Original Paper

Human Heredity

Hum Hered 2012;73:18–25
DOI: 10.1159/000334084

Received: April 20, 2011
Accepted after revision: September 24, 2011
Published online: December 30, 2011

Performance of Genotype Imputations Using Data from the 1000 Genomes Project

Yun Ju Sung[a] Lihua Wang[a] Tuomo Rankinen[b] Claude Bouchard[b] D.C. Rao[a]

[a]Division of Biostatistics, School of Medicine, Washington University in St. Louis, St. Louis, Mo., and
[b]Human Genomics Laboratory, Pennington Biomedical Research Center, Baton Rouge, La., USA

Key Words
1000 Genomes Project · HapMap Project · Genome-wide association study · Imputation performance

Abstract
Genotype imputations based on 1000 Genomes (1KG) Project data have the advantage of imputing many more SNPs than imputations based on HapMap data. It also provides an opportunity to discover associations with relatively rare variants. Recent investigations are increasingly using 1KG data for genotype imputations, but only limited evaluations of the performance of this approach are available. In this paper, we empirically evaluated imputation performance using 1KG data by comparing imputation results to those using the HapMap Phase II data that have been widely used. We used three reference panels: the CEU panel consisting of 120 haplotypes from HapMap II and 1KG data (June 2010 release) and the EUR panel consisting of 566 haplotypes also from 1KG data (August 2010 release). We used Illumina 324,607 autosomal SNPs genotyped in 501 individuals of European ancestry. Our most important finding was that both 1KG reference panels provided much higher imputation yield than the HapMap II panel. There were more than twice as many successfully imputed SNPs as there were using the HapMap II panel (6.7 million vs. 2.5 million). Our second most important finding was that accuracy using both 1KG panels was high and almost identical to accuracy using the HapMap II panel. Furthermore, after removing SNPs with MACH Rsq <0.3, accuracy for both rare and low frequency SNPs was very high and almost identical to accuracy for common SNPs. We found that imputation using the 1KG-EUR panel had advantages in successfully imputing rare, low frequency and common variants. Our findings suggest that 1KG-based imputation can increase the opportunity to discover significant associations for SNPs across the allele frequency spectrum. Because the 1KG Project is still underway, we expect that later versions will provide even better imputation performance.

Introduction

Genome-wide association (GWA) studies and meta-analyses have been successful in discovering common variants influencing many complex traits (http://www.genome.gov/gwastudies/) [1]. Genotype imputations provide inferences of untyped markers in a study sample by using the linkage disequilibrium among markers present in an outside reference panel, such as those from the HapMap Project. The public data from the International HapMap Consortium [2] provides over 3.1 million SNPs across populations. Typical GWA studies directly genotype fewer SNPs than HapMap SNPs and impute the untyped markers that were present in the HapMap panel [3, 4]. Therefore, imputation provides a

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/12/0731–0018$38.00/0

Accessible online at:
www.karger.com/hhe

Dr. Yun Ju Sung
Division of Biostatistics, Washington University School of Medicine
660 South Euclid Avenue, Campus Box 8067
St. Louis, MO 63110-1093 (USA)
Tel. +1 314 362 0053, E-Mail yunju@wubios.wustl.edu

higher resolution of association regions and can boost power when compared to using only the genotyped SNPs [5, 6]. Furthermore, when different studies use different genotyping platforms, imputation with a common set of SNPs, such as the HapMap reference panel, makes it possible for different studies to be used in meta-analysis (e.g. [7]).

Most imputations have used HapMap Phase II (HMII) data as a reference panel using several programs such as IMPUTE [8], MACH [9–11], BIMBAM [12] or BEAGLE [13, 14]. However, data from the 1000 Genomes (1KG) Project recently became available and provide a higher resolution of human genome sequence variation [15]. Using 1KG data as the reference panel, more genetic variants in GWA studies can be imputed than using HapMap data. Several recent papers have used 1KG data for imputation [16]. Using 1KG-based imputations with MACH, Sanna et al. [17] discovered association of *CBLB* gene variants with multiple sclerosis. Similarly using 1KG-based imputations with IMPUTE, Liu et al. [18] confirmed an effect on smoking quantity at a locus on 15q25. Using 1KG-based imputations, these and other investigations were able to observe stronger associations with SNPs that were not available in HMII data.

The major advantage of imputations based on the 1KG data, instead of the HapMap data, is the ability to impute a much larger number of SNPs. The 1KG Project aims to discover more than 95% of the variants with minor allele frequencies (MAF) as low as 0.01 across the genome and from 0.001 to 0.005 in gene regions [15]. Therefore, 1KG-based imputations should provide many more variants that are rare and low frequency than HapMap-based imputations. Browning [19] suggested that GWA studies using both imputed and observed genotypes increased the power for detecting rare causal variants. Li et al. [11, 20] also have shown that gain in power with imputation can be higher for rare variants than for common variants: for common variants (simulated MAF = 0.5) power slightly increased from 93.0 to 96.4%, whereas for rare variants (simulated MAF = 0.025) power dramatically increased from 24.4 to 56.2%. However, there are concerns about the data from the 1KG Project due to its potentially lower quality. The HapMap data were based on direct genotyping of previously discovered SNPs and have been thoroughly scrutinized, whereas currently available 1KG data were based on low-depth whole-genome sequence data and, hence, are expected to be of lower quality. Furthermore, only limited evaluations of the performance of 1KG-based imputations are available [21, 22].

In this paper, we empirically evaluated imputation performance using the 1KG data by comparing imputation results to those using the HMII data that have been widely used. Because the most used reference panel for imputing individuals of European ancestry is the CEU panel consisting of 120 haplotypes constructed from HMII data, we used two versions of reference panels constructed from the 1KG data. The CEU panel consisted of 120 haplotypes constructed from the 1KG sequencing data (June 2010 release) of the same 60 individuals as for CEU HapMap data. The EUR panel consisted of 566 haplotypes constructed from the 1KG sequencing data (August 2010 release) of 283 individuals of European ancestry. Due to the increased number of haplotypes of the EUR panel, it contains a substantially larger number of variants (11.4 million) than the other two panels. In particular, since both 1KG panels contain a large number of rare and low frequency variants, we investigated whether 1KG-based imputations provided good performance for these rare and low frequency variants.

Materials and Methods

Study Sample

We used Illumina 324,607 autosomal SNPs genotyped in 501 Caucasian individuals in the HERITAGE Family Study [23, 24]. The study recruited 503 Caucasian individuals from 99 nuclear families in the United States and Canada to investigate the genetic basis of cardiovascular and metabolic responses to exercise training. The participants were in good health and sedentary. GWA genotyping was performed using the Illumina HumanCNV370-Quad v3.0 BeadChips on Illumina BeadStation 500GX platform. The genotype calls were done with the Illumina GenomeStudio software and all samples were called in the same batch to eliminate batch-to-batch variation. Monomorphic SNPs and SNPs with only one heterozygote, as well as SNPs with more than 30% missing data were filtered out with GenomeStudio. Twelve samples were genotyped twice with 100% reproducibility across all SNPs. All GenomeStudio genotype calls with a GenTrain score <0.885 were checked and confirmed manually. Quality control of the GWA study SNP data confirmed all family relationships and found no evidence of DNA sample mix-ups. This study obtained informed consent from participants, and approval from the appropriate institutional review boards.

Reference Panels from the 1KG and HapMap

To impute individuals of European ancestry, we used 3 reference panels. The first panel, denoted by HMII-CEU, is the CEU reference panel consisting of 120 haplotypes constructed from HMII data (release 22, build 36). The second panel, denoted by 1KG-CEU, is the CEU panel consisting of 120 haplotypes constructed from the 1KG sequencing data (June 2010 release) of the same 60 individuals as for HMII-CEU. The third panel, denoted by 1KG-EUR, is the EUR panel consisting of 566 haplotypes con-

Table 1. Number of SNPs across the MAF spectrum in the three reference panels from HMII and 1KG projects

Reference panel	Rare		Low frequency		Common		Total	
	n	%	n	%	n	%	n	%
HMII-CEU	103,252	4.1	284,178	11.2	2,156,427	84.8	2,543,857	100.0
1KG-CEU	15,300	0.2	1,556,959	22.7	5,285,796	77.1	6,858,055	100.0
1KG-EUR	3,646,988	31.8	2,246,325	19.6	5,564,544	48.6	11,457,857	100.0

Rare variants: MAF ≤0.01; low frequency variants: 0.01 < MAF ≤ 0.05; common variants: MAF >0.05.

structed from the 1KG sequencing data (August 2010 release) of the 283 individuals of European ancestry. Both 1KG-CEU and 1KG-EUR were obtained from http://www.sph.umich.edu/csg/abecasis/MACH/download/, as directed by the 1KG Project. The 1KG-EUR is the most recent QC-ed reference panel of European ancestry from the 1KG Project.

The HapMap Project aimed to discover common variants (MAF >0.05) and the HMII-CEU panel was constructed using genotype data of SNPs that were previously discovered. In contrast, the 1KG Project aims to discover rare (MAF ≤0.01) and low frequency (0.01 <MAF ≤0.05) variants, and both 1KG-CEU and 1KG-EUR were constructed using low-depth whole-genome sequencing data, obtained with next-generation sequencers such as ABI's SOLiD, Roche's 454 and Illumina's Genome Analyzer. As shown in figure 1, HMII-CEU, 1KG-CEU and 1KG-EUR contained 2.5 million, 6.8 million and 11.4 million SNPs, respectively. Most SNPs in HMII-CEU panel were also present in 1KG panels.

The 1KG data contained a much larger number of variants than the HapMap data across the MAF spectrum. With sequencing data, 1KG-CEU contained a large number of additional low frequency and common SNPs that were not present in HMII-CEU (shown in table 1). Furthermore, with sequencing data from more individuals, 1KG-EUR contained an even larger number of additional rare and low frequency SNPs that were not present in either HMII-CEU or 1KG-CEU.

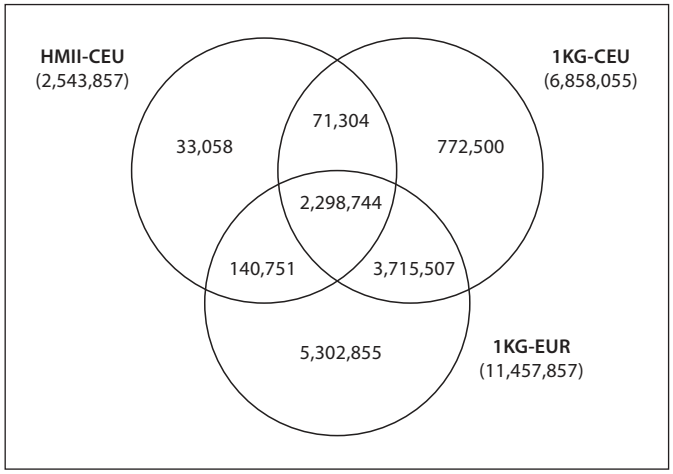

Fig. 1. Venn diagram showing 14,163,455 SNPs on chromosomes 1 through 22 across the three reference panels from HMII and 1KG Projects. For the overlap between 1KG-CEU and 1KG-EUR, the hg18 map positions of 1KG-CEU were converted into hg19 positions, using the liftOver program on the UCSC Genome Browser web site.

Imputations Using MACH

In this paper, we used MACH [10, 11] because MACH and IMPUTE [8], the two leading programs, have been shown to provide the most accurate results across various settings [25, 26]. Before imputation, the strand of GWA study SNPs that did not match with HapMap SNPs were flipped by using PLINK [27]. Following the developers' recommendation, we used a two-step procedure for running MACH (version 1.0.16). In the first step, the crossover map and error rate map were estimated using 50 rounds of iterations and 188 unrelated individuals. Then all subjects were used for genotype imputation in the second step. Imputations were performed separately by chromosomes.

To evaluate imputation performance, we masked 5, 25, 50 and 75% of SNPs in the HERITAGE GWA study. These masked SNPs were removed and imputation was performed using the remaining SNPs. Then the imputed results for these masked SNPs were compared with their real genotyped data to get imputation accuracy. To mask 5% of data, we removed an SNP in every 20th position in a physical map. Other masking was done similarly. We considered much higher masking rates than other papers that investigated imputation performance because we wanted to assess imputation performance for less desirable conditions, in particular, using the 1KG data. Most remaining SNPs that were used for imputation were present in all three reference panels (table 2). This is expected because most Illumina SNPs genotyped for the HERITAGE individuals were selected from HapMap SNPs. The panel 1KG-EUR contained a slightly smaller number of GWA SNPs than 1KG-CEU (317,110 vs. 320,366), although it included almost twice as many SNPs as shown in table 1. We observed that most of these differences occurred because the two panels sometimes used different rs IDs when there were multiple rs IDs at the same position.

To evaluate imputation performance, we used imputation yield and accuracy for each imputed data set and concordance among the three reference panels. Following the developers' recommendation, we applied a filtering rule that removed monomorphic SNPs and SNPs with MACH's quality measure Rsq <0.3.

Table 2. Number of Illumina SNPs in the HERITAGE original GWA and masked datasets and overlaps with the reference panels

Masking rate	Masked SNPs	Remaining SNPs	SNPs also in reference panels*		
			HMII-CEU	1KG-CEU	1KG-EUR
Original data	0	324,607	314,015	320,366	317,110
5%	16,231	308,376	298,308	304,348	301,293
25%	81,152	243,455	235,475	240,246	237,931
50%	162,304	162,303	157,010	160,139	158,649
75%	243,455	81,152	78,540	80,120	79,179

* SNPs also in reference panels correspond to the remaining HERITAGE SNPs that are used for imputation.

Table 3. Imputation yield, number of filtered SNPs (with MACH Rsq >0.3) across the MAF spectrum

Masking rate	Panel	Rare	Low frequency	Common	Total
5%	HMII-CEU	48,136	269,239	2,148,288	2,465,663
	1KG-CEU	67,122	982,959	4,941,873	5,991,954
	1KG-EUR	**394,061**	**1,228,568**	**5,120,445**	**6,743,074**
25%	HMII-CEU	45,207	263,495	2,138,067	2,446,769
	1KG-CEU	56,370	915,119	4,855,531	5,827,020
	1KG-EUR	**302,363**	**1,092,582**	**5,016,263**	**6,411,208**
50%	HMII-CEU	38,065	243,455	2,105,485	2,387,005
	1KG-CEU	35,880	743,918	4,623,292	5,403,090
	1KG-EUR	**149,272**	**788,592**	**4,715,170**	**5,653,034**
75%	HMII-CEU	**17,985**	165,944	1,893,608	2,077,537
	1KG-CEU	7,167	**315,203**	**3,559,825**	**3,882,195**
	1KG-EUR	15,717	233,130	3,352,765	3,601,612

Values in bold denote the highest accuracy rates for each masked data set. Figure 2 and supplementary figure 1 show the number of all imputed SNPs in gray color.

Rare variants: MAF ≤0.01; low frequency variants: 0.01 < MAF ≤ 0.05; common variants: MAF >0.05.

Hence, non-monomorphic SNPs with Rsq ≥0.3 were considered as successfully imputed SNPs with appropriate quality. This also corresponds to the current common practice for numerous imputed data sets that were used in GWA and meta-analysis published papers. We defined *imputation yield* as the number of filtered SNPs that remained after removing SNPs with low imputation quality measures.

We measured *imputation accuracy* using dosage Rsq: the squared correlation between the true genotype and continuous-valued imputed genotype dosage for each masked SNP. Concordance rates between true and imputed genotype calls are often very high for rare and low frequency SNPs and it is hard to compare imputation accuracy across SNPs with different MAFs. On the other hand, dosage Rsq is not confounded by MAF and can be used to compare accuracy for rare and common SNPs. We computed imputation accuracy for each panel using the mean of dosage Rsq values at filtered SNPs. Imputation accuracy was stratified within each MAF bin (rare, low frequency and common SNPs). To evaluate the effect of the filtering rule, we also computed accuracy using the mean of dosage Rsq values at all imputed SNPs.

Results

Imputation Yield across the MAF Spectrum

Imputations using 1KG data provided significantly higher yield (the number of filtered SNPs) than imputa-

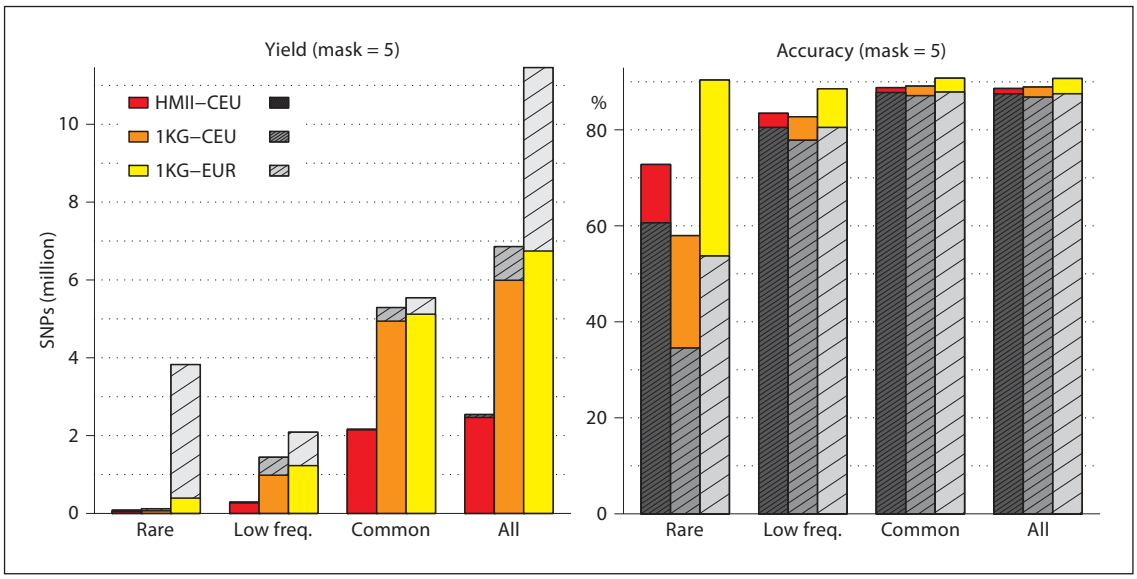

Fig. 2. Imputation yield (left) and accuracy (right) across the MAF spectrum for the 5% masked data using the three reference panels. Colored bars show yield and accuracy using filtered SNPs. Gray bars show total number of imputed SNPs and accuracy using all imputed SNPs. Online supplementary figure 1 shows the same information for all masked data.

tions using HapMap data. Table 3 shows overall yield and yields across the MAF spectrum for all four masked data sets. Figure 2 also shows the total number of imputed SNPs in gray color. Imputation yield was highest with 1KG-EUR, next highest with 1KG-CEU and lowest with HMII-CEU for most masking rates. For the 5% masked data, which corresponded to a typical imputation scenario, imputation yield was 2.5 million, 6.0 million and 6.7 million SNPs using HMII-CEU, 1KG-CEU and 1KG-EUR, respectively. In particular, for 1KG-EUR, only 6.7 million (out of 11.5 million) SNPs were classified as successfully imputed, with 4.7 million SNPs being filtered out (fig. 2). As expected, imputation yield dropped with higher masking rates. However, imputation yield dropped only slightly even for the 50% masked data (2.4 million, 5.4 million and 5.6 million SNPs). Decrease in imputation yield became significant only for the 80% masked data (2.1 million, 3.9 million and 3.6 million SNPs). This indicates a high tagging property of Illumina 330,000 SNPs because with 50% masking over 94% of SNPs in the HMII-CEU were successfully imputed. For 50% and higher masking rates, 1KG-CEU provided higher imputation yield than 1KG-EUR.

Imputation using 1KG-EUR provided the highest yield for rare and low frequency variants for most masking rates (figure 2; table 3). Improvement in imputation yield of 1KG-EUR over the other two panels was different across MAF spectrum. For 5% masked data, imputation using 1KG-EUR provided more than twice as many common SNPs as HMII-CEU (5.1 million vs. 2.1 million), four times as many low frequency SNPs (1.2 million vs. 0.3 million), and eight times as many rare SNPs (0.39 million vs. 0.05 million). Imputation yield using 1KG-CEU was similar to that using 1KG-EUR for common and low frequency SNPs. We expect that with non-masked data, an even larger number of rare and low frequency SNPs would be successfully imputed using the 1KG-EUR panel. However, imputation using 1KG-EUR provided 4.7 million SNPs that were classified as poorly imputed because they had MACH Rsq <0.3. Among these, 3.4 million were rare SNPs. With higher masking rates, imputation yield decreased more for rare and low frequency SNPs (see online supplementary figure 1, www.karger.com/doi/10.1159/000334084).

Imputation Accuracy across the MAF Spectrum

Overall imputation accuracy, the mean of dosage Rsq values at all imputed SNPs, using both 1KG panels was high and almost identical to the accuracy using the HMII-CEU panel for most masking rates. Table 4 shows accuracy before filtering (across all imputed SNPs) and after filtering (across filtered SNPs) for all masked data.

Table 4. Imputation accuracy (%), mean of dosage Rsq values, at all imputed SNPs and filtered SNPs (with MACH Rsq >0.3) across the MAF spectrum

Masking rate	Panel	All imputed SNPs				Filtered SNPs			
		rare	low frequency	common	total	rare	low frequency	common	total
5%	HMII-CEU	**60.6**	**80.5**	87.8	87.5	72.8	83.4	88.8	88.6
	1KG-CEU	34.6	77.8	87.1	86.8	58.0	82.7	89.1	88.9
	1KG-EUR	53.7	**80.5**	**87.9**	**87.5**	**90.4**	**88.5**	**90.8**	**90.7**
25%	HMII-CEU	**60.9**	**78.3**	**87.0**	**86.7**	70.8	81.6	88.1	87.9
	1KG-CEU	55.0	76.4	86.4	86.1	67.3	82.5	88.5	88.4
	1KG-EUR	50.4	77.6	86.9	86.5	**85.2**	**87.6**	**90.1**	**90.0**
50%	HMII-CEU	**53.3**	**70.8**	**81.9**	**81.5**	68.7	76.3	83.9	83.6
	1KG-CEU	44.5	68.9	81.2	80.8	61.5	79.0	85.0	84.8
	1KG-EUR	44.8	68.6	81.0	80.5	**82.2**	**84.6**	**86.3**	**86.3**
75%	HMII-CEU	**39.3**	**51.9**	**65.4**	**65.0**	64.0	67.3	73.2	73.1
	1KG-CEU	27.5	48.1	63.7	63.2	55.0	74.7	76.8	76.7
	1KG-EUR	27.3	45.5	61.6	61.0	**85.1**	**81.5**	**77.7**	**77.7**

Values in bold denote the highest accuracy rates for each masked data set.
Rare variants: MAF ≤0.01; low frequency variants: 0.01 < MAF ≤ 0.05; common variants: MAF >0.05.

Figure 2 shows accuracy before filtering (in gray) and after filtering (in color) for the 5% masked data. Before filtering, accuracy rates were 87.5, 86.8 and 87.5% using HMII-CEU, 1KG-CEU and 1KG-EUR, respectively, for the 5% masked data. Accuracy for common SNPs was similar to overall accuracy. Accuracy for rare variants was highest with HMII-CEU (61%), lowest with 1KG-CEU (35%) and in the middle with 1KG-EUR (54%).

After filtering SNPs with MACH Rsq <0.3, imputation with 1KG-EUR provided the highest imputation accuracy (90.7%). Accuracy was slightly lower with HMII-CEU and 1KG-CEU (88.6 and 88.9%). In particular, accuracy with 1KG-EUR was much improved for rare and low frequency variants (54 and 90% before and after filtering for the 5% masked data). Patterns were very similar across levels of masking (table 3; online suppl. fig. 1). This is very surprising because accuracy using HMII-CEU or 1KG-CEU was not improved nearly as much by the filtering rule.

We investigated why these three reference panels behaved differently with the same filtering rule. A possible explanation for their different behavior is shown in figure 3. The MACH Rsq value is supposed to be an unbiased estimate of the true dosage Rsq value, but this does not seem to be the case for the 1KG reference panels. The black line in the three panels is what the regression line would be if the MACH Rsq was actually unbiased. The magenta line is the actual regression line and the difference between the two is bias. Though there is little bias with the HMII-CEU, there is more with 1KG-CEU and still more with 1KG-EUR. Using the rule of removing SNPs with MACH Rsq <0.3 with the 1KG-EUR reference panel approximately corresponded to removing SNPs with true dosage Rsq <0.5. Because accuracy corresponds to the mean of dosage Rsq values at the filtered SNPs, this may account for the better performance of this filtering rule. This bias increased with higher masking rates.

Discussion

In this paper, we evaluated the performance of genotype imputations using the 1KG Project data, relative to imputations using the HMII data. Our most important finding was that both 1KG reference panels (1KG-CEU and 1KG-EUR) provided much higher imputation yield than the HMII panel. In particular, there were more than twice as many successfully imputed SNPs using 1KG-EUR as there were using HMII data (6.7 million vs. 2.5 million). There were twice as many common SNPs as HMII-CEU (5.1 million vs. 2.1 million), four times as many low frequency SNPs (1.2 million vs. 0.3 million), and eight times as many rare SNPs (0.39 million vs. 0.05 million). Our second most important finding was that ac-

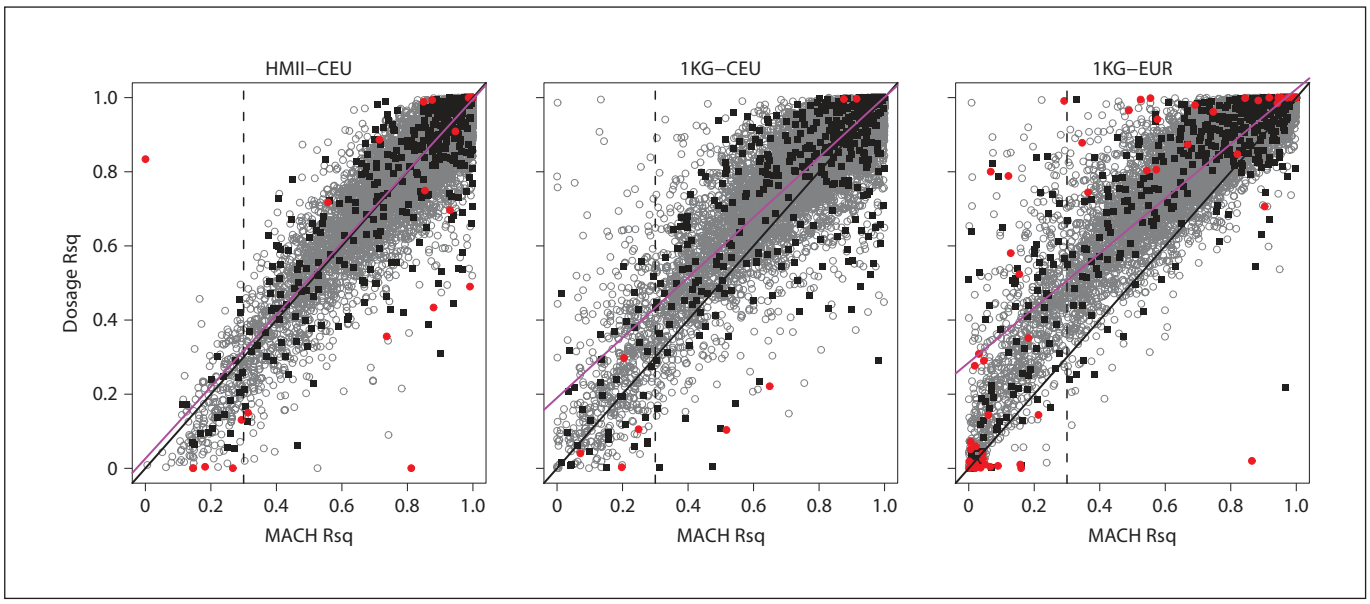

Fig. 3. Imputation accuracy (dosage Rsq) values versus MACH Rsq values for the 5% masked data. Red solid circles are rare SNPs and black solid squares are low frequency SNPs. The black line indicates where MACH Rsq equals dosage Rsq. The magenta line is the regression line. The vertical dashed line is the filtering rule that we used. Imputation accuracy (table 4, fig. 2) was computed as the average of dosage Rsq values shown in the Y-axis. Colors refer to the online version only.

curacy using both 1KG panels was high and almost identical to the accuracy using the HMII-CEU panel. Furthermore, after removing SNPs with MACH Rsq <0.3, accuracy for both rare and low frequency SNPs was very high and almost identical to accuracy for common SNPs.

Despite much higher SNP density, we expected the 1KG-based imputations to be of lower quality than HapMap-based imputations, due to low depth in terms of sequencing read depths (average of 4.6× in CEU). Our results were somewhat consistent in this regard in that 1KG-based imputations had a large fraction of SNPs with MACH Rsq <0.3. However, because the 1KG data contained a substantially larger number of SNPs, there were twice as many retained SNPs using the 1KG data as using the HapMap data. Furthermore, we observed that MACH Rsq values were consistently underestimating true dosage Rsq values for SNPs from 1KG-based imputations. As a consequence, accuracy of retained SNPs was highest with 1KG-based imputations.

We investigated whether using data from the 1KG has any advantage in terms of successfully imputing a large number of rare and low frequency variants. First, we found that rare variants in general had lower MACH Rsq values and were classified as poorly imputed. This was more obvious using the 1KG data. Our findings are consistent with other investigations [25, 28] which showed that rare variants were more difficult to tag than common SNPs and also confirm that imputation programs often perform poorly when imputing rare variants. Second, we found that imputations using 1KG-EUR had an advantage in imputing rare and low frequency variants. Our findings are consistent with those of Jostins et al. [29], who showed that HapMap3-based imputations provided highly accurate imputation of low frequency variants due to large and diverse reference sets [30]. Imputation using 1KG-CEU also had an advantage in imputing low frequency variants. Furthermore, we found that using both 1KG-CEU and 1KG-EUR had a big advantage in imputing a large number of common variants. Our findings suggest that imputation using data from the 1KG can increase the opportunity to discover significant associations for SNPs across the whole MAF spectrum.

In this paper, we used two versions of the 1KG Project Pilot 1 data. We emphasize here that even though these 1KG reference panels were based on low-depth sequence data, to better handle low-depth sequence data, these panels were constructed using three independently developed methods that combine sequence data across samples and HapMap3 data: (1) QCALL [31]; (2) Thunder [22], and (3) DePristo et al. [32] using Genome Analysis

Toolkit (GATK) [33] and BEAGLE [13, 14]. Furthermore, the 1KG Project is still underway, and genotype accuracy will be further improved due to increased sample sizes and a plan to directly genotype variants observed in the low-depth sequencing data. We expect that later versions will provide even better imputation performance.

Acknowledgements

We appreciate two anonymous reviewers for their constructive and insightful comments, which substantially improved the manuscript. The work was partly supported by NIH Grants GM 28719, HL 45670 and HL 54473.

References

1 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–9367.

2 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–861.

3 Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP: Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet 2008;83:112–119.

4 Hao K, Chudin E, McElwee J, Schadt EE: Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet 2009;10:27.

5 Spencer CCA, 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.

6 Marchini J, Howie B: Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499–511.

7 de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF: Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 2008;17:R122–R128.

8 Marchini 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.

9 Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al: A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007; 316:1341–1345.

10 Li Y, Abecasis GR: MACH 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet 2006;S79:2290.

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

12 Servin B, Stephens M: Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 2007;3:e114.

13 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.

14 Browning BL, Browning SR: A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009;84:210–223.

15 The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature 2010;467: 1061–1073.

16 Ellinghaus E, Ellinghaus D, Stuart PE, Nair RP, Debrus S, et al: Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat Genet 2010;42:991–995.

17 Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, et al: Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 2010;42:495–497.

18 Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, et al: Wellcome Trust Case Control Consortium, Mooser V, Francks C, Marchini J: Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 2010;42:436–440.

19 Browning SR: Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 2008;124: 439–450.

20 Li Y, Byrnes AE, Li M: To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet 2010;87:728–735.

21 Fridley BL, Jenkins G, Devo-Svendsen ME, Hebbring S, Freimuth R: Utilizing genotype imputation for the augmentation of sequence data. PLoS One 2010;5:e11018.

22 Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 2011;21:940–951.

23 Bouchard C, Leon AS, Rao DC, Skinner JS, Wilmore JH, Gagnon J: The HERITAGE family study. Aims, design, and measurement protocol. Med Sci Sports Exerc 1995;27: 721–729.

24 Bouchard C, Sarzynski MA, Rice TK, Kraus WE, Church TS, Sung YJ, Rao DC, Rankinen T: Genomic predictors of maximal oxygen uptake response to standardized exercise training programs. J Appl Physiol 2011;110: 1160–1170.

25 Pei YF, Li J, Zhang L, Papasian CJ, Deng HW: Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 2008;3:e3551.

26 Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: A comprehensive evaluation of SNP genotype imputation. Hum Genet 2009;125:163–171.

27 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.

28 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.

29 Jostins L, Morley KI, Barrett JC: Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur J Hum Genet 2011;19:662–666.

30 The International HapMap 3 Consortium: Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52–58.

31 Le SQ, Durbin R: SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res 2011;21:952–960.

32 DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–498.

33 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.