A Coalescent Model for Genotype Imputation

Ethan M. Jewett,*[,1] Matthew Zawistowski,[†,‡,1] Noah A. Rosenberg,*[,2] and Sebastian Zöllner[†,‡,§,2,3]
*Department of Biology, Stanford University, Stanford, California 94305, [†]Department of Biostatistics, [‡]Center for Statistical Genetics, and [§]Department of Psychiatry, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT The potential for imputed genotypes to enhance an analysis of genetic data depends largely on the accuracy of imputation, which in turn depends on properties of the reference panel of template haplotypes used to perform the imputation. To provide a basis for exploring how properties of the reference panel affect imputation accuracy theoretically rather than with computationally intensive imputation experiments, we introduce a coalescent model that considers imputation accuracy in terms of population-genetic parameters. Our model allows us to investigate sampling designs in the frequently occurring scenario in which imputation targets and templates are sampled from different populations. In particular, we derive expressions for expected imputation accuracy as a function of reference panel size and divergence time between the reference and target populations. We find that a modestly sized "internal" reference panel from the same population as a target haplotype yields, on average, greater imputation accuracy than a larger "external" panel from a different population, even if the divergence time between the two populations is small. The improvement in accuracy for the internal panel increases with increasing divergence time between the target and reference populations. Thus, in humans, our model predicts that imputation accuracy can be improved by generating small population-specific custom reference panels to augment existing collections such as those of the HapMap or 1000 Genomes Projects. Our approach can be extended to understand additional factors that affect imputation accuracy in complex population-genetic settings, and the results can ultimately facilitate improvements in imputation study designs.

GENOTYPE imputation is the estimation of genotypes at untyped markers using patterns of haplotype structure (Halperin and Stephan 2009; Li *et al.* 2009; Marchini and Howie 2010). Imputation is a powerful tool in modern genetic studies. It is routinely used to increase the fraction of the genome covered in genome-wide association studies (GWAS) conducted in human populations, thereby increasing power to detect risk variants through linkage disequilibrium (LD) mapping (Marchini *et al.* 2007; Becker *et al.* 2009; Hao *et al.* 2009; Spencer *et al.* 2009; Li *et al.* 2010). Imputation based on a shared set of markers in data sets genotyped on different platforms enables large-scale meta-analyses (Barrett *et al.* 2008; de Bakker *et al.* 2008; Zeggini *et al.* 2008). In sequencing studies of rare variation, imputation can improve the accuracy of genotype calls from sequencing reads (Li *et al.* 2010, 2011), and power for rare-variant tests of association can be increased by augmenting sequence data with imputed samples (Zawistowski *et al.* 2010).

Imputation procedures generally involve a set of target samples in which genotypes will be imputed, a reference panel of phased haplotypes from which genotypes are copied, and an algorithm for the copying procedure. Each target sample is genotyped for single-nucleotide polymorphisms (SNPs) whose genotypes serve as a scaffold for more complete haplotypes. Reference samples are more densely genotyped than the target samples, and might even be fully sequenced. Imputation algorithms typically employ a computationally intensive hidden Markov model that uses genotypes from the SNP scaffold of a target sample to choose haplotypes from the reference panel that are most similar (Scheet and Stephens 2006; Browning and Browning 2007; Marchini *et al.* 2007; Purcell *et al.* 2007; Li *et al.* 2010). Genotypes are then imputed in the target by locally copying reference haplotypes that provide the best match.

Analyses of the imputed genotypes, such as in marker-based association tests, depend on imputation accuracy (Huang *et al.* 2009b), which in turn depends on numerous factors, including the haplotypic diversity of the target

doi: 10.1534/genetics.111.137984
Manuscript received December 19, 2011; accepted for publication May 4, 2012
Available freely online through the author-supported open access option.
[1]These authors contributed equally to this work.
[2]These authors contributed equally to this work.
[3]Corresponding author: University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: szoellne@umich.edu

population, the size of the reference panel, and the genetic similarity of the reference and target individuals (Huang *et al.* 2009a, 2011; Browning and Browning 2009; International HapMap3 Consortium 2010; Jostins *et al.* 2011). Therefore, when performing imputation, and when prospectively designing imputation-based studies, it is important to understand how sampling approaches and population parameters affect imputation accuracy.

Currently, imputation accuracy is generally assessed by experiments on real or simulated data, in which known genotypes are masked, imputed, and compared to their true values. A thorough analysis of the effect of study design variables on imputation accuracy requires many experiments, a computationally expensive task that restricts the number of parameter combinations that can be tested. An alternative strategy is to develop a theoretical model that incorporates the study design variables as parameters and permits analytical calculations of imputation accuracy as a function of these variables. Such a model could enable faster predictions of imputation accuracy across a wider range of study designs than use of the empirical method. Moreover, analytical expressions for imputation accuracy as a function of model parameters can facilitate an intuitive understanding of the relationship between imputation accuracy and features of empirical studies.

In this article, we introduce a coalescent-based theoretical model of imputation. We consider a single target haplotype to be imputed for untyped markers and a reference panel of haplotypes, of which one will be chosen as the template for imputation. Our model relies on the premise that for a given target haplotype, an imputation algorithm ideally chooses as a template the reference haplotype whose number of sequence differences from the target is smallest. We take from the reference panel the haplotype with the closest genealogical history to the target, in terms of coalescence time, to serve as the template; this haplotype will have, on average, the fewest sequence differences from the target. By selecting the template haplotype in this way, our model mimics existing imputation approaches that attempt to select the haplotype, or set of haplotypes, with the closest genealogical relationship to the target (Pasaniuc *et al.* 2010; Howie *et al.* 2011). Under our imputation scheme, we quantify imputation accuracy as a function of reference panel size and demographic parameters for the populations containing the target and reference haplotypes.

The flexibility of the coalescent framework enables complex population-genetic models to be considered. Here, we use the model to investigate imputation-based study designs in the age of next-generation sequencing. In practice, researchers have typically relied on large, publicly available data sets to serve as reference panels. Although public panels are easily accessible and often result in sufficient imputation accuracy, the haplotypes available often derive from a different population than the study sample in which genotypes are imputed. Advances in sequencing technology now permit the creation of custom reference panels that contain genome sequences of individuals from the same source population as the study sample—perhaps even a subset of the study sample itself. Custom panels, however, will likely be smaller than public panels, owing to the sequencing cost required to create them. It is therefore of interest to determine the practical utility of custom panels by assessing the improvement in imputation accuracy for either replacing a public panel with a potentially smaller custom panel or augmenting the public panel with custom reference haplotypes.

Using our coalescent framework, we address this question by considering two potential reference panels for imputation in the same target population. The first is an "internal" reference panel of haplotypes drawn from the target population and is meant to represent a custom panel. The second is an "external" reference panel of haplotypes from a distinct population, such as one included in a public database. Our model predicts that an internal reference panel, even when considerably smaller than an existing external reference panel, nearly always improves imputation accuracy. We examine the dependence of this improvement on the relative sizes of the two panels and on demographic parameters such as divergence times and population growth rates. Our results suggest that in practice, augmenting an existing external reference panel (1000 Genomes, for example) with even relatively few haplotypes from the study population of interest can improve genotype imputation.

Overview of the Model

We use a coalescent framework to model genotype imputation at a nonrecombining genetic locus intended to represent a short region along a chromosome. We assume that genotypes for a single target haplotype T will be imputed. We define a reference panel to be a set of sequenced or densely genotyped haplotypes that does not include T. One haplotype from the reference panel will be chosen as an imputation template, and alleles from the template haplotype will be copied onto T. Assuming that mutations accumulate in proportion to time, the reference haplotypes with the fewest sequence differences from T are the descendants of the lineage with which T first coalesces. Under this assumption— which holds in expectation under the infinitely many-sites model (*e.g.*, Wakeley 2008)—an imputation algorithm that always selects the haplotype in the reference panel with the fewest sequence differences from T (or one such haplotype in case of a tie) is equivalent to the algorithm choosing the reference haplotype with the closest genealogical history to T. Thus, when predicting the accuracy produced by a reference panel, to a first approximation, it is reasonable to use a model that considers only coalescence times, rather than a more complete model with stochastic mutations. As we will see, considering only coalescence times is desirable because it makes analytical computations simple, fast, and straightforward to interpret.

By including haplotypes from multiple reference panels in our model, we can compare the performance of reference

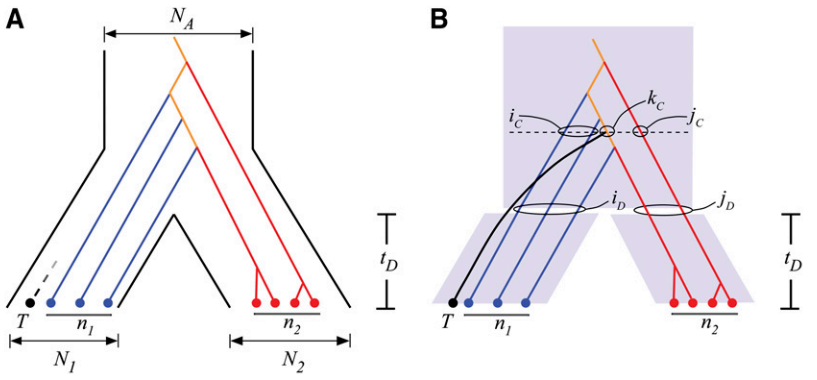

Figure 1 Two-population coalescent model for imputation reference panel selection. (A) Two populations, labeled 1 and 2, of sizes N_1 and N_2 diploid individuals, diverge from an ancestral population of size N_A at time t_D. A single haplotype T for which genotypes at untyped markers are to be imputed is sampled from population 1. We consider two possible reference panels for imputing T: an internal reference panel of n_1 haplotypes sampled from population 1 and an external reference panel of n_2 haplotypes sampled from population 2. If T first coalesces with a type 1 lineage (blue), then the internal panel is optimal for imputing T (event C_1). The external panel is optimal (event C_2) if T first coalesces with a lineage of type 2 (red). Finally, if T first coalesces with a type 1–2 lineage (orange), then the two reference panels are equivalent (event C_{12}). (B) To compute the probability of optimality for each reference panel, we condition on $\mathcal{D}$ (the event that T coalesces before the divergence), the quantities i_D and j_D (the numbers of lineages originating in populations 1 and 2, respectively, that remain at the time of divergence), and i_C, j_C, and k_C (the numbers of type 1, type 2, and type 1–2 lineages remaining at the instant when T first coalesces). In the realization pictured, T does not coalesce before the divergence time (event $\mathcal{D}^c$) and $i_D = 3$, $j_D = 2$, $i_C = 2$, and $j_C = k_C = 1$. Because T first coalesces with a type 1–2 lineage (event C_{12}), the two reference panels are equivalent for imputing T.

panels from different populations. Here, we consider a scenario with two possible reference panels for imputing T. The first is an internal reference panel consisting of n_1 haplotypes sampled from the same population as T. The second is an external reference panel consisting of n_2 haplotypes sampled from a distinct population. Defining the "optimal panel" as that which produces the highest accuracy for imputing T, we compute the probability of optimality for both the internal and external reference panels and quantify the gain in imputation accuracy obtained by using the optimal rather than the suboptimal panel.

We model the genealogical history of T and the reference haplotypes using a two-population coalescent model of divergence (Takahata and Nei 1985; Rosenberg 2003) (Figure 1A). The two populations are labeled 1 and 2, and T is sampled from population 1. The populations diverged from an ancestral population at time t_D in the past, and no migration has occurred between the descendant populations. Therefore, more recently than the divergence time ($t < t_D$), a lineage can coalesce only with other lineages from the same population. This assumption is reasonable for pairs of populations that are geographically isolated. More anciently than the split ($t > t_D$), all remaining lineages are assumed to belong to a homogeneous ancestral population, and any two lineages are allowed to coalesce. We assume effective population sizes of N_1 diploid individuals for population 1, N_2 for population 2, and N_A for the ancestral population.

At time $t = 0$, corresponding to the present, n_1 reference haplotypes in addition to the single target haplotype T are sampled from population 1, and n_2 reference haplotypes are sampled from population 2. The divergence time t_D and the panel sizes n_1 and n_2 are treated as model parameters. We refer to a lineage with descendants only in population 1 as a lineage of type 1. Similarly, a lineage with descendants only in population 2 has type 2, and a lineage with descendants in both populations has type 1–2. We assume that among available reference haplotypes, the optimal templates for imputing T are the descendants of the lineage with which T first coalesces. Thus, the internal reference panel is optimal if T first coalesces with a lineage of type 1 and the external panel is optimal if T first coalesces with a lineage of type 2. If T first coalesces with a lineage of type 1–2, then the two reference panels are equally appropriate, and we say that both are optimal.

Model

We use our coalescent model of genotype imputation to compute the probability of optimality for each reference panel and to quantify differences in imputation accuracy between potential reference panels. For the problem of imputing the target T, we first derive the probability of optimality for an internal reference panel of n_1 haplotypes and for an external reference panel of n_2 haplotypes, sampled from populations with a divergence time of t_D (in units of $2N_A$ generations). Let C_1 be the event that T first coalesces with a lineage of type 1, let C_2 be the event that T first coalesces with a lineage of type 2, and let C_{12} be the event that T first coalesces with a lineage of type 1–2. From our definition of optimality, it follows that $\mathbb{P}(C_1)$, the probability that the target T first coalesces with a lineage of type 1, is the probability that the internal reference panel is optimal for imputing T. Similarly, $\mathbb{P}(C_2)$ is the probability that the external reference panel is optimal, and $\mathbb{P}(C_{12})$ is the probability that the two reference panels are both optimal, with equal expected imputation accuracy.

In the case that exactly one reference panel is optimal, it is of interest to quantify the improvement in imputation accuracy achieved by using the optimal as opposed to the suboptimal reference panel. Assuming that mutations follow a Poisson process, under the infinitely many-sites model, the expected number of sites incorrectly imputed in T is proportional to the coalescence time between the target and the imputation template; the sites that produce imputation errors are precisely those sites at which mutations occur on either the target or the template branch more recently than their coalescence. Expected imputation accuracy for a given reference panel can then be quantified by the expected time that T first coalesces with a haplotype from the panel. We now derive several measurements of imputation error to evaluate the difference in imputation accuracy between the internal and external panels.

Reference panel optimality probabilities

We consider two approaches for obtaining optimality probabilities, a closed-form computation and a recursive computation.

Closed-form computation: Let C_1, C_2, and C_{12}, respectively, be the events that the internal reference panel is optimal for imputing T, the external panel is optimal, and the two panels are equally appropriate. To compute the probabilities for each of these events, we partition the coalescent model into three components: (1) the events in population 1 more recent than the divergence from the ancestral population, (2) the events in population 2 more recent than the divergence, and (3) the events in the ancestral population (Figure 1B). Because we assume that no migration occurs between populations 1 and 2, the coalescent processes in populations 1 and 2 in the period more recent than the divergence are independent. Conditional on the numbers of lineages remaining from populations 1 and 2 at the divergence time, the coalescence events in the ancestral population are independent of the events that occur more recently than the divergence time.

First, we consider the genealogy of haplotypes in population 1 from the present back to the divergence time t_{D}. Define $\mathcal{D}$ to be the event that lineage T coalesces more recently than t_{D} and $\mathcal{D}^c$ to be the event that T does not coalesce by t_{D}. Note that if T coalesces before t_{D}, then the lineage with which it first coalesces can have descendants only in population 1 and must therefore have type 1. It follows that if event $\mathcal{D}$ occurs, then C_1 also occurs, so that $\mathbb{P}(C_1, \mathcal{D}) = \mathbb{P}(\mathcal{D})$. If, however, T does not coalesce before the divergence time, then it enters the ancestral population, where it can also coalesce with lineages of types 2 and 1–2. In this scenario, to identify the optimal reference panel, we must consider the coalescence events in population 2 and the ancestral population.

The coalescent process in the ancestral population depends on the numbers of lineages from populations 1 and 2 that survive at the divergence time. Let i_{D}, $1 \le i_{\mathrm{D}} \le n_{1,}$ denote the number of lineages from population 1 (other than T) that survive to enter the ancestral population at the divergence time t_{D} (Figure 1B). Similarly, let j_{D}, $1 \le j_{\mathrm{D}} \le n_2$, be the number of lineages from population 2 that survive at the divergence time. Because the coalescent processes in populations 1 and 2 are independent, j_{D} is independent of both i_{D} and $\mathcal{D}$.

By conditioning on the number of lineages from each population remaining at the divergence time t_{D}, we can write the quantity $\mathbb{P}(C_1)$ as

$$\begin{aligned}\mathbb{P}(C_1) &= \mathbb{P}(C_1,\mathcal{D}) + \mathbb{P}(C_1,\mathcal{D}^c)\\ &= \mathbb{P}(\mathcal{D}) + \sum_{i_{\mathrm{D}}=1}^{n_1}\sum_{j_{\mathrm{D}}=1}^{n_2}\mathbb{P}(C_1,\mathcal{D}^c,i_{\mathrm{D}},j_{\mathrm{D}})\\ &= \mathbb{P}(\mathcal{D}) + \sum_{i_{\mathrm{D}}=1}^{n_1}\sum_{j_{\mathrm{D}}=1}^{n_2}\mathbb{P}(C_1|\mathcal{D}^c,i_{\mathrm{D}},j_{\mathrm{D}})\mathbb{P}(\mathcal{D}^c,i_{\mathrm{D}},j_{\mathrm{D}})\\ &= \mathbb{P}(\mathcal{D}) + \sum_{i_D=1}^{n_1}\sum_{j_{\mathrm{D}}=1}^{n_2}\mathbb{P}(C_1|\mathcal{D}^c,i_{\mathrm{D}},j_{\mathrm{D}})\mathbb{P}(\mathcal{D}^c,i_{\mathrm{D}})h_{n_2,j_{\mathrm{D}}}(t_{\mathrm{D}};N_2),\end{aligned} \tag{1}$$

where the last equality follows from independence between populations 1 and 2 and $h_{n,\ell}(t;\, N)$ is the probability that n lineages sampled from a diploid population with effective size N coalesce down to ℓ lineages at time t. Tavaré (1984) demonstrated that

$$h_{n,\ell}(t;N) = \sum_{m=\ell}^{n}\frac{(2m-1)(-1)^{m-\ell}\ell_{(m-1)}n_{[m]}}{\ell!(m-\ell)!n_{(m)}}e^{-\binom{m}{2}tN_{\mathrm{A}}/N}, \tag{2}$$

where $n_{[m]} = n(n-1)\cdots(n-m+1)$, $n_{(m)} = n(n+1)\cdots(n+m-1)$, and the factor of N_{A}/N in the exponent is due to the fact that time is measured in units of $2N_{\mathrm{A}}$ generations.

To obtain the probability $\mathbb{P}(\mathcal{D}^c,\, i_D)$, let $\mathcal{A}_n^{\ell}(t;N)$ be the event that n lineages in a diploid population of effective size N coalesce down to ℓ lineages at time t. Define $I_{n,\ell}$, where

$$I_{n,\ell} = \binom{n}{2}\binom{n-1}{2}\cdots\binom{\ell+1}{2} = \frac{n!(n-1)!}{2^{n-\ell}\ell!(\ell-1)!}, \tag{3}$$

is the number of ways that n lineages can coalesce to ℓ lineages (*e.g.*, Rosenberg 2003). Then $\mathbb{P}(\mathcal{D}^c,\, i_{\mathrm{D}})$ is derived by noting that for $\mathcal{D}^c$ and i_{D} to both occur, the $n_1 + 1$ total lineages originating in population 1 must coalesce to $i_{\mathrm{D}} + 1$ lineages at the divergence time t_{D}. This can occur in $I_{n_1+1,i_{\mathrm{D}}+1}$ ways. In $I_{n_1,i_{\mathrm{D}}}$ of these, the n_1 reference haplotypes coalesce to i_{D} lineages without T also coalescing. Thus,

$$\begin{aligned}\mathbb{P}(\mathcal{D}^c,i_{\mathrm{D}}) &= \mathbb{P}(\mathcal{D}^c,\mathcal{A}_{n_1+1}^{i_{\mathrm{D}}+1}(t_{\mathrm{D}};N_1))\\ &= \mathbb{P}(\mathcal{D}^c|\mathcal{A}_{n_1+1}^{i_{\mathrm{D}}+1}(t_{\mathrm{D}};N_1))\mathbb{P}(\mathcal{A}_{n_1+1}^{i_{\mathrm{D}}+1}(t_{\mathrm{D}};N_1))\\ &= \frac{I_{n_1,i_{\mathrm{D}}}}{I_{n_1+1,i_{\mathrm{D}}+1}}\mathbb{P}(\mathcal{A}_{n_1+1}^{i_{\mathrm{D}}+1}(t_{\mathrm{D}};N_1))\\ &= \frac{i_{\mathrm{D}}(i_{\mathrm{D}}+1)}{n_1(n_1+1)}h_{n_1+1,i_{\mathrm{D}}+1}(t_{\mathrm{D}};N_1).\end{aligned} \tag{4}$$

The probability $\mathbb{P}(\mathcal{D}^c)$ is obtained by summing over all possible values of i_{D},

$$\mathbb{P}(\mathcal{D}^c) = \sum_{i_{\mathrm{D}}=1}^{n_1}\mathbb{P}(\mathcal{D}^c,i_{\mathrm{D}}), \tag{5}$$

and the probability $\mathbb{P}(\mathcal{D})$ in Equation 1 can be computed as

$$\mathbb{P}(\mathcal{D}) = 1 - \mathbb{P}(\mathcal{D}^c) = 1 - \sum_{i_\mathrm{D}=1}^{n_1} \mathbb{P}(\mathcal{D}^c, i_\mathrm{D}). \tag{6}$$

The final term to derive in Equation 1, $\mathbb{P}(C_1|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$, is the probability that T first coalesces with a lineage of type 1 assuming that, in addition to T, i_D lineages from population 1 and j_D lineages from population 2 survive to the ancestral population.

To derive $\mathbb{P}(C_1|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ in closed form, let i_C, j_C, and k_C be the numbers of lineages of types 1, 2, and 1–2, respectively, remaining at the instant when T first coalesces. The probability $\mathbb{P}(C_1|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ is computed by summing over all possible values of i_C, j_C, and k_C,

$$\mathbb{P}(C_1|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) = \sum_{k_\mathrm{C}=0}^{\min\{i_\mathrm{D}, j_\mathrm{D}\}} \sum_{i_\mathrm{C}=\delta_{k_\mathrm{C},0}}^{i_\mathrm{D}-k_\mathrm{C}} \sum_{j_\mathrm{C}=\delta_{k_\mathrm{C},0}}^{j_\mathrm{D}-k_\mathrm{C}} \mathbb{P}(C_1, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}), \tag{7}$$

where $\delta_{a,b} = 1$ if $a = b$ and $\delta_{a,b} = 0$ otherwise.

To derive the probability $\mathbb{P}(C_1, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$, let $N(i_\mathrm{D}, j_\mathrm{D} \rightarrow i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C})$ be the number of ways in which i_D lineages of type 1 and j_D lineages of type 2 can coalesce down to i_C, j_C, and k_C lineages of types 1, 2, and 1–2, respectively. Then $\mathbb{P}(C_1, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ is given by

$$\mathbb{P}(C_1, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) = \frac{N(i_\mathrm{D}, j_\mathrm{D} \rightarrow i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}) i_\mathrm{C}}{I_{i_\mathrm{D}+j_\mathrm{D}+1, i_\mathrm{C}+j_\mathrm{C}+k_\mathrm{C}}}. \tag{8}$$

The quantity $N(i_D, j_D \rightarrow i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C})$ is derived in the Appendix.

The probability of optimality for the external reference panel, $\mathbb{P}(C_2)$, and the probability that the two reference panels are equally optimal, $\mathbb{P}(C_{12})$, are computed in a similar manner to Equation 1. Because the occurrence of $\mathcal{D}$ implies that the event C_1 has also occurred, however, $\mathbb{P}(C_2, \mathcal{D}) = \mathbb{P}(C_{12}, \mathcal{D}) = 0$. Thus, the probability that the external reference panel is optimal can be written as

$$\mathbb{P}(C_2) = \sum_{i_\mathrm{D}=1}^{n_1} \sum_{j_\mathrm{D}=1}^{n_2} \mathbb{P}(C_2|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) \mathbb{P}(\mathcal{D}^c, i_\mathrm{D}) h_{n_2, j_\mathrm{D}}(t_\mathrm{D}; N_2), \tag{9}$$

and the probability that the two reference panels are both optimal is

$$\mathbb{P}(C_{12}) = \sum_{i_\mathrm{D}=1}^{n_1} \sum_{j_\mathrm{D}=1}^{n_2} \mathbb{P}(C_{12}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) \mathbb{P}(\mathcal{D}^c, i_\mathrm{D}) h_{n_2, j_\mathrm{D}}(t_\mathrm{D}; N_2). \tag{10}$$

The closed-form expressions for $\mathbb{P}(C_2|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ and $\mathbb{P}(C_{12}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ in Equations 9 and 10 are similar to Equations 7 and 8. For $\mathbb{P}(C_2|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$, we compute

$$\mathbb{P}(C_2|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) = \sum_{k_\mathrm{C}=0}^{\min\{i_\mathrm{D}, j_\mathrm{D}\}} \sum_{i_\mathrm{C}=\delta_{k_\mathrm{C},0}}^{i_\mathrm{D}-k_\mathrm{C}} \sum_{j_\mathrm{C}=\delta_{k_\mathrm{C},0}}^{j_\mathrm{D}-k_\mathrm{C}} \mathbb{P}(C_2, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}), \tag{11}$$

where $\mathbb{P}(C_2, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ is given by

$$\mathbb{P}(C_2, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) = \frac{N(i_\mathrm{D}, j_\mathrm{D} \rightarrow i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}) j_\mathrm{C}}{I_{i_\mathrm{D}+j_\mathrm{D}+1, i_\mathrm{C}+j_\mathrm{C}+k_\mathrm{C}}}. \tag{12}$$

Similarly, for $\mathbb{P}(C_{12}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$, we compute

$$\mathbb{P}(C_2|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) = \sum_{k_\mathrm{C}=1}^{\min\{i_\mathrm{D}, j_\mathrm{D}\}} \sum_{i_\mathrm{C}=0}^{i_\mathrm{D}-k_\mathrm{C}} \sum_{j_\mathrm{C}=0}^{j_\mathrm{D}-k_\mathrm{C}} \mathbb{P}(C_{12}, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}), \tag{13}$$

where $\mathbb{P}(C_{12}, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$ is given by

$$\mathbb{P}(C_{12}, i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D}) = \frac{N(i_\mathrm{D}, j_\mathrm{D} \rightarrow i_\mathrm{C}, j_\mathrm{C}, k_\mathrm{C}) k_\mathrm{C}}{I_{i_\mathrm{D}+j_\mathrm{D}+1, i_\mathrm{C}+j_\mathrm{C}+k_\mathrm{C}}}. \tag{14}$$

Recursive computation: Computing the closed-form expressions (8), (12), and (14) is time–intensive for large i_D and j_D. Therefore, the probabilities $\mathbb{P}(C_1)$, $\mathbb{P}(C_2)$, and $\mathbb{P}(C_{12})$ are difficult to calculate using Equations 1, 9, and 10 with large panel sizes n_1 and n_2. In this section, we derive an efficient recursive approach for computing the probabilities $\mathbb{P}(C_1|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$, $\mathbb{P}(C_2|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$, and $\mathbb{P}(C_{12}|\mathcal{D}^c, i_\mathrm{D}, j_\mathrm{D})$.

Assume that at some time $t > t_\mathrm{D}$, in addition to lineage T, i lineages of type 1, j lineages of type 2, and k lineages of type 1–2 exist in the ancestral population. Conditional on this configuration, let $\tilde{\mathbb{P}}(C_1|i, j, k)$ denote the probability that T first coalesces with a lineage of type 1. We construct a recursive equation for $\tilde{\mathbb{P}}(C_1|i, j, k)$ by conditioning on the lineage pair involved in the next coalescence event and considering its effect on the subsequent coalescent process.

Let $m = i + j + k + 1$ be the total number of lineages remaining. Each of the $\binom{m}{2}$ pairs of lineages is equally likely to be the next to coalesce. Nine distinct pairs of lineage types can coalesce in the next event. For each pair, we compute the probability that the next coalescence will involve lineages of the specified types. Conditional on the lineage types that coalesce, we compute the subsequent probability that T first coalesces with a type 1 lineage. If the next coalescence involves T and a type 1 lineage, an event that occurs with probability $i/\binom{m}{2} = 2i/[m(m-1)]$, then event C_1 occurs. Alternatively, if the next coalescence occurs between T and either a type 2 or a type 1–2 lineage, events that occur with probabilities $j/\binom{m}{2} = 2j/[m(m-1)]$ and $k/\binom{m}{2} = 2k/[m(m-1)]$, respectively, then C_1 cannot occur. For the remaining lineage pairs, T is not involved in the next coalescence, and the probability of C_1 depends on the lineage pair that is involved in the event. For example, two type 1 lineages coalesce with probability $\binom{i}{2}/\binom{m}{2} = i(i-1)/[m(m-1)]$, reducing the number of type 1 lineages to $i - 1$. The coalescent process then restarts with $i - 1$, j, and k lineages of types 1, 2, and 1–2, respectively. Given this new configuration, the probability of event C_1 is $\tilde{\mathbb{P}}(C_1|i-1, k, j)$. The remaining cases for the recursion appear in Table 1.

By conditioning on the possible lineage pairs for the next coalescence, we obtain a recursion:

Table 1 Derivation of the recursion $\tilde{\mathbb{P}}(C_1|i,j,k)$

Lineage pair for next coalescence	Resulting lineage	Number of ways event can occur	$\mathbb{P}(C_1\|\text{event})$
T, 1	—	i	1
T, 2	—	j	0
T, 1–2	—	k	0
1, 1	1	$\binom{i}{2}$	$\tilde{\mathbb{P}}(C_1\|i-1,j,k)$
1, 2	1–2	ij	$\tilde{\mathbb{P}}(C_1\|i-1,j-1,k+1)$
1, 1–2	1–2	ik	$\tilde{\mathbb{P}}(C_1\|i-1,j,k)$
2, 1–2	1–2	jk	$\tilde{\mathbb{P}}(C_1\|i,j-1,k)$
2, 2	2	$\binom{j}{2}$	$\tilde{\mathbb{P}}(C_1\|i,j-1,k)$
1–2, 1–2	1–2	$\binom{k}{2}$	$\tilde{\mathbb{P}}(C_1\|i,j,k-1)$

Assume that in addition to lineage T, i lineages of type 1, j lineages of type 2, and k lineages of type 1–2 exist in the ancestral population at some time $t > t_D$. Conditional on this configuration, let $\tilde{\mathbb{P}}(C_1|i,j,k)$ denote the probability that T first coalesces with a lineage of type 1. Column 1 lists each possible lineage pair for the next coalescence event. Column 2 gives the resulting lineage type for the coalescence. Column 3 contains the number of ways each event can occur. Column 4 gives the probability that T first coalesces with a lineage of type 1, conditional on the pair of lineages in column 1 being the next to coalesce. The recursive equation for $\tilde{\mathbb{P}}(C_1|i,j,k)$ is obtained by conditioning on all the possible lineage pairs for the next coalescence.

$$\begin{aligned}\tilde{\mathbb{P}}(C_1|i,j,k) \;=\;& \frac{2i}{m(m-1)}+\frac{i(i-1)+2ik}{m(m-1)}\,\tilde{\mathbb{P}}(C_1|i-1,j,k)\\ &+\frac{j(j-1)+2jk}{m(m-1)}\,\tilde{\mathbb{P}}(C_1|i,j-1,k)\\ &+\frac{2ij}{m(m-1)}\,\tilde{\mathbb{P}}(C_1|i-1,j-1,k+1)\\ &+\frac{k(k-1)}{m(m-1)}\,\tilde{\mathbb{P}}(C_1|i,j,k-1).\end{aligned} \tag{15}$$

Equation 15 holds for $i > 0, j \geq 0, k \geq 0$. $\tilde{\mathbb{P}}(C_1|0,j,k) = 0$ for all $j,\, k \geq 0$ because there must be at least one lineage of type 1 for event C_1 to occur. The recursion is incorporated into Equation 1 by replacing $\mathbb{P}(C_1|\mathcal{D}^c,\, i_D,\, j_D)$ with $\tilde{\mathbb{P}}(C_1|i_D,j_D,0)$.

The terms $\mathbb{P}(C_2|\mathcal{D}^c,\, i_D, j_D)$ in Equation 9 and $\mathbb{P}(C_{12}|\mathcal{D}^c,\, i_D, j_D)$ in Equation 10 can also be evaluated recursively, following the same logic used to obtain Equation 15. Denote by $\tilde{\mathbb{P}}(C_2|i,j,k)$ the probability that T first coalesces with a type 2 lineage. Then

$$\begin{aligned}\tilde{\mathbb{P}}(C_2|i,j,k) \;=\;& \frac{2j}{m(m-1)}+\frac{i(i-1)+2ik}{m(m-1)}\,\tilde{\mathbb{P}}(C_2|i-1,j,k)\\ &+\frac{j(j-1)+2jk}{m(m-1)}\,\tilde{\mathbb{P}}(C_2|i,j-1,k)\\ &+\frac{2ij}{m(m-1)}\,\tilde{\mathbb{P}}(C_2|i-1,j-1,k+1)\\ &+\frac{k(k-1)}{m(m-1)}\,\tilde{\mathbb{P}}(C_2|i,j,k-1).\end{aligned} \tag{16}$$

Equation 16, which can replace $\mathbb{P}(C_2|\mathcal{D}^c,\, i_D,\, j_D)$ in Equation 9, holds for $i \geq 0,\ j > 0,\ k \geq 0$. $\tilde{\mathbb{P}}(C_2|i,0,k) = 0$ for all $i,\, k \geq 0$.

Finally, conditional on i, j, and k, denoting by $\tilde{\mathbb{P}}(C_{12}|i,j,k)$ the probability that T first coalesces with a lineage of type 1–2,

$$\begin{aligned}\tilde{\mathbb{P}}(C_{12}|i,j,k) \;=\;& \frac{2k}{m(m-1)}+\frac{i(i-1)+2ik}{m(m-1)}\,\tilde{\mathbb{P}}(C_{12}|i-1,j,k)\\ &+\frac{j(j-1)+2jk}{m(m-1)}\,\tilde{\mathbb{P}}(C_{12}|i,j-1,k)\\ &+\frac{2ij}{m(m-1)}\,\tilde{\mathbb{P}}(C_{12}|i-1,j-1,k+1)\\ &+\frac{k(k-1)}{m(m-1)}\,\tilde{\mathbb{P}}(C_{12}|i,\, j,k-1).\end{aligned} \tag{17}$$

The boundary condition for Equation 17 is $\tilde{\mathbb{P}}(C_{12}|i,0,0) = \tilde{\mathbb{P}}(C_{12}|0,j,0) = 0$ for $i,j \geq 0$, because production of a type 1–2 lineage requires at least one type 1 lineage and at least one type 2 lineage. The recursion is incorporated into Equation 10 by replacing $\mathbb{P}(C_{12}|\mathcal{D}^c,\, i_D,\, j_D)$ with $\tilde{\mathbb{P}}(C_{12}|i_D,j_D,0)$.

Expected coalescence times

In this section, we derive formulas that quantify and compare imputation accuracy for internal and external reference panels. Let S_1 be the number of sites that are incorrectly imputed when using an internal reference panel and let S_2 be the number of sites incorrectly imputed when using an external reference panel. We compute the expected numbers of incorrectly imputed sites, $E[S_1]$ and $E[S_2]$, conditional on reference panel sizes n_1 and n_2 and divergence time t_D.

Let T_1 be the random waiting time until T first coalesces with a lineage that has descendants in the internal reference panel, that is, a type 1 or type 1–2 lineage. Similarly, let T_2 be the waiting time until T first coalesces with a lineage that has descendants in the external reference panel, that is,

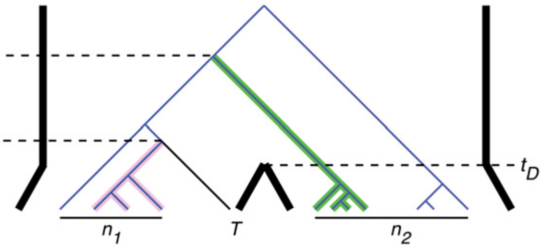

Figure 2 Coalescence times between the target T and the reference panels. T_1 indicates the time at which the target haplotype T first coalesces with a type 1 or type 1–2 lineage. We choose one of the descendant reference haplotypes from that coalescence event (highlighted in purple) to be the template from the internal reference panel. We assume that when using the internal panel, the number of mutations that result in incorrectly imputed sites follows a Poisson distribution with mean $2T_1\theta\omega/2$, where $2T_1$ is the total branch length separating the target T from the templates sampled from the internal panel in units of $2N_A$ generations. Here, $\theta = 4N_A\mu$ is the per-base population-scaled mutation rate, μ is the per-base per-generation mutation rate, and ω is the number of bases genotyped in the reference population that will be imputed in T. Similarly, T_2 is the time at which the target haplotype T first coalesces with a type 2 or type 1–2 lineage and $2T_2$ is the branch length between T and the set of potential templates from the external reference panel (the best external reference panel is highlighted in green).

a type 2 or type 1–2 lineage. Here, T_1 and T_2 are measured in units of $2N_A$ generations. The template haplotype is selected to minimize the coalescence time with lineage T (Figure 2). Thus, the total branch length separating the target from the template is $2T_1$ when the internal reference panel is used and $2T_2$ when the external reference panel is used.

We model mutation events using the infinitely many-sites model and assume that the number of mutations at genotyped sites along a branch of length t units of $2N_A$ generations follows a Poisson distribution with mean $t\theta\omega/2$, where $\theta = 4N_A\mu$ is the per-base population-scaled mutation rate, μ is the per-base per-generation mutation rate, and ω is the number of sites genotyped (or sequenced) in the reference haplotypes that will be imputed in the target. Under our model, mutations that occur along the branches separating the target haplotype and the template haplotype will be incorrectly imputed. Therefore, when the internal reference panel is used, the expected number of sites incorrectly imputed is

$$E[S_1] = E[E[S_1|T_1]] = E[2T_1\theta\omega/2] = \theta\omega E[T_1]. \tag{18}$$

Similarly, the expected number of sites incorrectly imputed using the external panel is

$$E[S_2] = \theta\omega E[T_2]. \tag{19}$$

It follows that the expected difference in the number of sites incorrectly imputed between the external and internal reference panels is

$$E[S_2 - S_1] = E[S_2] - E[S_1] = \theta\omega(E[T_2] - E[T_1]). \tag{20}$$

Thus, up to the scaling factor $\theta\omega$, which does not depend on the model parameters n_1, n_2, and t_D, deriving $E[T_1]$ and $E[T_2]$ is sufficient to determine the expected difference $E[S_2 - S_1]$.

Derivation of E[T*₁*]: To compute the expected waiting time $E[T_1]$ until T first coalesces with a lineage that has descendants in the internal reference panel, we condition on the population in which lineage T first coalesces:

$$E[T_1] = E[T_1|\mathcal{D}]\mathbb{P}(\mathcal{D}) + E[T_1|\mathcal{D}^c]\mathbb{P}(\mathcal{D}^c). \tag{21}$$

Here, $\mathbb{P}(\mathcal{D})$ and $\mathbb{P}(\mathcal{D}^c)$ are obtained using Equations 6 and 5, respectively. To obtain the expected waiting time $E[T_1|\mathcal{D}]$ until lineage T first coalesces, given that it coalesces in population 1, we integrate the conditional survival function $S_{T_1|\mathcal{D}}(t)$ of T_1 given $\mathcal{D}$:

$$E[T_1|\mathcal{D}] = \int_{t=0}^{t_D} S_{T_1|\mathcal{D}}(t)dt. \tag{22}$$

$S_{T_1|\mathcal{D}}(t)$ is calculated as follows:

$$\begin{aligned}
S_{T_1|\mathcal{D}}(t) &= \mathbb{P}(T_1 \geq t|\mathcal{D}) \\
&= 1 - \mathbb{P}(T_1 < t, \mathcal{D})/\mathbb{P}(\mathcal{D}) \\
&= 1 - \mathbb{P}(T_1 < \min\{t, t_D\})/\mathbb{P}(\mathcal{D}) \\
&= 1 - \frac{1}{\mathbb{P}(\mathcal{D})}\sum_{i=1}^{n_1+1}\mathbb{P}(T_1 < \min\{t, t_D\}|\mathcal{A}^i_{n_1+1}(\min\{t, t_D\}; N_1))\mathbb{P}(\mathcal{A}^i_{n_1+1}(\min\{t, t_D\}; N_1)) \\
&= 1 - \frac{1}{\mathbb{P}(\mathcal{D})}\sum_{i=1}^{n_1+1}\left[1 - \frac{I_{n_1,i-1}}{I_{n_1+1,i}}\right]h_{n_1+1,i}(\min\{t, t_D\}; N_1) \\
&= 1 - \frac{1}{\mathbb{P}(\mathcal{D})}\sum_{i=1}^{n_1+1}\left[1 - \frac{i(i-1)}{n_1(n_1+1)}\right]h_{n_1+1,i}(\min\{t, t_D\}; N_1).
\end{aligned} \tag{23}$$

In the fourth equality, $\mathcal{A}^k_n(t; N)$ is the event that n lineages coalesce to k lineages in time t in a diploid population of size N. In the fifth equality, by the same argument used to derive Equation 4, the probability that lineage T does not coalesce when the $n_1 + 1$ sampled lineages coalesce to i lineages is $I_{n_1,i-1}/I_{n_1+1,i}$. Therefore, the probability that T does coalesce is $1-I_{n_1,i-1}/I_{n_1+1,i}$. The term $\mathbb{P}(\mathcal{D})$ in Equation 23 is given by Equation 6. Although the integral in Equation 22 can be carried out analytically, we present the formula in its current form because it is easier to modify Equation 22 from the form given to account for exponential growth.

The quantity $E[T_1|\mathcal{D}^c]$ is the expected time until lineage T first coalesces in the ancestral population with a lineage that has descendants in population 1, given that it does not coalesce in population 1. This expected time can be found by conditioning on the number i_D of type 1 lineages that remain at the divergence time:

$$\begin{aligned}
E[T_1|\mathcal{D}^c] &= \sum_{i_D=1}^{n_1} E[T_1|i_D, \mathcal{D}^c]\mathbb{P}(i_D|\mathcal{D}^c) \\
&= \sum_{i_D=1}^{n_1}\left(t_D + \frac{4N_A}{i_D + 1}\right)\frac{\mathbb{P}(\mathcal{D}^c, i_D)}{\mathbb{P}(\mathcal{D}^c)}.
\end{aligned} \tag{24}$$

Here, we have used the fact that in a population of diploid size N, the expected sum of the lengths of all external branches in a genealogy is $4N$ (Fu and Li 1993). Thus, the mean length of an external branch in a genealogy with $n + 1$ lineages—and consequently, the expected time until a specific lineage coalesces with some lineage among a set of n—is $4N/(n + 1)$. $\mathbb{P}(\mathcal{D}^c, i_D)$ and $\mathbb{P}(\mathcal{D}^c)$ are found using Equations 4 and 5, respectively. The quantity $E[T_1]$ is then obtained by inserting Equations 5, 6, 22, and 24 into Equation 21.

Derivation of $\boldsymbol{E[T_2]}$***:*** To compute $E[T_2]$, we condition on the number j_D of type 2 lineages remaining at the divergence time:

$$\begin{aligned} E[T_2] &= \sum_{j_D=1}^{n_2} E[T_2 | j_D] h_{n_2,j_D}(t_D; N_2) \\ &= \sum_{j_D=1}^{n_2} \left(t_D + \frac{4N_A}{j_D + 1} \right) h_{n_2,j_D}(t_D; N_2). \end{aligned} \tag{25}$$

Exponential growth

Given the utility of population growth models for explaining properties of human genetic variation (*e.g.*, Schaffner *et al.* 2005; Coventry *et al.* 2010), we consider a model of exponential growth in populations 1 and 2 (Figure 3). Let $N_1(t)$ and $N_2(t)$ be functions that define the sizes of populations 1 and 2, respectively, at time t in the past, where t is measured in units of $2N_A$ generations. Here, we assume an ancestral population of constant size and set $N_1(t) = N_1(0)e^{-\alpha_1 t}$ and $N_2(t) = N_2(0)e^{-\alpha_2 t}$ for $t \in [0, t_D]$ and $\alpha_1, \alpha_2 > 0$. We compare the results from this model to those of the constant-size model to evaluate effects of exponential growth on imputation reference panel selection.

Exponential growth changes only the distribution of coalescent waiting times from our computations in the previous sections. All derived equations depend on the waiting times only through the quantity $h_{n,k}(t; N)$, the probability that n lineages coalesce to k lineages in time t in a diploid population with constant size N. Thus, for the exponential growth model, we replace $h_{n,k}(t; N)$ from the constant-size model by $h_{n,k}(t; N(0), \alpha)$, the probability that n lineages coalesce to k lineages in time t for a population with size $N(t) = N(0)e^{-\alpha t}$.

The probability $h_{n,k}(t; N(0), \alpha)$ is computed by solving for the value t' such that $h_{n,k}(t; N(0), \alpha) = h_{n,k}(t'; N_c)$ for a specified constant size N_c, and then computing $h_{n,k}(t'; N_c)$ using Equation 2. Here, we take $N_c = 1$ for simplicity. Let $g(t; N(0), \alpha)$ denote the transformation taking time t in the growing population to time t' in the population of constant size 1. Assuming the size of the growing population at time t is $N(t) = N(0)e^{-\alpha t}$, the transformation $g(t; N(0), \alpha)$ is

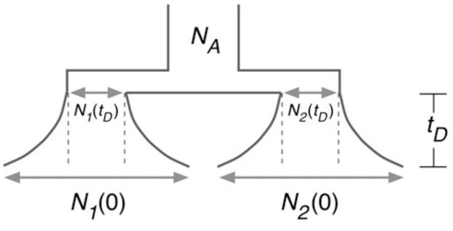

Figure 3 The two-population coalescent model of divergence, assuming exponential growth in the descendant populations. The sizes of populations 1 and 2 change over time according to $N_1(t) = N_1(0)e^{-\alpha_1 t}$ and $N_2(t) = N_2(0)e^{-\alpha_2 t}$, respectively, for $t \in [0, t_D]$. The quantities $\alpha_1, \alpha_2 > 0$ are growth rates, and $N_1(0)$ and $N_2(0)$ are the sizes of populations 1 and 2 in the present. At time t_D, populations 1 and 2 merge instantaneously into the ancestral population, which has constant size N_A. In our analysis, to explore the effect of exponential population growth on imputation accuracy, we vary $N_1(0)$ and $N_2(0)$ while holding $N_1(t_D)$ and $N_2(t_D)$ fixed.

$$g(t; N(0), \alpha) = \int_0^t 1/N(z)dz = \begin{cases} \dfrac{e^{\alpha t} - 1}{N(0)\alpha}, & \text{if } \alpha \neq 0, \\ t/N(0), & \text{if } \alpha = 0, \end{cases} \tag{26}$$

where the units of the transformed time $t' = g(t; N(0), \alpha)$ are the same as those of the untransformed time t (Griffiths and Tavaré 1994; Nordborg 2003). Then

$$\begin{aligned} h_{n,k}(t; N(0), \alpha) &= h_{n,k}(g(t; N(0), \alpha); 1) \\ &= \sum_{i=k}^{n} \frac{(2i-1)(-1)^{i-k} k_{(i-1)} n_{[i]}}{k!(i-k)! n_{(i)}} e^{-\binom{i}{2} g(t; N(0), \alpha) N_A}, \end{aligned} \tag{27}$$

where the factor N_A is needed because the transformed time t' is measured in units of $2N_A$ generations. The modified versions of all equations from the constant-size model appear in Table 2. Because the case in which populations 1 and 2 have constant size is the $\alpha = 0$ case of the growth scenario, results for the constant model can also be obtained using the formulas in Table 2.

Simulations

To validate our theoretical results, we carried out coalescent simulations. Given values of n_1, n_2, and t_D, in the constant-size model, following the method of Jewett and Rosenberg (2012), we simulated genealogies and estimated probabilities $\mathbb{P}(C_1)$, $\mathbb{P}(C_2)$, and $\mathbb{P}(C_{12})$ as the fractions of the simulated genealogies for which the events C_1, C_2, and C_{12} occurred. $\mathbb{P}(C_1)$, $\mathbb{P}(C_2)$, and $\mathbb{P}(C_{12})$ were obtained from the same set of 10^6 simulations.

Results

Agreement of closed-form, recursive, and simulation-based computations

We derived exact closed-form and recursive equations for the probability $\mathbb{P}(C_1)$ that the internal reference panel is optimal, the probability $\mathbb{P}(C_2)$ that the external reference

Table 2 Reformulation of the results of Equations (1) through (25) for the case of exponential growth

Index	Number	Quantity	Dependencies	Description
1	26	$g(t; N(0), \alpha) = \begin{cases} \frac{e^{\alpha t}-1}{N(0)\alpha}, & \text{if } \alpha \neq 0, \\ t/N(0), & \text{otherwise.} \end{cases}$	None	Conversion of elapsed time in units of $2N_A$ generations in a growing population of size $N(t) = N(0)e^{-\alpha t}$ to elapsed time in units of $2N_A$ generations in a diploid population of constant size $N = 1$
2	27	$h_{n,k}(t; N(0), \alpha) = \sum_{i=k}^{n} \frac{(2i-1)(-1)^{i-k} k_{(i-1)} n_{[i]}}{k!(i-k)! n_{(i)}} \exp\left\{-\binom{i}{2} g(t; N(0), \alpha) N_A\right\}$	1	Probability that k lineages remain at time t (in units of $N(0)$ generations) when n lineages are sampled at time 0
3	4	$\mathbb{P}(\mathcal{D}^c, i_D) = \frac{i_D(i_D+1)}{n_1(n_1+1)} h_{n_1+1, i_D+1}(t_D; N_1(0), \alpha_1)$	2	Joint probability that T does not coalesce by time t_D, and i_D lineages (not including T) remain at time t_D
4	5	$\mathbb{P}(\mathcal{D}^c) = \sum_{i_D=1}^{n_1} \mathbb{P}(\mathcal{D}^c, i_D)$	3	Probability that T does not coalesce before time t_D
5	6	$\mathbb{P}(\mathcal{D}) = 1 - \mathbb{P}(\mathcal{D}^c)$	4	Probability that T coalesces before time t_D
6	15	$\tilde{\mathbb{P}}(C_1 \mid i, j, k) = \frac{2i}{m(m-1)} + \frac{i(i-1)+2ik}{m(m-1)} \tilde{\mathbb{P}}(C_1 \mid i-1, j, k)$ $+ \frac{j(j-1)+2jk}{m(m-1)} \tilde{\mathbb{P}}(C_1 \mid i, j-1, k) + \frac{2ij}{m(m-1)} \tilde{\mathbb{P}}(C_1 \mid i-1, j-1, k+1)$ $+ \frac{k(k-1)}{m(m-1)} \tilde{\mathbb{P}}(C_1 \mid i, j, k-1)$ Boundary Condition : $\tilde{\mathbb{P}}(C_1 \mid 0, j, k) = 0$ for all $j, k \geq 0$	None	Probability that T first coalesces with a lineage of type 1, given that there are i lineages of type 1, j lineages of type 2 and k lineages of type 1–2 at time t_D
7	16	$\tilde{\mathbb{P}}(C_2 \mid i, k, j) = \frac{2j}{m(m-1)} + \frac{i(i-1)+2ik}{m(m-1)} \tilde{\mathbb{P}}(C_2 \mid i-1, j, k)$ $+ \frac{j(j-1)+2jk}{m(m-1)} \tilde{\mathbb{P}}(C_2 \mid i, j-1, k)$ $+ \frac{2ij}{m(m-1)} \tilde{\mathbb{P}}(C_2 \mid i-1, j-1, k+1)$ $+ \frac{k(k-1)}{m(m-1)} \tilde{\mathbb{P}}(C_2 \mid i, j, k-1)$ Boundary Condition : $\tilde{\mathbb{P}}(C_2 \mid i, 0, k) = 0$ for all $i, k \geq 0$	None	Probability that T first coalesces with a lineage of type 2, given that there are i lineages of type 1, j lineages of type 2 and k lineages of type 1–2 at time t_D
8	17	$\tilde{\mathbb{P}}(C_{12} \mid i, k, j) = \frac{2k}{m(m-1)} + \frac{i(i-1)+2ik}{m(m-1)} \tilde{\mathbb{P}}(C_{12} \mid i-1, j, k)$ $+ \frac{j(j-1)+2jk}{m(m-1)} \tilde{\mathbb{P}}(C_{12} \mid i, j-1, k)$ $+ \frac{2ij}{m(m-1)} \tilde{\mathbb{P}}(C_{12} \mid i-1, j-1, k+1)$ $+ \frac{k(k-1)}{m(m-1)} \tilde{\mathbb{P}}(C_{12} \mid i, j, k-1)$ Boundary Condition : $\tilde{\mathbb{P}}(C_{12} \mid i, 0, 0) = \tilde{\mathbb{P}}(C_{12} \mid 0, j, 0)$ for all $i, k \geq 0$	None	Probability that T first coalesces with a lineage of type 1-2, given that there are i lineages of type 1, j lineages of type 2 and k lineages of type 1-2 at time t_D
9	1	$\mathbb{P}(C_1) = \mathbb{P}(\mathcal{D}) + \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} \tilde{\mathbb{P}}(C_1 \mid i_D, j_D, 0) \mathbb{P}(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	6, 5, 3, 2	Probability that T first coalesces with a lineage of type 1
10	9	$\mathbb{P}(C_2) = \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} \tilde{\mathbb{P}}(C_2 \mid i_D, j_D, 0) \mathbb{P}(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	7, 3, 2	Probability that T first coalesces with a lineage of type 2
11	10	$\mathbb{P}(C_{12}) = \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} \tilde{\mathbb{P}}(C_{12} \mid i_D, j_D, 0) \mathbb{P}(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	8, 3, 2	Probability that T first coalesces with a lineage of type 1–2
12	23	$S_{T_1 \mid \mathcal{D}}(t) = 1 - \frac{1}{\mathbb{P}(\mathcal{D})} \sum_{i=1}^{n_1+1} \left[1 - \frac{i(i-1)}{n_1(n_1+1)}\right] h_{n_1+1, i}(\min\{t, t_D\}; N_1(0), \alpha_1)$	5, 2	Survival function of the time until lineage T coalesces, given that T coalesces before time t_D
13	22	$E[T_1 \mid \mathcal{D}] = \int_{t=0}^{t_D} S_{T_1 \mid \mathcal{D}}(t) dt$	12	Expected time until T coalesces, given that T coalesces before time t_D

(continued)

Table 2, *continued*

Index	Number	Quantity	Dependencies	Description
14	24	$E[T_1\|\mathcal{D}^c] \ = \sum_{i_D=1}^{n_1}\left(t_D + \frac{4N_A}{i_D+1}\right)\frac{\mathbb{P}(\mathcal{D}^c, i_D)}{\mathbb{P}(\mathcal{D}^c)}$	4, 3	Expected time until *T* coalesces with a lineage with descendants in population 1, given that *T* first coalesces after time t_D
15	21	$E[T_1] = E[T_1\|\mathcal{D}]\mathbb{P}(\mathcal{D}) + E[T_1\|\mathcal{D}^c]\mathbb{P}(\mathcal{D}^c)$	14, 13, 5, 4	Expected time until *T* coalesces with a lineage with descendants in population 1
16	25	$E[T_2] = \sum_{j_D=1}^{n_2}\left(t_D + \frac{4N_A}{j_D+1}\right)h_{n_2,j_D}(t_D; N_2(0), \alpha_2)$	2	Expected time until *T* first coalesces with a lineage with descendants in population 2

The derivation of each expression is the same as in the case of populations of constant size, except that $h_{n,k}(t; N(0),\alpha)$ is used, rather than $h_{n,k}(t; N)$. The quantities on which the expressions in the table depend are given in the Dependencies column. The numbers in the Dependencies column correspond to those in the Index column. The number of each equation—or its analog for the case of populations of constant size—is given in the Number column. Formulas for the case in which populations 1 and 2 have constant size are obtained by setting α_1 and α_2 equal to 0.

panel is optimal, and the probability $\mathbb{P}(C_{12})$ that both panels are equally optimal. Both the exact and recursive computations require the function $h_{n,k}(t; N)$ to be evaluated. However, Equation 2 for $h_{n,k}(t; N)$ is numerically unstable for small t and large n. Therefore, for small t and large n, we used an asymptotic approximation to $h_{n,k}(t; N)$ (Griffiths 1984). For n lineages sampled in the present, the distribution of the number of lineages at time t (expressed in units of $2N$ generations) is asymptotically normal with mean $\mu = 2\eta/t$ and variance $\sigma^2 = 2\eta t^{-1}(\eta + \beta)^2[1 + \eta/(\eta + \beta) - \eta/\alpha - \eta/(\alpha + \beta) - 2\eta]\beta^{-2}$ as $t \to 0$, $n \to \infty$, and $\frac{1}{2}nt \to \alpha < \infty$, where $\beta = -t/2$ and $\eta = \alpha\beta/\{\alpha(e^{\beta} - 1) + \beta e^{\beta}\}$. We used the asymptotic approximation to $h_{n,k}(t; N)$ when both $n \geq 40$ and $t_D \leq 0.1$, and we used Equation 1 otherwise.

Table 3 gives values of $\mathbb{P}(C_1)$ computed using the exact and recursive approaches at two divergence times t_D and several reference panel sizes n_1 and n_2. For larger n_1 or n_2, where closed-form evaluation of $\mathbb{P}(C_1)$ is computationally difficult, we report only values computed using the recursion (Equation 15). The closed-form and recursive probabilities agree at parameter values where a comparison is possible. To allow large reference panel sizes to be considered, we subsequently restrict our attention to the recursion.

The closed-form and recursive expressions also agree with the results of the simulations (Table 3). Because the simulations do not rely on the asymptotic approximation, when $t_D \leq 0.1$ and either $n_1 \geq 40$ or $n_2 \geq 40$, differences between analytical and simulation results are potentially attributable to errors in the approximation. However, these differences are generally negligible.

Comparison of internal and external reference panels

We report $\mathbb{P}(C_1)$, the optimality probability for an internal reference panel, and $E[S_2 - S_1]$, the expected difference in the number of incorrectly imputed sites between the external and internal reference panels, for a range of panel sizes and for large and small divergence times. The optimality probability $\mathbb{P}(C_1)$ is interpreted as the probability that a given locus is imputed more accurately when using the internal reference panel compared to the external panel. The relative accuracy for imputation using the external panel compared to the internal panel is quantified by $E[S_2 - S_1]$. A positive value of $E[S_2 - S_1]$ indicates that using the external reference panel will, on average, result in more imputation errors than using the internal panel. A negative value indicates that the external panel will result in fewer imputation errors on average, and a value of zero indicates that the two panels are expected to produce the same number of incorrectly imputed sites. $E[S_2 - S_1]$ is reported in units of the population-scaled mutation rate $\theta\omega = 4N_A\mu\omega$ for a set of ω imputed bases.

Populations of constant size: In our analyses of the constant-size model, we assumed a constant, equal size for all populations ($N_1 = N_2 = N_A$). Note that because we express imputation accuracy in terms of the population-scaled mutation rate $\theta\omega$, only the relative effective sizes of the populations and not their absolute sizes influence the results of our analyses. Under this model, if t_D is small, then t_D measured in units of $2N_A$ generations is related to the population differentiation statistic F_{st} through the approximation $t_D \approx -\log(1 - F_{st}) \approx F_{st}$ (Cavalli-Sforza 1969; Reynolds *et al.* 1983). We present results for divergence times of $t_D = 0.005$ ($F_{st} \approx 0.005$) to represent two populations within a continental region, and $t_D = 0.05$ ($F_{st} \approx 0.05$) to represent more strongly diverged populations.

Figure 4A shows $\mathbb{P}(C_1)$ for a range of reference panel sample sizes n_1 and n_2 at the smaller divergence time of $t_D = 0.005$. Each curve corresponds to a fixed external reference panel size n_2, and $\mathbb{P}(C_1)$ is plotted as a function of the internal reference panel size n_1. $\mathbb{P}(C_1)$ exceeds 0.5 over most of the range of values for n_1 and n_2, indicating that the internal panel is likely be optimal even when it is much smaller than the external panel. For example, $\mathbb{P}(C_1) > 0.7$ for internal reference panels of $n_1 \geq 400$ haplotypes even when considering an extremely large external panel of $n_2 =$ 10,000 haplotypes (light blue curve).

Figure 4B shows $\mathbb{P}(C_1)$ for the same values of n_1 and n_2 at the larger divergence time of $t_D = 0.05$. For fixed values

Table 3 Comparison of closed-form, recursive, and simulated probabilities

		$t_D = 0.01$			$t_D = 0.05$		
n_1	n_2	Closed form	Recursion	Simulation	Closed form	Recursion	Simulation
5	5	0.4069	0.4069	0.4069	0.5083	0.5083	0.5082
5	10	0.2807	0.2807	0.2808	0.4164	0.4164	0.4164
5	50	0.1188	0.1188	0.1191	0.3034	0.3034	0.3038
10	10	0.4392	0.4392	0.4392	0.6051	0.6051	0.6053
10	50	–	0.2146	0.2153	–	0.4830	0.4836
50	50	–	0.6078	0.6068	–	0.8778	0.8790
50	100	–	0.5404	0.5378	–	0.8670	0.8682
100	100	–	0.7268	0.7274	–	0.9494	0.9499

$\mathbb{P}(C_1)$ computed analytically using closed-form (Equations 7 and 8) and recursive (Equation 15) expressions, and estimated from coalescent simulations using 10^6 replicates.

of n_1 and n_2, $\mathbb{P}(C_1)$ is greater when $t_D = 0.05$ than when $t_D = 0.005$, indicating an increase in the probability of optimality for an internal panel at larger divergence times. Taken together, Figure 4, A and B, shows that under our model, smaller internal reference panels are more likely to be optimal than much larger external panels and that this improvement increases as the two source populations for the panels become more genetically distinct.

Next, we compute the expected difference $E[S_2 - S_1]$ in the number of incorrectly imputed sites between internal and external reference panels to quantify the improvement in accuracy (Figure 4C, $t_D = 0.005$; Figure 4D, $t_D = 0.05$). Increasing the size n_1 of the internal panel while holding the external panel size n_2 fixed results in an increase in $E[S_2 - S_1]$. Conversely, increasing n_2 while holding n_1 fixed leads to a decrease in $E[S_2 - S_1]$. Because the internal reference panel size has no effect on $E[S_2]$ and the external panel size does not affect $E[S_1]$, an increase in $E[S_2 - S_1]$ for fixed n_1 represents an increase in $E[S_2]$, while an increase of $E[S_2 - S_1]$ for fixed n_2 indicates a decrease of $E[S_1]$. The model thus predicts the intuitive result that increasing the size of a reference panel (internal or external) leads to fewer imputation errors. Increasing the number of haplotypes in a reference panel has the largest improvement when the panel is initially small. For example, once the internal panel has reached a certain size ($n_1 \approx 30$ for $t_D = 0.005$), including additional haplotypes in the panel produces only small increases in accuracy. Similarly, the addition of haplotypes to the external panel yields the greatest improvement in accuracy when the external panel is small; hence, a gap is visible between the lines for $n_2 = 50$ (orange) and $n_2 = 100$ (purple) in Figure 4, C and D, while the lines for $n_2 = 500$ (red) and $n_2 = 10{,}000$ (light blue) are nearly indistinguishable.

The magnitude of the divergence time affects the expected accuracies of the reference panels. Comparison of Figure 4, C and D, shows that the performance of the internal reference panel relative to the external panel improves for the larger divergence time. For example, at the smaller divergence time, $n_1 = 67$ haplotypes are required in the internal panel to acquire the same expected imputation accuracy as an external panel of $n_2 = 100$ haplotypes (Figure 4C, purple line). For the larger divergence time, an internal reference panel need only contain $n_1 = 17$ haplotypes to provide the same expected accuracy as $n_2 = 100$ haplotypes in an external panel (Figure 4D, purple line).

The results from Figure 4 can be directly interpreted in terms of events in the coalescent model. $\mathbb{P}(C_1)$ increases with t_D for fixed n_1 and n_2 because increasing the divergence time between the populations lengthens the amount of time that the target haplotype T remains in population 1, where it can coalesce only with type 1 lineages. In

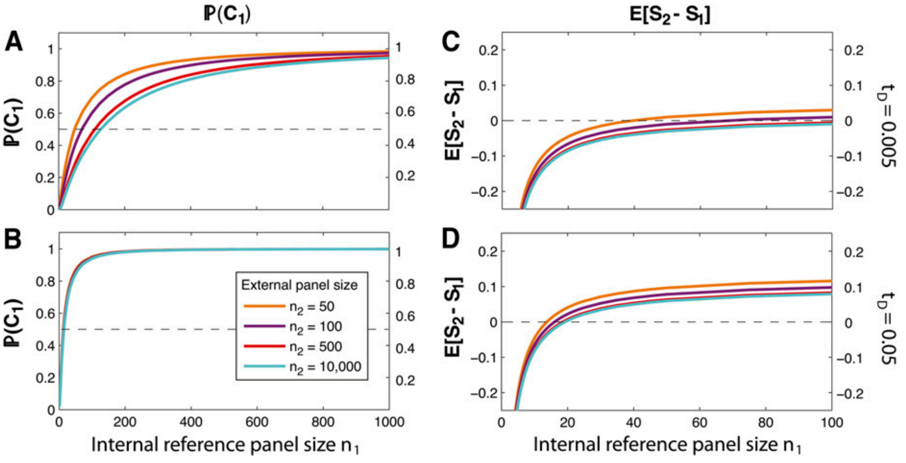

Figure 4 Imputation performance for the constant-size two-population model. For two different divergence times t_D, the figure shows the probability $\mathbb{P}(C_1)$ that the internal reference panel is optimal and the expectation $E[S_2 - S_1]$ of the number of additional imputation errors made when imputing using the external reference panel rather than the internal reference panel. (A) $\mathbb{P}(C_1)$, $t_D = 0.005$. (B) $\mathbb{P}(C_1)$, $t_D = 0.05$. (C) $E[S_2 - S_1]$, $t_D = 0.005$. (D) $E[S_2 - S_1]$, $t_D = 0.05$. $E[S_2 - S_1]$ is reported in units of the population-scaled mutation rate $\theta\omega = 4N_A\mu\omega$ for the imputed region of ω bases. Reference panel size is the number of haplotypes in the panel. For clarity, the scale of C and D differs from that of A and B.

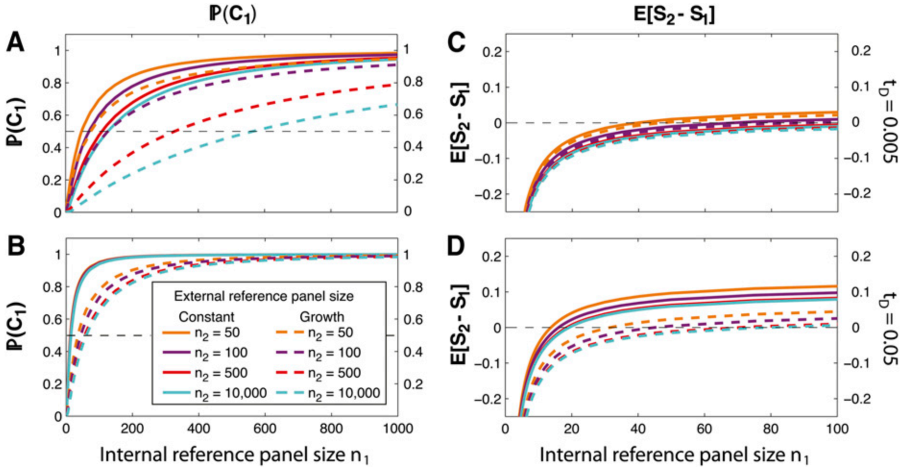

Figure 5 Imputation performance for the exponential-growth two-population model. For two different divergence times t_D, the figure shows the probability $\mathbb{P}(C_1)$ that the internal reference panel is optimal and the expectation $E[S_2 - S_1]$ of the number of additional imputation errors made when imputing using the external reference panel rather than the internal reference panel. Values for the exponential growth model are plotted with dashed lines and, for comparison, the corresponding values for a constant-size model are shown with solid lines. (A) $\mathbb{P}(C_1)$, $t_D = 0.005$. (B) $\mathbb{P}(C_1)$, $t_D = 0.05$. (C) $E[S_2 - S_1]$, $t_D = 0.005$. (D) $E[S_2 - S_1]$, $t_D = 0.05$. $E[S_2 - S_1]$ is reported in units of the population-scaled mutation rate $\theta\omega = 4N_A\mu\omega$ for the imputed region of ω bases. Reference panel size is the number of haplotypes in the panel. For clarity, the scale of C and D differs from that of A and B.

fact, for large divergence times or large n_1, the target T is likely to coalesce before reaching the ancestral population, explaining why $\mathbb{P}(C_1) \approx 1$ nearly independently of n_2 for larger t_D. $E[S_2 - S_1]$ levels off quickly as n_1 and n_2 increase because the expected time until a single lineage in a diploid population of constant size N coalesces with one of ℓ other lineages is $4N/(\ell + 1)$ (Fu and Li 1993). Thus, for our parameter values, the expected time until T coalesces with a type 1 lineage is $E[T_1] = 4N/(n_1 + 1)$ and the expected time until T coalesces with a type 2 lineage is $E[T_2] = t_D + \sum_{j_D=1}^{n_2} h_{n_2,j_D}(t_D; N)[4N/(j_D + 1)]$, where j_D is the number of lineages from population 2 that remain at time t_D. The quantity $4N/(\ell + 1)$ is small even when ℓ is only moderately large. Hence, changes in $E[T_1]$ and $E[T_2]$ with respect to n_1 and j_D, respectively, are small once n_1 or j_D exceeds around 50 lineages. Because j_D increases quickly with n_2 for small t_D, both $E[T_1]$ and $E[T_2]$ change little once n_1 and n_2 are moderate in size. $E[S_2 - S_1]$, which is proportional to $E[T_2] - E[T_1]$, therefore changes little as well.

Exponentially growing populations: We next examine $\mathbb{P}(C_1)$ and $E[S_2 - S_1]$ under a model of exponential growth in populations 1 and 2. Here we assume that both populations have size N_A at the divergence time t_D and size $100N_A$ in the present. We again show results for divergence times of $t_D = 0.005$ and 0.05, measured in units of $2N_A$ generations.

Figure 5A compares $\mathbb{P}(C_1)$ under the constant-size model (solid lines) and exponential growth model (dashed lines) at the smaller divergence time of $t_D = 0.005$. For fixed n_1 and n_2, the value of $\mathbb{P}(C_1)$ under the growth model is less than the corresponding $\mathbb{P}(C_1)$ value for the constant model. For instance, $\mathbb{P}(C_1)$ is 0.68 for $n_1 = 200$ and $n_2 = 500$ under the constant-size model, but it falls to 0.35 for the growth model. The differences in accuracy between the growth and constant-size models are similar although less pronounced for the larger divergence time (Figure 5B). Thus, recent exponential growth in the populations from which the reference and target haplotypes are sampled has the effect of reducing the optimality of the internal reference panel.

The expected difference $E[S_2 - S_1]$ in imputation accuracy between the two panels is also affected by exponential growth. When the divergence time is small ($t_D = 0.005$), the expected difference in accuracy for a given n_1 and n_2 decreases very slightly under exponential growth (Figure 5C). The change in $E[S_2 - S_1]$ between the constant-size and growth models is more extreme at the larger divergence time (Figure 5C). Thus, imputation accuracy for an external reference panel relative to an internal panel improves in the exponential growth model. Recall that for the larger divergence time, only $n_1 = 17$ internal haplotypes were required to achieve the same expected accuracy as $n_2 = 100$ external haplotypes under the constant model. Under the exponential growth model, the number of internal haplotypes needed to match the performance of the $n_2 = 100$ external haplotypes increases to $n_1 = 46$. Although smaller internal reference panels can still achieve accuracy similar to that of larger external panels in the presence of exponential growth, the number of internal haplotypes required to do so increases.

The effects of growth can be understood using intuition about the coalescent model. Including exponential growth in our coalescent model increases the mean waiting time for coalescent events compared to the constant-size case (Figure 6). Therefore, the probability that the target T coalesces more recently than the divergence time t_D decreases, and the number of type 2 lineages j_D that enter into the ancestral population increases. Both of these factors increase the probability that T will survive to the ancestral population and, therefore, the probability that T will coalesce first with a type 2 or type 1–2 lineage. This explains the reduction in $\mathbb{P}(C_1)$ observed for the growth model. Similarly, the decrease in $E[S_2 - S_1]$ under the exponential growth model compared to the constant-size model can be explained by the longer coalescent waiting times. The expected waiting time $E[T_1]$ until T coalesces with a type 1 lineage increases; however, the expected waiting time $E[T_2]$ until T coalesces with a type 2 or type 1–2 lineage decreases because the longer waiting times in population 2 result in a larger number of type 2

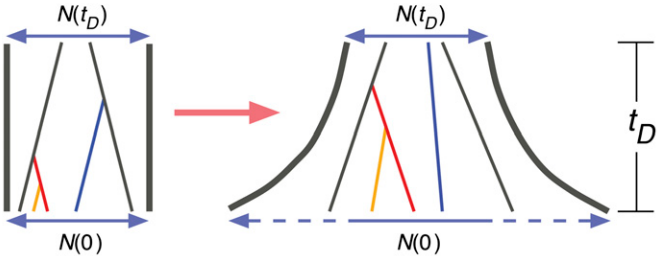

Figure 6 The effect of population growth on coalescent waiting times. Increasing the present-day size $N(0)$ of a population while holding the size $N(t_D)$ at time t_D fixed increases the mean waiting time for each coalescence event.

lineages surviving to the ancestral population. Both the increase in $E[T_1]$ and the decrease in $E[T_2]$ lead to a reduction in $E[S_2 - S_1]$ since $E[S_2 - S_1] \propto E[T_2] - E[T_1]$.

Discussion

We have introduced a theoretical model of genotype imputation accuracy, employing the coalescent framework to model the ancestry of an imputation target haplotype and a set of reference haplotypes. We examined an imputation algorithm that chooses the reference haplotype with the most similar genealogical history to the target as the template for imputation, and we used the expected coalescence time between the target and template haplotypes to predict the expected number of incorrectly imputed sites in the target.

Framing imputation in a coalescent model has two major benefits. First, the coalescent model enables the derivation of analytical formulas for imputation accuracy. These formulas provide a computationally fast method for studying accuracy across a range of imputation study design variables, and they facilitate the interpretation of observed relationships between imputation accuracy and demographic parameters in terms of well-understood properties of the coalescent model. Second, the coalescent allows complex modeling of population histories, so that a variety of relationships between the reference and target populations can be considered. Here we presented a simple model with a single study population and a single external population. However, this basic model can be extended to more complicated imputation scenarios by specifying different demographic settings for the coalescent. For example, the model can be reformulated to include multiple external populations, each with a unique number of available haplotypes and a distinctive divergence time from the study population. Equations analogous to the ones derived in this article can be computed for the more complicated model and used to determine the optimal imputation reference panel when several external panels are available.

We used the model to study the effect of population subdivision on imputation accuracy, employing a two-population divergence model to compare internal reference panels drawn from the same source population as the target to external panels from a distinct population. To quantify imputation accuracy, we focused on two quantities: (1) $\mathbb{P}(C_1)$, the probability that a target lineage on which genotypes are to be imputed coalesces first with a lineage from the internal panel, and (2) $E[S_2 - S_1]$, the expected difference in the number of imputation errors between the external and internal reference panels.

We have interpreted $\mathbb{P}(C_1)$ as the probability that the internal reference panel is optimal and results in fewer expected imputation errors at a locus than the external panel. $\mathbb{P}(C_1)$ has two additional interpretations. First, a reasonable imputation strategy is to augment an available external panel with internal reference haplotypes. In this setting, $\mathbb{P}(C_1)$ is the probability that the target lineage will coalesce first with one of the additional internal lineages. Thus, $\mathbb{P}(C_1)$ can be interpreted as the probability that imputation accuracy improves by augmenting the existing external reference panel with internal haplotypes. A second alternative interpretation of $\mathbb{P}(C_1)$ is obtained by considering imputation on a genome-wide scale, rather than at a single locus. In a genome-wide context, $\mathbb{P}(C_1)$ can be viewed as the fraction of sites in the genome that are more accurately imputed by the internal reference panel than by the external reference panel.

The model predicts that even when an internal reference panel is considerably smaller than an external panel, the internal panel is nearly always optimal in the sense that it contains the haplotype with the closest genealogical history to the target. Furthermore, the probability of optimality for the internal panel, $\mathbb{P}(C_1)$, increases as the divergence time increases. For populations of constant size, a large external reference panel can provide approximately the same accuracy as a modestly sized internal reference panel if the divergence time is small ($E[S_2] \approx E[S_1]$). As the divergence time increases, a small internal reference panel results in increasingly more accurate imputation compared to the large external panel. The expected improvement in imputation accuracy for a smaller internal panel becomes less pronounced if the populations experience exponential growth following the divergence. Thus, exponential growth attenuates the effect of the divergence time, improving the relative performance of an external reference panel in terms of both optimality probability and expected imputation accuracy when compared to populations of constant size.

The results from our model have implications for imputation strategies in population-based genetic association studies. Currently, large public data sets from the HapMap (International HapMap3 Consortium, 2010) and 1000 Genomes (1000 Genomes Project Consortium, 2010) projects are frequently used as external reference panels. However, advances in sequencing technology will enable investigators to create custom internal reference panels from the same source population as their study samples. Not only will custom internal panels enable successful imputation of rare variants private to the target population, our model predicts that a custom internal reference panel will often improve imputation accuracy in general, even if it is much smaller than an existing external reference panel. Under the model, the best strategy is to combine the internal haplotypes with an available external panel to create a single cosmopolitan reference panel. This approach combines the benefits contributed by the large sample size of the external panel and the greater genetic similarity of the internal panel.

Our equations for panel optimality and imputation accuracy rely on a rule that mimics computational imputation algorithms: the reference haplotype whose coalescence time with the target is minimal serves as the imputation template. In actual data, this rule might not always hold, as stochasticity of mutations and the use of small sets of "tagging" SNPs for estimating pairwise distances among haplotypes could cause a sequence whose coalescence time with the target is not minimal to be most genetically similar to the target. The problem may be more pronounced for rare variants that do not exist on a unique background of tagging SNPs. In addition, the rule for template selection might not be strictly followed by imputation software. We have assumed that the entire length of the target haplotype is imputed using the same reference haplotype; however, owing to past recombination events, a real target haplotype is likely to be composed of multiple segments, each with a distinct optimal template. Our model, therefore, implicitly assumes that an imputation algorithm will correctly jump between reference haplotypes when imputing a target. Deviations from this assumption provide a source of imputation error that was not treated in our analysis. Furthermore, as our analysis considered known haplotypes, we did not explicitly account for haplotype phasing errors. Haplotyping accuracy is known to increase with sample size, so that small internal reference panels may be more prone than large external panels to haplotyping errors that reduce imputation accuracy (Browning and Browning 2011). Consequently, our results can be interpreted as an approximation to the imputation error owing to use of a particular reference panel, without considering stochasticity in the choice of template haplotype and without considering phasing errors or errors that might occur from the imputation algorithm. In future research, it will be important to examine factors such as stochasticity in mutation, properties of sets of tagging SNPs, and recombination.

We have assumed that no migration occurs between the two populations in our model. Relaxing this assumption will likely reduce the magnitude of the difference in accuracy obtained on the basis of the internal and external panels. Migration allows the target haplotype to coalesce with a lineage ancestral to the external reference panel more recently than the divergence time, either if the target migrates to the population containing the external reference panel or if an ancestor of an external reference haplotype migrates to the population containing the target. Therefore, including migration in the model could reduce the expected coalescence time between the target and an ancestor of the external reference panel, leading both to an increase in accuracy and to an increase in the probability of optimality for an external reference panel. Determining rates of migration that lead to a noticeable effect requires further investigation.

Despite the fact that we have presented a simplified model without recombination or migration, our results provide mathematical formulas that allow us to estimate imputation accuracy for a variety of demographic and sampling scenarios. Our equations capture a variety of phenomena pertaining to imputation accuracy, and they facilitate the interpretation of these phenomena in terms of properties of the coalescent model. Genotype imputation is a valuable tool in genetic studies of complex disease, and optimizing imputation accuracy is important for conducting analyses with imputed data. The formulas we have derived are a step toward the development of more complicated models that can be used to make practical quantitative predictions about imputation accuracy, thereby facilitating sampling design.

Acknowledgments

The authors thank E. Buzbas, S. Gopalakrishnan, L. Huang, and two anonymous reviewers for helpful comments. Support was provided by National Institutes of Health grants R01 GM081441, R01 HG005855, and T32 HG000040, and by a grant from the Burroughs Wellcome Fund.

Literature Cited

1000 Genomes Project Consortium, 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

Andrews, G., 1984 *The Theory of Partitions*. Cambridge University Press, Cambridge, UK.

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr *et al.,* 2008 Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 40: 955–962.

Becker, T., A. Flaquer, F. F. Brockschmidt, C. Herold, and M. Steffens, 2009 Evaluation of potential power gain with imputed genotypes in genome-wide association studies. Hum. Hered. 68: 23–34.

Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84: 210–223.

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

Browning, S. R., and B. L. Browning, 2011 Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 12: 703–714.

Cavalli-Sforza, L., 1969 Human diversity. Proceedings of the 12th International Congress of Genetics, Tokyo, Vol. 3, pp. 405–416.

Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell *et al.,* 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. 1: 131.

de Bakker, P. I. W., M. A. R. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri *et al.,* 2008 Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. 17: R122–R128.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Griffiths, R. C., 1984 Asymptotic line-of-descent distributions. J. Math. Biol. 21: 67–75.

Griffiths, R., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. 344: 403–410.

Halperin, E., and D. A. Stephan, 2009 Maximizing power in association studies. Nat. Biotechnol. 27: 255–256.

Hao, K., E. Chudin, J. McElwee, and E. Schadt, 2009 Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet. 10: 27.

Howie, B., J. Marchini, and M. Stephens, 2011 Genotype imputation with thousands of genomes. G3: Genes, Genomes, Genetics 1: 457–470.

Huang, L., Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis *et al.,* 2009a Genotype-imputation accuracy across worldwide human populations. Am. J. Hum. Genet. 84: 235–250.

Huang, L., C. Wang, and N. A. Rosenberg, 2009b The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am. J. Hum. Genet. 85: 692–698.

Huang, L., M. Jakobsson, T. J. Pemberton, M. Ibrahim, T. Nyambo *et al.,* 2011 Haplotype variation and genotype imputation in African populations. Genet. Epidemiol. 35: 766–780.

International HapMap3 Consortium, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.

Jewett, E. M., and N. A. Rosenberg, 2012 iGLASS: an improvement to the GLASS method for estimating species trees from gene trees. J. Comput. Biol. 19: 293–315.

Jostins, L., K. I. Morley, and J. C. Barrett, 2011 Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur. J. Hum. Genet. 19: 662–666.

Li, Y., C. Willer, S. Sanna, and G. Abecasis, 2009 Genotype imputation. Annu. Rev. Genomics Hum. Genet. 10: 387–406.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34: 816–834.

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

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

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

Nordborg, M., 2003 Coalescent theory, pp. 602–635 in *Handbook of Statistical Genetics*, Ed. 2, edited by D. Balding, M. Bishop, and C. Cannings. Wiley, New York.

Pasaniuc, B., R. Avinery, T. Gur, C. Skibola, P. Bracci *et al.,* 2010 A generic coalescent-based framework for the selection of a reference panel for imputation. Genet. Epidemiol. 34: 773–782.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.,* 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Reynolds, J., B. S. Weir, and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767–779.

Rosenberg, N. A., 2003 The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57: 1465–1477.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.,* 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15: 1576–1583.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78: 629–644.

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5: e1000477.

Takahata, N., and M. Nei, 1985 Gene genealogy and variance of interpopulation nucleotide differences. Genetics 110: 325–344.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26: 119–164.

Wakeley, J., 2008 *Coalescent Theory: An Introduction*. Roberts, Greenwood Village, CO.

Zawistowski, M., S. Gopalakrishnan, J. Ding, Y. Li, S. Grimm *et al.,* 2010 Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am. J. Hum. Genet. 87: 604–617.

Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini *et al.,* 2008 Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. 40: 638–645.

Communicating editor: Y. S. Song

Appendix: The Quantity $N(i_D, j_D \rightarrow i_C, j_C, k_C)$

Here we derive the number of ways $N(i_{\mathrm{D}}, j_{\mathrm{D}} \rightarrow i_{\mathrm{C}}, j_{\mathrm{C}}, k_{\mathrm{C}})$ in which i_{D} type 1 lineages and j_{D} type 2 lineages can coalesce to $i_{\mathrm{C}}, j_{\mathrm{C}}$, and k_{C} lineages of types 1, 2, and 1–2. This quantity is used to obtain the closed forms of the probabilities $\mathbb{P}(C_1)$, $\mathbb{P}(C_2)$, and $\mathbb{P}(C_{12})$ (Equations 1, 9, and 10).

We first note that if k_{C} type 1–2 lineages remain, then at least k_{C} type 1 lineages, and at least k_{C} type 2 lineages, must coalesce together to produce these lineages. Let i^*_{D} and j^*_{D} be the numbers of type 1 and type 2 lineages, respectively, that combine to create the k_{C} lineages of type 1–2 (Figure A1). Further, let $i^*_{\mathrm{D}_r}$ type 1 lineages and $j^*_{\mathrm{D}_r}$ type 2 lineages combine to produce the rth type 1–2 lineage. The possible values of $i^*_{\mathrm{D}_1}, \ldots, i^*_{\mathrm{D}_{k_{\mathrm{D}}}}$ are given by all possible partitions of i^*_{D} objects into k_{C} nonempty subsets. Similarly, the possible values of $j^*_{\mathrm{D}_1}, \ldots, j^*_{\mathrm{D}_{k_{\mathrm{C}}}}$ are given by all possible partitions of j^*_{D} objects into k_{C} nonempty subsets.

Let $\psi(n, k)$ denote the number of partitions of an integer n into k positive integers (Andrews 1984, p. 16). Let $\pi^q(n,k) = (\pi^q_1(n,k), \ldots, \pi^q_k(n,k))$ denote the qth partition of this kind, with $\pi^q_r(n,k)$ denoting the rth part in the partition. We can permute the k parts of the qth partition in $k!$ ways. Denote the zth permutation of partition q by $\pi^{(q,z)}(n,k) = (\pi^{(q,z)}_1(n,k), \ldots, \pi^{(q,z)}_k(n,k))$. For simplicity of notation, denote the number of labeled histories among a set of n lineages—the number of sequences in which n lineages can coalesce to one lineage—by $F(n) \equiv I_{n,1}$, where $I_{n,1}$ is computed using Equation 3. Then the quantity $N(i_{\mathrm{D}}, j_{\mathrm{D}} \rightarrow i_{\mathrm{C}}, k_{\mathrm{C}}, j_{\mathrm{C}})$ is given by

$$\begin{aligned} N(i_{\mathrm{D}}, j_{\mathrm{D}} \;\rightarrow\; i_{\mathrm{C}}, k_{\mathrm{C}}, j_{\mathrm{C}}) &= \sum_{i^*_{\mathrm{D}}=k_{\mathrm{C}}}^{i_D - i_{\mathrm{C}}} \sum_{j^*_{\mathrm{D}}=k_{\mathrm{C}}}^{j_D - j_{\mathrm{C}}} \binom{i_D}{i^*_{\mathrm{D}}} \binom{j_D}{j^*_{\mathrm{D}}} \sum_{\eta=1}^{\psi(i^*_{\mathrm{D}},k_{\mathrm{C}})} \sum_{\gamma=1}^{\psi(j^*_{\mathrm{D}},k_{\mathrm{C}})} \alpha\left(i^*_{\mathrm{D}}, k_{\mathrm{C}}, \eta\right) \alpha\left(j^*_{\mathrm{D}}, k_{\mathrm{C}}, \gamma\right) I_{i_{\mathrm{D}}-i^*_{\mathrm{D}}, i_{\mathrm{C}}} I_{j_{\mathrm{D}}-j^*_{\mathrm{D}}, j_{\mathrm{C}}} \\ &\quad \times R(i^*_{\mathrm{D}}, j^*_{\mathrm{D}}, k_{\mathrm{C}}, \eta, \gamma) \binom{i_{\mathrm{D}} + j_{\mathrm{D}} - (i_{\mathrm{C}} + k_{\mathrm{C}} + j_{\mathrm{C}})}{i_{\mathrm{D}} - i^*_{\mathrm{D}} - i_{\mathrm{C}},\; j_{\mathrm{D}} - j^*_{\mathrm{D}} - j_{\mathrm{C}},\; i^*_{\mathrm{D}} + j^*_{\mathrm{D}} - k_{\mathrm{C}}}. \end{aligned} \tag{A1}$$

The quantity $\alpha(n, k, q)$ is the number of ways to distribute n distinguishable objects into k unordered, nonempty subsets whose sizes are those of the parts of the qth partition $\pi^q(n, k)$. To obtain an expression for $\alpha(n, k, q)$, define $a(\phi; \pi^q(n, k))$ to be the number of parts of the partition $\pi^q(n, k)$ that have size ϕ. Then $\alpha(n, k, q)$ is given by

$$\alpha(n,k,q) = \frac{\binom{n}{\pi^q_1(n,k), \ldots, \pi^q_k(n,k)}}{\prod_{\phi=1}^{k} a(\phi; \pi^q(n,k))!}, \tag{A2}$$

where $\binom{n}{\pi^q_1(n,k), \ldots, \pi^q_k(n,k)}$ is the number of ways to choose the elements in the k parts and $a(\phi; \pi^q(n, k))!$ is the number of ways to permute the parts of size ϕ. $\prod_{\phi=1}^{k} a(\phi; \pi^q(n,k))!$ is the factor by which we overcount $\alpha(n, k, q)$ because we take the partitions to be unordered, and there are $a(\phi; \pi^q(n, k))!$ arrangements of the subsets of size ϕ in which the same elements in these subsets are grouped together.

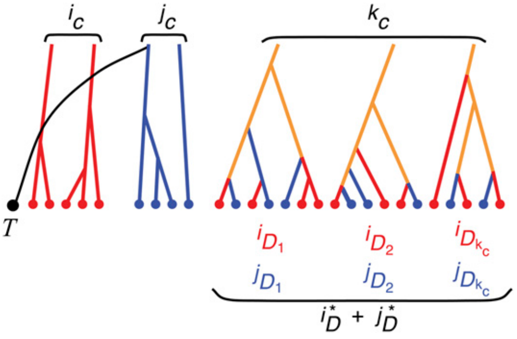

Figure A1 An illustration of the quantity $N(i_{\mathrm{D}}, j_{\mathrm{D}} \rightarrow i_{\mathrm{C}}, j_{\mathrm{C}}, k_{\mathrm{C}})$. One possible way in which i_{D} = 15 type 1 lineages, j_{D} = 12 type 2 lineages, and one target lineage T can coalesce to i_{C} = 2 type 1 lineages, j_{C} = 2 type 2 lineages, and k_{C} = 3 type 1–2 lineages. Type 1 lineages are in red, type 2 lineages are in blue, and type 1–2 lineages are in orange. In the scenario pictured, lineage T first coalesces with a lineage of type 2. Here $i^*_{\mathrm{D}} = 10$ is the number of type 1 lineages that coalesce with $j^*_{\mathrm{D}} = 8$ type 2 lineages to produce the k_C = 3 type 1–2 lineages. $i^*_{\mathrm{D}_1} = 4$ is the number of type 1 lineages that combine with $j^*_{\mathrm{D}_1} = 3$ type 2 lineages to create the first type 1–2 lineage. In general, $i^*_{\mathrm{D}_r}$ is the number of type 1 lineages that coalesce with $j^*_{\mathrm{D}_r}$ type 2 lineages to produce the rth type 1–2 lineage.

In Equation A1, $I_{i_{\mathrm{D}}-i^*_{\mathrm{D}}, i_{\mathrm{C}}}$ is the number of labeled histories for the $i_{\mathrm{D}}-i^*_{\mathrm{D}}$ lineages that coalesce to form type 1 lineages, and it is computed using Equation 3. Similarly, $I_{j_{\mathrm{D}}-j^*_{\mathrm{D}}, j_{\mathrm{C}}}$ is the number of labeled histories for the $j_{\mathrm{D}}-j^*_{\mathrm{D}}$ lineages that coalesce to form type 2 lineages.

The quantity $R(i, j, k, \eta, \gamma)$ in Equation A1 is the number of labeled histories for the $i^*_{\mathrm{D}} + j^*_{\mathrm{D}}$ lineages that ultimately coalesce to form the k_{C} lineages of type 1–2. Given i^*_{D} and j^*_D, consider a particular partition of the i^*_{D} lineages into k_{C} nonempty parts, and a particular partition of the j^*_{D} lineages into k_{C} nonempty parts. Each one of the k_C type 1–2 lineages is made by combining a part from the partition of the i^*_{D} lineages with a part from the partition of the j^*_{D} lineages. To find all possible ways to pair up parts, we fix the indices of the parts of the j^*_{D} lineages and we permute the parts of the i^*_{D} lineages. There are $k_C!$ ways to pair up the parts. We index these ways by z. For the zth way of permuting the parts, the lineages in part $\pi^{(\eta,z)}_r(i^*_{\mathrm{D}}, k_{\mathrm{C}})$ combine with the lineages in part $\pi^\gamma_r(j^*_{\mathrm{D}}, k_{\mathrm{C}})$ to produce the rth lineage of type 1–2.

The rth pair of parts of lineages undergoes $\pi^{(\eta,z)}_r(i^*_{\mathrm{D}}, k_{\mathrm{C}}) + \pi^\gamma_r(j^*_{\mathrm{D}}, k_{\mathrm{C}})-1$ coalescence events on its way down to a single lineage. Thus, there are

$$\begin{aligned} &W\left(i^*_{\mathrm{D}}, j^*_{\mathrm{D}}, k_{\mathrm{C}}, \pi^{(\eta,z)}\left(i^*_{\mathrm{D}}, k_{\mathrm{C}}\right), \pi^\gamma\left(j^*_{\mathrm{D}}, k_{\mathrm{C}}\right)\right) \\ &\equiv \binom{i^*_{\mathrm{D}} + j^*_{\mathrm{D}} - k_{\mathrm{C}}}{\pi^{(\eta,z)}_1\left(i^*_{\mathrm{D}}, k_{\mathrm{C}}\right) + \pi^\gamma_1\left(j^*_{\mathrm{D}}, k_{\mathrm{C}}\right) - 1, \ldots, \pi^{(\eta,z)}_{k_{\mathrm{C}}}\left(i^*_{\mathrm{D}}, k_{\mathrm{C}}\right) + \pi^\gamma_{k_{\mathrm{C}}}\left(j^*_{\mathrm{D}}, k_{\mathrm{C}}\right) - 1} \end{aligned} \tag{A3}$$

possible ways to order the coalescence events among all pairs of parts.

Because there are $F(\pi^{(\eta,z)}_r(i^*_{\mathrm{D}}, k_{\mathrm{C}}) + \pi^\gamma_r(j^*_{\mathrm{D}}, k_{\mathrm{C}}))$ labeled histories for the rth pair of parts as they coalesce to form the rth lineage of type 1–2, there are $W(i^*_{\mathrm{D}}, j^*_{\mathrm{D}}, k_{\mathrm{C}}, \pi^{(\eta,z)}(i^*_{\mathrm{D}}, k_{\mathrm{C}}), \pi^\gamma(j^*_{\mathrm{D}}, k_{\mathrm{C}})) \prod_{r=1}^{k_{\mathrm{C}}} F(\pi^{(\eta,z)}_r(i^*_{\mathrm{D}}, k_{\mathrm{C}}) + \pi^\gamma_r(j^*_{\mathrm{D}}, k_{\mathrm{C}}))$ labeled histories for all of the i_{D}^{*} and j_{D}^{*} lineages when paired in this way. Finally, summing over all $k_{\mathrm{C}}!$ possible ways to permute the partitions of the i_{D}^{*} lineages with respect to the partitions of the j_{D}^{*} lineages,

$$\begin{aligned} R\left(i_{\mathrm{D}}^{*}, j_{\mathrm{D}}^{*}, k_{\mathrm{C}}, \eta, \gamma\right) = & \sum_{z=1}^{k_{\mathrm{C}}!} W\left(i_{\mathrm{D}}^{*}, j_{\mathrm{D}}^{*}, k_{\mathrm{C}}, \pi^{(\eta,z)}\left(i_{\mathrm{D}}^{*}, k_{\mathrm{C}}\right), \pi^{\gamma}\left(j_{\mathrm{D}}^{*}, k_{\mathrm{C}}\right)\right) \\ & \times \prod_{r=1}^{k_{\mathrm{C}}} F\left(\pi_{r}^{(\eta,z)}\left(i_{\mathrm{D}}^{*}, k_{\mathrm{C}}\right) + \pi_{r}^{\gamma}\left(j_{\mathrm{D}}^{*}, k_{\mathrm{C}}\right)\right). \end{aligned} \tag{A4}$$

We have separately considered three parts of the labeled history of the lineages in the ancestral population: (1) the labeled history of the $i_{\mathrm{D}} - i_{\mathrm{D}}^{*}$ lineages that coalesce to form type 1 lineages, (2) the labeled history of the $j_{\mathrm{D}} - j_{\mathrm{D}}^{*}$ lineages that coalesce to form type 2 lineages, and (3) the labeled history of the $i_{\mathrm{D}}^{*} + j_{\mathrm{D}}^{*}$ lineages that coalesce to form type 1–2 lineages. To combine these components into one full history for all lineages, we must consider only how the coalescence times in each of these components relate to the coalescence times in the other components. Thus, the final quantity in Equation A1 is the number of ways to interweave the coalescence events in these labeled histories. There are $i_{\mathrm{D}} - i_{\mathrm{D}}^{*} - i_{\mathrm{C}}$ coalescence events among the lineages that ultimately have type 1, $j_{\mathrm{D}} - j_{\mathrm{D}}^{*} - j_{\mathrm{C}}$ coalescence events among the lineages that ultimately have type 2, and $i_{\mathrm{D}}^{*} + j_{\mathrm{D}}^{*} - k_{\mathrm{C}}$ coalescence events among the lineages that ultimately have type 1–2. The number of ways to interweave the coalescences for these three histories is equal to the number of ways to choose which of the $i_{\mathrm{D}} + j_{\mathrm{D}} - i_{\mathrm{C}} - j_{\mathrm{C}} - k_{\mathrm{C}}$ total coalescence events correspond to events within each of these different histories. This quantity is the trinomial coefficient

$$\binom{i_{\mathrm{D}} + j_{\mathrm{D}} - i_{\mathrm{C}} - k_{\mathrm{C}} - j_{\mathrm{C}}}{i_{\mathrm{D}} - i_{\mathrm{D}}^{*} - i_{\mathrm{C}},\, j_{\mathrm{D}} - j_{\mathrm{D}}^{*} - j_{\mathrm{C}},\, i_{\mathrm{D}}^{*} + j_{\mathrm{D}}^{*} - k_{\mathrm{C}}}. \tag{A5}$$