Human Genetics (2018) 137:431–436
https://doi.org/10.1007/s00439-018-1894-z

ORIGINAL INVESTIGATION

Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference

Yuan Lin[1] · Lu Liu[2,3] · Sen Yang[2,3] · Yun Li[4] · Dongxin Lin[5] · Xuejun Zhang[2,3] · Xianyong Yin[2,3,6]

Received: 13 February 2018 / Accepted: 26 May 2018 / Published online: 31 May 2018

Abstract

Genotype imputation is now routinely performed in genomic analysis. Reference panel size, that is, the number of haplotypes in the reference panel, has been well established to be one major driving factor of imputation accuracy. For that reason, huge efforts have been made worldwide to provide large reference panels, with the Haplotype Reference Consortium (HRC) being currently the largest available in the public domain. The imputation performance of HRC, whose major samples are Europeans, has been mainly evaluated in Europeans. We conducted whole-genome genotype imputation on two independent genome-wide genotyping datasets, one with 1000 European samples and the other with 1000 Han Chinese samples. We compared the results obtained using HRC with those using Phase III of the 1000 Genomes Project (1000G) reference panel. For the European dataset, using HRC improved imputation quality, especially for rare variants with minor allele-frequency (MAF) < 0.1%. However, 1000G demonstrates better performance in the Han Chinese dataset, in both imputation quality and number of well-imputed variants. We validated the performance of 1000G reference panel in a second, independent cohort of Han Chinese ($N = 2402$). Our study showcases the limitations of HRC for Han Chinese populations, strongly suggesting the necessity of building population-specific reference panels.

Introduction

Genotype imputation has now been routinely performed in genome-wide association studies (GWASs) to increase statistical power (Guan and Stephens 2008; Li et al. 2010a; Marchini et al. 2007), facilitate fine-mapping efforts (Liu et al. 2010) and enable meta-analysis of studies using different genotyping platforms (De Bakker et al. 2008). Conceptually, genotype imputation methods work by identifying haplotype segments shared between the study cohorts, typically genotyped on a commercial array that directly examines 10^5–10^6 single nucleotide polymorphisms (SNPs), and a reference panel such as those from the International HapMap Project (Frazer et al. 2007), the 1000 Genomes Project (Auton et al. 2015), Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (Vergara et al. 2018), and the Haplotype Reference Consortium (HRC) (McCarthy et al. 2016). Based on the inference regarding shared haplotype segments, imputation methods can estimate genotypes or genotype probabilities at 10^6–10^8 untyped markers.

Previous studies have shown that genotype imputation accuracy increases with the size of the reference panel (Auer et al. 2012; Duan et al. 2013; Li et al. 2009; Marchini and

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-018-1894-z) contains supplementary material, which is available to authorized users.

✉ Xianyong Yin
xianyongyin@gmail.com

1 Center for Bioinformatics, School of Life Sciences, Peking University, Beijing 100871, China

2 Institute of Dermatology and Department of Dermatology, First Affiliated Hospital, Anhui Medical University, Hefei, 81 Meishan Road, Hefei 230032, Anhui, China

3 Key Laboratory of Dermatology, Ministry of Education (Anhui Medical University), Hefei 230032, Anhui, China

4 Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27519, USA

5 Department of Etiology and Carcinogenesis, Cancer Institute and Hospital, Chinese Academy of Medical Sciences, Beijing 100021, China

6 Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Howie 2010). The large numbers of SNPs and increased haplotype diversity would allow researchers to impute SNPs with higher confidence, especially SNPs with low minor allele frequencies (MAFs), which will later enhance statistical power in subsequent association analysis. Therefore, large reference panels comprised of haplotypes from single or multiple populations have been built through international collaborations, including the HapMap Project (Frazer et al. 2007), the 1000 Genomes Project (Auton et al. 2015), the Genome of the Netherlands Consortium (GoNL) (Genome of the Netherlands Consortium 2014), and Singapore Genome Variation Project (SGVP) (Teo et al. 2009), Finnish (Fuchsberger et al. 2016), Icelandic (Gudbjartsson et al. 2015), Sardinian (Sidore et al. 2015), and African (Vergara et al. 2018) population-specific panels. Some argue that reference panel size could be more important than ancestry/ethnicity match between the reference panel and the study cohort for improving imputation accuracy (Marchini and Howie 2010a). Such argument has led to combining available datasets (mostly low-pass whole-genome sequencing data) from diverse ancestral sources. In 2015, the UK10K cohort project reported that the reference panel they built with 3,781 British samples produced higher quality than the commonly used 1000G reference panel, with greater gains at low-frequency variants (Huang et al. 2015). In addition, by combining the 1000G and the UK10K reference panels, the proportion of well-imputed variants with MAF 0.1%-0.5% in a North Chinese sample was further increased from 49.8 to 61.8% (Chou et al. 2016).

The recently released HRC (http://www.haplotype-reference-consortium.org/) reference panel (HRC v1.1) contains a set of variants collected from 20 different studies. HRC contains 64,976 haplotypes at 39,235,157 SNPs with minor allele counts (MAC) ≥ 5 (McCarthy et al. 2016). Although HRC v1.1 has clearly demonstrated its advantages over the 1000G reference panel for imputation in individuals of European ancestry (McCarthy et al. 2016), it has been rarely evaluated for non-European populations. In this study, we evaluated the performance of HRC v1.1 in the Han Chinese population, with an emphasis on imputation quality. We imputed two real Han Chinese datasets using HRC v1.1. For comparisons, we also imputed a European dataset using HRC and all three datasets using the 1000G Phase III reference panel.

Materials and methods

Three real data sets

The first data set (referred to as Anhui) consists of 1132 subjects of Han Chinese ancestry. These subjects are healthy controls recruited in multiple hospitals in China and genotyped with the Illumina Human 610-Quad BeadChips, as previously described (Zhang et al. 2009). The second data set (referred to as 58 BC) is the 1958 British Birth Cohort of 3000 subjects genotyped by the Wellcome Trust Case–Control Consortium (WTCCC2) using the Illumina 1.2M array (Craddock et al. 2010). A third data set (referred to as Beijing), used to validate our findings in the Anhui cohort, contains 2042 healthy participants from Beijing genotyped using the Affymetrix GeneChip Human Mapping 6.0 (Wu et al. 2011).

All data sets applied stringent quality controls (QC) on variants and samples. Only autosomal bi-allelic SNPs that have genotype call rate per sample and per variant > 95%, MAF > 1%, and no significant deviation from Hardy–Weinberg equilibrium (HWE) ($P > 10^{-6}$) were retained. From each of the Anhui and 58 BC data sets, we randomly selected 1000 unrelated samples with a pairwise genetic relationship coefficient less than 0.025. The genetic correlation matrix for each data set was estimated by GCTA v1.45 (Yang et al. 2011) using common variants with MAF > 10%. Imputation was performed using the 324,162 QC + genotyped variants with consistent allele coding in the 58 BC data set and in the Anhui data set; in the validation study, imputation was carried out using the 459,686 QC + genotyped variants shared between the 58 BC data set and the Beijing data set. Allele strand, position, and allele coding were all aligned and consistent with the HRC and 1000G reference panels through using HRC or 1000G imputation preparation and checking tools (http://www.well.ox.ac.uk/~wrayner/tools/).

Pre-phasing and Haplotype-based imputation

There are two publicly available imputation servers: Michigan (https://imputationserver.sph.umich.edu/) and Sanger (https://imputation.sanger.ac.uk/). All imputation in this study was conducted on the Sanger Imputation Server because it uses Positional Burrows-Wheeler Transform (PBWT) which supports efficient compression of large numbers of haplotypes (Durbin 2014). The Sanger Imputation Server provides both the HRC v1.1 and the 1000G reference panels. We used two pre-phasing methods, EAGLE v2 (Loh et al. 2016) and SHAPEIT v2 (Delaneau et al. 2013). The imputation performance was measured by mean imputation INFO r^2 value (Li et al. 2010b; Marchini and Howie 2010). Statistical significance in mean differences was evaluated using paired t test.

Results

The HRC panel consists of 32,470 subjects with predominantly European ancestry and also includes all 2504 subjects from the 1000 Genomes Project. The allele-frequency

differences of 30,556,873 bi-allelic SNPs shared by the HRC and the 1000G panels are relatively small in the European samples of both panels (Fig. S1). For the Anhui (Han Chinese) and the 58 BC (European) data sets, using the HRC and 1000G reference panels resulted in 39,284,340 and 82,303,958 imputed variants, respectively (Table 1 and S1). There were 48–58% of them well-imputed (INFO $r^2 \geq 0.8$) (Table 1 and S1). Regarding the 30,099,640 variants shared by both data sets, the imputation quality of the 58 BC dataset of European samples was considerably higher using the HRC reference panel than using the 1000G reference panel (Fig. 1a), and the mean imputation r^2 had a statistically significant increase from 0.60 to 0.71 ($P = 2.2 \times 10^{-6}$). The improvement was more profound for low-frequency (0.5% < MAF < 1%) and rare variants (MAF < 0.5%) than for common variants (MAF > 5%) (Fig. 1a and Table S2). When applying the two reference panels on the Anhui data set, however, the 1000G reference panel significantly outperformed the HRC reference panel ($P = 2.2 \times 10^{-6}$) regarding the mean imputation quality r^2 of imputed variants shared between the HRC and 1000G panels (Fig. 1b).

The superior performance of the 1000G reference panel in the Anhui dataset remained when we changed the pre-phasing methods from SHAPEIT v2 to EAGLE v2 (Fig. S2). We ran the same imputation pipeline on a second independent Chinese data set (Beijing) and achieved similar results

Table 1 Summary statistics of imputed variants in the 58 BC and Anhui populations[a]

Study	Panel[b]	N[c]	Number of variants within INFO Bin (%)[d]			Shared variants[e]	
			[0.3–0.5)	[0.5–0.8)	[0.8–1]	*N*	Mean INFO (SD)[f]
58 BC	HRC	39,284,340	1,723,432 (4.4%)	4,553,934 (11.6%)	22,742,918 (57.9%)	30,099,640	0.71 (0.36)
	1000G	82,303,958	5,233,553 (6.4%)	9,593,139 (11.7%)	39,593,513 (48.1%)	30,099,640	0.60 (0.36)
Anhui	HRC	39,284,340	3,405,925 (9.7%)	5,225,074 (13.3%)	16,307,450 (41.5%)	30,099,640	0.55 (0.37)
	1000G	82,303,958	6,563,591 (8.0%)	10,273,479 (12.5%)	36,233,411 (44.0%)	30,099,640	0.58 (0.36)

[a]The statistics in this table comes from the imputation using the pre-phasing approach of SHAPEIT v2

[b]Panel: reference panel

[c]*N*: total number of imputed variants

[d]*N* of variants within INFO Bin: number of imputed variant within the bin of imputation INFO r^2; the proportion is shown in parentheses

[e]Shared Variant: the shared imputed variants between outputs with the HRC and 1000G reference panels

[f]Mean INFO: mean imputation INFO r^2 value; SD: standard deviation

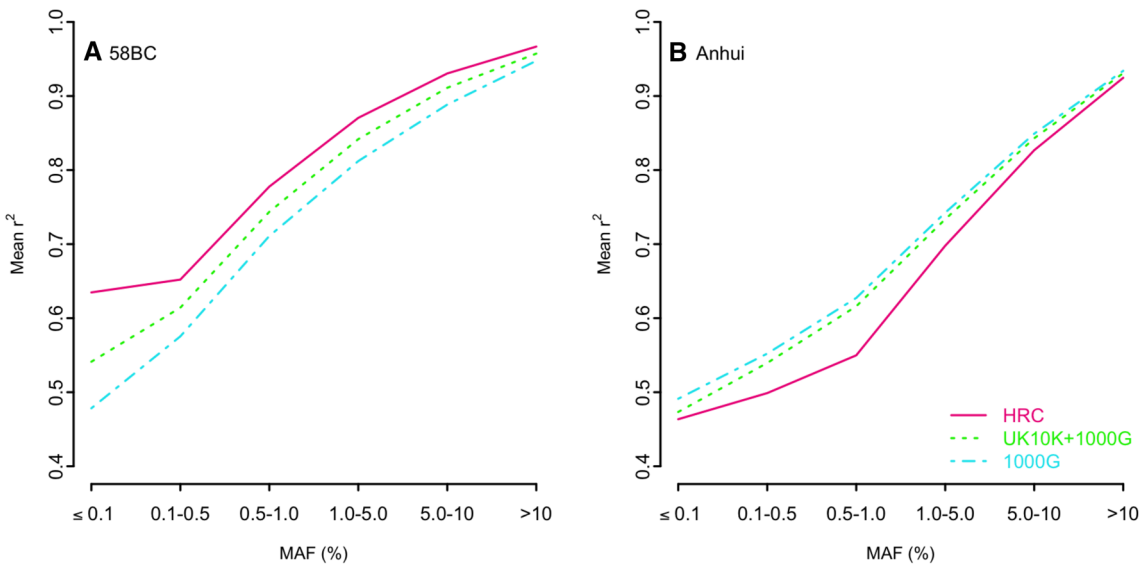

Fig. 1 Imputation performance of three reference panels in the 58 BC and the Anhui cohorts (pre-phased by SHAPEIT v2). **(a)** In the 58 BC cohort of Europeans. **b** In the Anhui cohort of Han Chinese samples. *MAF* minor allele frequency, *HRC* the Haplotype Reference Consortium reference panel (release 1.1), *1000G* the 1000 Genomes Projects Phase III reference panel; *UK10K + 1000G* UK10K plus 1000 Genomes Project Phase III reference panel. This figure was plotted based on the imputation qualities for the overlapped 30,099,640 genetic variants in the imputation results of three reference panels. The *y*-axis indicates the mean imputation quality r^2

(Fig. 2). Lastly, we randomly masked the genotypes of 100,000 common SNPs in both Anhui and Beijing data sets, one at a time, and imputed the masked genotypes using our imputation pipeline. We computed the Pearson r^2 between the true and the imputed genotypes as a measure of imputation accuracy. Again, we observed the same pattern (Fig. S3).

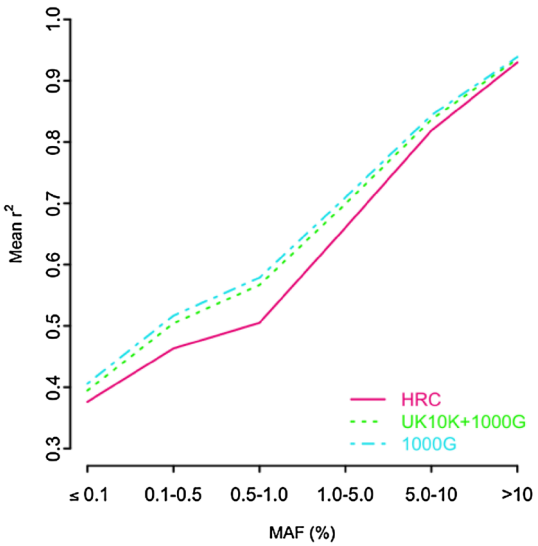

Fig. 2 Imputation performance of three reference panels in the Beijing cohort (pre-phased by SHAPEIT v2). This figure was plotted based on the imputation qualities for the overlapped 30,099,640 genetic variants in the imputation results of three reference panels. The *y*-axis indicates the mean imputation quality r^2

Not only did the 1000G reference panel result in higher imputation qualities in the Han Chinese datasets, it also produced more and more well-imputed variants for both the Han Chinese population and the European population. The 1000G reference panel contains much more SNPs than the HRC panel, 81,706,022 vs. 38,913,048 to be more specific. In particular, there are 10,053,291 East Asian population-specific variants available in the 1000G panel, but absent in the HRC panel (Table S3). Thus, using the 1000G reference panel resulted in 2,217,856 and 2,929,358 variants with MAF ≥ 0.1% being additionally imputed in the 58 BC and the Anhui data sets, respectively; of these variants, 1,283,315 and 1,127,653 were well-imputed with INFO $r^2 \geq 0.8$ (Fig. 3 and Fig. S4). Admittedly, the HRC panel contains exclusive genetic sites too, but the number of exclusive variants is relatively small (only 8,373,325) and most of them (99.2%, 8,305,075/8,373,325) are rare variants (MAF < 0.5%) (Fig. S5). The 1000G reference uniquely imputed variants contain 1439 or 16.5% of the total 8,741 putative functional causal variants for autoimmune diseases (Farh et al. 2015), while only 11 or 0.13% of these variants were uniquely imputed using the HRC panel. The variants that were not imputed would be missed for downstream association analysis.

Discussion

The present study provides evidence for the utilities of the HRC and the 1000G reference panels in current genotype imputation practices but also raises cautions about the

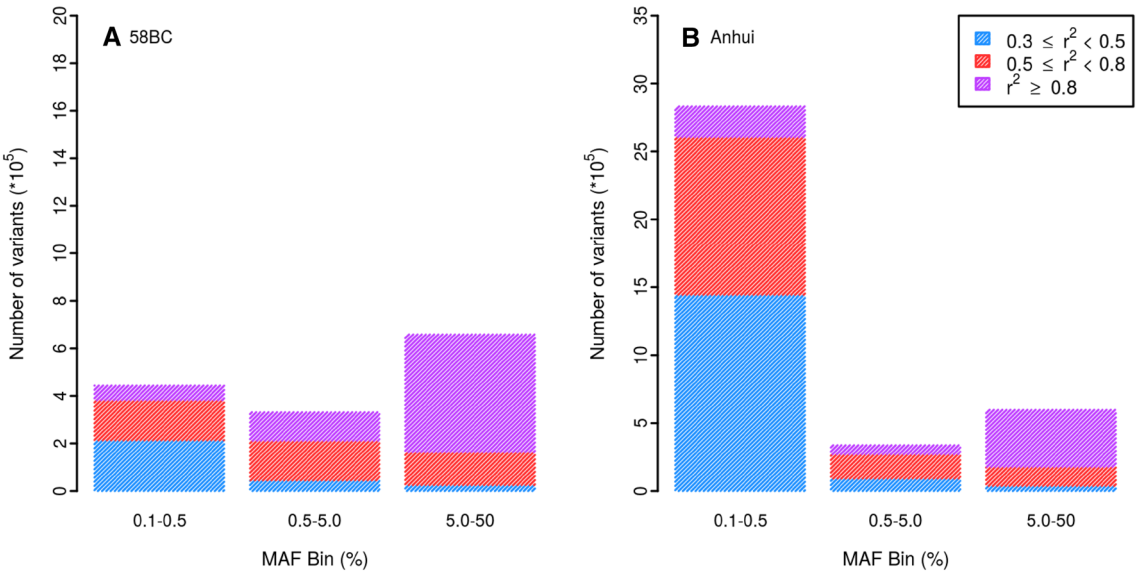

Fig. 3 Distribution of additional imputed variants in the 1000 Genomes Project reference panel. **a** In the 58 BC cohort of Europeans; **b** In the Anhui cohort of Han Chinese samples. *MAF* minor allele frequency, r^2 imputation INFO r^2 value. The SHAPEIT v2 was used to pre-phase the haplotype before genomic imputation. The study-specific MAF was used to group genetic variants in this figure

application of the HRC panel in the Han Chinese population. Genotype imputation as a statistical technique has been routinely adopted in GWASs. The choice of reference panel is crucial for imputation accuracy, as well as the number and the spectrum of imputed variants, which subsequently influence downstream analysis. Some study claims that reference panel size could be more important than ancestry similarity between the study samples and the reference samples for improving imputation accuracy (Howie et al. 2011). For that reason, many recent efforts have been made to increase the size of reference panels by merging existing data sets regardless of the samples' ancestries. As the most recent and the largest reference panel, HRC v1.1 resulted in more accurate imputation for European samples than panels such as 1000G, with greater accuracy at low-frequency variants (McCarthy et al. 2016).

Our independent imputation study on the 58 BC cohort of European samples confirmed the above findings. When we added UK10K samples to the 1000G reference panel or used the HRC reference panel that has much more European samples than the 1000G panel, the imputation qualities of variants across a wide span of MAFs were remarkably improved. However, for the Chinese samples, we found the 1000G reference panel performed significantly better than the HRC reference panel, in terms of both imputation quality and the number of well-imputed variants, and such outperformance was consistent across the two Chinese cohorts we attempted. Although the underlying reasons are yet to be elucidated and beyond the scope of this paper, we speculate the phenomenon might be caused by the haplotype phasing biased against Chinese samples due to the predominance of Europeans in the HRC panel or the relatively low variant density of the HRC panel compared to the 1000G panel. If inaccurate phasing was the sole problem, HRC would outperform 1000G on all counts in imputing the 58 BC cohort, as it contains much more European samples and probably more European-specific haplotypes. However, our study showed that imputation with the 1000G reference panel produces more well-imputed variants not only for the two Chinese cohorts, but also for the European cohort. Since the HRC panel only kept variants with minor allele counts great than 5, a large number of rare genetic variants, particularly non-European population-specific SNPs, could be underrepresented in or excluded from the HRC panel, but well represented in the 1000G panel. We demonstrated that a lot more functional causal variants for autoimmune diseases would be missed in genotype imputation using the HRC reference panel than using the 1000G panel. During the comparison, we used the SNPs shared by the 58 BC and the Anhui datasets as backbone for genotype imputation to avoid introducing bias through the SNP density difference of the datasets being imputed (Nelson et al. 2013). It is worth mentioning that both HRC and 1000G did a less satisfactory job on imputing Chinese samples than they did on imputing European samples. As shown in Fig. 1, the best imputation qualifies achieved for variants with MAF $\geq$ 5% in the Chinese cohort was onefold lower than in the European cohort.

Based on these results, we have a few suggestions for future genotype imputation practices. First, when imputing samples with European ancestry, we would recommend the HRC panel for higher imputation quality, and the 1000G reference panel for a larger number of imputed variants. Second, for imputation of Chinese samples, we would recommend using the 1000G Phase III reference panel over HRC v1.1. Using the much larger HRC panel would incur substantially more computational costs without noticeable gains. An on-going effort of Phase 2 of HRC will contain a large scale of whole-genome sequencing data for Chinese population (Cai et al. 2017). Improved imputation accuracy would be expected in genotype imputation for Chinese cohorts using the new phase of HRC reference panel. In addition, we also call for collaborations among government, academia, and industry to build a large-scale Chinese population-specific reference panel. The benefits of a population-specific reference panel have been suggested in genotype imputation despite the availability of HRC reference panel (Zhou et al. 2017). This is feasible as more and more whole-genome sequencing data of Chinese samples have been generated and made available (Cai et al. 2017). A large Chinese specific reference panel would not only improve the overall imputation accuracy, but also provide good estimation for population-specific variants.

Acknowledgements We acknowledge data sharing from European Genome-phenome archive (EGAD00000000024). This research was funded by Normal Project (81370044), Youth Project (81000692), Key Project of National Natural Science Foundation of China (81130031), Anhui Excellent Youth Fund (1808085J08), Anhui High Education Talent Fund (X. Yin), Anhui Medical University Ph.D. Fund (XJ201429). We thank Dr. Kimberly Robasky at University of North Carolina at Chapel Hill for valuable comments.

Author contributions XYY and YL conceived the study. XYY and YL designed the research strategy and conducted the analysis. SY, DXL, and XJZ participated in sample procurement. XYY and YL prepared the manuscript.

Conflict of interest The authors declare that they have no conflict of interest.

References

Auer PL et al (2012) Imputation of exome sequence variants into population-based samples and blood–cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing. Project Am J Hum Genet 91:794–808. https://doi.org/10.1016/j.ajhg.2012.08.031

Auton A et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Cai N et al (2017) 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Sci Data 4:170011. https://doi.org/10.1038/sdata.2017.11

Chou WC et al (2016) A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. Sci Rep 6:39313. https://doi.org/10.1038/srep39313

Craddock N et al (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3000 shared controls. Nature 464:713–720. https://doi.org/10.1038/nature08979

De Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:R122-R128

Delaneau O, Zagury JF, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10:5–6. https://doi.org/10.1038/nmeth.2307

Duan Q et al (2013) Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. Bioinformatics. (Oxford England) 29:2744–2749. https://doi.org/10.1093/bioinformatics/btt477

Durbin R (2014) Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics 30:1266–1272. https://doi.org/10.1093/bioinformatics/btu014

Farh KK et al (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518:337–343. https://doi.org/10.1038/nature13835

Frazer KA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861. https://doi.org/10.1038/nature06258

Fuchsberger C et al (2016) The genetic architecture of type 2 diabetes. Nature 536:41–47. https://doi.org/10.1038/nature18642

Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46:818–825 https://doi.org/10.1038/ng.3021

Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. PLoS genetics 4:e1000279

Gudbjartsson DF et al (2015) Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47:435–444. https://doi.org/10.1038/ng.3247

Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes G3 (Bethesda) 1:457–470 https://doi.org/10.1534/g3.111.001198

Huang J et al (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 6:8111. https://doi.org/10.1038/ncomms9111

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010a) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834. https://doi.org/10.1002/gepi.20533

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010b) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiol 34:816–834

Liu JZ et al (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 42:436–440

Loh PR et al (2016) Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. https://doi.org/10.1038/ng.3679

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511. https://doi.org/10.1038/nrg2796

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

McCarthy S et al (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48:1279–1283. https://doi.org/10.1038/ng.3643

Nelson SC et al (2013) Imputation-based genomic coverage assessments of current human genotyping arrays G3: Genes, Genomes, Genetics:g3. 113.007161

Sidore C et al (2015) Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet 47:1272–1281. https://doi.org/10.1038/ng.3368

Teo YY et al (2009) Singapore genome variation project: a Haplotype map of three Southeast Asian populations. Genome Res 19:2154–2162. https://doi.org/10.1101/gr.095000.109

Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, Duggal P (2018) Genotype imputation performance of three reference panels using African ancestry individuals. Hum Genet 137:281–292. https://doi.org/10.1007/s00439-018-1881-4

Wu C et al (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nat Genet 43:679–684. https://doi.org/10.1038/ng.849

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zhang XJ et al (2009) Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. Nat Genet 41:205–210. https://doi.org/10.1038/ng.310

Zhou W et al (2017) Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. Genetic Epidemiol 41:744–755. https://doi.org/10.1002/gepi.22067