

1 Imputation aware tag SNP selection to 2 improve power for multi-ethnic 3 association studies

4 Genevieve L. Wojcik¹, Christian Fuchsberger², Daniel Taliun², Ryan Welch², Alicia R Martin^{1, #a},
5 Suyash Shringarpure^{1, #b}, Christopher S. Carlson³, Goncalo Abecasis², Hyun Min Kang², Michael
6 Boehnke², Carlos D. Bustamante^{1,4}, Christopher R. Gignoux¹ and Eimear E. Kenny^{5,6,7,8}

7

8 1. *Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA*

9 2. *Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of*
10 *Michigan, Ann Arbor, MI, USA*

11 3. *Fred Hutchinson Cancer Center, University of Washington, Seattle, WA, USA*

12 4. *Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA*

13 5. *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY,*
14 *USA*

15 6. *The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New*
16 *York, NY, USA*

17 7. *The Icahn Institute of Multiscale Biology and Genomics, Icahn School of Medicine at Mount Sinai, New*
18 *York, NY, USA*

19 8. *The Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY, USA*

20

21 #a *Current address: Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA,*
22 *USA*

23 #b *Current address: 23andMe, Mountain View, CA, USA*

24

1 Abstract

2 The emergence of very large cohorts in genomic research has facilitated a focus on
3 genotype-imputation strategies to power rare variant association. Consequently, a new generation
4 of genotyping arrays are being developed designed with tag single nucleotide polymorphisms
5 (SNPs) to improve rare variant imputation. Selection of these tag SNPs poses several challenges
6 as rare variants tend to be continentally- or even population-specific and reflect fine-scale linkage
7 disequilibrium (LD) structure impacted by recent demographic events. To explore the landscape of
8 tag-able variation and guide design considerations for large-cohort and biobank arrays, we
9 developed a novel pipeline to select tag SNPs using the 26 population reference panel from Phase
10 3 of the 1000 Genomes Project. We evaluate our approach using leave-one-out internal validation
11 via standard imputation methods that allows the direct comparison of tag SNP performance by
12 estimating the correlation of the imputed and real genotypes for each iteration of potential array
13 sites. We show how this approach allows for an assessment of array design and performance that
14 can take advantage of the development of deeper and more diverse sequenced reference panels.
15 We quantify the impact of demography on tag SNP performance across populations and provide
16 population-specific guidelines for tag SNP selection. We also examine array design strategies that
17 target single populations versus multi-ethnic cohorts, and demonstrate a boost in performance for
18 the latter can be obtained by prioritizing tag SNPs that contribute information across multiple
19 populations simultaneously. Finally, we demonstrate the utility of improved array design to provide
20 meaningful improvements in power, particularly in trans-ethnic studies. The unified framework
21 presented will enable investigators to make informed decisions for the design of new arrays, and
22 help empower the next phase of rare variant association for global health.

1 Introduction

2 The vast majority of human genomic variation is rare ¹, and an appreciable fraction of rare variants
3 are likely to be functionally consequential. ² The gold standard approach to assay rare variation is
4 via sequencing. So far, large-scale sequencing studies have had some, but limited, success for
5 discovery of rare variant associations³⁻⁶. There is a new appreciation that studies of hundreds of
6 thousands or millions of individuals will be needed to drive well-powered discovery efforts. ^{7,8}
7 Currently, genome sequencing on this scale is prohibitively expensive and computationally
8 burdensome. In contrast, genome-wide genotyping arrays are inexpensive, with far less
9 bioinformatic overhead compared to sequencing. The past decade of genomic research has seen
10 the development of myriad commercial high-throughput genotyping arrays. ^{9,10} While initially
11 designed to capture common variants ¹¹, in recent years arrays have been leveraged to capture
12 variation at the rare end of the frequency spectrum. One strategy is to ascertain rare variants
13 directly on arrays, which is restricted to a very narrow subset of the rare variant spectrum due to
14 array size limits. ¹²⁻¹⁴ Another strategy is to leverage the haplotype structure determined by common
15 variants on the array, which form a 'scaffold', for accurate inference of un-genotyped variation
16 through multi-marker imputation into sequenced reference panels of whole genomes. The strategy
17 of genotyping, followed by imputation, has the potential to recover rare untyped variants in very
18 large cohorts of arrayed samples at no additional experimental cost. ^{15,16} This model bridging
19 genotyping and imputation has prompted efforts to build deep reference sequence databases and a
20 renewed interest in methods for improving genome-wide scaffold design. ^{6,17,18}

21 Genotype array scaffolds have been designed historically using algorithms that select
22 tagging single nucleotide polymorphisms (tag SNPs) that are in linkage disequilibrium (LD) with a
23 maximal number of other SNPs. Tag SNP algorithms are optimized to maximize this score, typically
24 described as pairwise coverage. However, imputation tools increasingly incorporate sophisticated
25 haplotype information to impute unobserved variants. ¹⁹⁻²¹ Consequently, it is not clear that tag
26 SNPs that maximize pairwise coverage will be tag SNP's that provide, in aggregate, the best GWAS
27 scaffold for accurate imputation. ²² Further, most tag SNP selection algorithms use LD architecture
28 in a single population, ^{23,24}, while we know LD patterns can vary extensively between populations.¹⁷
29 Historically, many commercial arrays were designed by selecting tag SNPs from European
30 populations, although arrays targeting some other populations have recently entered the market. ^{9,10}
31 The number of SNPs tagged by a tag SNP can vary appreciably between populations due to
32 demographic forces of migration, population expansion, and genetic drift. This may diminish GWAS
33 scaffold performance in populations other than those in which the tag SNPs were selected, which in
34 turn, can lead to reduced power for imputation-based association. This is a particularly pernicious
35 problem in populations for which no targeted commercial array is available, in studies with multi-

1 ethnic populations, and for accurate estimation of the transferability of genetic risk across
2 populations.

3 As association studies grow larger and increasingly diverse, there is a need to reassess
4 design criteria for GWAS scaffolds and arrays.^{25,26} On the one hand, tag SNPs that tag lower
5 frequency variants are likely to be on the lower end of the site frequency spectrum and,
6 consequentially, more geospatially restricted.²⁷⁻³⁰ On the other hand, as studies grow very large,
7 cohort heterogeneity is likely to increase substantially for both discovery and replication populations.
8^{31,32} Given finite GWAS scaffold density, examining the trade-off between lowering the frequency
9 threshold for accurate imputation and extending utility to multiple populations will become important.
10^{33,34} In this manuscript, we describe a framework for developing well-powered tag SNP selection
11 leveraging thousands of whole genomes from diverse populations for balanced cross-population
12 coverage. In our study, genomic coverage is evaluated based on genome-wide imputation accuracy
13 as measured by mean imputed r^2 at untyped sites, rather than pairwise linkage disequilibrium.
14 Moving beyond pairwise metrics allows us to account for haplotype diversity across the genome and
15 demonstrates population-specific biases from pairwise estimates. Assessing accuracy using leave-
16 one-out cross-validation yields a real-world estimate of genomic coverage. We examine the effect of
17 allele frequency, correlation thresholds, and population diversity on the selection of tag SNP and on
18 the landscape of tag-able variation. This work demonstrates that, while there may be limits given
19 current reference panels, improving GWAS scaffold design is an underused means to increase
20 power in association studies.

21 **Results**

22 **Selecting tag SNPs from a single population results in suboptimal tagging 23 performance across populations impacted by different patterns of 24 demography**

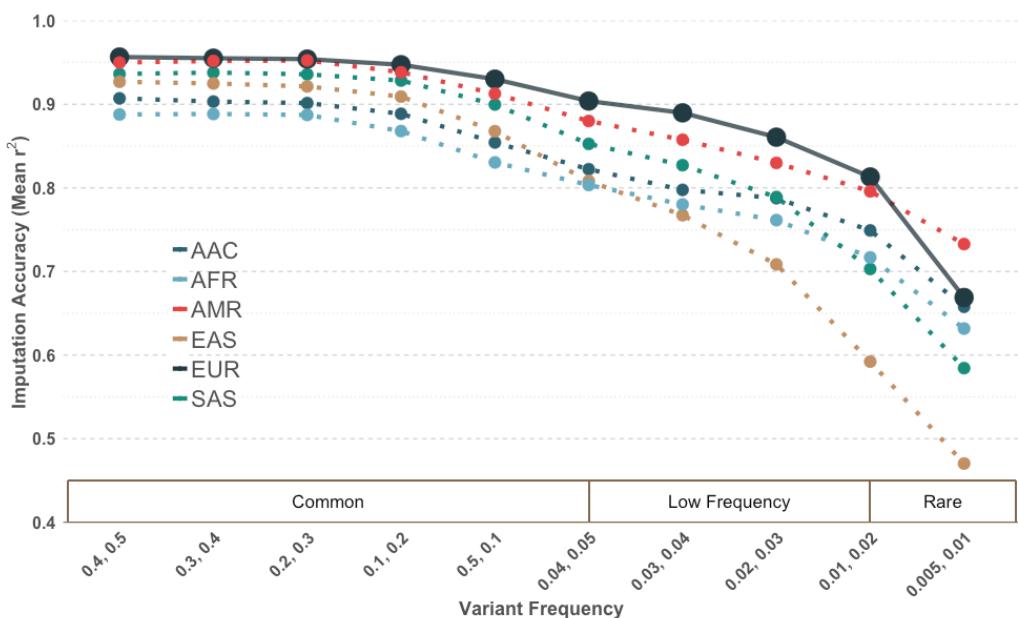
25 First we designed an experiment to assess imputation accuracy performance comparing tag
26 SNP selection from different populations. This experiment mimics the current design of many
27 commercial arrays, in which tag SNPs were selected to capture the primarily variation in a single
28 population or a closely related group of populations. We built a pipeline using the 26 population
29 reference panel from Phase 3 of the 1000 Genomes Project and the Tagit algorithm for tag SNP
30 selection.²³ (**Supplementary Table 1**) Individuals were split into mutually exclusive “super
31 populations.” These included the Admixed American (AMR), East Asian (EAS), European (EUR),
32 and South Asian (SAS) populations as described in Auton et al.¹⁷ In addition, we divided the African
33 super population into two groups: four populations from Africa (AFR) and two populations of African

1 descent in the Americas (AAC) (see **Methods**). Initially, to mimic the design of many arrays, tag
2 SNPs were only selected from a single super population. We assumed a genome-wide allocation of
3 500,000 tag SNPs, however analyses for a single population tagging strategy were only conducted
4 on chromosome 9 with the allocation of 21,107 sites proportional to the physical distance of
5 chromosome 9 compared to all chromosomes combined. Potential tags were required to have a
6 minor allele frequency (MAF) $\geq 1\%$ and be in pairwise LD with the tagged target site with a $r^2 \geq 0.5$.

7 The current generation of phase-based imputation algorithms (BEAGLE, IMPUTE2,
8 Minimac3) leverage local haplotype information and sequenced reference panels to improve
9 accuracy of variant inference compared to tag SNP approaches.^{19,21,35-38} Therefore, optimal array
10 design depends not only on tag SNP selection, but also on empirical evaluation of imputation
11 performance. For each of the population-specific GWAS scaffolds, imputation accuracy was
12 assessed in all six super populations by MAF bins (common, MAF = 0.05-0.5; low frequency, MAF =
13 0.01-0.05; and rare, MAF < 0.01) by comparing the imputed dosages to the real genotypes through
14 leave-one-out internal validation. (see Methods)

15 Consistently across all super populations, the population from which the tags were
16 ascertained had the highest imputation accuracy in the common bin. (**Supplementary Figure 1**)
17 Trends in imputation accuracy follow known patterns of demography. For example, if the tags were
18 ascertained in European populations, imputation accuracy was best in Europeans (EUR), followed
19 by out-of-Africa populations (AMR, SAS, EAS), and worst in African ancestry populations (AFR,
20 AAC). (**Figure 1**) If the tags were ascertained in African populations, the inverse was observed.
21 (**Supplementary Figure 1**) As expected, the same trend of reduced imputation accuracy in non-
22 ascertained populations was exacerbated in the low frequency bin. Imputation of low frequency
23 variants in East Asian populations (EAS) was consistently most challenging; even when tag SNPs
24 were selected from EAS, accuracy of low frequency imputation was the same or better in other
25 populations. This can be explained by evidence of a recent tight bottleneck followed by rapid
26 population growth in EAS, resulting in a large proportion of rare variants that are difficult to tag due
27 to lower LD, especially with a limited scaffold of 500,000 sites.²⁷ In contrast, the imputation
28 performance of tag SNPs ascertained in AFR, AMR, and AAC populations is the same or better
29 compared to the performance in out-of-Africa populations. This is likely due to increased allelic
30 heterogeneity in African ancestry populations, which results in greater haplotypic diversity and a
31 higher chance that a rare variant is well tagged by a haplotype for imputation.¹⁷ The imputation
32 accuracy of AMR higher in the rare frequency bin (MAF 0.5-1%), independent of the ascertainment
33 population, is likely due to longer haplotypes resulting from recent admixture, allowing the rare
34 variation to be captured accurately given the limited allocation.³⁹ Importantly, in each case we
35 observe a notable drop-off in performance across most of the frequency spectrum when examining

1 imputation coverage in populations diverging from the one used for tag SNP selection
2 (Supplementary Figure 1).



3

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

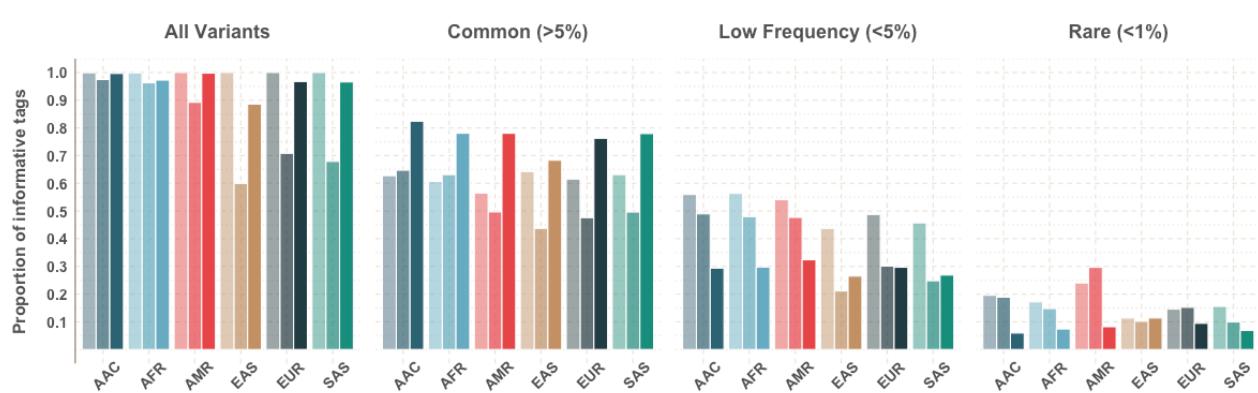
7 Comparing single versus cross population tag SNP selection strategies

8 The first criterion of developing a useful genotyping platform is whether or not the variants
9 assayed segregate in the population of interest and contribute tagging ability by being in LD (high r^2)
10 with untagged sites. For example, using Illumina's OmniExpress platform within the 1000 Genomes
11 Project data, over 99.7% of the sites will be polymorphic ($MAF > 0.5\%$) in the overall dataset.
12 However, when we stratify by super population, each group has a differential loss. While AFR loses
13 <1% of sites for having a $MAF < 0.5\%$, EUR and EAS lose 4.4% and 9.2% of variants, respectively.
14 This will lead to a loss of statistical power dependent on ancestry and could limit analyses. This is
15 quantified as "informativeness", or the ability of a tag SNP to both segregate in the population and
16 provide LD information ($r^2 > 0.5$ with at least one untagged site). When working in multiple
17 populations, it is essential to have balanced representation of variation across all groups.

18 To explore different approaches for GWAS scaffold design we compared three strategies for
19 selecting tag SNPs; single population tag SNP ascertainment, in which all tags are selected from a
20 single population; a 'naïve' approach, in which all populations are combined and tags are selected
21 based on composite statistics derived from this multi-population pool; and a 'cross-population
22 prioritization' approach, in which tags are prioritized if they are both informative in multiple
23 populations and by the number of unique sites targeted across all groups (see **Methods** and
24 **Supplementary Figure 2**). We generated lists of tags per method assuming a total genome-wide

1 allocation of 500,000 sites and minimum thresholds of $r^2 > 0.5$ and minor allele frequency (MAF) \geq
2 1%. Using these parameters, an exhaustive set of tag SNPs were selected using the naïve
3 approach with tags ranked by the absolute number of sites tagged across the 6 super populations,
4 regardless of how many super populations had LD between tags and targets. We then re-ranked
5 them using the cross-population prioritization approach (**Supplementary Figure 2**).

6 To compare the three approaches, we tallied the number of informative tags per population
7 for each method to investigate the added value of tags contributing information in multiple
8 populations. (**Figure 2**) This was done for all 22 autosomes. As per the design, all the single-
9 population tags were informative within the super population from which tag SNPs were selected.
10 Comparing the naïve and cross-population approaches that selected tag SNPs across all
11 populations, the cross-population prioritization approach increased the number of informative tag
12 SNPs in all populations relative to the naïve approach. In the naïve approach, we observed that the
13 majority of tag SNPs were selected from the AFR population, followed by AAC, due to African-
14 descent populations having more polymorphic sites across the genome with lower linkage
15 disequilibrium.^{17,40} Whereas in the cross-population prioritization approach, variation specific to a
16 single population is down-weighted leading to more balanced representation between all 6 super
17 populations. By leveraging cross-population information the largest boost in the proportion of tag
18 SNPs contributing linkage disequilibrium information compared to the naïve approach was observed
19 in non-African descent populations (10.5%, 28.6%, 25.9%, and 28.7% in AMR, EAS, EUR and SAS,
20 respectively). Even the African descent populations (AFR and AAC), which dominate the naïve
21 approach, have a higher proportion of tags in linkage disequilibrium with target sites with the cross-
22 population prioritization approach (a 2.2% and 1.0% boost for AAC and AFR, respectively).



23
24 **Figure 2: Proportion of tags that are informative by population with the three methods.** (Left, lightest)
25 tags selected from only a single population, (Center) tags selected by pooling all populations agnostically, and
26 (Right) tags selected with the cross-population prioritization approach. Tag SNPs were informative if they
27 were in linkage disequilibrium ($r^2 > 0.5$) with at least one untagged site.
28

1 To assess performance across the frequency spectrum we also stratified our accuracy
2 estimates by super population-specific MAF into common, low frequency, and rare bins, as
3 previously described. We observed that the cross-prioritization approach results in a larger
4 proportion of tags being informative compared to both the single-population and naïve for common
5 tag SNPs (MAF>0.05) in all super populations. This is likely because the cross-prioritization
6 approach prioritizes potential tag SNPs that provide LD information across multiple populations,
7 therefore prioritizing common variants tagging common variation. However, by limiting tag SNP
8 selection to these common variants only, the proportion of tags that provide LD information for low
9 frequency variants is decreased compared to the single population approach, which had the highest
10 proportion of informative tag SNPs in low and rare frequency in the target population. For example,
11 when tags were ascertained using only AAC LD information, 19.5% of the 500,000 SNP scaffold
12 were informative for rare variation (MAF<1%) and 62.8% for common variation MAF>5%) within
13 AAC populations. When the cross-population approach was used, ensuring the prioritization of
14 common variation, the proportion of tag SNPs informative for rare variation dropped to 6% while the
15 proportion informative for common variation jumped up to 82.4%. This is consistent with low
16 frequency and rare variants being population-specific, therefore not tagged by cosmopolitan
17 common variation present in multiple populations. A notable exception is that the naïve approach
18 contributes the most LD information for rare variants in the AMR super population. This is consistent
19 with our previous findings showing highest imputation accuracy in the rare variation within AMR,
20 even when the population from which tag SNPs were ascertained was different. The AMR on
21 average exhibit longer haplotype lengths from the recently admixed populations in the Americas.
22^{17,39} Because of the long haplotype tract lengths, more limited haplotypic diversity, and the limited
23 allocation of tag SNPs, a naïve approach emphasizing the absolute number of unique sites up-
24 weights variation that is informative for at least one of the ancestral components present in these
25 populations.

26 **Cross population prioritization of tag SNPs increases imputation accuracy for
27 all groups across frequency spectrum compared to naïve approach**

28 Beyond being polymorphic and providing LD information across global populations, an
29 efficient tag SNP scaffold must also optimize the LD structure beyond individual tag SNPs' pairwise.
30 The goal of tag SNP selection is to inform the unmeasured haplotypes, and therefore their
31 performance must be evaluated as a collaborative unit. One way to assess this is through
32 imputation accuracy. Hence, following the observation that cross-population prioritization selects a
33 higher proportion of informative common tag SNPs for each population, even compared to the
34 single population approach, we next assessed what impact this would have on imputation accuracy.
35 We deployed the same leave-one-out internal cross validation approach as before using the 1000
36 Genomes Project populations (see **Methods**). We assumed a genome-wide scaffold of 500,000

1 sites and tags had to have a $MAF > 1\%$ and $r^2 > 0.5$ with tagged sites. Imputation accuracy was
2 highest across all population-specific minor allele frequency bins when ascertaining in the target
3 population in non-African non-admixed descent continental populations (EAS, EUR, and SAS).
4 **(Supplementary Figure 3)** For the two African descent groups (AAC and AFR), the cross-
5 population prioritization approach had the highest imputation accuracy across all sites. When
6 stratified by MAF bins, the increase in informative tag SNPs for common variants with the cross-
7 population approach yielded higher imputation accuracy for common variation in all super
8 populations. As previously seen, the population-specific nature of low frequency and rare variants
9 led to decreased imputation accuracy in non-African descent populations for both the cross-
10 population and naïve approach when compared to targeted single-population ascertainment. The
11 cross-population prioritization approach had higher imputation accuracy than the naïve approach for
12 all MAF bins.

13 As scaffold size can dramatically affect imputation accuracy⁴¹, we additionally examined
14 allocations of 250,000, 500,000, 1,000,000, 1,500,000, and 2,000,000 genome-wide tags, which
15 were all selected with $r^2 > 0.5$ and $MAF > 0.01$. The cross-population prioritization scheme performed
16 better with higher imputation accuracy than the naïve method for all super populations across all
17 minor allele frequency bins with tags selected. **(Figure 3)** The biggest improvement came with the
18 smaller array sizes. The most marked improvement was found in EAS, which originally had the
19 lowest imputation accuracy of the 6 super populations with the naive approach. Within EAS groups,
20 the cross-population approach increased imputation accuracy overall by 9.8% (from 67.3% to
21 77.1%) for a tag scaffold of 250,000 sites. For a scaffold of 500,000 sites, an overall improve of
22 6.2% was observed (from 77.4% to 83.6%). Improvements were largely consistent with the increase
23 of informative tag SNPs. **(Figure 2)** As with the naive prioritization approach SNPs were
24 disproportionately informative within AFR and AAC, consistent with admixed ancestry reflected by
25 reference panels. For the smaller sizes (250K), the greatest increase in performance incorporating
26 cross-population information was found within common SNPs ($MAF > 5\%$). However, the larger sized
27 scaffolds (1-2 million) showed the most improvement within the low frequency bins ($MAF < 5\%$).

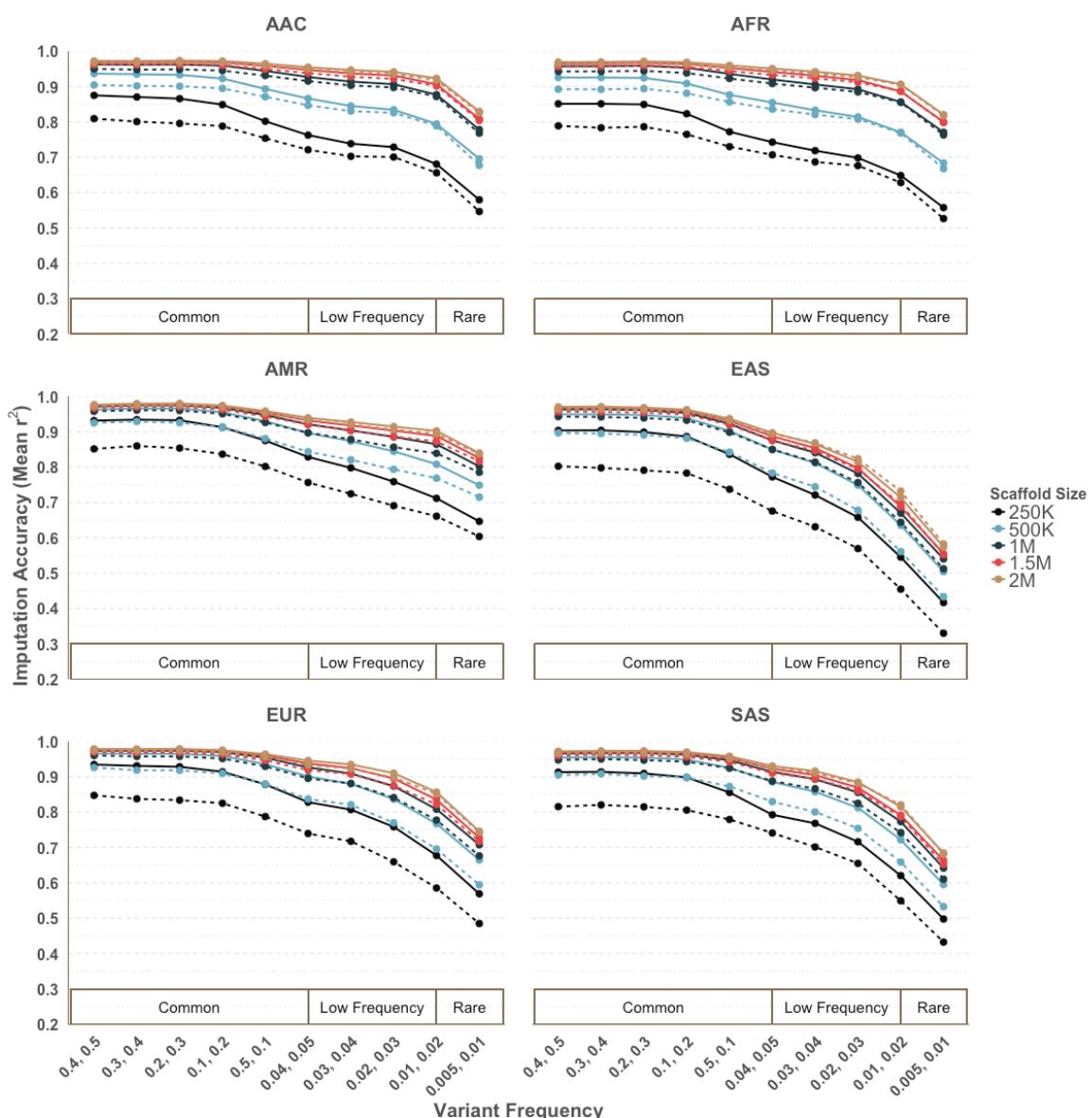


Figure 3: Increased imputation accuracy with cross-population prioritization (solid line) versus naïve approach (dashed line) for a minimum pairwise correlation threshold of $r^2 > 0.5$ and MAF > 1% across different scaffold sizes. Imputation accuracy was calculated separately within minor allele frequency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

Imputation accuracy varies by local ancestry background in admixed individuals

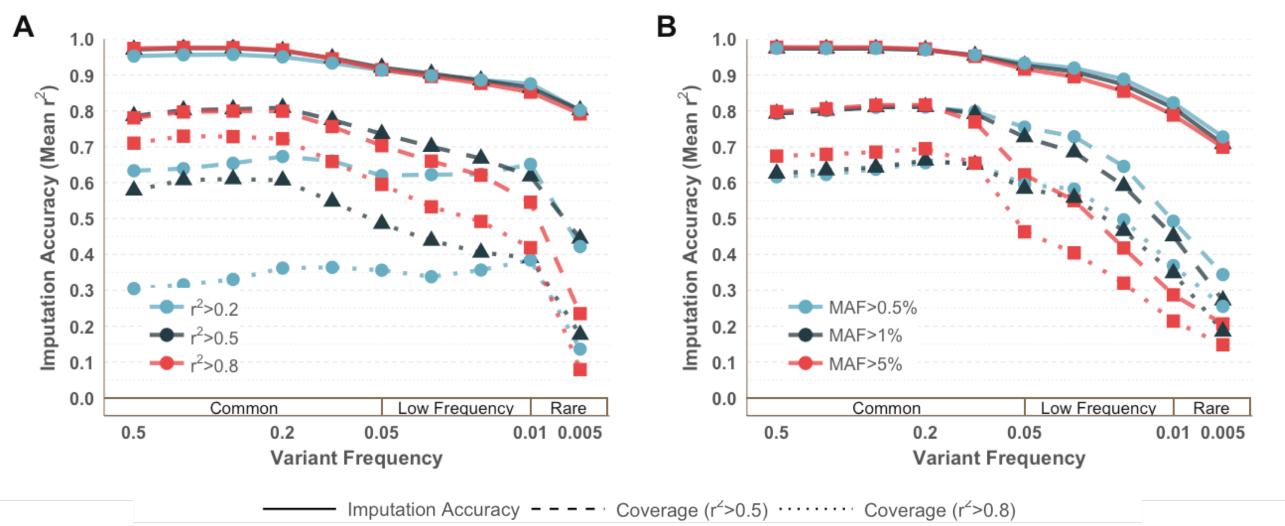
We also assessed imputation ancestry stratified by local ancestry diplotype in the two admixed populations, the AAC and AMR, for a genome-wide allocation of 500,000 tag SNPs. First, using phased data, we inferred haploid tracts of African, European, and Native American local ancestry along the genomes of all individuals in the AMR and AAC populations (see **Methods**, ^{17,42}). 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

1 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
2 accuracy when compared to the naïve approach. When looking at ASW population (Americans of
3 African ancestry in South West US), performance was high overall with all diploid tracts having
4 imputation accuracies of 92.8-96.8% for all sites with minor allele frequency above 1%.
5 **(Supplementary Figure 4)** The lowest imputation accuracy was found in AFR-AFR tracts,
6 especially at the lower end of the frequency spectrum. The highest imputation accuracy was found
7 in EUR-EUR tracts (94% overall for ASW). In AMR populations, by contrast, the NAT-NAT tracts
8 had the lowest performance of all. An example can be seen in the MXL population (Mexican
9 Ancestry from Los Angeles), where the highest imputation accuracy was found in the AFR-EUR
10 tracts (overall imputation accuracy of 90.1% for all SNPs with $MAF > 0.5\%$) and the lowest within
11 NAT-NAT tracts (74.8% for all SNPs with $MAF > 0.5\%$). **(Supplementary Figure 4B)** These
12 performances are reflective of the relative availability of reference data relevant to these specific
13 ancestral components.

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

16 Previous standards in scaffold design have considered minimum linkage disequilibrium (r^2)
17 and minor allele frequency (MAF) thresholds when prioritizing possible tag SNPs. However, the
18 impact of these thresholds are often evaluated through pairwise coverage. We explored varying the
19 minimum r^2 threshold (0.2, 0.5, 0.8) and MAF (0.5%, 1%, 5%) to assess their impacts on imputation
20 accuracy, as well as pairwise coverage, assuming a genome-wide allocation of one million tags. For
21 common variants, a higher minimum r^2 threshold ($r^2 > 0.8$) resulted in slightly higher imputation
22 accuracy. **(Figure 4A)** However, the sites in the low and rare bin demonstrate population-specific
23 accuracy only. **(Supplementary Figure 5)** For AFR, SAS, and EAS, a less stringent threshold of
24 $r^2 > 0.2$ had the worst imputation accuracy across all frequency bins. Low frequency and rare
25 variation had higher imputation accuracy for an r^2 threshold of 0.5 compared to 0.8. Within AAC,
26 AMR, and EUR, the low frequency variation had improved imputation accuracy with the lowest r^2
27 threshold of 0.2. However, the imputation accuracy within this low threshold was notably
28 compromised for common variants. This indicates that low frequency variation is better captured by
29 weak correlation structure, but at a cost to common variation in these populations. Analyses
30 performed with $r^2 > 0.5$ had the best balance of performance across all frequency bins with the
31 highest overall imputation accuracy in all super populations except for EAS. **(Supplementary Table**
32 **2)** Overall, there was very small differences in imputation accuracy between the different r^2
33 thresholds. There were much larger differences in coverage, including both coverage evaluated with
34 minimum r^2 (LD) of 0.5 and 0.8. **(Figure 4A)** Additionally, the best “performance” using pairwise
35 coverage was highly dependent on the definition of coverage. Specifically, if pairwise coverage was

1 calculated as the proportion of sites that are in LD with $r^2 > 0.5$, then the best minimum r^2 threshold in
2 tag SNP selection will be 0.5. This holds true for $r^2 > 0.8$ as well.



3
4 **Figure 4: Influence of (A) minimum r^2 threshold and (B) lower MAF threshold on imputation accuracy**
5 **and coverage($r^2 > 0.5$ and $r^2 > 0.8$) within populations from the Americas with an allocation of 1M sites.**

6
7 The impact of minimum minor allele frequency threshold was negligible across variants with
8 MAF > 5% for all non-African populations (**Supplementary Figure 6**). Within populations of African
9 descent, limiting tags to variants with MAF > 5% resulted in increased imputation accuracy for all
10 frequency bins, especially for common variants. Lowering the MAF to 0.5% reduced accuracy in
11 African-descent populations across all frequency bins. For EUR, SAS, and AMR, tags with MAF > 1%
12 had decreased accuracy for variants with MAF 0.5-1% compared to when tags are limited to
13 MAF > 0.5%. (**Figure 4B**) The lowest limit of MAF (0.5%) showed increased accuracy for rare
14 variation but at a slight cost to the accuracy for common sites (MAF > 5%). We concluded that the
15 best balance for tag SNP selection across all populations among these was MAF > 1% within the
16 population being tagged, as the imputation accuracy was best for MAF > 5% for half of the groups
17 (AAC, AFR, EAS) and best for MAF > 0.5% for the other half (AMR, EUR, SAS). (**Supplementary**
18 **Table 2**) However, the overall differences in imputation accuracy was minimal, with less than 1%
19 between all lower MAF thresholds across all sites. Again, we observed large differences in pairwise
20 coverage, despite negligible differences when performance is evaluated by imputation accuracy.
21 (**Supplementary Figure 6**) This is particularly striking for African-descent populations (ASW and
22 AFR), where there were large gains of pairwise coverage for MAF > 1%, compared to MAF > 0.5%
23 and MAF > 5%. As previously described, African populations have shorter LD blocks and a greater
24 absolute number of polymorphic variants compared to other populations.¹⁷ Therefore, pairwise
25 coverage underestimates performance compared to imputation accuracy, as addressed below.

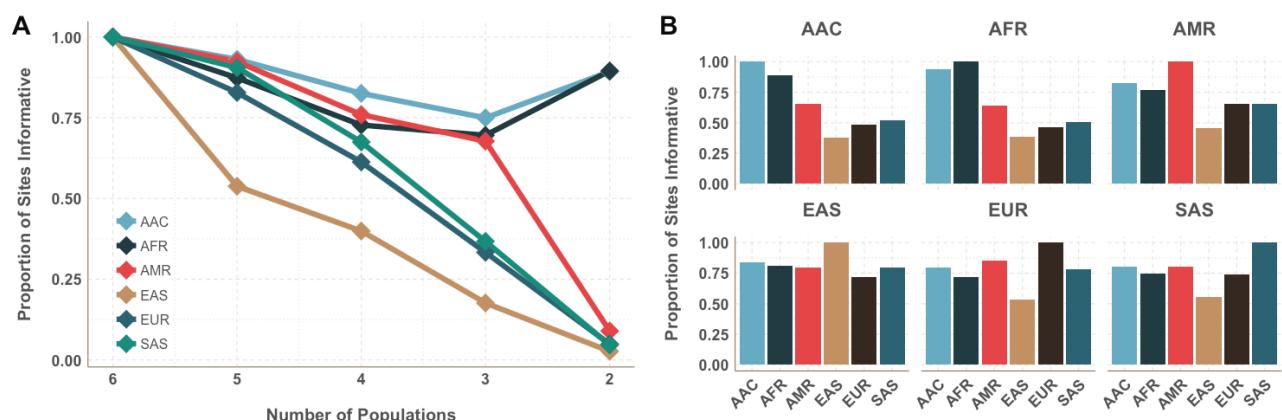
26

27

1 Tagging potential differs between populations

2 Efficient tag SNP selection is an opportunity to boost power in downstream analyses. African
3 and out-of-Africa populations exhibit distinct genetic architecture which resulted in different
4 performance trends. There is therefore a need to balance the representation from all global
5 populations. However, even when cross-population performance was prioritized, it did not
6 guarantee equal representation of all population groups within the tag SNP set. To determine the
7 contribution of each population, we again focused on chromosome 9 (42,215 tags), equivalent to
8 one million sites genome-wide, selected with our novel cross-population prioritization scheme. This
9 tag SNP allocation resulted in including all tags that were informative in at least 3 to all 6
10 populations in the scaffold. Out of all tags for chromosome 9, 17.96% were informative in all 6
11 populations. (**Supplementary Table 3**) No tags were included that were informative in only one or
12 two populations. Of tags that were informative in 5 out of the 6 super-populations, only 54% were in
13 LD with any target sites within EAS populations, while 93% were informative in AAC populations.
14 (**Figure 5A**) This trend is consistent with cross-population tags tending to be less informative in EAS
15 populations compared to the other populations. When tags are informative in 3 out of 6 groups, only
16 18% were informative in EAS, while 75% were informative in AAC. Tags informative in only 2 of the
17 6 groups were likely informative in AAC and AFR, the African descent populations, while very few of
18 them were informative for non-African descent groups, consistent with capturing differential LD
19 patterns in African populations.⁴³ When tags are stratified by MAF (0.5-1%, 1-5%, and >5%), these
20 trends are exaggerated in the rare and very rare MAF bins. (**Supplementary Figure 7**) As
21 expected, the very rare variation (0.5-1% MAF) was highly population-specific with no sites in this
22 frequency bin being informative across all populations, or even 5 out of the 6 populations.²⁷ For rare
23 variation (1-5%), tags were the least informative within EAS, with only 36% of the tags informative in
24 5 out of 6 populations.

25 Conditional performance, or the ability of a tag which is informative in the index population
26 also being informative in an additional population, was also examined and found to be consistent
27 with known population histories. Of tags that are informative within AFR, 94% were informative
28 within AAC, while only 38% were informative within EAS. (**Figure 5B**) However, among tags that
29 were informative within EAS, 81% were informative within African populations. Once again, the
30 stratified analyses show exaggerated trends for the very rare and rare MAF bins. (**Supplementary**
31 **Figure 8**) For the very rare variation (0.5-1%), only a very small percentage (<10%) of tags are
32 informative in other populations (AMR, EAS, EUR, SAS) if they were informative within African-
33 descent populations (AFR and AAC). The high level of sharing between AFR and AAC is expected
34 due to the high proportion of African ancestry within African-American and Afro-Caribbean
35 populations. Of tags informative within EUR, 78% are also informative within AMR, largely due to
36 the high proportion of European ancestry within some Hispanic/Latino populations.^{39,44,45}



1
2 **Figure 5: Tag SNPs informativeness across population.** (A) Proportion of sites informative ($r^2 > 0.5$,
3 MAF > 0.01, 1M site scaffold) across a number of populations, with lines corresponding to the index population.
4 (B) Proportion of sites shared across populations, conditional on index population.

5
6 The tags were also not equally informative in each population when it comes to the number
7 of sites they tag with $r^2 > 0.5$. For chromosome 9, it would take 81,416 tags to capture all possible
8 tag-able variation with an $r^2 > 0.5$ within AFR populations, while it would take only 28,473 tags within
9 EAS populations to saturate coverage. However, each tag within the AFR populations captures on
10 average 7.17 other sites, whereas for EAS populations, each tag captures on average 10.27 other
11 SNPs. When restricting the design to a million tag SNP scaffold, each tag captures on average
12 16.16 other SNPs within EAS populations and 12.16 other SNPs in AFR populations. (Table 1) This
13 reflects the different underlying genetic architecture of these different groups.

Population	All Possible Tags		One Million Tag Scaffold	
	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

14
15 **Table 1: Performance per tag SNP to capture all variation possible with $r^2 > 0.8$ on chromosome 9, as**

16 well as within a million site genome-wide scaffold allocation through cross-population prioritization.

17 Limits of tagging and imputation

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

1 AMR. Although these sites have no notable pairwise correlation with any other site in the human
2 genome, haplotypes may be informative and allow the recovery of information for imputation. We
3 again assumed a one million genome-wide tag SNP scaffold allocation with minimum MAF of 1%
4 and minimum r^2 threshold of 0.5 and imputed to the entire 1000 Genomes reference panel. As
5 expected, imputation accuracy and ability to recover information was population-specific. The
6 imputation accuracy within AAC was an outlier when compared to other populations, with 80.72% of
7 lone sites being imputed with at least the accuracy of $r_{acc}^2 \geq 0.5$ and over 50% of sites being imputed
8 with even higher accuracy ($r_{acc}^2 \geq 0.8$). Many of these lone sites within AAC were captured with
9 pairwise and haplotype LD within other populations, primarily AFR and to a lesser extent EUR.
10 While there were likely insufficient allele counts for accurate correlation estimation within AAC due
11 to the small sample size, this information could be recovered using a global reference panel. The
12 number of unrecoverable “dark sites”, which had no pairwise correlation and were not recoverable
13 with imputation using haplotype information, was the largest in EAS and is consistent with known
14 demography and population history yielding an excess of highly rare variation compared to other
15 populations.²⁷

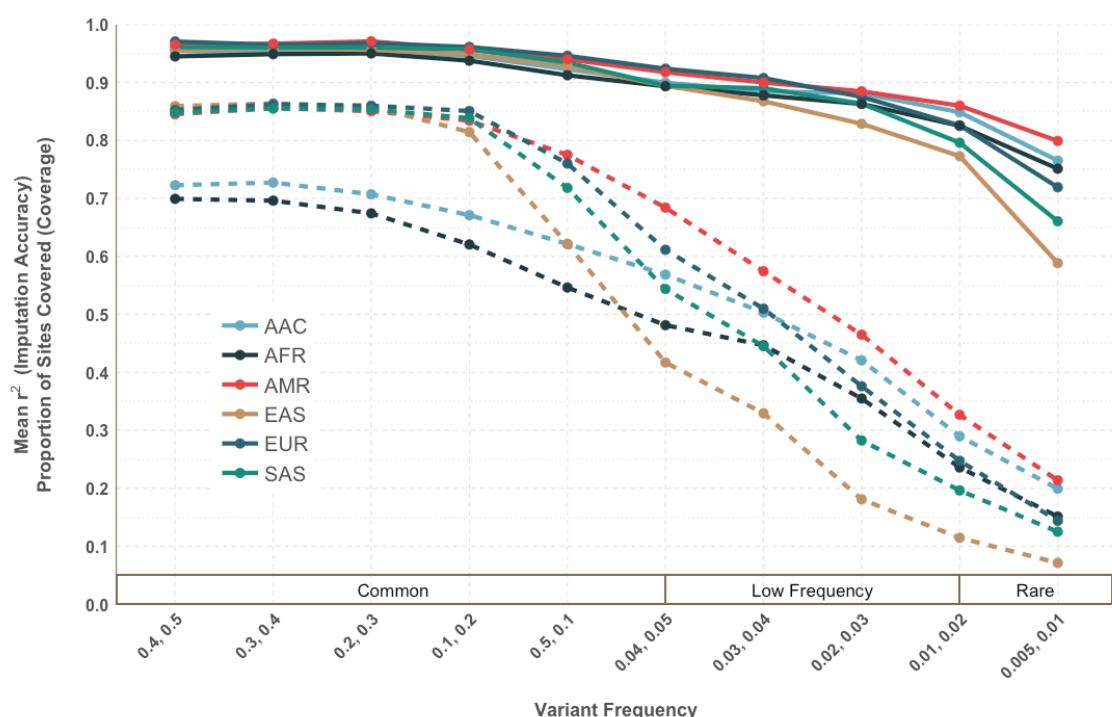
Population	Number of Individuals	Number of Lone Sites	Imputation Accuracy Quality			Number Unrecoverable with $r_{acc}^2 \geq 0.2$ (%)
			$r_{acc}^2 \geq 0.2$	$r_{acc}^2 \geq 0.5$	$r_{acc}^2 \geq 0.8$	
AAC	156	7,509	90.79%	80.72%	51.72%	691 (9.2%)
AFR	495	4,497	63.29%	38.73%	7.03%	1,651 (36.7%)
AMR	341	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%)

16 **Table 2: Lone sites by super population and their imputation accuracy for a 1M site scaffold.**
17

18 Pairwise coverage versus imputation accuracy

19 When evaluating the performance of a GWAS scaffold, there are numerous factors to take
20 into consideration. These include the number of sites you have allocated to tag SNPs and what your
21 priorities are for balanced representation. To a lesser extent, the benefits and pitfalls of prioritizing
22 low-frequency variants must be weighed. However, we have demonstrated that the influence of
23 these components are highly dependent on how performance is measured. The notion of genomic
24 “coverage” has historically been estimated using pairwise correlations, and therefore this term will
25 be used to denote the proportion of polymorphic sites that are in pairwise LD (r^2 threshold) with at
26 least one tag SNP. We calculated coverage separately per super population at an r^2 threshold of 0.5
27 and 0.8 within minor allele frequency bins identical to the imputation accuracy estimation analyses,
28 assuming a genome-wide tag SNP set of 500,000 and 1,000,000. (**Table 3**) For a tag SNP set of
29 one million sites, coverage was lowest in AFR with an overall average of 59.15% for all sites with
30 MAF>0.5% and $r^2>0.5$. (**Supplementary Figure 9**) When the r^2 threshold is raised to 0.8, the

1 proportion of sites in linkage disequilibrium with at least one tag SNP lowers to 28%. (**Figure 6**) The
 2 highest coverage was found in populations from the Americas (AMR) and East Asia (EAS). For a
 3 lower r^2 threshold of 0.5, 79.9% of AMR sites with MAF>0.5% were covered. When using the higher
 4 r^2 threshold of 0.8, East Asian populations had the highest coverage with 63.08% of sites in LD with
 5 at least one tag SNP. This difference is even more marked when looking at a smaller tag SNP set of
 6 500,000 sites. (**Supplementary Figure 10-11**) African populations now have an overall coverage of
 7 33.17% with $r^2>0.5$ and 14.10% with $r^2>0.8$. East Asian populations have the highest coverage with
 8 73.16% of sites covered with $r^2>0.5$ and 55.09% with $r^2>0.8$.



9
 10 **Figure 6: Coverage versus Imputation Accuracy, assuming a genome-wide scaffold size of one million**
 11 **tags.** Coverage is shown with an $r^2>0.8$. While pairwise tagging values are low, particularly in African-descent
 12 populations, multi-marker imputation accuracy remains high across groups.
 13

Super population	Total Number of Polymorphic Sites	Scaffold of 1,000,000 tags		Imputation Accuracy	Scaffold of 500,000 tags		Imputation Accuracy		
		Coverage			Coverage				
		$r^2>0.5$	$r^2>0.8$		$r^2>0.5$	$r^2>0.8$			
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%		

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

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

1 consistent with other populations (imputation accuracy of 89.62%) among SNPs with a minor allele
2 frequency above 0.5%. This contrast is also found in East Asian populations, which had one of the
3 highest proportion of polymorphic SNPs with $r^2 > 0.5$ for coverage (76.95%), but the lowest
4 imputation accuracy (86.28%). (**Table 3**) When sites are stratified by minor allele frequency bins,
5 the differences in trends are even more striking. (**Figure 6, Supplemental Figure 9**) For example,
6 within the lowest frequency bin (0.5% to 1%) for admixed populations of African-descent, the
7 coverage of sites for a set of 500,000 tag SNPs with $r^2 > 0.8$ falls below 10%, however the imputation
8 accuracy remains relatively high at 77.82%. These trends are consistent and more dramatic when
9 evaluated within a tag SNP set of 500,000 sites. (**Supplemental Figures 10-11**) These
10 observations reinforce the necessity of examining imputation accuracy, instead of pairwise
11 coverage, when evaluating the performance of tag SNPs.

12 Discussion

13 As larger and larger whole genome reference panels come into availability for imputation, it
14 is important to design arrays with this ultimate goal in mind. There are currently two accepted
15 methods of evaluating the performance of a tag SNPs: pairwise LD “coverage” and imputation
16 accuracy. Coverage has historically been used as a term to denote the proportion of polymorphic
17 sites that are in linkage disequilibrium with at least one tag marker above a certain r^2 threshold.⁴⁶⁻⁴⁹
18 Genotyping arrays are typically compared using this score averaged across the genome. However,
19 as others and we have demonstrated, restricting performance assessment to this definition of
20 pairwise coverage is limited by removing multimarker information.^{33,34} Evaluating imputation
21 accuracy, particularly via leave-one-out cross validation, is highly computationally intensive, but
22 provides a better assessment of how well untyped variation can be recaptured and a more realistic
23 depiction of array performance. Imputation accuracy is also a more useful statistic in a practical
24 sense, especially with the development of deeper and more diverse reference panels,^{17,18,50-52} as
25 performing GWAS with imputed variants is now the expectation. Emerging evidence suggests that
26 very rare variants that are poorly tagged by an individual tag SNP will be accessible via imputation,
27 due to added haplotype information, particularly as sample sizes move beyond the thousands into
28 the tens of thousands.^{19,34}

29 In addition, previous tagging strategies have predominantly focused on optimizing
30 performance in a single population. In prioritizing potential tags by their ability to provide linkage
31 disequilibrium information across multiple populations, we were able to demonstrate that cross
32 population tag SNP selection outperforms single population selection. This boost in imputation
33 accuracy exists across all populations and frequency bins. We simulated tag SNP sets for a range
34 of sizes (250,000-2 million), as well as for several minimum minor allele frequencies (0.5%, 1%, 5%)

1 and minimum r^2 thresholds (0.2, 0.5, 0.8). For investigators with limited real estate or budget for tag
2 SNP selection, we found that the biggest improvement in imputation accuracy provided with our
3 cross population approach was with the smaller array sizes (250K) when compared to a naïve
4 design or biased population ascertainment. As expected, the influence of MAF and r^2 threshold was
5 population-specific. For African-descent populations, including tag SNPs with a low threshold of $r^2 \geq$
6 0.2 resulted in lower imputation accuracy across all bins, while in other populations (EUR, AMR,
7 SAS) tags at $r^2 \geq 0.2$ led to increased imputation accuracy for low frequency variants to the
8 detriment of common variation. This is due to the lower LD patterns overall in African haplotypes,
9 requiring denser coverage. The best balance was found with a moderate r^2 threshold of ≥ 0.5 for
10 those seeking to perform well across all populations. This compromise is also present in choosing
11 the lower MAF threshold. Limiting tag SNP selection to common variants with $MAF \geq 5\%$ produced
12 the highest imputation accuracy across all frequency bins within African-descent populations.
13 However, this threshold decreased imputation accuracy for low frequency and rare variants in all
14 other populations. Therefore, the best balance is once again found in the moderate value of $MAF \geq$
15 1%. Investigators will need to take their priorities into account when selecting the correct thresholds
16 for their populations and if they have a specific target frequency bin. We chose to prioritize all
17 populations equally to provide a design of broad global utility, but if the study is comprised of mostly
18 one ancestral group then the investigators should choose the appropriate thresholds tailored for
19 their study.

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

1 SNP selection and imputation accuracy, and we expect imputation accuracy to improve for all
2 populations and across the frequency spectrum.¹⁹

3 As cosmopolitan biobanks and large-scale multi-ethnic epidemiological studies become
4 more commonplace, the available technology to capture genetic variation must keep pace. It is
5 important to rely on a platform that is as equitable as possible in providing information about the
6 groups of interest when conducting genetic association studies within diverse populations. There
7 are various considerations that an investigator must consider when selecting tag SNPs to customize
8 or build an array, or evaluate an available commercial array. Assessing performance through
9 imputation accuracy, as performed here, is the most apt comparison between tag SNP sets,
10 allowing a real-world look at the extent of variation a set of tags can capture based on haplotype
11 structure. We have presented an improved tagging algorithm and evaluation pipeline that prioritizes
12 cross-population performance for increased imputation accuracy across multiple populations and
13 the full range of MAF $\geq 0.5\%$. We also provide recommendations and context for other researchers
14 interested in similar goals.

15 The power to identify relevant loci is inherently constrained by sample size and genome
16 coverage. Imputation improves this by providing increased effective coverage across the genome. It
17 is important to note that algorithmic development both on association testing and imputation
18 methods has been a productive avenue of research since GWAS began, with new methods
19 providing incremental improvements in statistical power. Here, we demonstrate with a fixed
20 allocation, methods to improve statistical power by tailored SNP selection in the initial array, with
21 sometimes dramatic improvements in imputation accuracy. With the expansion and improvement of
22 global reference panels, genotyping arrays will be able to capture an increased amount of variation,
23 especially when cross-population performance is prioritized. The unified framework presented will
24 enable investigators to make informed decisions in the development and selection of genotyping
25 arrays for future large-scale multi-ethnic epidemiological studies. This increased representation of
26 multi-ethnic genetic variation will promote the investigation of the genetics of complex disease and
27 the improvement of global health in the next phase of GWAS.

28 Competing Interests

29 CDB is an SAB member of Liberty Biosecurity, Personalis, Inc., 23andMe Roots into the Future,
30 Ancestry.com, IdentifyGenomics, LLC, Etalon, Inc., and is a founder of CDB Consulting, LTD. CRG
31 owns stock in 23andMe, Inc. All other authors declare that they have no competing interests.

1 Author Contributions

2 GLW, CF, RKW, CC, GA, HMK, MB, CDB, CRG, and EEK conceived of and designed the
3 experiments. GLW performed the data analysis. CR, DT, RW, ARM, SS, and HMK contributed
4 analysis tools/materials. GLW wrote the manuscript with comments from CF, ARM, LH, CC, GA,
5 CRG, and EEK. All authors read and approved the final manuscript.

6 Acknowledgments

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

12 Materials and Methods

13 **Genetic Data:** The genetic data are from the 1000 Genomes Project (1000 Genomes) Phase 3 data
14 release, version 2 (7/8/2014) containing whole genome sequences for 2,535 individuals from 26
15 global populations.⁵³ Sequence data were in VCFv4.1 format, mapped to the forward strand and
16 variants annotated as reference or alternate alleles. Only biallelic SNPs were included in this
17 analysis (77,224,748 SNPs total). A list of known cryptically related individuals was obtained from
18 the 1000 Genomes FTP site, and one individual from each related pair were subsequently removed
19 (n=62). Individuals were assigned to their super populations according to the original 1000
20 Genomes assignments (EAS=East Asian, EUR=European, AFR=African, SAS=South Asian,
21 AMR=Americas, comprising 503, 501, 495, 477, and 341 individuals, respectively). Two populations
22 of admixed African ancestry (ASW and ACB) were removed from the African super population and
23 formed a separate African American/Caribbean (AAC) super population (n=156).

24 **Tag SNP Selection:** Allele frequency was estimated within super population for each SNP using
25 Plink v1.9.⁵⁴ Linkage Disequilibrium (LD) was also calculated within each super population using
26 Plink v1.9 and settings for pairwise linkage with a minimum r^2 of 0.2 within a maximum distance of 1
27 megabase (mb). Tag SNP selection was performed per chromosome in the program TagIT²³, with
28 frequency and LD files for each super population as input. The TagIT algorithm analyzed each
29 super population separately. After filtering based on the minor allele frequency (set as either 0.5%,
30 1% or 5%), TagIT annotates the tag SNP that has the highest number of LD pairs with r^2 above a
31 minimum threshold (set as either 0.2, 0.5, or 0.8). The selected tag SNP and all of its linked SNPs
32 are masked and TagIT finds the next tag SNP with the highest number of LD pairs. The output for

1 each super population included for each index tag SNP the number of sites in LD, as well as the
2 number of unique sites that weren't already tagged by a previously chosen tag. The number of
3 unique SNPs tagged across all populations per tag SNP was tallied in the final output.

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

15 **Metric of Performance:** Coverage and imputation accuracy were assessed using all polymorphic
16 biallelic sites within the 1000 Genomes Phase 3 data release, version 2. Sites were categorized into
17 ten discrete minor allele frequency bins: (0.005-0.01], (0.01-0.02], (0.03-0.04], (0.04-0.05], (0.05-
18 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
19 proportion of untyped sites that had at least one tag SNP with pairwise r^2 greater than a certain
20 threshold (0.2, 0.5, or 0.8). Imputation accuracy was determined through a leave-one-out internal
21 validation approach with the 1000 Genomes Project Phase 3 data using a modified version of
22 Minimac.¹⁹ Correlation was calculated comparing the estimated dosages to the true genotypes from
23 the original vcf files.

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

29 **Local Ancestry:** Local ancestry was estimated using RFMix ⁵⁵ assuming three ancestral
30 backgrounds: African, European, and Native American, and is described in detail in ⁴². Tracts were
31 dropped if smaller than 20 cM to improve accuracy in local ancestry estimation. Diploid ancestry
32 with three ancestral backgrounds yielded six categories of variation. Imputation accuracy was then
33 calculated separately per diploid tract category, with all other sections masked out. Results were
34 aggregated across all chromosomes to calculate the genome-wide performance per diploid
35 ancestry. Tracts were removed from analysis if the ancestral diplotype was found in fewer than 5

1 individuals. This included AFR-NAT and EUR-NAT within ACB which only occurred in 2 individuals
2 each, NAT-NAT diplotypes in ASW which occurred in one individual, and AFR-AFR diplotypes in
3 MXL which occurred in 3 individuals.

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

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

21 **Bibliography**

- 22 1. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes
23 sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
- 24 2. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
25 genetic variants. *Nat Genet* **46**, 310–315 (2014).
- 26 3. Emond, M. J. *et al.* Exome sequencing of extreme phenotypes identifies DCTN4 as a
27 modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat Genet* **44**, 886–
28 889 (2012).
- 29 4. Lohmueller, K. E. *et al.* Whole-exome sequencing of 2,000 Danish individuals and the role of
30 rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.* **93**, 1072–1086 (2013).
- 31 5. SIGMA Type 2 Diabetes Consortium *et al.* Association of a low-frequency variant in HNF1A
32 with type 2 diabetes in a Latino population. *JAMA* **311**, 2305–2314 (2014).
- 33 6. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease.
34 *Nature* **526**, 82–90 (2015).
- 35 7. Lindquist, K. J., Jorgenson, E., Hoffmann, T. J. & Witte, J. S. The impact of improved
36 microarray coverage and larger sample sizes on future genome-wide association studies.
37 *Genet. Epidemiol.* **37**, 383–392 (2013).
- 38 8. Kosmicki, J. A., Churchhouse, C. L., Rivas, M. A. & Neale, B. M. Discovery of rare variants
39 for complex phenotypes. *Hum Genet* **135**, 625–634 (2016).

- 1 9. Hoffmann, T. J. *et al.* Design and coverage of high throughput genotyping arrays optimized
- 2 for individuals of East Asian, African American, and Latino race/ethnicity using imputation
- 3 and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422–430 (2011).
- 4 10. Hoffmann, T. J. *et al.* Next generation genome-wide association tool: Design and coverage
- 5 of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89 (2011).
- 6 11. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796
- 7 (2003).
- 8 12. McCarthy, N. S. *et al.* Exome array analysis suggests an increased variant burden in families
- 9 with schizophrenia. *Schizophr. Res.* (2016). doi:10.1016/j.schres.2016.12.007
- 10 13. Wessel, J. *et al.* Low-frequency and rare exome chip variants associate with fasting glucose
- 11 and type 2 diabetes susceptibility. *Nature Communications* **6**, 5897 (2015).
- 12 14. Igartua, C. *et al.* Ethnic-specific associations of rare and low-frequency DNA sequence
- 13 variants with asthma. *Nature Communications* **6**, 5965 (2015).
- 14 15. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K
- 15 haplotype reference panel. *Nature Communications* **6**, 8111 (2015).
- 16 16. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals
- 17 identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373–380 (2015).
- 18 17. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- 19 18. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*
- 20 *Genet* **48**, 1279–1283 (2016).
- 21 19. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation.
- 22 *Bioinformatics* (2014).
- 23 20. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate
- 24 genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*
- 25 **44**, 955–959 (2012).
- 26 21. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples.
- 27 *Am. J. Hum. Genet.* **98**, 116–126 (2016).
- 28 22. de Bakker, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nature*
- 29 *Genetics* **37**, 1217–1223 (2005).
- 30 23. Weale, M. E. *et al.* Selection and evaluation of tagging SNPs in the neuronal-sodium-channel
- 31 gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73**,
- 32 551–565 (2003).
- 33 24. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide
- 34 polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*
- 35 **74**, 106–120 (2004).
- 36 25. Carlson, C. S. *et al.* Generalization and dilution of association results from European GWAS
- 37 in populations of non-European ancestry: the PAGE study. *PLoS Biol* **11**, e1001661 (2013).
- 38 26. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
- 39 27. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations.
- 40 *PNAS* **108**, 11983–11988 (2011).
- 41 28. Bustamante, C. D., Burchard, E. G. & La Vega, De, F. M. Genomics for the world. *Nature*
- 42 **475**, 163–165 (2011).
- 43 29. Nelson, M. R. *et al.* The Population Reference Sample, POPRES: A Resource for Population,
- 44 Disease, and Pharmacological Genetics Research. *The American Journal of Human*
- 45 *Genetics* **83**, 347–358 (2008).
- 46 30. Mathieson, I. & McVean, G. Demography and the Age of Rare Variants. *PLoS Genetics* **10**,
- 47 e1004528 (2014).
- 48 31. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in
- 49 the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*
- 50 **200**, 1285–1295 (2015).
- 51 32. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature*
- 52 (2017). doi:10.1038/nature21039

1 33. Martin, A. R., Tse, G., Bustamante, C. D. & Kenny, E. E. Imputation-based assessment of
2 next generation rare exome variant arrays. *Pac Symp Biocomput* **3**, 241–252 (2014).

3 34. Nelson, S. C. *et al.* Imputation-based genomic coverage assessments of current human
4 genotyping arrays. *G3 (Bethesda)* **3**, 1795–1807 (2013).

5 35. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data
6 inference for whole-genome association studies by use of localized haplotype clustering.
7 *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

8 36. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for
9 genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913
10 (2007).

11 37. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature
12 Reviews Genetics* **11**, 499–511 (2010).

13 38. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation
14 method for the next generation of genome-wide association studies. *PLoS Genetics* **5**,
15 e1000529 (2009).

16 39. Gravel, S. *et al.* Reconstructing Native American Migrations from Whole-Genome and
17 Whole-Exome Data. *PLoS Genet* **9**, e1004023 (2013).

18 40. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the
19 mutation load in human genomes. *Nature Reviews Genetics* **16**, 1–11 (2015).

20 41. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing Genome-Wide Association
21 Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS
22 Genetics* **5**, e1000477 (2009).

23 42. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across
24 diverse populations. (2016). doi:10.1101/070797

25 43. Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for
26 modern humans. *PNAS* **108**, 5154–5162 (2011).

27 44. Moreno-Estrada, A. *et al.* Reconstructing the Population Genetic History of the Caribbean.
28 *PLoS Genet* **9**, e1003925 (2013).

29 45. Moreno-Estrada, A. *et al.* The genetics of Mexico recapitulates Native American
30 substructure and affects biomedical traits. *Science* **344**, 1280–1285 (2014).

31 46. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies.
32 *Nature Genetics* **38**, 659–662 (2006).

33 47. Bhangale, T. R., Rieder, M. J. & Nickerson, D. A. Estimating coverage and power for genetic
34 association studies using near-complete variation data. *Nature Genetics* **40**, 841–843 (2008).

35 48. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using
36 fixed marker sets. *Nature Genetics* **38**, 663–667 (2006).

37 49. Li, M., Li, C. & Guan, W. Evaluation of coverage variation of SNP chips for genome-wide
38 association studies. *European Journal of Human Genetics* **16**, 635–643 (2008).

39 50. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide
40 range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).

41 51. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains.
42 *Nature* **505**, 43–49 (2014).

43 52. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa.
44 *Nature* **517**, 327–332 (2015).

45 53. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
46 *Nature* **526**, 68–74 (2015).

47 54. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
48 datasets. *GigaSci* **4**, 7–16 (2015).

49 55. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling
50 approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288
51 (2013).

52