The Effect of Reference Panels and Software Tools on Genotype Imputation

Kwangsik Nho, PhD[1,2], Li Shen, PhD[2,3], Sungeun Kim, PhD[2,3], Shanker Swaminathan, BTech[2,4], Shannon L. Risacher, BS[2], Andrew J. Saykin, PsyD[2,3,4], and the Alzheimer's Disease Neuroimaging Initiative (ADNI)

[1]Regenstrief Institute and Indiana University School of Medicine, Indianapolis, IN
[2]Center for Neuroimaging, Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN
[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN
[4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN

Abstract

Genotype imputation is increasingly employed in genome-wide association studies, particularly for integrative and cross-platform analysis. Several imputation algorithms use reference panels with a larger set of genotyped markers to infer genotypes at ungenotyped marker locations. Our objective was to assess which method and reference panel was more accurate when carrying out imputation. We investigated the influence of choice of two most popular imputation methods, IMPUTE and MACH, on two reference panels from the HapMap and the 1000 Genomes Project. Our results indicated that for the HapMap, MACH consistently yielded more accurate imputation results than IMPUTE, while for the 1000 Genomes Project, IMPUTE performed slightly better. The best imputation results were achieved by IMPUTE with the combined reference panel (HapMap + 1000 Genomes Project). IMPUTE with the combined reference panel is a promising strategy for genotype imputation, which should facilitate fine-mapping for discovery as well as known disease-associated candidate regions.

Introduction

Due to the advance in low-cost and high-throughput genotyping techniques, genome-wide association studies (GWAS) have successfully identified numerous susceptibility loci strongly associated with a trait or disease of interest.[1] Recently, GWAS have been widely employed for discovery of novel disease loci in very large sets of cases and controls, such as Alzheimer's Disease(AD).[2-3] However, many single nucleotide polymorphisms (SNPs) identified from previous studies were not replicated even in these large-scale studies. Thus, GWAS still needs larger sample sizes to identify and replicate genetic variation of modest effects at adequate power. Consequently, combining genotype data from multiple studies is one way to increase sample sizes and thus the detection power of GWAS. However, GWAS projects have usually used different genotyping platforms containing distinct sets of markers. When combining the results across two or more studies that have different sets of genetic markers, it is more powerful to combine the genotype data from all studies and then analyze them together than to simply investigate the top results from each individual study. In addition, in the fine-mapping of known disease-associated regions, having denser genotype data by imputing additional ungenotyped markers in the same regions can allow one to localize the disease-associated regions more precisely, and some of the candidate SNPs identified by a pathway analysis are often ungenotyped in a given study and need to be imputed for further analysis. As a result, genotype imputation methods have increasingly become popular.[4]

Recently, many imputation methods have been proposed, and their performance has been compared by investigating the effect of linkage disequilibrium (LD; defined formally below), minor allele frequency (MAF), and reference population on imputation accuracy rate.[5-7] MACH (www.sph.umich.edu/csg/abecasis/MACH)[8] and IMPUTE (https://mathgen.stats.ox.ac.uk/impute/impute.html)[9] produced similar accurate results and these two methods are consistently superior to other methods. These imputation algorithms use a reference panel with very dense genotyping to effectively impute genotypes at unobserved markers by using the pattern of LD in the reference panel. A reference panel consists of a number of individuals genotyped at all markers of interest. To date, most imputation-based association studies have been conducted using diverse reference population samples only from the HapMap (www.hapmap.org) as reference panels.[10] Very recently, the 1000 Genome Project (www.1000genomes.org) released high quality genotype data with denser markers which may help yield improved imputation results.

Furthermore, IMPUTE can use a combined set of haplotypes from both the HapMap and the 1000 Genomes Project as a reference panel for a single imputation. The choice of a reference panel may play an important role in influencing the accuracy of imputation methods. Therefore, key components for a successful imputation include not only a promising imputation method but also an appropriate reference panel.

In this study, our goal was to examine two highly popular genotype imputation software packages, IMPUTE v2 and MACH v1, by investigating their performances using two independent reference panels from the HapMap and the 1000 Genomes Project. It is necessary to assess the influence of choice of reference panels on imputation accuracy rate in order to address which combination of the method and the reference panel is optimal for imputation prior to GWAS.

Methods

Study samples

822 participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) (www.adni-info.org) were used in this study. The ADNI was launched in 2004 to help researchers and clinicians develop new treatments for MCI (mild cognitive impairment) and early AD, monitor their effectiveness, and decrease the time and cost of clinical trials. Neuroimaging and biological markers were used to achieve the goal of the ADNI study.

This multi-year multi-site longitudinal study was started by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations. The ADNI participants consist of AD, MCI, and elderly healthy control individuals. They were aged 55-90 years and recruited from 59 sites across the U.S. and Canada. Written informed consent was obtained from all 822 participants and the study was conducted with prior Institutional Review Boards approval.

For the clinical diagnosis of AD, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria, mini mental state examination (MMSE) scores, and clinical dementia rating (CDR) were used. Demographic information, *APOE* genotype, neuropsychological test scores, and diagnosis were downloaded from the ADNI database (www.loni.ucla.edu\ADNI).

Genotyping

A majority of ADNI participants (818 out of 822) were genotyped using the Illumina Human610-Quad BeadChip, which contains 620,901 markers. Their genotyping was conducted as described previously.[11-12] ADNI genotyping data are publicly available at the ADNI database.

Quality Control

We performed standard quality control procedures for genetic markers and subjects using PLINK v1.06 (pngu.mgh.harvard.edu/~purcell/plink).[13] All the copy number variation (CNV) markers were excluded. SNPs were excluded using the following marker exclusion criteria: (1) call rate ≤ 90%, (2) minor allele frequency (MAF) ≤ 5% , and (3) Hardy-Weinberg equilibrium (HWE) test $P < 1 \times 10^{-6}$ in healthy control participants only.[11-12]

After excluding markers, we removed participants with overall genotyping call rates ≤ 90% and then compared the gender in the clinical database with the gender determined through the heterozygosity of SNPs on the X chromosome to exclude gender mismatches from the analysis. We evaluated the pair-wise identity by descent (IBD) for all subjects to identify pairs with estimated proportional IBD > 0.125 and removed one subject from each pair who appeared to be relatives closer than first-cousin or sample duplications. Since population stratification is known to cause spurious association in disease studies, we restricted our analyses to non-Hispanic Caucasian participants.[11-12]

Consequently, 733 individuals and 530,992 SNPs passed all quality control tests, and the total genotyping rate in the remaining subjects was > 99.5%.

Imputation Methods

The present study investigated two of the most widely used software packages: MACH v1[8] and IMPUTE v2[9].

MACH[8] implements a Markov chain-based algorithm to infer possible pairs of haplotypes for each individual's genotypes and accurately impute missing genotypes on the basis of LD information. MACH works by successively updating the phase of each individual's genotype data conditional on the current haplotype estimates of all the other individuals. It carries out a two-stage procedure: it first estimates unknown parameters to be used in the second stage

using a subset of individuals and then carries out genotype imputation based on the first-stage maximum-likelihood estimates of the crossover map and the error rate map. MACH produces several output files containing the dosages of reference allele and the posterior probability for the most likely genotype at each marker for each individual.[8]

IMPUTE[9] employs a hidden Markov chain Monte Carlo method to compare the set of genotype for each individual in the given study dataset to the reference haplotypes. SNPs are first divided into two sets: a set T that is genotyped in both the study sample and reference panel, and a set U that is ungenotyped in the study sample but genotyped in the reference panel. IMPUTE produces posterior probabilities of missing genotypes by estimating haplotypes at SNPs in T and then imputing alleles at SNPs in U conditional on the current estimated haplotypes.[9]

We determined discrete imputed genotypes using the posterior probabilities obtained from MACH and IMPUTE: for a given individual, imputed genotypes at given marker loci with posterior probabilities greater than a threshold value were accepted, otherwise classified as missing.

Reference Population

MACH and IMPUTE use a reference panel with very dense genotyping to compare the potential haplotypes for each individual in a given study with all other observed haplotypes from the reference panel. A reference panel consists of a number of individuals genotyped at all markers of interest. For this purpose, the HapMap data (www.hapmap.org) have successfully been employed in most imputation-based studies. Very recently, sets of haplotypes from the pilot phase of the 1000 Genomes Project (www.1000genomes.org) have been made available.

The reference panel should be representative of the study sample population. We used non-Hispanic Caucasian participants in this study. Therefore, we used the CEU (Utah residents with northern and western European ancestry from the CEPH collection) panel of HapMap3 release 2 data and the CEU panel of the pilot 1 data of the 1000 Genomes Project as reference panels for inferring missing genotypes. The CEU panel of HapMap 3 release 2 has 234 haplotypes in phased files and about 1.39 million SNPs, and the CEU panel of the pilot 1 data of the 1000 Genomes Project has 120 haplotypes and about 7.9 million SNPs. In particular, IMPUTE can use a set of combined haplotypes from both the pilot 1 data of the 1000 Genomes Project and the HapMap3 data as a reference panel for a single imputation. Henceforth, we refer the CEU panel of HapMap 3 release 2 as HM3CEU, and the CEU panel of the pilot 1 data of the 1000 Genomes Project as G1KCEU.

Imputation Performance

MACH and IMPUTE produce the posterior probabilities of the imputed genotypes at ungenotyped marker loci for each individual. In order to assess the quality of imputation, we determined discrete imputed genotypes by accepting an imputed genotype if its posterior probability reached a pre-specified threshold or classifying it as missing otherwise. To make comparisons about imputation accuracy across MACH and IMPUTE using two different reference datasets, HM3CEU and G1KCEU, we used 7,991 SNPs from chromosome 22 of the ADNI data to reduce computation time. First, 300 genotyped SNPs were selected on chromosome 22 by picking a SNP every 26 SNPs among the 7,991 SNPs. These 300 SNPs were then removed and subsequently imputed for all 733 subjects. The total genotyping rate of the 300 SNPs is 0.9954. Imputation accuracy was calculated as the concordance rate between the imputed and observed genotypes.

We calculated the dependence of imputation accuracy rates on a weighted average of linkage disequilibrium (LD) and a SNPs' minor allele frequency (MAF). Linkage disequilibrium (LD) is defined as the nonrandom association of alleles of SNPs residing near one another on a chromosome and D' denotes the measure of LD.[1] We determined a weighted average of the pairwise D' between each imputed SNP and all other SNPs in the same chromosome with weights as

$$wD' = \sum_{i=1}^{n} D_i' \exp(-d_i),$$

where d_i is the physical distance between the imputed SNP and the ith SNP in 100 kb, and D_i' is the estimate of LD between the same two SNPs.[14]

Results

In Figure 1(a,b), we present the dependence of imputation accuracy rates on a weighted average of LD at the default threshold value (qc=0.9), where we impose a posterior probability equal to 0.90 as a threshold value to accept the imputed genotypes.

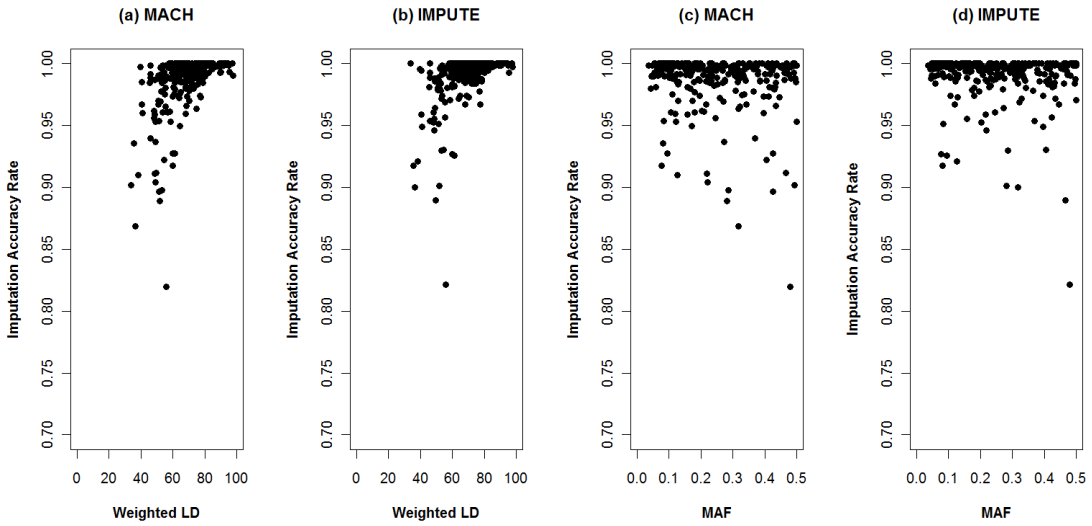

Fig. 1 Imputation accuracy rate as a function of a weighted average of linkage disequilibrium (LD) ((a) MACH and (b) IMPUTE) and minor allele frequency (MAF) ((c) MACH and (d) IMPUTE) at the default confidence threshold value (qc=0.9) using HapMap 3 CEU population.

With a reference panel (HM3CEU), MACH and IMPUTE were used to impute SNPs that are not genotyped in the sample but that are genotyped in the reference panel. As expected, imputing genotypes at SNPs that are in strong LD with genotyped markers is much more likely to produce correct genotypes. Imputation accuracy strongly depends on a weighted average of LD.

In addition, Figure 1(c,d) shows the accuracy of imputed genotypes as a function of the SNPs' minor allele frequency (MAF). Imputation of SNPs with a lower MAF appears to be more accurate than imputation of SNPs with a higher MAF. Overall, IMPUTE and MACH had a similar performance.

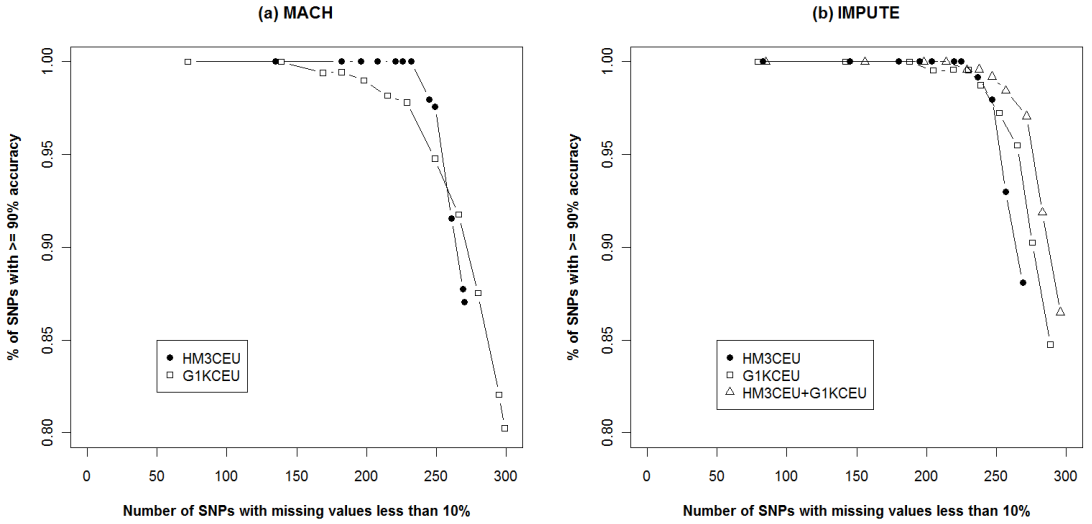

Fig. 2 Proportion of SNPs that have imputation accuracy rates equal to or exceeding 90% as a function of the number of SNPs with missing values ≤ 10% at different threshold values: (a) MACH and (b) IMPUTE.

Figure 2 shows the proportion of SNPs that have imputation accuracy rates equal to or exceeding 90% as a function of the number of SNPs with missing values less than 10% at different threshold values of the posterior probability. In general, the imputation accuracies increase if the threshold value for the posterior probabilities of genotypes is

raised with the expense of more missing data. Within two compared reference panels (HM3CEU and G1KCEU), HM3CEU produced better imputation accuracy at different threshold values for MACH. Within three compared reference panels (HM3CEU, G1KCEU, and HM3CEU+G1KCEU), imputation accuracy rates were highest when using IMPUTE and the combined reference panel (HM3CEU+G1KCEU). This result suggests that the large increase in the number of both SNPs and samples in the reference panel allows more accurate imputation of most ungenotyped SNPs.

We compared the imputation accuracy of two imputation methods (MACH, IMPUTE) within the same reference panel. The results are shown in Figure 3. For HM3CEU, MACH consistently yields higher imputation accuracy rates than IMPUTE. By contrast, for G1KCEU, MACH has slightly lower imputation accuracy rates.

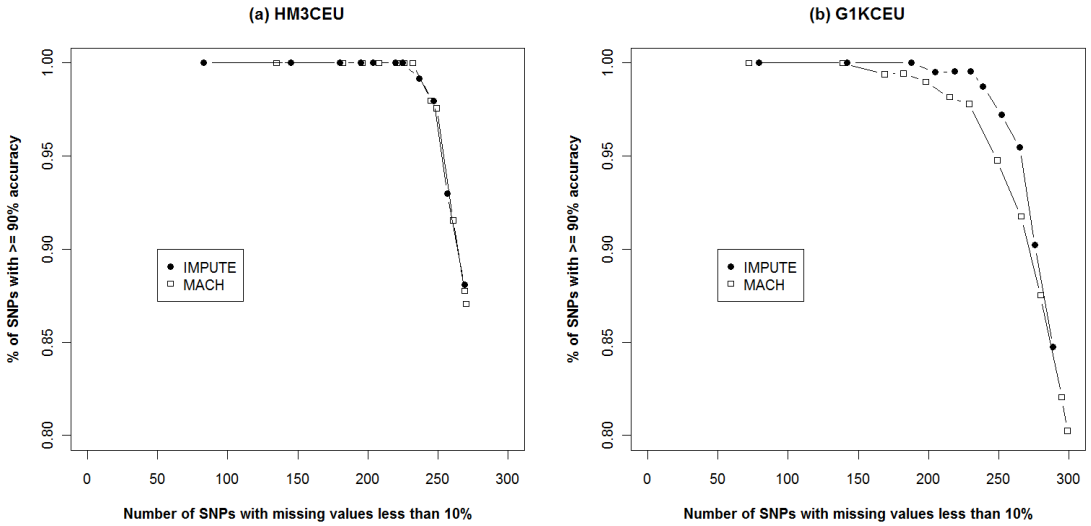

Fig. 3 Proportion of SNPs that have imputation accuracy rates equal to or exceeding 90% as a function of the number of SNPs with missing values less than 10% at different threshold values: (a) HM3CEU and (b) G1KCEU.

Discussion

In this study, we investigated the effect of the reference panels on the imputation accuracy of ungenotyped SNPs using two commonly used imputation methods: IMPUTE v2 and MACH v1. Until now, nearly all imputation-based association studies have been performed using HapMap haplotypes as reference panels. Our study assessed for the first time which combination of genotype imputation method and reference panel could yield the most accurate results. In our analyses, for the HapMap 3 data, MACH consistently yielded higher imputation accuracy rates than IMPUTE, while for the 1000 Genomes Project data, IMPUTE performed slightly better. The best results was achieved by IMPUTE coupled with a combined reference panel (HapMap 3 + 1000 Genomes Project). In addition, we observed that the imputation accuracy was dependent on the extent of LD between the ungenotyped marker and the neighboring genotyped markers as well as its minor allele frequency. However, compared to LD, MAF has a weaker effect on imputation accuracy rate.

Each of the imputation methods, MACH and IMPUTE, has its own strengths and weaknesses. MACH is more user-friendly in terms of data handling, yet MACH requires high memory and CPU time especially for larger chromosomes. A new MACH-based imputation software package, minimac (http://genome.sph.umich.edu/wiki/Minimac), has recently been released. Minimac is a low memory, computationally efficient implementation of the MACH algorithm for genotype imputation that supports multi-threading. IMPUTE can also reduce the computation time and memory requirements, in this case by dividing larger chromosomes into smaller segments of several mega bases. IMPUTE, however, is less user-friendly in handling data.

In summary, in order to maximize the imputation accuracy, IMPUTE coupled with the combined reference data (HapMap + 1000 Genome Project) appears to be a particularly promising strategy to support especially fine-mapping for GWAS and analysis of candidate regions. We note that imputation algorithms and reference panels are

a rapidly evolving target. These issues will require frequent re-evaluation given the current status and prospects for further improvement.

Acknowledgements

Data collection and sharing was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; PI: Michael Weiner; NIH grant U01 AG024904). ADNI is funded by the National Initiative on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contribution from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, the Alzheimer's Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation by the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuro Imaging at the University of California, Los Angeles. This work also was funded by grant U24AG021886 from National Cell Repository for Alzheimer's Disease.

Data analysis was supported in part by the following grants: 5T 15 LM007117-14 from the National Library of Medicine, NIBIB R03 EB008674, NIA R01 AG19771, NCI R01 CA101318 and U54 EB005149 from the NIH, Foundation for the NIH, and grant #87884 from the Indiana Economic Development Corporation (IEDC).

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9:356-69.
2. Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, et al. Common variants at MS4A4/MS4A6E, CD2Ap, CD33, EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011; 43(5):436-41.
3. Hollingworth P, Harold D, Sims R, Gerrish A, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33, and CD2AP are associated with Alzheimer's disease. Nat Genet. 2011; 43(5):429-35.
4. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499-511.
5. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 2008;124:439-50.
6. Nothnagel M, Ellinghaus D, Schreiber S, Krawezak M, Franke A. A comprehensive evaluation of SNP genotype imputation. Hum Genet. 2009; 125:163-171.
7. Pei YF, Zhang L, Li J, Deng HW. Analysis and Comparison of Imputation-Based Association Methods. PLoS One. 2010;5:e10827.
8. Li Y, Abecasis GR. Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. Am J Hum Genet. 2006;S79: 2290.
9. Machini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39: 906-913.
10. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. Genotype-Imputation Accuracy across Worldwide Human Populations. Am J Hum Genet. 2009; 84:235-50.
11. Shen L, Kim S, Risacher SL, Nho K, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. Neuroimage. 2010;53(3):1051-63
12. Saykin AJ, Shen L, Foroud TM, Potkin SG, et al. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. Alzheimers Dement. 2010;6(3):265-73
13. Purcell S, Neale B, Todd-Brown K, Thomas L, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559-75.
14. Zhao Z, Timofeev N, Hartley S, Chui DHK, et al. Imputation of missing genotypes: an empirical evaluation of IMPUTE. BMC Genet. 2008; 9:85.