G3 Genes | Genomes | Genetics

INVESTIGATION

Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies

Genevieve L. Wojcik,* Christian Fuchsberger,[†,‡,1] Daniel Taliun,[†] Ryan Welch,[†] Alicia R Martin,[*,2] Suyash Shringarpure,[*,3] Christopher S. Carlson,[§] Goncalo Abecasis,[†] Hyun Min Kang,[†] Michael Boehnke,[†] Carlos D. Bustamante,[*,] Christopher R. Gignoux,[*,4,5] and Eimear E. Kenny[††,‡‡,§§,***,5]**

*Department of Genetics and **Department of Biomedical Data Science, Stanford University School of Medicine, 365 Lasuen Street, Littlefield Center MC2069, Stanford, CA 94305, [†]Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, [‡]Center for Biomedicine, European Academy of Bolzano/Bozen (EURAC), affiliated with the University of Lübeck, Bolzano, Bozen, 39100, Italy, [§]Fred Hutchinson Cancer Center, University of Washington, 1100 Fairview Ave. N., Seattle, WA 98109, [††]Department of Genetics and Genomic Sciences, [‡‡]The Charles Bronfman Institute for Personalized Medicine, [§§]The Icahn Institute of Multiscale Biology and Genomics and ***The Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, NY 10029

ORCID IDs: 0000-0001-7206-8088 (G.L.W.); 0000-0003-0241-3522 (A.R.M.); 0000-0001-9728-6567 (C.R.G.); 0000-0001-9198-759X (E.E.K.)

ABSTRACT The emergence of very large cohorts in genomic research has facilitated a focus on genotype-imputation strategies to power rare variant association. These strategies have benefited from improvements in imputation methods and association tests, however little attention has been paid to ways in which array design can increase rare variant association power. Therefore, we developed a novel framework to select tag SNPs using the reference panel of 26 populations from Phase 3 of the 1000 Genomes Project. We evaluate tag SNP performance *via* mean imputed r^2 at untyped sites using leave-one-out internal validation and standard imputation methods, rather than pairwise linkage disequilibrium. Moving beyond pairwise metrics allows us to account for haplotype diversity across the genome for improve imputation accuracy and demonstrates population-specific biases from pairwise estimates. We also examine array design strategies that contrast multi-ethnic cohorts *vs.* single populations, and show a boost in performance for the former can be obtained by prioritizing tag SNPs that contribute information across multiple populations simultaneously. Using our framework, we demonstrate increased imputation accuracy for rare variants (frequency $< 1\%$) by 0.5–3.1% for an array of one million sites and 0.7–7.1% for an array of 500,000 sites, depending on the population. Finally, we show how recent explosive growth in non-African populations means tag SNPs capture on average 30% fewer other variants than in African populations. The unified framework presented here will enable investigators to make informed decisions for the design of new arrays, and help empower the next phase of rare variant association for global health.

KEYWORDS
Genomics
Statistical Genetics
Imputation
tag SNPs
array design

There is a growing recognition in genomic research of the need for very large-scale associations studies and genome-wide arrays are often the cost-efficient technology of choice. In this study we explore ways to improve array design for rare variant imputation, an underused means to increase power in association studies. We describe a pipeline in which an array is empirically evaluated based on genome-wide imputation accuracy, rather than pairwise linkage disequilibrium, to improve tagging and give real-world estimates of array performance. We explore the impact of patterns of demography on array performance, and discuss the trade-off between accurate rare variant imputation and *trans*-ethnic utility. This work provides a framework and insights that can guide the next generation of array development.

The vast majority of human genomic variation is rare (Nelson *et al.* 2012), and an appreciable fraction of rare variants are likely to be functionally consequential. (Kircher *et al.* 2014) The gold standard approach to assay rare variation (MAF $< 1\%$) is via deep sequencing. So far, large-scale sequencing studies have had some, but limited, success for discovery of rare variant associations (Emond *et al.* 2012;

Lohmueller *et al.* 2013; SIGMA Type 2 Diabetes Consortium *et al.* 2014; UK10K Consortium *et al.* 2015). There is a new appreciation that studies of hundreds of thousands or millions of individuals will be needed to drive well-powered discovery efforts. (Lindquist *et al.* 2013; Kosmicki *et al.* 2016) Currently, genome sequencing on this scale is prohibitively expensive and computationally burdensome. In contrast, genome-wide genotyping arrays are inexpensive, with far less bioinformatic overhead compared to sequencing. The past decade of genomic research has seen the development of myriad commercial high-throughput genotyping arrays.(Hoffmann *et al.* 2011a; Hoffmann *et al.* 2011b) While initially designed to capture common variants (International HapMap Consortium 2003), in recent years arrays have been leveraged to capture variation at the rare end of the frequency spectrum. One strategy is to ascertain rare variants directly on arrays, which is restricted to a very narrow subset of the rare variant spectrum due to array size limits. (Igartua *et al.* 2015; Wessel *et al.* 2015; McCarthy *et al.* 2017) Another strategy is to leverage the haplotype structure determined by common variants on the array, which form a 'scaffold', for accurate inference of un-genotyped variation through multi-marker imputation into sequenced reference panels of whole genomes. The strategy of genotyping, followed by imputation, has the potential to recover rare untyped variants in very large cohorts of arrayed samples at no additional experimental cost. (Huang *et al.* 2015; Michailidou *et al.* 2015) Imputation increases the effective sample size, leading to increased statistical power. (Pritchard and Przeworski 2001) This model bridging genotyping and imputation has prompted efforts to build deep reference sequence databases and a renewed interest in methods for improving genome-wide scaffold design. (1000 Genomes Project Consortium *et al.* 2015; UK10K Consortium *et al.* 2015; McCarthy *et al.* 2016).

Genotype array scaffolds have historically been designed using algorithms that select tagging single nucleotide polymorphisms (tag SNPs) that are in linkage disequilibrium (LD) with a maximal number of other SNPs. Tag SNP algorithms are optimized to maximize this score, typically described as pairwise coverage. However, imputation tools increasingly incorporate sophisticated haplotype information to impute unobserved variants. (Howie *et al.* 2012; Fuchsberger *et al.* 2014; Browning and Browning 2016) Consequently, it is not clear that tag SNPs that maximize pairwise coverage will be tag SNP's that provide, in aggregate, the best GWAS scaffold for accurate imputation. (de Bakker *et al.* 2005) Further, most tag SNP selection algorithms use LD architecture in a single population (Weale *et al.* 2003; Carlson *et al.* 2004), while we know LD patterns can vary extensively between populations. (1000 Genomes Project Consortium *et al.* 2015) Historically, many commercial arrays were designed by selecting tag SNPs from European populations, although arrays targeting some other populations have recently entered the market. (Hoffmann *et al.* 2011a; Hoffmann *et al.* 2011b) The number of SNPs tagged by a tag SNP can vary appreciably between populations due to demographic forces of migration, population expansion, and genetic drift. This may diminish GWAS scaffold performance in populations other than those in which the tag SNPs were selected, which in turn, can lead to reduced power for imputation-based association. This is a particularly pernicious problem in populations for which no targeted commercial array is available, in studies with multi-ethnic populations, and for accurate estimation of the transferability of genetic risk across populations.

As association studies grow larger and increasingly diverse, there is a need to reassess design criteria for GWAS scaffolds and arrays. (Carlson *et al.* 2013; Fuchsberger *et al.* 2016) On the one hand, tag SNPs that tag lower frequency variants are likely to be on the lower end of the site frequency spectrum and, consequentially, more geospatially restricted. (Nelson *et al.* 2008; Bustamante *et al.* 2011; Gravel *et al.* 2011; Mathieson and McVean 2014) On the other hand, as studies grow very large, cohort heterogeneity is likely to increase substantially. (Banda *et al.* 2015; Marouli *et al.* 2017) Given finite GWAS scaffold density, examining the trade-off between lowering the frequency threshold for accurate imputation and extending utility to multiple populations will become important. (Nelson *et al.* 2013; Martin *et al.* 2014) In this manuscript, we describe a framework for developing well-powered tag SNP selection leveraging thousands of whole genomes from diverse populations for balanced cross-population coverage. In our study, genomic coverage is evaluated based on genome-wide imputation accuracy as measured by mean imputed r^2 at untyped sites, rather than pairwise linkage disequilibrium. Moving beyond pairwise metrics allows us to account for haplotype diversity across the genome and demonstrates population-specific biases from pairwise estimates. Assessing accuracy using leave-one-out cross-validation yields a real-world estimate of genomic coverage. We examine the effect of allele frequency, correlation thresholds, and population diversity on the selection of tag SNP and on the landscape of tag-able variation. This work demonstrates that, while there may be limits given current reference panels, improving GWAS scaffold design is an underused means to increase power in association studies.

MATERIALS AND METHODS

Genetic Data

The genetic data are from the 1000 Genomes Project (1000 Genomes) Phase 3 data release, version 2 (7/8/2014) containing whole genome sequences for 2,535 individuals from 26 global populations. (1000 Genomes Project Consortium *et al.* 2015) Sequence data were in VCFv4.1 format, mapped to the forward strand and variants annotated as reference or alternate alleles. Only biallelic SNPs were included in this analysis (77,224,748 SNPs total). A list of known cryptically related individuals was obtained from the 1000 Genomes FTP site, and one individual from each related pair were subsequently removed (n = 62). Individuals were assigned to their super populations according to the original 1000 Genomes assignments (EAS = East Asian, EUR = European, AFR = African, SAS = South Asian, AMR = Americas, comprising 503, 501, 495, 477, and 341 individuals, respectively). Two populations of admixed African ancestry (ASW and ACB) were

doi: https://doi.org/10.1534/g3.118.200502
Manuscript received June 14, 2018; accepted for publication August 3, 2018; published Early Online August 25, 2018.

Supplemental material available at Figshare: https://doi.org/10.25387/g3.6626762.
[1]Current address: Institute of Biomedicine, EURAC research, Viale Druso, 1, 39100 Bolzano/Bozen, Italy
[2]Current address: Analytic and Translational Genetics Unit, Massachusetts General Hospital, Richard B. Simches Research Center, 185 Cambridge Street, CPZN-6818, Boston, MA 02114
[3]Current address: 23andMe, 899 W. Evelyn Ave, Mountain View, CA 94041
[4]Current address: Colorado Center for Personalized Medicine and Department of Biostatistics, University of Colorado Anschutz Medical Campus, 13001 E 17th Pl, Aurora, CO 80045
[5]Corresponding Authors: Center for Population Genomic Health, Icahn School of Medicine at Mount Sinai, Box 1003, One Gustave L Levy Place, NY 10029. Email: Christopher R Gignoux (chris.gignoux@ucdenver.edu) and Eimear E. Kenny (eimear.kenny@mssm.edu)

removed from the African super population and formed a separate African American/Caribbean (AAC) super population (n = 156).

Tag SNP Selection

Allele frequency was estimated within super population for each SNP using Plink v1.9. (Chang *et al.* 2015) Linkage Disequilibrium (LD) was also calculated within each super population using Plink v1.9 and settings for pairwise linkage with a minimum r^2 of 0.2 within a maximum distance of 1 megabase (mb). Tag SNP selection was performed per chromosome in the program TagIT (Weale *et al.* 2003) (https://github.com/statgen/TagIt), with frequency and LD files for each super population as input. The TagIT algorithm analyzed each super population separately. After filtering based on the minor allele frequency (set as either 0.5%, 1% or 5%), TagIT annotates the tag SNP that has the highest number of LD pairs with r2 above a minimum threshold (set as either 0.2, 0.5, or 0.8). The selected tag SNP and all of its linked SNPs are masked and TagIT finds the next tag SNP with the highest number of LD pairs. The output for each super population included for each index tag SNP the number of sites in LD, as well as the number of unique sites that weren't already tagged by a previously chosen tag. The number of unique SNPs tagged across all populations per tag SNP was tallied in the final output.

Cross-population tag SNP ranking and scoring

The naive approach ranked potential tags by the absolute number of unique SNPs that are tagged across all super populations. From this list, the top SNPs were selected for the appropriate allocation. To ensure performance of the tags across multiple populations, the cross-population prioritization schema first ranks tags by the number of populations in which they are informative, meaning they tag at least one site (Supplementary Figure 1). This ensures that the top ranked SNPs are not biased to a super population with large LD blocks or high SNP density in which one tag can contribute information about many other SNPs. Within each one of these categories (all 6 super populations down to only 1 super population), the tags are ranked by the number of unique tags across all six super populations, as was done in the original approach. The appropriate allocation is selected from the top of this list, scaled to the size of the chromosome of interest.

Metric of Performance

Coverage and imputation accuracy were assessed using all polymorphic biallelic sites within the 1000 Genomes Phase 3 data release, version 2. Sites were categorized into ten discrete minor allele frequency bins: (0.005-0.01], (0.01-0.02], (0.03-0.04], (0.04-0.05], (0.05-0.1], (0.1-0.2], (0.2-0.3], (0.3-0.4], and (0.4-0.5]. The term "coverage" is used to denote the proportion of untyped sites that had at least one tag SNP with pairwise r^2 greater than a certain threshold (0.2, 0.5, or 0.8). Imputation accuracy was determined through a leave-one-out internal validation approach with the 1000 Genomes Project Phase 3 data using a modified version of Minimac. (Fuchsberger *et al.* 2014) For this approach, each individual within the 1000Genomes data had the appropriate tag SNPs denoted as 'genotyped', with all other sites set as missing. These missing sites are then imputed using the rest of the 1000Genomes panel as a reference. Correlation was calculated comparing the estimated dosages from this imputation to the true genotypes from the original VCF files. While this internal validation approach may introduce overfitting of the data and an upwards bias of imputation accuracy, we sought the relative imputation accuracy for different methods and do not see any bias altering described trends.

Ascertainment Bias Analyses

Population-specific tags were selected separately through TagIT for each super population with a genome-wide allocation of 500,000 sites. All tags had a minimum MAF of 1% and a minimum r^2 threshold of 0.5. Each of the single population ascertained tag lists assessed for imputation accuracy in all six super populations, including their index population. Imputation accuracy was calculated as previously described and limited to chromosome 9.

Local Ancestry

Local ancestry was estimated using RFMix (Maples *et al.* 2013) assuming three ancestral backgrounds: African, European, and Native American, and is described in detail in (Martin *et al.* 2016) Tracts were dropped if smaller than 20 cM to improve accuracy in local ancestry estimation. Diploid ancestry with three ancestral backgrounds yielded six categories of variation. Imputation accuracy was then calculated separately per diploid tract category, with all other sections masked out. Results were aggregated across all chromosomes to calculate the genome-wide performance per diploid ancestry. Tracts were removed from analysis if the ancestral diplotype was found in fewer than 5 individuals. This included AFR-NAT and EUR-NAT within ACB which only occurred in 2 individuals each, NAT-NAT diplotypes in ASW which occurred in one individual, and AFR-AFR diplotypes in MXL which occurred in 3 individuals.

Cross-population patterns of linkage disequilibrium

To determine how many sites were in LD with tag SNPs across all 6 super populations, we selected one million SNPs for a GWAS scaffold using a minimum r^2 of 0.5 and a minimum MAF of 0.01 on chromosome 9. We calculated the number of polymorphic sites (MAF $>$ 0.5%) and the proportion of these sites that were in LD ($r^2 > 0.5$ or $r^2 > 0.8$) with at least one tag marker. To determine sharing of tags across multiple populations, we calculated the proportion of tag markers that were informative in other populations, conditional upon them being informative in the index population. The proportion of sites shared among multiple populations was calculated as the proportion of tag SNPs that performed in a certain number of populations (from 1 to 6 super populations) per super population.

Tagging Potential

Tag SNPs were selected with a minimum r^2 of 0.5 and a minimum MAF of 0.01 on chromosome 9. The potential for tagging was determined assuming an infinite site scaffold, using all possible tags until every pairwise relationship with r^2 above 0.5 was captured. The average number of sites captured per tag was calculated in each super population separately, using only the tags that were informative within that population. We also calculated these trends assuming a scaffold of one million sites, following the same procedures. The "dark sites" were calculated as sites in which there was no pairwise correlation with any other site with $r^2 > 0.2$, determined separately for each super population.

Data Availability

The input data from 1000 Genomes Project, Phase 3 is publicly available at the following link: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/. The program TagIt is available on github (https://github.com/statgen/TagIt), as well as a tutorial for how to select tag SNPs as detailed in this manuscript (https://github.com/chrisgene/crosspoptagging). Supplemental material available at Figshare: https://doi.org/10.25387/g3.6626762.

RESULTS

Assessing population-specific imputation accuracy With standard GWAS scaffold design

First we designed an experiment to assess imputation accuracy performance comparing tag SNP selection from different populations. This

experiment mimics the current design of many commercial arrays, in which tag SNPs were selected to capture the primarily variation in a single population or a closely related group of populations. We built a pipeline using the 26 population reference panel from Phase 3 of the 1000 Genomes Project and the Tagit algorithm(Taliun) for tag SNP selection. (Weale *et al.* 2003) (Supplementary Table 1) Individuals were split into mutually exclusive "super populations." These included the Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) populations as described in 1000 Genomes Project Consortium *et al.* (2015) In addition, we divided the African super population into two groups: four populations from Africa (AFR) and two populations of African descent in the Americas (AAC) (see **Methods**). Initially, to mimic the design of many arrays, tag SNPs were selected from a single super population. We assumed a genome-wide allocation of 500,000 tag SNPs, however analyses for a single population tagging strategy were only conducted on chromosome 9 with the allocation of 21,107 sites proportional to the physical distance of chromosome 9 compared to all chromosomes combined. Potential tags were required to have a minor allele frequency (MAF) $\geq$ 1% and be in pairwise LD with the tagged target site with a $r^2 \geq 0.5$.

The current generation of phase-based imputation algorithms (BEAGLE, IMPUTE2, Minimac3) leverage local haplotype information and sequenced reference panels to improve accuracy of variant inference compared to tag SNP approaches. (Marchini *et al.* 2007; Browning and Browning 2007; Howie *et al.* 2009; Marchini and Howie 2010; Fuchsberger *et al.* 2014; Browning and Browning 2016) Therefore, optimal array design depends not only on tag SNP selection, but also on empirical evaluation of imputation performance. For each of the population-specific GWAS scaffolds, imputation accuracy was assessed in all six super populations by MAF bins (common, MAF = 0.05-0.5; low frequency, MAF = 0.01-0.05; and rare, MAF $<$ 0.01) by comparing the imputed dosages to the real genotypes through leave-one-out internal validation. (see Methods).

Consistently across all super populations, the population from which the tags were ascertained had the highest imputation accuracy in the common bin. (S1 Fig) Trends in imputation accuracy follow known patterns of demography. For example, if the tags were ascertained in European populations, imputation accuracy was best in Europeans (EUR), followed by out-of-Africa populations (AMR, SAS, EAS), and worst in African ancestry populations (AFR, AAC). (Figure 1) If the tags were ascertained in African populations, the inverse was observed. (S1 Fig) As expected, the same trend of reduced imputation accuracy in non-ascertained populations was exacerbated in the low frequency bin. Imputation of low frequency variants in East Asian populations (EAS) was consistently most challenging; even when tag SNPs were selected from EAS, accuracy of low frequency imputation was the same or better in other populations. This can be explained by evidence of a recent tight bottleneck followed by rapid population grown in EAS, resulting in a large proportion of rare variants that are difficult to tag due to lower LD, especially with a limited scaffold of 500,000 sites. (Gravel *et al.* 2011) In contrast, the imputation performance of tag SNPs ascertained in AFR, AMR, and AAC populations is the same or better compared to the performance in out-of-Africa populations. This is likely due to increased allelic heterogeneity in African ancestry populations, which results in greater haplotypic diversity and a higher chance that a rare variant is well tagged by a haplotype for imputation. (1000 Genomes Project Consortium *et al.* 2015) The imputation accuracy of AMR higher in the rare frequency bin (MAF 0.5–1%), independent of the ascertainment population, is likely due to longer haplotypes resulting from recent admixture, allowing the rare variation to be captured accurately given the limited allocation. (Gravel *et al.* 2013) Importantly, in each case we observe a notable drop-off in performance across most of the frequency spectrum when examining imputation coverage in populations diverging from the one used for tag SNP selection. (S1 Fig).

Comparing single *vs.* cross population tag SNP selection strategies

When developing a genotyping platform, it is useful to assess whether selected tag SNPs segregate in the population of interest and contribute to tagging by being in LD (high r^2) with untagged sites. For example, using Illumina's OmniExpress platform (Illumina) within the 1000 Genomes Project data, over 99.7% of the sites will be polymorphic (MAF $>$ 0.5%) in the overall dataset. However, when we stratify by super population, each group has a differential loss due to monomorphic sites. AFR loses only $<$1% of sites with a MAF $<$ 0.5%, whereas EUR and EAS lose 4.4% and 9.2% of variants, respectively. Reduction in tagging can result in loss of statistical power for downstream analysis. We quantify this as "informativeness", or the ability of a tag SNP to both segregate in the population and provide LD information ($r^2 > 0.5$ with at least one untagged site). Balancing representation of variation across all groups becomes very important in multi-ethnic studies.

To explore different approaches for GWAS scaffold design we compared three strategies for selecting tag SNPs; single population tag SNP ascertainment, in which all tags are selected from a single population; a 'naïve' approach, in which all populations are combined and tags are selected based on composite statistics derived from this multi-population pool; and a 'cross-population prioritization' approach, in which tags are prioritized if they are both informative in multiple populations and by the number of unique sites targeted across all groups (see Methods and S2 Fig). We generated lists of tags per method assuming a total genome-wide allocation of 500,000 sites and minimum thresholds of $r^2 > 0.5$ and minor allele frequency (MAF) $\geq$ 1%. Using these parameters, an exhaustive set of tag SNPs were selected using the naïve approach with tags ranked by the absolute number of sites tagged across the 6 super populations, regardless of how many super populations had LD between tags and targets. We then re-ranked them using the cross-population prioritization approach (S2 Fig).

To compare the three approaches, we tallied the number of informative tags per population for each method to investigate the added value of tags contributing information in multiple populations. (Figure 2) This was done for all 22 autosomes. As per the design, all the single-population tags were informative within the super population from which tag SNPs were selected. Comparing the naïve and cross-population approaches that selected tag SNPs across all populations, the cross-population prioritization approach increased the number of informative tag SNPs in all populations relative to the naïve approach. In the naïve approach, we observed that the majority of tag SNPs were selected from the AFR population, followed by AAC, due to African-descent populations having more polymorphic sites across the genome with lower linkage disequilibrium. (Henn *et al.* 2015; 1000 Genomes Project Consortium *et al.* 2015) Whereas in the cross-population prioritization approach variation specific to a single population is down-weighted, leading to more balanced representation between all 6 super populations. By leveraging cross-population information the largest boost in the proportion of tag SNPs contributing linkage disequilibrium information compared to the naive approach was observed in non-African descent populations (10.5%, 28.6%, 25.9%, and 28.7% in AMR, EAS, EUR and SAS, respectively). Even the African descent populations (AFR and AAC), which dominate the naïve approach, have a higher proportion of tags in linkage disequilibrium with target sites with the cross-population prioritization approach (a 2.2% and 1.0% boost for AAC and AFR, respectively).

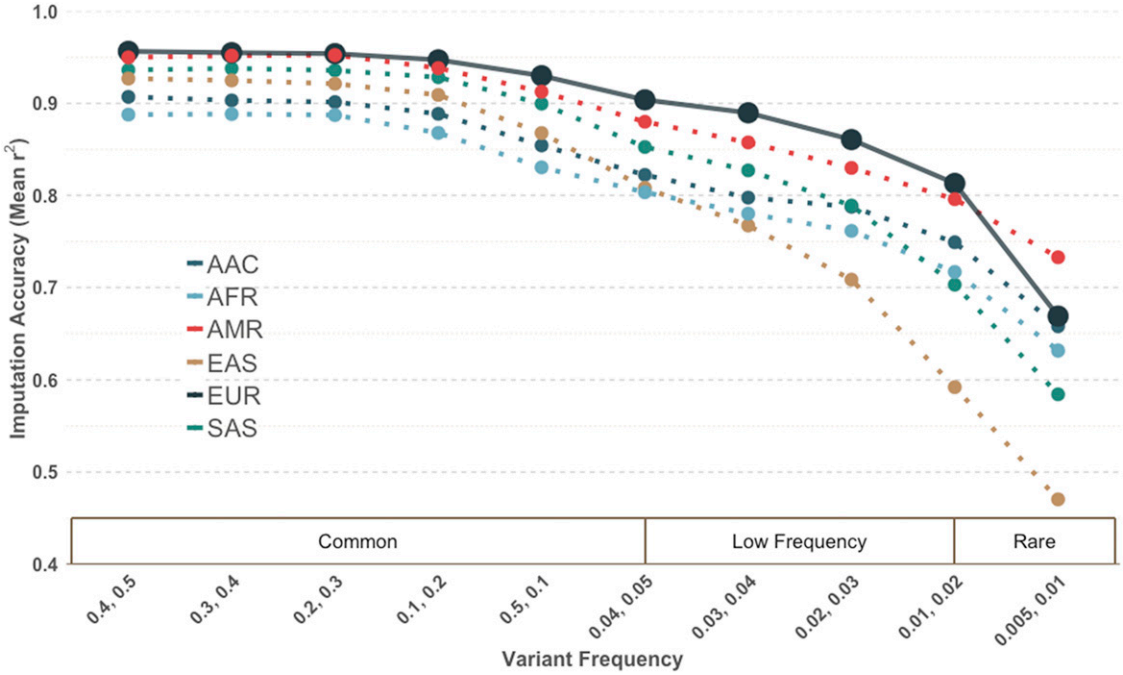

Figure 1 Imputation Accuracy by super population of tags selected in European populations for a scaffold assuming 500,000 genome-wide variants. Tags were required to have a MAF ≥ 1% and r^2 ≥ 0.5 with target sites. This trend is observed across all super populations *(S1 Fig)*.

To assess performance across the frequency spectrum we also stratified our accuracy estimates by super population-specific MAF into common, low frequency, and rare bins, as previously described. We observed that the cross-prioritization approach results in a larger proportion of tags being informative compared to both the single-population and naïve for common tag SNPs (MAF > 0.05) in all super populations. This is likely because the cross-prioritization approach prioritizes potential tag SNPs that provide LD information across multiple populations, therefore prioritizing common variants tagging common variation. However, by limiting tag SNP selection to these common variants only, the proportion of tags that provide LD information for low frequency variants is decreased compared to the single population approach, which had the highest proportion of informative tag SNPs in low and rare frequency in the target population. For example, when tags were ascertained using only AAC LD information, 19.5% of the 500,000 SNP scaffold were informative for rare variation (MAF < 1%) and 62.8% for common variation MAF > 5%) within AAC populations. When the cross-population approach was used, ensuring the prioritization of common variation, the proportion of tag SNPs informative for rare variation dropped to 6% while the proportion informative for common variation jumped up to 82.4%. This is consistent with low frequency and rare variants being population-specific, therefore not tagged by cosmopolitan common variation present in multiple populations. A notable exception is that the naïve approach contributes the most LD information for rare variants in the AMR super population. This is consistent with our previous findings showing highest imputation accuracy in the rare variation within AMR, even when the population from which tag SNPs were ascertained was different. The AMR on average exhibit longer haplotype lengths from the recently admixed populations in the Americas. (Gravel *et al.* 2013; 1000 Genomes Project Consortium *et al.* 2015) Because of the long haplotype tract lengths, more limited haplotypic diversity,

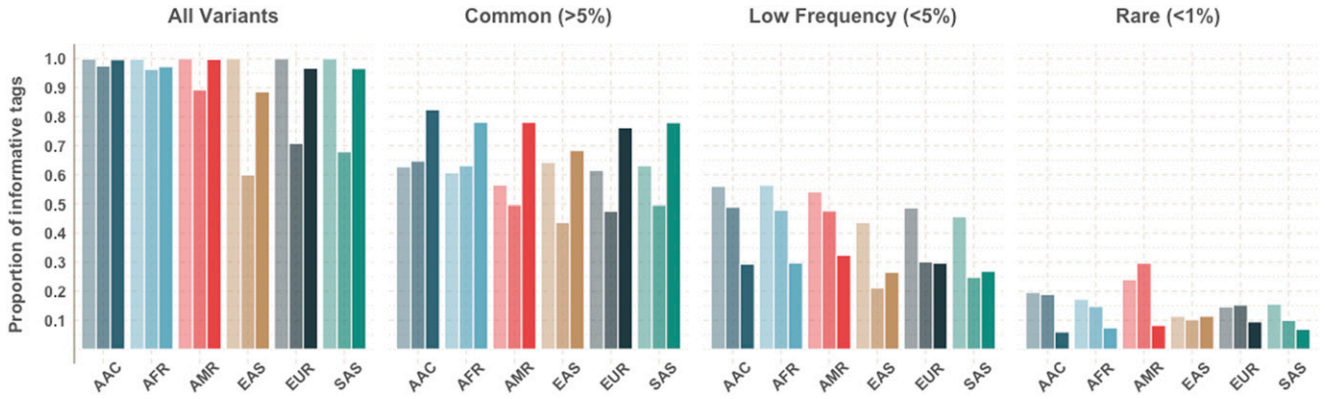

Figure 2 Proportion of tags that are informative by population with the three methods. (Left, lightest) tags selected from only a single population, (Center) tags selected by pooling all populations agnostically, and (Right) tags selected with the cross-population prioritization approach. Tag SNPs were informative if they were in linkage disequilibrium (r^2 > 0.5) with at least one untagged site.

and the limited allocation of tag SNPs, a naïve approach emphasizing the absolute number of unique sites up-weights variation that is informative for at least one of the ancestral components present in these populations.

Cross population prioritization of tag SNPs increases imputation accuracy for all groups Across frequency spectrum compared to naïve approach

The goal of tag SNP selection is to inform the unmeasured haplotypes, and therefore their performance must be evaluated in aggregate. One way to assess this is through imputation accuracy. Following the observation that cross-population prioritization selects a higher proportion of informative common tag SNPs for each population, even compared to the single population approach, we next assessed what impact this would have on imputation accuracy. We deployed the same leave-one-out internal cross validation approach as before using the 1000 Genomes Project populations (see Methods). We again assumed a genome-wide scaffold of 500,000 sites and tags had to have a MAF $>$ 1% and $r^2 > 0.5$ with tagged sites. Imputation accuracy was highest across all population-specific minor allele frequency bins when ascertaining in the target population in non-African non-admixed descent continental populations (EAS, EUR, and SAS). (S3 Fig) For the two African descent groups (AAC and AFR), the cross-population prioritization approach had the highest imputation accuracy across all sites. When stratified by MAF bins, the increase in informative tag SNPs for common variants with the cross population approach yielded higher imputation accuracy for common variation in all super populations. As previously seen, the population-specific nature of low frequency and rare variants led to decreased imputation accuracy in non-African descent populations for both the cross-population and naïve approach when compared to targeted single-population ascertainment. The cross-population prioritization approach had higher imputation accuracy than the naïve approach for all MAF bins.

As scaffold size can dramatically affect imputation accuracy(Spencer *et al.* 2009), we additionally examined allocations of 250,000, 500,000, 1,000,000, 1,500,000, and 2,000,000 genome-wide tags, which were all selected with $r^2 > 0.5$ and MAF $>$ 0.01. These allocations approximate the size range of many commercially available arrays. The cross-population prioritization scheme performed better with higher imputation accuracy than the naïve method for all super populations across all minor allele frequency bins with tags selected. (Figure 3) The biggest improvement came with the smaller array sizes. The most marked improvement was found in EAS, which originally had the lowest imputation accuracy of the 6 super populations with the naive approach. Within EAS groups, the cross-population approach increased imputation accuracy overall by 9.8% (from 67.3 to 77.1%) for a tag scaffold of 250,000 sites. For a scaffold of 500,000 sites, an overall improve of 6.2% was observed (from 77.4 to 83.6%). Improvements were largely consistent with the increase of informative tag SNPs. (Figure 2) As with the naive prioritization approach SNPs were disproportionately informative within AFR and AAC, consistent with admixed ancestry reflected by reference panels. For the smaller sizes (250K), the greatest increase in performance incorporating cross-population information was found within common SNPs (MAF $>$ 5%). However, the larger sized scaffolds (1-2 million) showed the most improvement within the low frequency bins (MAF $<$ 5%).

Imputation accuracy varies by local ancestry background in admixed individuals

We also assessed imputation ancestry stratified by local ancestry diplotype in the two admixed populations, the AAC and AMR, for a genome-wide allocation of 500,000 tag SNPs. First, using phased data, we inferred haploid tracts of African, European, and Native American local ancestry along the genomes of all individuals in the AMR and AAC populations (see Methods, (1000 Genomes Project Consortium *et al.* 2015; Martin *et al.* 2016)). Then each variant was inferred to be on one of six ancestral diploid tracts; European-European (EUR-EUR), European-African (EUR-AFR), European-Native American (EUR-NAT), African-Native American (AFR-NAT), African-African (AFR-AFR) and Native American-Native American (NAT-NAT). In all local ancestry strata the cross-population prioritization yielded improved imputation accuracy when compared to the naïve approach. When looking at ASW population (Americans of African ancestry in South West US), performance was high overall with all diploid tracts having imputation accuracies of 92.8–96.8% for all sites with minor allele frequency above 1%. (S4 Figure) The lowest imputation accuracy was found in AFR-AFR tracts, especially at the lower end of the frequency spectrum. The highest imputation accuracy was found in EUR-EUR tracts (94% overall for ASW). In AMR populations, by contrast, the NAT-NAT tracts had the lowest performance of all. An example can be seen in the MXL population (Mexican Ancestry from Los Angeles), where the highest imputation accuracy was found in the AFR-EUR tracts (overall imputation accuracy of 90.1% for all SNPs with MAF $>$ 0.5%) and the lowest within NAT-NAT tracts (74.8% for all SNPS with MAF $>$ 0.5%). (S4B Fig) These performances could be reflective of the relative availability of reference data relevant to these specific ancestral components.

Evaluating impact of r^2 and MAF thresholds on tag SNP performance

Previous standards in scaffold design have considered minimum linkage disequilibrium (r^2) and minor allele frequency (MAF) thresholds when prioritizing possible tag SNPs. However, the impact of these thresholds are often evaluated through pairwise coverage. We explored varying the minimum r^2 threshold (0.2, 0.5, 0.8) and MAF (0.5%, 1%, 5%) to assess their impacts on imputation accuracy, as well as pairwise coverage, assuming a genome-wide allocation of one million tags. For common variants, a higher minimum r^2 threshold ($r^2 > 0.8$) resulted in slightly higher imputation accuracy. (Figure 4A) However, the sites in the low and rare bin demonstrate population-specific accuracy only. (S5 Fig) For AFR, SAS, and EAS, a less stringent threshold of $r^2 > 0.2$ had the worst imputation accuracy across all frequency bins. Low frequency and rare variation had higher imputation accuracy for an r^2 threshold of 0.5 compared to 0.8. Within AAC, AMR, and EUR, the low frequency variation had improved imputation accuracy with the lowest r^2 threshold of 0.2. However, the imputation accuracy within this low threshold was notably compromised for common variants. This indicates that low frequency variation is better captured by weak correlation structure, but at a cost to common variation in these populations. Analyses performed with $r^2 > 0.5$ had the best balance of performance across all frequency bins with the highest overall imputation accuracy in all super populations except for EAS. (S2 Table) Overall, there were very small differences in imputation accuracy between the different r^2 thresholds. There were much larger differences in coverage, including both coverage evaluated with minimum r^2 (LD) of 0.5 and 0.8. (Figure 4A) Additionally, the best "performance" using pairwise coverage was highly dependent on the definition of coverage. Specifically, if pairwise coverage was calculated as the proportion of sites that are in LD with $r^2 > 0.5$, then the best minimum r^2 threshold in tag SNP selection will be 0.5. This holds true for $r^2 > 0.8$ as well.

The impact of minimum minor allele frequency threshold was negligible across variants with MAF $>$ 5% for all non-African populations (S6 Fig). Within populations of African descent, limiting tags to

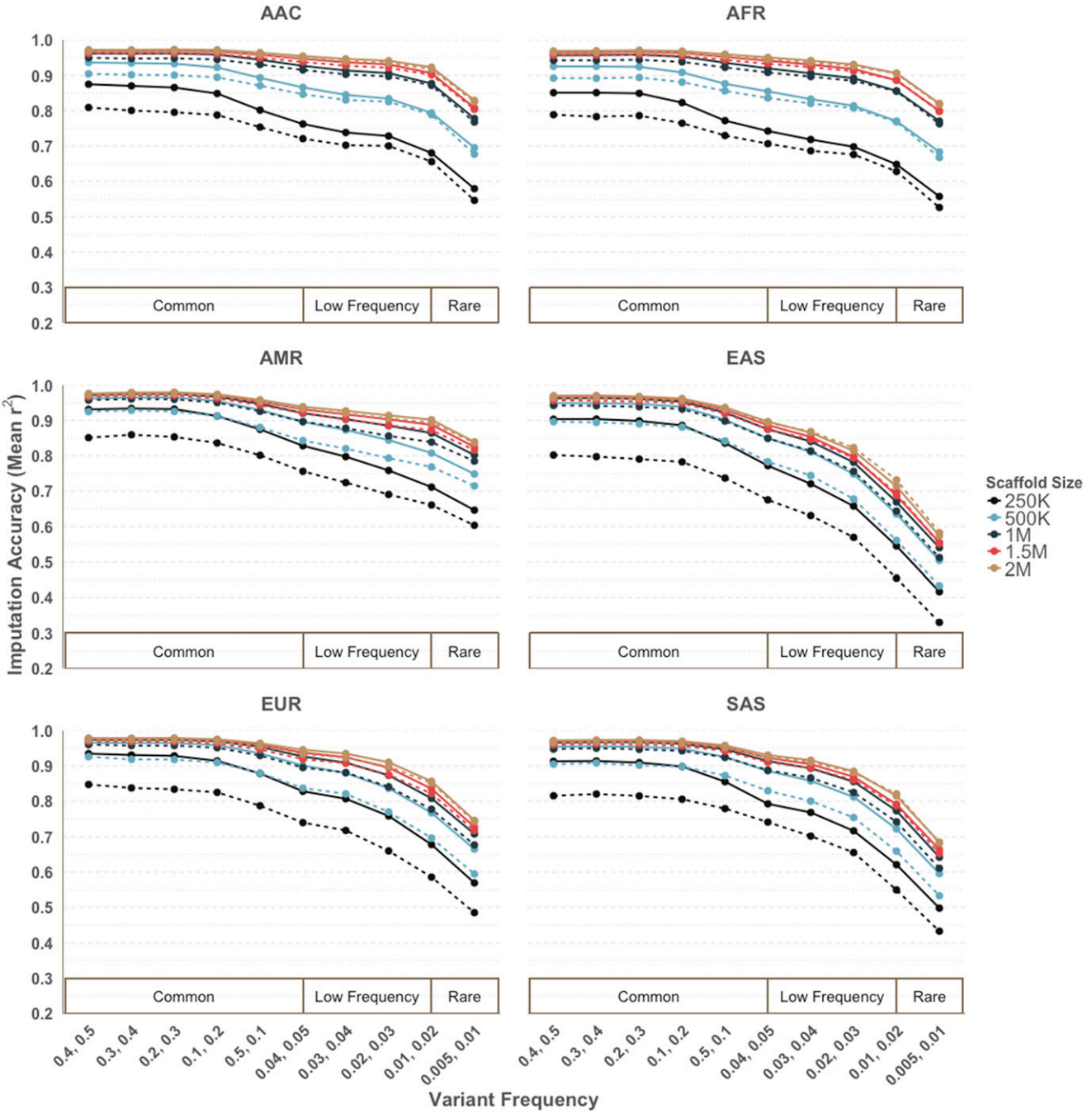

Figure 3 Increased imputation accuracy with cross-population prioritization (solid line) *vs.* naïve approach (dashed line) for a minimum pairwise correlation threshold of $r^2 > 0.5$ and MAF > 1% across different scaffold sizes. Imputation accuracy was calculated separately within minor allele frequency bins for each super population.

variants with MAF > 5% resulted in increased imputation accuracy for all frequency bins, especially for common variants. Lowering the MAF to 0.5% reduced accuracy in African-descent populations across all frequency bins. For EUR, SAS, and AMR, tags with MAF > 1% had decreased accuracy for variants with MAF 0.5–1% compared to when tags are limited to MAF > 0.5%. (Figure 4B) The lowest limit of MAF (0.5%) showed increased accuracy for rare variation but at a slight cost to the accuracy for common sites (MAF > 5%). We concluded that the best balance for tag SNP selection across all populations among these was MAF > 1% within the population being tagged, as the imputation accuracy was best for MAF > 5% for half of the groups (AAC, AFR, EAS) and best for MAF > 0.5% for the other half (AMR, EUR, SAS). (S2 Table) However, the overall differences in imputation accuracy was minimal, with less than 1% between all lower MAF thresholds across all sites. Again, we observed large differences in pairwise coverage, despite negligible differences when performance is evaluated by imputation accuracy. (S6 Fig) This is particularly striking for African-descent populations (ASW and AFR), where there were large gains of pairwise coverage for MAF > 1%, compared to MAF > 0.5% and MAF > 5%. As previously described, African populations have shorter

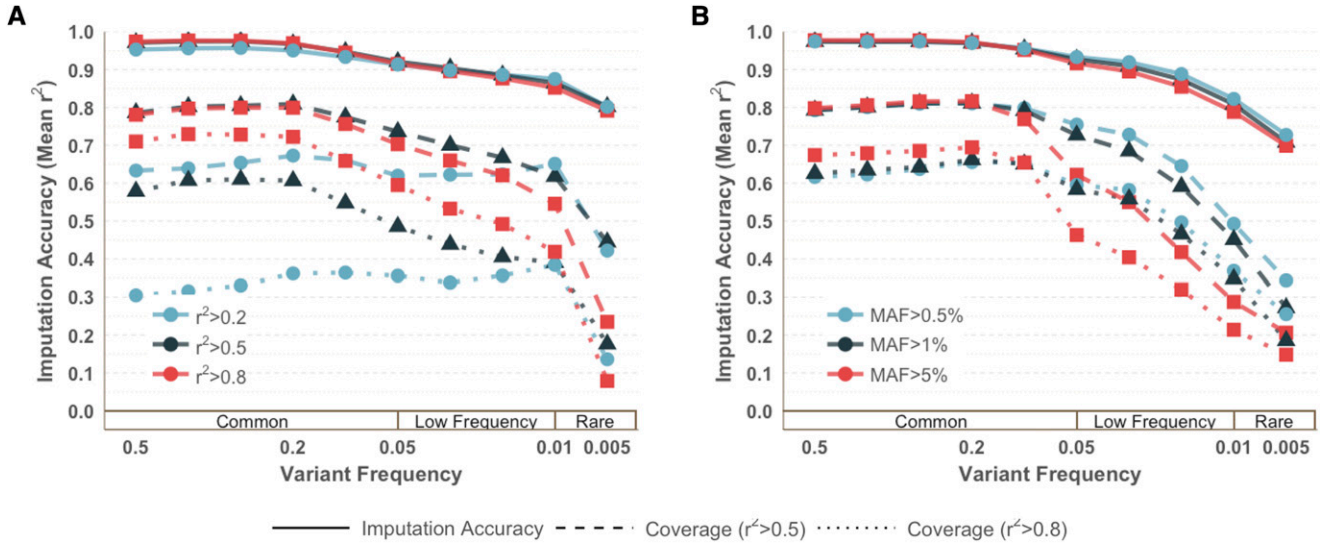

Figure 4 Influence of (A) minimum r^2 threshold and (B) lower MAF threshold on imputation accuracy and coverage ($r^2 > 0.5$ and $r^2 > 0.8$) within populations from the Americas with an allocation of 1M sites.

LD blocks and a greater absolute number of polymorphic variants compared to other populations. (1000 Genomes Project Consortium *et al.* 2015) Therefore, pairwise coverage underestimates performance compared to imputation accuracy, as addressed below.

Tagging potential differs between populations

Efficient tag SNP selection is an opportunity to boost power in downstream analyses. In our study, African and out-of-Africa populations exhibited distinct genetic architectures, which resulted in different performance trends. Even when cross-population performance was prioritized, it did not guarantee equal representation of all population groups within the tag SNP set. To determine the contribution of each population, we focused on chromosome 9 (42,215 tags), equivalent to one million sites genome-wide, selected with our novel cross-population prioritization scheme. This tag SNP allocation resulted in including all tags that were informative in at least 3 to all 6 populations in the scaffold. Out of all tags for chromosome 9, 17.96% were informative in all 6 populations. (S3 Table) No tags were included that were informative in only one or two populations. Of tags that were informative in 5 out of the 6 super-populations, only 54% were in LD with any target sites within EAS populations, while 93% were informative in AAC populations. (Figure 5A) This trend is consistent with cross-population tags tending to be less informative in EAS populations compared to the other populations. When tags are informative in 3 out of 6 groups, only 18% were informative in EAS, while 75% were informative in AAC. Tags informative in only 2 of the 6 groups were likely informative in AAC and AFR, the African descent populations, while very few of them were informative for non-African descent groups, consistent with capturing differential LD patterns in African populations.(Henn *et al.* 2011) When tags are stratified by MAF (0.5–1%, 1–5%, and >5%), these trends are exaggerated in the low frequency and rare MAF bins. (S7 Fig) As expected, the rare variation (0.5–1% MAF) was highly population-specific with no sites in this frequency bin being informative across all populations, or even 5 out of the 6 populations. (Gravel *et al.* 2011) For low frequency variation (1–5%), tags were the least informative within EAS, with only 36% of the tags informative in 5 out of 6 populations.

Conditional performance, or the ability of a tag which is informative in the index population also being informative in an additional population, was also examined and found to be consistent with known population histories. Of tags that are informative within AFR, 94% were informative within AAC, while only 38% were informative within EAS. (Figure 5B) However, among tags that were informative within EAS, 81% were informative within African populations. Once again, the stratified analyses show exaggerated trends for the low frequency and rare MAF bins. (S8 Fig) For the rare variation (0.5–1%), only a very small percentage (<10%) of tags are informative in other populations (AMR, EAS, EUR, SAS) if they were informative within African-descent populations (AFR and AAC). The high level of sharing between AFR and AAC is expected due to the high proportion of African ancestry within African-American and Afro-Caribbean populations. Of tags informative within EUR, 78% are also informative within AMR, largely due to the high proportion of European ancestry within some Hispanic/Latino populations.(Moreno-Estrada *et al.* 2013; Gravel *et al.* 2013; Moreno-Estrada *et al.* 2014).

The tags were also not equally informative in each population when it comes to the number of sites they tag with $r^2 > 0.5$. For chromosome 9, it would take 81,416 tags to capture all possible tag-able variation with an $r^2 > 0.5$ within AFR populations, while it would take only 28,473 tags within EAS populations to saturate coverage. However, each tag within the AFR populations captures on average 7.17 other sites, whereas for EAS populations, each tag captures on average 10.27 other SNPs. When restricting the design to a million tag SNP scaffold, each tag captures on average 16.16 other SNPs within EAS populations and 12.16 other SNPs in AFR populations. (Table 1) This reflects the different underlying genetic architecture of these different groups.

Limits of tagging and imputation

Not all of the human genome can be captured through pairwise tagging given existing reference panels. For each super population, we filtered for sites that were polymorphic (MAF > 0.5%) and had no pairwise correlation ($r^2 > 0.2$) with any other site within one megabase. The number of these "lone sites" without any pairwise correlation was dependent upon population. AAC had the greatest number of lone sites, but that is likely due to the significantly decreased sample size compared to the other populations. (Table 2) The lowest number of lone sites was found within AMR. Although these sites have no notable pairwise correlation

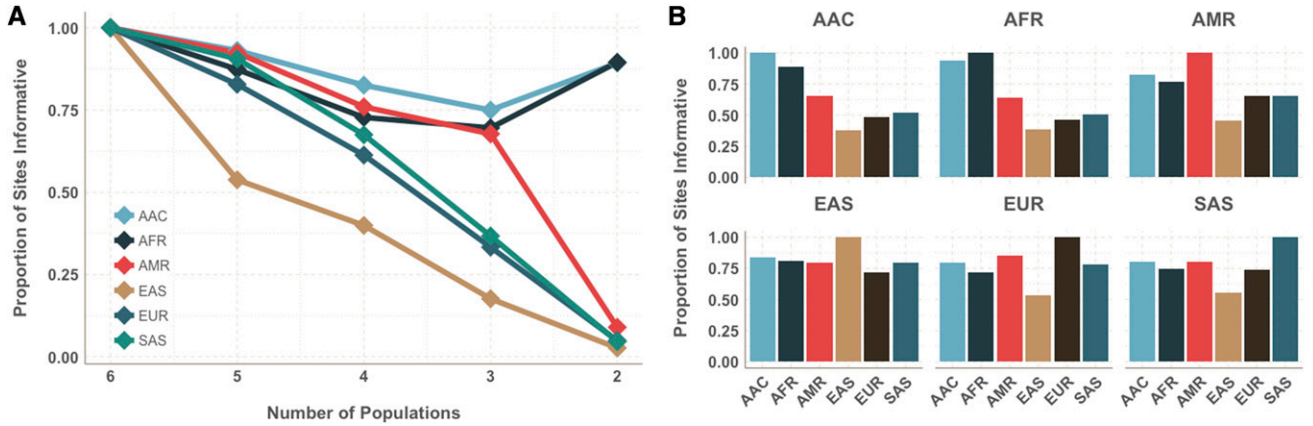

Figure 5 Tag SNPs informativeness across population. (A) Proportion of sites informative ($r^2 > 0.5$, MAF > 0.01, 1M site scaffold) across a number of populations, with lines corresponding to the index population. For example, for sites that are informative ($r^2 > 0.5$ with any untyped SNP in genome) in five out of the six populations, only slightly more than half are informative in East Asian populations while greater than 90% are informative in African populations. (B) Proportion of sites shared across populations, conditional on index population. For example, for sites informative in African populations, less than half are informative in East Asian, European, and South Asian populations.

with any other site in the human genome, haplotypes may be informative and allow the recovery of information for imputation. We again assumed a one million genome-wide tag SNP scaffold allocation with minimum MAF of 1% and minimum r^2 threshold of 0.5 and imputed to the entire 1000 Genomes reference panel. As expected, imputation accuracy and ability to recover information was population-specific. The imputation accuracy within AAC was an outlier when compared to other populations, with 80.72% of lone sites being imputed with at least the accuracy of $r^2_{acc} \geq 0.5$ and over 50% of sites being imputed with even higher accuracy ($r^2_{acc} \geq 0.8$). Many of these lone sites within AAC were captured with pairwise and haplotype LD within other populations, primarily AFR and to a lesser extent EUR. While there were likely insufficient allele counts for accurate correlation estimation within AAC due to the small sample size, this information could be recovered using a global reference panel. The number of unrecoverable "dark sites", which had no pairwise correlation and were not recoverable with imputation using haplotype information, was the largest in EAS and is consistent with known demography and population history yielding an excess of highly rare variation compared to other populations.(Gravel *et al.* 2011)

Pairwise coverage *vs.* imputation accuracy

When evaluating the performance of a GWAS scaffold, there are numerous factors to take into consideration. These include the number of sites you have allocated to tag SNPs and what your priorities are for balanced representation. To a lesser extent, the benefits and pitfalls of prioritizing low-frequency variants must be weighed. However, we have demonstrated that the influence of these factors is highly dependent on how performance is measured. The notion of genomic "coverage" has historically been estimated using pairwise correlations, and therefore this term will be used to denote the proportion of polymorphic sites that are in pairwise LD (r^2 threshold) with at least one tag SNP. We calculated coverage separately per super population at an r^2 threshold of 0.5 and 0.8 within minor allele frequency bins identical to the imputation accuracy estimation analyses, assuming a genome-wide tag SNP set of 500,000 and 1,000,000. (Table 3) For a tag SNP set of one million sites, coverage was lowest in AFR with an overall average of 59.15% for all sites with MAF > 0.5% and $r^2 > 0.5$. (S9 Fig) When the r^2 threshold is raised to 0.8, the proportion of sites in linkage disequilibrium with at least one tag SNP lowers to 28%. (Figure 6) The highest coverage was found in populations from the Americas (AMR) and East Asia (EAS). For a lower r^2 threshold of 0.5, 79.9% of AMR sites with MAF > 0.5% were covered. When using the higher r^2 threshold of 0.8, East Asian populations had the highest coverage with 63.08% of sites in LD with at least one tag SNP. This difference is even more marked when looking at a smaller tag SNP set of 500,000 sites. (S10 Fig, S11 Fig) African populations now have an overall coverage of 33.17% with $r^2 > 0.5$ and 14.10% with $r^2 > 0.8$. East Asian populations have the highest coverage with 73.16% of sites covered with $r^2 > 0.5$ and 55.09% with $r^2 > 0.8$.

These trends are in striking contrast to those we observed in imputation accuracy. When comparing a tag SNP set of 1 million, pairwise LD coverage is the lowest in populations of African descent (59% with $r^2 > 0.5$) yet imputation's ability to recover un-typed sites is on average high and consistent with other populations (imputation accuracy of 89.62%) among SNPs with a minor allele frequency above

Table 1 Performance per tag SNP to capture all variation possible with $r^2 > 0.8$ on chromosome 9, as well as within a one million site genome-wide scaffold allocation through cross-population prioritization

	All Possible Tags		One Million Tag Scaffold	
Population	*Number of Tags*	*Sites Captured per Tag*	*Number of Tags*	*Sites Captured per Tag*
AAC	74,255	8.04	36,336	12.97
AFR	81,416	7.17	34,548	12.16
AMR	43,065	9.40	28,691	12.80
EAS	28,473	10.27	16,457	16.16
EUR	35,027	9.48	22,111	13.63
SAS	37,644	9.28	23,480	13.33

Table 2 Lone sites by super population and their imputation accuracy for a one million site scaffold

Population	Number of Individuals	Number of Lone Sites	Imputation Accuracy Quality			Number Unrecoverable with $r^2_{acc} \geq 0.2$ (%)
			$r^2_{acc} \geq 0.2$	$r^2_{acc} \geq 0.5$	$r^2_{acc} \geq 0.8$	
AAC	156	7,509	90.79%	80.72%	51.72%	691 (9.2%)
AFR	495	4,497	63.29%	38.73%	7.03%	1,651 (36.7%)
AMR	341	2,701	48.98%	25.88%	3.78%	1,378 (51.02%)
EAS	503	4,947	44.37%	12.41%	2.14%	2,752 (55.63%)
EUR	501	3,881	51.07%	23.22%	3.74%	1,899 (48.93%)
SAS	477	4,293	51.01%	18.77%	2.26%	2,103 (48.99%)

0.5%. This contrast is also found in East Asian populations, which had one of the highest proportion of polymorphic SNPs with $r^2 > 0.5$ for coverage (76.95%), but the lowest imputation accuracy (86.28%). (Table 3) When sites are stratified by minor allele frequency bins, the differences in trends are even more striking. (Figure 6, S9 Fig) For example, within the lowest frequency bin (0.5–1%) for admixed populations of African-descent, the coverage of sites for a set of 500,000 tag SNPs with $r^2 > 0.8$ falls below 10%, however the imputation accuracy remains relatively high at 77.82%. These trends are consistent and more dramatic when evaluated within a tag SNP set of 500,000 sites. (S10 Fig, S11 Fig) These observations reinforce the necessity of examining imputation accuracy, instead of pairwise coverage, when evaluating the performance of tag SNPs.

DISCUSSION

As genomic researchers shift their focus to rare variant association in large and increasingly heterogeneous populations, it is important to design arrays with this ultimate goal in mind. There are currently two accepted methods of evaluating the performance of a tag SNPs: pairwise LD "coverage" and imputation accuracy. Coverage has historically been used as a term to denote the proportion of polymorphic sites that are in linkage disequilibrium with at least one tag marker above a certain r^2 threshold. (Barrett and Cardon 2006; Pe'er *et al.* 2006; Li *et al.* 2008; Bhangale *et al.* 2008) Genotyping arrays are typically compared using this score averaged across the genome. However, as we and others have demonstrated, restricting performance assessment to this definition of pairwise coverage is limited by removing multi-marker information. (Nelson *et al.* 2013; Martin *et al.* 2014) Evaluating imputation accuracy, particularly via leave-one-out cross validation, is highly computationally intensive, but provides a better assessment of how well untyped variation can be recaptured and a more realistic depiction of array performance than pairwise coverage. Imputation accuracy is also a more useful statistic in a practical sense, especially with the development of deeper and more diverse reference panels, (Prüfer *et al.* 2014; Gurdasani *et al.* 2015; Sudlow *et al.* 2015; 1000 Genomes Project Consortium *et al.* 2015; McCarthy *et al.* 2016) as performing GWAS with imputed variants is now the expectation. Emerging evidence suggests that rare variants (MAF < 1%) that are poorly tagged by an individual tag SNP will be accessible via imputation, due to added haplotype information, particularly as sample sizes move beyond the thousands into the tens or hundreds of thousands. (Nelson *et al.* 2013; Fuchsberger *et al.* 2014).

Previous tagging strategies have predominantly focused on optimizing performance in a single population. In prioritizing potential tags by their ability to provide linkage disequilibrium information across multiple populations, we were able to demonstrate that cross population tag SNP selection outperforms single population selection. This boost in imputation accuracy exists across all populations and frequency bins. We simulated tag SNP sets for a range of sizes (250,000-2 million), as well as for several minimum minor allele frequencies (0.5%, 1%, 5%) and minimum r^2 thresholds (0.2, 0.5, 0.8). For investigators with limited real estate or budget for tag SNP selection, we found that the biggest improvement in imputation accuracy provided with our cross population approach was with the smaller array sizes (250,000) when compared to a naïve design or biased population ascertainment. As expected, the influence of MAF and r^2 threshold was population-specific. For African-descent populations, including tag SNPs with a low threshold of $r^2 \geq 0.2$ resulted in lower imputation accuracy across all bins, while in other populations (EUR, AMR, SAS) tags at $r^2 \geq 0.2$ led to increased imputation accuracy for low frequency variants to the detriment of common variation. This is due to the lower LD patterns overall in African haplotypes, requiring denser coverage. The best balance was found with a moderate r^2 threshold of ≥ 0.5 for those seeking to perform well across all populations. This compromise is also present in choosing the lower MAF threshold. Limiting tag SNP selection to common variants with MAF $\geq$ 5% produced the highest imputation accuracy across all frequency bins within African-descent populations. However, this threshold decreased imputation accuracy for low frequency and rare variants in all other populations. Therefore, the best balance is once again found in the moderate value of MAF $\geq$ 1%. Investigators will need to take their priorities into account when selecting the correct thresholds for their populations and if they have a

Table 3 Coverage of 1 million and 500,000 tag SNP set by super population for all polymorphic sites on chromosome 9 with MAF > 0.5%

Super population	Total Number of Polymorphic Sites	Scaffold of 1,000,000 tags			Scaffold of 500,000 tags		
		Coverage		Imputation Accuracy	Coverage		Imputation Accuracy
		$r^2 > 0.5$	$r^2 > 0.8$		$r^2 > 0.5$	$r^2 > 0.8$	
AAC	780896	63.64%	30.27%	*90.59%*	34.03%	14.07%	*84.85%*
AFR	777207	59.15%	28.05%	*89.62%*	33.17%	14.10%	*83.32%*
AMR	503804	79.90%	53.60%	*92.77%*	61.00%	37.02%	*90.09%*
EAS	367189	76.95%	63.08%	*86.28%*	73.16%	55.09%	*84.16%*
EUR	414184	78.77%	62.65%	*91.02%*	72.87%	52.86%	*88.90%*
SAS	455573	74.84%	56.97%	*88.09%*	67.28%	45.91%	*85.46%*

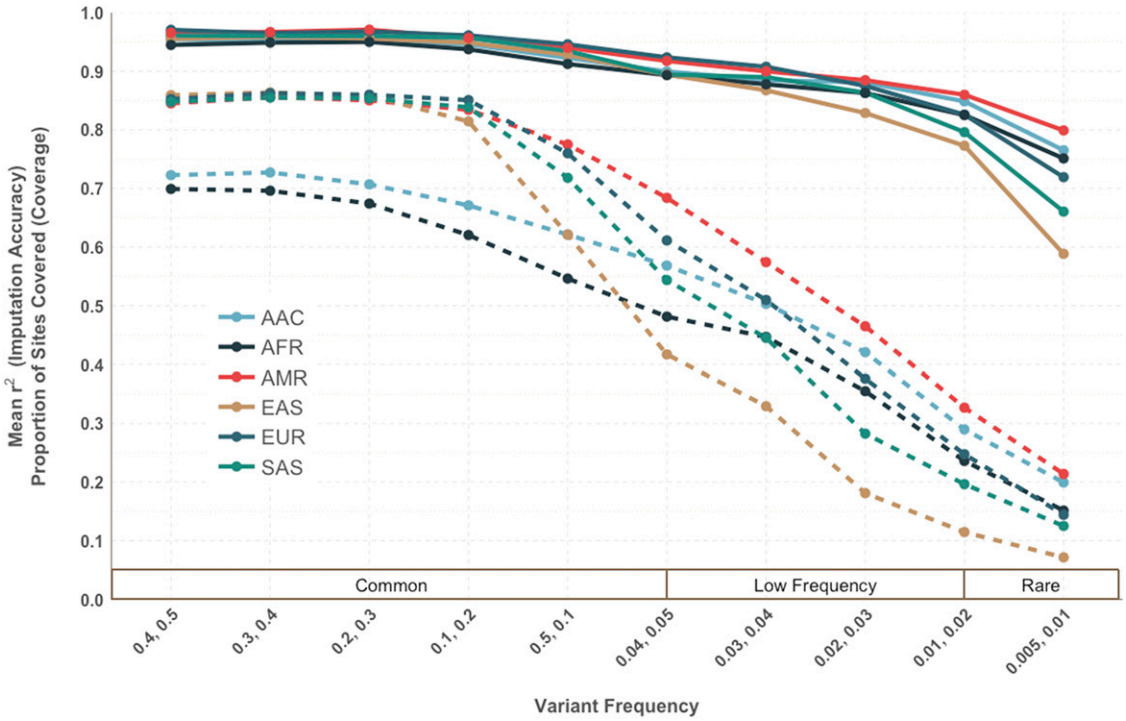

Figure 6 Coverage (dashed lines) *vs.* Imputation Accuracy (solid lines), assuming a genome-wide scaffold size of one million tags. Coverage is shown with an $r^2 > 0.8$. While pairwise tagging values are low, particularly in African-descent populations, multi-marker imputation accuracy remains high across groups.

specific target frequency bin. We chose to prioritize all populations equally to provide a design of broad global utility, which was adopted to construct the GWAS scaffold for Illumina Infinium Multi-Ethnic Global Arrays (Illumina) and Global Screening Arrays (Illumina). If a study is comprised of mostly one ancestral group, then the investigators should choose the appropriate thresholds tailored for their study.

Consistent with demographic history, the potential to capture variation with a limited allocation is unequal between the different populations in the 1000 Genomes Project. The naïve tagging approach will bias tag SNP selection to be primarily informative within African-descent populations. The absolute number of polymorphic sites within African populations is much larger than other populations, and while LD tends to be lower than in other populations, the high number of potential tags and pairwise correlations overwhelms the other populations' contributions without controlling for this unique pattern. By prioritizing potential tags that provide information across all populations, the population-level contributions are more balanced without detriment to the African-descent groups (Figure 4). The absolute number of rare variants (MAF $<$ 1%) is larger in African populations, but the frequency spectrum is more skewed toward rare variants in populations with recent bottlenecks and exponential population expansion, such as in East Asians. Contrasting these two populations (AFR and EAS), East Asian populations require fewer sites to saturate coverage, with each potential tag being in LD with more sites. However, far more polymorphic sites across the genome cannot be captured with either pairwise linkage disequilibrium or through haplotype information with imputation accuracy within these populations due to a dearth of LD information. This is amplified by the lack of comprehensive reference panels for many populations, such as East and South Asia. As reference panels are expanded, more variation will be captured to inform tag SNP selection and imputation accuracy, and we expect imputation accuracy to improve for all populations and across the frequency spectrum. (Fuchsberger *et al.* 2014).

The power to identify relevant disease loci is inherently constrained by sample size and genome coverage. It is important to note that algorithmic development both on association testing and imputation methods have been a productive avenue of research since GWAS began, with new methods providing incremental improvements in statistical power. Here, we demonstrate a complementary strategy to improve statistical power by designing arrays optimized for imputation accuracy. Also, as cosmopolitan biobanks and large-scale multi-ethnic epidemiological studies become more commonplace, it will be important to have available platforms with built *in trans*-ethnic utility. As global reference panels become deeper and more diverse, more variation will be available for array design. The unified framework presented here will enable investigators to make informed decisions in the development and selection of GWAS scaffolds for future large-scale multi-ethnic studies. This increased representation of multi-ethnic genetic variation will promote the investigation of the genetics of complex disease and the improvement of global health in the next phase of GWAS.

ACKNOWLEDGMENTS

Research reported in this paper was supported by the Office of Research Infrastructure under award number S10OD018522 and the National Human Genome Research Institute under award numbers U01HG007376, U01HG007417, U01HG007419, U01HG009080 and R01HG000376 of the National Institutes of Health. CRG was supported partially by T32HG00044. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

LITERATURE CITED

1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74. https://doi.org/10.1038/nature15393

Banda, Y., M. N. Kvale, T. J. Hoffmann, S. E. Hesselson, D. Ranatunga *et al.*, 2015 Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics 200: 1285–1295. https://doi.org/10.1534/genetics.115.178616

Barrett, J. C., and L. R. Cardon, 2006 Evaluating coverage of genome-wide association studies. Nat. Genet. 38: 659–662. https://doi.org/10.1038/ng1801

Bhangale, T. R., M. J. Rieder, and D. A. Nickerson, 2008 Estimating coverage and power for genetic association studies using near-complete variation data. Nat. Genet. 40: 841–843. https://doi.org/10.1038/ng.180

Browning, B. L., and S. R. Browning, 2016 Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. 98: 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084–1097. https://doi.org/10.1086/521987

Bustamante, C. D., E. G. Burchard, and F. M. De La Vega, 2011 Genomics for the world. Nature 475: 163–165. https://doi.org/10.1038/475163a

Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak *et al.*, 2004 Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74: 106–120. https://doi.org/10.1086/381000

Carlson, C. S., T. C. Matise, K. E. North, C. A. Haiman, M. D. Fesinmeyer *et al.*, 2013 Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. 11: e1001661. https://doi.org/10.1371/journal.pbio.1001661

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci 4: 7–16. https://doi.org/10.1186/s13742-015-0047-8

de Bakker, P. I. W., R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly *et al.*, 2005 Efficiency and power in genetic association studies. Nat. Genet. 37: 1217–1223. https://doi.org/10.1038/ng1669

Emond, M. J., T. Louie, J. Emerson, W. Zhao, R. A. Mathias *et al.*, 2012 Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat. Genet. 44: 886–889. https://doi.org/10.1038/ng.2344

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds, 2014 minimac2: faster genotype imputation. Bioinformatics.

Fuchsberger, C., J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala *et al.*, 2016 The genetic architecture of type 2 diabetes. Nature 536: 41–47. https://doi.org/10.1038/nature18642

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108: 11983–11988. https://doi.org/10.1073/pnas.1019276108

Gravel, S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio *et al.*, 2013 Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genet.* 9: e1004023. https://doi.org/10.1371/journal.pgen.1004023

Gurdasani, D., T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou *et al.*, 2015 The African Genome Variation Project shapes medical genetics in Africa. Nature 517: 327–332. https://doi.org/10.1038/nature13997

Henn, B. M., L. R. Botigué, C. D. Bustamante, A. G. Clark, and S. Gravel, 2015 Estimating the mutation load in human genomes. Nat. Rev. Genet. 16: 333–343. https://doi.org/10.1038/nrg3931

Henn, B. M., C. R. Gignoux, M. Jobin, J. M. Granka, J. M. Macpherson *et al.*, 2011 Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. USA 108: 5154–5162. https://doi.org/10.1073/pnas.1017511108

Hoffmann, T. J., M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino *et al.*, 2011a Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. Genomics 98: 79–89. https://doi.org/10.1016/j.ygeno.2011.04.005

Hoffmann, T. J., Y. Zhan, M. N. Kvale, S. E. Hesselson, J. Gollub *et al.*, 2011b Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics 98: 422–430. https://doi.org/10.1016/j.ygeno.2011.08.007

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, 2012 Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44: 955–959. https://doi.org/10.1038/ng.2354

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5: e1000529. https://doi.org/10.1371/journal.pgen.1000529

Huang, J., B. Howie, S. McCarthy, Y. Memari, K. Walter *et al.*, 2015 Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. 6: 8111. https://doi.org/10.1038/ncomms9111

Igartua, C., R. A. Myers, R. A. Mathias, M. Pino-Yanes, C. Eng *et al.*, 2015 Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nat. Commun. 6: 5965. https://doi.org/10.1038/ncomms6965

Illumina Infinium Global Screening Array Information Sheet.

Illumina Infinium Multi-Ethnic Global BeadChip Information Sheet.

Illumina Infinium OmniExpress-24 v1.2 BeadChip Information Sheet.

International HapMap Consortium, 2003 The International HapMap Project. Nature 426: 789–796. https://doi.org/10.1038/nature02168

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper *et al.*, 2014 A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46: 310–315. https://doi.org/10.1038/ng.2892

Kosmicki, J. A., C. L. Churchhouse, M. A. Rivas, and B. M. Neale, 2016 Discovery of rare variants for complex phenotypes. Hum. Genet. 135: 625–634. https://doi.org/10.1007/s00439-016-1679-1

Li, M., C. Li, and W. Guan, 2008 Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur. J. Hum. Genet. 16: 635–643. https://doi.org/10.1038/sj.ejhg.5202007

Lindquist, K. J., E. Jorgenson, T. J. Hoffmann, and J. S. Witte, 2013 The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. Genet. Epidemiol. 37: 383–392. https://doi.org/10.1002/gepi.21724

Lohmueller, K. E., T. Sparsø, Q. Li, E. Andersson, T. Korneliussen *et al.*, 2013 Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. Am. J. Hum. Genet. 93: 1072–1086. https://doi.org/10.1016/j.ajhg.2013.11.005

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93: 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

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

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39: 906–913. https://doi.org/10.1038/ng2088

Marouli, E., M. Graff, C. Medina-Gomez, K. S. Lo, A. R. Wood *et al.*, 2017 Rare and low-frequency coding variants alter human adult height. Nature 542: 186–190. https://doi.org/10.1038/nature21039

Martin, A. R., C. R. Gignoux, R. K. Walters, G. Wojcik, B. M. Neale *et al.*, 2016 Human demographic history impacts genetic risk prediction across diverse populations. https://doi.org/10.1016/j.ajhg.2017.03.004

Martin, A. R., G. Tse, C. D. Bustamante, and E. E. Kenny, 2014 Imputation-based assessment of next generation rare exome variant arrays. Pac. Symp. Biocomput. 3: 241–252.

Mathieson, I., and G. McVean, 2014 Demography and the Age of Rare Variants. *PLoS Genet.* 10: e1004528. https://doi.org/10.1371/journal.pgen.1004528

McCarthy, N. S., P. E. Melton, S. V. Ward, S. M. Allan, M. Dragovic *et al.*, 2017 Exome array analysis suggests an increased variant burden in families with schizophrenia. Schizophr. Res. 185: 9–16. https://doi.org/10.1016/j.schres.2016.12.007

McCarthy, S., S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood *et al.*, 2016 A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48: 1279–1283. https://doi.org/10.1038/ng.3643

Michailidou, K., J. Beesley, S. Lindstrom, S. Canisius, J. Dennis *et al.*, 2015 Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat. Genet. 47: 373–380. https://doi.org/10.1038/ng.3242

Moreno-Estrada, A., C. R. Gignoux, J. C. Fernández-López, F. Zakharia, M. Sikora *et al.*, 2014 The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science 344: 1280–1285. https://doi.org/10.1126/science.1251688

Moreno-Estrada, A., S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes *et al.*, 2013 Reconstructing the Population Genetic History of the Caribbean. *PLoS Genet.* 9: e1003925. https://doi.org/10.1371/journal.pgen.1003925

Nelson, M. R., K. Bryc, K. S. King, A. Indap, A. R. Boyko *et al.*, 2008 The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. Am. J. Hum. Genet. 83: 347–358. https://doi.org/10.1016/j.ajhg.2008.08.005

Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100–104. https://doi.org/10.1126/science.1217876

Nelson, S. C., K. F. Doheny, E. W. Pugh, J. M. Romm, H. Ling *et al.*, 2013 Imputation-based genomic coverage assessments of current human genotyping arrays. G3 (Bethesda) 3: 1795–1807. https://doi.org/10.1534/g3.113.007161

Pe'er, I., P. I. W. de Bakker, J. Maller, R. Yelensky, D. Altshuler *et al.*, 2006 Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat. Genet. 38: 663–667. https://doi.org/10.1038/ng1816

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69: 1–14. https://doi.org/10.1086/321275

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman *et al.*, 2014 The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49. https://doi.org/10.1038/nature12886

SIGMA Type 2 Diabetes Consortium, Estrada, K., I. Aukrust, L. Bjørkhaug, N. P. Burtt *et al.*, 2014 Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. JAMA 311: 2305–2314. https://doi.org/10.1001/jama.2014.6511

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genetics* 5: e1000477. https://doi.org/10.1371/journal.pgen.1000477

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton *et al.*, 2015 UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12: e1001779. https://doi.org/10.1371/journal.pmed.1001779

UK10K ConsortiumWalter, K., J. L. Min, J. Huang, L. Crooks *et al.*, 2015 The UK10K project identifies rare variants in health and disease. Nature 526: 82–90. https://doi.org/10.1038/nature14962

Weale, M. E., C. Depondt, S. J. Macdonald, A. Smith, P. S. Lai *et al.*, 2003 Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am. J. Hum. Genet. 73: 551–565. https://doi.org/10.1086/378098

Wessel, J., A. Y. Chu, S. M. Willems, S. Wang, H. Yaghootkar *et al.*, 2015 Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat. Commun. 6: 5897. https://doi.org/10.1038/ncomms6897

Communicating editor: R. Hernandez