MetaGLIMPSE: Meta Imputation of Low Coverage Sequencing Data for Modern and Ancient Genomes

Kiran H. Kumar[1], Simone Rubinacci[2], Sebastian Zöllner[1]*
[1] *Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA*
[2] *Institute of Molecular Medicine Finland, University of Helsinki, Helsinki, Finland*
*Corresponding author: kiranhk@umich.edu

Abstract

The advent of efficient and accurate imputation for low coverage sequencing offers an unbiased alternative to SNP array imputation, increasing the accuracy of rare variant imputation across all populations. Since imputation accuracy generally increases with larger reference panel size and closer ancestry match between target and reference samples, leveraging imputation from multiple reference panels would facilitate better imputation accuracy; however, individual reference panel genotypes are often privacy protected. We present a novel meta-imputation method, MetaGLIMPSE, that combines estimates from multiple reference panels for low coverage sequencing imputation. We directly combine single panel imputation results using weights estimated for each individual and imputed marker. Across all our scenarios, MetaGLIMPSE outperforms the best single panel imputation for coverages 0.5x - 8x and across all minor allele frequencies, and for certain coverages and scenarios, performs equally to the combined panel imputation.

Introduction

Genomic imputation is a cost-effective tool that powers downstream analysis such as GWAS, fine mapping, and meta-analyses of cohort studies (Browning and Browning 2007; Das et al. 2018; Howie et al. 2012; Fuchsberger et al. 2015; Li et al. 2010; Marcini et al. 2010; Yu et al. 2022). Earlier papers developed genomic imputation algorithms for SNP array data (Browning and Browning et al. 2007; Fuchsberger et al. 2015; Howie et al. 2012). Nowadays, low coverage sequencing is a growing alternative to SNP arrays due to decreasing sequencing costs (Boltz et al. 2024; Martin et al. 2021). Low coverage sequencing utilizes shotgun sequencing at low mean depths between 0.1x and 8x (Boltz et al. 2024; Martin et al. 2021; Rubinacci et al. 2021). While such read data is often insufficient to confidently call genotypes, imputation accuracy for low pass sequencing at coverages as low as 0.5x is comparable to commonly used SNP arrays, and a coverage of 1x suffices for confident rare variant detection. (Martin et al. 2021; Rubinacci et al. 2021). Moreover, low coverage sequencing is especially useful for non-European populations as it provides unbiased coverage across the genome and eliminates ascertainment bias, a drawback of SNP arrays (Boltz et al. 2024; Martin et al. 2021).

The challenge of low coverage imputation is that individual genotypes are more uncertain, which precludes pre-phasing, a computation time saving strategy used in the SNP array imputation (Howie et al. 2012). Thus, an algorithm for low coverage imputation must handle both phasing and imputation with computational efficiency. A commonly used method for low coverage imputation (Allentoft et al. 2024; Boltz et al. 2024; Erven et al. 2024; Nakamura et al. 2024; Ringbauer et al. 2024; Santos et al. 2024; Sousa da Mota et al. 2023) is GLIMPSE (Rubinacci et al. 2021) and its companion GLIMPSE2 (Rubinacci et al. 2022). GLIMPSE and GLIMPSE2 iteratively phase and impute using a Gibbs Sampler, resulting in phased genotypes.

Aside from algorithmic considerations, the main determinants of imputation accuracy are ancestry match with target samples and reference panel size. The latter has a large benefit, in particular, on the imputation accuracy of rare variants, which are associated with substantial functional implications (Das et al. 2018). For this reason, the size of reference panels has grown over time with TopMed and Human Reference Consortium (HRC) representing 133,597, and 32,470 (Das et al. 2016; McCarthy et al. 2016; Taliun et al. 2021). However, for underrepresented populations imputing with an ancestry specific panel may result in more accurate genotypes than imputing with a large, primarily European ancestry reference panel (Carlson et al. 2023; Das et al. 2018; Jewett et al. 2012; Yu et al. 2022). The optimal method leverages genotypes from multiple reference panels. The most straightforward solution, typically, is to combine the smaller panels into a ‘mega’ panel and then impute once with this mega panel. However, creating such a panel is often not possible because the individual level data is privacy protected. Furthermore, merging two reference panels, which are genotyped and quality controlled using different algorithms and parameters, is not trivial, as seen in the creation of the HGDP+1KG panel (Koenig et al. 2023). Finally, combining a small population-specific reference panel with a large nonspecific panel may cause rare population-specific haplotypes to be ignored in favor of more frequent haplotypes, reducing imputation accuracy. To address this challenge for imputation from individual-level genotyping data, Yu et al. 2022 proposed a “meta-imputation” method, MetaMinimac2, that combines imputation results generated from chip-genotypes with multiple reference panels. As they assume the input data is genotypes, their approach models no uncertainty in genotypes and requires pre-phased data. Thus, MetaMinimac2 cannot be applied to low coverage sequencing data.

In this paper, we introduce a novel method for meta-imputation for low coverage sequencing data, called MetaGLIMPSE, which enables imputation of low coverage sequencing data with multiple reference panels. Through meta imputation, we bypass accessing individual reference panel samples and instead directly combine the results of single panel imputations.

We evaluate our method with down-sampled high coverage sequencing data from 1000 Genomes. First, we evaluate the case in which the target samples and reference panels originate from the same ancestry, modeling combining imputation results from two reference panels that have the same ancestry as the target sample. Second, we evaluate the case in which the target samples are an admixture of the ancestries in each reference panel, modeling imputing admixed individuals with two reference panels, each containing a different source population. Third, we evaluate the case where the target samples are from a different ancestry than either reference panels, modeling when there are no available reference panels that contain the target samples' ancestry. In all these cases, our method outperforms the best-available single reference panel imputation for mean coverages 0.5x to 8x, especially at rare variants. We observe, for extremely low mean coverages, e.g. 0.1x, contrary to expectation, that larger reference panels do not necessarily guarantee higher imputation accuracy, and that, at coverages 0.5x to 8x, MetaGLIMPSE consistently outperforms the best single panel imputation across a wide variety of scenarios for all MAFs. In summary, MetaGLIMPSE delivers imputed genotypes that are more accurate than single panel imputation at standard mean coverages (0.5x - 8x) without knowing, *a priori,* which reference panel is the best.

Subjects and Methods

MetaGLIMPSE combines multiple low-coverage sequencing imputation results by calculating each imputation result's fit with the sequencing data. The imputation results, along with the genotype likelihoods, are inputs to a hidden Markov model (HMM) that estimates weights for each panel at every imputed polymorphism. We leverage these weights to combine the single-panel imputation estimates into meta-imputed genotypes.

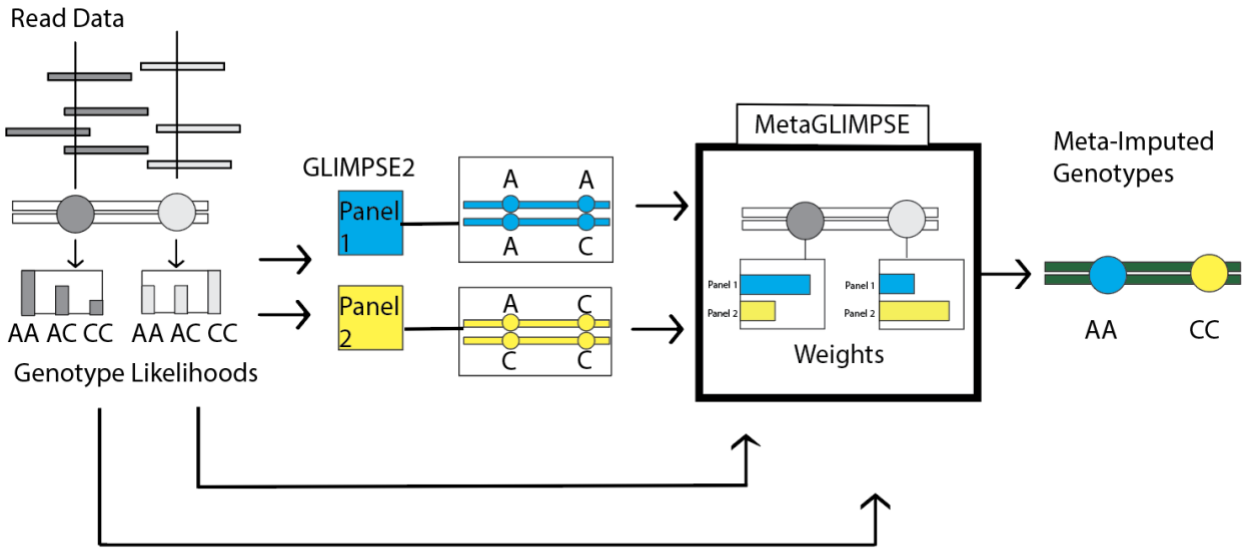


Figure 1 Model Overview of MetaGLIMPSE. The workflow for MetaGLIMPSE is shown here for a single target sample with two markers in dark and light grey, respectively. The workflow starts from the read data, from which genotypes likelihoods are called. Here, the most likely genotypes are A/A, C/C, for the first and second markers, respectively. Then, we phase and impute haplotypes based on the genotype likelihoods with two different reference panels, 1 in blue, and 2, in yellow via GLIMPSE2. Here we show two panels; however, MetaGLIMPSE can combine any number of reference panels. The first panel imputes A|A, A|C, whereas the second panel imputes A|C, C|C. MetaGLIMPSE takes the imputed haplotypes and genotype likelihoods as inputs and then estimates weights, customized for each individual and marker using a HMM. In this example, the first marker has a higher weight for panel 1 and the second marker has a higher weight for panel 2. Finally, MetaGLIMPSE outputs meta-imputed genotypes by generating a weighted average of each of the imputation estimates. In the example, MetaGLIMPSE imputes A\A, C\C as the best guess genotypes, which derive from the higher weights on panel 1 and panel 2 respectively.

Model Description

In a dataset where variants have been called at M loci, we model sequencing data for each individual as genotype likelihoods $GL_m = [GL(0), GL(1), GL(2)]$ at each called variant m. Let there be K distinct reference panels, and let $(A_{1,k,m}, A_{2,k,m})$ denote the imputation output for marker m ($1 \leq m \leq M$) with panel k ($1 \leq k \leq K$), where $A_{i,k,m} \in \{0,1\}$. $(A_{1,k,m}, A_{2,k,m})$ are the outputs of any low coverage sequencing algorithm, e.g. GLIMPSE2. $A_{1,k,m} + A_{2,m}$ is the unphased imputed genotype for a given marker and panel. Finally, let $G_m \in \{0,1,2\}$ be the meta-imputed genotypes, obtained through estimating marker and panel-pair specific weights via HMM.

Defining the HMM

Hidden and Observed States

We define an HMM across all imputed genotypes m where the genotype likelihoods GL_m are the observed states, and the "best" imputation results $A^*=(A_{*,1,m}, A_{*,2,m}) \in \{(A_{1,k,m}, A_{2,k,m}): 1 \leq k \leq K, 1 \leq m \leq M\}$ are the hidden states.

Emission and Transition Probabilities

The emission probability captures how well each imputation result fits with the observed sequencing data (Equation 1).

$$P\left(GL_m \mid A_{1,k_1,m},\ A_{2,k_2,m}\right) = \sum_{g \in G} GL_m(g) * P(G_m = \mathrm{g} \mid (A_{1,k1,m},\ A_{1,k2,m})) \quad (1)$$

Due to the genotype uncertainty at each marker, we marginalize over the possible unphased genotypes in the emission probability. We also sum over phase since the imputation results are phased whereas the genotype likelihoods are not.

The transition probability is a function of marker distance based on the Li and Stephens transition probabilities for diploid data (Ausmees et al. 2023; Li et al. 2010).

$$P\left((A_{1,k_1,m}, A_{2,k_2,m})\middle|(A_{1,k_1',m-1}, A_{2,k_2',m-1}\right) = \begin{cases} \frac{\lambda_m{}^2}{|K^*|^2}, & if\ k_1 \neq k_1'\ and\ k_2 \neq k_2' \\ \frac{\lambda_m(1-\lambda_m)}{|K^*|} + \frac{\lambda_m{}^2}{|K^*|^2}, & if\ k_1 \neq k_1'\ and\ k_2 = k_2'\ or\ k_1 = k_1'\ and\ k_2 \neq k_2' \\ (1-\lambda_m)^2 + \frac{2(\lambda_m - \lambda_m{}^2)}{|K^*|} + \frac{\lambda_m{}^2}{|K^*|^2}, & if\ k_1 = k_1'\ and\ k_2 = k_2' \end{cases} \quad (2)$$

Here $\lambda_m = 1 - e^{-\theta * d(m-1,\ m)}$, and $\theta = 2 * 10^{-7}$ is the recombination rate between the variants (Yu et al. 2022). $d(m-1, m)$ is the one dimensional distance in base pairs between the mth and $m-1^{th}$ markers. The probability of a transition between distinct hidden states increases as the distance between adjacent markers increases. We fix θ according to values found in the meta-imputation literature (Yu et al. 2022).

Weight Estimation and Meta Imputed Genotypes

We apply the forward-backward algorithm (Baum 1970) to this HMM, which gives us the marginal probability of each imputation panel pair being optimal at variant m. These probabilities serve as the weights, denoted as $W_{(A_{1,k_1,m}, A_{2\ k_2,m})}$ (Equation 3), where the subscripts indicate the reference panels k_1 and k_2 used for the two haplotypes.

$$W_{(A_{1,k_1,m}, A_{2\ k_2,m})} = P\left((A_{1,k_1,m}, A_{2,k_2,m}) \mid GL(g)_1 \ldots GL(g)_M\right) \quad (3)$$

Using these weights, we compute the meta-imputed genotype at each marker by averaging the imputed genotypes from all possible panel pairs, weighted by their probabilities. Formally, for each marker m, we calculate G_m as follows: (Equation 4).

$$G_m = \sum_{k_1,k_2 \in K^*: k_1 \neq k_2} W_{(A_{1,k_1,m},A_{2\,k_2,m})} \left(A_{1,k_1,m} + A_{2,k_2,m}\right) \quad (4)$$

This derives an expected genotype value that reflects a consensus across all panels, giving more weight to the panels that best fit the observed sequencing data.

Two Versions of MetaGLIMPSE

Depending on how we define the hidden states in the HMM, MetaGLIMPSE can either ignore or leverage the phasing in GLIMPSE2 or any other low-coverage sequencing imputation algorithm. Therefore, we have two versions of MetaGLIMPSE. The default version, MetaGLIMPSE, leverages phasing, whereas "MetaGLIMPSE-plain" does not.

Formally, let $\mathcal{S}$ be the set of all hidden states. For MetaGLIMPSE-plain, $\mathcal{S}=\{(A_{11},A_{12}),(A_{21},A_{22}),\ldots,(A_{k1},A_{k2})\}$, requiring that A_{*1} and A_{*2} originate from the same reference panel. For the default MetaGLIMPSE, we define $\mathcal{S}=\{(A_{1i},A_{2j}) : 1 \leq i,j \leq k\}$, thus allowing A_{*1} to originate from a different reference panel than A_{*2}. MetaGLIMPSE-plain has $|\mathcal{S}| = K$ hidden states whereas MetaGLIMPSE has $|\mathcal{S}| = \binom{2K}{2}$ hidden states.

Evaluation

To assess the performance of MetaGLIMPSE and MetaGLIMPSE-plain in a variety of scenarios we use chromosome 20 from the unrelated 2504 samples in 1000 Genomes Phase 3 resequenced at 30x (Byrska-Bishop et al. 2022) as the source dataset for both target samples and reference panels.

First, from the 2504 total samples we remove samples that have any parents in the dataset, using the pedigree file from Byrska Bishop et al. 2022, leaving 2495 total unrelated samples. From these remaining samples, we generate low coverage target samples by downsampling to a range of mean coverages {0.1x, 0.5x, 1x, 2x, 4x, 6x, 8x} using samtools view v1.19.2. For each sample and coverage, we calculate genotype likelihoods using bcftools mpileup v1.19 at all biallelic sites for the entire 3202 individuals in 1000 Genomes (Byrska-Bishop et al. 2022).

We then create five reference panels by subsetting the samples for the following populations based on the 1000 Genomes ancestry classification: Europeans half A (n = 249), Europeans half B (n = 249), Europeans (n = 498), Africans (n = 599), and Africans + Europeans (n = 1098). We generate European half A and European half B by randomly sampling half of the full European panel. We exclude admixed African-Americans and remove any singletons in the 2495 samples. We also remove non-biallelic SNPs from each reference panel, resulting in ~344k and ~281K variants that are polymorphic in both the Europeans half A and Europeans half B panel and both the African and European panel, respectively. When imputing a target sample, we exclude that sample from the reference panel.

We impute the downsampled East Asians and African-American with the African, European, and African + European-mega panels. We impute the downsampled Europeans with each sample's distinct European A, European B, and full European panels. We impute in 6cM length chunks using GLIMPSE 2.0.1 software -AP option branch for 19 burn-in iterations and 1 main

iteration for each reference panel described above. We then meta-impute East Asians, African-Americans, and Europeans using their respective single panel imputations via MetaGLIMPSE.

To compare the performance of both single panel imputation, mega-imputation, and meta-imputation, we identify the polymorphic markers in both single-panel imputations and, for these markers, calculate the accordance of each imputation result with the 30x genotypes. We bin variants by MAF and calculate in each bin the aggregated Pearson R^2 between the 30x genotypes and imputed genotype estimates. Aggregated R^2 combines the true genotypes into one vector and the estimated genotypes into a second vector and calculates the correlation between these two vectors (Rubinacci et al. 2021). This metric accounts for the relatively small number of target samples we impute, since the Pearson R^2per marker is not defined when the variance of either the true or imputed genotypes is zero (Li et al. 2023). We use aggRSquare v1.0.0 to compute the aggregated R^2. The allele frequency of a variant corresponds to its value across all unrelated samples in the 30x resequenced Phase 3 1000 Genomes dataset.

Results

We evaluate MetaGLIMPSE in situations with two available reference panels, yielding four distinct imputed datasets for the target samples. Initially, two imputed cohorts are generated, each using a single reference panel. A third cohort is then created through mega-imputation, where both panels are merged into a unified reference. The fourth cohort is produced by applying MetaGLIMPSE to combine the two individually imputed cohorts.

The accuracy of these four strategies is then assessed across sequencing coverages ranging from 0.1x to 8x, considering a variety of realistic scenarios. These scenarios include instances where (i) target samples share ancestry with the reference panels; (ii) where admixed individuals are present, with each panel representing only one ancestral background; (iii) where samples come from ancestries not well-represented in either panel; and finally, (iv) where target samples are sequenced at ultra-low coverage, such as 0.1x.

Meta Imputing with reference panels from the same population

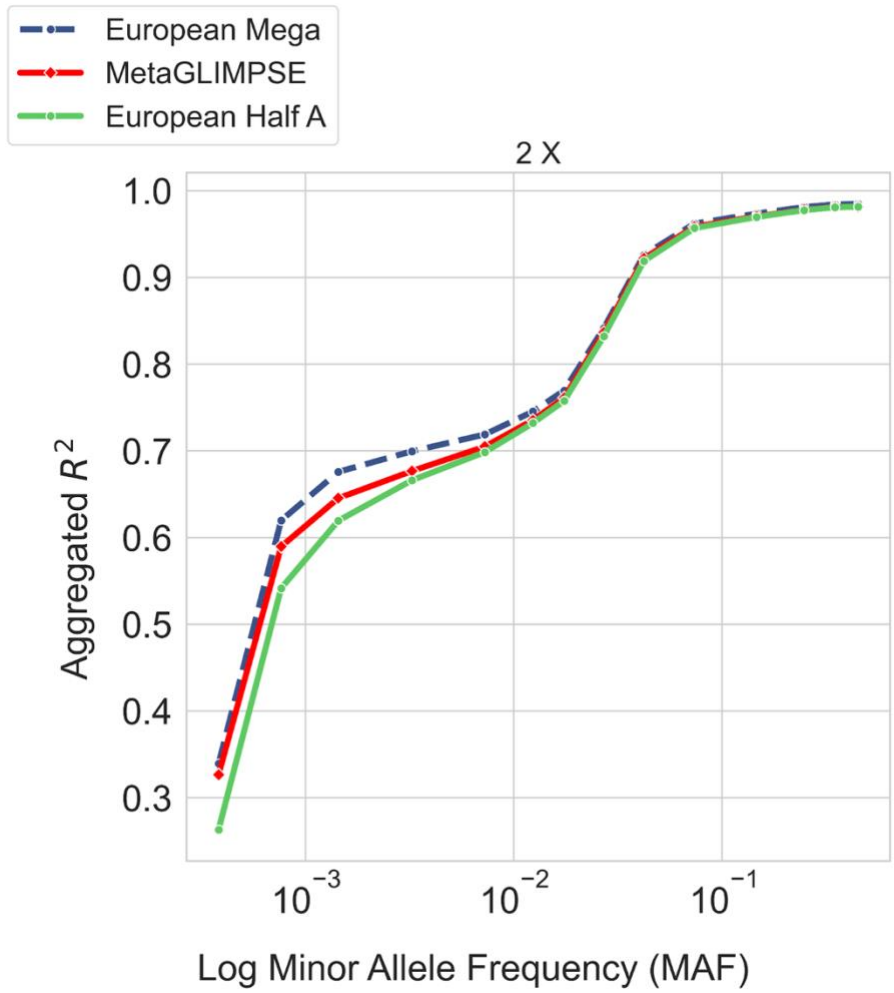
A common use of meta-imputation is to combine imputation results from several reference panels such as TopMed and HRC that cannot be merged in a mega-imputation framework due to privacy reasons. We model this challenge by splitting the unrelated 30x European samples (n = 499) in 1000 Genomes (Byrska-Bishop et al. 2022) into two reference panels while excluding the target sample from either panel. We imputed each of our 499 target samples with its distinct set of three reference panels separately: European mega panel (n = 498), European half A (n = 249) and European half B (n = 249). We then combined the results from European half A imputation and European half B imputation using MetaGLIMPSE. We compared MetaGLIMPSE to the two single half-panel imputations and the one mega-imputation.

While mega-imputation, unsurprisingly, achieves the highest overall aggregated R^2, and the individual European panels, at an almost identical aggregated R^2, perform the worst. MetaGLIMPSE provides more accurate imputation results than either of the individual European panels across all MAFs and all sequencing depths (Supplementary Figure 1a-b).

At 2x coverage, particularly for rare (MAF<1%) variants, mega-imputation is advantageous as it has a higher aggregated R^2 than the European single panel imputations, whereas for MAF > 2%, all imputations converge to a similar aggregated R^2 regardless of the reference panel (Figure 2). Specifically, for variants with MAF between 0.1% and 0.2%, mega-imputation (aggregated R^2 = 0.676) has 9.1% higher aggregated R^2 than the European half A imputation (aggregated R^2 = 0.620). MetaGLIMPSE (aggregated R^2 = 0.645) recovers 45% of the 9.1% performance increase in aggregated R^2 by the mega-imputation. Overall, MetaGLIMPSE improves the aggregated R^2 of the single panel imputation by 4.1%.

As coverage increases from 0.5x to 8x, MetaGLIMPSE recovers more of the performance increase by mega-imputation for rare variants. At 0.5x and 2x, MetaGLIMPSE recovers 26% and 45% of the performance increase at MAF between 0.1% and 0.2%, respectively, whereas at 8x, it recovers 66% of the performance increase. At 8x, for MAF > 1%, the performance of mega-imputation and MetaGLIMPSE are practically equal (Supplementary Figure 1a-b).

A



B

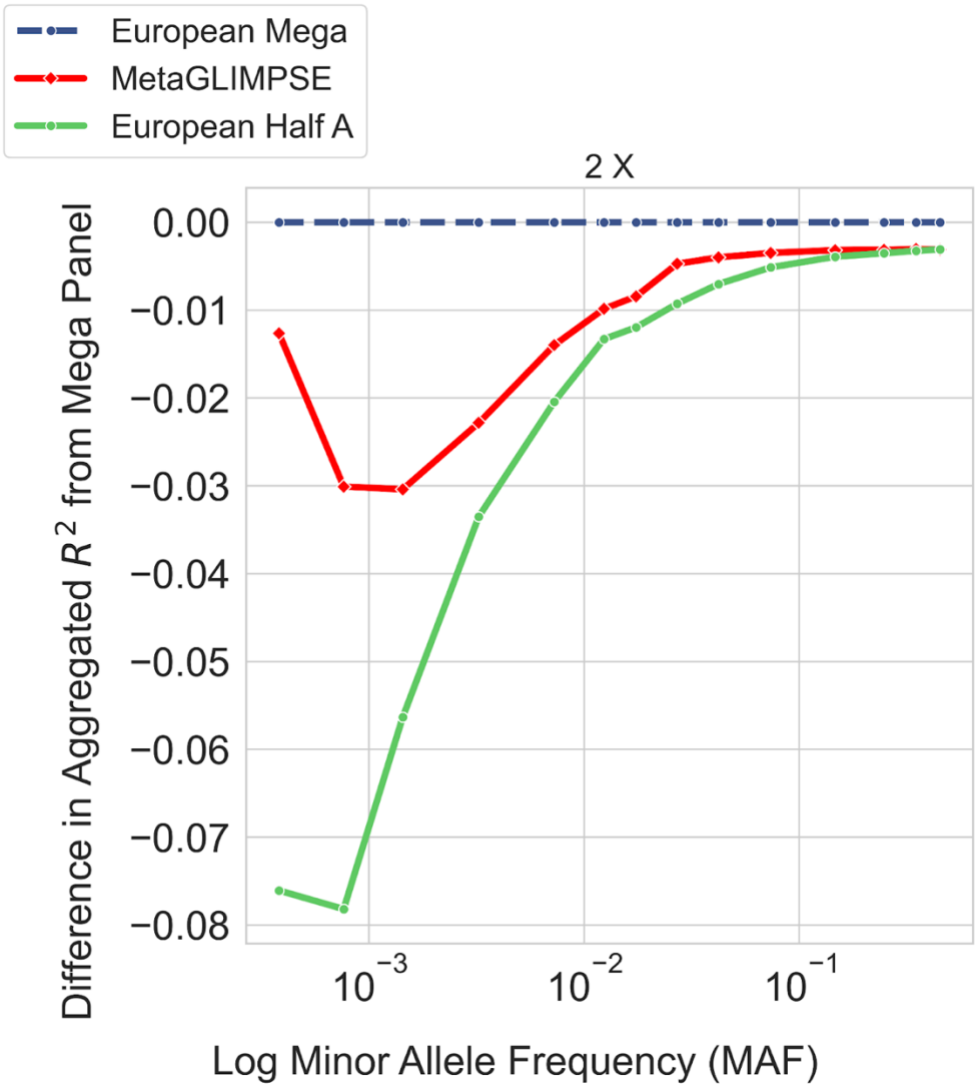


Figure 2A and B Meta Imputation of European target samples with two European reference panels. We impute 499 Europeans from 1000G downsampled to 2x using two European reference panels (n=249). We evaluate MetaGLIMPSE by meta-imputing the results and calculating the aggregated R^2 between the (meta-)imputed genotypes and true genotypes. We plot the aggregated R^2 for each imputation by MAF. We report the aggregated R^2 (A) and the difference between the aggregated $R^2$from mega imputation and the aggregated $R^2$for each of the other imputations (B). Both single reference panels, European A and European B, have roughly identical aggregated R^2s, so we show only the European A imputation in the main text.

Meta Imputing Admixed Samples

Meta-imputation facilitates imputation for admixed populations by combining imputation results from reference panels that each contain a different source population. We model imputation for admixed targets by imputing 57 unrelated African Americans from 1000 Genomes using three imputation panels: an African population without African-Americans (n = 599), a European

population (n = 499), and a mega-panel (n=1098), which is a concatenation of the African and European reference panels. We generate meta-imputation results by combining the African panel imputation results and European panel imputation results through MetaGLIMPSE.

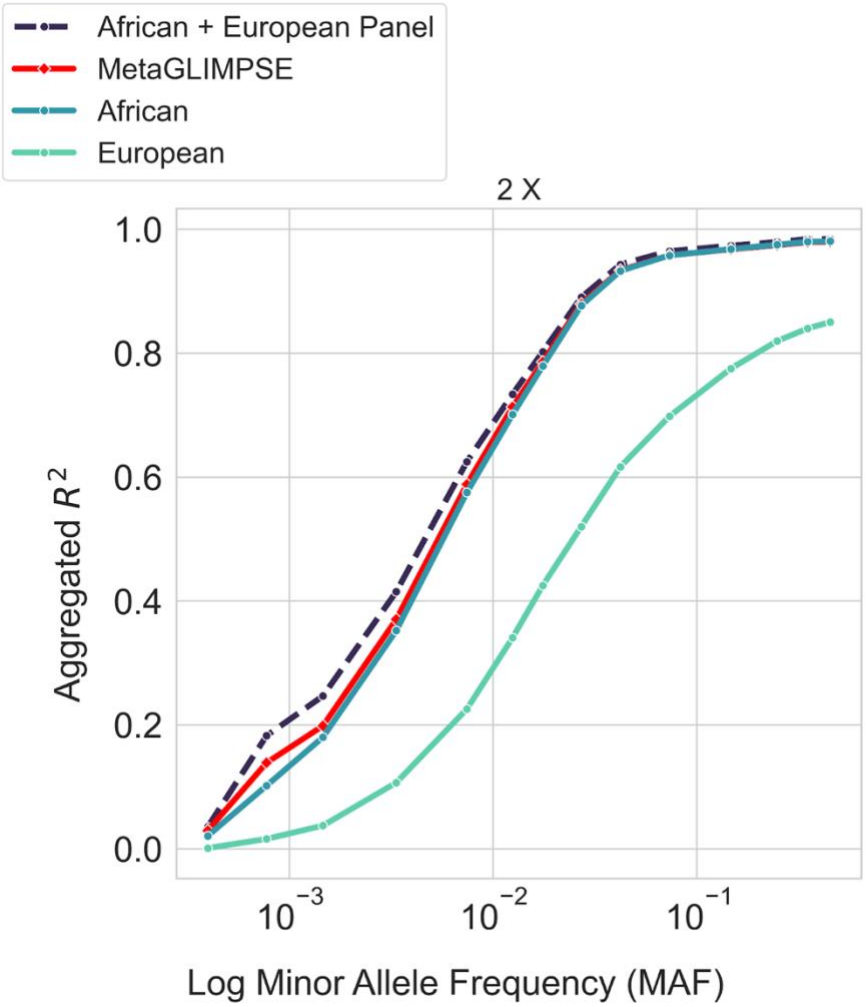
Across all MAFs and coverages, the mega-imputation has the highest aggregated R^2 and the European panel has the lowest aggregated R^2. Across all coverages, for MAFs > 2%, the performance of MetaGLIMPSE is indistinguishable from the performance of the African panel. However, for MAFs < 2%, MetaGLIMPSE performs substantially better than the African panel alone, which concordes with the MAFs where mega-imputation provides the greater performance increase (Supplementary Figure 2).

At 2x coverage, for MAFs between 0.05% and 0.1%, mega-imputation (aggregated R^2 = 0.183) has a 79% higher aggregated R^2 than the best single-panel imputation (aggregated R^2 = 0.102). MetaGLIMPSE (aggregated R^2= 0.138) has a 36% higher aggregated R^2 than imputing with the best single-panel imputation and recovers 46% of the performance increase by mega-imputation (Figure 3a-b).

MetaGLIMPSE's increased performance compared to the African panel imputation is roughly consistent regardless of coverage. For instance, at 0.5x and 8x, for MAFs between 0.05% and 0.1%, MetaGLIMPSE has a 26% and 24% higher aggregated R^2, respectively, than the best single panel imputation

We also find that phasing has an important role in meta-imputation. Across all coverages, when MetaGLIMPSE leverages phasing, it has a higher aggregated R^2 compared to when it does not utilize phasing, denoted MetaGLIMPSE-plain. For example, at 2x, for variants with MAF between 0.05% and 0.1%, MetaGLIMPSE has a 23% higher aggregated R^2 than MetaGLIMPSE-plain (aggregated R^2 = 0.114) (Supplementary Figure 2a-b).

A


African + European Panel
MetaGLIMPSE
African
European
2 X
1.0
0.8
0.6
0.4
0.2
0.0
Aggregated R^2
10^{-3}
10^{-2}
10^{-1}
Log Minor Allele Frequency (MAF)

B

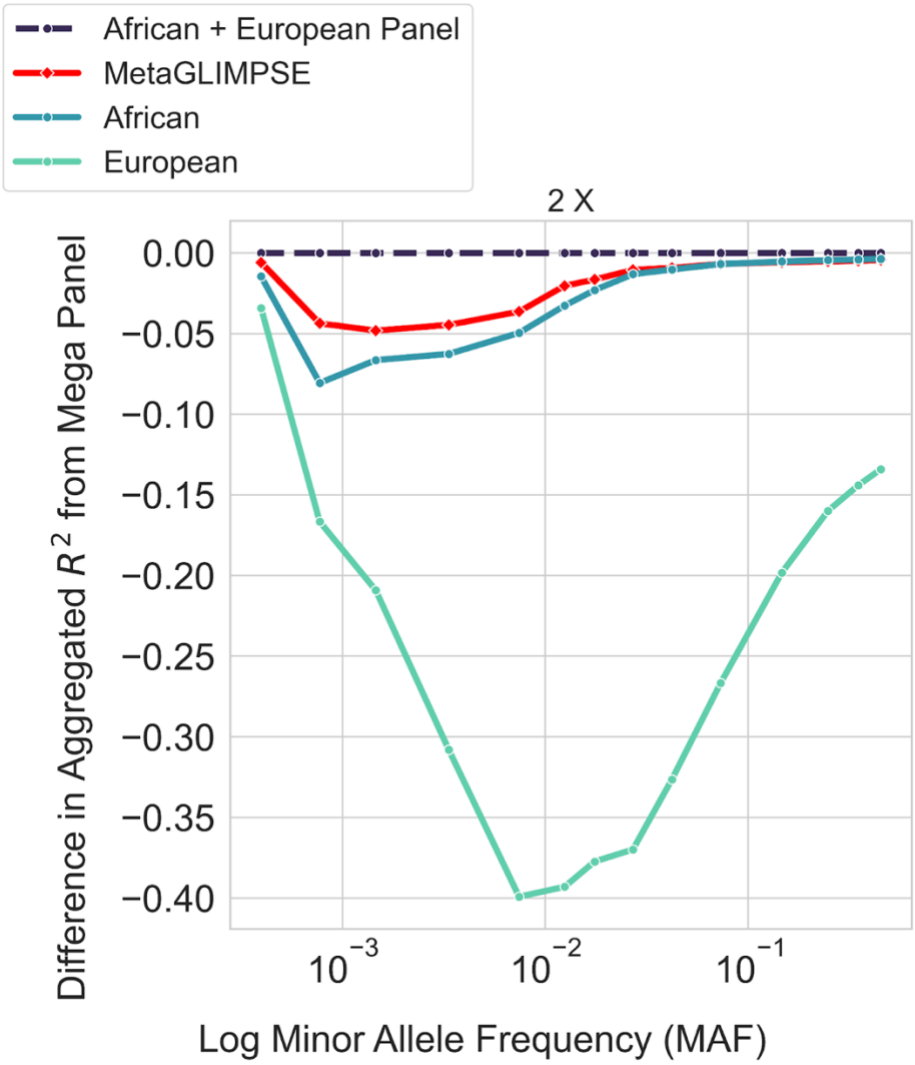


Figure 3A and B Meta Imputation of African-American target samples with an African and a European reference panel. We impute 57 African-Americans from 1000G downsampled to 2x using an African (n=599) and a European reference panel (n = 499). We evaluate MetaGLIMPSE by meta-imputing the results and calculating the aggregated R^2 between the (meta-)imputed genotypes and true genotypes. We plot the aggregated R^2 for each imputation by MAF. We report the aggregated R^2 (A) and the difference between the aggregated R^2 from mega imputation and the aggregated R^2 for each of the other imputations (B).

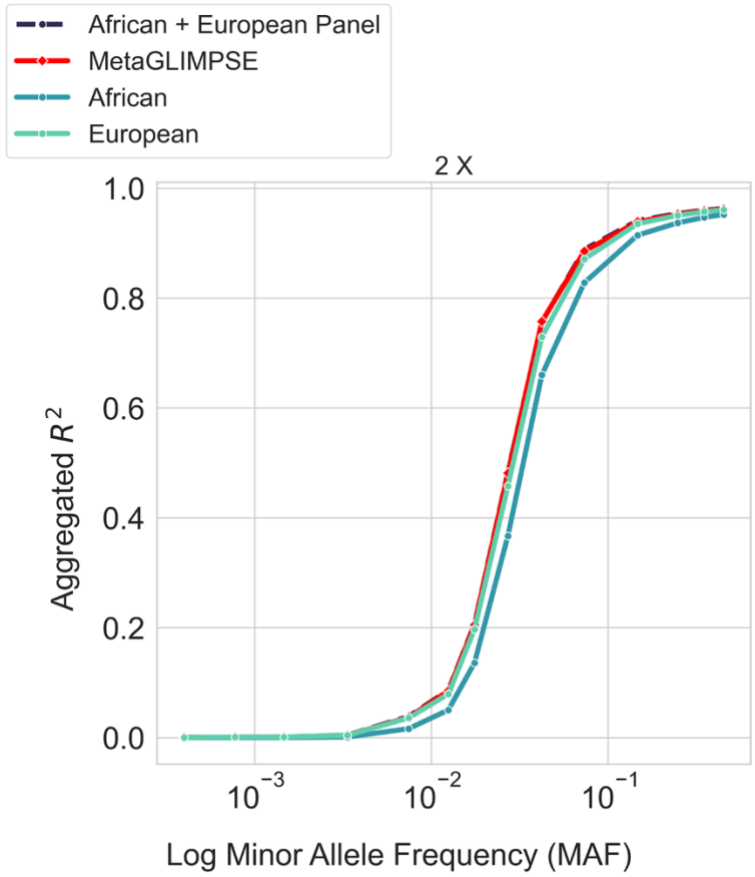
Meta Imputing Populations Absent from Reference Panels

Next, we consider a scenario that is applicable to a sequenced population that either does not exist in any reference panel or is not in an accessible reference panel. We mimic this scenario by imputing a target population using two reference panels that are not closely related to the target population. Specifically, we meta-impute 504 East Asians from 1000 Genomes with the same African and European reference panels described above. We compare MetaGLIMPSE to the mega-imputation and the two single panel imputations.

For coverages between 0.5x and 8x, mega imputation has the highest aggregated R^2 followed by MetaGLIMPSE, the European panel imputation and the African panel imputation, respectively. Across all coverages, the aggregated R^2 is overall much smaller for lower frequency alleles (MAF < 1%) than in the other scenarios. (Supplementary Figure 3a-b).

Across all MAFs and coverages of 0.5x and greater, MetaGLIMPSE practically matches the aggregated R^2 of the mega-imputation. For instance, at 2x, for MAF between 3.5% and 5%, MetaGLIMPSE matches the aggregated R^2 of the mega-imputation (aggregated R^2= 0.757). This result is consistent across coverages. At 0.5x, MetaGLIMPSE (aggregated R^2 = 0.599) has a slightly higher aggregated R^2 than mega-imputation (aggregated R^2 = 0.595), and at 8x, the aggregated R^2 is the same for mega-imputation and MetaGLIMPSE (aggregated R^2 = 0.861).

A



B

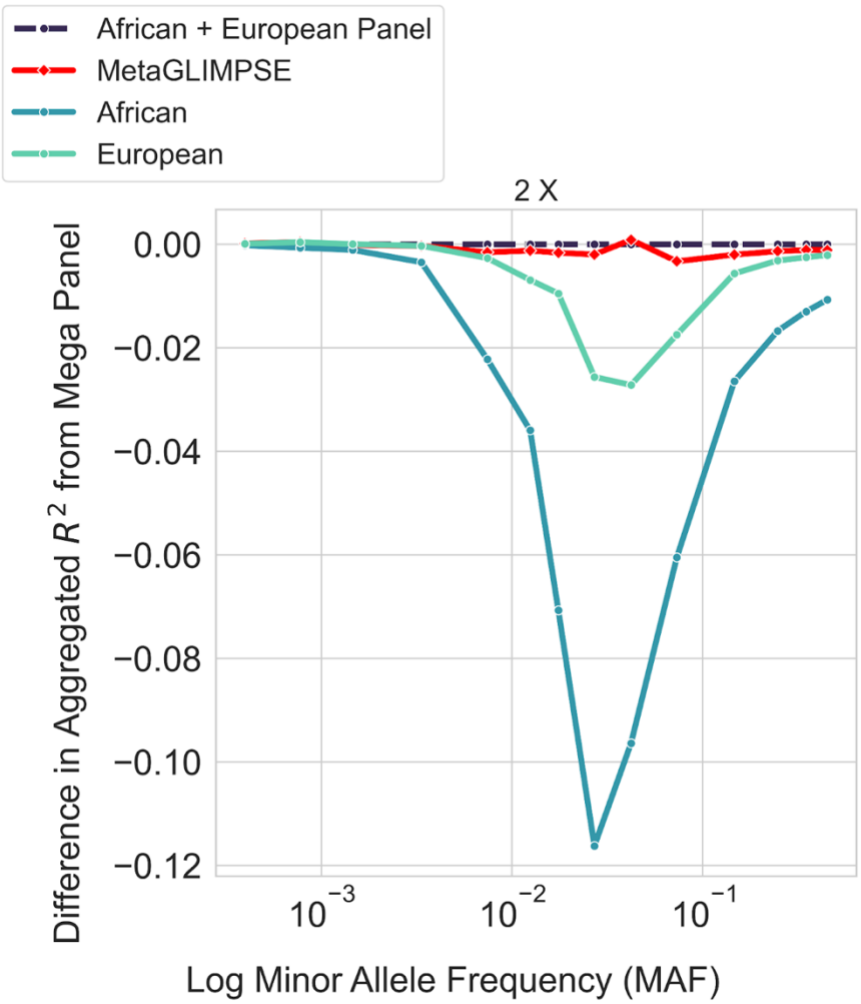


Figure 4A and B Meta Imputation of East Asian target samples with an African and a European reference panel. We impute 504 East Asians from 1000G downsampled to 2x using an African (n=599) and a European reference panel (n = 499). We evaluate MetaGLIMPSE by meta-imputing the results and calculating the aggregated R^2 between the (meta-)imputed genotypes and true genotypes. We plot the aggregated R^2 for each imputation by MAF. We report the aggregated R^2 (A) and the difference between the aggregated R^2 from mega imputation and the aggregated R^2 for each of the other imputations (B).

Meta-Imputing at Ultra Low Coverages

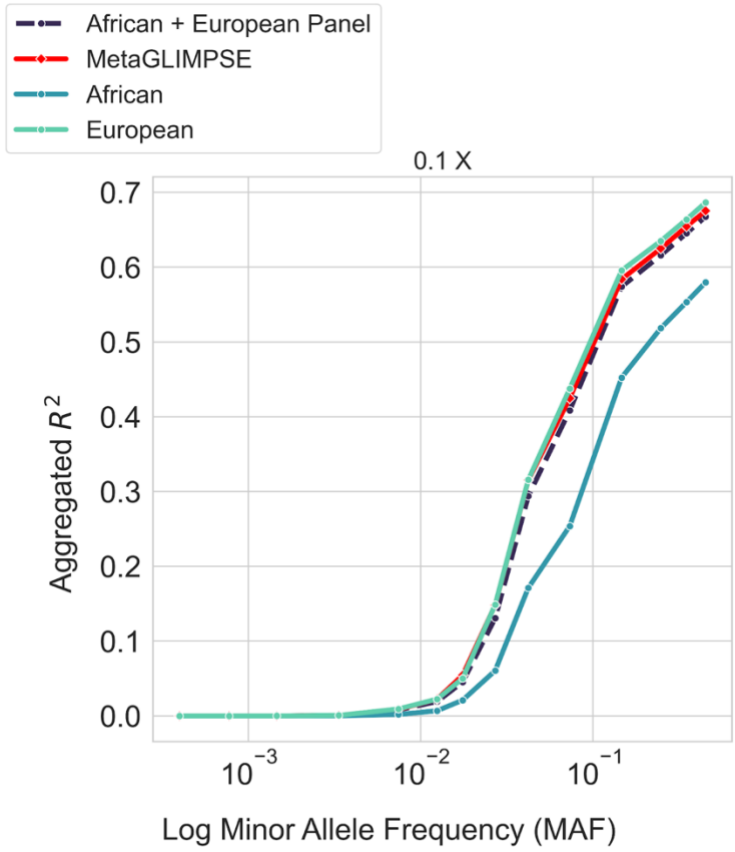
Finally, we consider meta-imputation at ultra-low coverages (e.g. 0.1x), which are common in fields like ancient DNA (Orlando 2021). We simulate this scenario by downsampling our target samples from the above three simulations to 0.1x. We then compare MetaGLIMPSE for these samples at 0.1x to mega-imputation and single-panel imputations.

For the scenario where both reference panels are from the same ancestry as the target and for the scenario where the reference panels are the source populations to an admixed target, the

patterns we observed persist across all coverages, and the mega-imputation is most accurate, followed by meta-imputation and then the individual panels (Supplementary Figure 4a-b). However, in the scenario where both reference panels are selected from different populations than the target, the results at 0.1x differ from the patterns observed for higher coverages (Figure 5).

For this scenario, the relative performance of the different imputations varies by MAF. For MAFs less than 1.5%, MetaGLIMPSE achieves a comparable aggregated R^2 to the best single-panel imputation, which is the European panel imputation, while the mega-imputation performs equal to or worse than the European panel imputation. For MAFs between 1.5% and 5%, MetaGLIMPSE achieves an equal or higher aggregated R^2 than the European panel imputation and the mega-imputation, which performs worse than the European panel imputation. For MAFs higher than 5%, MetaGLIMPSE outperforms the mega-panel imputation, but not the single-panel European imputation.

A



B

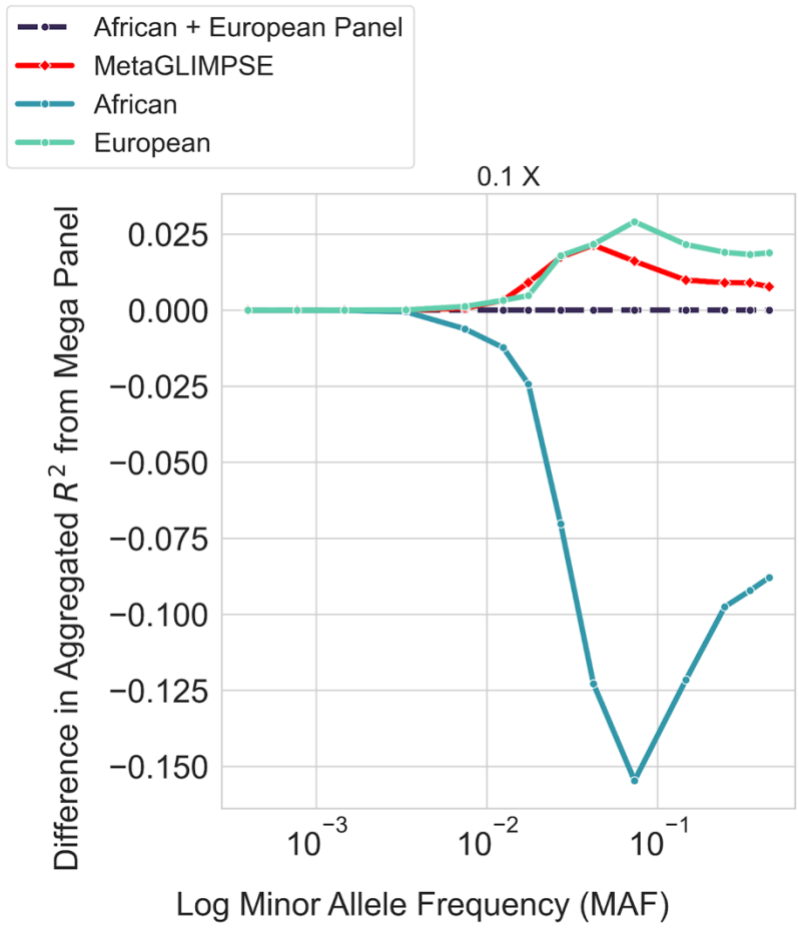


Figure 5A and 5B Meta Imputation of East Asian target samples at 0.1x with an African and a European reference panel. We impute 504 East Asians from 1000G downsampled to 0.1x using an African (n=599) and a European reference panel (n = 499). We evaluate MetaGLIMPSE by meta-imputing the results and calculating the aggregated R^2 between the (meta-)imputed genotypes and true genotypes. We plot the aggregated R^2 for each imputation by MAF. We report the aggregated R^2 (A) and the difference between the aggregated R^2 from mega imputation and the aggregated R^2 for each of the other imputations (B).

Computation Time

Computation time is an important consideration in imputation and meta-imputation algorithms (Das et al. 2018; Howie et al. 2012; Yu et al. 2022) due to the large number of markers processed. For HMMs, computation time is linear with the number of markers and quadratic with the number of reference panel pairs. Therefore, for *K* reference panels and *M* markers, the computation time, without file reading and writing, for MetaGLIMPSE is $O(K^4M)$, whereas it is $O(K^2M)$ for MetaGLIMPSE-plain because the MetaGLIMPSE has more possible configurations for reference panel pairs than MetaGLIMPSE-plain.

We report the computational times for MetaGLIMPSE, including file reading and writing time according to the number of samples and markers. MetaGLIMPSE computes the underlying HMM in 30,000 marker chunks to allow for parallelization, which significantly speeds up computation time. MetaGLIMPSE with 1 sample and 900,000 markers takes a total time of 4.42 seconds on an Intel Xeon Platinum 8176 CPU @ 2.10 GHz with 1 core, if we run all 30 chunks in parallel. 900,000 markers are roughly the size of the union set of polymorphic markers between the African and European panels on chromosome 20. Since chromosome 20 comprises roughly 2% of the genome, the time for a whole genome will be 3 minutes and 41 seconds. For 50 samples, the time for a whole genome is 2 hours 21 minutes and 51 seconds, and for 500 samples, the time for a whole genome is 26 hours 45 minutes 42 seconds.

	M = 900,000	M = 1 Genome
N= 1	4.42 seconds	3 minutes 41 seconds
N= 50	2 minutes 50 seconds	2 hours 21 minutes 51 seconds
N = 500	32 minutes 7 seconds	26 hours 45 min 42 seconds

Table 1: Grid of computation times by number of samples (N) and number of markers (M) while running all the chunks in parallel.

Discussion:

We present a novel meta-imputation framework for low coverage sequencing that enables imputation with multiple reference panels. MetaGLIMPSE is efficient and improves accuracy, achieving consistently better aggregated R^2 than the best single panel imputation for a wide range of coverages and allele frequencies. Meta-imputation benefits rare variants the most, consistent with observations elsewhere that the imputation quality of rare variants is dependent on reference panel size and the extent of shared ancestry with target samples (Das et al. 2018). Our method achieves better imputation accuracy than single-panel imputation for a wide variety of scenarios, including meta-imputing samples with the same ancestry as the reference panels, admixed populations with the source populations as reference panels, and populations that are not closely related to the individuals in any reference panel. These scenarios cover a wide range of applications. First, the scenario with target samples and reference panels both belonging to the same ancestry group, in this case Europeans, would translate to combining imputation results for European target samples from several reference panels such as TopMed and HRC that cannot be used in a mega-imputation framework. Another application involves meta-imputing European targets sequenced using blended genome exome (BGE) (Botlz et al. 2024). Since BGE is sequenced at 2-4x coverage (Boltz et al. 2024), and at 4x, in this scenario, meta-imputation is nearly equal to mega-imputation. Secondly, MetaGLIMPSE would also facilitate imputing admixed samples, such as those in the Mexican Biobank, where the best reference panel for imputing a given sample depends on its global ancestry composition (Ziyatdinov et al. 2023). Thirdly, the imputation of East Asians with Africans and Europeans mimics the imputation of newly sequenced populations or of ancient DNA (aDNA), for which the available reference panels consist of only modern-day humans (Allentoft et al. 2024; Ringbauer et al. 2024; Sousa da Mota et al 2023). One third of aDNA of the 10,000+ aDNA samples are not of European ancestry. (Mallick et al. 2023). Meta imputing aDNA with 1000 Genomes as well as a more population specific panel more closely related to the target sample could improve the imputation accuracy of aDNA. Finally, our method is easily adaptable to other low coverage sequencing imputation algorithms other than GLIMPSE2, such as Beagle 4.1 (Browning et al.

2016) or QUILT (Davies et al. 2021). MetaGLIMPSE could also meta-impute samples imputed with multiple low coverage sequencing algorithms. Some algorithms may impute certain variants better than others (Rui et al. 2021), motivating the need for meta-imputation.

We consider a wide range of coverages for low pass sequencing and observe broadly consistent patterns in the results. Only at coverages of 0.1X, when the target population is not represented in the reference panel, we find that mega imputation is less accurate than the single best reference panel imputation. This matches observations from SNP array imputation that larger reference panels do not always outperform the smaller, more population specific reference panels (Carlson et al. 2023; Das et al. 2018). We believe this is driven by the challenge of GLIMPSE2 choosing the correct template haplotype with such limited information. This suggests that, at extremely low coverages, researchers more carefully choose their reference panel because a larger reference panel is not necessarily better.

Our results are relevant for aDNA, which is imputed with, most commonly, 1000 Genomes as a reference panel (Allentoft et al. 2024; Ringbauer et al. 2024; Sousa da Mota et al. 2023). The current literature on imputing aDNA (Allentoft et al. 2024; Ausmees et al. 2022; Ausmees et al. 2023; Biddanda et al. 2022; Hui et al. 2020; Sousa da Mota et al. 2023) does not exploit the 10,000+ and growing set of sequenced ancient DNA samples (Mallick et al. 2023), from which one could create an aDNA reference panel. Meta imputing aDNA with both a small aDNA-only panel and large modern-day panel would leverage these aDNA samples. Due to their lower coverage and higher genotyping error rate (Ausmees et al. 2023; Hui et al. 2020; Garrido Marques et al. 2024) than modern reference panel genomes, aDNA might not be selected as the best template haplotype by imputation algorithms. MetaGLIMPSE, which explicitly allows for aDNA as templates in imputation algorithms, could provide a substantial benefit–especially for older ancient genomes, which have undergone many subsequent population replacements (Orlando et al. 2021).

We made several choices during our analysis. First, we do not present meta-imputation with large reference panels such as HRC and TopMed, since they are not publicly available for imputation using algorithms designed for low coverage sequencing data. Therefore, many low coverage sequencing studies use 1000 Genomes as their reference panel (Allentoft et al. 2024; Santos et al. 2024; Sousa da Mota 2023). While we do not explicitly model the size of the reference panel in MetaGLIMPSE, we expect that since our method takes a weighted average of single panel imputation estimates, it should extend to larger reference panels. Second, we handle variants missing in only one of the two reference panels by setting its genotype to the reference. Deciding how to handle variants that are not polymorphic in all reference panels is a challenging consideration for meta-imputation (Yu et al. 2022). The advantage of setting the missing variant to the reference is that this is correct in the vast majority of cases. However, individuals of the same ancestry as those in the reference panel that have a non-reference allele might not have been sequenced. Future work would involve constructing a more robust strategy to handle such variants. Finally, we evaluate our results by subdividing a single reference panel; however, combining multiple reference panels, each phased differently, can sometimes lead to some hard to interpret results.

In summary, our method outperforms the best single panel imputation for a wide range of scenarios—including target samples and reference panels from the same ancestry, admixed target samples with reference panels from each source population, and target samples whose ancestry is not represented in the available reference panels. MetaGLIMPSE is especially useful for increasing the imputation accuracy of rare variants, and we expect our method will

help increase the imputation accuracy for samples of many different ancestries by allowing for privacy protected imputation across multiple reference panels.

Code and Data Availability

The code for MetaGLIMPSE along with tutorial and instructions is publicly available at https://github.com/karinkumar/MetaGLIMPSE. The links to the software used are as follows: bcftools and samtools (https://www.htslib.org/) and GLIMPSE2 branch ap-option (https://github.com/odelaneau/GLIMPSE/tree/ap-field). This branch of GLIMPSE2 provides the inputs necessary to run MetaGLIMPSE. The 2504 Phase 3 1000 Genomes resequenced at 30x are available at the European Nucleotide Archive under accession number PRJEB31736.

Acknowledgements

This work was funded by the NIH Grants R01 HG011031 and R01 HG005855 to SZ. We acknowledge Ketian Yu and Jean Morrison for helpful conversations and support.

Author Contributions

K.H.K. and S.Z. developed the ideas. K.H.K. and S.Z. wrote the manuscript with contributions from S.R.; S.R. developed the inputs for the software. K.H.K. implemented the software and performed the analyses.

References

Allentoft, M. E., Sikora, M., Refoyo-Martínez, A., Irving-Pease, E. K., Fischer, A., Barrie, W., Ingason, A., Stenderup, J., Sjögren, K. G., Pearson, A., Sousa da Mota, B., Schulz Paulsson, B., Halgren, A., Macleod, R., Jørkov, M. L. S., Demeter, F., Sørensen, L., Nielsen, P. O., Henriksen, R. A., … Willerslev, E. (2024). Population genomics of post-glacial western Eurasia. *Nature*, *625*(7994), 301–311. https://doi.org/10.1038/s41586-023-06865-0

Ausmees, K., & Nettelblad, C. (2023). Achieving improved accuracy for imputation of ancient DNA. *Bioinformatics (Oxford, England)*, *39*(1), 1–5. https://doi.org/10.1093/bioinformatics/btac738

Ausmees, K., Sanchez-Quinto, F., Jakobsson, M., & Nettelblad, C. (2022). An empirical evaluation of genotype imputation of ancient DNA. *G3: Genes, Genomes, Genetics*, *12*(6). https://doi.org/10.1093/g3journal/jkac089

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Boltz, T. A., Chu, B. B., Liao, C., Sealock, J. M., Ye, R., & Majara, L. (2024). A blended genome and exome sequencing method captures genetic variation in an unbiased , high-quality , and cost- effective manner. *BioRxiv*, 1–23.

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics*, *98*(1), 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Eichler, E. E., Korbel, J. O., Lee, C., Marschall, T., Devine, S. E., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Carlson, J. C. (2023). Improving imputation quality in Samoans through the integration of population-specific1 sequences into existing reference panels. *MedRxiv*, *1*(165), 1–13.

Connell, J. O., Yun, T., Moreno, M., Li, H., Litterman, N., Kolesnikov, A., Noblin, E., Chang, P., Shastri, A., Dorfman, E. H., Shringarpure, S., Auton, A., Carroll, A., & Mclean, C. Y. (n.d.). *A population-speci fi c reference panel for improved genotype imputation in African Americans*. *2021*, 1–9. https://doi.org/10.1038/s42003-021-02777-9

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype imputation from large reference panels. *Annual Review of Genomics and Human Genetics*, *19*, 73–96. https://doi.org/10.1146/annurev-genom-083117-021602

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Davies, R. W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C. M., Chan, Y. F., & Myers, S. (2021). Rapid genotype imputation from sequence with reference panels. *Nature Genetics*, *53*(7), 1104–1111. https://doi.org/10.1038/s41588-021-00877-0

Erven Jolijn A.M. , Scheu Amelie, Verdugo Pereira Marta, Cassidy, Lara, Chen Ningbo, Gehlen Birgit, Street Martin, Madsen Ole, M. V. E. (2024). A high coverage Mesolithic aurochs genome and effective leveraging of ancient cattle genomes using whole genome imputation. *BioRxiv*.

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). Minimac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784. https://doi.org/10.1093/bioinformatics/btu704

Garrido Marques, A., Rubinacci, S., Malaspinas, A. S., Delaneau, O., & Sousa da Mota, B. (2024). Assessing the impact of post-mortem damage and contamination on imputation performance in ancient DNA. *Scientific Reports*, *14*(1), 1–13. https://doi.org/10.1038/s41598-024-56584-3

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L., & Kivisild, T. (2020). Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific Reports*, *10*(1), 1–8. https://doi.org/10.1038/s41598-020-75387-w

Jewett, E. M., Zawistowski, M., Rosenberg, N. A., & Zöllner, S. (2012). A Coalescent Model for Genotype Imputation. *Genetics*, *191*(4), 1239–1255. https://doi.org/10.1534/genetics.111.137984

Koenig, Z., Yohannes, M. T., Nkambule, L. L., Goodrich, J. K., Kim, H. A., Zhao, X., Wilson, M. W., Tiao, G., Hao, S. P., Sahakian, N., Chao, K. R., Consortium, gnomAD P., Rehm, H. L., Neale, B. M., Talkowski, M. E., Daly, M. J., Brand, H., Karczewski, K. J., Atkinson, E. G., & Martin, A. R. (2023). A harmonized public resource of deeply sequenced diverse human genomes. *BioRxiv*, 2023.01.23.525248. https://www.biorxiv.org/content/10.1101/2023.01.23.525248v3%0Ahttps://www.biorxiv.org/content/10.1101/2023.01.23.525248v3.abstract

Li, J. H., Liu, A., Buerkle, C. A., Palmer, W., Belbin, G. M., Ahangari, M., Gibson, M. J. S., & Flagel, L. (2023). *The effects of reference panel perturbations on the accuracy of genotype imputation*. 1–20.

Li, Na (Department of Biostatistics, University of Washington, Seattle, W. 98195), & Stephens, Matthew (Department of Statistics, University of Washington, Seattle, W. 98195). (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, *165*(December), 2213–2233.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834. https://doi.org/10.1002/gepi.20533

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F. K., Atwoli, L., Bowers, T., Chibnik, L. B., Daly, M. J., DeSmet, T., Dodge, S., Fekadu, A., Ferriera, S., Gelaye, B., Gichuru, S., … Zingela, Z. (2021). Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *American Journal of Human Genetics*, *108*(4), 656–668. https://doi.org/10.1016/j.ajhg.2021.03.012

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Nakamura, W., Hirata, M., Oda, S., Chiba, K., Okada, A., Mateos, R. N., Sugawa, M., Iida, N., Ushiama, M., Tanabe, N., Sakamoto, H., Sekine, S., Hirasawa, A., Kawai, Y., Tokunaga, K., Tsujimoto, S. I., Shiba, N., Ito, S., Yoshida, T., … Shimanuki, H. (2024). Assessing the efficacy of target adaptive sampling long-read sequencing through hereditary cancer patient genomes. *Npj Genomic Medicine*, *9*(1), 1–13. https://doi.org/10.1038/s41525-024-00394-z

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00011-0

Ringbauer, H., Huang, Y., Akbari, A., Mallick, S., Olalde, I., Patterson, N., & Reich, D. (2024). Accurate detection of identity-by-descent segments in human ancient DNA. *Nature Genetics*, *56*(1), 143–151. https://doi.org/10.1038/s41588-023-01582-w

Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., & Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, *53*(1), 120–126. https://doi.org/10.1038/s41588-020-00756-0

Santos, R., Moreno-Torres, V., Pintos, I., Corral, O., de Mendoza, C., Soriano, V., & Corpas, M. (2024). Low-coverage whole genome sequencing for a highly selective cohort of severe COVID-19 patients. In *GigaByte* (Vol. 2024). https://doi.org/10.46471/gigabyte.127

Sousa da Mota, B., Rubinacci, S., Cruz Dávalos, D. I., Carlos, C. E., Sikora, M., Johannsen, N. N., Szmyt, M. H., Włodarczak, P., Szczepanek, A., Przybyła, M. M., Schroeder, H., Allentoft, M. E., Willerslev, E., Malaspinas, A. S., & Delaneau, O. (2023). Imputation of ancient human genomes. *Nature Communications*, *14*(1), 1–17. https://doi.org/10.1038/s41467-023-39202-0

Yu, K., Das, S., LeFaive, J., Kwong, A., Pleiness, J., Forer, L., Schönherr, S., Fuchsberger, C., Smith, A. V., & Abecasis, G. R. (2022). Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *American Journal of Human Genetics*, *109*(6), 1007–1015. https://doi.org/10.1016/j.ajhg.2022.04.002

Ziyatdinov, A., Torres, J., Alegre-díaz, J., Backman, J., Mbatchou, J., Turner, M., Gaynor, S. M., Joseph, T., Zou, Y., Liu, D., Wade, R., Staples, J., Panea, R., Popov, A., Bai, X., Balasubramanian, S., Habegger, L., Lanche, R., Lopez, A., … Berumen, J. (2023). *Genotyping , sequencing and analysis of 140 , 000 adults from Mexico City*. https://doi.org/10.1038/s41586-023-06595-3