Published in final edited form as:
Hum Genet. 2012 January ; 131(1): 111–119. doi:10.1007/s00439-011-1054-1.

Artifact due to differential error when cases and controls are imputed from different platforms

Jennifer A. Sinnott and
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA

Peter Kraft
Department of Epidemiology and Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA, Phone: 617 432 4271

Jennifer A. Sinnott: jsinnott@hsph.harvard.edu; Peter Kraft: pkraft@hsph.harvard.edu

Abstract

Including previously-genotyped controls in a genome-wide association study can provide cost-savings, but can also create design biases. When cases and controls are genotyped on different platforms, the imputation needed to provide genome-wide coverage will introduce differential measurement error and may lead to false positives. We compared genotype frequencies of two healthy control groups from the Nurses' Health Study genotyped on different platforms (Affymetrix 6.0 [n=1,672] and Illumina HumanHap550 [n=1,038]). Using standard imputation quality filters, we observed 9,841 SNPs out of 2,347,809 (0.4%) significant at the 5×10^{-8} level. We explored three methods for controlling for this Type I error inflation. One method was to remove platform effects using principal components; another was to restrict to SNPs of highest quality imputation; and a third was to genotype some controls alongside cases to exclude SNPs that are statistical artifact. The first method could not reduce the Type I error rate; the other two could dramatically reduce the error rate, although both required that a portion of SNPs be excluded from analysis. Ideally, the biases we describe would be eliminated at the design stage, by genotyping sufficient numbers of cases and controls on each platform. Researchers using imputation to combine samples genotyped on different platforms with severely unbalanced case-control ratios should be aware of the potential for inflated Type I error rates and apply appropriate quality filters. Every SNP found with genome-wide significance should be validated on another platform to verify that its significance is not an artifact of study design.

Keywords

Genome-wide association study; Imputation; GWAS quality control

Introduction

A population-based genome-wide association (GWA) study requires thousands of cases and controls in order to detect moderate associations between SNPs and disease, and each person genotyped can cost hundreds of dollars. Thus, when researchers plan numerous GWA studies for different diseases, it would be attractive to use the same healthy control group for more than one disease if all cases are being drawn from the same underlying population. The Wellcome Trust Case Control Consortium (WTCCC) demonstrated the effectiveness of this approach by comparing case groups of 7 major diseases to a shared control group (Wellcome Trust Case Control Consortium 2007). Additionally, researchers may want to bring in publicly available controls to increase power without increasing cost. Zhuang et al. (2010) advocated this approach, and Ho and Lange (2010) did extensive simulations in this

vein that demonstrate the potential improvement in power. Ho and Lange provided some examples of studies that have augmented their control groups with publicly available controls (Hom et al. 2008; Wrensch et al. 2009).

A complication in the reuse of control groups or the inclusion of external controls arises when investigators wish to genotype cases on a platform different from the one used for controls. This may easily happen as genotyping technology changes and new chips with new pricing plans become available. It can appear necessary when funding is too limited to support a sufficiently powered study with both cases and controls genotyped together. Moreover, even if funding exists to genotype or re-genotype a control group on a particular chip, there may be limited biological samples available for use, or a desire to conserve such samples. However, while each platform genotypes a collection of tagging SNPs, different platforms choose these tagging SNPs in different ways. For example, Illumina uses patterns of linkage disequilibrium in the HapMap to choose its tagging SNPs, while Affymetrix (Affy) provides a large but less determinate collection of SNPs designed to give good coverage of the entire genome. There is not necessarily much overlap between the SNPs genotyped on two different platforms. For example, there were 140,325 SNPs in the overlap between the 508,123 markers on the Illumina HumanHap550 chip and the 606,625 markers on the Affymetrix Genome-Wide Human 6.0 array we use in this study. Thus, if we restricted to SNPs in the overlap, we would drop about three-quarters of the SNPs we have available on each of these chips.

When pooling genotype data from different platforms, investigators could impute the SNPs missing on each platform to get a data set with comparable variables. This approach has been suggested as a way of combining study cases and controls with publicly available controls genotyped on a different chip (Zhuang et al. 2010). Fallin et al. (2010) used imputation to combine their case-control study, genotyped on Illumina, with a publicly available case-control study genotyped on Affy. A number of imputation methods exist, and they have been shown to be very accurate in the typical setting where cases and controls are genotyped together on the same platform (Li et al. 2010; Howie et al. 2009). However, their performance in the setting we are discussing here, when cases and controls have been genotyped on different platforms, has been largely unexplored.

After imputation, investigators run association tests as usual, producing *p*-values for each SNP and looking for the most significant SNPs. However, the imputation has introduced differential measurement error: for example, some SNPs are measured almost perfectly (through actual genotyping) among the controls, but measured imperfectly (through imputation based on nearby measured SNPs) among the cases. Furthermore, the imputation itself may introduce bias. Many imputation programs base the imputation on a database of known genomes, such as the HapMap. If the minor allele frequency (MAF) of a SNP in the HapMap differs substantially from the MAF in study data, imputation in cases only or controls only can yield very different MAFs in cases and controls. This setting has been recognized as potentially problematic. For example, when discussing combining data from studies using different genotyping platforms, Li et al. (2010) recommends imputing and doing association tests within platform and then combining the results using a meta-analysis approach, which cannot be implemented unless each platform has at least some cases and controls.

Differential error induced by imputation may yield SNPs that appear to differ substantially between cases and controls purely as a result of the imputation. Past studies have shown that differential genotyping error between cases and controls can inflate Type I error rates (e.g. Moskvina et al. 2006). A recent study by Sebastiani et al. (2010) which built a model using 150 SNPs to predict longevity has been criticized for not controlling for different chips used

with different frequencies between cases and controls. Critics suspect that many of the significant SNPs it identified are artifact of differential genotyping errors between these different chips (Alberts 2010; Carmichael 2010).

In this paper, we are concerned with problems occurring one step further down the pipeline. Under the assumption that markers actually genotyped by each chip are being genotyped with good accuracy, we investigate how well Type I error rates are maintained after imputation in a study where cases and controls are genotyped on different platforms. To do this, we used the healthy control groups from two studies nested within the Nurses' Health Study: a Type 2 Diabetes (T2D) study genotyped on Affy, and a Breast Cancer (BrCa) study genotyped on Illumina. After imputation within each study, we label the T2D controls "cases" and the BrCa controls "controls," and fit a logistic regression predicting this case-control status from each SNP. We expect there to be no substantial genetic differences between these two groups –so any significant differences we see reflect a Type I error rate higher than expected.

When we did in fact observe inflated Type I error after applying standard imputation quality filters, we explored a number of ways to lessen the inflation. We first considered controlling for platform effect as we would control for population stratification: by using principal components (PCs) as covariates in logistic regression. However, the platform effect was so strong and confounded with case-control status that we could not fit the models. Then we considered restricting to SNPs imputed with good accuracy. This approach yields excellent results, but reduces power by reducing the number of SNPs we can test. Finally, we considered the possibility of genotyping a small number of additional controls alongside cases on the new platform, who could be compared to the original controls in a preliminary analysis to identify aberrant SNPs. This approach yields good results, but requires the additional expense of genotyping more subjects.

Methods

The BrCa and T2D studies have been described elsewhere (Hunter et al. 2007; Qi et al. 2010). Both studies were restricted to women of European ancestry. Genotyping in the BrCa study was done on the Illumina HumanHap550 chip, while the T2D study was genotyped on the Affymetrix Genome-Wide Human 6.0 array. We imputed missing genotypes separately within each study using MaCH 1.0, which relies on Markov chain haplotyping (http://www.sph.umich.edu/csg/yli/mach/index.html) (Li et al. 2009, 2010). We present results from imputation done separately in the two studies; when the two control groups were pooled first and then the imputation was done, results were similar. The imputations used HapMap Release 22 (NCBI build 36) as a reference panel. For each unmeasured SNP, we considered both a *soft call*, or *dosage*, imputation, which gives the expected number of rare alleles given the other SNPs available for that individual and takes values on a continuum between 0 and 2, and a *hard call* imputation, which gives the best integral guess for the number of rare alleles, either 0, 1, or 2. We had available 1,038 BrCa controls, which we labeled "controls," and 1,672 T2D controls, which we labeled "cases." SNPs with MAF < 0.025 (calculated using both groups after imputation) or imputation quality $R^2 < 0.30$ (calculated in either group) were removed.

We ran a logistic regression for each of m SNPs, modeling the log-odds of being a "case" ($Y = 1$) as a linear function of the number of rare alleles at the locus. That is, for the i^{th} SNP, $i = 1,\ldots,m$, with A_i copies of the rare allele, we fit

$$\log\left\{\frac{P(Y_i=1)}{1-P(Y_i=1)}\right\}=\beta_0+\beta_1 A_i$$

where β_1 is the effect of SNP i and β_0 is an intercept term. We stored the p-value and the χ^2 test statistic for the Wald test of β_1. For the soft call genotypes, where A_i is the expected number of rare alleles given the observed data ($0 \le A_i \le 2$), the software mach2dat was used (http://www.sph.umich.edu/csg/yli/mach/index.html) (Li et al. 2009, 2010). For the hard call genotypes, where $A_i \in \{0, 1, 2\}$, we used the software PLINK version 1.07 (http://pngu.mgh.harvard.edu/purcell/plink) (Purcell et al. 2007). Figures were generated in the statistical software R version 2.9.0 (R Development Core Team 2009).

We grouped the SNPs into four categories: SNPs genotyped on both chips; SNPs genotyped on Affy and imputed for the Illumina controls; SNPs genotyped on Illumina and imputed for the Affy controls; and SNPs imputed for both groups. The false positives found among SNPs genotyped on both platforms can be thought of as a baseline error rate against which to compare the other three groups. For each group of SNPs we summarized the error rates using two quantities: the Genomic Control λ and the percentage of SNPs with p-value less than 5×10^{-8}. For χ^2 test statistics X_i, $i = 1, \ldots, m$, the Genomic Control λ is defined as

$$\lambda=\underset{i=1,\ldots,m}{\text{median}}\{X_i\}/0.455$$

where 0.455 is approximately the theoretical median of a χ^2_1 distribution (Devlin and Roeder 1999). Our model assumes the null distribution of each X_i is χ^2_1, so if this assumption is valid, we should have $\lambda \approx 1$. A value of $\lambda > 1$ suggests that the observed variance of the test statistic is larger than the theoretical variance, which will tend to increase the number of false positives. We also calculated the percentage of SNPs significant at the 5×10^{-8} significance level, a standard significance level used for GWA studies (McCarthy et al. 2008). Assuming the genotype is measured accurately, we don't expect genotype frequency differences between our cases and controls, because they are both samples of healthy women used as control groups for other studies. Thus, we should see very few SNPs with such significant p-values (approximately 1 out of every 20,000,000 independent tests).

When $\lambda > 1$ and the percentage of SNPs significant at the 5×10^{-8} level was more than expected in our null setting, we explored 3 methods for controlling for the error inflation:

Method 1

We investigated whether we could capture the platform effect using PCs. To do this, we used EIGENSTRAT (http://genepath.med.harvard.edu/~reich/Software.htm) (Patterson et al. 2006; Price et al. 2006). In a typical application of this program, the first few PCs are calculated and included as covariates in logistic regression to capture and control for population stratification. An example in Price et al. (2006) suggests the possibility of some components capturing lab and batch effects as well. We calculated the first ten PCs and assessed how well they correlated with platform effect. Then we attempted to include these components as covariates in logistic regression models predicting case-control status from each SNP. We did this in two ways: first, we calculated the PCs using all measured and imputed SNPs; second, we restricted to SNPs in each of the four categories, and calculated PCs using only those SNPs (e.g., using only SNPs measured on one chip and imputed in the other).

Method 2

When missing genotypes are imputed by MaCH, each SNP has an R^2 value associated with it that quantifies the quality of the imputation. The R^2 value is an estimate of the squared correlation between the imputed genotype and the actual genotype, so a higher R^2 corresponds to a SNP imputed with more certainty. Standard advice is to restrict to SNPs with $R^2 > 0.3$, which we did (Scott et al. 2007). It is expected that this will remove 70% of poorly imputed SNPs while keeping 99.5% of better imputed SNPs (Li et al. 2010). To reduce the error inflation in our less standard setting, we considered restricting to SNPs imputed at even higher quality.

Focusing on SNPs measured on one chip and imputed in the other, we considered removing SNPs with imputation $R^2 < 0.5,0.75,0.9,0.95$ and 0.99. After thresholding by each value of R^2, we calculated λ and the percentage of SNPs with $p < 5\times10^{-8}$. We kept track of the number of SNPs still available for analysis at each threshold.

We also constructed an ROC curve to assess the discriminatory ability of this method. We labeled SNPs with $p < 5 \times10^{-8}$ as "problematic." As we varied the R^2 threshold between 0 and 1, we compared how many problematic SNPs were being detected (sensitivity) to how many non-problematic SNPs were being excluded due to low R^2 (1 – specificity).

Method 3

The genotype distributions for some SNPs may differ markedly across platforms due to genotyping artifact or differences in imputation quality. These differences may be identified even in relatively small samples. We explored the possibility of genotyping a small number of additional controls along with the cases, which could be used to identify and eliminate the problematic SNPs. Researchers would perform a preliminary analysis comparing the additional controls to the original controls, and any SNP significant in this preliminary analysis would be discarded. Researchers could then perform standard association tests between cases and controls using the remaining SNPs.

We randomly selected 1000 subjects from the 1,038 on Illumina to serve as controls, and 1000 subjects from the 1,672 on Affy to serve as cases. Then from the remaining 672 subjects on Affy, we selected n additional subjects to serve as controls genotyped alongside cases on the Affy platform. We first performed a screening step, in which we compared these n Affy controls to the 1000 Illumina controls and eliminated SNPs significant at level α. Then, restricting to SNPs that passed this screening, we performed the main analysis, comparing the 1000 Illumina controls to the 1000 Affy cases, and calculated the Genomic control λ and the percentage of SNPs with $p < 5 \times 10^{-8}$ in this main analysis. We did this calculation for $n = 100,300$ and 500, and for $\alpha = 0.001,0.01,0.1$, and 0.2. We also constructed ROC curves to assess the discriminatory ability of this method while varying α, the screening threshold. That is, as we varied the α screening threshold between 0 and 1, we compared how many problematic SNPs (in the main analysis of 1000 Illumina controls vs. 1000 Affy cases) were being detected to how many non-problematic SNPs were being excluded.

Results

Figure 1 summarizes the results of a standard logistic regression analysis, where SNPs are grouped by MAF. For each collection of SNPs, we found the Genomic Control λ (in black) and the percentage of SNPs with $p < 5 \times 10^{-8}$ (in gray). Results from the soft call analysis are shown in solid lines, while those from the hard call analysis are shown in dashed lines. In Figure 1a, we see that $\lambda \approx 1$ among the 139,732 SNPs measured on both chips, and the percentage of highly significant SNPs is close to 0 across all MAFs; the error measures in

this setting are virtually identical whether we use hard call or soft call imputation. Thus, when we consider only the SNPs measured on both chips, we have no evidence from these two measures that the distribution of the test statistics deviates from the null.

However, among the 357,361 SNPs measured on Illumina and imputed on Affy (Figure 1b), we see an overall increase in λ to 1.6. We see an increase in the percentage of highly significant SNPs to 1.3% when using soft call genotypes, and to 2.1% when using hard call genotypes. Thus, when using hard call genotypes, 7,644 SNPs are being declared significant at the 5×10^{-8} level. These increases are most prominent among SNPs with low MAF, as shown in the Figure. The Type I error inflation is also apparent, though less dramatic, among the 458,034 SNPs measured on Affy and imputed on Illumina (Figure 1c) where λ = 1.3 overall, and where we are seeing 0.4% highly significant SNPs when using soft calls and 0.8% highly significant SNPs when using hard calls; we see similar numbers among the 1,392,682 SNPs imputed in both (Figure 1d). Results were largely unchanged when we first pooled the two groups and then imputed.

To try to correct these problems, we applied the three described methods. Here, we present results for the SNPs measured on Illumina and imputed on Affy for simplicity; results were similar in the other two problematic cases.

Method 1

We found the first ten PCs using hard call genotypes because those are currently supported by EIGENSTRAT. We did this once using all SNPs, and once restricting to SNPs measured on Illumina and imputed on Affy. Results were similar in the two approaches, and results from the latter are shown. The top three PCs are plotted against one another in Figure 2. We see that the second PC completely separates the cases (i.e., the Affy controls) and controls (the Illumina controls). Thus, when these PCs are included in a logistic regression predicting case-control status, we get a complete separation of data points, and the models cannot be fit.

Method 2

We considered restricting to SNPs imputed with increasingly higher quality, as quantified by the imputation R^2. Results for the soft call genotypes are shown in Table 1. As the R^2 threshold was increased, our summary measures improved; however, this happened at the expense of losing SNPs for analysis, which reflects some loss of power. It should also be noted that even at the most stringent threshold listed, $R^2 > 0.99$, when we've excluded nearly 70% of the SNPs, there remain 57 SNPs with $p < 5 \times 10^{-8}$. Figure 3 shows the discriminatory ability of this method as we vary the R^2 threshold.

Method 3

For various thresholds (α = 0.001,0.01,0.1,0.2) and various numbers of additional controls on Affy (n = 100,300,500) we removed SNPs significant at level α in a preliminary analysis comparing the n additional Affy controls to the 1000 original Illumina controls. We then performed standard logistic regressions comparing the 1000 Illumina controls to the 1000 Affy cases using the remaining SNPs. The genomic control λ and the percentage of highly significant SNPs were calculated; results for the soft call genotypes are shown in Table 2. As n increased and α increased, our summary measures improved. This again happened at the expense of losing SNPs for analysis, but not as quickly as in Method 2. Figure 4 shows the discriminatory ability of this method for each n as we vary the α screening threshold.

Discussion

We observed a large number of highly significant SNPs after imputation in a study comparing two healthy control groups genotyped on different platforms. Because both control groups are nested in the NHS and chosen using similar criteria, we expect no SNPs to significantly distinguish the two groups in the absence of measurement error, and we expect no differential population substructure. Thus, statistically significant SNPs are false positives, and must be due to genotyping or imputation error. Furthermore, because we see almost no inflation in Type I error among SNPs actually genotyped on both chips (Figure 1a), the false positives do not appear to result from genotyping error. Rather, the inflation in Type I error is seen among SNPs measured in one group and imputed in the other (and among SNPs imputed in both). In this setting, it would be detrimental to avoid imputation altogether since only about a quarter of the SNPs genotyped on each platform overlap, so that three-quarters of the SNPs on each chip would be unusable without any imputation. Thus, we need to understand the errors being introduced by imputation and attempt to control for them.

We believe that the inflation in Type I error is due to bias introduced by the differential imputation. The imputation uses individuals in the HapMap as a reference panel, and it seems plausible that estimates in the HapMap, particularly for rare alleles, may diverge from the allele frequencies observed in our population. Thus, if a rare allele has similar frequencies in our cases and controls but is not well covered in the HapMap, the *p*-value calculated when the SNP is measured in one group and imputed in the other will tend to be smaller than the *p*-value that would arise if that SNP were measured in both groups. Moreover, among SNPs with low MAF, Moskvina et al. (2006) showed that even modest differential errors in genotype calling can yield an inflation in Type I error. Generalized to our setting, this suggests that even slight differential errors in imputation among SNPs with low MAF would lead to false positive associations. This is borne out by our results, where we see larger numbers of highly significant *p*-values among SNPs with low MAF, as shown in Figure 1.

The percentage of highly significant SNPs is noticeably larger in the hard call analysis than in the soft call analysis. This is because the soft call imputations better account for uncertainty in the imputed values. We recommend using soft calls, or another technique that accounts for imputation uncertainty, in order to reduce false positives. It is worth considering whether we could somehow alter the imputation methods themselves to avoid these false positives altogether; however, it is unclear to what extent this is possible. Imputation algorithms are limited by the information they are provided. For some platforms, the genotyped SNPs provide enough information to accurately infer an unobserved SNP; for other platforms, they do not, regardless of the imputation algorithm. Moreover, current imputation methods have good accuracy, particularly for SNPs with higher imputation R^2 (Li et al. 2010), yet even SNPs with high R^2 appear among our false positives. This suggests that even well-imputed SNPs can be falsely significant when the imputation error is differential.

The inflation in Type I error appears to be most dramatic among SNPs measured in Illumina and imputed in Affy. We suspect that this is because Illumina uses HapMap for SNP selection, and we used HapMap for SNP imputation. When we considered SNPs common to both chips, the distribution of test statistics was what we expect under the null, suggesting that the actual genotyping across the two chips is in good agreement.

When we attempted to reduce the error inflation using PCs, in Method 1, we observed a complete separation of the two control groups. This complete separation shows the difficulty

of controlling for platform effect by simply adjusting for PCs. Including the PCs as covariates in the model is equivalent to including case-control status as a covariate, and thus there does not appear to be a direct way to use those PCs to resolve the error inflation problem. Furthermore, any method using the PCs would likely wash out all differences between cases and controls in a non-null setting. Thus, it makes sense to focus on approaches that filter out problematic SNPs and exclude them from subsequent analysis. Methods 2 and 3 are two such approaches.

In Method 2, we used imputation quality to filter SNPs before performing any association tests. This approach improved the results and does not require genotyping any additional controls. It reduces the number of SNPs available for analysis, but still allows the use of more SNPs than just those actually genotyped on both platforms. However, in our example of SNPs genotyped on Illumina and imputed on Affy, even after filtering to SNPs imputed with $R^2 > 0.99$ (allowing us to retain only 30% of SNPs), we are left with 57 SNPs with highly significant *p*-values out of 112,249 remaining SNPs. So if this method is used, researchers should be prepared to sift through many false positives in a second stage analysis to find any true associations. Furthermore, this method will tend to reduce power to detect SNPs in regions with low linkage disequilibrium. Beecham et al. (2010) demonstrated this problem by pooling two case-control GWA studies for Alzheimer disease which had been genotyped on different chips, and testing for associations in the *APOE* gene, which is known to be strongly associated with risk. They used imputation to produce commensurable data sets, and filtered out SNPs according to imputation quality. They found that even though each study separately found strong associations in the *APOE* gene, there was no association in the pooled analysis, because many SNPs had been excluded due to low imputation quality measures caused by weak linkage disequilibrium in the region.

In Method 3, we propose genotyping a small number of additional controls alongside the cases and performing a preliminary step of filtering SNPs by comparing these additional controls to the original controls. This approach also improves results, but at increased monetary cost. It should, however, retain more non-artifactual SNPs while reducing the number of artifactual SNPs. In our example of 1000 cases and 1000 controls, it appeared that genotyping 300 additional controls alongside cases would allow researchers to filter out most of the false positives — with $\alpha = 0.2$, only 5 highly significant SNPs were left among the SNPs genotyped on Illumina and imputed on Affy, with 264,519 (74%) remaining for analysis. We believe these results would be the same if we had new cases and controls on Illumina and a separate control group on Affy — we merely consider this setting because it made best use of the subjects available on each chip. This method is in line with the discussion in McCarthy et al. (2008) regarding the use of historical controls. McCarthy et al. listed many possible sources of systematic error that might arise in the use of historical controls, and recommended always genotyping some ethnically matched controls alongside cases on the same platform.

It may also be worth considering a related study design in which very little error inflation was seen, which was considered by Howie et al. (2009). In their setting, a central control group in the WTCCC was genotyped on both Affy and Illumina, while different case groups from different disease studies were genotyped on just one of these platforms. The authors were interested in whether imputing SNPs missing in cases using both the HapMap and the central control group as a reference panel led to inflated Type I error. To assess this, they compared the central control group with another control group genotyped on Affy alone. They imputed SNPs missing in this new control group and then performed association tests. They found very few significant results, which demonstrated minimal inflation of Type I error in this setting. Their methods differ slightly from ours; however, we believe that the most important difference was the nested structure of their design – that is, that their central

control group had SNPs from both Affy and Illumina chips, rather than Illumina alone. A comparison of their results and ours suggests that if a central control group is going to be reused for different diseases, it may be wise to invest in genotyping the central control group on multiple platforms. A similar conclusion is offered by Marchini and Howie (2010).

Researchers can make use of accumulating genetic resources to more economically and more powerfully understand the effects of genes on complex diseases. However, our findings add to a familiar refrain about GWA studies – that every step must be done with extreme care to avoid spurious results (McCarthy et al. 2008). More work needs to be done to determine the best approaches for combining cases and controls obtained from different sources. In any case-control study, cases and controls should be comparable, and recent studies have discussed how to control for differential population substructure when using publicly available controls (Zhuang et al. 2010; Luca et al. 2008). Our work emphasizes the need to control for technical errors caused by integrating data from different chips. Researchers attempting to use the sort of data we describe, in which cases and controls are genotyped on different chips, need to be aware of the high potential for false positives after imputation, and must guard against it or control for it. In particular, it is vitally important to technically validate any SNPs that appear significant before reporting them, by regenotyping those SNPs on an independent platform – considered best practice in any GWA study, it is all the more important here where the chance of false positive results due to differential imputation is so high.

Acknowledgments

We would like to thank Constance Chen and Marilyn C. Cornelis for their assistance with programming. JAS was supported by the National Institutes of Health (NIH) grant T32 GM074897. PK was supported by the NIH grant U01 CA098233. The T2D GWAS was funded by NIH grant U01 HG004399 as part of the Gene Environment-Association Studies (GENEVA) under the NIH Genes, Environment and Health Initiative (GEI)

References

Alberts B. Editorial expression of concern. Science. 2010; 330(6006):912.10. 1126/science. 330.6006.912-b [PubMed: 21071647]

Beecham GW, Martin ER, Gilbert JR, Haines JL, Pericak-Vance MA. APOE is not associated with alzheimer disease: a cautionary tale of genotype imputation. Ann Hum Genet. 2010; 74(3):189–94.10.1111/j.1469-1809.2010.00573.x [PubMed: 20529013]

Carmichael, M. The little flaw in the longevity-gene study that could be a big problem. 2010. URL http://www.newsweek.com/2010/07/07/the-little-flaw-in-the-longevity-gene-study-that-could-be-a-big-problem.html

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004. [PubMed: 11315092]

Fallin MD, Szymanski M, Wang R, Gherman A, Bassett SS, Avramopoulos D. Fine mapping of the chromosome 10q11-q21 linkage region in Alzheimer's disease cases and controls. Neurogenetics. 2010; 11(3):335–48.10. 1007/s10048-010-0234-9 [PubMed: 20182759]

Ho LA, Lange EM. Using public control genotype data to increase power and decrease cost of case-control genetic association studies. Hum Genet. 2010; 128(6):597–608.10.1007/s00439-010-0880-x [PubMed: 20821337]

Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S, Lee AT, Chung SA, Ferreira RC, Pant PVK, Ballinger DG, Kosoy R, Demirci FY, Kamboh MI, Kao AH, Tian C, Gunnarsson I, Bengtsson AA, Rantapaa-Dahlqvist S, Petri M, Manzi S, Seldin MF, Ronnblom L, Syvanen AC, Criswell LA, Gregersen PK, Behrens TW. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med. 2008; 358(9):900–909.10.1056/NEJMoa0707865 [PubMed: 18204098]

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5(6):e1000, 529.10.1371/journal.pgen.1000529

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JFJ, Hoover RN, Thomas G, Chanock SJ. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007; 39(7):870–874.10.1038/ng2075 [PubMed: 17529973]

Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10:387–406.10.1146/annurev.genom.9.081307.164242 [PubMed: 19715440]

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34(8):816–834.10.1002/gepi.20533 [PubMed: 21058334]

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet. 2008; 82(2):453–63.10.1016/j.ajhg.2007.11.003 [PubMed: 18252225]

Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010; 11(7):499–511.10.1038/nrg2796 [PubMed: 20517342]

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9(5):356–369.10.1038/nrg2344 [PubMed: 18398418]

Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC. Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered. 2006; 61(1):55–64.10.1159/000092553 [PubMed: 16612103]

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2(12):e190.10.1371/journal.pgen.0020190 [PubMed: 17194218]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–909.10.1038/ng1847 [PubMed: 16862161]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–575.10.1086/519795 [PubMed: 17701901]

Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Pare G, Sun Q, Girman CJ, Laurie CC, Mirel DB, Manolio TA, Chasman DI, Boerwinkle E, Ridker PM, Hunter DJ, Meigs JB, Lee CH, Hu FB, van Dam RM. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet. 2010; 19(13):2706–2715.10.1093/hmg/ddq156 [PubMed: 20418489]

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2009. URL http://www.R-project.org

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science. 2007; 316(5829):1341–1345.10.1126/science.1142382 [PubMed: 17463248]

Sebastiani P, Solovieff N, Puca A, Hartley S, Melista E, Andersen S, Dworkis D, Wilk J, Myers R, Steinberg M, Montano M, Baldwin C, Perls T. Genetic signatures of exceptional longevity in humans. Science. 201010.1126/science.1190532

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–678.10.1038/nature05911 [PubMed: 17554300]

Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S, Giannini C, Halder C, Kollmeyer TM, Kosel ML, LaChance DH, McCoy L, O'Neill BP, Patoka J, Pico AR, Prados M, Quesenberry C, Rice T, Rynearson AL, Smirnov I, Tihan T, Wiemels J, Yang P, Wiencke JK. Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. Nat Genet. 2009; 41(8):905–908.10.1038/ng.408 [PubMed: 19578366]

Zhuang JJ, Zondervan K, Nyberg F, Harbron C, Jawaid A, Cardon LR, Barratt BJ, Morris AP. Optimizing the power of genome-wide association studies by using publicly available reference samples to expand the control group. Genet Epidemiol. 2010; 34(4):319–326.10.1002/gepi.20482 [PubMed: 20088020]

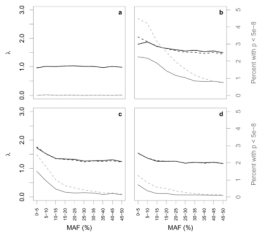

Fig. 1.

Black and gray lines represent λ values and the percentages of p-values less than 5×10^{-8}, respectively, for SNPs grouped by minor allele frequency (MAF) in four settings: **a** SNPs genotyped on both Affy and Illumina platforms; **b** SNPs genotyped on Illumina platform and imputed for the Affy controls; **c** SNPs genotyped on Affy platform and imputed for the Illumina controls; and **d** SNPs imputed for both groups. Solid lines are from soft call analysis and dashed lines are from hard call analysis. Note that in some places (particularly in panel **a**) the solid and dashed lines are indistinguishable because the results from the soft call and hard call analyses were very similar.

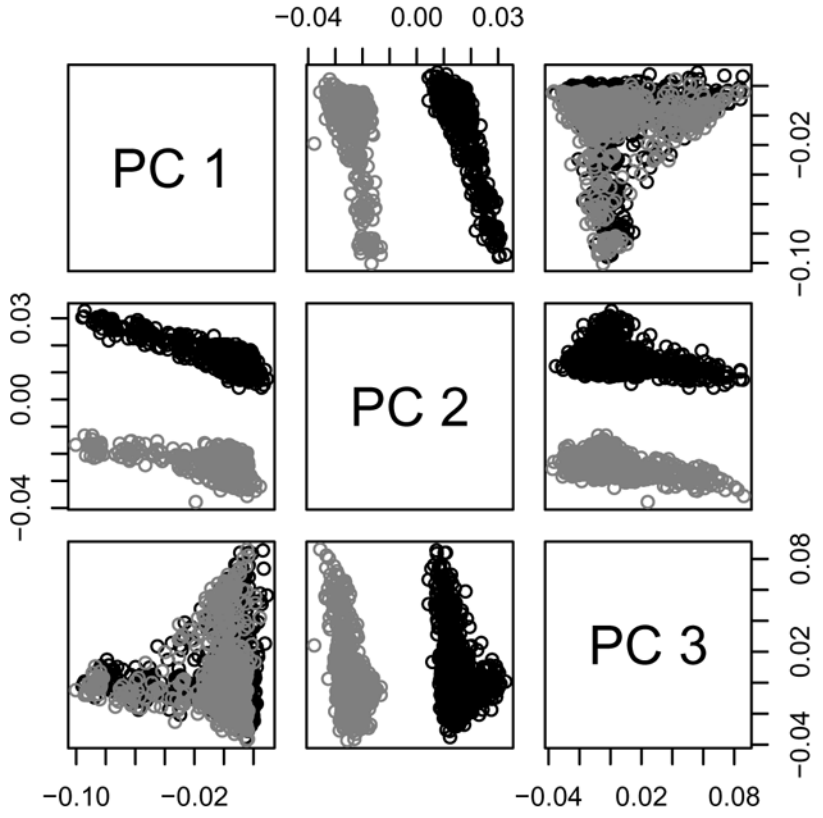

Fig. 2.
Top 3 principal components (PCs), among SNPs genotyped in the Illumina controls and imputed using hard calls in the Affy controls, plotted against one another. Affy samples are plotted in black; Illumina samples are plotted in gray.

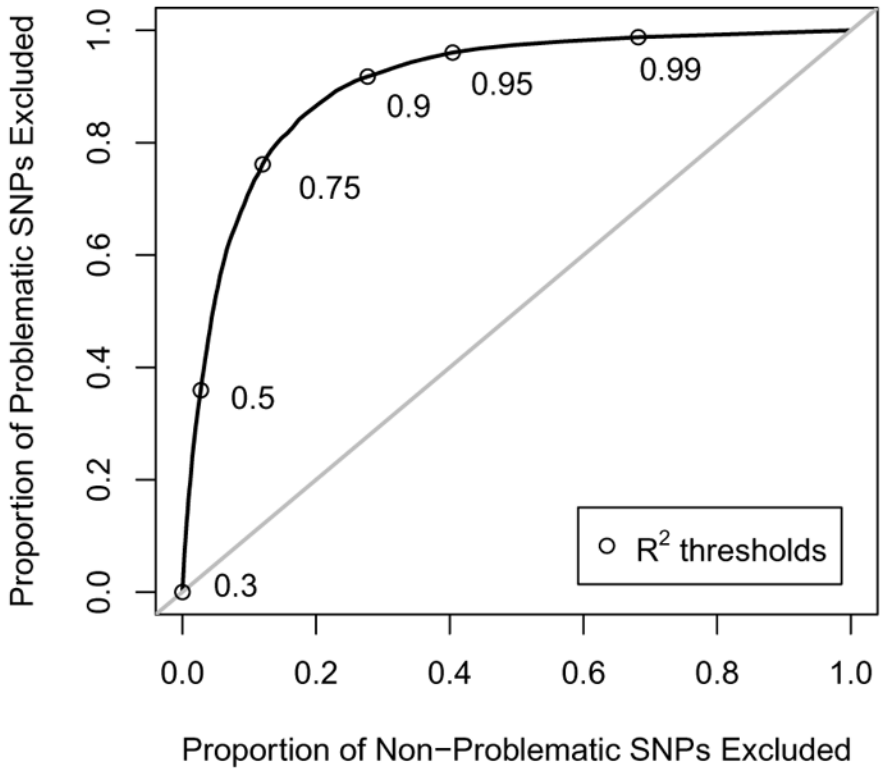

Fig. 3.

Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, discrimination of the R^2 criterion described in Method 2, as the R^2-threshold varies. The *y*-axis is the sensitivity, the proportion of highly significant SNPs which are excluded; the *x*-axis is 1–specificity, the proportion of non-significant SNPs which are excluded. R^2 threshold choices between 0.3 and 0.99 are pointed out along the curve.

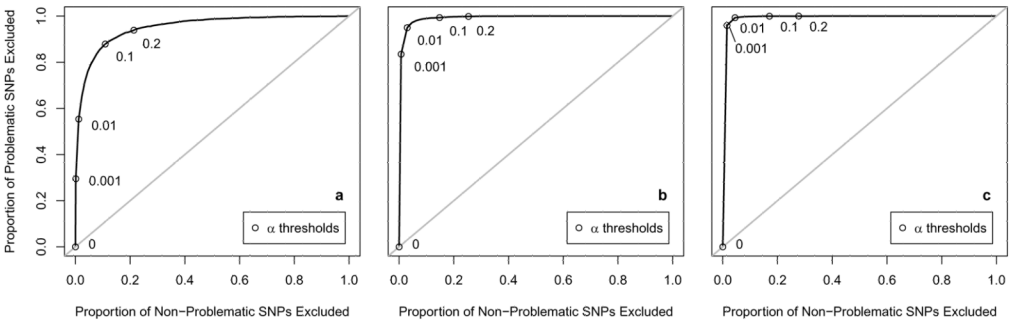

Fig. 4.

Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, discrimination of the preliminary screening criterion described in Method 3, as the α-threshold varies. The *y*-axis is the sensitivity, the proportion of highly significant SNPs which are excluded; the *x*-axis is 1– specificity, the proportion of non-significant SNPs which are excluded. Plots shown are for **a** $n = 100$, **b** $n = 300$, and **c** $n = 500$ additional controls. α threshold choices between 0.001 and 0.2 are pointed out along the curves.

Table 1

Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, values of λ and percentages of SNPs with $p < 5\times10^{-8}$ when we restrict to SNPs with imputation quality R^2 larger than the given thresholds, as detailed in Method 2. Also listed are the percentages of total SNPs remaining for analysis at each threshold.

	R^2 threshold					
	0.3	**0.5**	**0.75**	**0.9**	**0.95**	**0.99**
λ	1.6	1.6	1.4	1.2	1.1	1.04
% SNPs with $p < 5\times10^{-8}$	1.3	0.87	0.36	0.15	0.09	0.05
% SNPs remaining for analysis	100	97	87	71	59	31

Hum Genet. Author manuscript; available in PMC 2013 January 1.

Table 2

Results from the main analysis comparing 1000 Affy cases and 1000 Illumina controls, among SNPs remaining after a preliminary screen in which we compare n = 100, 300, or 500 additional controls on Affy to 1000 controls on Illumina and remove SNPs significant at level α, as detailed in Method 3. Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, values of λ and percentages of SNPs with $p < 5 \times 10^{-8}$ are presented, along with the percentages of total SNPs remaining for analysis at each threshold.

		α threshold				
		0	**0.001**	**0.01**	**0.1**	**0.2**
100 additional controls	λ	1.5	1.5	1.5	1.4	1.3
	% SNPs with $p < 5 \times 10^{-8}$	0.83	0.59	0.38	0.11	0.07
	% SNPs remaining for analysis	100	100	98	88	78
300 additional controls	λ	1.5	1.5	1.4	1.2	1.2
	% SNPs with $p < 5 \times 10^{-8}$	0.83	0.14	0.04	0.01	0.002
	% SNPs remaining for analysis	100	99	96	85	74
500 additional controls	λ	1.5	1.4	1.4	1.2	1.1
	% SNPs with $p < 5 \times 10^{-8}$	0.83	0.03	0.01	0.001	0.0004
	% SNPs remaining for analysis	100	98	95	82	72