Genomics & Informatics Vol. 7(2) 136-140, June 2009

A Scheme for Filtering SNPs Imputed in 8,842 Korean Individuals Based on the International HapMap Project Data

Kichan Lee and Sangsoo Kim*

Department of Bioinformatics & Life Science, Soongsil University, Seoul 156-743, Korea

Abstract

Genome-wide association (GWA) studies may benefit from the inclusion of imputed SNPs into their dataset. Due to its predictive nature, the imputation process is typically not perfect. Thus, it would be desirable to develop a scheme for filtering out the imputed SNPs by maximizing the concordance with the observed genotypes. We report such a scheme, which is based on the combination of several parameters that are calculated by PLINK, a popular GWA analysis software program. We imputed the genotypes of 8,842 Korean individuals, based on approximately 2 million SNP genotypes of the CHB+JPT panel in the International HapMap Project Phase II data, complementing the 352k SNPs in the original Affymetrix 5.0 dataset. A total of 333,418 SNPs were found in both datasets, with a median concordance rate of 98.7%. The concordance rates were calculated at different ranges of parameters, such as the number of proxy SNPs (NPRX), the fraction of successfully imputed individuals (IMPUTED), and the information content (INFO). The poor concordance that was observed at the lower values of the parameters allowed us to develop an optimal combination of the cutoffs (IMPUTED≥0.9 and INFO≥0.9). A total of 1,026,596 SNPs passed the cutoff, of which 94,364 were found in both datasets and had 99.4% median concordance. This study illustrates a conservative scheme for filtering imputed SNPs that would be useful in GWA studies.

Keywords: genome-wide association, HapMap, PLINK, SNP imputation

Introduction

Genome-wide association (GWA) studies that employ ultrahigh-density microarray chips and over several thousand samples have been successful in mapping loci that are associated with diseases or epidemiological traits. However, in the course of the analyses, these studies have generated a large number of false positives, due to the multiple-testing nature in the statistical analysis or some bias in sampling. If multiple datasets of the same trait in different samples are available, the statistical power can be greatly improved by combining the studies through meta-analysis (de Bakker *et al.*, 2008). On the other hand, the results from a single dataset are typically validated via expensive replication studies. It would be of great help if the associations could be corroborated, based on some other information, such as linkage disequilibrium, that is readily available prior to the replication studies.

Linkage disequilibrium (LD), the nonrandom association of alleles at different loci, is the result of the evolutionary history of a population, involving mutations, selection, recombination, population bottlenecks, and random genetic drift (Xiong & Jin 2007). A haplotype is a set of co-occurring polymorphic alleles on the same chromosome. The haplotype block model has particular implications in the study of dense markers, such as the single nucleotide polymorphism (SNP), because it implies that a smaller number of markers (tagging SNPs) are necessary to uniquely distinguish different haplotypes. On the other hand, if a marker shows a strong association with a trait, then the other markers within the same haplotype block should show the same association. This idea has been implemented in GWA analysis software programs, including PLINK (Purcell *et al.*, 2007), in various ways: direct association of each haplotype, proxy association, or LD-based clumping. These approaches require an accurate haplotype model. The quality of the haplotype models in a particular GWA dataset can be improved by phasing the haplotypes, based on the background genotypes of the ultra-high-density International HapMap data.

The International HapMap Project aims to produce such valuable haplotype information of the human genome (The International HapMap Consortium 2003). A total of 270 samples, 90 samples from each of three major ethnic groups-African, Asian, and European-were genotyped, phased, and released to the public freely (Thorisson *et al.*, 2005). The second phase of the project released the genotypes for over 2 million SNPs. The genotypes of the markers that were not included in the study dataset but were included in the reference dataset, the International HapMap dataset, can be inferred if

*Corresponding author: E-mail sskimb@ssu.ac.kr
Tel +82-2-820-0457, Fax +82-2-824-4383
Accepted 10 May 2009

the surrounding haplotypes of the study dataset are compatible with those of the reference. This is called imputation, because it is similar to filling in missing information in statistical analysis. Due to its predictive nature, imputation is not perfect and is bound to generate some errors. Hence, we need a heuristic scheme to reduce the error. One may estimate the imputation error by calculating the concordance of the genotypes of a marker in the study dataset with those that are imputed for the same individual.

Recently, a GWA study of quantitative traits with 8,842 Korean individuals (Korea Association Resource (KARE)) was reported (Cho *et al.*, 2009). In that study, SNP imputing and a subsequent association study were carried out with IMPUTE (Marchini *et al.*, 2007) and SNPTEST (Marchini *et al.*, 2007), respectively, and no filtering scheme was reported. For the same dataset, we imputed HapMap SNPs using PLINK; evaluated the concordance rate for various ranges of the parameters that were reported by PLINK, a popular GWA analysis software; and developed heuristic cutoffs of those parameters for filtering out potentially poorly imputed markers.

Methods

Genotype and HapMap data

The genotype data that were used in this study were previously reported. Briefly, after standard quality control steps, a total of 352,228 SNPs for 8842 individuals were obtained. We called it the KARE (Korean Association Resource) dataset. A total of 351,677 SNP IDs of the dataset were converted from those of Affymetrix to those of dbSNP RefSNP IDs using an annotation file (Mapping250K_Nsp.na26.annot.csv) that was downloaded from the Affymetrix website. Among them, 176,059 underwent strand flipping in order to conform to the +ve strand of NCBI human genome build 36. The genotypes of the JPT+CHB panel of International HapMap Phase II (The International HapMap Consortium 2005) were downloaded from the PLINK website (http://pngu.mgh.harvard.edu/~purcell/plink/dist/hapmap_JPT_CHB_r23a_filtered.zip). The dataset contains 2.2 million markers that had been filtered for those that had a MAF greater than 0.01 and a genotyping rate in the CEU panel greater than 95%. These two datasets were subsequently merged into a single set, followed by a split into each chromosome for parallel job submissions.

SNP imputation

PLINK, one of the most popular GWA analysis software programs, was used for SNP imputation. It works by phasing the haplotypes of the reference panel and using it to infer the alleles of the study panel. For phasing haplotypes, PLINK uses an EM algorithm. The SNPs of the study panel that are in LD with the imputed SNP are called proxies. If the genotypes of these proxies in some individuals in the study panel are not compatible with any of the haplotypes of the reference panel, the genotypes for these individuals cannot be imputed. PLINK reports the number of individuals who have successful imputation for each marker. The quality of imputation is often estimated by the variance of the imputed allele dosage relative to that expected from the reference panel. PLINK reports this parameter as INFO (de Bakker *et al.*, 2008). PLINK also reports the concordance rate for each maker in the study panel . All of the PLINK jobs were performed on the 128-CPU cluster at Korean Bioinformation Center (http://www.kobic.re.kr). The results from PLINK runs were analyzed using locally developed R scripts.

Results

The KARE genotype data were merged with those of the International HapMap Phase II JPT+CHB panel, resulting in a total of 2,168,896 SNPs (CJK dataset). The genotypes of the KARE data were imputed using PLINK, based on the background haplotypes of the reference panel (the PLINK option "--proxy-impute all"). Among 351,766 SNPs in the KARE dataset (called "OBSERVED"), 333,418 were also found in the reference panel (called "OVERLAP"). The concordance rates for these SNPs were calculated, wherein the first quartile and median were 0.964 and 0.987, respectively. We surveyed the concordance rates at various intervals of key parameters that were reported by PLINK. First, we examined the trend by the Hardy-Weinberg equilibrium (HWE) p value (P_{HWE}). Fig. 1A shows the boxplot of the concordance rates at each bin of -log$_{10}$ (P_{HWE}). The interquartile range (IQR) that corresponds to the vertical dimension of the box increases as the deviation from HWE increases. While the third quartile remains virtually the same, the first quartile drops significantly with the increased deviation from HWE. The combination of IQR and the first quartile is represented by the lower "whisker," which also shows a dramatic decrease at severe deviation from HWE. At the rightmost bin, where $P_{HWE} < 10^{-7}$, the concordance was extremely poor for 702 OVERLAP markers in this bin. We decided to exclude markers that had $P_{HWE} < 10^{-6}$, where the lower "whisker" of the concordance rate dropped substantially. It should be noted that the original KARE analysis also used a similar HWE cutoff (Cho *et al.*, 2009).

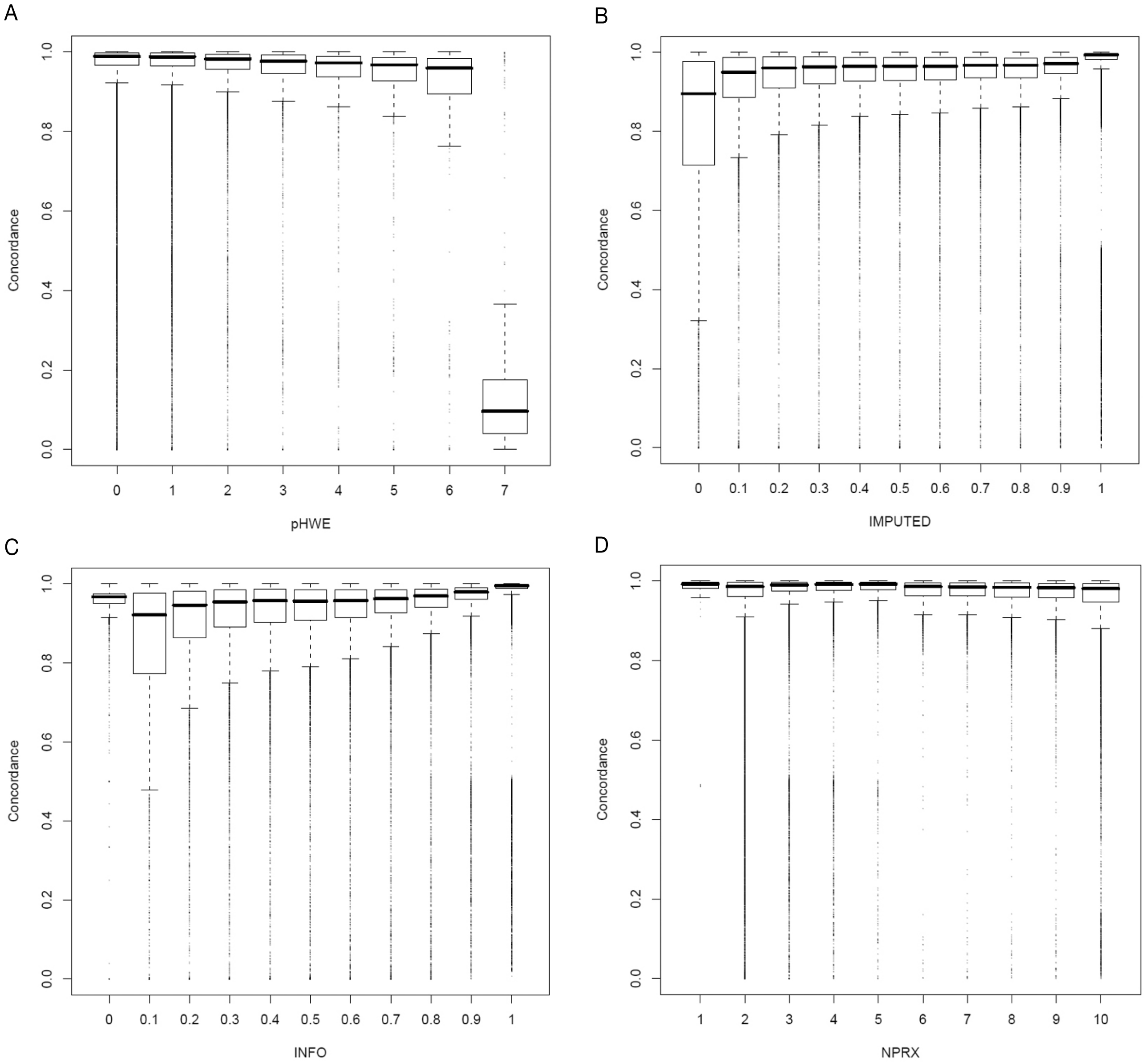

Fig. 1. Distribution of the concordance rates at each bin of the parameters reported by PLINK. The concordance between the observed genotypes and imputed ones are viewed via boxplots for the bins of (A) $-\log_{10}$ (P_{HWE}), where P_{HWE} is the Hardy-Weinberg Equilibrium p value; (B) IMPUTED, the fraction of imputed individuals in the KARE panel; (C) INFO, the relative variance of allele frequency of the imputed alleles; and (D) NPRX, the number of proxy SNPs for each observed marker.

The imputation process in PLINK is based on the haplotype that is formed by the neighboring markers within the LD block (so-called "proxies"). By comparing the genotypes of the proxies from the study panel (KARE, in this case) with those of the reference panel (HapMap), the compatible haplotypes are chosen and the alleles are assigned. If no compatible haplotypes are found for some of the individuals, their alleles cannot be assigned. PLINK provides a parameter called IMPUTED, which is the fraction of individuals whose genotypes were able to be imputed. It is implied that the haplotype structures of the study panel are less compatible with those of the reference panel as IMPUTED drops below 1. With the increase of the incompatible fraction, the concordance is expected to drop. The concordance rate at each bin of IMPUTED is summarized in Fig. 1B. While the concordance at the rightmost bin was excellent, substantial deterioration in concordance was noticed as IMPUTED dropped. The relative variance of the imputed alleles of a marker is given as a parameter called INFO

Table 1. Distribution of concordance rates for various cutoffs

Cutoff	Imputed[a]	Overlap[b]	Lower whisker[c]	1st quartile[c]	Median[c]	3rd quartile[c]	Higher whisker[c]
None	2,168,896	333,418	0.916	0.964	0.987	0.996	1.000
$P_{HWE} > 10^{-6}$	1,910,471	332,631	0.916	0.964	0.987	0.996	1.000
$P_{HWE} > 10^{-6}$, IMPUTED ≥ 0.9	1,323,616	229,716	0.952	0.979	0.992	0.997	1.000
$P_{HWE} > 10^{-6}$, INFO ≥ 0.9	1,098,100	194,492	0.965	0.984	0.993	0.997	1.000
$P_{HWE} > 10^{-6}$, IMPUTED ≥ 0.9, INFO ≥ 0.9	1,026,596	94,364	0.966	0.985	0.994	0.998	1.000

[a] Number of all the SNP markers imputed.

[b] Number of markers overlapping between KARE and HapMap Phase II JPT+CHB datasets.

[c] Boxplot statistics of the concordance rates.

(de Bakker *et al.*, 2008). Similar to IMPUTED, INFO is also expected to show the compatibility in the haplotype structures of the two panels. Indeed, we noticed significant deterioration in concordance at lower INFO values (Fig. 1C). One might anticipate that the more proxies that are used, the more reliable the imputation is. Interestingly, the concordance was somewhat poorer with more proxies (Fig. 1D). Currently we have no explanation for this, except that it might manifest increased errors in haplotype phasing by the EM algorithm for cases that have many proxies.

The monotonous change in the concordance rate along these parameters allows development of cutoffs for filtering out poorly imputed markers. Application of these cutoffs inevitably loses some well-imputed markers and retains some poorly imputed ones. A combination of the cutoffs is expected to improve the overall concordance rate even more. Table 1 shows the concordance statistics at various cutoffs. Introduction of a cutoff that was based on P_{HWE} alone did not demonstrate the improvement in the statistics, because the number of markers that were removed among OVERLAP was merely 787 (0.2%). However, it removed approximately 10% of all of the imputed markers. An additional condition of IMPUTED ≥ 0.9 dramatically improved the lower whisker from 0.916 to 0.952, at the expense of about 1/3 of markers. The cutoff that was based on INFO had an even higher impact. A combination of all three cutoffs showed the most improvement, although the statistics themselves improved marginally. One might argue that IMPUTED and INFO are correlated and that only one of them would be necessary. On the contrary, the combination of all three cutoffs retained only about 1/3 of the original markers. If both parameters had been well correlated, we would not have seen such a remarkable reduction in surviving markers. In fact, the correlation coefficient between INFO and IMPUTED for the 1,026,596 markers that survived the combined cutoffs was 0.41353.

Discussion

This study intended to evaluate the performance of PLINK in imputing SNPs for the KARE population based on International HapMap Phase II JPT+CHB. The imputation performance was measured in terms of the concordance between the observed and imputed genotypes. Correlations between the concordance rate and several parameters that were reported by PLINK were noticed, allowing us to develop a heuristic cutoff that reasonably improved the concordance rate. However, the application of such a cutoff inevitably caused the elimination of nearly half of the markers. If all of the HapMap SNPs had been retained and well imputed, the marker density would have been increased by 6-fold from the original KARE marker density. Consequently, the increase in the marker density at this filtering was only 3-fold, which is still substantially denser than the original dataset. Such an expanded dataset would allow us to analyze the haplotype structure around a marker that shows an association with traits. The subsequent analysis, based on haplotype association, would be a powerful tool, filtering out unlikely association signals that might be caused by some non-genotyping errors. Stable association of the haplotype that encompasses the original association signal from the unimputed dataset would corroborate that original association. Another application of the imputed dataset would be to discover an imputed marker, if any, that is nearby the original one and shows a stronger association than it. However, any association signal from the imputed markers should be scrutinized carefully, because they are not directly observed experimentally and the imputation process cannot be perfect.

It should be noted that the imputation work that is presented here has some limitations. The fact that half of the imputed SNPs had to be filtered out indicates a substantial difference in haplotype structures between the HapMap and KARE population. It would not emphasize the ethnic differences between Chinese, Japanese, and Koreans, because Chinese and Japanese are genetically more distant from each other than from Koreans (Ahn *et al*., 2009). The haplotype difference may be attributed to the difference in sample size. Merging 45 Chinese and Japanese each to form 90-individual genotype data may not capture the full spectra of haplotypes that are present in 8842 individuals of the KARE data. In other words, the haplotype structures that are formed by the HapMap JPT+CHB would be accurate, considering the extremely high-density SNP markers. Although they may capture major haplotypes in an East Asian population, they may miss some low-frequency haplotypes that are present in the East Asian population, due to limited sample size. Further study that compares the haplotype structures of these two datasets would clarify this point.

Acknowledgments

The KARE genotype and epidemiological data were gratefully made available by the National Institute of Health, Korea Center for Disease Control, which supported this work through a grant to the KARE Analysis Consortium. SK would like to thank the help and support of all of the staff at KNIH. The computing cluster that was used in this study was kindly granted from the Korean Bio-Information Center (KOBIC).

References

Ahn, S.M., Kim, T.H., Lee, S., *et al*. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* published in advance.

de Bakker, P.I.W., Ferreira, M.A.R., Xioming, J., Neale, B.M., Raychaudhuri, S., and Voicht, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 17, R122-128.

Cho, Y.S., Go, M.J., Kim, Y.J., *et al*. (2009). A large-scale genome-wide association study of Asian populations uncover genetic factors influencing eight quantitative traits. *Nat. Genet.* 41, 527-534.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906-913.

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

The International HapMap Consortium. (2003). The International HapMap Project. *Nature* 426, 789-796.

The International HapMap Consortium. (2005). A Haplotype Map of the Human Genome. *Nature* 437, 1299-1320.

Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. (2005). The International HapMap Project Web site. *Genome Res.* 15, 1591-1593.

Xiong, M., and Jin, L. (2007). Association Studies of Complex Diseases. In *Bioinformatics - From Genomes to Therapies* Vol. 3, T. Lengauer, ed. (Wiley-VCH, Germany), pp.1375-1426