Theoretical Population Biology 87 (2013) 62–74

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Genotype imputation in a coalescent model with infinitely-many-sites mutation

Lucy Huang [a], Erkan O. Buzbas [b], Noah A. Rosenberg [b,*]

[a] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[b] Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Available online 16 October 2012

Keywords:
Coalescent
Imputation
Population divergence

ABSTRACT

Empirical studies have identified population-genetic factors as important determinants of the properties of genotype-imputation accuracy in imputation-based disease association studies. Here, we develop a simple coalescent model of three sequences that we use to explore the theoretical basis for the influence of these factors on genotype-imputation accuracy, under the assumption of infinitely-many-sites mutation. Employing a demographic model in which two populations diverged at a given time in the past, we derive the approximate expectation and variance of imputation accuracy in a study sequence sampled from one of the two populations, choosing between two reference sequences, one sampled from the same population as the study sequence and the other sampled from the other population. We show that, under this model, imputation accuracy—as measured by the proportion of polymorphic sites that are imputed correctly in the study sequence—increases in expectation with the mutation rate, the proportion of the markers in a chromosomal region that are genotyped, and the time to divergence between the study and reference populations. Each of these effects derives largely from an increase in information available for determining the reference sequence that is genetically most similar to the sequence targeted for imputation. We analyze as a function of divergence time the expected gain in imputation accuracy in the target using a reference sequence from the same population as the target rather than from the other population. Together with a growing body of empirical investigations of genotype imputation in diverse human populations, our modeling framework lays a foundation for extending imputation techniques to novel populations that have not yet been extensively examined.

1. Introduction

The field of human genetics has recently witnessed an explosion in the number of published genome-wide association (GWA) studies, revealing hundreds of novel disease-associated genes (Donnelly, 2008; Manolio et al., 2008; Hindorff et al., 2009, 2011). The considerable potential of GWA studies—which examine thousands to millions of genetic markers in samples of unrelated individuals with the goal of uncovering genotype–phenotype correlations—to ultimately improve human health has been widely recognized (e.g., Hardy and Singleton, 2009; Manolio, 2010; Stranger et al., 2011).

Among factors contributing to the success of GWA studies has been the advent of genotype-imputation methods that use chromosomal segments shared among subjects to predict, or impute, genotypes at marker positions not directly measured in individual GWA studies (Li et al., 2006; Nicolae, 2006; Browning and Browning, 2007; Marchini et al., 2007; Servin and Stephens, 2007). In imputation studies, the haplotypes of "reference" individuals that have been genotyped at a higher density than GWA individuals targeted for imputation often serve as template sequences on the basis of which unknown genotypes in the targets are inferred. Because imputation increases the number of markers that can be interrogated for disease associations and permits larger sample sizes by enabling data sets typed on different platforms to be merged, it can increase the statistical power of GWA studies (e.g., Li et al., 2009; Marchini and Howie, 2010). This important role for imputation is likely to persist as technology advances; when whole-genome sequencing of at least a portion of GWA samples becomes routinely feasible, the power of sequence-based GWA studies can be improved by imputation in genotyped individuals using sequenced individuals as templates (Li et al., 2011).

Recent studies have empirically examined the determinants of genotype-imputation accuracy in globally distributed human populations (Guan and Stephens, 2008; Pei et al., 2008; Huang et al., 2009, 2011; Li et al., 2009; Fridley et al., 2010; Surakka et al., 2010). These investigations have shown that, in imputation-based GWA studies, population-genetic factors play an important role in determining levels of imputation accuracy attainable in a study population. Factors such as the level of linkage disequilibrium in

[*] Corresponding author.
E-mail address: noahr@stanford.edu (N.A. Rosenberg).

0040-5809/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2012.09.006

a study population and the degree of genetic similarity between a study population and a reference population whose members serve as templates have been found in imputation experiments to be prominent drivers of imputation accuracy (Egyud et al., 2009; Huang et al., 2009, 2011; Paşaniuc et al., 2010; Shriner et al., 2010). Though empirical work on genotype imputation has provided some understanding of the population-genetic factors that affect imputation accuracy, theoretical work exploring the influence of these factors on imputation accuracy has been limited.

A theoretical approach to studying genotype imputation under a population-genetic model offers the potential for producing a variety of insights. First, by obtaining expressions for the mean and the variance of the imputation accuracy as a function of population-genetic parameters, we can explain patterns of imputation accuracy observed in empirical studies in terms of the population-genetic factors that affect the underlying genealogical relationship between study and reference individuals. Second, using simple expressions, imputation accuracy can be evaluated with less computation than in simulation-based approaches, enabling investigators to predict imputation accuracy under a model rather than implement computationally intensive simulations. Third, unlike targeted simulations specific to particular populations of interest, a general modeling framework can be adapted for organisms beyond humans in which imputation-based association studies and large-scale genomic resources have begun to emerge (e.g., Atwell et al., 2010; Druet et al., 2010; Kirby et al., 2010; Badke et al., 2012; Hickey et al., 2012).

Jewett et al. (2012) introduced a theoretical model for evaluating imputation accuracy as a function of population-genetic parameters. Using a coalescent framework, they analytically studied the effect of reference-panel size on imputation accuracy, as well as the degree to which the use of reference haplotypes from the same population as a target sequence (an "internal" reference panel) improves the accuracy of imputation compared to the use of reference haplotypes from a separate population (an "external" reference panel). In order to incorporate a large sample size in obtaining their analytical results, however, Jewett et al. (2012) did not account for randomness in the mutation process. Instead, their treatment of mutation amounted to an assumption that mutation is a deterministic process, in which mutations accumulate along a genealogical branch in direct proportion to the branch length. Consequently, under this assumption, the best template for imputation is always a haplotype whose coalescence time with the target sequence on which genotypes are to be imputed is smallest.

Here, we consider a coalescent model of genotype imputation that, at the cost of examining only a small sample size, allows for randomness in the imputation process resulting from the stochasticity of mutation. Assuming the infinitely-many-sites mutation model, we derive the approximate expectation and variance of imputation accuracy under a straightforward imputation scheme, conditioning on a mutation parameter (θ), a proportion of markers genotyped in a given length of a chromosome (p), and a time to divergence between the target population and an external reference population (t_d). As in Jewett et al. (2012), our derivations account for randomness in the genealogy by considering the distribution of genealogies under a model in which study and reference individuals are sampled from two populations that diverged at time t_d in the past. We pose the following questions: (1) What are the influences of θ, p, and t_d on the expectation and variance of imputation accuracy? (2) What is the expected gain in imputation accuracy in a study sequence targeted for imputation by using a reference sequence from the same population as the target rather than from a different population? Answers to these questions provide information on the factors that affect genotype-imputation accuracy, with implications for the design of imputation-based association studies and the expansion of genomic databases.

2. Theory

In this section, we introduce a theoretical framework that permits the computation of the approximate expectation and variance of imputation accuracy in a target sequence on the basis of one of two reference sequences. The framework has four parts: a coalescent model for the genealogical relationships among lineages, a mutation model, a decision rule that guides the selection of the reference sequence for the imputation, and an imputation scheme that specifies how the imputation is performed and its accuracy evaluated. The computations employ three approximations, including a Monte Carlo step for the evaluation of certain integrals.

2.1. A coalescent model

Consider two populations P_1 and P_2 that diverged from an ancestral population P_A at time t_d in the past. Further, consider three haploid individuals—a study individual targeted for imputation (henceforth simply identified as a *target* and denoted by I) and two reference individuals (denoted by R_1 and R_2). Individuals R_1 and I are from population P_1, and individual R_2 is from population P_2. In a diploid organism, the haploid individuals can be viewed as single haplotypes.

For the set of three individuals, let $\mathcal{G}$ denote a labeled gene tree topology. We denote the times during which three and two lineages exist in the genealogy, and the divergence time between populations P_1 and P_2, by T_3, T_2, and t_d, respectively. We assume that the diploid effective population size, denoted by N_e, is the same for populations P_1, P_2, and P_A. All times are measured in units of $2N_e$ generations. For convenience, we refer to $\mathcal{G}$ as the genealogy and use the notation $\mathbf{T} = (T_3, T_2, t_d)$.

The genealogy $\mathcal{G}$ can have one of four possible genealogical types G (Fig. 1): three in which the first coalescence event occurs more anciently than the population-divergence time t_d ($g = A, B, C$), and one in which the first coalescence occurs more recently than t_d ($g = D$). For each type, we label the external branches for the lineages of reference individuals R_1 and R_2 and target individual I by 1, 2, and 3, respectively (Fig. 1). We examine the genealogy backward in time, combining the external and internal branches immediately descended from the root into one branch that takes on the label for the external branch. Thus, for instance, branch 2 in genealogy A has length $t_d + t_3 + 2t_2$. Note that, as shown in Fig. 1, in genealogies A, B, and C, t_3 measures from t_d back in time, whereas, in genealogy D, t_3 measures from the present.

Under standard coalescent theory, the time (in units of $2N_e$ generations) for k lineages in a population to coalesce to $k-1$ lineages follows an exponential distribution with parameter $\binom{k}{2}$ (Kingman, 1982a,b). For our model of sequences R_1, R_2, and I, the probability density function of T_2 for genealogies $g = A, B, C, D$ is

$$f_{T_2}(t_2|G = g, t_d) = e^{-t_2}, \quad g = A, B, C, D. \tag{1}$$

For $g = A, B, C$, the probability density function of T_3 is

$$f_{T_3}(t_3|G = g, t_d) = 3e^{-3t_3}, \quad g = A, B, C. \tag{2}$$

For genealogy D, however, the lineages for reference sequence R_1 and target sequence I are constrained to coalesce no more anciently than the divergence time t_d. Thus, the probability density function for the time T_3 during which all three lineages exist is

$$f_{T_3}(t_3|G = D, t_d) = \frac{e^{-t_3}}{1 - e^{-t_d}}\mathbf{1}_{\{t_3 < t_d\}}, \tag{3}$$

where $\mathbf{1}_{\{\mathcal{B}\}}$ is the indicator function that takes a value of 1 if condition $\mathcal{B}$ holds and is 0 otherwise.

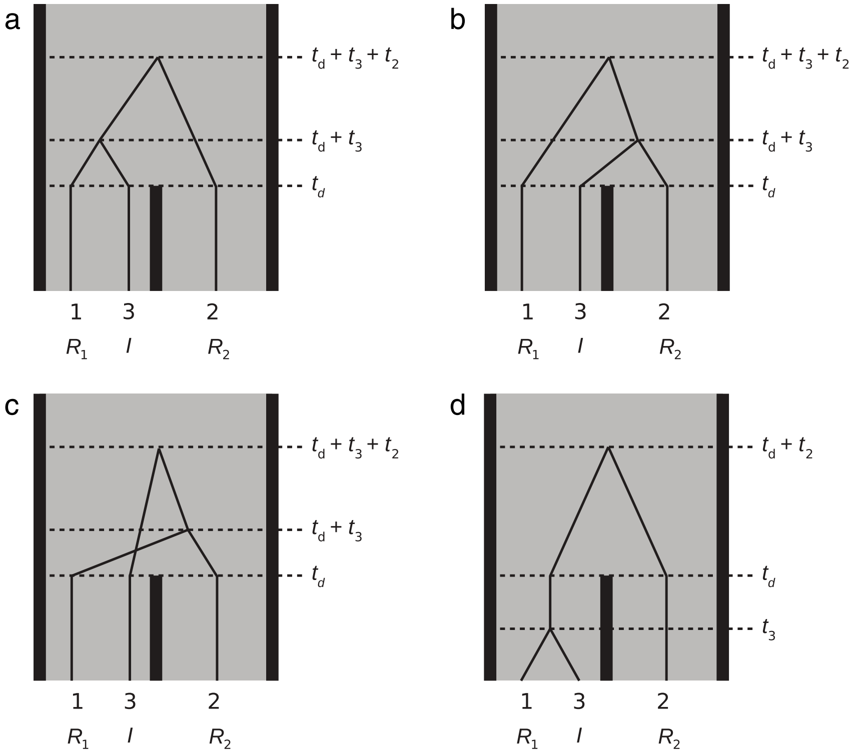


Fig. 1. Four possible genealogical types for a set of three sequences: a candidate reference sequence R_1 and a sequence I targeted for imputation from one population, and another candidate reference sequence R_2 from a second population. The two populations diverged from an ancestral population at time t_d in the past. Times t_3 and t_2 are coalescence times for sets of three and two distinct lineages. For genealogies A, B, and C, t_3 measures from the divergence time, whereas, for genealogy D, t_3 measures from the present.

We compute the probability $\mathbb{P}(G = g|t_d)$ for $g = A, B, C, D$ by conditioning on the time interval of the coalescence of reference sequence R_1 and target sequence I. Considering the lineages backward in time, we define $\mathcal{E}$ to be the event that R_1 and I do not coalesce by t_d and $\bar{\mathcal{E}}$ to be the event that R_1 and I do coalesce by t_d. Because each pair of lineages in the same population has the same probability of being the first to coalesce, conditional on $\mathcal{E}$, genealogies A, B, and C occur with equal probability. Therefore,

$$\mathbb{P}(G = g|t_d) = \mathbb{P}(G = g|\mathcal{E}, t_d)\,\mathbb{P}(\mathcal{E}|t_d) + \mathbb{P}(G = g|\bar{\mathcal{E}}, t_d)\,\mathbb{P}(\bar{\mathcal{E}}|t_d)$$
$$= \begin{cases} \left(\dfrac{1}{3}\right) e^{-t_d} + (0)(1 - e^{-t_d}) = \dfrac{1}{3} e^{-t_d} \\ \quad \text{if } g = A, B, C \\ (0)e^{-t_d} + (1)(1 - e^{-t_d}) = 1 - e^{-t_d} \\ \quad \text{if } g = D. \end{cases} \qquad (4)$$

2.2. Stochastic mutation

We examine imputation at a locus evolving according to the infinitely-many-sites mutation model, with no recombination (Watterson, 1975). We consider only the polymorphic sites in a sample of three sequences, ignoring all sites that are not polymorphic in the sample. Under the infinitely-many-sites model, the number of polymorphic sites in a sample is the same as the number of mutations in its gene genealogy, and we use the terms "polymorphic sites" and "mutations" interchangeably. We denote the population-scaled mutation parameter by $\theta = 4N_e\mu L$, where μ is the mutation rate per base per generation, and L is the length (in bases) of the sequence under consideration.

For our computations of the mean and the variance of the imputation accuracy, we will need for each genealogical type g various distributions related to numbers of mutations that occur on a random genealogy. Let X_i be the total number of mutations that occur on branch i under the neutral coalescent model with infinitely-many-sites mutation ($i = 1, 2, 3$). We assume that, with probability p, a given site is genotyped in the target, and that sites are chosen independently for genotyping. Reference sequences R_1 and R_2 are assumed to be genotyped at all sites at which the set of three lineages $\{R_1, R_2, I\}$ is polymorphic. Let Y_i be the random number among the X_i mutations on branch i that are genotyped in the target. Let $h_i(\mathbf{T}; g)$ denote the length of branch i for a given genealogy assumed to have time $\mathbf{T}$ and type g.

We assume that the total number of mutations X_i on a branch, conditional on its branch length $h_i(\mathbf{T}; g)$, follows a Poisson distribution with parameter $h_i(\mathbf{T}; g)\,\theta/2$. That is,

$$X_i|\mathbf{T}, g \sim \text{Poisson}(h_i(\mathbf{T}; g)\,\theta/2), \qquad (5)$$

where $h_i(\mathbf{T}; g)$ is specified in Table 1 for $g = A, B, C, D$ and $i = 1, 2, 3$. Because individual sites are genotyped in the target independently of each other, each with probability p, conditional on the total number of mutations X_i on branch i, the random number of mutations Y_i on branch i that are genotyped in the target follows a binomial distribution with parameters X_i and p,

$$Y_i|X_i \sim \text{Bin}(X_i, p). \qquad (6)$$

Eqs. (5) and (6) imply that, conditional on the coalescence and population-divergence times $\mathbf{T}$, the number of mutations Y_i on branch i that are typed in the target follows a Poisson distribution:

$$Y_i|\mathbf{T}, g \sim \text{Poisson}(h_i(\mathbf{T}; g)\,\theta p/2). \qquad (7)$$

Similarly, the number of mutations $X_i - Y_i$ on branch i that are untyped in the target follows a binomial distribution with parameters X_i and $1 - p$:

$$(X_i - Y_i)|X_i \sim \text{Bin}(X_i, 1 - p). \qquad (8)$$

Eqs. (5) and (8) then imply that

$$(X_i - Y_i)|\mathbf{T}, g \sim \text{Poisson}(h_i(\mathbf{T}; g)\,\theta(1 - p)/2). \qquad (9)$$

Table 1
Branch lengths $h_i(\mathbf{T}; g)$. For genealogical types $g = A, B, C, D$ and branches $i = 1, 2, 3$ under the two-population model in Fig. 1, the branch lengths are given in units of $2N_e$ generations.

Genealogical type	Branch		
	1	2	3
A	$t_d + t_3$	$t_d + t_3 + 2t_2$	$t_d + t_3$
B	$t_d + t_3 + 2t_2$	$t_d + t_3$	$t_d + t_3$
C	$t_d + t_3$	$t_d + t_3$	$t_d + t_3 + 2t_2$
D	t_3	$2t_d - t_3 + 2t_2$	t_3

Finally, conditional on $\mathbf{T}$, the numbers of mutations on any two branches Y_i and Y_j that are genotyped in the target are independent. The difference between two independent Poisson-distributed variables is described by the Skellam distribution (Johnson and Kotz, 1969). Thus, for $i, j \in \{1, 2, 3\}$ and $i \neq j$,

$$(Y_i - Y_j)|\mathbf{T}, g \sim \text{Skellam}(h_i(\mathbf{T}; g)\,\theta p/2, h_j(\mathbf{T}; g)\,\theta p/2), \tag{10}$$

with mean $(h_i(\mathbf{T}; g) - h_j(\mathbf{T}; g))\theta p/2$, variance $(h_i(\mathbf{T}; g) + h_j(\mathbf{T}; g))\theta\, p/2$, and probability mass function

$$f_{\text{Sk}}(y_i - y_j; h_i(\mathbf{T}; g)\,\theta p/2, h_j(\mathbf{T}; g)\,\theta p/2) = e^{-\frac{h_i(\mathbf{T};g)+h_j(\mathbf{T};g)}{2}\theta p}$$

$$\times \left(\frac{h_i(\mathbf{T}; g)}{h_j(\mathbf{T}; g)}\right)^{\frac{y_i - y_j}{2}} I_{|y_i - y_j|}\left(\theta p\sqrt{h_i(\mathbf{T}; g)\, h_j(\mathbf{T}; g)}\right). \tag{11}$$

$I_\alpha(x)$ is the modified Bessel function of the first kind.

2.3. A decision rule

Recall that in our sample of three haploid sequences—two references R_1 and R_2, and a target I—we consider only polymorphic sites. The target sequence is assumed to be genotyped at only a subset of the sites that are polymorphic in the set of three sequences. We now further assume that missing genotypes at untyped markers in the target are imputed by copying the corresponding genotypes in a chosen reference sequence that has been genotyped at all of the sites. The choice of a reference is specified by a decision rule δ that we introduce below.

Accurate imputation relies on the occurrence of chromosomal segments that are shared identically by descent between target and reference sequences. Similar sequences are more likely to descend from the same ancestral sequence, and therefore we generally expect imputation accuracy in a target sequence to increase with increased genetic similarity between the target and reference sequences. We define a distance statistic d_i between reference sequence R_i ($i = 1, 2$) and target sequence I to be the number of pairwise differences between the two sequences at positions genotyped in the target. Because we associate R_1, R_2, and I with branches 1, 2, and 3, respectively, we have

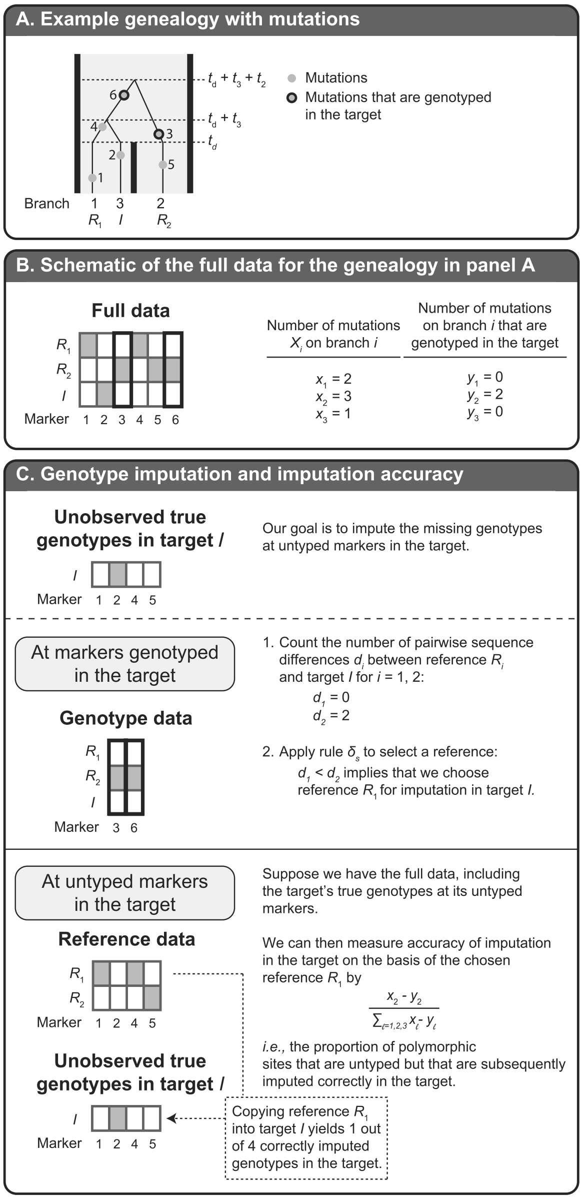
$$d_i = Y_i + Y_3.$$

Smaller values of d_i indicate greater genetic similarity—measured at the genotyped positions in the target—between reference sequence R_i and the target.

We now present a decision rule δ—based on the distance statistic d_i—that we use to select one of the two reference sequences, R_1 or R_2, for imputation in the target (Box 1). In short, rule δ selects the genetically more similar reference sequence to the target, as measured by d_i.

Box 1. Rule δ

- If $d_1 < d_2$, use reference R_1.
- If $d_2 < d_1$, use reference R_2.
- If $d_1 = d_2$, with probability 1/2, use reference R_1, and, with probability 1/2, use reference R_2.


A. Example genealogy with mutations
$t_d + t_3 + t_2$
$t_d + t_3$
t_d
6
4
3
2
5
1
Branch 1 3 2
R_1 I R_2
Mutations
Mutations that are genotyped in the target
B. Schematic of the full data for the genealogy in panel A
Full data
R_1
R_2
I
Marker 1 2 3 4 5 6
Number of mutations X_i on branch i
Number of mutations on branch i that are genotyped in the target
$x_1 = 2$
$x_2 = 3$
$x_3 = 1$
$y_1 = 0$
$y_2 = 2$
$y_3 = 0$
C. Genotype imputation and imputation accuracy
Unobserved true genotypes in target I
I
Marker 1 2 4 5
At markers genotyped in the target
Genotype data
R_1
R_2
I
Marker 3 6
At untyped markers in the target
Reference data
R_1
R_2
Marker 1 2 4 5
Unobserved true genotypes in target I
I
Marker 1 2 4 5

Our goal is to impute the missing genotypes at untyped markers in the target.

1. Count the number of pairwise sequence differences d_i between reference R_i and target I for i = 1, 2:
   $d_1 = 0$
   $d_2 = 2$
2. Apply rule δ_s to select a reference:
   $d_1 < d_2$ implies that we choose reference R_1 for imputation in target I.

Suppose we have the full data, including the target's true genotypes at its untyped markers.

We can then measure accuracy of imputation in the target on the basis of the chosen reference R_1 by

$$\frac{x_2 - y_2}{\sum_{\ell=1,2,3} x_\ell - y_\ell}$$

i.e., the proportion of polymorphic sites that are untyped but that are subsequently imputed correctly in the target.

Copying reference R_1 into target I yields 1 out of 4 correctly imputed genotypes in the target.

Fig. 2. Schematic of the imputation procedure. (A) An example genealogy with mutations. (B) The full data for the genealogy in (A). (C) The imputation process. In (B) and (C), each row represents a sequence, and each column represents a site. White and gray boxes indicate the two allelic types at a site, and thick black lines indicate positions genotyped in the target.

2.4. An imputation scheme

Once a reference sequence is chosen for imputation in the target, we substitute missing genotypes at untyped markers in the target by those at corresponding positions in the reference. We illustrate the reference selection and the imputation procedure in Fig. 2.

Imputation accuracy is assessed as the proportion of polymorphic sites untyped in the target that are imputed correctly on the basis of a chosen reference sequence. Let R_i, $i = 1, 2$, denote the chosen sequence, and let R_j, $j \neq i$ and $j = 2, 1$, denote the reference sequence that is not chosen. If R_1 is chosen as the template for imputation, the imputation accuracy obtained is

$$Z = \frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}. \tag{12}$$

Alternatively, if R_2 is chosen as the template, then the imputation accuracy is

$$Z = \frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}. \tag{13}$$

In both cases, the numerator is $X_j - Y_j$, because, under the infinitely-many-sites mutation model, polymorphic sites produced by mutations on the branch for reference sequence R_j are exactly where reference R_i and target I have identical genotypes. Thus, to count the number of sites imputed correctly in the target when using reference sequence R_i, one simply counts the number of mutations on the branch for reference sequence R_j that are not genotyped in the target but that are imputed. The denominator $\sum_{\ell=1}^{3}(X_\ell - Y_\ell)$ corresponds to the total number of untyped polymorphic sites in the target that are subsequently imputed; in the case that $\sum_{\ell=1}^{3}(X_\ell - Y_\ell) = 0$, Z is undefined because there are no genotypes to impute.

2.5. Expectation and variance of imputation accuracy

At sites genotyped in both reference and target sequences, the number d_i of pairwise differences between reference R_i $(i = 1, 2)$ and target I is observable. Given d_1 and d_2, we apply rule δ in Box 1 to select a reference sequence for imputing missing genotypes at untyped markers in the target. In this section, conditioning on the model parameters—the mutation parameter θ, the proportion p of polymorphic sites genotyped in the target, and the population-divergence time t_d—we derive the approximate expectation and variance of imputation accuracy Z by averaging over all possible genealogical types G and coalescence times T_3 and T_2.

To compute the expectation $\mathbb{E}[Z|\theta, p, t_d]$, we consider three possible scenarios that can occur when we apply rule δ to a genealogy: reference sequence R_1 is selected as the template sequence for imputation in target sequence I because $d_1 < d_2$, reference sequence R_2 is selected because $d_1 > d_2$, and a choice is made randomly between references R_1 and R_2 because $d_1 = d_2$. Let $\mathcal{S}_1$ be the scenario in which $d_1 < d_2$ (i.e., $Y_1 - Y_2 < 0$), let $\mathcal{S}_2$ be the scenario in which $d_1 > d_2$ (i.e., $Y_1 - Y_2 > 0$), and let $\mathcal{S}_3$ be the scenario in which $d_1 = d_2$ (i.e., $Y_1 - Y_2 = 0$). We can obtain $\mathbb{E}[Z|\theta, p, t_d]$ by taking a weighted average of its expectation conditional on the genealogical type g and the scenario $\mathcal{S}_w$, where $g = A, B, C, D$ and $w = 1, 2, 3$, and where the weight is the joint probability of the genealogical type $G = g$ and the scenario $\mathcal{S}_w$:

$$\mathbb{E}[Z|\theta, p, t_d] = \sum_{g=A,B,C,D} \sum_{w=1}^{3} \mathbb{E}[Z|g, \mathcal{S}_w, \theta, p, t_d]\, \mathbb{P}(g, \mathcal{S}_w|\theta, p, t_d). \tag{14}$$

We first derive the conditional expectations $\mathbb{E}[Z|g, \mathcal{S}_w, \theta, p, t_d]$ and the probabilities $\mathbb{P}(g, \mathcal{S}_w|\theta, p, t_d)$ for $g = A, B, C, D$ and $w = 1, 2, 3$, and we then obtain the expectation $\mathbb{E}[Z|\theta, p, t_d]$ using Eq. (14).

Many quantities in Sections 2.5.1 and 2.5.2 are conditioned on θ, p, and t_d, but, for notational convenience, these parameters are suppressed.

2.5.1. Derivation of $\mathbb{E}[Z|g, \mathcal{S}_w]$ in Eq. (14)

Let B be a Bernoulli random variable with parameter 1/2. For any genealogical type g, we can write the expectation $\mathbb{E}[Z|g, \mathcal{S}_w]$ under a specific scenario $\mathcal{S}_w$ for $w = 1, 2, 3$:

$$\mathbb{E}[Z|g, \mathcal{S}_w] = \begin{cases} \mathbb{E}\left[\left. \dfrac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} \right| g, \mathcal{S}_w \right] \\ \text{if } w = 1, 2, \text{ where } j = 3 - w \\ \mathbb{E}\left[B \left. \dfrac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} + (1 - B) \dfrac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} \right| g, \mathcal{S}_w \right] \\ \text{if } w = 3. \end{cases} \tag{15}$$

We make two approximations to obtain an expression for $\mathbb{E}[Z|g, \mathcal{S}_w]$. First, we use the first-order Taylor approximation that treats the expectation of a quotient as a quotient of expectations; although this approximation is not accurate in general, we will see later that, in our analysis, it is not unreasonable. Next, we approximate $\mathbb{E}[X_i - Y_i|g, \mathcal{S}_w]$ by $\mathbb{E}[X_i - Y_i|g]$; this approximation amounts to assuming for $i \in \{1, 2, 3\}$ that the number of untyped mutations on branch i is independent of which reference is closer to the target at sites genotyped in the target. Although these quantities are not independent, we will see that this assumption is also reasonable.

Applying the approximations, for $g = A, B, C, D$ and $w \in \{1, 2\}$,

$$\begin{aligned} \mathbb{E}[Z|g, \mathcal{S}_w] &= \mathbb{E}\left[\left. \frac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} \right| g, \mathcal{S}_w \right] \\ &\approx \frac{\mathbb{E}[X_j - Y_j|g, \mathcal{S}_w]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_w\right]} \\ &\approx \frac{\mathbb{E}[X_j - Y_j|g]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]}, \end{aligned} \tag{16}$$

where $j = 3 - w$. For $g = A, B, C, D$ and $w = 3$,

$$\begin{aligned} &\mathbb{E}[Z|g, \mathcal{S}_w] \\ &= \frac{1}{2}\left(\mathbb{E}\left[\left. \frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} \right| g, \mathcal{S}_w \right] + \mathbb{E}\left[\left. \frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} \right| g, \mathcal{S}_w \right] \right) \\ &\approx \frac{1}{2}\left(\frac{\mathbb{E}[X_2 - Y_2|g]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]} + \frac{\mathbb{E}[X_1 - Y_1|g]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]} \right). \end{aligned} \tag{17}$$

In Eqs. (16) and (17), for $g = A, B, C, D$ and $i = 1, 2, 3$, the expectation $\mathbb{E}[X_i - Y_i|g]$ can be found by conditioning on the coalescence times T_3 and T_2 and then integrating over their distributions:

$$\begin{aligned} &\mathbb{E}[X_i - Y_i|g] \\ &= \int_{t_2=0}^{\infty} \int_{t_3=0}^{a} \mathbb{E}[X_i - Y_i|t_3, t_2, g] f_{T_3,T_2}(t_3, t_2|g)\, \mathrm{d}t_3\, \mathrm{d}t_2. \end{aligned} \tag{18}$$

Table 2
Mean numbers of untyped mutations $\mathbb{E}[X_i - Y_i|g]$, for genealogical types $g = A, B, C, D$ and branches $i = 1, 2, 3$. The definitions of $E_\circ$, $E_\times$, $E_\diamond$, and $E_{\surd}$—each of which is a function of θ, p, and t_d—appear in Eqs. (23)–(26).

Genealogical type	Branch		
	1	2	3
A	$E_\circ$	$E_\times$	$E_\circ$
B	$E_\times$	$E_\circ$	$E_\circ$
C	$E_\circ$	$E_\circ$	$E_\times$
D	$E_\diamond$	$E_{\surd}$	$E_\diamond$

The upper limit of the inner integral depends on the genealogical type:

$$a = \begin{cases} \infty & \text{if } g = A, B, C \\ t_d & \text{if } g = D. \end{cases} \tag{19}$$

The formula for the expectation of a Poisson random variable gives

$$\mathbb{E}[X_i - Y_i|t_3, t_2, g] = h_i(\mathbf{T}; g)\,\theta(1-p)/2. \tag{20}$$

In any genealogy, by the independence of coalescence times under the coalescent model,

$$f_{T_3,T_2}(t_3, t_2|g) = f_{T_3}(t_3|g)\,f_{T_2}(t_2|g). \tag{21}$$

Using Eqs. (1)–(3),

$$f_{T_3,T_2}(t_3, t_2|g) = \begin{cases} 3e^{-3t_3-t_2} & \text{if } g = A, B, C \\ \dfrac{e^{-t_3-t_2}}{1-e^{-t_d}}\mathbf{1}_{\{t_3<t_d\}} & \text{if } g = D. \end{cases} \tag{22}$$

Considering all 12 choices of (g, i) with $g = A, B, C, D$ and $i = 1, 2, 3$, the 12 cases in Eq. (18) evaluate to four possible quantities, which we denote as follows:

$$E_\circ = \frac{1}{6}\theta(1-p)(3t_d+1) \tag{23}$$

$$E_\times = \frac{1}{6}\theta(1-p)(3t_d+7) \tag{24}$$

$$E_\diamond = \frac{1}{2}\theta(1-p)\left[\frac{1-(t_d+1)e^{-t_d}}{1-e^{-t_d}}\right] \tag{25}$$

$$E_{\surd} = \frac{1}{2}\theta(1-p)\left[\frac{2t_d+1-(t_d+1)e^{-t_d}}{1-e^{-t_d}}\right]. \tag{26}$$

Table 2 indicates which among the 12 integrals equal $E_\circ$, $E_\times$, $E_\diamond$, and $E_{\surd}$.

We can then insert the appropriate quantities from Eqs. (23)–(26) into Eqs. (16) and (17) to obtain the 12 terms $\mathbb{E}[Z|g, \mathcal{S}_w]$ that appear in Eq. (14) (Table 3). This completes the derivation of $\mathbb{E}[Z|g, \mathcal{S}_w]$.

2.5.2. Derivation of $\mathbb{P}(g, \mathcal{S}_w)$ in Eq. (14)

We obtain the probability $\mathbb{P}(g, \mathcal{S}_w)$ by jointly considering the marginal distribution of g and the conditional distribution of $\mathcal{S}_w$ given a genealogical type $G = g$:

$$\mathbb{P}(g, \mathcal{S}_w) = \mathbb{P}(g)\,\mathbb{P}(\mathcal{S}_w|g). \tag{27}$$

The probability $\mathbb{P}(g)$ is given in Eq. (4). As in the derivation of $\mathbb{E}[X_i - Y_i|g]$, to compute $\mathbb{P}(\mathcal{S}_w|g)$, we condition on the coalescence times T_3 and T_2 and then integrate over their distributions:

$$\mathbb{P}(\mathcal{S}_w|g) = \int_{t_2=0}^{\infty}\int_{t_3=0}^{a} \mathbb{P}(\mathcal{S}_w|t_3, t_2, g)\,f_{T_3,T_2}(t_3, t_2|g)\,dt_3\,dt_2, \tag{28}$$

Table 3
The computation of $\mathbb{E}[Z|g, \mathcal{S}_w]$ for $g = A, B, C, D$ and $w = 1, 2, 3$. Each initial expression is obtained using Eqs. (16) and (17), and then simplified using Eqs. (23)–(26).

Quantity	Initial expression	Simplified expression
$\mathbb{E}[Z\|A, \mathcal{S}_1]$	$\frac{E_\times}{2E_\circ+E_\times}$	$\frac{3t_d+7}{9t_d+9}$
$\mathbb{E}[Z\|B, \mathcal{S}_1]$	$\frac{E_\circ}{2E_\circ+E_\times}$	$\frac{3t_d+1}{9t_d+9}$
$\mathbb{E}[Z\|C, \mathcal{S}_1]$	$\frac{E_\circ}{2E_\circ+E_\times}$	$\frac{3t_d+1}{9t_d+9}$
$\mathbb{E}[Z\|D, \mathcal{S}_1]$	$\frac{E_{\surd}}{2E_\diamond+E_{\surd}}$	$\frac{2t_d+1-(t_d+1)e^{-t_d}}{2t_d+3-3(t_d+1)e^{-t_d}}$
$\mathbb{E}[Z\|A, \mathcal{S}_2]$	$\frac{E_\circ}{2E_\circ+E_\times}$	$\frac{3t_d+1}{9t_d+9}$
$\mathbb{E}[Z\|B, \mathcal{S}_2]$	$\frac{E_\times}{2E_\circ+E_\times}$	$\frac{3t_d+7}{9t_d+9}$
$\mathbb{E}[Z\|C, \mathcal{S}_2]$	$\frac{E_\circ}{2E_\circ+E_\times}$	$\frac{3t_d+1}{9t_d+9}$
$\mathbb{E}[Z\|D, \mathcal{S}_2]$	$\frac{E_\diamond}{2E_\diamond+E_{\surd}}$	$\frac{1-(t_d+1)e^{-t_d}}{2t_d+3-3(t_d+1)e^{-t_d}}$
$\mathbb{E}[Z\|A, \mathcal{S}_3]$	$\frac{E_\circ+E_\times}{2(2E_\circ+E_\times)}$	$\frac{3t_d+4}{9t_d+9}$
$\mathbb{E}[Z\|B, \mathcal{S}_3]$	$\frac{E_\circ+E_\times}{2(2E_\circ+E_\times)}$	$\frac{3t_d+4}{9t_d+9}$
$\mathbb{E}[Z\|C, \mathcal{S}_3]$	$\frac{E_\circ}{2E_\circ+E_\times}$	$\frac{3t_d+1}{9t_d+9}$
$\mathbb{E}[Z\|D, \mathcal{S}_3]$	$\frac{E_\diamond+E_{\surd}}{2(2E_\diamond+E_{\surd})}$	$\frac{(t_d+1)(1-e^{-t_d})}{2t_d+3-3(t_d+1)e^{-t_d}}$

where $f_{T_3,T_2}(t_3, t_2|g)$ is calculated with Eq. (22) and a is given in Eq. (19). In principle, $\mathbb{P}(\mathcal{S}_w|t_3, t_2, g)$ can be obtained by considering the difference $Y_1 - Y_2$ and using Eq. (11):

$$\mathbb{P}(\mathcal{S}_1|t_3, t_2, g) = \sum_{y_1=0}^{\infty}\sum_{y_2=y_1+1}^{\infty} f_{\text{Sk}}(y_1-y_2;\, h_1(\mathbf{T}; g)\,\theta p/2,\, h_2(\mathbf{T}; g)\,\theta p/2) \tag{29}$$

$$\mathbb{P}(\mathcal{S}_2|t_3, t_2, g) = \sum_{y_2=0}^{\infty}\sum_{y_1=y_2+1}^{\infty} f_{\text{Sk}}(y_1-y_2;\, h_1(\mathbf{T}; g)\,\theta p/2,\, h_2(\mathbf{T}; g)\,\theta p/2) \tag{30}$$

$$\mathbb{P}(\mathcal{S}_3|t_3, t_2, g) = f_{\text{Sk}}(0;\, h_1(\mathbf{T}; g)\,\theta p/2,\, h_2(\mathbf{T}; g)\,\theta p/2). \tag{31}$$

In practice, the sums are unwieldy, and, instead of computing them, we use a Monte Carlo approach to evaluate Eq. (28), as described in Section 3. This completes the derivation of the expectation $\mathbb{E}[Z|\theta, p, t_d]$ in Eq. (14).

2.5.3. Derivation of $\text{Var}[Z|\theta, p, t_d]$

$\text{Var}[Z|\theta, p, t_d]$ is obtained as

$$\text{Var}[Z|\theta, p, t_d] = \mathbb{E}[Z^2|\theta, p, t_d] - \mathbb{E}[Z|\theta, p, t_d]^2, \tag{32}$$

where $\mathbb{E}[Z|\theta, p, t_d]$ has already been derived (Eq. (14)). It remains to evaluate $\mathbb{E}[Z^2|\theta, p, t_d]$.

As in the derivation of $\mathbb{E}[Z|\theta, p, t_d]$, we obtain $\mathbb{E}[Z^2|\theta, p, t_d]$ by conditioning on the genealogical type g and the scenario $\mathcal{S}_w$:

$$\mathbb{E}[Z^2|\theta, p, t_d] = \sum_{g=A,B,C,D}\sum_{w=1}^{3}\mathbb{E}[Z^2|g, \mathcal{S}_w, \theta, p, t_d]\,\mathbb{P}(g, \mathcal{S}_w|\theta, p, t_d). \tag{33}$$

For $g = A, B, C, D$ and $w = 1, 2, 3$,

$$\mathbb{E}[Z^2|g, \mathcal{S}_w, \theta, p, t_d] = \text{Var}[Z|g, \mathcal{S}_w, \theta, p, t_d] + \mathbb{E}[Z|g, \mathcal{S}_w, \theta, p, t_d]^2. \tag{34}$$

We again suppress θ, p, and t_d. The probability $\mathbb{P}(g, \mathcal{S}_w)$ in Eq. (33) and the expectation $\mathbb{E}[Z|g, \mathcal{S}_w]$ in Eq. (34) have already been derived (Eqs. (27), (16), (17)). To obtain an expression for $\text{Var}[Z|g, \mathcal{S}_w]$ in Eq. (34), we apply the same two approximations used for obtaining $\mathbb{E}[Z|g, \mathcal{S}_w]$ in Section 2.5.1. The first-order

Taylor-series approximation for the variance $\mathrm{Var}[Z|g,\mathcal{S}_w]$ (Casella and Berger, 2001, p. 245) is followed by an additional approximation that disregards the dependence on $\mathcal{S}_w$.

Applying the approximations, for $g = A, B, C, D$ and $w \in \{1, 2\}$,

$$\begin{aligned}
\mathrm{Var}[Z|g,\mathcal{S}_w] &= \mathrm{Var}\left[\left.\frac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_w\right] \\
&\approx \left(\frac{\mathbb{E}[X_j - Y_j|g,\mathcal{S}_w]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w\right]}\right)^2 \\
&\quad \times \left(\frac{\mathrm{Var}[X_j - Y_j|g,\mathcal{S}_w]}{\mathbb{E}[X_j - Y_j|g,\mathcal{S}_w]^2} + \frac{\mathrm{Var}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w\right]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w\right]^2}\right. \\
&\quad \left. - \frac{2\mathrm{Cov}\left(X_j - Y_j, \sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w\right)}{\mathbb{E}[X_j - Y_j|g,\mathcal{S}_w]\,\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w\right]}\right) \\
&\approx \left(\frac{\mathbb{E}[X_j - Y_j|g]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]}\right)^2 \\
&\quad \times \left(\frac{\mathrm{Var}[X_j - Y_j|g]}{\mathbb{E}[X_j - Y_j|g]^2} + \frac{\mathrm{Var}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]}{\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]^2}\right. \\
&\quad \left. - \frac{2\mathrm{Cov}\left(X_j - Y_j, \sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right)}{\mathbb{E}[X_j - Y_j|g]\,\mathbb{E}\left[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g\right]}\right), \qquad (35)
\end{aligned}$$

where $j = 3 - w$. For $g = A, B, C, D$ and $w = 3$,

$$\begin{aligned}
&\mathrm{Var}[Z|g,\mathcal{S}_w] \\
&= \mathrm{Var}\left[\left. B\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} + (1-B)\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_w\right] \\
&= \mathbb{E}\left[\left.\left(B\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} + (1-B)\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right)^2\right| g, \mathcal{S}_w\right] \\
&\quad - \mathbb{E}\left[\left. B\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} + (1-B)\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_w\right]^2. \qquad (36)
\end{aligned}$$

Because B is Bernoulli, $B^2 = B$, $(1-B)^2 = (1-B)$, and $B(1-B) = 0$. Using the independence of B from the imputation accuracy Z, we can simplify Eq. (36) to

$$\begin{aligned}
\mathrm{Var}[Z|g,\mathcal{S}_3] &= \frac{1}{2}\left(\mathrm{Var}\left[\left.\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_3\right]\right. \\
&\quad \left. + \mathrm{Var}\left[\left.\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_3\right]\right) \\
&\quad + \frac{1}{4}\left(\mathbb{E}\left[\left.\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_3\right]\right. \\
&\quad \left. - \mathbb{E}\left[\left.\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\right| g, \mathcal{S}_3\right]\right)^2. \qquad (37)
\end{aligned}$$

When $w = 1, Z = (X_2 - Y_2)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)$, and when $w = 2, Z = (X_1 - Y_1)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)$ (Eqs. (12) and (13)). As we have elsewhere used an approximation that, conditional on a genealogical type g, terms $X_i - Y_i$ do not depend on whether w is equal to 1, 2, or 3 (Eqs. (16) and (17)), we here make a similar approximation that, conditional on g, $(X_i - Y_i)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)$ does not depend on $\mathcal{S}_w$. Thus, instead of conditioning on $w = 3$, we can condition on $w = 1$ or $w = 2$. If $w = 3$ is replaced with $w = 1$ for terms $\mathbb{E}[(X_2 - Y_2)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w]$ and $\mathrm{Var}[(X_2 - Y_2)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w]$, and with $w = 2$ for terms $\mathbb{E}[(X_1 - Y_1)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w]$ and $\mathrm{Var}[(X_1 - Y_1)/\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g,\mathcal{S}_w]$, we obtain

$$\begin{aligned}
\mathrm{Var}[Z|g,\mathcal{S}_3] &\approx \frac{\mathrm{Var}[Z|g,\mathcal{S}_1] + \mathrm{Var}[Z|g,\mathcal{S}_2]}{2} \\
&\quad + \frac{(\mathbb{E}[Z|g,\mathcal{S}_1] - \mathbb{E}[Z|g,\mathcal{S}_2])^2}{4}. \qquad (38)
\end{aligned}$$

This approximation reduces $\mathrm{Var}[Z|g,\mathcal{S}_3]$ to simpler computations conditional on $\mathcal{S}_1$ and $\mathcal{S}_2$.

In Eq. (35), for $g = A, B, C, D$ and $i = 1, 2, 3$, the expectation $\mathbb{E}[X_i - Y_i|g]$ is computed using Eqs. (18)–(19), and the variance $\mathrm{Var}[X_i - Y_i|g]$ can be found using the conditional variance identity (Casella and Berger, 2001, p. 167). Applying the identity, conditioning on the coalescence times T_3 and T_2, and integrating over the coalescence-time distributions, we have

$$\begin{aligned}
\mathrm{Var}[X_i - Y_i|g] &= \mathbb{E}[\mathrm{Var}[X_i - Y_i|t_3, t_2, g]] \\
&\quad + \mathrm{Var}[\mathbb{E}[X_i - Y_i|t_3, t_2, g]] \\
&= \int_{t_2=0}^{\infty}\int_{t_3=0}^{a} \mathrm{Var}[X_i - Y_i|t_3, t_2, g] \\
&\quad \times f_{T_3,T_2}(t_3, t_2|g)\, dt_3\, dt_2 \\
&\quad + \int_{t_2=0}^{\infty}\int_{t_3=0}^{a} (\mathbb{E}[X_i - Y_i|t_3, t_2, g] \\
&\quad - \mathbb{E}[X_i - Y_i|g])^2 f_{T_3,T_2}(t_3, t_2|g)\, dt_3\, dt_2. \qquad (39)
\end{aligned}$$

In Eq. (39), the upper limit a of the inner integral is given in Eq. (19), and the joint density function $f_{T_3,T_2}(t_3, t_2|g)$ is evaluated using

Table 4
Variances of the number of untyped mutations $\mathrm{Var}[X_i - Y_i|g]$, for genealogical types $g = A, B, C, D$ and branches $i = 1, 2, 3$. The definitions of $V_\circ$, $V_\times$, $V_\diamond$, and $V_{\surd}$ – each of which is a function of θ, p, and t_d – appear in Eqs. (41)–(44).

Genealogical type	Branch		
	1	2	3
A	$V_\circ$	$V_\times$	$V_\circ$
B	$V_\times$	$V_\circ$	$V_\circ$
C	$V_\circ$	$V_\circ$	$V_\times$
D	$V_\diamond$	$V_{\surd}$	$V_\diamond$

Eq. (22). The formula for the variance of a Poisson random variable gives,

$$\begin{aligned}\mathrm{Var}[X_i - Y_i|t_3, t_2, g] &= \mathbb{E}[X_i - Y_i|t_3, t_2, g] \\ &= h_i(\mathbf{T}; g)\,\theta(1-p)/2, \end{aligned} \tag{40}$$

where $h_i(\mathbf{T}; g)$ appears in Table 1. Consequently, in Eq. (39), the first term is $\mathbb{E}[\mathrm{Var}[X_i - Y_i|t_3, t_2, g]] = \mathbb{E}[\mathbb{E}[X_i - Y_i|t_3, t_2, g]] = \mathbb{E}[X_i - Y_i|g]$, which has already been obtained in Eq. (18).

For the second term in Eq. (39), the expectations $\mathbb{E}[X_i - Y_i|t_3, t_2, g]$ and $\mathbb{E}[X_i - Y_i|g]$ are given in Eqs. (20) and (18), respectively. Considering all 12 choices of (g, i) with $g = A, B, C, D$ and $i = 1, 2, 3$, the 12 cases in Eq. (39) evaluate to four possible quantities:

$$V_\circ = \frac{1}{6}\theta(1-p)(3t_d+1) + \frac{1}{36}\theta^2(1-p)^2 \tag{41}$$

$$V_\times = \frac{1}{6}\theta(1-p)(3t_d+7) + \frac{37}{36}\theta^2(1-p)^2 \tag{42}$$

$$\begin{aligned} V_\diamond = {} & \frac{1}{2}\theta(1-p)\left[\frac{1-(t_d+1)e^{-t_d}}{1-e^{-t_d}}\right] \\ & + \frac{1}{4}\theta^2(1-p)^2\left[\frac{1-2e^{-t_d}-t_d^2e^{-t_d}+e^{-2t_d}}{1-2e^{-t_d}+e^{-2t_d}}\right] \end{aligned} \tag{43}$$

$$\begin{aligned} V_{\surd} = {} & \frac{1}{2}\theta(1-p)\left[\frac{2t_d+1-(t_d+1)e^{-t_d}}{1-e^{-t_d}}\right] \\ & + \frac{1}{4}\theta^2(1-p)^2\left[\frac{5-10e^{-t_d}-t_d^2e^{-t_d}+5e^{-2t_d}}{1-2e^{-t_d}+e^{-2t_d}}\right]. \end{aligned} \tag{44}$$

Table 4 indicates which among the 12 integrals equal $V_\circ$, $V_\times$, $V_\diamond$, and $V_{\surd}$.

To complete the calculation of $\mathrm{Var}[Z|g, \mathcal{S}_w]$, we compute the covariance $\mathrm{Cov}(X_j - Y_j, \sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g)$ in Eq. (35). For any genealogy, conditional on the coalescence times T_3 and T_2, $(X_i - Y_i)$ and $(X_j - Y_j)$ are independent for any i and j with $j \neq i$. Then

$$\mathrm{Cov}\left(X_j - Y_j, \sum_{\ell=1}^{3}(X_\ell - Y_\ell)\middle|\, g\right) = \mathrm{Var}[X_j - Y_j|g], \tag{45}$$

which has been obtained in Eq. (39).

For $g = A, B, C, D$, we obtain the eight terms $\mathrm{Var}[Z|g, \mathcal{S}_1]$ and $\mathrm{Var}[Z|g, \mathcal{S}_2]$ in Eq. (33) by inserting the appropriate quantities from Eqs. (23)–(26) and (41)–(44) into Eq. (35). We obtain the remaining four terms $\mathrm{Var}[Z|g, \mathcal{S}_3]$ using Eq. (38) (Table 5). The resulting quantities are unwieldy, and we do not list the full expressions for $\mathrm{Var}[Z|g, \mathcal{S}_w]$. This completes the derivation of $\mathbb{E}[Z^2|\theta, p, t_d]$ and thus of $\mathrm{Var}[Z|\theta, p, t_d]$.

Table 5
The computation of $\mathrm{Var}[Z|g, \mathcal{S}_w]$ for $g = A, B, C, D$ and $w = 1, 2, 3$. Each initial expression is obtained using Eqs. (35) and (38), and then simplified using Eqs. (23)–(26) and (41)–(44). We omit the simplified expressions, as they are unwieldy.

Quantity	Initial expression
$\mathrm{Var}[Z|A, \mathcal{S}_1]$	$\left(\frac{E_\times}{2E_\circ+E_\times}\right)^2\left[\frac{V_\times}{E_\times^2} + \frac{2V_\circ+V_\times}{(2E_\circ+E_\times)^2} - \frac{2V_\times}{E_\times(2E_\circ+E_\times)}\right]$
$\mathrm{Var}[Z|B, \mathcal{S}_1]$	$\left(\frac{E_\circ}{2E_\circ+E_\times}\right)^2\left[\frac{V_\circ}{E_\circ^2} + \frac{2V_\circ+V_\times}{(2E_\circ+E_\times)^2} - \frac{2V_\circ}{E_\circ(2E_\circ+E_\times)}\right]$
$\mathrm{Var}[Z|C, \mathcal{S}_1]$	$\left(\frac{E_\circ}{2E_\circ+E_\times}\right)^2\left[\frac{V_\circ}{E_\circ^2} + \frac{2V_\circ+V_\times}{(2E_\circ+E_\times)^2} - \frac{2V_\circ}{E_\circ(2E_\circ+E_\times)}\right]$
$\mathrm{Var}[Z|D, \mathcal{S}_1]$	$\left(\frac{E_{\surd}}{2E_\diamond+E_{\surd}}\right)^2\left[\frac{V_{\surd}}{E_{\surd}^2} + \frac{2V_\diamond+V_{\surd}}{(2E_\diamond+E_{\surd})^2} - \frac{2V_{\surd}}{E_{\surd}(2E_\diamond+E_{\surd})}\right]$
$\mathrm{Var}[Z|A, \mathcal{S}_2]$	$\left(\frac{E_\circ}{2E_\circ+E_\times}\right)^2\left[\frac{V_\circ}{E_\circ^2} + \frac{2V_\circ+V_\times}{(2E_\circ+E_\times)^2} - \frac{2V_\circ}{E_\circ(2E_\circ+E_\times)}\right]$
$\mathrm{Var}[Z|B, \mathcal{S}_2]$	$\left(\frac{E_\times}{2E_\circ+E_\times}\right)^2\left[\frac{V_\times}{E_\times^2} + \frac{2V_\circ+V_\times}{(2E_\circ+E_\times)^2} - \frac{2V_\times}{E_\times(2E_\circ+E_\times)}\right]$
$\mathrm{Var}[Z|C, \mathcal{S}_2]$	$\left(\frac{E_\circ}{2E_\circ+E_\times}\right)^2\left[\frac{V_\circ}{E_\circ^2} + \frac{2V_\circ+V_\times}{(2E_\circ+E_\times)^2} - \frac{2V_\circ}{E_\circ(2E_\circ+E_\times)}\right]$
$\mathrm{Var}[Z|D, \mathcal{S}_2]$	$\left(\frac{E_\diamond}{2E_\diamond+E_{\surd}}\right)^2\left[\frac{V_\diamond}{E_\diamond^2} + \frac{2V_\diamond+V_{\surd}}{(2E_\diamond+E_{\surd})^2} - \frac{2V_\diamond}{E_\diamond(2E_\diamond+E_{\surd})}\right]$
$\mathrm{Var}[Z|A, \mathcal{S}_3]$	$\frac{\mathrm{Var}[Z|A,\mathcal{S}_1]+\mathrm{Var}[Z|A,\mathcal{S}_2]}{2} + \frac{(\mathbb{E}[Z|A,\mathcal{S}_1]-\mathbb{E}[Z|A,\mathcal{S}_2])^2}{4}$
$\mathrm{Var}[Z|B, \mathcal{S}_3]$	$\frac{\mathrm{Var}[Z|B,\mathcal{S}_1]+\mathrm{Var}[Z|B,\mathcal{S}_2]}{2} + \frac{(\mathbb{E}[Z|B,\mathcal{S}_1]-\mathbb{E}[Z|B,\mathcal{S}_2])^2}{4}$
$\mathrm{Var}[Z|C, \mathcal{S}_3]$	$\frac{\mathrm{Var}[Z|C,\mathcal{S}_1]+\mathrm{Var}[Z|C,\mathcal{S}_2]}{2} + \frac{(\mathbb{E}[Z|C,\mathcal{S}_1]-\mathbb{E}[Z|C,\mathcal{S}_2])^2}{4}$
$\mathrm{Var}[Z|D, \mathcal{S}_3]$	$\frac{\mathrm{Var}[Z|D,\mathcal{S}_1]+\mathrm{Var}[Z|D,\mathcal{S}_2]}{2} + \frac{(\mathbb{E}[Z|D,\mathcal{S}_1]-\mathbb{E}[Z|D,\mathcal{S}_2])^2}{4}$

3. Methods of computation and simulation

To calculate the expectation $\mathbb{E}[Z|\theta, p, t_d]$ (Eq. (14)), we computed $\mathbb{E}[Z|g, \mathcal{S}_w, \theta, p, t_d]$ using Eqs. (16) and (17). We obtained Monte Carlo estimates of $\mathbb{P}(\mathcal{S}_w|g, \theta, p, t_d)$ included in the expression for $\mathbb{P}(g, \mathcal{S}_w|\theta, p, t_d)$, using 10^5 draws from the Skellam distribution defined in Eq. (11). Each of these draws was obtained by first sampling t_3 and t_2 from their respective distributions, conditional on g (and t_d). Next, we evaluated the difference between two simulated Poisson random variables, with parameters $h_1(\mathbf{T}; g)\,\theta p/2$ and $h_2(\mathbf{T}; g)\,\theta p/2$, respectively. These Poisson variates were sampled using the GNU Scientific Library function `gsl_ran_Poisson`. Thus, the expectation was obtained using three approximations: a Taylor approximation, an approximation that disregards a dependence on $\mathcal{S}_w$, and a Monte Carlo approximation for integrals associated with the Skellam distribution.

The computation of the variance $\mathrm{Var}[Z|\theta, p, t_d]$ used some of the same approximations used in evaluating the mean $\mathbb{E}[Z|\theta, p, t_d]$. The variance computation incorporated the two steps of approximation for $\mathbb{E}[Z|g, \mathcal{S}_w, \theta, p, t_d]$. Additionally, the same Monte Carlo samples of $\mathbb{P}(\mathcal{S}_w|g, \theta, p, t_d)$ employed in evaluating the mean were used in the variance computation. Beyond the Taylor approximation and omission of the conditioning on $\mathcal{S}_w$ that were required in obtaining the mean, the variance computation applied corresponding approximations in obtaining $\mathbb{E}[Z^2|g, \mathcal{S}_w, \theta, p, t_d]$.

Given θ, p, and t_d, we also performed stochastic simulations under the coalescent to estimate the mean and the variance of the imputation accuracy by summing over all simulations, employing Monte Carlo integration as described in Box 2 with $M = 10^5$ simulation replicates. To verify the expressions for $\mathbb{E}[Z|\theta, p, t_d]$ and $\mathrm{Var}[Z|\theta, p, t_d]$ in Eqs. (14) and (32), we then compared the simulated means and variances of the imputation accuracy to our formula-based estimates.

Fig. 3 shows the mean and the variance of the imputation accuracy computed both using our formulas and using the simulations. For the parameter values that we considered, the theoretical approximations of $\mathbb{E}[Z|\theta, p, t_d]$ obtained using Eq. (14) closely match the simulated mean imputation accuracy (Fig. 3, top row). Except when θ is small and p is large, the theoretical estimates of $\mathrm{Var}[Z|\theta, p, t_d]$ obtained using Eq. (32) closely match the simulated variances (Fig. 3, bottom row).

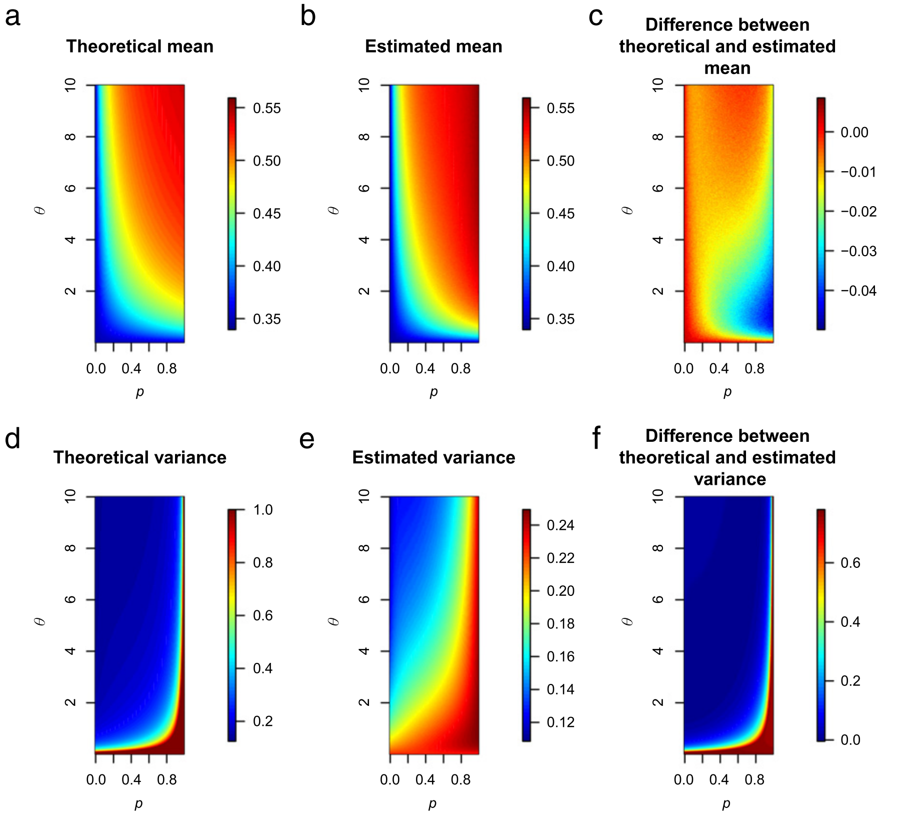


Fig. 3. Mean and variance of imputation accuracy as functions of the mutation parameter θ and the proportion p of polymorphic sites that are genotyped in the target. (A) Mean imputation accuracy obtained by Eq. (14). (B) Mean imputation accuracy obtained using the simulation algorithm in Box 2. (C) The difference between (A) and (B). (D) Variance of imputation accuracy obtained by Eq. (32). (E) Variance of imputation accuracy obtained using the simulation algorithm in Box 2. (F) The difference between (D) and (E). For all plots, t_d is fixed at 0.1. In part (D), values of the approximate theoretical variance that exceed 1 are set to 1.

Box 2. Simulation algorithm for estimating $\mathbb{E}[Z|\theta, p, t_d]$ ***and*** $\mathrm{Var}[Z|\theta, p, t_d]$

```
1. Set parameter values for $\theta$, $p$, and $t_d$.
2. For $m = 1$ to $M$:
   (a) Generate a genealogical type $G$ using a uniformly distributed random variable $U \sim \mathrm{Uniform}(0, 1)$. If $u < 1 - e^{-t_d}$, set $g = D$. Otherwise, generate $u'$ from Uniform(0, 1), independently of $u$. Set $g = A$ if $u' \in (0, 1/3)$, $g = B$ if $u' \in [1/3, 2/3)$, and $g = C$ if $u' \in [2/3, 1)$.
   (b) Generate a coalescence time $T_2 \sim \mathrm{Exp}(1)$.
   (c) If $g \in \{A, B, C\}$, generate a coalescence time $T_3 \sim \mathrm{Exp}(3)$. Otherwise, generate $T_3$ from the probability density function in eq. 3.
   (d) For $i = 1, 2, 3$, generate a total number of mutations $X_i \sim \mathrm{Poisson}(h_i(\mathbf{T}; g)\,\theta/2)$ on branch $i$, where $\mathbf{T} = (t_3, t_2, t_d)$ and $h_i(\mathbf{T}; g)$ is specified in Table 1.
   (e) For $i = 1, 2, 3$, given $X_i$, sample the number of mutations on branch $i$ that are genotyped in the target as $Y_i|X_i = x_i \sim \mathrm{Binomial}(x_i, p)$.
   (f) If $\sum_{i=1}^{3}(x_i - y_i) = 0$, return to (a); otherwise, continue.
   (g) If $y_1 - y_2 < 0$ (i.e., if $d_1 < d_2$), compute $z_{(m)} = \frac{x_2 - y_2}{\sum_{\ell=1}^{3}(x_\ell - y_\ell)}$.
   (h) If $y_1 - y_2 > 0$ (i.e., if $d_2 < d_1$), compute $z_{(m)} = \frac{x_1 - y_1}{\sum_{\ell=1}^{3}(x_\ell - y_\ell)}$.
   (i) If $y_1 - y_2 = 0$ (i.e., if $d_1 = d_2$), generate $B \sim \mathrm{Bernoulli}(1/2)$. Compute $z_{(m)} = \frac{x_2 - y_2}{\sum_{\ell=1}^{3}(x_\ell - y_\ell)}$ if $b = 1$ and $z_{(m)} = \frac{x_1 - y_1}{\sum_{\ell=1}^{3}(x_\ell - y_\ell)}$ if $b = 0$.
3. Compute the sample mean $\bar{z} = \frac{1}{M}\sum_{m=1}^{M} z_{(m)}$ and the sample variance $s^2 = \frac{1}{M-1}\sum_{m=1}^{M}(z_{(m)} - \bar{z})^2$ that respectively represent simulation-based estimates of $\mathbb{E}[Z|\theta, p, t_d]$ and $\mathrm{Var}[Z|\theta, p, t_d]$.
```

4. The role of the parameters

As the formulas in Eqs. (14) and (32) provide reasonable approximations to the mean and the variance of the imputation accuracy, we next examined the effects of the parameters on the mean and the variance.

4.1. The mutation parameter θ

Fig. 4(A) shows the mean imputation accuracy as a function of θ, demonstrating that, for each of three nonzero choices of p at a fixed t_d, the mean imputation accuracy increases with θ. As θ increases, more mutations are likely to occur, and more sites are genotyped in the target. Consequently, the genotyped sites are less likely to suggest a misleading closer relationship between the target sequence and one reference sequence when the other reference is in fact more genetically similar to the target. The probability of making an incorrect decision with the genotyped sites, and subsequently using the reference sequence that has larger pairwise distance to the target sequence, decreases with increasing θ.

When $p = 0$, no sites are genotyped in the target, and no information exists for deciding which reference sequence is genetically more similar to the target. Each reference sequence might be selected with equal probability, independent of θ, and the mean imputation accuracy does not depend on θ.

For fixed p and t_d, the variance of imputation accuracy decreases with increasing θ (Fig. 5(A)). An increase in θ increases the number of polymorphic sites that need to be imputed in the target, increasing the sample size for the imputation process, and thereby producing a smaller variance in imputation accuracy (Fig. 5(A)).

4.2. The proportion p of polymorphic sites that are genotyped in the target

For θ and t_d fixed, the mean imputation accuracy increases with increasing p (Fig. 4(B)). As in the case of θ, increasing p increases the number of polymorphic sites available in the target for identifying

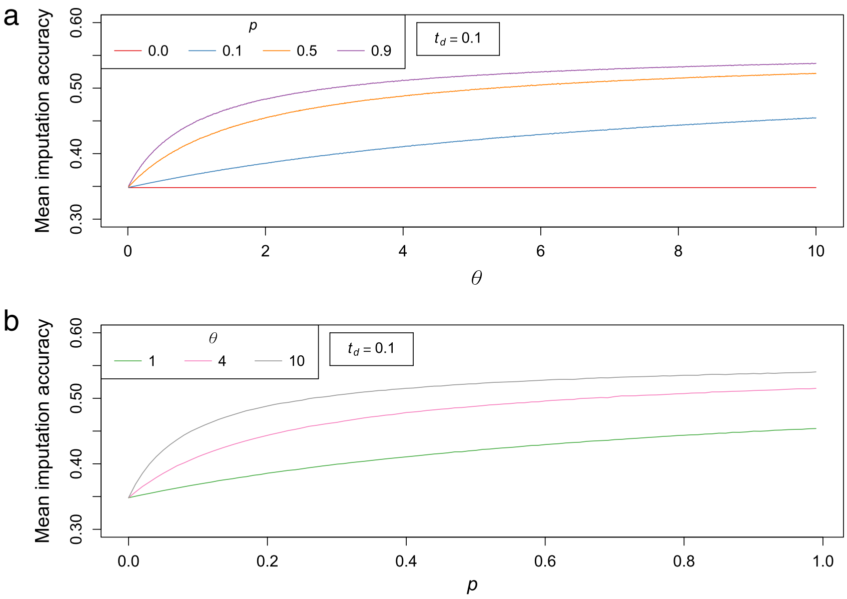


Fig. 4. Mean imputation accuracy for various values of the mutation parameter θ and the proportion p of polymorphic sites that are genotyped in the target sequence. For t_d fixed at 0.1, we obtained $\mathbb{E}[Z|\theta, p, t_d]$ using Eq. (14). (A) Mean imputation accuracy as a function of θ for fixed values of p. (B) Mean imputation accuracy as a function of p for fixed values of θ. The values plotted in both panels are extracted from Fig. 3(A).

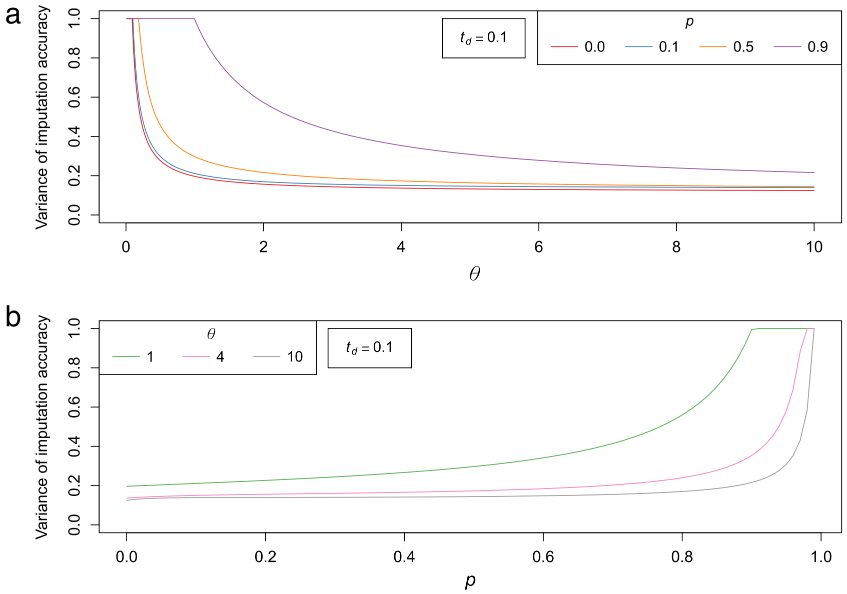


Fig. 5. Variance of imputation accuracy for various values of the mutation parameter θ and the proportion p of polymorphic sites that are genotyped in the target sequence. For t_d fixed at 0.1, we obtained $\mathrm{Var}[Z|\theta, p, t_d]$ using Eq. (32). (A) Variance of imputation accuracy as a function of θ for fixed values of p. (B) Variance of imputation accuracy as a function of p for fixed values of θ. The values plotted in both panels are extracted from Fig. 3(D).

the reference sequence that it most closely resembles. Thus, an increase in p increases the probability that the more appropriate of the two reference sequences is identified.

Unlike in the case of θ, the variance of imputation accuracy increases with increasing p (Fig. 5(B)). Increasing p reduces the number of polymorphic sites imputed in the target sequence, decreasing the sample size for the imputation process, and thereby producing a larger variance in imputation accuracy (Fig. 5(B)).

4.3. The divergence time t_d

Fig. 6 shows the mean imputation accuracy as a function of t_d for different values of θ and p. Both for θ fixed and for p fixed, the mean imputation accuracy increases as t_d increases. Increasing t_d enables more mutations to accumulate along branch 2 of the genealogy, increasing the probability that references R_1 and R_2 differ substantially in their similarity to target I. The sites genotyped in I are then more likely to correctly identify the more suitable reference sequence to serve as an imputation template, and thus, to lead to a greater imputation accuracy.

The variance of imputation accuracy exhibits a slight decrease as a function of t_d, for fixed θ and p (Fig. 7). Increasing t_d lengthens the genealogy of the three sequences, so that more mutations occur along its branches. The larger number of polymorphic sites then leads to a greater sample size in the imputation process, and consequently, to a smaller variance in imputation accuracy.

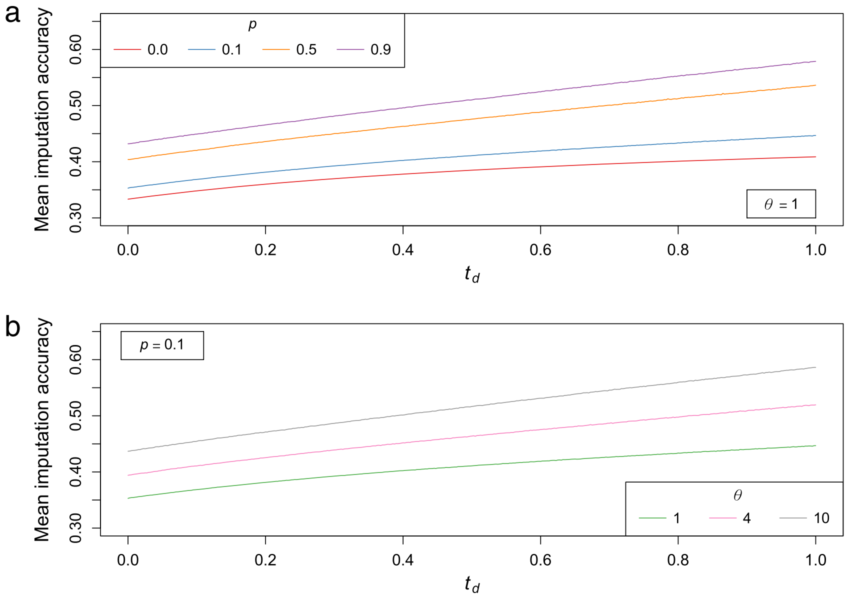


Fig. 6. Mean imputation accuracy as a function of divergence time t_d. For fixed values of θ and p, we obtained $\mathbb{E}[Z|\theta, p, t_d]$ using Eq. (14). (A) Mean imputation accuracy as a function of t_d for fixed values of p, with $\theta = 1$. (B) Mean imputation accuracy as a function of t_d for fixed values of θ, with $p = 0.1$.

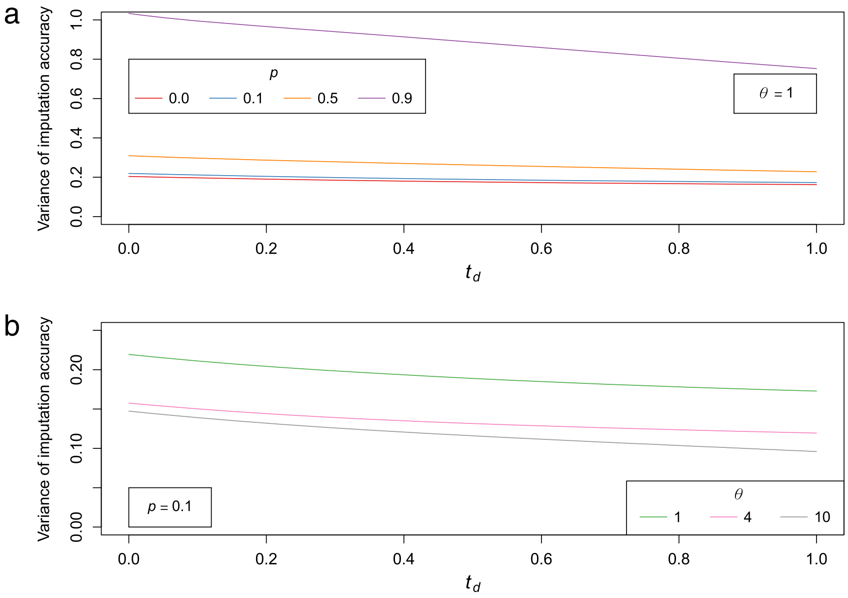


Fig. 7. Variance of imputation accuracy as a function of divergence time t_d. For fixed values of θ and p, we obtained $\mathrm{Var}[Z|\theta, p, t_d]$ using Eq. (32). (A) Variance of imputation accuracy as a function of t_d for fixed values of p, with $\theta = 1$. (B) Variance of imputation accuracy as a function of t_d for fixed values of θ, with $p = 0.1$.

4.4. Expected gain in imputation accuracy from the use of an internal reference

Our numerical studies have found that the choice of parameter values affects the choice of reference sequence, which in turn affects the imputation accuracy. To assess the expected gain in imputation accuracy in a target sequence by using R_1 rather than R_2, we plotted simulated mean imputation accuracies in the target for each of several scenarios with the reference sequence fixed (Fig. 8(A)). We modified the simulation procedure, replacing steps (g)–(i) by a single step, in which $z_{(m)} = (x_2 - y_2)/[\sum_{\ell=1}^{3}(x_\ell - y_\ell)]$ is computed for the case in which R_1 is used as the template and $z_{(m)} = (x_1 - y_1)/[\sum_{i=1}^{3}(x_\ell - y_\ell)]$ is computed if R_2 is used. The difference in mean accuracies between the imputations performed using R_1 and those performed using R_2 is largely constant across a wide range of values for θ and p (Fig. 8(B)).

Denote the expected difference in accuracy—mean imputation accuracy based on R_1 minus mean imputation accuracy based on R_2—by Δ_{t_d}. Averaging across 40 mean differences, considering (θ, p) with θ in $\{1, 2, \ldots, 10\}$ and p in $\{0, 0.1, 0.5, 0.9\}$, $\Delta_{0.1}$ has mean 0.0829, and $\Delta_{0.01}$ has mean 0.0099. We can analytically estimate this expected difference quite accurately using a simple formula. In particular, Δ_{t_d} can be computed by a weighted sum of its conditional expectations given the four values of g, with weights equal to the values of $\mathbb{P}(G = g|t_d)$. The cases $g = A$ and $g = B$ are equiprobable, and differ only in that the roles of R_1 and R_2 are exchanged in the genealogy. Thus, the contributions of these cases to Δ_{t_d} sum to 0. If $g = C$, R_1 and R_2 have the same genealogical distance from the target I, and the value of Δ_{t_d} is 0. Thus, only the case when $g = D$ contributes to the difference in mean imputation accuracy between references R_1 and R_2.

For $g = D$, the fraction of sites imputed correctly with reference R_1 is the fraction of sites due to mutations on branch 2, and

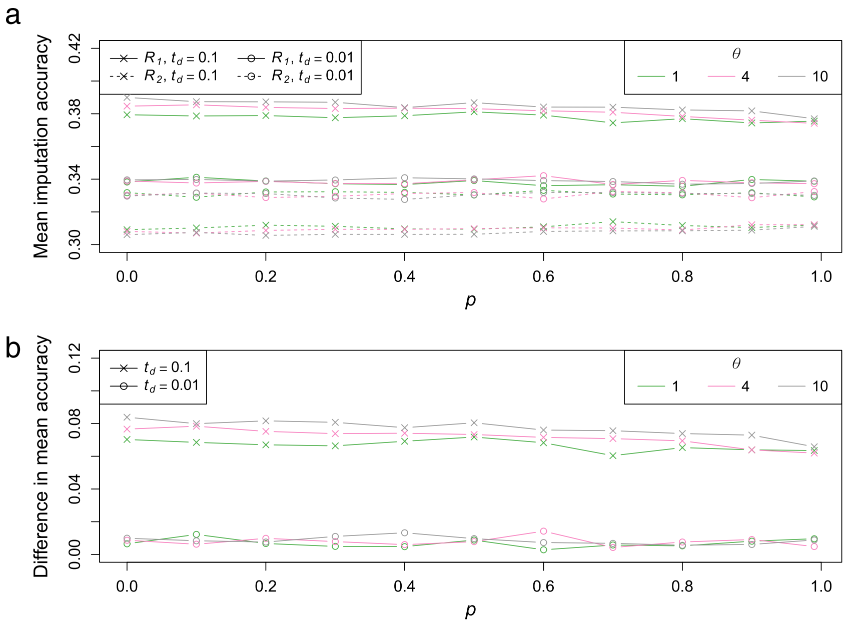


Fig. 8. Simulated mean imputation accuracy for imputations separately performed using R_1 and R_2 as the reference sequence. (A) Mean imputation accuracies based on R_1 and R_2, as a function of p. (B) The pointwise difference in (A) between mean imputation accuracies based on R_1 and R_2, as a function of p. Each point relies on 10^5 simulation replicates.

the fraction of sites imputed correctly with reference R_2 is the fraction of sites due to mutations on branch 1. As the expected number of mutations on a branch is proportional to the branch length, the expected proportion of mutations that lie on branch 2 is approximated by the ratio of the expected length of branch 2 to the expected length of a genealogy with type D. Similarly, the expected proportion of mutations that lie on branch 1 is approximated by the ratio of the expected lengths of branch 1 and the whole genealogy. Thus, we obtain

$$\widehat{\Delta}_{t_d} = \frac{2\mathbb{E}[T_2|G = D, t_d] + 2t_d - 2\mathbb{E}[T_3|G = D, t_d]}{2\mathbb{E}[T_2|G = D, t_d] + 2t_d + \mathbb{E}[T_3|G = D, t_d]} \times \mathbb{P}(G = D|t_d), \tag{46}$$

where $\mathbb{E}[T_2|G = D, t_d]$ and $\mathbb{E}[T_3|G = D, t_d]$ are computed using Eqs. (1) and (3), respectively, and where $\mathbb{P}(G = D|t_d) = 1 - e^{-t_d}$ is given in Eq. (4). For any $t_d \geq 0$, $\mathbb{E}[T_2|G = D, t_d] = 1$, and

$$\mathbb{E}[T_3|G = D, t_d] = \int_0^{t_d} t_3 \frac{e^{-t_3}}{1 - e^{-t_d}}\, \mathrm{d}t_3 = \frac{1 - (1 + t_d)e^{-t_d}}{1 - e^{-t_d}}. \tag{47}$$

We can simplify to obtain

$$\widehat{\Delta}_{t_d} = \frac{2t_d(1 - e^{-t_d})}{2t_d + 3(1 - e^{-t_d} - t_d e^{-t_d})}. \tag{48}$$

Evaluating $\widehat{\Delta}_{t_d}$ at $t_d = 0.1$ and $t_d = 0.01$, we have $\widehat{\Delta}_{0.1} = 0.0889$ and $\widehat{\Delta}_{0.01} = 0.0099$, close to the values observed in simulations.

Note that by definition of genealogical type D, $t_d \geq \mathbb{E}[T_3|G = D, t_d]$. Thus, it is easily seen that Eq. (48) is nonnegative, and use of reference R_1 always results in higher imputation accuracy in the target, on average, than use of reference R_2.

5. Discussion

This paper has introduced a theoretical framework for investigating the population-genetic factors that affect genotype-imputation accuracy. Our framework includes a two-population coalescent model for three sequences, as well as a mutation model to account for stochasticity in the mutation process and thus in the choice of imputation template. Using the model, we have derived approximate expressions for the expectation and variance of imputation accuracy in the target sequence using a reference sequence chosen on the basis of genetic similarity to the target at genotyped positions.

In measuring imputation accuracy as the proportion of polymorphic sites that are missing but subsequently imputed correctly in the target sequence, we found that the mean imputation accuracy increases with increasing θ, p, and t_d. We also observed that the variance in imputation accuracy decreases with increasing θ and t_d but increases with increasing p. Additionally, we found that, under the model, the expected gain in accuracy when reference R_1 rather than reference R_2 is used as a template for imputing the target sequence can be accurately predicted by a formula relating the expected difference in imputation accuracy to the expected difference in branch lengths of R_1 and R_2 (Eqs. (46) and (48)).

Our results on trends in mean imputation accuracy can be explained intuitively by considering the amount of information available for determining which of the two reference sequences, R_1 or R_2, is genetically more similar to the target sequence. For fixed p and t_d, examining genotypes in sequences with more mutations (greater θ) increases the probability that the reference sequence that is genetically more similar to the target at typed and untyped markers alike can be identified. For fixed θ and t_d, increasing the proportion p of sites genotyped has a similar effect. For fixed θ and p, increasing the divergence time t_d increases the relative similarity of the internal rather than external reference sequence to the target. For all three parameters, an increase in the probability of correctly identifying the genetically more similar reference sequence leads to an increase in mean imputation accuracy.

The results have implications for empirical studies of imputation. First, the model develops a framework for simultaneously considering the roles of θ, p, and t_d in the design of an imputation study. To maximize imputation accuracy, investigators might select a higher value of p in low-θ genomic regions than elsewhere in the genome. In examining target populations for which reference sequences are unavailable, the value of t_d to a reference population has a substantial influence on imputation accuracy. The model provides a perspective in which the increases in imputation accuracy from increasing marker density in a sample (increasing p) and from development of a more appropriate reference panel (decreasing t_d) can be compared in terms of their relative cost.

We note that, because of the complexity of the computations, we restricted our attention to a simple demographic model, with only two reference sequences. We have assumed a straightforward imputation scheme that copies an entire genomic region of interest in a template reference sequence into the corresponding positions in the target sequence, rather than allowing different templates in different genomic regions. Further, the reference sequences are assumed to be fully known without error. Each of these assumptions does not account for the full complexity of practical studies in human populations. Nevertheless, the simplicity of our modeling framework has enabled us to study patterns of imputation accuracy that provide insights into how individual population-genetic factors influence imputation accuracy. These insights, along with the complementary coalescent model of Jewett et al. (2012) on large samples but without mutation, can help contribute to further development of advanced strategies for the design of imputation-based association studies in humans and other organisms.

Acknowledgments

We are grateful to M. Edge and E. Jewett for comments on the manuscript, and to M. Zawistowski and S. Zöllner for helpful discussions. We acknowledge support from the Burroughs Wellcome Fund and NIH grants GM081441 and HG005855.

References

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J., Nordborg, M., 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465, 627–631.

Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., Steibel, J.P., 2012. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13, 24.

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. The American Journal of Human Genetics 81, 1084–1097.

Casella, G., Berger, R.L., 2001. Statistical Inference. Duxbury Press, Belmont, CA.

Donnelly, P., 2008. Progress and challenges in genome-wide association studies in humans. Nature 456, 728–731.

Druet, T., Schrooten, C., de Roos, A.P.W., 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science 93, 5443–5454.

Egyud, M.R.L., Gajdos, Z.K.Z., Butler, J.L., Tischfield, S., Le Marchand, L., Kolonel, L.N., Haiman, C.A., Henderson, B.E., Hirschhorn, J.N., 2009. Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation. Human Genetics 125, 295–303.

Fridley, B.L., Jenkins, G., Deyo-Svendsen, M.E., Hebbring, S., Freimuth, R., 2010. Utilizing genotype imputation for the augmentation of sequence data. PLoS One 5, e11018.

Guan, Y., Stephens, M., 2008. Practical issues in imputation-based association mapping. PLoS Genetics 4, e1000279.

Hardy, J., Singleton, A., 2009. Genomewide association studies and human disease. New England Journal of Medicine 360, 1759–1768.

Hickey, J.M., Crossa, J., Rabu, R., de los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science 52, 654–663.

Hindorff, L.A., MacArthur, J., Wise, A., Junkins, H.A., Hall, P.N., Klemm, A.K., Manolio, T.A., A catalog of published genome-wide association studies. www.genome.gov/gwastudies/ (accessed: September 6, 2011).

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106, 9362–9367.

Huang, L., Jakobsson, M., Pemberton, T.J., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J.K., Tishkoff, S.A., Rosenberg, N.A., 2011. Haplotype variation and genotype imputation in African populations. Genetic Epidemiology 35, 766–780.

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A., Scheet, P., 2009. Genotype-imputation accuracy across worldwide human populations. The American Journal of Human Genetics 84, 235–250.

Jewett, E., Zawistowski, M., Rosenberg, N.A., Zöllner, S., 2012. A coalescent model for genotype imputation. Genetics 191, 1239–1255.

Johnson, N.L., Kotz, S., 1969. Discrete Distributions. Wiley, New York.

Kingman, J.F.C., 1982a. The coalescent. Stochastic Processes and their Applications 13, 235–248.

Kingman, J.F.C., 1982b. On the genealogy of large populations. Journal of Applied Probability 19A, 27–43.

Kirby, A., Kang, H.M., Wade, C.M., Cotsapas, C., Kostem, E., Han, B., Furlotte, N., Kang, E.Y., Rivas, M., Bogue, M.A., Frazer, K.A., Johnson, F.M., Beilharz, E.J., Cox, D.R., Eskin, E., Daly, M.J., 2010. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics 185, 1081–1095.

Li, Y., Ding, J., Abecasis, G.R., 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. The American Journal of Human Genetics 79, S2290.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., Abecasis, G.R., 2011. Low-coverage sequencing: implications for design of complex trait association studies. Genome Research 21, 940–951.

Li, Y., Willer, C.J., Sanna, S., Abecasis, G.R., 2009. Genotype imputation. Annual Review of Genomics and Human Genetics 10, 387–406.

Manolio, T.A., 2010. Genomewide association studies and assessment of the risk of disease. New England Journal of Medicine 363, 166–176.

Manolio, T.A., Brooks, L.D., Collins, F.S., 2008. A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Investigation 118, 1590–1605.

Marchini, J., Howie, B., 2010. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11, 499–511.

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

Nicolae, D.L., 2006. Testing untyped alleles (TUNA)—applications to genome-wide association studies. Genetic Epidemiology 30, 718–727.

Paşaniuc, B., Avinery, R., Gur, T., Skibola, C.F., Bracci, P.M., Halperin, E., 2010. A generic coalescent-based framework for the selection of a reference panel for imputation. Genetic Epidemiology 34, 773–782.

Pei, Y.-F., Li, J., Zhang, L., Papasian, C.J., Deng, H.-W., 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 3, e3551.

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

Shriner, D., Adeyemo, A., Chen, G., Rotimi, C.N., 2010. Practical considerations for imputation of untyped markers in admixed populations. Genetic Epidemiology 34, 258–265.

Stranger, B.E., Stahl, E.A., Raj, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187, 367–383.

Surakka, I., Kristiansson, K., Anttila, V., Inouye, M., Barnes, C., Moutsianas, L., Salomaa, V., Daly, M., Palotie, A., Peltonen, L., Ripatti, S., 2010. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Research 20, 1344–1351.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7, 256–276.