the top 30 SNPs from stage 1 and ran an independent univariate logistic regression analysis on an additional 1000 cases and 1000 controls. (We chose 30 SNPs because that is on order with the numbers followed up in recently published GWA studies (Alberts 2002, Hung, et al. 2008, Shete, et al. 2009, Sladek, et al. 2007, Thorgeirsson, et al. 2008, Wu, et al. 2009). We used Fisher's method for meta analyses (Fisher 1932) to combine the stage 1 and stage 2 p-values, selecting SNPs with genome-wide significance (p-values less than 10^{-7}).

We performed the targeted resequencing analysis in two different ways: with and without ascertainment. We used the following ascertainment procedure. The SNPs with p-values less than 10^{-7} were denoted as the risk-associated SNPs for the purpose of ascertainment, and their minor alleles were denoted as the risk-associated alleles. Then, from the subset of cases carrying the risk-associated alleles, we ascertained 96 individuals for targeted resequencing. For the targeted resequencing analysis without ascertainment, which was considered the standard method, we randomly selected 96 cases. For each set of cases (with and without ascertainment), we selected SNPs from within a 300-kb window around each SNP that passed genome-wide significance. For each SNP found, we compared allele frequencies with those from the HapMap dataset using the normal approximation test for proportions, as in (Yamada, et al. 2009). Then for each set we examined the distance, in base pairs, between the top-ranking SNP from the resequencing analysis and the true disease SNP. We also compared the rank of the p-value of the causal SNP in the ascertainment set with its rank in the standard targeted resequencing set.

To show the effect of varying LD on our ascertainment method, we analyzed two different simulated populations. The first population of replicates was simulated with the disease SNP being in LD with a SNP on the Hap550 chip on the order of 0.95. The second population of replicates was simulated with a disease SNP in LD with SNPs on the Hap550 chip on the order of 0.80. We refer to the first analysis as the "high-LD analysis," since it included one marker SNP in tight LD with the causal SNP ($r^2 \approx 0.95$), and to the second analysis as the "moderate-LD analysis." We present a picture of the LD pattern of the stage 2 SNPs in Fig. 1. We performed both analyses with and without ascertainment.

To fully evaluate the ascertained targeted resequencing, we also simulated 100 null sets, using a method similar to that described in (Swartz, Yu and Shete 2008), such that there is no association between the SNPs and disease. This method disrupts the disease-gene association but on average maintains the case-control ratio. We then followed the same analysis protocol as above, analyzing the data using the same targeted resequencing analysis (with and without ascertainment) after two-stage GWAS.

3. RESULTS

3.1 High-LD analysis

In the high-LD analysis, 90 of 100 simulated genetic disease replicates had SNPs that were significant at the

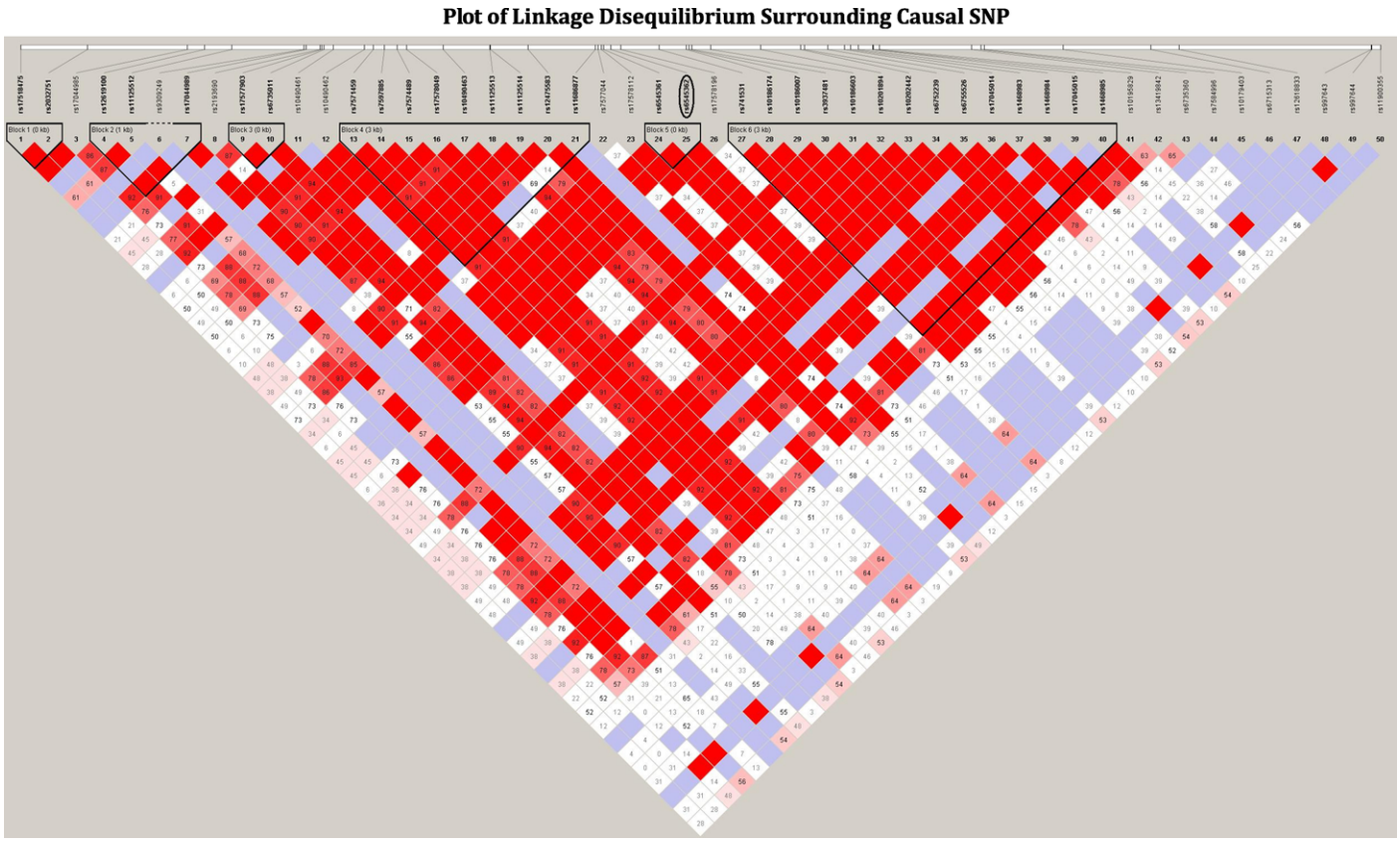

Figure 1. Plot of the Linkage Disequilibrium (LD) pattern in the 50 SNPs surrounding the causal SNP (circled) in the stage 2 data, computed from Haploview.

genome-wide level. In these 90 replicates, we compared the differences in rankings of the top-ranked SNP and the true SNP between the ascertained and the standard targeted sequence analyses. We found that with the non-ascertained data, the causal SNP was ranked as low as 536th among the SNPs investigated, whereas with ascertainment, the lowest ranking was 34th. On average, the ranking of the p-values for the causal SNP was much higher with ascertainment (mean rank = 2.68, standard deviation (SD) = 4.26) than without ascertainment (mean rank = 40.8, SD = 78). Figure 2 plots the ranks of the p-values for the causal SNPs comparing the allele frequencies with HapMap frequencies across the 90 replicates that underwent targeted resequencing analysis. Note that of the non-ascertained replicates, roughly half of them ranked the true SNP outside of the 10 most significant SNPs, while almost all of the ascertained targeted sequencing replicates (89 of the 90) ranked the true SNP within the top 10 most significant SNPs. Also note the rankings were much more widely dispersed among the replicates using non-ascertainment than among those using ascertainment. Furthermore, with ascertainment, the most significant SNP was typically closer to the true SNP. On average, the p-values comparing the allele frequencies from the targeted resequencing analysis with ascertainment (mean = 0.000074, SD = 0.00071) were much lower than the p-values from the targeted resequencing analysis without ascertainment (mean = 0.1065, SD = 0.132). Additionally, with ascertained targeted resequencing, 50% of the iterations reported the true SNP with a p-value less than 10^{-20}, while the standard method reported p-values for the true SNP greater than 10^{-5} across all replicates.

Likewise, we compared the distance (in base pairs) between the top-ranking SNP and the causal SNP. Figure 3 shows these distances and how they varied across the 90 replicates that warranted targeted resequencing. Note that the most significant SNPs were located much closer on the chromosome to the causal SNP when ascertainment was used. The bottom line is that ascertaining the targeted resequencing sample resulted in higher-ranking p-values for the true SNP and in the detection of SNPs closer to the true SNP location than did non-ascertainment targeted resequencing.

3.2 Moderate-LD analysis

In the case of reduced LD between the causal SNP and the markers, 87 simulated replicates had SNPs significant at or beyond the genome-wide significance level ($<10^{-7}$). In the presence of lower LD, the spread of the difference in

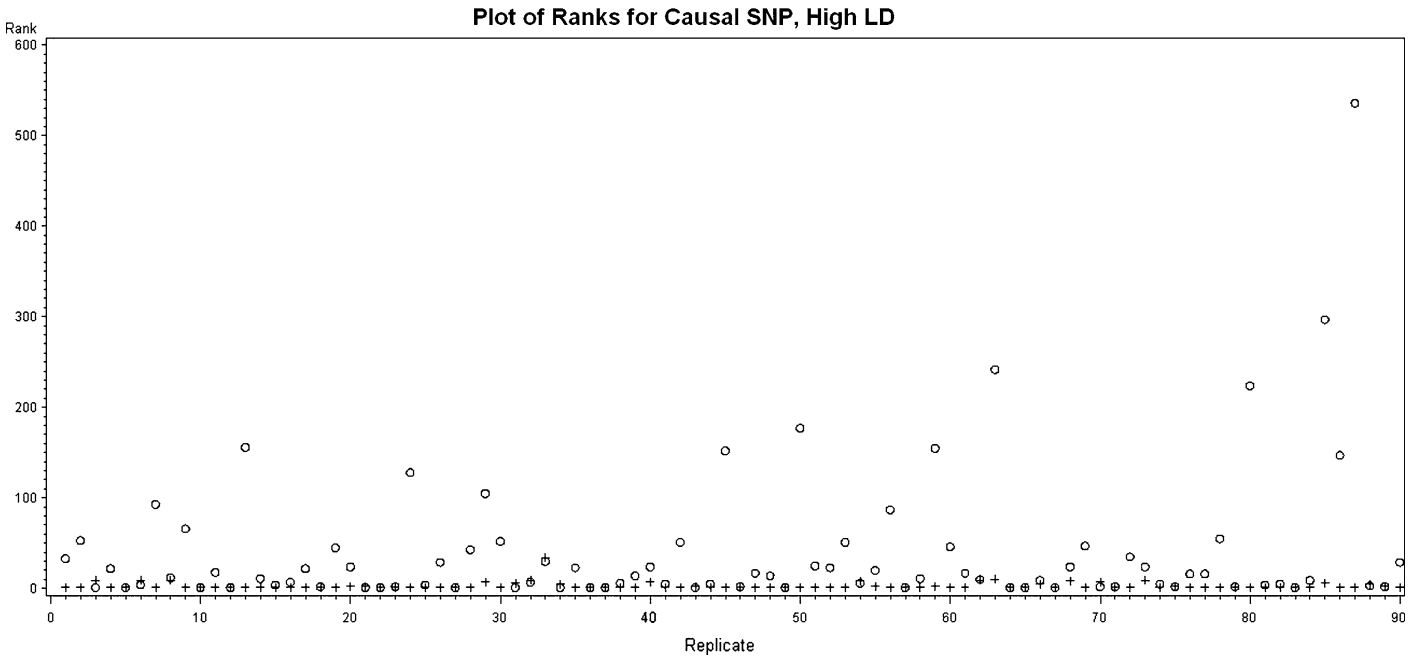

Figure 2. This plot shows the differences in ranking of the causal SNP relative to the top-ranked SNP (top ranked minus causal) for the ascertained targeted sequencing (+) and the non-ascertained targeted sequencing (O), when analyzing the higher-LD replicates.

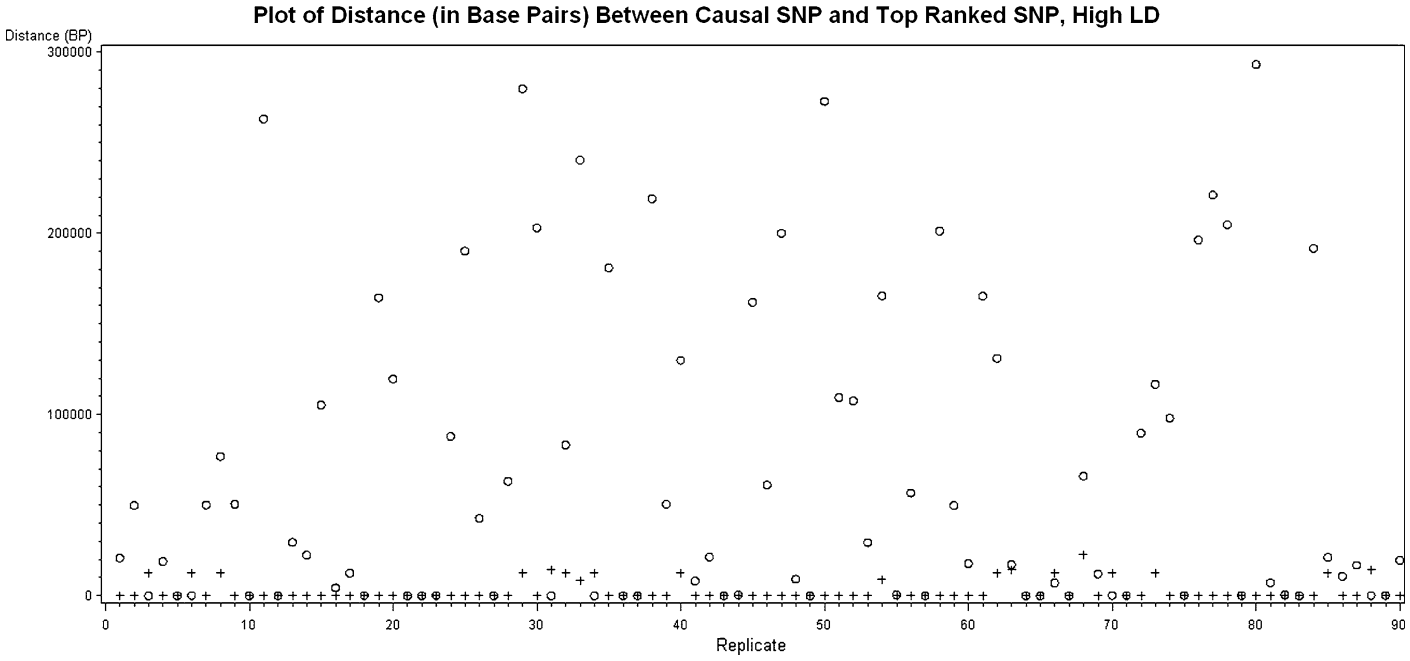

Figure 3. Plot of distances from the most significant SNP to the causal SNP. This plot shows the number of base pairs from the top-ranked SNP to the causal SNP for each replicate (iteration) of the design, comparing the ascertained targeted resequencing analysis (+) to the standard resequencing analysis (O) for higher LD.

rank of the true SNP was larger, but still, in the majority of replicates using ascertainment, the true SNP was ranked higher than in those under non-ascertainment. Since the LD was lower for both methods, we saw fewer replicates ranking the true SNP within the top 20, let alone the top 10, yet the true SNP was more often among the 20 most significant

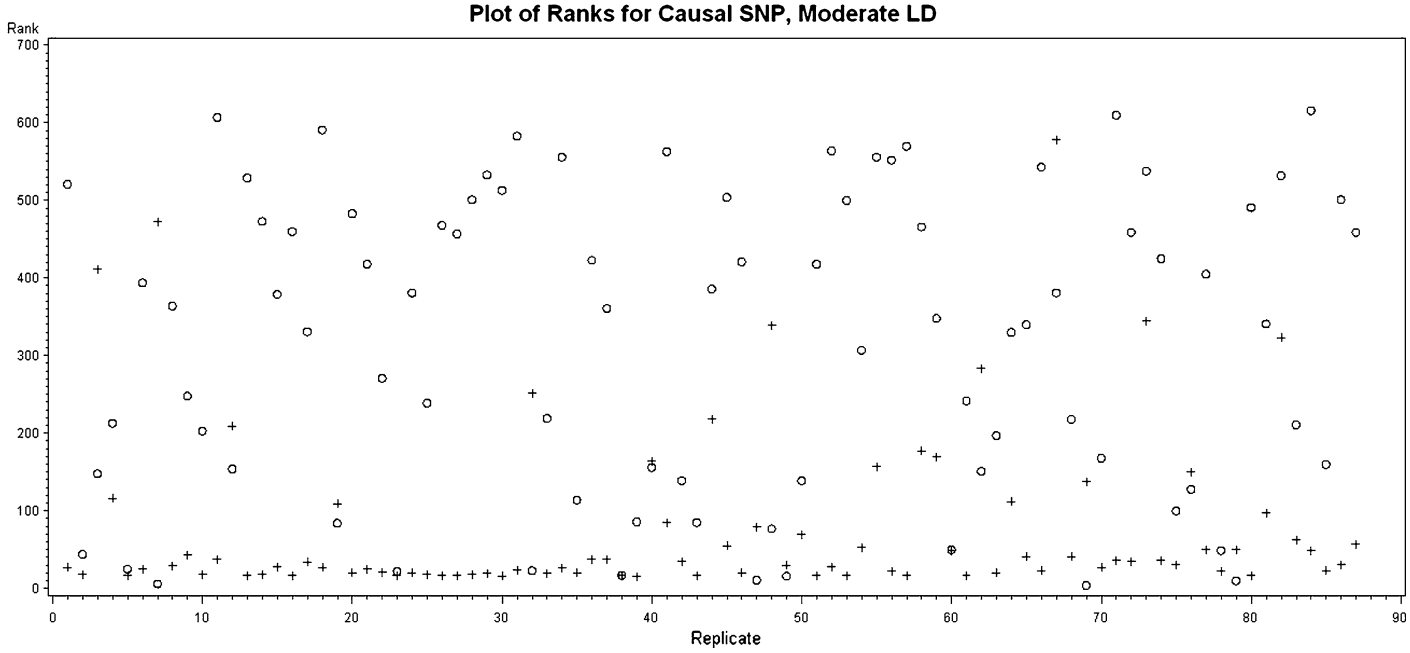

Figure 4. This plot shows the differences in ranking of the causal SNP relative to the top-ranked SNP (top ranked minus causal) for the ascertained targeted sequencing (+) and the non-ascertained targeted sequencing (O), when analyzing the lower-LD replicates.

SNPs when ascertained targeted resequencing was used. On average, using ascertainment for targeted resequencing samples led to better ranking of the causal SNP (mean rank = 78.6, SD = 109.9) than non-ascertained targeted resequencing (mean rank = 320, SD = 192.9).

Figure 4 plots the p-value rankings for the causal SNP computed by comparing the sample allele frequency with that from HapMap for ascertained and non-ascertained targeted resequencing samples across the 87 replicates that warranted targeted resequencing analysis. With ascertainment, 28 replicates reported the true SNP within the 20 most significant SNPs, while without ascertainment, only 6 replicates reported the true SNP within the 20 most significant SNPs. Figure 5 plots the distances between the SNP with the smallest computed p-value and the causal SNP for all 87 replicates using both ascertained and non-ascertained targeted resequencing. The distances from the SNP with the smallest p-value to the true SNP across replicates were more widely dispersed than in the higher-LD analysis, but still the distances under ascertainment were typically smaller than the distances under non-ascertainment. On average, the ascertained targeted resequencing continued to report lower p-values (mean = 0.09, SD = 0.17) than the standard targeted resequencing (mean = 0.51, SD = 0.24) under lower LD. Thus, even with lower LD, similar to that typical of tagging SNPs, using ascertainment for targeted resequencing generally resulted in the most significant SNP being closer to the true SNP, with a smaller p-value, than typically found with non-ascertainment.

For the 100 null replicates, both under high and low LD, we did not find any SNPs significant at the genome-wide level after stage 2. Therefore, we did not perform targeted resequencing analysis, either with or without ascertainment, on the null replicates.

4. DISCUSSION

Our simulation studies show that applying ascertainment to select the sample used for targeted resequencing greatly increases the power to detect the causal SNP. We show that the greatest increase in power occurs when the causal SNP is in high LD with tagging SNPs on the standard high-throughput SNP chips used for GWAS. More importantly, we show that the increase in power persists with lower LD, which is more typical of the minimum LD found among tagging SNPs. Therefore, ascertaining on the basis of the minor allele of the best candidate SNPs at the conclusion of a two-stage GWAS can boost the signal to find the causal SNP with targeted resequencing.

Ascertaining on the basis of a tagging SNP allows us to use the LD between the causal SNP and tagging SNP to increase the power by enriching the sample with the causal allele. Since we base our ascertainment on all tagging SNPs that are significant on the genome-wide level, we are potentially selecting multiple SNPs (1-3 SNPs) in LD with the causal SNP. When multiple SNPs are tagging the causal locus, the ascertainment scheme increases the probability of capturing the causal SNP for resequencing even more. If the

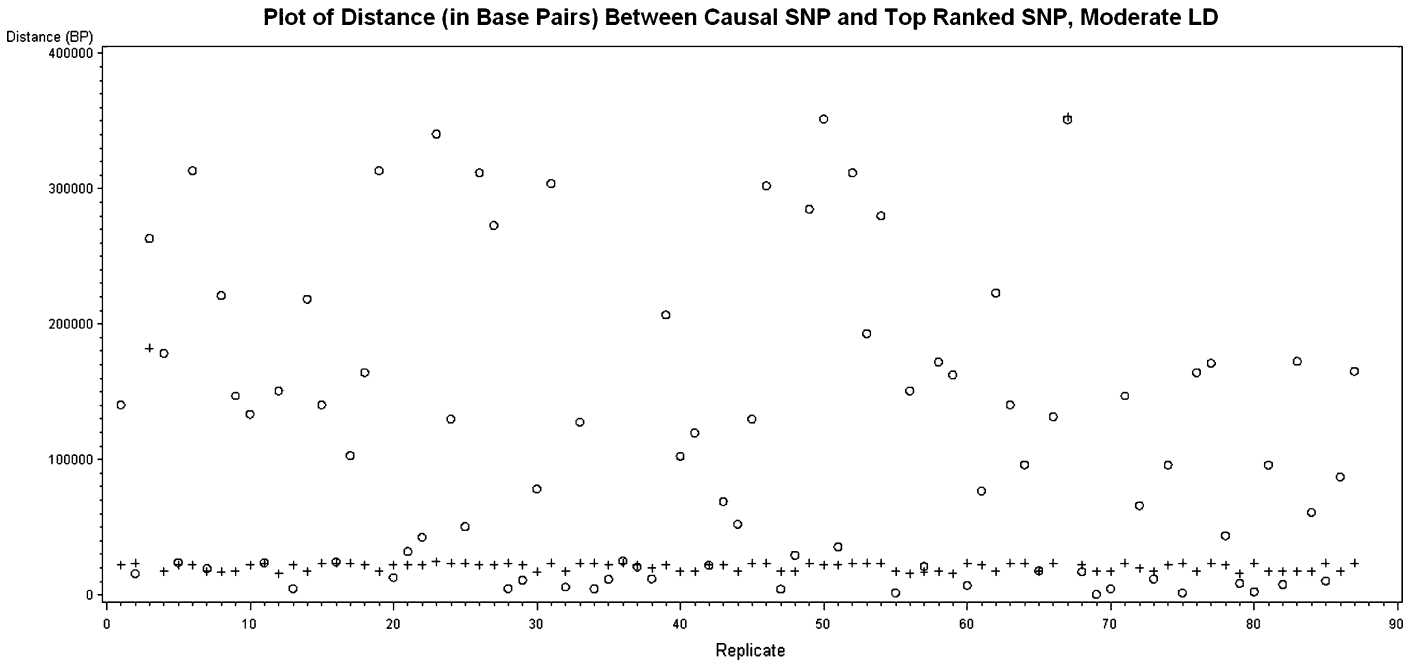

Figure 5. Plot of distances from the most significant SNP to the causal SNP. This plot shows the number of base pairs from the top-ranked SNP to the causal SNP for each replicate (iteration) of the design, comparing the ascertained targeted resequencing analysis (+) to the standard resequencing analysis (O) for lower LD.

LD is close to 0, however, tagging SNPs would not be detected in the two-stage design, and there would not be any SNPs to follow up with resequencing.

Although Appendix A shows that tagging SNPs based on r^2 measures of LD won't tag a rare variant with a major allele, it is possible for a major allele to have $D' = 1$ with a minor allele, even though the r^2 is low. This may result in a non-causal allele being tagged with higher r^2 than the causal locus and may inflate false positives. However, our simulation with moderate LD (r^2) between the causal SNP and the tagging SNP, the causal SNP is still ranked higher on average, and closer to the top ranked SNP than without ascertainment. This implies that although ascertainment can "enrich" for those non-causal SNPs in tighter LD with the tagging SNP than the causal SNP, it still enriches for the causal SNP as well. Referring to Fig. 4, we see that under ascertainment, the causal ranked SNP is ranked more often in the top 100 most significant SNPs than under non-ascertainment. Therefore, the false positive rate is not inflated over that of non-ascertainment, even if the tagging SNP is also tagging a non-causal allele.

Appendix A also guarantees improvement in power when the rare causal variant is in positive LD with the tagging SNP's minor allele. However, it is possible for protective alleles to be tagged by minor alleles. When the association of the tagging SNP is in the protective direction, conceivably, ascertaining the cases on the major allele should provide some improvement of power, however, the statistical properties of how to ascertain when the association is in the protective directions still needs to be investigated.

As a discovery method, the real false positive control comes from the two stage design. This ascertainment method will not correct for any false positive results that remain through the two stage design. The focus of this method is to increase the power to discover the causal SNP by enriching the sample frequency of potentially causal SNPs based on the two stage design. Therefore, some enriching will occur if a false positive was ascertained at the end of the two stage design. Furthermore, true positives and false positives can be determined by further biological validation of the sequences. Therefore, it is not surprising that when comparing the performance of the ascertainment method with the standard method on a data set with no disease association, the false positive rates were similar.

In this paper, we used a type 1 error control and compared ascertaining versus not ascertaining individuals for targeted resequencing. We used the typical Bonferonni-type multiple comparison correction for both the two stage GWAS and the targeted resequencing p-values. This ascertainment method easily generalizes to be used in conjunction with any type 1 error control of choice, for example, selecting SNPs on the false discovery rate correction (Benjamini and Hochberg 1995).

The method described here applies after completion of a two-stage GWAS. There are multiple issues regarding SNP coverage of the genome and power to detect a signal in a

GWAS (Anderson, et al. 2008, Barrett and Cardon 2006, Eberle, et al. 2007), and all these issues still apply when considering this method. Additionally, there has been discussion about selecting the proportion of the overall sample to be used in stage 1 and its effect on power (Gail, et al. 2008, Gail, et al. 2008). Our method can be applied after those decisions are made, and the current study assesses the additional effects on power of ascertainment after such important design decisions have been made.

Since this method describes sampling based on the significance of an association study, an inherent limitation, common to all associations studies, is that the probability of detecting the underlying causal SNP depends on the underlying LD between the causal SNP and tagging SNPs. If the proband tagging SNP for ascertainment is in stronger LD with non-causal SNPs, this method may result in some non-causal SNPs having a higher ranking than the causal SNP. However, the causal SNP will be among the top ranked SNPs, if not the top ranked SNP, by nature of the association study.

This study develops a novel design for targeted resequencing, tested on realistic simulated data sets. One of the strengths of this study is that we resampled haplotypes, and used odds ratios for disease as found in the literature. A minor limitation of this study is that we only simulate one disease locus. However, most current two stage GWAS only detect one major locus. The fact that we ascertained on multiple SNPs suggest this strategy would generalize to multiple loci.

With the development of denser SNP chips (such as the 1 million SNP chip), there will come a day when the SNP chip for genome-wide studies will most likely include a marker in high LD with the causal SNP, or even the causal SNP itself. There is also some discussion in the field regarding GWAS with low-resolution sequence data, which would involve even denser SNPs, and again, the high LD required to maximize the benefit of this ascertainment method will be available.

Essentially, our proposed targeted resequencing design with ascertainment can increase the power to find the causal SNP, following a two-stage GWAS, without increasing false positives. This simulation study supports using ascertainment in a targeted resequencing design to increase the power of genome-wide association studies, especially as the SNP chips become denser.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health, National Cancer Institute (K07 CA123109-01, 1R03CA141998-01, M.D.S; 1R01CA131324, S.S.; R01CA133996 and PO1CA 34936-14A1, B.P.) and the National Institute of Environmental Health Sciences (P30ES007784, B.P.).

AUTHOR CONTRIBUTIONS

M.D.S. helped develop the method, prepared the simulations, performed the analyses, and wrote the manuscript. B.P. simulated the data and helped write part of the manuscript. C. R-G helped with the revision of the manuscript and the response to reviewers. S.S. conceived of and helped to develop the method, directed the analyses, and participated in writing and revising the manuscript.

APPENDIX A. ASCERTAINING INCREASES THE PROBABILITY OF SEQUENCING THE DISEASE SNP

Here we show that ascertaining cases for resequencing based on the minor allele of the tagging SNP increases the probability of finding the causal SNP, assuming that the tagging SNP used for ascertainment actually tags the causal SNP. We consider the haplotypes consisting of the disease locus (D) and the Tag locus (T). We denote the minor alleles with lower-case letters and the major alleles with upper-case letters. Let δ denote the linkage disequilibrium parameter, and let p denote the sample allele frequency of the d allele, and q denote the sample allele frequency of the t allele. Then we can define the haplotype frequencies as:

$$\begin{aligned} H_{dt} &= pq + \delta \\ H_{dT} &= p(1-q) - \delta \\ H_{Dt} &= (1-p)q - \delta \\ H_{DT} &= (1-p)(1-q) + \delta. \end{aligned}$$

From the above haplotype frequencies, we can construct the diplotype frequencies of observed individuals in the sample. We have 10 different diplotypes, denoted as G.

$$\begin{aligned} G_1 &= H_{dt}H_{dt} \\ G_2 &= 2H_{dt}H_{dT} \\ G_3 &= 2H_{dt}H_{Dt} \\ G_4 &= 2H_{dt}H_{DT} \\ G_5 &= H_{dT}H_{dT} \\ G_6 &= 2H_{dT}H_{Dt} \\ G_7 &= 2H_{dT}H_{DT} \\ G_8 &= H_{Dt}H_{Dt} \\ G_9 &= 2H_{Dt}H_{DT} \\ G_{10} &= H_{DT}H_{DT}. \end{aligned}$$

Given the above diplotype definitions, the probability of unconditionally sampling the disease SNP for targeted resequencing can be computed by:

$$\begin{aligned} (1) \quad & P(d \in \text{resequencing sample}) \\ & = G_1 + G_2 + \frac{1}{2}G_3 + \frac{1}{2}G_4 + G_5 + \frac{1}{2}G_6 + \frac{1}{2}G_7 = p. \end{aligned}$$

While the probability of sampling the disease SNP for targeted resequencing ascertained with the tagging SNP is given by:

$$
\begin{aligned}
(2) \qquad & P(d \in \text{resequencing sample} \mid t) \\
&= \frac{G_1 + G_2 + \frac{1}{2}G_3 + \frac{1}{2}G_4 + \frac{1}{2}G_6}{1 - (G_5 + G_7 + G_{10})} \\
&= \frac{-\delta - 2pq + \delta q + pq^2}{q(q-2)}.
\end{aligned}
$$

It is straightforward to see that in the absence of linkage disequilibrium (LD), $P(d \in \text{resequencing sample} \mid t) = P(d \in \text{resequencing sample}) = p$. Assuming LD between the tagging SNP and the causal SNP, we next show that

$$
\begin{aligned}
&P(d \in \text{resequencing sample} \mid t) \\
&\quad > P(d \in \text{resequencing sample}).
\end{aligned}
$$

Subtracting the unconditional probability from the conditional probability yields

$$
(3) \qquad \frac{\delta(q-1)}{q(q-2)}.
$$

Upon inspection of (3), we see that the quantity is not affected by the value of the disease allele frequency, p. Therefore, for $\delta > 0$, and any $q \in (0,1)$, the difference will be positive, and we have greater probability of including the causal SNP in our resequencing sample when we ascertain on the basis of the minor allele of a tagging SNP.

However, notice that the probability of detecting the causal allele is smaller under ascertainment if $\delta < 0$. Consider $\delta = H_{dt} - pq$, then by definition, H_{dt} is constrained to be in the interval $[\max(0, p+q-1), \min(p,q)]$ (VanLiere and Rosenberg 2008). Since d and t are both minor alleles, then $p < 0.5$, and $q < 0.5$, which implies $p+q < 1$, reducing the interval to $[0, \min(p,q)]$.

Also, since $\delta = H_{dt} - pq$, then $\delta < 0$ when H_{dt} is in the interval $[0, pq)$, and $\delta > 0$ when H_{dt} is in the interval $[pq, \min(p,q)]$. We have already shown above that under the latter case, ascertainment improves power. When $\delta < 0$, consider r^2 computed by

$$
(4) \qquad r^2 = \frac{(H_{dt} - pq)^2}{p(1-p)q(1-q)} = \frac{\delta^2}{p(1-p)q(1-q)}.
$$

When $H_{dt} \in [0, pq]$, then r^2 is maximum when $H_{dt} = 0$. Furthermore, $H_{dt} = 0$ implies

$$
(5) \qquad \max_{\delta<0}(r^2) = \frac{p^2 q^2}{p(1-p)q(1-q)}.
$$

Rearranging equation (5), we get

$$
(6) \qquad \max_{\delta<0}(r^2) = \left(\frac{p}{1-p}\right)\left(\frac{q}{1-q}\right).
$$

From equation (6), it is easy to see that since both t and d are minor alleles, for r^2 to be big enough for a tagging SNP, p and q have to be close to 0.5. For tagging SNPs, we are interested in the $r^2 > c$. Additionally, since d is our disease allele, we are interested in how large p has to be for r^2 large enough to be a tagging SNP. The smallest admissible values for p would occur when $q = 0.5$. Then we consider

$$
(7) \qquad \max_{\delta<0}(r^2) = \frac{p}{1-p} > c,
$$

which implies that $p > c/(1-c)$. Typically, for tagging SNPs, we use $c = 0.8$, which implies $p > 0.44$. If $c = 0.5$, then $p > 0.33$, which is still high for a disease allele frequency, when the allele is assumed to be rare. Therefore, for a typical rare disease allele, when $\delta < 0$ between the disease allele and a marker allele, it won't be tagged by that marker.

This generalizes to the marker SNP major allele as well. When $\delta < 0$ between the minor allele of the disease locus and the minor allele of the marker, it implies $\delta > 0$ between the minor allele of the disease locus and the major allele of the marker locus. Therefore, we can say that the likelihood of a rare disease allele being tagged by a common marker allele is also rare, and ascertaining on the minor allele will improve the power for most studies, especially using tagging SNPs with $r^2 > 0.8$.

Received 1 June 2010

REFERENCES

Alberts, B. (2002). *Molecular Biology of the Cell.* Garland Science, New York.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y. Y., Chen, W. V., Shete, S., Spitz, M. R. and Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40** 616–622.

Anderson, C. A., Pettersson, F. H., Barrett, J. C., Zhuang, J. J., Ragoussis, J., Cardon, L. R. and Morris, A. P. (2008). Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* **83** 112–119.

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38** 659–662.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57** 289–300. MR1325392

Eberle, M. A., Ng, P. C., Kuhn, K., Zhou, L., Peiffer, D. A., Galver, L., Viaud-Martinez, K. A., Lawley, C. T., Gunderson, K. L., Shen, R. and Murray, S. S. (2007). Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet.* **3** 1827–1837.

Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Olama, A. A., Guy, M., Jugurnauth, S. K., Mulholland, S., Leongamornlert, D. A., Edwards, S. M., Morrison, J., Field, H. I., Southey, M. C., Severi, G., Donovan, J. L., Hamdy, F. C., Dearnaley, D. P., Muir, K. R., Smith, C., Bagnato, M., Ardern-Jones, A. T., Hall, A. L., O'Brien, L. T., Gehr-Swain, B. N., Wilkinson, R. A., Cox, A., Lewis, S., Brown, P. M., Jhavar, S. G., Tymrakiewicz, M., Lophatananon, A., Bryant, S. L., Horwich, A.,

Huddart, R. A., Khoo, V. S., Parker, C. C., Woodhouse, C. J., Thompson, A., Christmas, T., Ogden, C., Fisher, C., Jamieson, C., Cooper, C. S., English, D. R., Hopper, J. L., Neal, D. E. and Easton, D. F. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40** 316–321.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, London.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A. M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M. R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N., Doney, A. S., Morris, A. D., Smith, G. D., Hattersley, A. T. and McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316** 889–894.

Gail, M. H., Pfeiffer, R. M., Wheeler, W. and Pee, D. (2008). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics* **9** 201–215.

Gail, M. H., Pfeiffer, R. M., Wheeler, W. and Pee, D. (2008). Probability that a two-stage genome-wide association study will detect a disease-associated snp and implications for multistage designs. *Ann. Hum. Genet.* **72** 812–820.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6** 95–108.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., Chen, C., Goodman, G., Field, J. K., Liloglou, T., Xinarianos, G., Cassidy, A., McLaughlin, J., Liu, G., Narod, S., Krokan, H. E., Skorpen, F., Elvestad, M. B., Hveem, K., Vatten, L., Linseisen, J., Clavel-Chapelon, F., Vineis, P., Bueno-de-Mesquita, H. B., Lund, E., Martinez, C., Bingham, S., Rasmuson, T., Hainaut, P., Riboli, E., Ahrens, W., Benhamou, S., Lagiou, P., Trichopoulos, D., Holcatova, I., Merletti, F., Kjaerheim, K., Agudo, A., Macfarlane, G., Talamini, R., Simonato, L., Lowry, R., Conway, D. I., Znaor, A., Healy, C., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Matsuda, F., Blanche, H., Gut, I., Heath, S., Lathrop, M. and Brennan, P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452** 633–637.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37** 413–417.

Peng, B. and Kimmel, M. (2005). simuPOP: A forward-time population genetics simulation environment. *Bioinformatics* **21** 3686–3687.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* **25** 149–157.

Shete, S., Hosking, F. J., Robertson, L. B., Dobbins, S. E., Sanson, M., Malmer, B., Simon, M., Marie, Y., Boisselier, B., Delattre, J. Y., Hoang-Xuan, K., El Hallani, S., Idbaih, A., Zelenika, D., Andersson, U., Henriksson, R., Bergenheim, A. T., Feychting, M., Lonn, S., Ahlbom, A., Schramm, J., Linnebank, M., Hemminki, K., Kumar, R., Hepworth, S. J., Price, A., Armstrong, G., Liu, Y., Gu, X., Yu, R., Lau, C., Schoemaker, M., Muir, K., Swerdlow, A., Lathrop, M., Bondy, M. and Houlston, R. S. (2009). Genome-wide association study identifies five susceptibility loci for glioma. *Nature Genetics* **41** 899–904.

Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38** 209–213.

Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2007). Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* **31** 776–788.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445** 881–885.

Swartz, M. D., Yu, R. K. and Shete, S. (2008). Finding factors influencing risk: Comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. *Statistics in Medicine* **27** 6158–6174. MR2522315

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Jr., Hoover, R., Hayes, R. B., Hunter, D. J. and Chanock, S. J. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40** 310–315.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsater, A., Flex, A., Aben, K. K., de Vegt, F., Mulders, P. F., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., van Rij, A. M., Dieplinger, B., Haltmayer, M., Mayordomo, J. I., Kiemeney, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfingsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A. and Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452** 638–642.

VanLiere, J. M. and Rosenberg, N. A. (2008). Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical Population Biology* **74** 130–137.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., de Villena, F. P., Sullivan, P. F., Wilhelmsen, K. C. and Zou, F. (2007). Simulating association studies: A data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* **23** 2581–2588.

Wu, X., Ye, Y., Kiemeney, L. A., Sulem, P., Rafnar, T., Matullo, G., Seminara, D., Yoshida, T., Saeki, N., Andrew, A. S., Dinney, C. P., Czerniak, B., Zhang, Z. F., Kiltie, A. E., Bishop, D. T., Vineis, P., Porru, S., Buntinx, F., Kellen, E., Zeegers, M. P., Kumar, R., Rudnai, P., Gurzau, E., Koppova, K., Mayordomo, J. I., Sanchez, M., Saez, B., Lindblom, A., de Verdier, P., Steineck, G., Mills, G. B., Schned, A., Guarrera, S., Polidoro, S., Chang, S. C., Lin, J., Chang, D. W., Hale, K. S., Majewski, T., Grossman, H. B., Thorlacius, S., Thorsteinsdottir, U., Aben, K. K., Witjes, J. A., Stefansson, K., Amos, C. I., Karagas, M. R. and Gu, J. (2009). Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nature Genetics* **41** 991–995.

Yamada, Y., Fuku, N., Tanaka, M., Aoyagi, Y., Sawabe, M., Metoki, N., Yoshida, H., Satoh, K., Kato, K., Watanabe, S., Nozawa, Y., Hasegawa, A. and Kojima, T. (2009). Identification of CELSR1 as a susceptibility gene for ischemic stroke in Japanese individuals by a genome-wide association study. *Atherosclerosis* **207** 144–149.

M. D. Swartz
Division of Biostatistics
The University of Texas Health
Science Center at Houston (UT Health)
School of Public Health
1200 Pressler Blvd RAS W920
Houston, TX 77030, USA
Phone: 713-500-9570
Fax: 713-500-9329
E-mail address: michael.d.swartz@gmail.com

B. Peng
Department of Epidemiology
The University of Texas
M. D. Anderson Cancer Center
Houston, TX 77030, USA

C. Reyes-Gibby
Department of Epidemiology
The University of Texas
M. D. Anderson Cancer Center
Houston, TX 77030, USA

S. Shete
Department of Epidemiology
The University of Texas
M. D. Anderson Cancer Center
Houston, TX 77030, USA