UNIVERSITÀ DEGLI STUDI DI SASSARI

Scuola di Dottorato in Scienze Biomediche

Direttore: Prof. Andrea Piana

XXVII CICLO DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE
INDIRIZZO IN GENETICA MEDICA, MALATTIE METABOLICHE E NUTRIGENOMICA

Responsabile di indirizzo: Prof. Francesco Cucca

Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs

Relatori:

Francesco Cucca

Serena Sanna

Dottorando:

Giorgio Pistis

Anno accademico 2013/2014

The research activity included in this thesis has been conducted during my PhD studies and it has been published in European Journal of Human Genetics. The complete citation of the publication is:

Pistis G., et al, Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs.
Eur J Hum Genet. 2014 Oct 8. doi: 10.1038/ejhg.2014.216. PMID: 25293720

Index

1. Introduction

The utility of genotype imputation in genome-wide association studies is increasing as progressively larger reference panels are improved and expanded through whole-genome sequencing. Developing general guidelines for optimally cost-effective imputation, however, requires evaluation of performance issues that include the relative utility of study-specific compared with general/multi-population reference panels. This study thus provides general guidelines for researchers planning large-scale genetic studies.

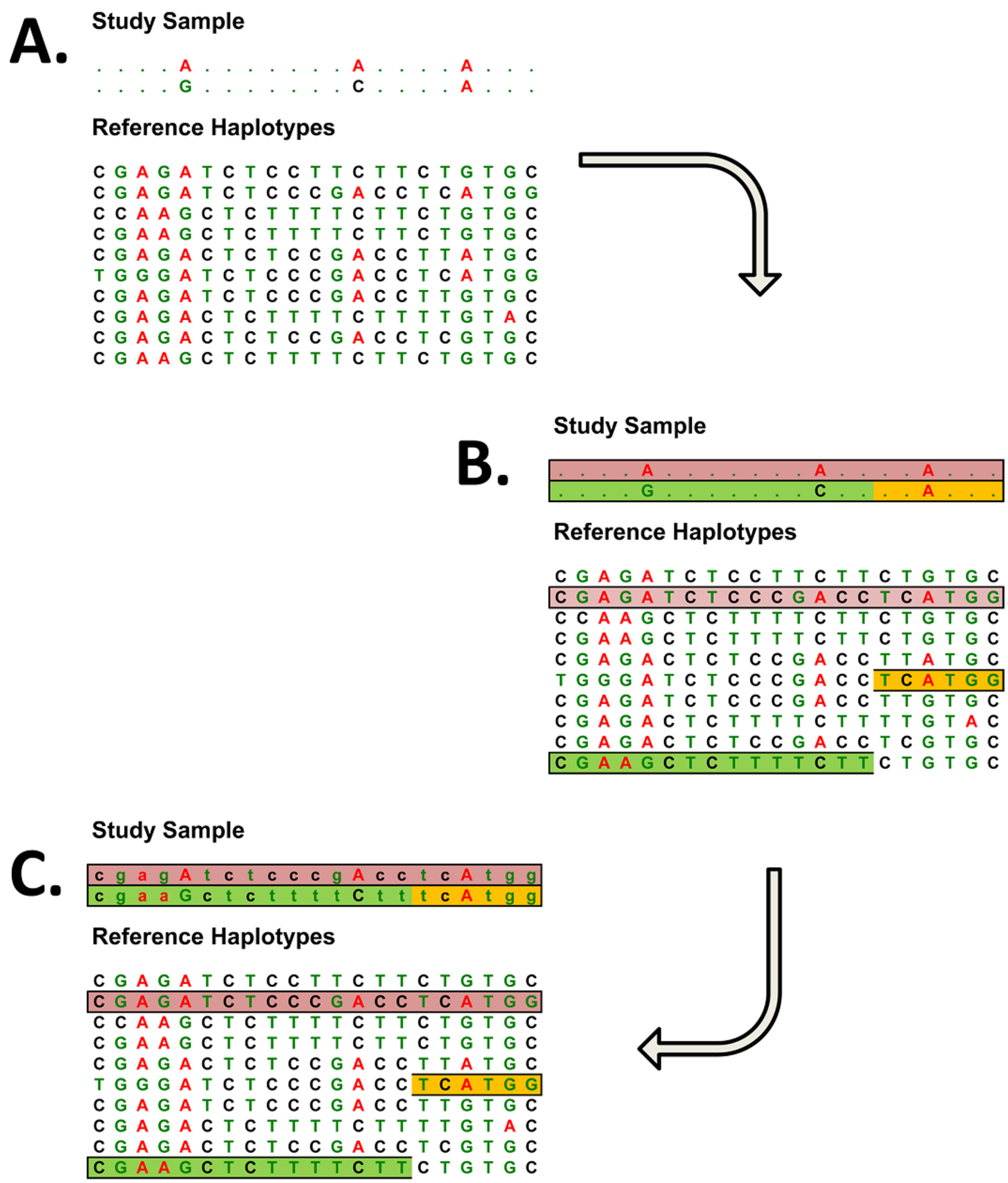
1.1. Genome-wide association studies

Genome-wide association studies (GWASs) have successfully identified thousands of common, single-nucleotide polymorphisms (SNPs) associated with complex traits. The amount of the SNPs assessed in a GWAS is one of the key factors in determining the power of these studies. In fact, the greater the number of markers, the higher the probability to identify a novel association. In the past decade, the majority of GWAS were carried out using a limited number of SNPs experimentally derived by commercial genotyping arrays. However, while the design of such arrays has evolved to target up to 2.5 Million of SNPs, they still survey only a limited repertoire of sequence variation, and underrepresent rare and population specific variants. Much more complete extraction of genetic variation is now accessible using next-generation sequencing (NGS) technologies, but efficient detection of rare and low frequency variants requires sequencing hundreds to thousands of individuals and could be therefore very expensive and so unfeasible for the majority of the studies[1]. An alternative cost-effective approach to enlarge the frequency spectrum of variants assessed in GWASs capitalizes on publicly available sequencing reference panels, especially the 1000 Genomes Project (1000G) reference panels. Indeed, 'probabilistic' sequenced genomes can be reconstructed by means of genotyping imputation methods, inferring untyped variants by combining partial haplotypes found in a study sample with the full haplotypes available in a more densely characterized reference set.

1.2. The genotype imputation method

The term genotype imputation indicates the process of predicting (or imputing) genotypes that are not directly assayed in a sample of individuals. Genotype imputation most often refers to the situation in which a reference panel of haplotypes at a dense set of SNPs is used to impute into a study sample of individuals that have been genotyped at a subset of the SNPs. The fundamental idea is that short stretches of haplotypes can be shared even between unrelated individuals from distant common ancestors. These stretches can be identified using genotypes for a given set of SNPs. Alleles for SNPs that are measured in the reference panel, but not the study samples, can be imputed. In a typical scenario, the study sample is genotyped with a commercial genotyping platform for hundreds of thousands to millions of SNPs located across the entire genome while the reference panel consists in a group of sequenced samples. An overview of this process is given in Figure 1[2]. Several different statistical descriptions of genotype imputation procedures have now been published and implemented in a number of software. In principle, any of the software typically used to estimate missing genotypes is based on a simple heuristic[3], or on an E-M algorithm[4], or on more sophisticated coalescent models[5]. These tools typically provide convenient summaries of the uncertainty surrounding each genotype estimate. The imputation quality is commonly measured with a parameter called Rsq, i.e., the estimate of the squared correlation between imputed and true genotypes or, in other words, the ratio of the variances of imputed and true allele counts. In this context, it should be noted that the accuracy of predicted Rsq values is, in general, high for common variants, but rapid performance degradation is seen for lower minor allele frequencies, thereby limiting the applicability of such methods, especially for rare variants. The performance depends on multiple factors, including: choice of baseline array, quality of input genotypes/haplotypes and limited representation of reference haplotypes carrying rare alleles. Also and very importantly, differences in linkage disequilibrium (LD) patterns and allele frequency spectrum significantly decrease the quality of imputation overall, especially when using public reference panels for ancestral or geographically isolated populations[6,7]. It has, however, been unclear how well general reference panels represent variation in populations that were poorly or not at all represented in projects like 1000 Genomes. Furthermore, even for well-represented populations, a complete evaluation is needed to assess the benefits of sequencing more study samples for successfully imputing rare or low frequency variants.

Figure 1. Schematic representation of genotype imputation method

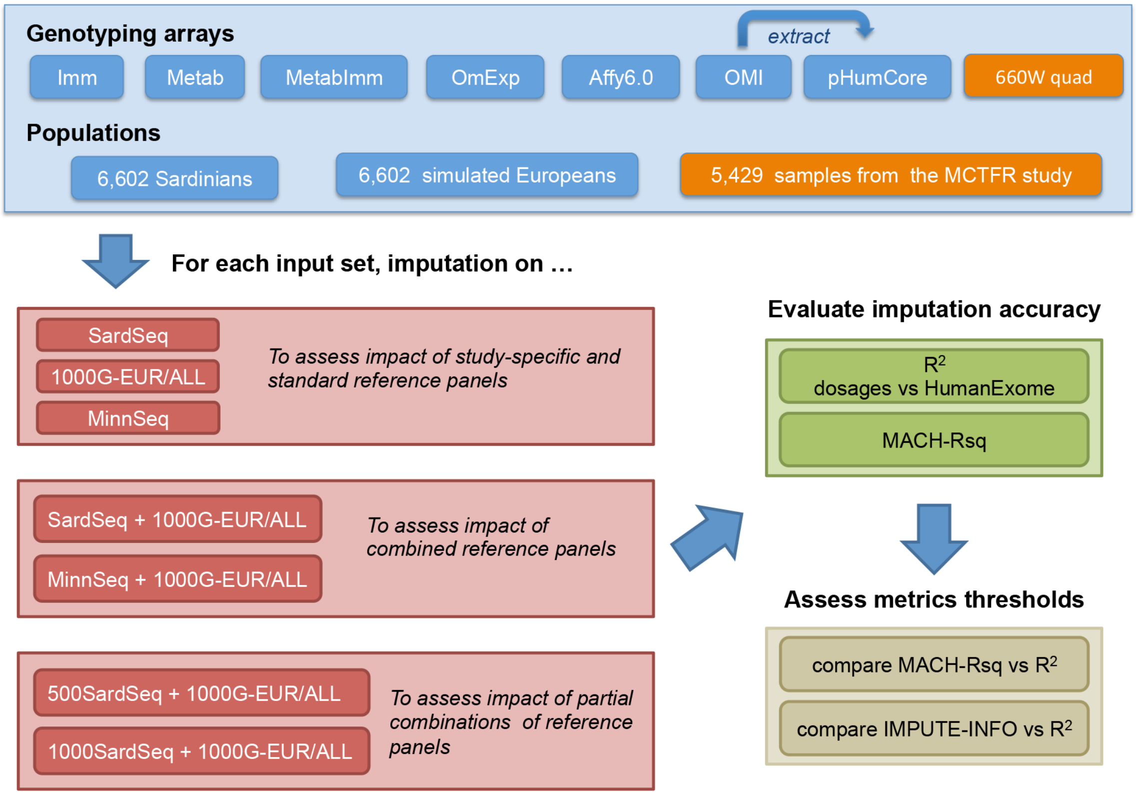


Panel A illustrates the genotypes at a modest number of genetic markers in each sample being studied. Panel B illustrates the process of identifying regions of chromosome shared between a study sample and individuals in the reference panel. In Panel C, observed genotypes and haplotype sharing information have been combined to fill in a series of unobserved genotypes in the study sample. (Figure by Yun Li et al, Annu Rev Genomics Hum Genet. 2009).

1.3. Aim and project overview

How can imputation be further improved? To dissect this question, we compared imputation quality using three complementary sets of reference panels: 1488 Sardinians from Sardinia, Italy; 1325 individuals of Northern European ancestry from Minnesota, USA; and 1092 individuals from the 1000 Genomes project. These reference panels permit comparison of the relative efficiency of study-specific imputation in founder (i.e. Sardinia) and continental (i.e. Northern European) populations that have also been genotyped, and contrast those results with the current standard approach (i.e. 1000 Genomes). We also examined different combinations of genome-wide and custom arrays for baseline genotypes. Finally, we evaluated the efficiency of the conventional quality thresholds to discard poorly imputed rare and low-frequency variants, focusing on metrics defined by the two most commonly used imputation software, MACH[8] and IMPUTE[9]. Figure 2 shows a schematic representation of the study.

Figure 2. Graphical representation of the analyses and study aims



The figure shows a scheme of analyses carried out. For each genotype input set we carried out several imputation runs (genome-wide for SardiNIA, and on chromosome 20 for other European populations) with different reference panels. We assessed imputation quality of each genotyping array/reference panel combination by looking at the mean imputation quality (MACH-Rsq) and by comparing imputed markers with those directly typed with the HumanExome array (R^2). Finally, we assessed the efficiency of standard thresholds at the commonly used accuracy metrics (MACH-Rsq/IMPUTE-INFO) in filtering badly imputed markers.

1.4. The two studied cohorts: SardiNIA and MCTFR

1.4.1. The SardiNIA cohort

The SardiNIA project is a genetic and epidemiological study that aims to identify the biological and genetic mechanisms involved in age-related traits and diseases. It is based in Sardinia, the second largest island in the Mediterranean sea. The targeted area is the Lanusei Valley in the Ogliastra region, located in the middle-east area of Sardinia. The SardiNIA project enrolled 6921 individuals, representing >60% of the adult population of four villages in the Lanusei Valley. All individuals were genotyped using four different Illumina genotyping chip array: the HumanOmniExpress GWAS array, containing ~750,000 markers, and three different custom arrays: the Cardio-MetaboChip, the ImmunoChip and the HumanExome, each containing ~200,000 markers[10,11]. Among the 6921 volunteers, 1122 were also whole-genome sequenced within the SardiNIA Medical Sequencing Discovery Project (dbGaP Study Accession: phs000313.v1.p1) (see ‘Sequencing’ paragraph for further details).

1.4.2. The MCTFR cohort

The Minnesota Center for Twin and Family Research (MCTFR[12,13]) at the University of Minnesota specializes in the use of genetically informative family cohorts to investigate the etiology of behavioral and psychiatric phenotypes. The MCTFR consists of two complementary cohorts. One is a population-based cohort of twins and their parents, and the other is a family adoption study. Volunteers leave in the Minnesota State, and were all of north European origin. The full MCTFR cohort was genotyped with the Illumina 660W-quad array containing ~600,000 markers. The full sample was also genotyped with the Illumina HumanExome array. Furthermore, 1328 individuals from 602 families were whole-genome sequenced (see ‘Sequencing’ paragraph for further details).

Both SardiNIA and the MCTFR studies were approved by the corresponding institutional review boards and a signed informed consent was obtained from every volunteer.

2. Materials and Methods

2.1. Genotyping

In the SardiNIA cohort genotype calling was performed using the Illumina GenCall algorithm (Illumina, San Diego, CA, USA), and an additional 2968 rare variants were called for HumanExome using Zcall[14]. A subset of 1072 samples was also previously genotyped with Affymetrix 6.0 (Affymetrix, Santa Clara, CA, USA)[15]. For the Illumina arrays, we discarded samples with a genotyping call rate <95% in HumanOmniExpress, or <98% in the other arrays. SNP genotypes were carefully assessed though several quality control checks. In particular, we analyzed the four Illumina arrays independently and removed markers with call rate <98%, deviating strongly from Hardy Weinberg Equilibrium ($p < 1x10^{-6}$), monomorphic (or with minor allele frequency (MAF) <1% for HumanOmniExpress), or leading to an excess of Mendelian errors (defined as >1% of the families or >1 for HumanExome SNPs called with Zcall). In addition, we removed SNPs in common between the chips that showed a high level of discordance or that generated >1% discrepancies when comparing genotypes across 13 twin pairs. For full details on array quality checks see Supplementary Table S1. After performing quality control checks we used the quality-checked (QCed) autosomal markers from the HumanOmniExpress, ImmunoChip and Cardio-MetaboChip arrays as baseline genotypes to impute variants detected through sequencing, as described below. In order to have fully comparable data sets for all analyses described here, we considered only the 6602 samples for which all four Illumina arrays were successfully genotyped. Data from the Affymetrix 6.0 array were instead not combined with the Illumina arrays, given the smaller number of samples available (1072 vs 6602); for this set, quality control filters have been already described[16]. From the QCed set of markers, we extracted a subset of 227 745 SNPs representing most of the content of the Illumina HumanCore array (78.9% prior QC), a low-density genome-wide array. Given the extensive overlap, and considering that after quality control filtering the effective content of an array is always reduced, we treated this subset of markers as an approximation of the genomic content accessible with the HumanCore array that we refer to here as ‘pseudo-HumanCore’.

Genotyping protocols and quality control procedure for the MCTFR study have been described previously[13,17]. In short, the full MCTFR study sample was genotyped with the Illumina 660W-quad array, with 7278 (97.8%) samples and 527 829 (94.3%) markers passing quality control filters. The full sample was also genotyped with the Illumina HumanExome array, with 7244 (97.4%) samples and 144 075 (58.1%) markers passing quality control filters. We initially used 6610 individuals of European ancestry, and noticed that the inclusion of the 1181 individuals who were also in the reference panel biased accuracy estimates at rare variants because of perfect match of haplotypes. (Supplementary Table S2). We therefore restricted the analyses to the 5429 samples not overlapping with the reference panel.

2.2. Sequencing

Samples to be sequenced were selected in trios, taking advantage of their highly informative content for haplotypes reconstruction. Trios (or parent–offspring pairs for incomplete trios) were selected starting from the founders of all available families to assure the representation of all haplotypes that have been propagated within families (using ExomePicks, see URLs of Web Resources). For the Sardinians, 2120 samples from 695 nuclear families were sequenced to an average coverage of 4.16-fold. Of these, 1122 samples were part of the SardiNIA project[18], whereas the other 998 were individuals enrolled in case–control studies of multiple sclerosis and type I diabetes[19,20]. The Sardinian samples were sequenced in part at the CRS4 center (Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna) in Pula, Italy, and in part at the DNA Sequencing Core center at the University of Michigan, USA, with Illumina Genome Analyzer IIx, Illumina Hiseq 2000 and Illumina HiSeq 2500 instruments. The sequencing effort has been described in part previously[21]. Sequenced samples were analyzed as single-end or paired-end reads, depending on the success of the paired-end procedure, and the majority were produced by 100+100 cycles of paired-end runs. An average of 4.16-fold coverage was obtained across all sequenced samples. Raw sequencing data was 1) aligned to the GRCh37 assembly available from the 1000 Genomes Project website, using an average

Phred score of 15 as a threshold to accept or trim sequence bases. The resulting SAM file was 2) sorted and 3) indexed, and 4) PCR duplicates were removed. Those steps were implemented with BWA 0.5.9 (Burrows Wheeler Alignment tool)[22] (step 1), SAMTOOLS 0.1.18[23] (step 2 and 3), and MarkDuplicates from Picard Tools 1.52[23] (step 4). For the paired-end data produced at CRS4, steps 1, 2, and 4 were performed using the Seal toolkit version 0.3.1[24], whereas indexing (step 3) was performed with SAMTOOLS 0.1.18. Finally, at both sites quality scores were recalibrated by taking into account alignment information and known polymorphisms in the Single Nucleotide Polymorphisms database release 132 (dbSNP132), using the GATK 1.1-35-ge253f6[25]. Each sample has been verified for sample contamination and swaps by comparing genotype likelihoods in the alignment file with the genotypes available from genotyping arrays using verifyBamID (see URLs of Web Resources).

In the MCTFR study, 1328 individuals from 602 families were sequenced to an average coverage of 10.4-fold. Three samples gave unacceptable sequence quality, leaving 1325 total sequenced samples for analysis. The MCTFR samples were sequenced at the University of Michigan Sequencing Core (1024 samples, with 150bp paired-end reads) and at the HudsonAlpha Institute for Biotechnology (304 samples, with 100bp paired-end reads). Reads were aligned to with BWA-MEM version 0.7.4-r385[26], duplicates removed with Picard version 1.91 (see URLs of Web Resources), and recalibrated with GATK version 1.1-35-ge253f6f. All indexing and sorting was done in SAMTOOLS 0.1.18.

2.3. Variant calling

Variant calling was performed in both studies using GotCloud[27], a variant call pipeline developed at the University of Michigan (see URLs of Web Resources). Briefly, the procedure consisted of detecting an initial set of polymorphic sites (SNP detection) and performing an LD-aware genotype refinement using BEAGLE software[28]. The variant site detection algorithm uses a population based approach to increase the power of calling variants even in sites with shallow coverage. Variants are then screened to remove false positives and mapping artifacts by using a two steps approach. A first step consists of hard filtering based on fixed thresholds applied to coverage depth, strand bias, inflation of other alleles and allele balance. This is followed by a second step of filtering using non-linear thresholds based on training sets (HumanOmni 2.5M and HapMap) and machine learning approach (SVM, support vector machine). After this phase, the genotype of each individual is called according to the recalibrated likelihoods generated by the sequencing. Finally, genotypes are refined with BEAGLE[28], which generates a set of haplotypes for each individual.

Sequencing yielded 17.6 and 27.1 million autosomal bi-allelic SNPs in Sardinian and Minnesota samples, respectively, of which 30.6 and 48.4% were not described in dbSNP135.

2.4. Genotype imputation

Genotype imputations for all scenarios were performed on haploid data using Minimac (see URLs of Web Resources), a modified version of the MACH software. For SardiNIA, phased haplotypes were generated using MACH (-phase option) with 400 states and 30 rounds by subdividing the variants in 344 groups of 2500 with an overlap of 500, and imputation was subsequently performed independently on each phased chunk (for a description of the code, see the '1000G imputation cookbook' in URLs of Web Resources). Imputation performance was evaluated on seven different input genotype data sets: (1) HumanOmniExpress (OmExp), (2) Cardio-MetaboChip (Metab), (3) ImmunoChip (Imm), (4) Cardio-MetaboChip and ImmunoChip (MetabImm), (5) HumanOmniExpress, Cardio-MetaboChip and ImmunoChip (OMI), (6) pseudo-HumanCore (pHumCore) and (7) Affymetrix 6.0 (Affy 6.0). For simplicity, we phased the Cardio-MetaboChip, ImmunoChip and HumanOmniExpress arrays jointly, and then extracted haplotypes at relevant SNPs to perform imputation for each particular genotyping set. In actual practice, Cardio-MetaboChip and ImmunoChip will be phased without the additional support of a genome-wide array, and hence we assessed the impact of our procedure by phasing separately each SNP set, for chromosome 20. We noticed that only imputations performed with the SardSeq panel or its combination with 1000G were slightly overestimated (see Supplementary Table S3, and 'Impact of different phasing strategies' paragraph). In the MCTFR study, haplotypes were phased using SHAPEIT2 (v2.644)[29] with the following model options: -thread 8 -burn 10 -prune 8 -main 20 -states 200. Imputation was performed using Minimac and the Illumina 660W-quad array as baseline genotypes. We used as reference panels the 1000G-ALL (1092 samples) and 1000G-EUR (379 samples) data sets from the 1000 Genomes March 2012 release; the full MCTFR sequencing data (1325 samples, named MinnSeq in the text); a subset of the Sardinian sequencing data (1488 samples, named SardSeq in the text); and combinations of those (see 'Combination of reference panels' paragraph). Considering the overall high inbreeding in Sardinia, the SardSeq reference panel was created by selecting only haplotypes of parents at each sequenced trios to avoid overrepresentation of rare variants. We also performed imputation with IMPUTE2 (newest release of IMPUTE), to test a different approach for reference panels' combination (see 'Combination of reference panels' paragraph) and to assess the efficiency of its imputation accuracy metric INFO (see 'Evaluation of imputation accuracy' paragraph).

2.5. Simulation of European haplotypes

Because the Minnesota samples were genotyped with different arrays from those used for Sardinians, they could not be used to assess relative efficiency of arrays in genotype imputation. We therefore generated, by simulation with the HAPGEN[30] software and 1000G-EUR as reference, 6602 unrelated individuals of European ancestry for SNPs present in each different genotyping array considered in the SardiNIA study. For simplicity, we focused only on chromosome 20. Haplotypes were phased using MACH (-phase option) with 400 states and 30 rounds, and imputation performed using Minimac, as in the SardiNIA and MCTFR data sets. This simulated data set was only used for assessing the efficiency of different genotyping arrays and reference panels in genotype imputation.

2.6. Combination of reference panels

We used VCFtools[31] to combine the SardSeq and the MinnSeq panels with 1000G-EUR and 1000G-ALL reference panels for chromosome 20. The variants in each set were 331 799, 602 317, 851 702 and 377 494 for SardSeq, MinnSeq, 1000G-ALL and 1000G-EUR, respectively. During the merging procedure, we removed the variants present only in one panel, leading to SardSeq+1000G-ALL, SardSeq+1000G-EUR, MinnSeq+1000G-ALL and MinnSeq+1000G-EUR reference panels containing 249 624, 227 405, 304 899 and 267 550 variants, respectively. Imputation was then performed using Minimac, as for single reference panels. For combinations with 1000G and SardSeq panels, we also performed imputation with IMPUTE2 using the -merge_ref_panels option that imputes variants unique to one panel into the other, prior imputation. We observed no difference in imputation accuracy at all frequency ranges when using this approach, which should be preferable for research studies, allowing imputation of all available variants, including those that are study specific, in the same run (Supplementary Table S4). In addition, to assess the impact of adding a smaller number of population-specific haplotypes, we created two additional reference panels using 500 and 1000 randomly chosen samples from the SardSeq reference panel and merging them with 1000G reference panels (500SardSeq+1000G and 1000SardSeq+1000G, respectively). This analysis was restricted to the SardSeq panel and the SardiNIA cohort, because the advantage in accuracy was substantial for this population.

2.7. Evaluation of imputation accuracy

Imputation accuracy was assessed using both the MACH Rsq metric and the squared Pearson's correlation (R^2)[8] between dosages and the real genotypes (considered as allele count) available for the same individuals, extracted from the HumanExome array. The Rsq metric is also known as variance ratio, being calculated as the proportion of the empirically observed variance (based on the imputation) to the expected binomial variance p(1-p), where p is the minor allele frequency. In SardiNIA we tested 21,398

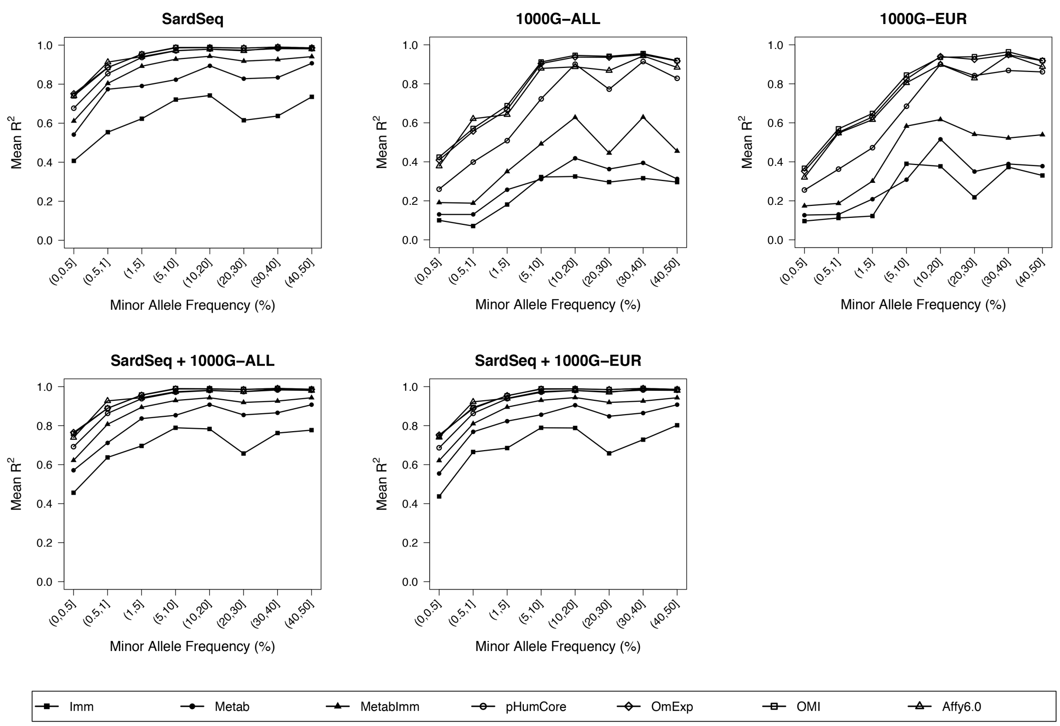
SNPs across autosomes for genome-wide evaluation of imputation accuracy and tested a subset of 558 SNPs for comparisons restricted to chromosome 20. For the MCTFR study, as the baseline array was different, we used a subset of 541 SNPs. The number of SNPs tested for comparing imputation with SardSeq versus 500SardSeq+1000G and 1000SardSeq+1000G was reduced to 517 because 41 SNPs (MAF range 0.0008-0.0072%) were not detected in the selected subset of sequenced samples. We also assessed efficiency in discriminating between well and poorly imputed markers of the imputation accuracy metrics estimated by MACH (Rsq) and IMPUTE (INFO). The INFO metric, also known as imputed information score (INFO), is a measure of the relative statistical information about the SNP allele frequency from the imputed data. We defined good- and bad-quality imputed SNPs as in the original MACH paper, that is, those with $R^2 > 0.5$ and with $R^2 < 0.2$, respectively, and stratified imputed SNPs based on their Rsq and INFO scores. This analysis was restricted to chromosome 20, and performed using as baseline genotypes the OmExp for the SardiNIA study and the Illumina 660W-quad for the MCTFR cohort.

3. Results

3.1. Effect of baseline genotyping array

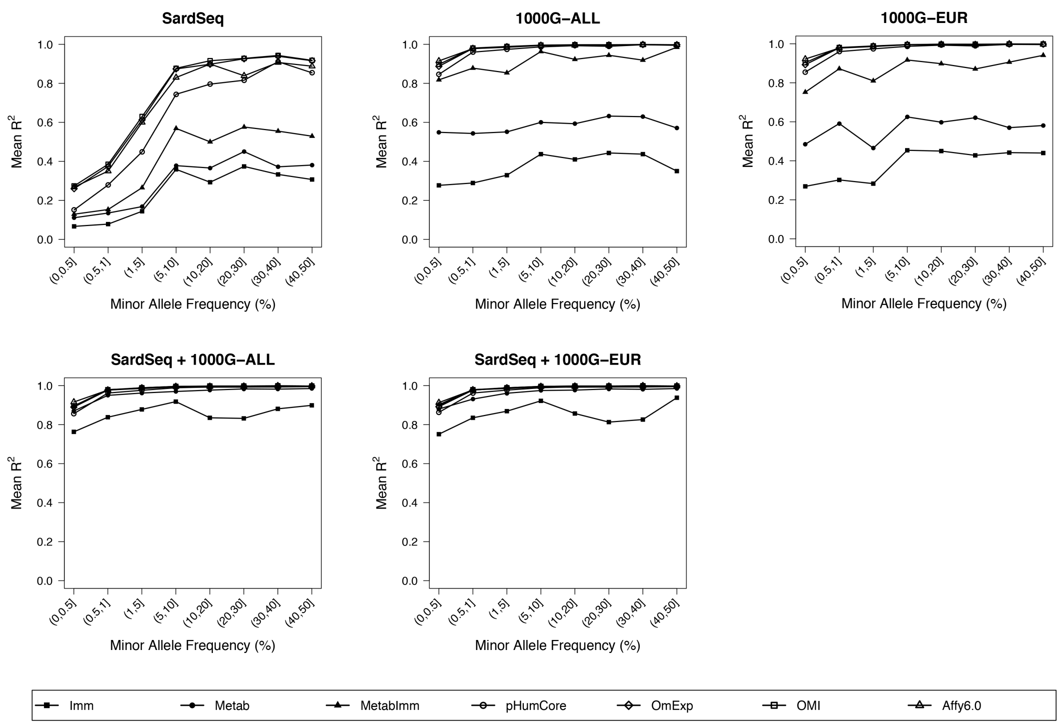
This subsection is restricted to the SardiNIA study and the simulated European haplotypes, because the MCTFR study used only one array. We found clear differences in imputation performance depending on the baseline genotyping set. Comparable differences were seen when assessments were done with either the Rsq metric - the imputation quality metric from MACH - or the R^2 metric, the squared Pearson correlation, between dosages and real genotypes[8] (Table 1). When using the 1000G reference panels for Sardinians, the two custom arrays (Cardio-MetaboChip and ImmunoChip) provided very limited information for imputation and far less accuracy than the genome-wide arrays, reflecting their low marker density. However, the Cardio-MetaboChip array performed very well when imputing with the SardSeq panel, allowing accurate inference of the rest of the genome (mean Rsq = 0.62, and mean R^2 = 0.70 at HumanExome SNPs). The relative efficiency was similar when considering all autosomes (Table 1 and Supplementary Table S5) or focusing only on chromosome 20 (Figure 3 and Supplementary Table S6). The extended LD in the population and the increased genetic similarity of the reference panel aid in haplotype reconstruction when using a relatively small set of markers. The addition of the two custom arrays to the OmExp genome-wide array (OmExp+Metab+Imm, called OMI here) did not improve quality for common or low-frequency variants compared with that reached using OmExp alone. Thus, such arrays provide direct genotyping of low-frequency and rare variants in genes of interest but do not contribute to an overall improvement in imputation accuracy. We also observed negligible differences in imputation accuracy between the two tested Illumina genome-wide arrays, OmExp and pHumCore (Table 1 and Supplementary Tables S5 and S6), when imputing the SardSeq panel. In particular, we noticed that the low-density genome-wide array pHumCore provided only slightly less accuracy than the denser OmExp array when the SardSeq sequencing panel was used for imputation (mean R^2 = 0.85 and 0.87, for pHumCore and OmExp, respectively, at HumanExome SNPs, Supplementary Table S5) and a very similar genomic coverage (92.6 and 91.8% of markers imputed with Rsq >0.3, Table 1). Of note, performance was patently lower for both arrays and more significantly for pHumCore when imputation was performed with the 1000G panels (mean R^2 = 0.54 and 0.64, for pHumCore and OmExp, respectively, imputing with the 1000G-ALL; Table 1, Supplementary Tables S5 and S6, and Figure 3). In contrast, in the simulated European data, the Cardio-MetaboChip performed poorly, with insufficient genomic coverage. Contrarily to previous observations[32], the pHumCore was fairly comparable in efficiency to the OmExp array (Figure 4 and Supplementary Table S7), but we expect performance to be overestimated (because the genotypes were simulated based on 1000 Genomes). In fact, when we extracted subset of SNPs that are present in HumanOmniExpress and HumanCore from the MCTFR genotypes, the difference between the two arrays was clearly evident (Supplementary Table S8). This difference has also been observed for another European population[33]. Thus, in founder populations it appears that highly accurate imputation can be achieved with cost-effective sparse genotyping arrays when a population-specific reference panel is available.

Figure 3. Mean R^2 for each particular genotyping array/reference panel in the SardiNIA cohort



The figure shows the mean R^2 at different allele frequency ranges for each particular genotyping array/reference panel combination, including the combination of SardSeq and 1000G panels. Results are restricted to chromosome 20.

Figure 4. Mean R^2 for each combination of genotyping array/reference panel in the European simulated dataset



The figure shows the mean R^2 at different allele frequency ranges for each particular genotyping array/reference panel, including the combination of SardSeq and 1000G panels. Results are restricted to chromosome 20.

Table 1. Basic imputation statistics on the SardiNIA samples for different panels/genotyping arrays

		Whole imputed SNP set		*Rsq > 0.3*		*Shared imputed SNPs*
Array	*Reference Panel*	*No of SNPs*	*Mean (SD) Rsq*	*% SNPs*	*Mean (SD) Rsq*	*Mean (SD) Rsq*
Imm	SardSeq	15,071,719	0.258 (0.312)	33.33	0.652 (0.213)	0.299 (0.321)
	1000G-ALL	37,798,002	0.037 (0.134)	3.90	0.638 (0.232)	0.099 (0.213)
	1000G-EUR	16,873,087	0.085 (0.203)	9.68	0.647 (0.231)	0.115 (0.232)
Metab	SardSeq	15,069,660	0.617 (0.335)	76.91	0.777 (0.181)	0.685 (0.301)
	1000G-ALL	37,782,741	0.064 (0.170)	7.20	0.614 (0.217)	0.175 (0.260)
	1000G-EUR	16,878,099	0.149 (0.253)	18.05	0.634 (0.219)	0.201 (0.282)
MetabImm	SardSeq	14,977,409	0.734 (0.300)	86.51	0.835 (0.163)	0.808 (0.239)
	1000G-ALL	37,721,853	0.100 (0.218)	11.71	0.644 (0.221)	0.272 (0.311)
	1000G-EUR	16,781,983	0.219 (0.303)	27.12	0.667 (0.222)	0.297 (0.328)
OmExp	SardSeq	14,580,754	0.861 (0.256)	92.61	0.924 (0.131)	0.935 (0.161)
	1000G-ALL	37,424,729	0.297 (0.382)	33.61	0.796 (0.224)	0.742 (0.322)
	1000G-EUR	16,453,325	0.543 (0.406)	60.89	0.84 (0.206)	0.729 (0.341)
OMI	SardSeq	14,319,695	0.862 (0.256)	92.57	0.925 (0.131)	0.937 (0.159)
	1000G-ALL	37,211,511	0.300 (0.385)	34.00	0.799 (0.131)	0.753 (0.318)
	1000G-EUR	16,255,689	0.549 (0.406)	61.50	0.842 (0.206)	0.739 (0.337)
pHumCore	SardSeq	15,020,615	0.840 (0.264)	91.81	0.908 (0.139)	0.913 (0.179)
	1000G-ALL	37,793,052	0.234 (0.341)	26.66	0.759 (0.221)	0.614 (0.354)
	1000G-EUR	16,825,817	0.455 (0.398)	52.64	0.802 (0.207)	0.615 (0.367)
Affy6.0	SardSeq	14,550,658	0.798 (0.342)	84.51	0.937 (0.116)	0.905 (0.232)
	1000G-ALL	37,328,716	0.263 (0.379)	29.55	0.814 (0.217)	0.721 (0.341)
	1000G-EUR	16,350,040	0.515 (0.416)	57.63	0.843 (0.205)	0.708 (0.357)

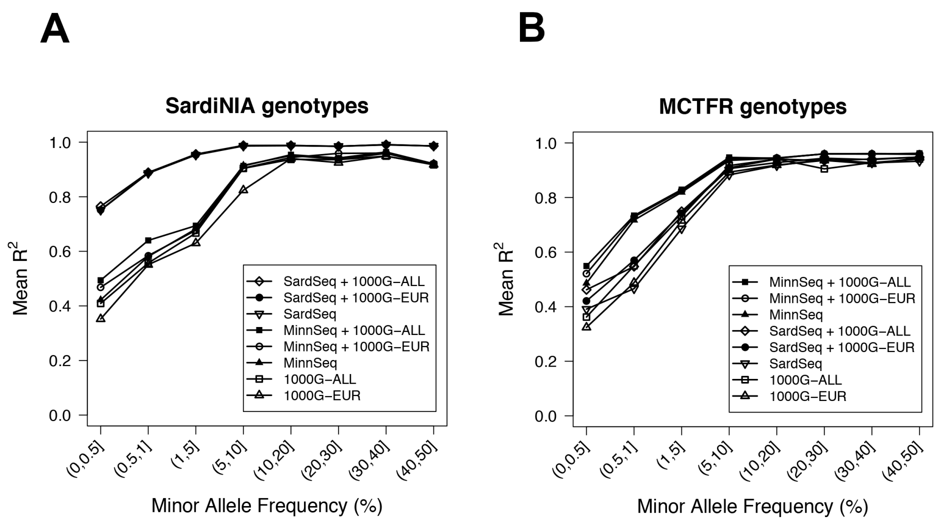
The table shows, for each genotyping array/reference panel combination, the number of imputed SNPs and the corresponding mean Rsq and SD, the percentage of SNPs with Rsq >0.3, with the corresponding mean Rsq and SD, mean Rsq and SD evaluated for 8 842 944 SNPs that were imputed in all genotyping array/reference panel combinations (called ‘Shared imputed SNPs’).

3.2. Effect of study-specific reference panels

Study-specific reference panels increased the accuracy and completeness of coverage in both Sardinian and Minnesota samples, but the gain in accuracy was greater for the Sardinia founder population. In Sardinians, the 1000G-ALL reference panel provided the highest number of imputed variants - ~37 million including both indels and SNPs vs ~15 million SNPs for the SardSeq panel - but the majority were of poor quality and were subsequently discarded. For example, for the Metab/SardSeq combination, 11.5 million imputed SNPs passed the standard Rsq >0.3 filter, but only 2.7 million and 3.0 million reached that threshold for Metab/1000G-ALL and Metab/1000G-EUR, respectively. The gap was less striking but still marked when denser genotype data sets were considered, and was still noticeable even considering only SNPs present in all reference panels (which are enriched for high-frequency variants; Table 1). Consistent results were seen for the OmExp, OMI, pHumCore and Affy6.0 data sets, with accuracy consistently better when using SardSeq (Figure 3). The benefit in overall accuracy was clear at all frequency ranges and even greater for low-frequency and rare variants. For example, using the OMI data set, the average R^2 for SNPs with MAF ranging from 0.5 to 1% is 0.91, 0.57 and 0.52 when using SardSeq, 1000G-ALL and 1000G-EUR reference panels, respectively (Supplementary Table S5). This reinforces the finding that on average, low-frequency variants are hard to impute in founder populations when using external reference panels because these variants appear in fewer haplotypes[6]. Of note, the results remained the same after removing 646 Sardinian samples that appear in both the genotyping set and the SardSeq reference panel (Supplementary Table S2). To assess whether the advantage with the SardSeq panel was attributable to the lower number of European haplotypes present in the 1000 Genomes reference, we performed imputation using the MinnSeq panel. There was no appreciable gain in accuracy within Sardinians compared with 1000G-based imputations (Figure 5a, Supplementary Tables S9A and S10). Similar to results with Sardinians, the MinnSeq panel outperformed the 1000G panels in the MCTFR study at all frequency ranges (Figure 5b and Supplementary Table S9B). However, the gain in accuracy was far less than that observed in Sardinians with the SardSeq panels. For example, for variants with MAF ranging from 1 to 5%, we observed

11% and 42% additional gain in mean R^2 for Minnesotans and Sardinians, respectively. Of note, in both cohorts the study-specific panel also yielded a higher number of SNPs useful for analyses (considering an Rsq >0.3) even when the other reference sets contain more SNPs (Supplementary Table S10).

Figure 5. Impact of cross-studies reference panels



The figure shows the mean R^2 at different allele frequency ranges for the chromosome 20 of OmExp genotyping array for SardiNIA (A) and the Illumina 660W-quad array for the MCTFR (B) study, when using different reference panels, including combination of SardSeq/MinnSeq and 1000G panels and cross-studies references.

3.3. Effect of combined reference panels

We also evaluated the impact on imputation accuracy of extended panels created by combining the two study-specific panels and 1000G haplotypes. The combined SardSeq+1000G panels provided only marginally higher accuracy at rarer shared SNPs in Sardinians (Figure 3 and Supplementary Tables S5 and S6). Slight increase in accuracy was also observable for more frequent variants (except for the two custom arrays (Metab and Imm), for which the improvement was substantial across all frequency ranges (Figure 3 and Supplementary Tables S5 and Table S6). Thus, for Sardinians, the inclusion of 1000G haplotypes would only be beneficial for very rare variants if a genome-wide array was used for baseline imputation. In the simulated European set, the addition of SardSeq haplotypes to the 1000G panels remarkably increased imputation accuracy for custom genotyping arrays (Metab and Imm) for both common and rare variants (Figure 4 and Supplementary Table S7). For example, for variants with MAF >40% and MAF ≤50% the mean R^2 is 0.57 and 0.98, when imputing with 1000G-ALL and SardSeq+1000G-ALL and using the Metab data set (Figure 4 and Supplementary Table S7). The impact of a combined panel was instead negligible for the more comprehensive genotype data (OmExp, OMI, pHumCore and Affy6.0). However, imputation on simulated data could give slight overestimations, and this could mask the advantage of adding SardSeq to 1000G panels. Indeed, when considering the MCTFR study, the combined SardSeq+1000G-ALL panel provided benefit at all frequency ranges compared with 1000G-ALL imputation, and for MAF ≤0.5% variants accuracy becomes fairly similar to that observed when using the MCTFR-specific panel (Figure 5 and Supplementary Table S9). Thus, the Sardinian panel could be generally useful to increase the overall accuracy in population cohorts other than Sardinians, especially where only custom array genotyping is available or when a study-specific reference is not available. Compared with imputation with MinnSeq alone, the addition of the 1000G haplotypes to the MinnSeq reference panel was useful only for rare variants in Minnesotans. The difference in accuracy was >4-fold higher than that seen in Sardinians comparing imputations with SardSeq and SardSeq+1000G panels. Thus, for Europeans, the inclusion of 1000G haplotypes in a study-specific panel is sensitively beneficial for very rare variants. Of note, for the Minnesotans, genotype imputation at the full spectrum of frequency ranges never reaches the same accuracy as in SardiNIA with the SardSeq panel, even when using the combined MinnSeq+1000G with almost twice as many individuals as there are in the SardSeq panel. Given the great utility of the Sardinian haplotypes, we further examined whether the advantage achieved by imputing with the SardSeq panel could have been reached sequencing a smaller number of samples and merging their haplotypes with the 1000G panels. For simplicity, we again focused on chromosome 20 and the OmExp array. Only for variants with MAF >5% does adding 500 Sardinian samples to the 1000G panels provide the same accuracy as the SardSeq panel alone. Instead, adding 1000 Sardinians to the 1000G panels provides the same accuracy given by the SardSeq panel for all frequency bins, with only a modest difference in accuracy for the very rare variants (MAF <0.5%) (Supplementary Figure S1 and Supplementary Table S11). Thus, sequencing a smaller number of individuals and combining their haplotypes with the 1000G panels could give imputation accuracy that is highly comparable to a panel comprising a large number of samples. However, the caveat remains that the genotype accuracy and variant discovery in low-pass sequencing is highly dependent on the number of sequenced samples. Consequently, sequencing only 500 samples would not provide genotypes as precise as those obtained by randomly selecting 500 samples from a set of 2000 sequenced genomes. For example, when we performed variant calling on a subset of 508 samples, the heterozygous error rate increased from 2.6 to 11.3% at rare sites (Supplementary Table S12).

3.4. Impact of different phasing strategies

Phasing accuracy is affected by many parameters, including sample size and marker density[34]. As pre-phasing is a key step in genotype imputation, any distortion in this part could potentially distort imputation accuracy. We assessed the magnitude of the expected decrease in accuracy when Cardio-MetaboChip and ImmunoChip, as well as their combination, are phased separately from HumanOmniExpress. As reported in

Supplementary Table S3, there is only a slight decrease in accuracy at rarer sites if imputation is performed using 1000 Genomes panels. Results are instead less accurate if imputation is performed with the SardSeq panel, especially for very rare variants (MAF <0.5%). We also noticed that the differences between the two sets of results are larger for ImmunoChip than Cardio-MetaboChip, confirming the lower informative content of this array even at the phasing level. Of note, even when Cardio-MetaboChip is phased separately, it provides accurate genotype imputation over the full genome of Sardinian samples when the SardSeq panel is used.

3.5. Performance of imputation quality metrics

To determine whether the commonly used MACH-Rsq threshold >0.3 and IMPUTE-INFO >0.4 can be applied to all frequency ranges (and if not, to infer appropriate cutoffs), we investigated how well imputation quality metrics can predict true imputation accuracy, especially for rare and less common variants. We found that for MAF ≥1%, imputation accuracy and therefore concordance between real genotypes and dosages using study-specific panels was almost perfect in both Sardinians and Minnesotans (Tables 2 and 3 and Supplementary Figure S2). At these frequency ranges, high but clearly less concordance was also seen when imputing with the 1000G panels. Whatever the reference panel used and the population under study, the standard Rsq cutoff of >0.3 efficiently discarded most badly imputed markers while keeping most of those imputed well (see Materials and Methods). In particular, imputation was so accurate overall that even an Rsq cutoff of >0 would leave no badly imputed markers on chromosome 20 (Tables 2a and 3a) (and only 8 over the entire genome in Sardinians, Supplementary Table S13). Similarly for the INFO metrics, the standard >0.4 threshold was efficient to discriminate between well and poorly inferred genotypes at this range of frequency (Tables 2b and 3b). In contrast, for MAF <1%, we noticed that both metrics were slightly overestimated when using the study-specific panels, possibly because of the inclusion of relatives with similar haplotypes in the target data set; but overall concordance was better than 1000G imputation for this range of frequency as well. Specifically, in this range and when imputation was performed with the 1000G panels, the threshold of Rsq >0.3 was less efficient, aggressively discarding some well-imputed variants (eliminating 7-18% and 7-25% of the well-imputed markers for ALL and EUR panels) and retaining an excess of the badly imputed ones (Tables 2a and 3a and Supplementary Tables S14 and S15). The INFO >0.4 threshold instead worked efficiently on selecting well-imputed variants, but was too lenient on discarding those of poor quality (Tables 2b and 3b and Supplementary Tables S14 and S15). Nevertheless, Rsq >0.3 and INFO >0.4 still remain the optimal thresholds. When imputation was performed with the study-specific panels, both the Rsq and INFO thresholds were more efficient in capturing all well-imputed markers, but less efficient in discarding the poorly imputed. In such cases, i.e. for MAF <1% and when imputation is performed with a reference panel that is genetically close to the study population, an Rsq threshold of >0.6 and INFO >0.7 should be preferred in lieu of the standard thresholds of 0.3 and 0.4, respectively.

Table 2. Efficiency of imputation quality metrics in the SardiNIA cohort

A	MAF < 1%				MAF ≥ 1%			
	SardSeq		1000G-ALL		SardSeq		1000G-ALL	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (14)	100 (222)	100 (98)	100 (124)	0 (0)	100 (301)	100 (20)	100 (255)
> 0.1	92.86 (13)	100 (222)	44.9 (44)	92.74 (115)	0 (0)	100 (301)	90 (18)	99.61 (254)
> 0.2	85.71 (12)	100 (222)	19.39 (19)	86.29 (107)	0 (0)	100 (301)	75 (15)	99.61 (254)
> 0.3	78.57 (11)	100 (222)	11.22 (11)	81.45 (101)	0 (0)	100 (301)	65 (13)	98.43 (251)
> 0.4	71.43 (10)	100 (222)	5.1 (5)	70.16 (87)	0 (0)	100 (301)	45 (9)	97.25 (248)
> 0.5	64.29 (9)	99.55 (221)	3.06 (3)	62.9 (78)	0 (0)	100 (301)	30 (6)	94.9 (242)
> 0.6	42.86 (6)	95.95 (213)	2.04 (2)	50.81 (63)	0 (0)	100 (301)	20 (4)	89.41 (228)
> 0.7	28.57 (4)	91.89 (204)	0 (0)	43.55 (54)	0 (0)	100 (301)	15 (3)	82.35 (210)
> 0.8	14.29 (2)	83.78 (186)	0 (0)	33.87 (42)	0 (0)	100 (301)	0 (0)	72.16 (184)
> 0.9	7.14 (1)	58.11 (129)	0 (0)	23.39 (29)	0 (0)	98.01 (295)	0 (0)	59.22 (151)
> 1	0 (0)	0.9 (2)	0 (0)	0 (0)	0 (0)	3.65 (11)	0 (0)	2.75 (7)

B	MAF < 1%				MAF ≥ 1%			
	SardSeq		1000G-ALL		SardSeq		1000G-ALL	
INFO	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (7)	100 (189)	100 (81)	100 (83)	0 (0)	100 (307)	100 (32)	100 (251)
> 0.1	100 (7)	100 (189)	100 (81)	98.8 (82)	0 (0)	100 (307)	100 (32)	100 (251)
> 0.2	100 (7)	99.47 (188)	90.12 (73)	97.59 (81)	0 (0)	100 (307)	100 (32)	100 (251)
> 0.3	100 (7)	99.47 (188)	62.96 (51)	96.39 (80)	0 (0)	100 (307)	96.88 (31)	100 (251)
> 0.4	100 (7)	99.47 (188)	48.15 (39)	93.98 (78)	0 (0)	100 (307)	96.88 (31)	100 (251)
> 0.5	100 (7)	99.47 (188)	27.16 (22)	89.16 (74)	0 (0)	100 (307)	84.38 (27)	99.2 (249)
> 0.6	100 (7)	98.94 (187)	17.28 (14)	85.54 (71)	0 (0)	100 (307)	59.38 (19)	98.01 (246)
> 0.7	71.43 (5)	97.35 (184)	11.11 (9)	73.49 (61)	0 (0)	100 (307)	37.5 (12)	95.62 (240)
> 0.8	42.86 (3)	92.59 (175)	3.7 (3)	60.24 (50)	0 (0)	100 (307)	15.62 (5)	88.84 (223)
> 0.9	14.29 (1)	76.19 (144)	0 (0)	44.58 (37)	0 (0)	99.35 (305)	6.25 (2)	72.91 (183)
> 1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

The table shows the number and the percentage of poorly imputed and well-imputed SNPs (see Materials and Methods) that are captured for each Rsq (A) and INFO (B) threshold. Imputation was performed on chromosome 20 HumanOmniExpress SNPs, using the SardSeq and 1000G-ALL panels. Statistics are reported separately for common and rare variants.

Table 3. Efficiency of imputation quality metrics in the MCTFR cohort

A	MAF < 1%				MAF ≥ 1%			
	MinnSeq		1000G-ALL		MinnSeq		1000G-ALL	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (38)	100 (129)	100 (80)	100 (92)	0 (0)	100 (284)	100 (4)	100 (258)
> 0.1	81.58 (31)	100 (129)	72.5 (58)	96.74 (89)	0 (0)	100 (284)	100 (4)	100 (258)
> 0.2	73.68 (28)	100 (129)	41.25 (33)	95.65 (88)	0 (0)	100 (284)	25 (1)	100 (258)
> 0.3	57.89 (22)	100 (129)	26.25 (21)	92.39 (85)	0 (0)	100 (284)	25 (1)	99.61 (257)
> 0.4	47.37 (18)	100 (129)	17.5 (14)	83.7 (77)	0 (0)	100 (284)	25 (1)	98.84 (255)
> 0.5	28.95 (11)	96.9 (125)	10 (8)	72.83 (67)	0 (0)	100 (284)	0 (0)	96.12 (248)
> 0.6	21.05 (8)	92.25 (119)	3.75 (3)	59.78 (55)	0 (0)	95.07 (270)	0 (0)	87.21 (225)
> 0.7	2.63 (1)	72.09 (93)	1.25 (1)	48.91 (45)	0 (0)	89.79 (255)	0 (0)	77.13 (199)
> 0.8	0 (0)	51.94 (67)	0 (0)	31.52 (29)	0 (0)	79.58 (226)	0 (0)	62.02 (160)
> 0.9	0 (0)	28.68 (37)	0 (0)	17.39 (16)	0 (0)	59.51 (169)	0 (0)	48.45 (125)
> 1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

B	MAF < 1%				MAF ≥ 1%			
	MinnSeq		1000G-ALL		MinnSeq		1000G-ALL	
INFO	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (38)	100 (96)	100 (82)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.1	100 (38)	100 (96)	100 (82)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.2	100 (38)	100 (96)	97.56 (80)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.3	97.37 (37)	100 (96)	95.12 (78)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.4	94.74 (36)	100 (96)	69.51 (57)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.5	73.68 (28)	100 (96)	41.46 (34)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.6	50 (19)	98.96 (95)	14.63 (12)	100 (67)	0 (0)	100 (277)	44.44 (4)	100 (241)
> 0.7	23.68 (9)	95.83 (92)	7.32 (6)	92.54 (62)	0 (0)	99.28 (275)	0 (0)	99.59 (240)
> 0.8	5.26 (2)	77.08 (74)	0 (0)	71.64 (48)	0 (0)	91.7 (254)	0 (0)	87.97 (212)
> 0.9	2.63 (1)	40.62 (39)	0 (0)	46.27 (31)	0 (0)	72.2 (200)	0 (0)	63.49 (153)
> 1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

The table shows the number and the percentage of poorly imputed and well-imputed SNPs (see Materials and Methods) that are captured for each Rsq (A) and INFO (B) threshold. Imputation was performed on chromosome 20 Illumina 660W-quad array SNPs, using MinnSeq and 1000G-ALL as reference panels. Statistics are reported separately for common and rare variants.

4. Concluding remarks

We used different reference panels and genotype input sets to investigate the effects on imputation in founder and non-founder populations of European ancestry. We found that a study-specific reference panel considerably improved imputation accuracy and genomic coverage compared with external equally large reference panels, regardless of the genotyping array, especially for rare variants. However, the benefit was strikingly higher in the founder population of Sardinians, with a precision that was not obtainable in Europeans even with a reference panel twice the size. In fact, in such homogenous populations each sequenced genome provides information that can be extended to distant relatives as well, whereas in continental Europeans, haplotypes carrying rare variants can only inform closely related samples. We also observed that in Sardinians a study-specific panel boosts imputation even for low-coverage genotyping array(s) like the Cardio-MetaboChip that are barely informative when imputing with the 1000G panels alone, or for the HumanCore that becomes highly comparable for all frequency ranges to the wider HumanOmniExpress. Given the low cost of the sparser arrays, accurate population-scale imputation is more feasible in the Sardinian founder population than in non-founder populations when combined with large-scale sequencing. This is true also for custom arrays, which can be specifically designed to include population-specific variants in order to increase the power of the association tests for these variants. For example, at current cost schedules, with an investment of 500 000 dollars one could genotype ~8300 Sardinian samples with the HumanCore array instead of ~4500 with the HumanOmniExpress. The power to detect association for variants accounting for 0.5% of the trait variance thereby rises from 24 to 84%. Finally, we observed that standard thresholds on metrics for evaluating accuracy, estimated by two commonly used imputation software, are somewhat imprecise for rare variants. We propose that all cohorts using study-specific reference panels for imputation consider adopting different thresholds for common and rare variants to filter inaccurate genotypes. Taken together, these imputation-based analyses can guide genetic studies, and complement recent reports[32,35] with several novel aspects that can improve performance:

- They exploit imputation accuracy with the two larger study-specific reference panels so far published, including one that is population specific.

- They also provide the first evaluation of imputation performance of the 1000 Genomes Project haplotypes in an isolated population.

- They include analyses of large cohorts coupled with the use of HumanExome array, allowing appropriate assessment of results for less frequent and rare variants.

- Using real data sets, they based analyses on a subset of quality-controlled SNPs instead of the full list of markers present on an array (excluding many that are likely to be imperfectly genotyped in a case study).

- They evaluate two widely used custom genotyping arrays, Cardio-MetaboChip and ImmunoChip, providing information for cohorts that are limited to that source of genotypes.

- They also evaluate for rare variants the efficiency of accuracy metric thresholds that were previously suggested for common variants.

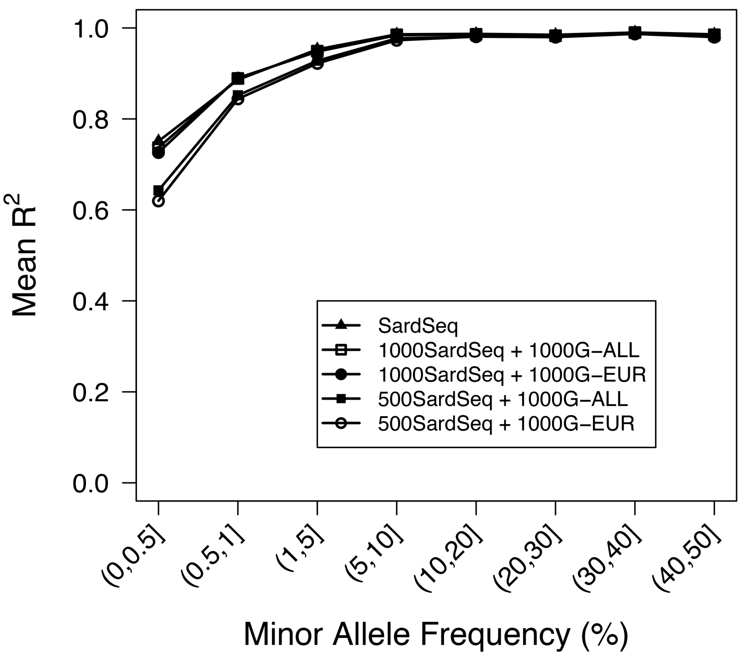
Ultimately, full genome sequencing could make imputation methods superfluous, but the timescale remains indeterminate. It should be considered that increasing sample size can augment genome-wide power to assess rare variants more than increasing array density - even up to full genotyping of the complete 1000 Genomes Project variant set[32,35]. Thus, aids to imputation are increasingly valuable, because most studies are likely to be collecting increasing numbers of samples and using this inferential process rather than sequencing full genomes.

Overall, population-specific panels might have been thought to be ‘private’, with potential discoveries limited to that population. Instead, the effectiveness of population-specific reference panels can be appreciable for other populations, but will vary depending on the size of the panels and the demographic history of the isolate. Intuitively in Europe, their value may be greater for populations like Basques and Greeks, who are relatively genetically distant from the European samples selected for the 1000 Genomes Project. Here, we show that sequencing efforts from the Sardinian founder population can, when coupled with available panels, improve rare variant imputation accuracy in other population backgrounds as well. This reinforces the value of isolated populations for discovery of variants that are locally enriched but rarer and thus harder to detect in international surveys[36].

5. Supplementary Informations

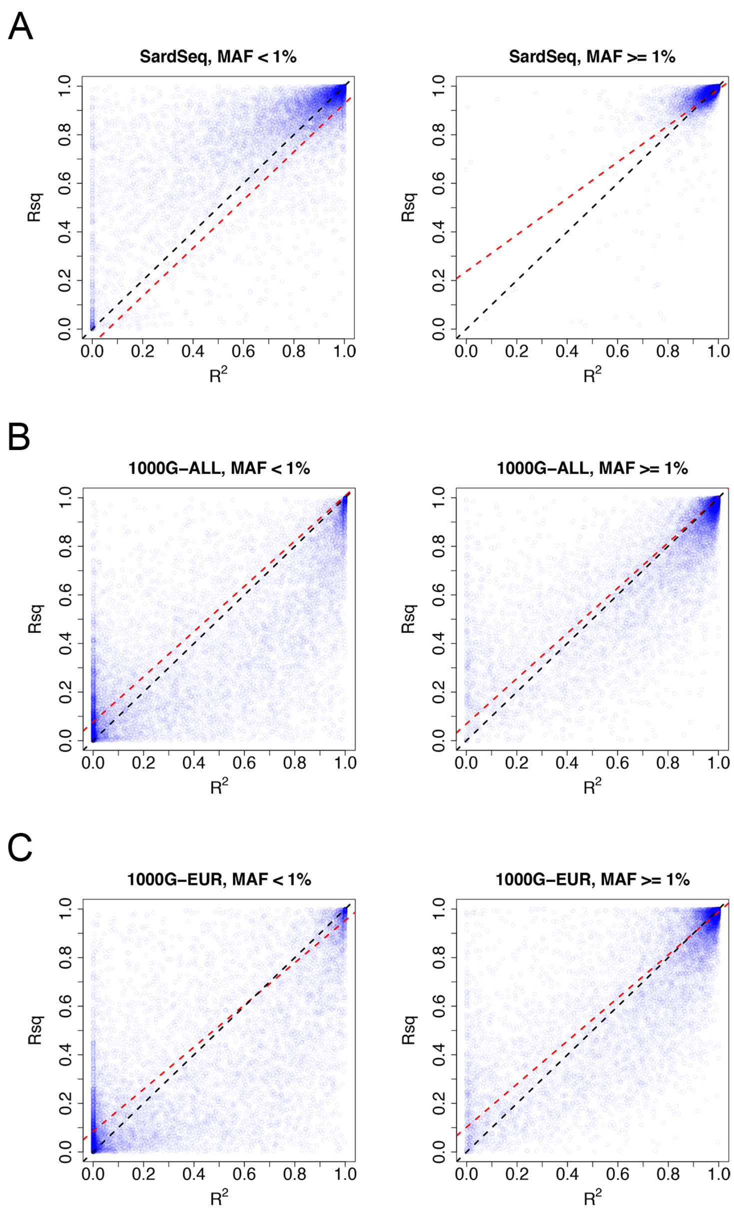
5.1. Supplementary Figures

Figure S1. Impact of combined 1000G and subsets of SardSeq reference panels on the SardiNIA cohort



The figure shows the mean R^2 at different allele frequencies ranges, for the chromosome 20 of OmExp genotyping array and different reference panels, including combination of SardSeq and 1000G panels.

Figure S2. Imputation accuracy: Rsq values versus R^2



The figure shows a scatterplot of estimated Rsq (y-axis) versus R^2 (x-axis) for autosomal SNPs on OmExp genotyping array in the SardiNIA samples. The black line represents the diagonal line, while the red line is the estimated fitted correlation line. A different plot is given for three different minor allele frequency ranges and for each imputation panel: SardSeq (A), 1000G-ALL (B) and 1000G-EUR (C).

5.2. Supplementary Tables

Table S1. Features of genotyped and derived arrays

Array	N SNPs	N Autosomal SNPs	N QCed Autosomal SNPs	% QCed Autosomal SNPs		
				MAF < 1%	1% ≤ MAF < 5%	MAF ≥ 5%
HumanOmniExpress	730,525	709,358	607,038	0.00	9.36	90.64
Cardio-MetaboChip	196,725	196,474	141,231	18.96	13.58	67.46
ImmunoChip	196,524	192,403	150,979	16.42	13.73	69.84
HumanExome	247,870	242,296	79,980	56.48	12.08	31.44
pseudo-HumanCore	298,930	288,675	227,745	0.27	5.36	94.37
Affymetrix 6.0	934,970	895,351	723,763	0.44	11.62	87.94

The table shows for each genotyped array, the total number of SNPs, the number of those in autosomes, the number of autosomal QCed SNPs and their frequencies. The row corresponding to pHumCore shows the total number of SNPs and those in autosomes for the non-genotyped array.

Table S2. Impact of overlapping individuals between target genotyping set and reference panel

MAF bins (%)	SardiNIA genotypes & SardSeq		MCTFR genotypes & MinnSeq	
	Full set N=6602	**Restricted set N=5956**	**Full set N=6610**	**Restricted set N=5429**
(0,0.5]	0.751	0.734	0.557	0.490
(0.5,1]	0.886	0.890	0.759	0.711
(1,5]	0.953	0.948	0.856	0.825
(5,10]	0.986	0.986	0.952	0.942
(10,20]	0.987	0.987	0.965	0.958
(20,30]	0.985	0.984	0.976	0.972
(30,40]	0.990	0.990	0.979	0.976
(40,50]	0.986	0.985	0.960	0.953

The table compares the mean R^2, at different frequency ranges, in the SardiNIA and MCTFR cohorts when imputation is performed with the full study set or after removing individuals that are also included in the respective reference panels. The analysis is restricted to chromosome 20, and uses as baseline genotypes the OmExp for the SardiNIA study and the Illumina 660W-quad for the MCTFR cohort.

Table S3. Impact of array phasing procedure

		Mean R^2 - separated phasing					Difference with Table S6				
	MAF bins (%)	Sard Seq	1000G-ALL	1000G-EUR	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR	Sard Seq	1000G-ALL	1000G-EUR	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR
Imm	(0,0.5]	0.347	0.100	0.092	0.456	0.437	0.060	0.000	0.004	0.053	0.055
	(0.5,1]	0.510	0.071	0.110	0.637	0.665	0.044	0.007	0.002	0.038	0.038
	(1,5]	0.588	0.181	0.120	0.696	0.685	0.035	0.001	0.002	0.034	0.026
	(5,10]	0.691	0.322	0.387	0.789	0.789	0.030	0.015	0.003	0.014	0.032
	(10,20]	0.708	0.325	0.377	0.783	0.788	0.034	-0.007	0.001	0.031	0.024
	(20,30]	0.596	0.296	0.222	0.657	0.658	0.019	0.002	-0.004	0.008	-0.018
	(30,40]	0.623	0.316	0.372	0.762	0.728	0.014	0.001	0.001	0.012	0.031
	(40,50]	0.714	0.296	0.326	0.777	0.802	0.021	0.001	0.004	0.013	0.020

		Mean R^2 - separated phasing					Difference with Table S6				
	MAF bins (%)	Sard Seq	1000G-ALL	1000G-EUR	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR	Sard Seq	1000G-ALL	1000G-EUR	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR
Metab	(0,0.5]	0.489	0.117	0.112	0.571	0.555	0.053	0.013	0.015	0.057	0.060
	(0.5,1]	0.753	0.125	0.126	0.712	0.768	0.021	0.005	0.004	0.014	0.034
	(1,5]	0.768	0.254	0.207	0.836	0.823	0.023	0.003	0.002	0.031	0.026
	(5,10]	0.803	0.315	0.320	0.853	0.856	0.020	-0.004	-0.012	0.029	0.025
	(10,20]	0.880	0.417	0.501	0.908	0.905	0.014	0.001	0.015	0.016	0.016
	(20,30]	0.809	0.360	0.348	0.855	0.848	0.019	0.003	0.002	0.019	0.017
	(30,40]	0.816	0.396	0.388	0.866	0.865	0.018	-0.002	0.001	0.022	0.026
	(40,50]	0.891	0.306	0.377	0.907	0.907	0.016	0.006	0.001	0.019	0.010

		Mean R^2 - separated phasing					Difference with Table S6				
	MAF bins (%)	Sard Seq	1000G-ALL	1000G-EUR	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR	Sard Seq	1000G-ALL	1000G-EUR	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR
MetabImm	(0,0.5]	0.586	0.191	0.165	0.622	0.621	0.025	0.008	0.009	0.016	0.028
	(0.5,1]	0.770	0.188	0.184	0.807	0.810	0.034	0.005	0.003	0.028	0.021
	(1,5]	0.875	0.350	0.299	0.894	0.895	0.017	0.005	0.002	0.016	0.018
	(5,10]	0.907	0.492	0.594	0.929	0.930	0.021	0.000	-0.011	0.014	0.013
	(10,20]	0.932	0.628	0.604	0.943	0.944	0.011	-0.003	0.013	0.010	0.011
	(20,30]	0.904	0.445	0.536	0.919	0.919	0.014	0.004	0.005	0.013	0.013
	(30,40]	0.912	0.629	0.521	0.926	0.926	0.014	0.002	0.001	0.012	0.012
	(40,50]	0.931	0.455	0.537	0.943	0.943	0.010	0.001	0.002	0.011	0.011

The table shows the mean R^2 when phasing prior imputation is performed separately for Imm, Metab and MetabImm genotyping sets (only chromosome 20 and only for the SardiNIA study). The right side of the table reports differences with the mean R^2 obtained when phasing is performed for all arrays jointly, as reported in **Table S6**.

Table S4. Mean R^2 values for different imputation approaches

MAF bins (%)	1000G-ALL	1000G-EUR	SardSeq	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR
(0,0.5]	0.386	0.356	0.766	0.762	0.751
(0.5,1]	0.522	0.486	0.885	0.886	0.882
(1,5]	0.600	0.602	0.944	0.948	0.947
(5,10]	0.842	0.772	0.983	0.983	0.983
(10,20]	0.920	0.923	0.984	0.985	0.984
(20,30]	0.906	0.900	0.982	0.982	0.982
(30,40]	0.955	0.948	0.987	0.987	0.987
(40,50]	0.900	0.921	0.984	0.984	0.984

The table shows the mean true R^2, for different frequency ranges, when IMPUTE2 is used to perform imputation, as well as panels combinations, in the SardiNIA study. Analysis was restricted to chromosome 20 and used the OmExp as baseline genotypes. Statistics for imputation performed with Minimac for single reference panels and for panels combined with VCF tools are reported in **Table S6**.

Table S5. Mean R^2 in the SardiNIA sample, for different allele frequency ranges, for all autosomes

A

	Imm			**Metab**			**MetabImm**			**OmExp**		
MAF bins (%)	**Sard Seq**	**1000G-ALL**	**1000G-EUR**	**Sard Seq**	**1000G-ALL**	**1000G-EUR**	**Sard Seq**	**1000G-ALL**	**1000G-EUR**	**Sard Seq**	**1000G-ALL**	**1000G-EUR**
(0,0.5]	0.322	0.074	0.073	0.492	0.102	0.097	0.571	0.160	0.150	0.691	0.379	0.337
(0.5,1]	0.487	0.122	0.106	0.732	0.175	0.145	0.823	0.264	0.226	0.908	0.552	0.501
(1,5]	0.555	0.182	0.154	0.788	0.247	0.216	0.884	0.351	0.314	0.951	0.677	0.633
(5,10]	0.592	0.242	0.267	0.824	0.328	0.365	0.910	0.472	0.512	0.974	0.863	0.845
(10,20]	0.590	0.272	0.318	0.839	0.402	0.468	0.922	0.536	0.586	0.983	0.934	0.926
(20,30]	0.606	0.304	0.340	0.852	0.466	0.497	0.930	0.597	0.623	0.989	0.944	0.944
(30,40]	0.577	0.359	0.328	0.849	0.499	0.479	0.927	0.629	0.609	0.988	0.948	0.949
(40,50]	0.591	0.264	0.310	0.863	0.424	0.481	0.933	0.558	0.607	0.990	0.953	0.950
all variants	0.486	0.171	0.173	0.703	0.246	0.246	0.787	0.344	0.339	0.871	0.643	0.614

B

	OMI			**pHumCore**			**Affy6.0**		
MAF bins (%)	**Sard Seq**	**1000G-ALL**	**1000G-EUR**	**SardSeq**	**1000G-ALL**	**1000G-EUR**	**SardSeq**	**1000G-ALL**	**1000G-EUR**
(0,0.5]	0.696	0.392	0.350	0.652	0.269	0.244	0.733	0.381	0.347
(0.5,1]	0.910	0.568	0.515	0.885	0.415	0.371	0.896	0.536	0.494
(1,5]	0.952	0.688	0.645	0.936	0.540	0.498	0.937	0.635	0.595
(5,10]	0.974	0.868	0.852	0.961	0.738	0.740	0.964	0.817	0.807
(10,20]	0.984	0.937	0.930	0.975	0.853	0.855	0.975	0.893	0.895
(20,30]	0.989	0.948	0.946	0.983	0.894	0.902	0.982	0.921	0.921
(30,40]	0.988	0.953	0.951	0.982	0.908	0.906	0.983	0.925	0.921
(40,50]	0.991	0.954	0.952	0.985	0.907	0.910	0.984	0.930	0.933
all variants	0.873	0.653	0.624	0.849	0.536	0.517	0.834	0.598	0.576

Table S6. Mean R^2 in the SardiNIA sample, for different allele frequency ranges, focused on chromosome 20, as represented in Figure 3

A

	Imm					Metab					MetabImm				
MAF bins (%)	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**
(0,0.5]	0.407	0.100	0.096	0.456	0.437	0.542	0.130	0.127	0.571	0.555	0.611	0.191	0.174	0.622	0.621
(0.5,1]	0.554	0.071	0.112	0.637	0.665	0.774	0.130	0.130	0.712	0.768	0.804	0.188	0.187	0.807	0.810
(1,5]	0.623	0.181	0.122	0.696	0.685	0.791	0.257	0.209	0.836	0.823	0.892	0.350	0.301	0.894	0.895
(5,10]	0.721	0.322	0.390	0.789	0.789	0.823	0.311	0.308	0.853	0.856	0.928	0.492	0.583	0.929	0.930
(10,20]	0.742	0.325	0.377	0.783	0.788	0.894	0.418	0.516	0.908	0.905	0.943	0.628	0.617	0.943	0.944
(20,30]	0.615	0.296	0.218	0.657	0.658	0.828	0.363	0.350	0.855	0.848	0.918	0.445	0.541	0.919	0.919
(30,40]	0.637	0.316	0.373	0.762	0.728	0.834	0.394	0.389	0.866	0.865	0.926	0.629	0.522	0.926	0.926
(40,50]	0.735	0.296	0.330	0.777	0.802	0.907	0.312	0.378	0.907	0.907	0.941	0.455	0.539	0.943	0.943

B

	OmExp					OMI					pHumCore				
MAF bins (%)	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**
(0,0.5]	0.751	0.409	0.352	0.765	0.752	0.740	0.424	0.367	0.759	0.747	0.677	0.260	0.255	0.692	0.686
(0.5,1]	0.886	0.557	0.551	0.889	0.888	0.885	0.571	0.569	0.889	0.893	0.854	0.399	0.362	0.863	0.863
(1,5]	0.953	0.667	0.630	0.956	0.954	0.954	0.688	0.648	0.957	0.953	0.937	0.509	0.472	0.938	0.938
(5,10]	0.986	0.904	0.824	0.988	0.987	0.988	0.912	0.845	0.990	0.989	0.971	0.723	0.685	0.971	0.971
(10,20]	0.987	0.937	0.940	0.988	0.988	0.988	0.946	0.936	0.989	0.989	0.979	0.901	0.900	0.980	0.980
(20,30]	0.985	0.936	0.925	0.985	0.985	0.985	0.941	0.939	0.986	0.985	0.974	0.772	0.842	0.974	0.974
(30,40]	0.990	0.949	0.949	0.991	0.991	0.990	0.955	0.964	0.991	0.991	0.982	0.914	0.868	0.983	0.982
(40,50]	0.986	0.917	0.918	0.986	0.986	0.986	0.919	0.919	0.987	0.986	0.981	0.828	0.861	0.981	0.981

Table S6 (Continued)

C

	Affy6.0				
MAF bins (%)	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**
(0,0.5]	0.740	0.378	0.320	0.739	0.740
(0.5,1]	0.913	0.621	0.547	0.927	0.922
(1,5]	0.941	0.642	0.616	0.943	0.940
(5,10]	0.972	0.879	0.805	0.974	0.975
(10,20]	0.979	0.887	0.899	0.981	0.980
(20,30]	0.971	0.868	0.829	0.974	0.971
(30,40]	0.988	0.943	0.947	0.988	0.988
(40,50]	0.980	0.884	0.888	0.980	0.980

Table S7. Mean R2 values in the simulated data set of Europeans, for different allele frequency ranges, as represented in Figure 4

A

	Imm					**Metab**					**MetabImm**				
MAF bins (%)	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**
(0,0.5]	0.067	0.277	0.269	0.763	0.751	0.111	0.549	0.485	0.870	0.880	0.129	0.818	0.752	0.895	0.902
(0.5,1]	0.078	0.289	0.302	0.838	0.835	0.135	0.543	0.591	0.950	0.931	0.153	0.878	0.872	0.980	0.980
(1,5]	0.144	0.329	0.283	0.878	0.869	0.168	0.551	0.465	0.962	0.961	0.265	0.854	0.810	0.985	0.985
(5,10]	0.359	0.437	0.454	0.918	0.922	0.378	0.600	0.625	0.970	0.975	0.569	0.963	0.917	0.993	0.993
(10,20]	0.293	0.410	0.450	0.835	0.857	0.366	0.593	0.598	0.977	0.977	0.500	0.923	0.898	0.992	0.992
(20,30]	0.374	0.443	0.428	0.832	0.813	0.450	0.632	0.621	0.983	0.983	0.576	0.944	0.871	0.993	0.993
(30,40]	0.333	0.437	0.442	0.881	0.826	0.372	0.629	0.570	0.982	0.981	0.555	0.919	0.906	0.993	0.993
(40,50]	0.307	0.350	0.440	0.899	0.938	0.381	0.571	0.581	0.985	0.985	0.529	0.985	0.941	0.995	0.995

B

	OmExp					**OMI**					**pHumCore**				
MAF bins (%)	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**
(0,0.5]	0.260	0.887	0.893	0.890	0.892	0.274	0.897	0.903	0.896	0.896	0.151	0.846	0.855	0.856	0.863
(0.5,1]	0.374	0.979	0.979	0.977	0.977	0.385	0.981	0.981	0.979	0.979	0.279	0.960	0.960	0.962	0.962
(1,5]	0.613	0.988	0.988	0.989	0.989	0.630	0.989	0.989	0.989	0.989	0.449	0.974	0.974	0.976	0.977
(5,10]	0.874	0.995	0.995	0.996	0.996	0.877	0.996	0.996	0.996	0.996	0.744	0.987	0.987	0.989	0.989
(10,20]	0.898	0.997	0.998	0.997	0.998	0.916	0.998	0.998	0.998	0.998	0.796	0.993	0.993	0.993	0.994
(20,30]	0.926	0.997	0.997	0.998	0.998	0.928	0.998	0.998	0.998	0.998	0.816	0.990	0.989	0.992	0.992
(30,40]	0.939	0.999	0.999	0.999	0.999	0.943	0.999	0.999	0.999	0.999	0.911	0.998	0.997	0.998	0.998
(40,50]	0.916	0.998	0.998	0.998	0.998	0.918	0.998	0.999	0.998	0.998	0.854	0.995	0.996	0.996	0.996

Table S7 (Continued)

C

	Affy6.0				
MAF bins (%)	**Sard Seq**	**1000G -ALL**	**1000G -EUR**	**SardSeq + 1000G-ALL**	**SardSeq + 1000G-EUR**
(0,0.5]	0.268	0.915	0.922	0.916	0.913
(0.5,1]	0.350	0.978	0.978	0.977	0.978
(1,5]	0.600	0.985	0.986	0.986	0.986
(5,10]	0.830	0.994	0.995	0.995	0.995
(10,20]	0.899	0.995	0.995	0.995	0.995
(20,30]	0.839	0.994	0.993	0.995	0.995
(30,40]	0.907	0.998	0.999	0.999	0.999
(40,50]	0.888	0.997	0.997	0.998	0.998

Table S8. Relative efficiency of HumanOmniExpress and HumanCore arrays in the MCTFR study

MAF bins (%)	**HumanOmniExpress**	**HumanCore**
(0,0.5]	0.478	0.420
(0.5,1]	0.727	0.613
(1,5]	0.809	0.677
(5,10]	0.928	0.782
(10,20]	0.953	0.837
(20,30]	0.945	0.840
(30,40]	0.962	0.901
(40,50]	0.945	0.875

The table compares imputation accuracy in the MCTFR cohort when imputation is performed with MinnSeq panel and baseline genotypes are subset of SNPs overlapping between the Illumina 660W-quad, directly typed, and the HumanOmniExpress and HumanCore arrays (7944 and 3236 SNPs, respectively, representing 50% and 46% of the original content).

Table S9. Mean R^2 values in the SardiNIA and MCTFR studies, for different allele frequency ranges, as represented in Figure 5

A

SardiNIA genotypes

MAF bins (%)	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR	Sard Seq	MinnSeq + 1000G-ALL	MinnSeq + 1000G-EUR	Minn Seq	1000G-ALL	1000G-EUR
(0,0.5]	0.765	0.752	0.751	0.494	0.468	0.422	0.409	0.352
(0.5,1]	0.889	0.888	0.886	0.64	0.584	0.581	0.557	0.551
(1,5]	0.956	0.954	0.953	0.694	0.677	0.682	0.667	0.63
(5,10]	0.988	0.987	0.986	0.913	0.905	0.915	0.904	0.824
(10,20]	0.988	0.988	0.987	0.954	0.944	0.951	0.937	0.94
(20,30]	0.985	0.985	0.985	0.942	0.959	0.937	0.936	0.925
(30,40]	0.991	0.991	0.99	0.964	0.96	0.959	0.949	0.949
(40,50]	0.986	0.986	0.986	0.921	0.921	0.917	0.917	0.918

B

MCTFR genotypes

MAF bins (%)	MinnSeq + 1000G-ALL	MinnSeq + 1000G-EUR	Minn Seq	SardSeq + 1000G-ALL	SardSeq + 1000G-EUR	Sard Seq	1000G-ALL	1000G-EUR
(0,0.5]	0.549	0.521	0.486	0.462	0.421	0.391	0.362	0.324
(0.5,1]	0.734	0.729	0.717	0.547	0.57	0.466	0.55	0.488
(1,5]	0.829	0.825	0.819	0.749	0.742	0.686	0.731	0.715
(5,10]	0.947	0.942	0.936	0.909	0.906	0.883	0.917	0.893
(10,20]	0.944	0.945	0.944	0.941	0.928	0.917	0.941	0.919
(20,30]	0.96	0.96	0.96	0.935	0.944	0.94	0.905	0.936
(30,40]	0.959	0.961	0.96	0.941	0.939	0.927	0.928	0.923
(40,50]	0.962	0.96	0.958	0.948	0.948	0.933	0.941	0.946

The table shows the mean R^2 at different allele frequencies ranges, as shown in **Figure 5**, for SardiNIA (A) and MCTFR (B) studies, using HumanOmniExpress and Illumina 660W-quad arrays as baseline genotypes, respectively.

Table S10. Basic statistics for imputations in the SardiNIA and MCTFR cohorts using different reference panels

	Whole imputed SNPs set		Rsq > 0.3		Shared imputed SNPs
Reference Panel	**N SNPs**	**Mean (SD) Rsq**	**% SNPs**	**Mean (SD) Rsq**	**Mean (SD) Rsq**
SardiNIA					
SardSeq	315,846	0.877 (0.235)	0.94	0.927 (0.125)	0.918 (0.138)
1000G-ALL	835,114	0.289 (0.377)	0.33	0.791 (0.225)	0.593 (0.351)
1000G-EUR	361,269	0.543 (0.405)	0.61	0.838 (0.206)	0.568 (0.367)
SardSeq + 1000G-ALL	232,683	0.933 (0.163)	0.98	0.954 (0.096)	0.924 (0.128)
SardSeq + 1000G-EUR	211,183	0.948 (0.131)	0.99	0.959 (0.088)	0.920 (0.136)
MinnSeq	586,293	0.385 (0.401)	0.43	0.811 (0.225)	0.677 (0.321)
MinnSeq + 1000G-ALL	286,098	0.667 (0.369)	0.75	0.852 (0.200)	0.703 (0.300)
MinnSeq + 1000G-EUR	251,003	0.715 (0.348)	0.80	0.866 (0.192)	0.696 (0.307)
MCTFR					
SardSeq	318,468	0.636 (0.356)	0.74	0.814 (0.217)	0.669 (0.292)
1000G-ALL	838,326	0.330 (0.375)	0.39	0.752 (0.232)	0.620 (0.306)
1000G-EUR	364,132	0.607 (0.365)	0.71	0.805 (0.219)	0.592 (0.335)
SardSeq + 1000G-ALL	235,268	0.791 (0.269)	0.91	0.853 (0.189)	0.704 (0.268)
SardSeq + 1000G-EUR	213,757	0.829 (0.236)	0.94	0.867 (0.181)	0.700 (0.275)
MinnSeq	588,975	0.574 (0.338)	0.73	0.738 (0.228)	0.780 (0.224)
MinnSeq + 1000G-ALL	288,661	0.790 (0.263)	0.92	0.844 (0.188)	0.797 (0.199)
MinnSeq + 1000G-EUR	253,564	0.831 (0.229)	0.96	0.863 (0.178)	0.794 (0.201)

The table shows, for SardiNIA and MCTFR cohorts and for each reference panel, the number of SNPs included in the reference set, the corresponding mean Rsq and standard deviation, the percentage of SNPs with Rsq >0.3, with the corresponding mean Rsq and standard deviation, and the mean Rsq and standard deviation evaluated for SNPs that were imputed with all reference panels.

Table S11. Mean R^2 values for combined 1000G and subsets of SardSeq reference panels, for different allele frequency ranges, as represented in Figure S1

MAF bins (%)	SardSeq	500SardSeq + 1000G-ALL	1000SardSeq + 1000G-ALL	500SardSeq + 1000G-EUR	1000SardSeq + 1000G-EUR
(0,0.5]	0.751	0.643	0.737	0.620	0.726
(0.5,1]	0.886	0.852	0.889	0.844	0.890
(1,5]	0.953	0.928	0.949	0.922	0.948
(5,10]	0.986	0.977	0.985	0.973	0.984
(10,20]	0.987	0.981	0.986	0.981	0.986
(20,30]	0.985	0.980	0.983	0.980	0.983
(30,40]	0.990	0.987	0.990	0.987	0.989
(40,50]	0.986	0.980	0.985	0.980	0.985

The table shows the mean R^2 at different allele frequencies ranges, for each particular genotyping array/reference panel combination, as shown in **Figure S1**. Imputation was performed on the SardiNIA samples and restricted to chromosome 20 and OmExp genotypes.

Table S12. Genotype accuracy for low-pass sequencing

Variants grouping	N samples analyzed	Number of Variants in common between sequencing and Cardio-MetaboChip	Low pass results: Overall Discordance Rate (%)	Low pass results: Heterozygote Discordance Rate (%)
All variants genotyped using arrays				
All variants	508	2490	0.47	1
	1146	2542	0.29	0.73
	2120	2573	0.20	0.52
All variants genotyped using arrays, stratified by frequency among low pass samples				
MAF < 0.5%	508	77	0.12	11.3
	1146	124	0.04	10.42
	2120	142	0.02	2.59
MAF 0.5 - 5.0%	508	484	0.27	2.05
	1146	498	0.10	1.65
	2120	502	0.9	1.28
MAF > 5%	508	1929	0.52	0.96
	1146	1920	0.36	0.69
	2120	1929	0.24	0.49

The table compares accuracy of genotypes detected through sequencing when performing variants calling in the full set (2120 individuals) or in subsets of samples (508 and 1146 individuals). Accuracy was evaluated as the percentage of discordant genotypes with those available from Cardio-MetaboChip array. Discordance rate is reported for all genotypes as well as for heterozygous sites.

Table S13. Efficiency of Rsq quality metric on SardiNIA genotypes

A MAF < 1%

	SardSeq		1000G-ALL		1000G-EUR	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good left (n)
> 0	99.88 (825)	100.00 (7705)	99.95 (4016)	100.00 (4106)	99.91 (4248)	100.00 (3457)
> 0.1	73.24 (605)	99.83 (7692)	46.52 (1869)	96.10 (3946)	38.01 (1616)	94.13 (3254)
> 0.2	63.20 (522)	99.61 (7675)	24.69 (992)	90.60 (3720)	19.90 (846)	87.88 (3038)
> 0.3	56.30 (465)	99.38 (7657)	13.69 (550)	84.27 (3460)	11.45 (487)	81.52 (2818)
> 0.4	47.94 (396)	98.97 (7626)	7.84 (315)	75.89 (3116)	7.38 (314)	73.30 (2534)
> 0.5	40.19 (332)	98.03 (7553)	4.26 (171)	66.63 (2736)	4.61 (196)	65.63 (2269)
> 0.6	29.78 (246)	96.20 (7412)	2.71 (109)	57.84 (2375)	3.08 (131)	58.11 (2009)
> 0.7	20.58 (170)	91.69 (7065)	1.37 (55)	48.17 (1978)	2.12 (90)	49.67 (1717)
> 0.8	11.50 (95)	82.91 (6388)	0.70 (28)	37.26 (1530)	1.32 (56)	40.21 (1390)
> 0.9	3.87 (32)	60.95 (4696)	0.27 (11)	23.94 (983)	0.56 (24)	27.48 (950)
> 1	0.00 (0)	0.92 (71)	0.00 (0)	0.17 (7)	0.00 (0)	0.20 (7)

B MAF ≥ 1%

	SardSeq		1000G-ALL		1000G-EUR	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good left (n)
> 0	88.89 (8)	100.00 (11898)	99.88 (814)	100.00 (10081)	99.82 (1087)	100.00 (10094)
> 0.1	77.78 (7)	99.96 (11893)	85.89 (700)	99.83 (10064)	74.75 (814)	99.82 (10076)
> 0.2	44.44 (4)	99.93 (11890)	64.91 (529)	99.50 (10031)	49.86 (543)	99.32 (10025)
> 0.3	44.44 (4)	99.91 (11887)	45.77 (373)	98.75 (9955)	33.24 (362)	98.01 (9893)
> 0.4	44.44 (4)	99.86 (11881)	28.83 (235)	97.37 (9816)	21.03 (229)	95.88 (9678)
> 0.5	44.44 (4)	99.79 (11873)	17.18 (140)	94.75 (9552)	14.14 (154)	92.95 (9382)
> 0.6	44.44 (4)	99.72 (11865)	11.04 (90)	91.11 (9185)	8.91 (97)	89.04 (8988)
> 0.7	22.22 (2)	99.57 (11847)	6.01 (49)	85.83 (8653)	5.14 (56)	83.98 (8477)
> 0.8	22.22 (2)	99.18 (11800)	3.07 (25)	78.60 (7924)	3.49 (38)	77.87 (7860)
> 0.9	22.22 (2)	96.13 (11438)	1.10 (9)	66.96 (6750)	2.11 (23)	67.22 (6785)
> 1	0.00 (0)	3.80 (452)	0.00 (0)	2.28 (230)	0.00 (0)	2.28 (230)

The table shows the percentage and the number of poorly and well-imputed SNPs (see Materials and Methods) when using OmExp array as baseline for imputation, for each Rsq threshold. Panel A focus on variants with MAF <1%, and panel B on those with MAF ≥1%. Reported statistics are genome-wide.

Table S14. Efficiency of imputation quality metrics in the SardiNIA cohort, for 1000G-EUR based imputations

A	MAF < 1%		MAF ≥ 1%	
	1000G-EUR		1000G-EUR	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (109)	100 (105)	100 (30)	100 (255)
> 0.1	36.7 (40)	95.24 (100)	80 (24)	100 (255)
> 0.2	14.68 (16)	81.9 (86)	53.33 (16)	100 (255)
> 0.3	9.17 (10)	74.29 (78)	43.33 (13)	98.82 (252)
> 0.4	5.5 (6)	67.62 (71)	30 (9)	97.25 (248)
> 0.5	2.75 (3)	57.14 (60)	10 (3)	90.98 (232)
> 0.6	0.92 (1)	50.48 (53)	10 (3)	84.71 (216)
> 0.7	0.92 (1)	45.71 (48)	6.67 (2)	78.82 (201)
> 0.8	0 (0)	40 (42)	3.33 (1)	71.37 (182)
> 0.9	0 (0)	25.71 (27)	3.33 (1)	58.43 (149)
> 1	0 (0)	0 (0)	0 (0)	1.57 (4)

B	MAF < 1%		MAF ≥ 1%	
	1000G-EUR		1000G-EUR	
INFO	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (86)	100 (80)	100 (37)	100 (247)
> 0.1	89.53 (77)	100 (80)	100 (37)	100 (247)
> 0.2	62.79 (54)	98.75 (79)	100 (37)	100 (247)
> 0.3	46.51 (40)	97.5 (78)	86.49 (32)	100 (247)
> 0.4	31.4 (27)	92.5 (74)	70.27 (26)	99.6 (246)
> 0.5	18.6 (16)	82.5 (66)	51.35 (19)	99.6 (246)
> 0.6	11.63 (10)	72.5 (58)	29.73 (11)	97.57 (241)
> 0.7	4.65 (4)	62.5 (50)	13.51 (5)	95.14 (235)
> 0.8	3.49 (3)	51.25 (41)	8.11 (3)	87.45 (216)
> 0.9	0 (0)	38.75 (31)	5.41 (2)	70.04 (173)
> 1	0 (0)	0 (0)	0 (0)	0 (0)

The table shows the number and the percentage of poorly and well-imputed SNPs (see Materials and Methods) that are captured for each Rsq (A) and INFO (B) threshold. Imputation was performed on chromosome 20 OmniExpress SNPs, using the 1000G-EUR panel. Statistics are reported separately for common and rare variants.

Table S15. Efficiency of imputation quality metrics in the MCTFR cohort, for 1000G-EUR based imputations

A	MAF < 1%		MAF ≥ 1%	
	1000G-EUR		1000G-EUR	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (96)	100 (83)	100 (2)	100 (256)
> 0.1	51.04 (49)	98.8 (82)	50 (1)	100 (256)
> 0.2	28.12 (27)	96.39 (80)	0 (0)	100 (256)
> 0.3	18.75 (18)	92.77 (77)	0 (0)	98.83 (253)
> 0.4	9.38 (9)	86.75 (72)	0 (0)	96.48 (247)
> 0.5	4.17 (4)	78.31 (65)	0 (0)	92.58 (237)
> 0.6	3.12 (3)	63.86 (53)	0 (0)	83.59 (214)
> 0.7	2.08 (2)	54.22 (45)	0 (0)	73.83 (189)
> 0.8	0 (0)	42.17 (35)	0 (0)	63.28 (162)
> 0.9	0 (0)	26.51 (22)	0 (0)	50 (128)
> 1	0 (0)	0 (0)	0 (0)	0 (0)

B	MAF<1%		MAF ≥ 1%	
	1000G-EUR		1000G-EUR	
INFO	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (80)	100 (62)	100 (8)	100 (241)
> 0.1	98.75 (79)	100 (62)	100 (8)	100 (241)
> 0.2	92.5 (74)	100 (62)	100 (8)	100 (241)
> 0.3	71.25 (57)	100 (62)	100 (8)	100 (241)
> 0.4	42.5 (34)	100 (62)	100 (8)	100 (241)
> 0.5	25 (20)	98.39 (61)	50 (4)	100 (241)
> 0.6	7.5 (6)	93.55 (58)	12.5 (1)	100 (241)
> 0.7	2.5 (2)	80.65 (50)	0 (0)	95.02 (229)
> 0.8	0 (0)	59.68 (37)	0 (0)	83.82 (202)
> 0.9	0 (0)	35.48 (22)	0 (0)	61.83 (149)
> 1	0 (0)	0 (0)	0 (0)	0 (0)

The table shows the number and the percentage of poorly and well-imputed SNPs (see Materials and Methods) that are captured for each Rsq (A) and INFO (B) threshold. Imputation was performed on chromosome 20 Illumina 660W-quad SNPs, using the 1000G-EUR panel. Statistics are reported separately for common and rare variants.

6. URLs of Web Resources

- **Data access:** Part of the sequenced samples is available under the SardiNIA Medical Sequencing Discovery Project, dbGaP Study Accession: phs000313.v3.p2
- **1000 Genomes: http://www.1000genomes.org/home**
- **1000Genomes imputation cookbook:** http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook
- **HumanExome Design: http://genome.sph.umich.edu/wiki/Exome_Chip_Design**
- **ExomePicks: http://genome.sph.umich.edu/wiki/ExomePicks**
- **GotCloud: http://genome.sph.umich.edu/wiki/GotCloud**
- **Picard: http://picard.sourceforge.net.**
- **VerifyBamID: http://genome.sph.umich.edu/wiki/VerifyBam**

7. References

1. 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.

2. Li Y, Willer C, Sanna S, Abecasis GR: Genotype Imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406. doi: 10.1146/annurev.genom.9.081307.164242.

3. Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol. 1990 Mar;7(2):111-22.

4. Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995 Sep;12(5):921-7.

5. Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001 Apr;68(4):978-89. Epub 2001 Mar 9.

6. Porcu E, Sanna S, Fuchsberger C, Fritsche LG: Genotype imputation in genome-wide association studies. Curr Protoc Hum Genet 2013; Chapter 1: Unit1.25.

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

8. 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.

9. 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.

10. Voight BF, Kang HM, Ding J et al: The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 2012; 8: e1002793.

11. Cortes A, Brown MA: Promise and pitfalls of the Immunochip. Arthritis Res Ther 2011; 13: 101.

12. Iacono WG, McGue M, Krueger RF: Minnesota Center for Twin and Family Research. Twin Res Hum Genet 2006; 9: 978–984.

13. Miller MB, Basu S, Cunningham J et al: The Minnesota Center for Twin and Family Research genome-wide association study. Twin Res Hum Genet 2012; 15:

767–774.

14. Goldstein JI, Crenshaw A, Carey J et al: zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics 2012; 28: 2543–2545.

15. Scuteri A, Sanna S, Chen WM et al: Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 2007; 3: e115.

16. Naitza S, Porcu E, Steri M et al: A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet 2012; 8: e1002480.

17. Vrieze SI, Feng S, Miller MB et al: Rare nonsynonymous exonic variants in addiction and behavioral disinhibition. Biol Psychiatry 2013; 75: 783–789.

18. Pilia G, Chen WM, Scuteri A et al: Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet 2006; 2: e132.

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

20. Pitzalis M, Zavattari P, Murru R et al: Genetic loci linked to type 1 diabetes and multiple sclerosis families in Sardinia. BMC Med Genet 2008; 9: 3.

21. Orru V, Steri M, Sole G et al: Genetic variants regulating immune cell levels in health and disease. Cell 2013; 155: 242–256.

22. Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26: 589-595.

23. Li H, Handsaker B, Wysoker A et al: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078-2079.

24. Pireddu L, Leo S, Zanetti G: SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics 2011; 27: 2159-2160.

25. McKenna A, Hanna M, Banks E et al: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20: 1297-1303.

26. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997v2, 2013.

27. Jun G, Wing MK, Abecasis GR, Kang HM: An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA

sequence data. Genome Res 2014 (submitted).

28. 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.

29. Delaneau O, Zagury JF, Marchini J: Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 2013; 10: 5–6.

30. Su Z, Marchini J, Donnelly P: HAPGEN2: simulation of multiple disease SNPs. Bioinformatics 2011; 27: 2304–2305.

31. Danecek P, Auton A, Abecasis G et al: The variant call format and VCFtools. Bioinformatics 2011; 27: 2156–2158.

32. Nelson SC, Doheny KF, Pugh EW et al: Imputation-based genomic coverage assessments of current human genotyping arrays. G3 (Bethesda) 2013; 3: 1795–1807.

33. Francioli LC, Menelaou A, Pulit SL et al: Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 2014; 46: 818–825.

34. Browning SR, Browning BL: Haplotype phasing: existing methods and new developments. Nat Rev Genet 2011; 12: 703-714.

35. Lindquist KJ, Jorgenson E, Hoffmann TJ, Witte JS: The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. Genet Epidemiol 2013; 37: 383–392.

36. Holm H, Gudbjartsson DF, Sulem P et al: A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 2011; 43: 316–320.