Original Paper

Human Heredity

Hum Hered 2017/2018;83:107–116
DOI: 10.1159/000489758

Published online: January 22, 2019

Comprehensive Assessment of Genotype Imputation Performance

Shuo Shi[a–c] Na Yuan[b] Ming Yang[d] Zhenglin Du[b] Jinyue Wang[a–c]
Xin Sheng[a–c] Jiayan Wu[a] Jingfa Xiao[a–c]

[a]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China; [b]Big Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China; [c]University of Chinese Academy of Sciences, Beijing, China; [d]Office of General Affairs, Chinese Academy of Sciences, Beijing, China

Keywords
Genotype imputation · Minor allele frequency · Genome-wide association study · Whole genome sequencing

Abstract
Genotype imputation is a process of estimating missing genotypes from the haplotype or genotype reference panel. It can effectively boost the power of detecting single nucleotide polymorphisms (SNPs) in genome-wide association studies, integrate multi-studies for meta-analysis, and be applied in fine-mapping studies. The performance of genotype imputation is affected by many factors, including software, reference selection, sample size, and SNP density/sequencing coverage. A systematical evaluation of the imputation performance of current popular software will benefit future studies. Here, we evaluate imputation performances of Beagle4.1, IMPUTE2, MACH+Minimac3, and SHAPEIT2+IMPUTE2 using test samples of East Asian ancestry and references of the 1000 Genomes Project. The result indicated the accuracy of IMPUTE2 (99.18%) is slightly higher than that of the others (Beagle4.1: 98.94%, MACH+Minimac3: 98.51%, and SHAPEIT2+IMPUTE2: 99.08%). To achieve good and stable imputation quality, the minimum requirement of SNP density needs to be >200/Mb. The imputation accuracies of IMPUTE2 and Beagle4.1 were under the minor influence of the study sample size. The contribution extent of reference to genotype imputation performance relied on software selection. We assessed the imputation performance on SNPs generated by next-generation whole genome sequencing and found that SNP sets detected by sequencing with 15× depth could be mostly got by imputing from the haplotype reference panel of the 1000 Genomes Project based on SNP data detected by sequencing with 4× depth. All of the imputation software had a weaker performance in low minor allele frequency SNP regions because of the bias of reference or software. In the future, more comprehensive reference panels or new algorithm developments may rise up to this challenge.

Introduction

Next-generation sequencing and single nucleotide polymorphism (SNP) arrays are now two main methods to reveal the genotype information. The detection of more loci requires a larger sample size, larger sequencing depth for whole-genome sequencing, and a denser SNP array for microarray-based genotyping. Genotype imputation can be used to solve this dilemma by predicting

E-Mail karger@karger.com
www.karger.com/hhe

Jingfa Xiao or Jiayan Wu
Beijing Institute of Genomics, Chinese Academy of Sciences
No. 1 Beichen West Road
Beijing 100101 (China)
E-Mail xiaojingfa@big.ac.cn or wujy@big.ac.cn

Table 1. Brief introduction of software

Software	URL	Platform	Function
Beagle4.1	https://faculty.washington.edu/browning/beagle/beagle.html	Linux, Mac, Windows	phasing, imputation
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Linux, Mac	phasing, imputation
MACH	http://csg.sph.umich.edu/abecasis/mach/	Linux, Mac, Windows	phasing, imputation
Minimac3	http://genome.sph.umich.edu/wiki/Minimac3	Linux	imputation
SHAPEIT2	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html	Linux, Mac	phasing

untyped genotypes from the haplotype reference panel. It is now being widely used in genome-wide association studies (GWASs) to find novel risk alleles [1, 2], in fine-mapping to get a high-resolution view to increase the likelihood of identifying causal variants [3], and in integrating studies across different platforms for meta-analysis [4, 5]. Now, national large-scale whole-genome sequencing projects are being carried out in many countries, including the United Kingdom [6], Iceland [7], Japan [8], the Netherlands [9], and Singapore [10]. Phasing and imputation have been widely used in these projects to form and assess the haplotype reference panel. Furthermore, the haplotype reference panel generated from these projects can also be used in future imputation studies [11–14].

The performance of genotype imputation is affected by many factors, such as software, reference selection, SNP density (see respective section in "Methods"), sample size, and sequencing coverage. A series of assessments of genotype imputation performance has previously been performed. Huang et al. [16] found that the reference selection, all or part of the International HapMap Project haplotype reference panel, should consider the haplotype number of the reference and the population similarity between the reference and the study set when using MACH [15] for imputation. Nho et al. [18] showed that the effect of reference selection on imputation performance varies across software (MACH and IMPUTE2 [17]) for populations of the USA and Canada. Biernacka et al. [19] reported that imputation accuracy was proportional to the linkage of SNPs. Gao et al. [20] and Zheng et al. [21] found that imputation accuracy was lower in low minor allele frequency (MAF) regions for MACH and IMPUTE2. Zhang et al. [22] found that the sample untype rate of a study set was correlated to imputation accuracy. Here, we added the newly developed and widely used software Beagle4.1 and Minimac3 to our evaluation. Popular software such as Beagle4.1 [23], IMPUTE2, MACH, Minimac3 [24], and SHAPEIT2 [25] has been used in the 1000 Genomes Project [11] and many national large-scale whole-genome sequencing projects [6–10] (Table 1).

Comparisons of software can guide the future of imputation studies. Critical factors of reference selection are haplotype number in the reference panel and the similarity of populations between the reference and study [16]. We chose the haplotype reference of the 1000 Genomes Project as it has more populations and haplotype numbers than HapMap. The sample size and sequencing coverage/SNP density are also related to imputation accuracy. We estimated the performance of the current popular imputation software with all these factors: reference selection, sample size, SNP density, and sequencing coverage. As the low-frequency genetic variants in a population prefer to associate with disease [26], we also evaluated the imputation performance for low MAF SNPs compared to common ones. Furthermore, we analyzed the reasons behind the different performance.

Methods

Study Data

Study sets were selected from Chinese individuals in the International HapMap Project and the 1000 Genomes Project. Only chromosome 1 and 22 were chosen to evaluate the performance of genotype imputation with considering efficiency. 43 individuals were genotyped in all three phases of HapMap, with a denser SNP density (Table 2). 130 individuals were genotyped from Illumina Human1M and Affymetrix SNP 6.0 in HapMap phase III. The National Center for Biotechnology Information (NCBI) genome build of study sets was transformed from 36 to 37 (see below). 10 individuals were sequenced in the 1000 Genomes Project with an

Table 2. Sequencing chips used in the International HapMap Project

Institute	Chip
Affymetrix	GeneChip500K
Affymetrix	GenomeWideSNP_6.0
Affymetrix	genotype_0002
Affymetrix	genotype_protocol_1
Bcm	genotype_0002
Broad	GenomeWideSNP_6.0
Broad	genotype_protocol_1
Illumina	Golden_Gate_1.1.0
Illumina	Infinium_genotyping_1.0.0
Illumina	Infinium_genotyping_2.0.0
Mcgill-gqic	Golden_Gate_1.0.0
Mcgill-gqic	Golden_Gate_1.1.0
Perlegen	Genotyping_1.0.0
Sanger	Golden_Gate_1.0.0
Sanger	Human_1M_BeadChip

averaging coverage of 15× (12× to 16×). Raw sequencing reads were downloaded and aligned to the reference genome (GRCh37) using Burrows-Wheeler Aligner [27]. PCR duplications were removed by Picard [28]. Genome Analysis Toolkit [29] was used to generate high-quality variants. Study data of low-coverage sequencing were simulated by subsampling the raw reads of the 10 individuals' alignment file using SAMtools [28]. Two alignment files of average sequencing coverage 4× (3× to 4×) and 7× (6× to 8×) were generated, followed by variant calling.

Quality Control

Quality control was performed on data using PLINK [30]. The sample filter criterion was: sample call rate <97%. SNP filter criteria were as follows: (i) SNP call rate <97%, (ii) Hardy-Weinberg p value $<10^{-6}$, and (iii) inconsistent SNP sites with the 1000 Genomes Project. The quality control was performed for all samples in the study set. All samples passed the filtration. After quality control step, 183,676 SNPs for chr1 and 31,796 SNPs for chr22 of the 43 samples and 103,797 SNPs for chr1 and 18,510 SNPs for chr22 of the 130 samples in the study set remained. For 10 sequencing samples, only the inconsistent SNPs and SNPs deviating from the Hardy-Weinberg equilibrium were removed.

Imputation Strategy

Several phasing and imputation software programs have been developed to achieve a high imputation accuracy and limited computational burden. Popular software such as Beagle4.1, IMPUTE2, MACH, Minimac3, and SHAPEIT2 has been used in the 1000 Genomes Project and several national large-scale whole-genome sequencing projects in the United Kingdom, Iceland, Japan, the Netherlands, and Singapore. The main algorithm of Beagle4.1 is a hidden Markov model (HMM), which uses a clustering graphical model on haplotypes. IMPUTE2 method is a two-step algorithm comprising phasing and imputation. First, it infers haplotype conditioning from information of the study sample, reference, and recombination rate using a Markov Chain Monte Carlo approach. Second, it uses an HMM to impute the missing genotypes based on the haplotypes of the study sample inferred from the phasing step. IMPUTE2 iterates these two steps to maximize the posterior probabilities of the missing alleles for imputation. Minimac3 uses state space reduction HMM to reduce the running time. MACH is a phasing method using an HMM model. SHAPEIT2 is a haplotype estimation method using an HMM model on a graph structure of haplotypes. These tools are the most commonly used in genotype imputation. As the previous studies suggested, MACH can be used with Minimac3 for pre-phasing [24], and IMPUTE2 can use SHAPEIT2 for pre-phasing [31]. Thus, we performed four imputation methods, which were Beagle4.1, IMPUTE2, SHAPEIT2+IMPUTE2, and MACH+Minimac3. All software was run with default parameters.

As for the reference, we should consider sample number and population similarity between the reference and the study set simultaneously. Compared to the International HapMap Project, the 1000 Genomes Project with next-generation sequencing data is now the primary source for reference panels with more variants, individuals, and various populations. The 1000 Genomes Project Phase 3 contains 84.4 million variants of 2,504 individuals from 26 populations. We employed the 1000 Genomes Project as a reference panel (1KG_ALL). Considering the similarity of the populations, we introduced 504 East Asian individuals from the 1000 Genomes Project as another reference panel (1KG_EAS) to evaluate the influence of reference selection on imputation performance.

In addition to software and reference selection, we also considered other factors that affected imputation performance. In order to evaluate the effect of SNP density in study sets, we subsampled SNPs by sorting SNPs based on position and extracted m SNPs (m = 1, 1.5, 2, 2.5, 3–6, 9, 12, 15, 18, 21, 24) for every 36 SNPs to generate 14 subsets with different SNP density levels. After extracting 24 SNPs for every 36 SNPs, the remaining SNPs were selected as test set. Considering software, reference selection, and SNP density, these three factors formed 112 different combinations; here, we evaluated their influence on imputation. After comprehending the correlation result of these three factors, we further estimated the effect of sample size on imputation using 6 levels (10, 20, 40, 80, 100, and 130). The study sets with different sample sizes were sampled from 130 HapMap phase III without replacement. The bigger sample size study sets were generated by adding new individuals to smaller one. As next-generation sequencing data is a mainstream method for large-scale population sequencing, we also evaluated the imputation performance with different sequencing coverage (Fig. 1).

Imputation Measurement

The accuracy was represented by the concordance rate, the percentage of correctly imputed genotypes of the test set. We also used r^2 [squared Pearson correlation coefficient between imputed genotype dosages in (0–2) and masked sequence genotypes in (0, 1, 2)] to represent the imputation accuracy. In addition to the direct comparison of imputation accuracy, we also analyzed the discordance rate versus the missing rate under different thresholds of imputation genotype posterior probabilities. Imputation accuracy was assessed for different MAFs. We emphasized the assessment of the imputation performances between common (≥ 5%) and low (<5%) frequency groups. In addition to comparing the imputation accuracy directly, we calculated the sensitivity and false-positive rate (FPR) of these two groups (see below). We ran all imputation strategies twice on the same Linux server, which has two Intel Xeon

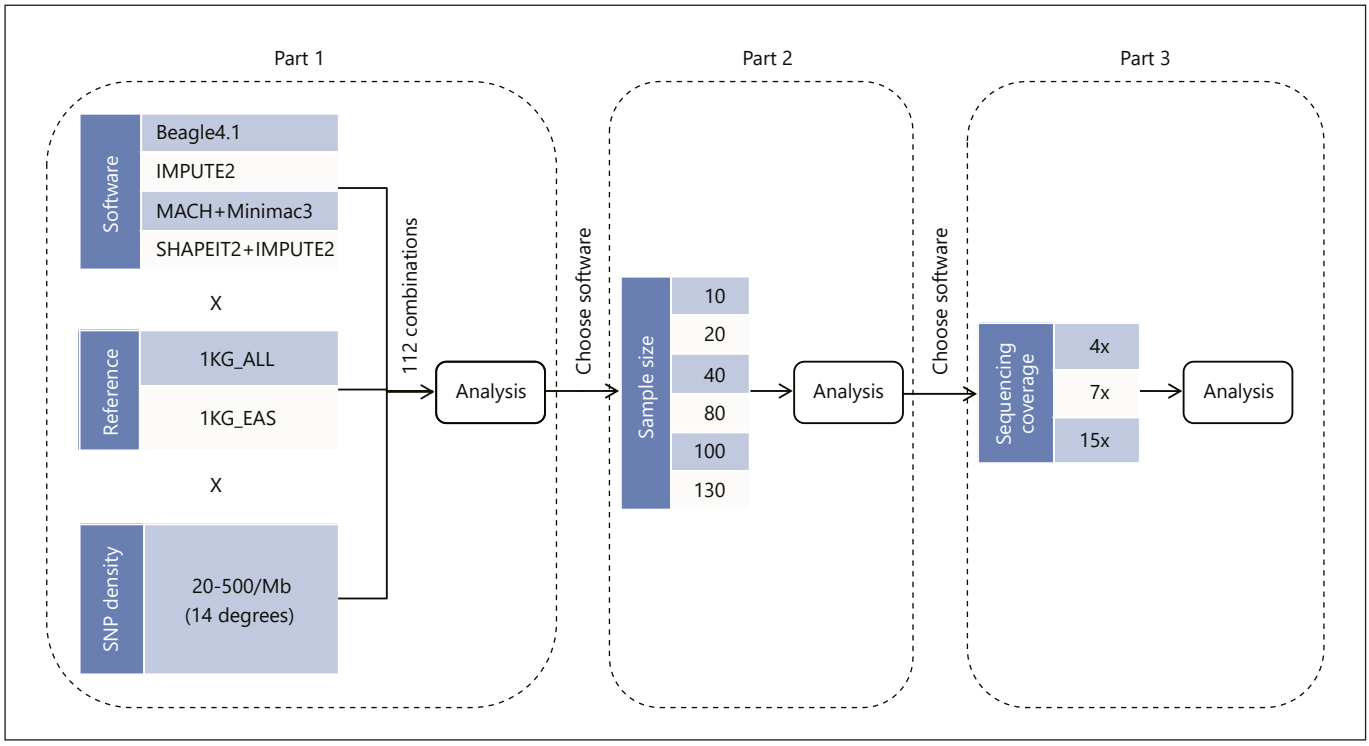

Fig. 1. Flow chart of imputation performance assessment.

Table 3. 2 × 2 contingency table

		Imputed genotype	
		1	0
Masked genotype	1	true positive (TP)	false negative (FN)
	0	false positive (FP)	true negative (TN)

Sensitivity = TP/(TP + FN); false-positive rate = FP/(FP + TN).

X5650 processors (running at 2.66 GHz, with a 12 MB cache, and using a 64-bit architecture) and used the average value to assess the computational burdens, including running time and computational memory.

SNP Density

$$\text{SNP density} = \frac{\text{snp number}}{\text{chromosome length}}$$

SNP density is the percentage of SNP number per megabase (MB).

Converting HapMap Build36 Data to Build37

First, Genome Analysis Toolkit was used to convert the HapMap txt file into a vcf file. Then, liftover was used to convert the build36 vcf file into a build37 vcf file.

Imputation Quality Measurement

The discordance rate versus the missing rate on different threshold of imputation genotype posterior probabilities is one of the measurements to evaluate the performance. The discordance rate is the percentages of discordance between imputed genotypes and masked genotypes. The missing rate is the percentage of no calls made. Sensitivity and FPR are formulated in a 2 × 2 contingency table (Table 3).

Results

Imputation Performance Affected by Software, Reference Selection, and SNP Density

Imputations were performed on 43 individuals using Beagle4.1, IMPUTE2, MACH+Minimac3, or SHAPEIT2+IMPUTE2 under different reference selection and SNP density conditions (Fig. 1, Part 1). The imputation accuracy generated by each piece of software was higher than 85% on the test sets, chr1, and chr22 (Fig. 2a, b). The results from chr1 and chr22 are consistent; therefore, only the results for chr1 are shown below. IMPUTE2 performed slightly better than the others. The advantages of IMPUTE2 are reflected in using the reference informa-

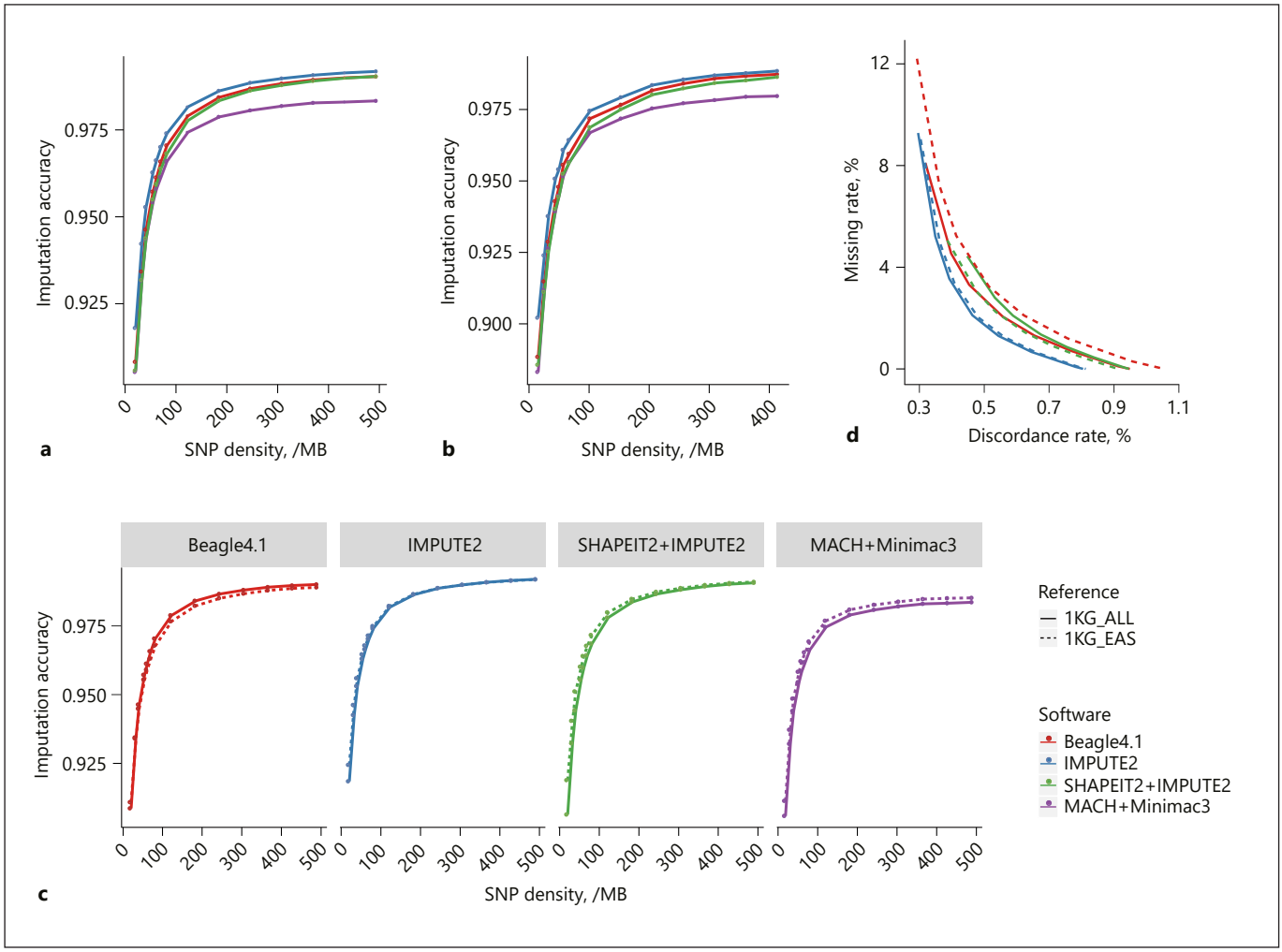

Fig. 2. a, **b** Imputation accuracy of the 112 kinds of strategies for chr1 and chr22. The x-axis represents SNP density. The y-axis represents the imputation accuracy, which is the rate of consistent sites between imputed genotypes and masked genotypes. **c** Imputation accuracy of Beagle4.1, IMPUTE2, MACH+Minimac3, and SHAPEIT2+IMPUTE2 with two references (1KG_ALL and 1KG_EAS) for chr1. **d** Percentage discordance versus percentage missing genotypes for calling thresholds ranged from 0.33 to 0.99 for chr1.

tion both in phasing and imputation compared to SHAPEIT2+IMPUTE2, directly applying an HMM on haplotypes compared to Beagle4.1, and choosing the closest subset of haplotypes with the study set compared to MACH.

These are probable causes for IMPUTE2's better performance. Different software programs preferred different reference panels: 1KG_EAS was better suited to MACH+Minimac3 and SHAPEIT2+IMPUTE2, and 1KG_ALL was better suited to Beagle4.1 (Fig. 2c, d). IMPUTE2 performed stably on both 1KG_EAS and 1KG_ALL. The results suggested that software with a pre-phasing method, MACH+Minimac3 and SHAPEIT2+IMPUTE2, performed better with higher population similarity between the study and the reference. As SNP density increased, the imputation accuracy increased gradually in all software for microarray-based genotyping. The imputation result from chr1 and chr22 claimed that when SNP density reached 200/Mb, the accuracy tended to saturate.

As for computational burden, IMPUTE2 was the most time-consuming, while SHAPEIT2+IMPUTE2 was the fastest (Fig. 3a). Except for Beagle4.1, the running times of the other software programs were influenced only by SNP number and sample number of the reference, instead

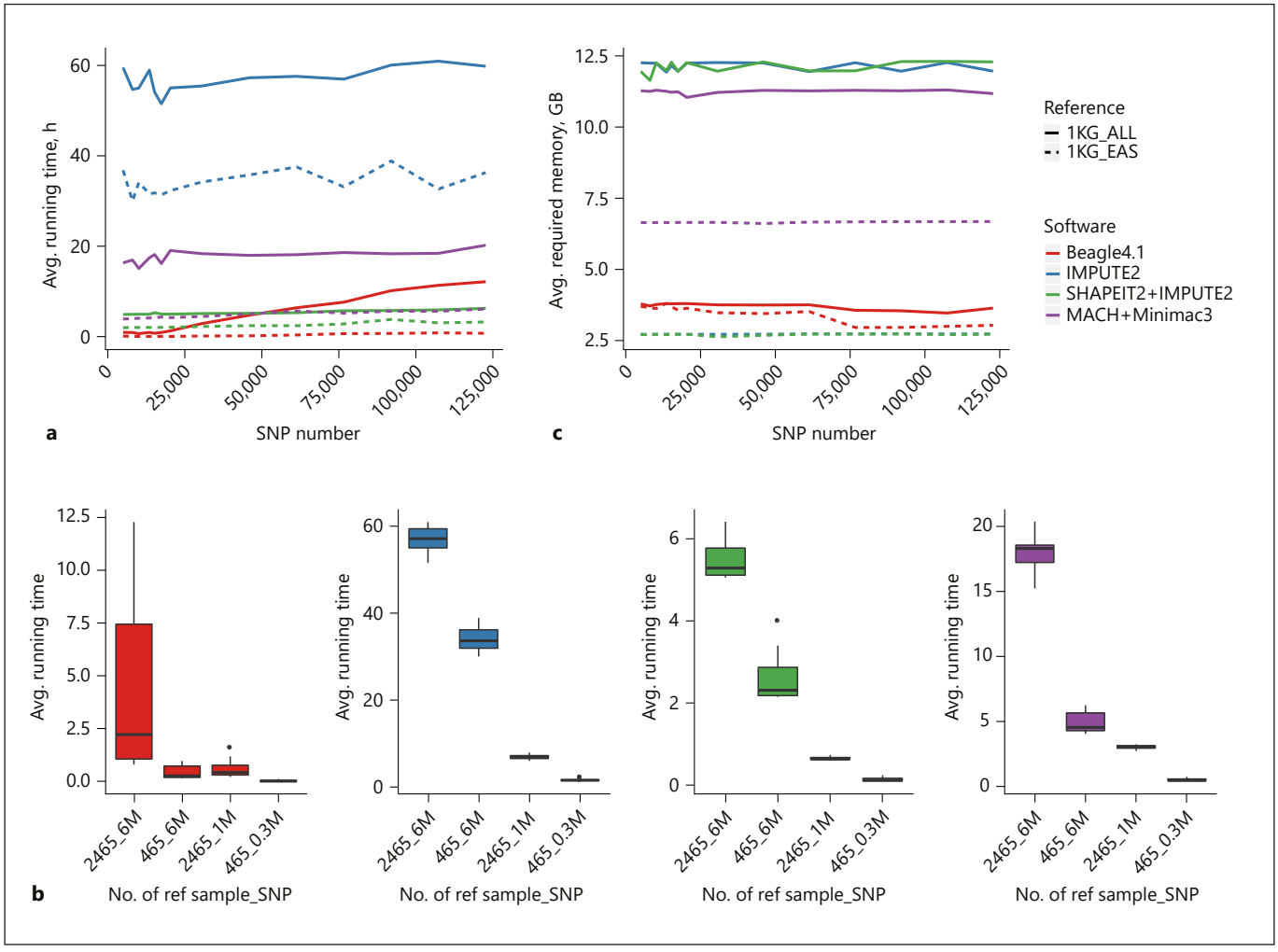

Fig. 3. a Average running time of all 112 kinds of strategies for chr1. **b** Average running time of software on different references (sample number_snp number). **c** Average computer memory of all 112 kinds of strategies for chr1.

of SNP density in the study set (Fig. 3b). This limitation is because the SNP number and sample number of the reference are much larger than those of the study set. The required computer memory for the software was not affected by SNP density and was sensitive to the reference (Fig. 3a, c).

Effect of Sample Size

After a general assessment of the software, we chose better-performing software, IMPUTE2, Beagle4.1, and SHAPEIT2+IMPUTE2 with an SNP density 278/Mb, to evaluate the influence of sample size. IMPUTE2 and Beagle4.1 were generally not affected by sample size as compared to SHAPEIT2+IMPUTE2 (Fig. 4a). With expanded sample size, IMPUTE2 was still the best performing software and was stable on both references (Fig. 4b). As for the running time of the software, Beagle4.1 and SHAPEIT2+IMPUTE2 were much more insensitive to sample size (Fig. 4c).

Imputation Accuracy with MAF

In GWASs, low-allele frequency SNPs tend to be disease-associated [32, 33]. Thus, the imputation of low-frequency alleles is more important. As the MAF decreased, the accuracies of all software also decreased (Fig. 5a). IMPUTE2 and Beagle4.1 worked better with low-frequency

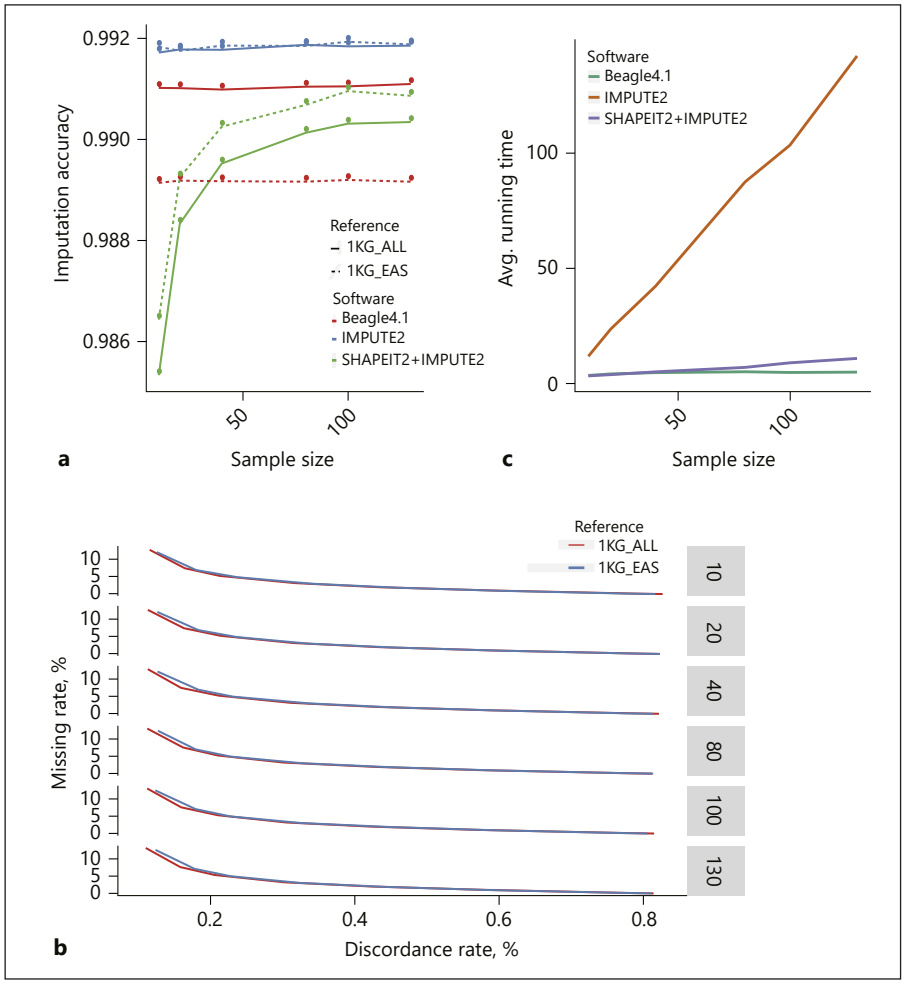

Fig. 4. The influence of sample size on imputation for chr1. **a** The average imputation accuracy of different sample sizes and methods. **b** Percentage discordance versus percentage missing genotypes for calling thresholds ranged from 0.33 to 0.99 of IMPUTE2 for chr1. **c** The corresponding average running time.

SNPs. Usually, software utilizes linkage disequilibrium of SNPs to impute missing genotypes, and imputation accuracy is proportional to SNP linkages [34]. Here, we divided SNPs into two groups by their MAF: <5% and ≥5%. Comparing the linkage disequilibrium rate between the two groups, we found the average linkage disequilibrium rate (the mean of the correlation coefficients r^2 between SNPs calculated by PLINK) of the <5% group was not lower than that of the ≥5% group (0.691 vs. 0.670). Therefore, in our study, low imputation accuracy of low MAF SNP was not caused by the linkage disequilibrium rate.

The study set or reference is actually a subset of all populations. Sampling error can cause discordance between study set and reference. In mathematics, sampling error is proportion to the variability of the data set. As rare SNPs have more variability, the discordance between study set and reference of rare SNPs is bigger than that between study set and reference of common SNPs. So, this might be the reason why imputation accuracies of rare SNPs are poorer, the bias of reference rather than the bias of the linkage disequilibrium rate. The sensitivity and FPR in the <5% group were both lower than in the ≥5% group, which suggested that the software preferred to impute the low MAF alleles to be the same allele as the human genome reference, no matter what the genotypes really were (Fig. 5b and Table 4). Therefore, software bias could also be the reason for the low accuracy of low MAF SNPs.

Effect of Sequencing Coverage

Imputation is an essential part of genotype detection of low-coverage next-generation sequencing projects. Therefore, we assessed the imputation performance with different sequencing depths in a low-coverage sequencing project. The identified SNPs of 10 individuals with 4× and 7× coverage were mostly consistent with 15× (Fig. 6).

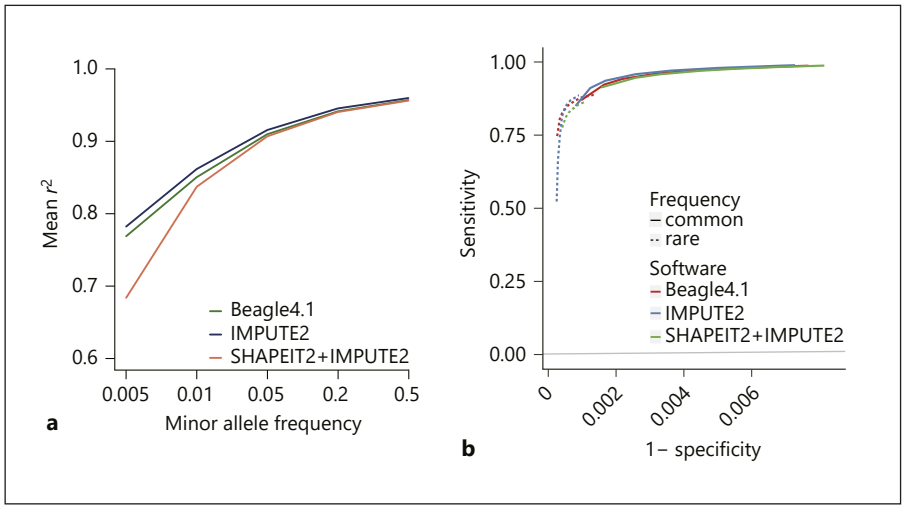

Fig. 5. a Imputation accuracy as a function of minor allele frequency. r^2 Pearson correlation coefficient between imputed genotype dosages and masked sequence genotypes. **b** Receiver operating characteristic curve (ROC). Sensitivity and false-positive rate (1 – specificity) were calculated by calling thresholds ranging from 0.33 to 0.99. The missing sites default to be 0/0.

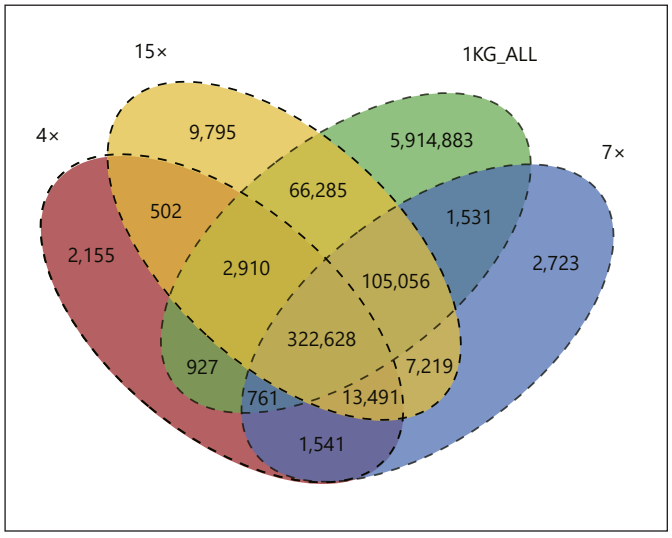

Fig. 6 Venn diagram of SNPs discovered in three sets of average sequencing coverage (4×, 7×, and 15×) and 1KG_ALL.

Table 4. Sensitivity and false-positive rate (FPR) of Beagle4.1, IMPUTE2, and SHAPEIT2+IMPUTE2 on minor allele frequency (MAF) <5% and ≥5% separately without calling threshold

	MAF <5%		MAF ≥5%	
	sensitivity, %	FPR, %	sensitivity, %	FPR, %
Beagle4.1	88.68	0.13	98.69	0.77
IMPUTE2	88.49	0.09	98.87	0.73
SHAPEIT2+IMPUTE2	86.46	0.11	98.81	0.84

Compared with 4× or 7×, most of the 15× specified SNP could be imputed from 1KG_ALL. In theory, the SNP data set with 15× could be inferred based on data of 4× or 7×. Imputation was performed on SNP data of 4× and 7× from 1KG_ALL using IMPUTE2 because of its better performance on 10 individuals. The SNP density of 4× and 7× is 1,384/Mb and 1,825/Mb, respectively. The average imputation accuracies of IMPUTE2 at 4× and 7× coverage were 90.30 and 90.56%, respectively. With a study set SNP density over 200/Mb, imputation accuracy with 4× or 7× coverage in next-generation sequencing was lower than the microarray study. The reasons might be the lower sample typed rate (78.95% of 4× and 94.94% of 7×) [22] and the influence of inconsistently called SNP loci (1.56% between 4× and 15×, and 1.44% between 7× and 15×).

Discussion

In this study, we systematically estimated the effect of software, reference selection, sample size, SNP density/sequencing coverage, and MAF on imputation. When SNP density reached 200/Mb for microarray-based genotyping, imputation accuracy tended to be stable and Beagle4.1, IMPUTE2, MACH+Minimac3, and SHAPEIT2+IM-PUTE2 all achieved high imputation accuracy. IMPUTE2 provided the best accuracy but was also the most time-consuming and used the most memory, since it adopts as much biological information as possible in the phasing and imputation steps. SHAPEIT2+IMPUTE2 was

the fastest and most memory-saving imputation method with little sacrifices to accuracy compared with IMPUTE2, which might make it more practical for large data sets generated by next-generation sequencing [31]. As for reference selection, a reference that has a high population similarity with the study set was more suitable for MACH+Minimac3 and SHAPEIT2+IMPUTE2. Beagle4.1 performed better with a reference that has a large sample number. IMPUTE2 was relatively stable regardless of reference selection. The performances of IMPUTE2 and Beagle4.1 were stable regardless of sample size.

We found that all tested software worked poorly for the imputation of low MAF SNPs. The cause was not the linkage disequilibrium rate but the bias of references and software. For low MAF SNPs, discordance information provided by the reference duo to the bias could cause imputation accuracy to decrease. According to sensitivity and FPR, the software preferred to impute the missing genotypes to be the same as the human genome reference at low MAF SNPs. The problem with the reference can be eased by increasing the sample number of the reference, especially the population-specific samples [16]. Fortunately, population-specific haplotype reference panels have been constructed in several countries. For the software, the problem of improving the imputation accuracy at low MAF SNPs is challenging and warrants considerable attention. A new algorithm needs to be developed to improve the accuracy on low MAF SNPs.

As next-generation sequencing is now the main way for GWASs, imputation on sequencing-provided data needs extra attention [13]. The imputation on low-coverage sequencing is challenging mostly because of the low sample typed rate and inconsistently called SNP loci. The inconsistently called SNPs may be filtered by strict quality control standards. For the low sample typed rate of a study set, there are two ways to solve the problem: deeper sequencing or higher quality references. Increased sequencing depth comes with more sequencing costs, and the existing references are limited. Therefore, improvement of imputation software and some newly developed algorithm could be a direct solution to this problem.

Conclusions

The effects of software, reference selection, SNP density, and sample size on imputation performance were evaluated to give a general and practical guidance for future imputation studies. We found that IMPUTE2 was the most accuracy and stable method. SHAPEIT2+IMPUTE2 was the fastest method with some accuracy sacrifice. To get a reliable imputation result, SNP density should be >200/Mb for microarray-based genotyping. Imputation performance on SNPs generated by next-generation whole genome sequencing was assessed, we found that SNP sets detected by higher sequencing depth (15×) could be mostly available by imputing from the haplotype reference panel of the 1000 Genomes Project based on SNP data of lower sequencing depth (4×). All of the imputation software had a weaker performance in low MAF SNP regions because of the bias of reference or software, so more comprehensive reference panels or new algorithm developments may rise up to this challenge.

Acknowledgements

The authors thank Dr. Jun Yu and Dr. Songnian Hu for their valuable discussions on this work.

Statement of Ethics

The work used open access data of the 1000 Genomes Project and the International HapMap Project.

Disclosure Statement

The authors declare that they have no competing interests.

Funding Sources

This work was supported by the National Key Research Program of China (2016YFB0201702; 2017YFC0907503 to J.X. and 2016YFC0901903 to Z.D.); the National Natural Science Foundation of China (31471248 to J.X.); the Key Program of the Chinese Academy of Sciences (KJZD-EW-L14 to J.X.); the Youth Innovation Promotion Association CAS (grant to J.W.); and funding for open access charge: the National High-tech R and D Program (2015AA020101 to Z.D.).

References

1 Spencer CC, Su Z, Donnelly P, Marchini J: Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 2009; 5:e1000477.
2 Marchini J, Howie B: Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499–511.
3 Browning SR, Browning BL: Haplotype phasing: existing methods and new developments. Nat Rev Genet 2011;12:703–714.

4 Chang BL, Cramer SD, Wiklund F, Isaacs SD, Stevens VL, Sun J, Smith S, Pruett K, Romero LM, Wiley KE, Kim ST, Zhu Y, Zhang Z, Hsu FC, Turner AR, Adolfsson J, Liu W, Kim JW, Duggan D, Carpten J, Zheng SL, Rodriguez C, Isaacs WB, Gronberg H, Xu J: Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. Hum Mol Genet 2009;18:1368–1375.
5 Chen F, Chen GK, Millikan RC, et al: Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. Hum Mol Genet 2011;20:4491–4503.
6 Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng HF; UK10K Consortium, Gambaro G, Richards JB, Durbin R, Timpson NJ, Marchini J, Soranzo N: Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 2015;6:8111.
7 Gudbjartsson DF, Helgason H, Gudjonsson SA, et al: Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 2015;47:435–444.
8 Nagasaki M, Yasuda J, Katsuoka F, et al: Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 2015;6:8018.
9 Genome of the Netherlands Consortium: Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 2014;46:818–825.
10 Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, Lam KK, Pillai NE, Sim KS, Xu H, Sim NL, Teo SM, Foo JN, Tan LW, Lim Y, Koo SH, Gan LS, Cheng CY, Wee S, Yap EP, Ng PC, Lim WY, Soong R, Wenk MR, Aung T, Wong TY, Khor CC, Little P, Chia KS, Teo YY: Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet 2013;92:52–66.
11 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: A global reference for human genetic variation. Nature 2015; 526:68–74.
12 Ling Y, Jin Z, Su M, Zhong J, Zhao Y, Yu J, Wu J, Xiao J: VCGDB: a dynamic genome database of the Chinese population. BMC Genomics 2014;15:265.
13 McCarthy S, Das S, Kretzschmar W, et al: A reference panel of 64,976 haplotypes for genotype imputation. Nature Genet 2016;48: 1279–1283.
14 BIG Data Center Members: The BIG Data Center: from deposition to integration to translation. Nucleic Acids Res 2017;45:D18–D24.
15 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;34:816–834.
16 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: 235–250.
17 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:e1000529.
18 Nho K, Shen L, Kim S, Swaminathan S, Risacher SL, Saykin AJ; Alzheimer's Disease Neuroimaging Initiative (ADNI): The effect of reference panels and software tools on genotype imputation. AMIA Annu Symp Proc 2011;2011:1013–1018.
19 Biernacka JM, Tang R, Li J, McDonnell SK, Rabe KG, Sinnwell JP, Rider DN, de Andrade M, Goode EL, Fridley BL: Assessment of genotype imputation methods. BMC Proc 2009; 3(suppl 7):S5.
20 Gao X, Haritunians T, Marjoram P, McKean-Cowdin R, Torres M, Taylor KD, Rotter JI, Gauderman WJ, Varma R: Genotype imputation for Latinos using the HapMap and 1000 Genomes Project reference panels. Front Genet 2012;3:117.
21 Zheng HF, Ladouceur M, Greenwood CM, Richards JB: Effect of genome-wide genotyping and reference panels on rare variants imputation. J Genet Genomics 2012;39:545–550.
22 Zhang B, Zhi D, Zhang K, Gao G, Limdi NN, Liu N: Practical consideration of genotype imputation: sample size, window size, reference choice, and untyped rate. Stat Interface 2011;4:339–352.
23 Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084–1097.
24 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;44:955–959.
25 Delaneau O, Marchini J, Zagury JF: A linear complexity phasing method for thousands of genomes. Nat Methods 2011;9:179–181.
26 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010;11:415–425.
27 Li H, Durbin R: Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25:1754–1760.
28 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–2079.
29 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20: 1297–1303.
30 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.
31 Delaneau O, Zagury JF, Marchini J: Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 2013;10:5–6.
32 Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009;324:387–389.
33 Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH: Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. New Engl J Med 2006;354: 1264–1272.
34 Biernacka JM, Tang R, Li J, McDonnell SK, Rabe KG, Sinnwell JP, Rider DN, de Andrade M, Goode EL, Fridley BL: Assessment of genotype imputation methods. BMC Proc 2009; 3:S5.