Supplementary Information for “Rapid genotype imputation from sequence without reference panels”

Davies et al.

May 11, 2016

Contents

1 Supplementary Note ................................................. 2
   1.1 Pseudo-haploid model ............................................ 3
   1.2 Maximization and parameter updating .......................... 6
      1.2.1 Useful Variables ............................................ 7
      1.2.2 Haploid model ............................................... 8
      1.2.3 Pseudo-haploid model ................................... 10
      1.2.4 Diploid model ............................................. 11
   1.3 Efficient calculation of forward backward variables .......... 14
   1.4 Initialization .................................................. 16
   1.5 Parameter bounding ............................................ 16
   1.6 Heuristics ..................................................... 16
   1.7 Guidance behind parameter options ........................... 17

2 Supplementary Tables ................................................ 19
1 Supplementary Note

Definitions for commonly used variables

| Symbol | Definition |
|--------|------------|
| $K$    | Number of ancestral or founder haplotypes |
| $T$    | Number of SNPs in region |
| $N$    | Number of sample individuals |
| $G$    | Number of generations since population founding |
| $R_r$  | Read with index $r$ which spans $J_r$ SNPs, with SNP indices $u_r$, sequenced bases $s_r$ and base qualities $b_r$, or $R_r = \{u_r, s_r, b_r\}$ |
| $J_r$  | Number of SNPs spanned by read $R_r$ |
| $c_r$  | Central SNP for read $R_r$ |
| $O_t$  | Set of reads with central SNP $t$, $O_t = \{R_r|c_r = t\}$ |
| $O$    | Set of observations for each SNP $t$ on the chromosome, $O = \{O_t|t = 1, ..., T\}$ |
| $u_{r,j}$ | For SNP $j$ in read $R_r$, its index with respect to the chromosomal listing of SNPs (e.g., If the physical position of SNP $t$ in the region is $L_t$ for $t = 1, ..., T$, then SNP $j$ in read $R_r$ has physical position $L_{u_{r,j}}$) |
| $s_{r,j}$ | Sequencing base for SNP $j$ in read $R_r$, with $s_{r,j} = 1$ for the alternate base and 0 for the reference base |
| $b_{r,j}$ | Base quality for SNP $j$ in read $R_r$ |
| $R_{r,j}$ | Subset of read $R_r$ for SNP $j$, or $R_{r,j} = \{u_{r,j}, s_{r,j}, b_{r,j}\}$ |
| $\phi_{r,j}^i$ | Probability of SNP $j$ from read $R_r$ coming from an underlying genotype $i$, or $P(s_{r,j}|g = i)$ |
| $I_t$  | Variable counting the number of recombinations that take place between SNPs $t$ and $t + 1$ |
| $H_{r,t}$ | Variable that takes value 1 if SNP $j$ from read $R_r$ is the alternate base and value 0 if it is the reference base |
| $H_r$  | Variable that takes value 1 if read $R_r$ comes from the maternal haplotype and 2 if it comes from the paternal haplotype |
| $\pi_k$ | Probability of starting in state $k$ at the first SNP |
| $\sigma_t$ | Recombination distance between SNPs $t$ and $t + 1$ |
| $\alpha_{t,k}$ | Probability of switching into state $k$ between SNPs $t$ and $t + 1$ |
| $\theta_{t,k}$ | Probability that haplotype $k$ emits the alternate base at SNP $t$ |
| $\lambda$ | Parameters of the model $\lambda = \{\pi, \sigma, \alpha, \theta\}$ |

In the main text, we described the model by showing one would simulate it, and more formally laid out the details necessary to generate probabilities under the model. Here, we further specify the model by describing how the expectation of the complete data likelihood can be used in an EM framework to provide updated parameters $\lambda^{t+1}$ which guarantee no decrease in the likelihood of the observed data. Doing this requires state space augmentation and calculating expectations over hidden states in the Markov model. Here we show how these
expectations are calculated for the haploid, diploid and pseudo-haploid cases. Later, initialization, bounding, and heuristics of the model are given as well.

First, we give a brief review of notation (and see list above). We consider a genomic region with \( T \) SNPs, and sequencing reads from \( N \) individuals. For each individual, we index their reads with \( r \) so we speak of read \( R_r \). We define a central SNP \( c_r \) for read \( R_r \), so that for each SNP in the region, we observe a set of reads, \( O_t = \{ R_r | c_r = t \} \). Read \( R_r \) consists of a triplet of vectors: \( u_r \), the indices; \( s_r \) the reference (0) or alternate (1) bases; and the base qualities \( b_r \). From this we use \( \phi_r^{i,j} \), the probability SNP \( j \) in read \( R_r \) has underlying genotype \( i \).

We model our population as having been founded \( G \) generations ago with \( K \) ancestral haplotypes. Sampling a set of observations for an individual can be thought of as 1) choosing an initial haplotype \( k \) according to \( \pi_k \), the prior probability of starting in state \( k \); 2) choosing where to switch states according to \( \sigma_t \), the genetic distance between SNPs \( t \) and \( t+1 \); 3) choosing which haplotype \( k \) to sample at recombination breakpoints according to \( \alpha_{t,k} \), the local probability of switching into haplotype \( k \) at SNP \( t+1 \); and 4) sampling reads by i) choosing read breakpoints and determining \( u_{r,j} \), the indices of the SNPs in the read; ii) obtaining \( b_{r,j} \), the base qualities of the SNPs in the read; iii) choosing the real bases of the SNPs in the read according to \( \theta_{u_{r,j},k} \), the probability that haplotype \( k \) emits the alternate base at SNP \( u_{r,j} \); iv) observing sequenced bases \( s_{r,j} \) according to \( b_{r,j} \) and the real bases.

In the unaugmented hidden state space, the haploid model corresponds to a set of \( k_t \in \{1, \ldots, K\} \forall t = 1, \ldots, T \). For the diploid model, this consists of a set of pairs of states \( (k_{t,1}, k_{t,2}) \), while for the pseudo-haploid mode, it is two hidden states \( k_{t,1} \) and \( k_{t,2} \). In the augmented hidden state space, we further consider knowledge of: how many recombinations occur between SNPs \( t \) and \( t+1 \), defined by variable \( I_t \); whether base \( j \) of read \( R_r \) is a reference or alternate base, defined by variable \( H_j \); and whether read \( R_r \) comes from the maternal or paternal haplotype, defined by variable \( H_r \). Utilization of the augmented hidden state space is necessary for updating parameters, as explained below.

### 1.1 Pseudo-haploid model

The diploid model presented here and used in fastPHASE and other similar algorithms suffers from a quadratic computational complexity due to the need to sum over \( K^2 \) possible diploid states at each site. With sequencing reads, the observed data fundamentally comes from either the first (e.g. maternal) or second (e.g. paternal) haplotype. If we had labels for each read as to whether they came from the maternal or paternal haplotype, we would have separable likelihoods, and could use the maternal reads to infer the maternal states, and likewise for the paternal reads and paternal states, which would have computational cost proportional to \( 2 \times K \) as opposed to \( K^2 \).

In the diploid EM algorithm, we use the current set of parameters to generate the posterior probability of the pair of hidden states given the observations, and use these to generate a new set of parameters that maximize the
likelihood. An alternative approach is to average over sampled hidden states realized through a hypothetical Gibbs sampler that i) samples labels conditional on states, observations, and parameters, and ii) samples states conditional on labels, observations and parameters. Implementing such a Gibbs sampler in reality would be computationally unwise, as it would likely take at least as long as the original diploid EM. However, with certain assumptions about the posterior distribution of the labels, we can approximate the posterior distribution of the hidden states quickly.

Let \( q_1 \) be the full hidden state for haplotype 1, the maternal haplotype. Let \( H_r \) be the label for read \( r \) with \( H_r = 1 \) corresponding to the maternal haplotype and \( H_r = 2 \) corresponding to the paternal haplotype. Let \( O = \{ R_r \} \) be the set of all reads, with \(|O|\) reads in total, and let \( H \) correspond to an assignment of labels \( H \in \mathcal{H} = \{1, 2\}^{\mid O\mid} \). Let \( R_h = \{ R_r \mid H_r = h \} \) be the set of reads with label \( h \). Then we have

\[
P(q_1 \mid O, \lambda) = \sum_{H \in \mathcal{H}} P(q_1, H \mid O, \lambda) \tag{1}
\]

\[
= \sum_{H \in \mathcal{H}} P(q_1 \mid H, O, \lambda) P(H \mid O, \lambda) \tag{2}
\]

\[
= \sum_{H \in \mathcal{H}} P(q_1 \mid H, O, \lambda) \prod_{r=1}^{|O|} P(H_r \mid O, \lambda) \tag{3}
\]

where the last equality requires the approximation that the probability of the labels are independent of each other. Now, the probability of a state given labels and reads can be further written as

\[
P(q_1 \mid H, O, \lambda) = \frac{P(O \mid H, q_1, \lambda) P(q_1 \mid H, \lambda)}{P(O \mid H, \lambda)} \tag{4}
\]

\[
= \frac{\left( \prod_{r:H_r=1} P(R_r \mid q_1, \lambda) \right) P(R_2 \mid \text{hap2}, \lambda) P(q_1 \mid \lambda)}{P(R_1 \mid \text{hap1}, \lambda) P(R_2 \mid \text{hap2}, \lambda)} \tag{5}
\]

where we use \( P(q_1 \mid H, \lambda) = P(q_1 \mid \lambda) \), since labels don’t affect state probabilities without observations, and where \( (R_1 \mid \text{hap1}, \lambda) \) is the probability of observing the set of reads labeled as coming from haplotype 1, conditional on their having come from haplotype 1. If we further approximate \( P(R_1 \mid \text{hap1}, \lambda) = \prod_{r:H_r=1} P(R_r \mid \text{hap1}, \lambda) \), and approximate \( P(R_r \mid \text{hap1}, \lambda) = P(R_r \mid \lambda) \), we get that

\[
P(q_1, H, O, \lambda) = P(q_1 \mid \lambda) \prod_{r:H_r=1} \frac{(P(R_r \mid q_1 \mid \lambda))}{(P(R_r \mid \lambda))} \tag{6}
\]
This gives us that

$$P(q_1|O, \lambda) = \left( \sum_{H \in \mathcal{H}} P(q_1|\lambda) \left( \prod_{r:H_r=1} \frac{P(R_r|q_1, \lambda)}{P(R_r|\lambda)} \right) \left( \prod_{r=1}^{|O|} P(H_r|O, \lambda) \right) \right)$$  \hspace{1cm} (7)

$$= P(q_1|\lambda) \sum_{H \in \mathcal{H}} \prod_{r=1}^{|O|} \left( P(H_r|O, \lambda) \left( I\{H_r = 1\} \frac{P(R_r|q_1, \lambda)}{P(R_r|\lambda)} + I\{H_r = 2\} \right) \right)$$ \hspace{1cm} (8)

$$= P(q_1|\lambda) \prod_{r=1}^{|O|} \left( P(H_r = 1|O, \lambda) \frac{P(R_r|q_1, \lambda)}{P(R_r|\lambda)} + P(H_r = 2|O, \lambda) \right)$$ \hspace{1cm} (9)

Therefore, we get that read $r$ contributes $P(H_r = 1|O, \lambda)P(R_r|q_1, \lambda) + P(H_r = 2|O, \lambda)P(R_r|\lambda)$ to the likelihood, after multiplying by the constant $P(R_r|\lambda)$, as opposed to $P(R_r|q_1, \lambda)$ as it would under a fully separable model. When testing on real data, we found that we achieved marginally but consistently better performance using $P(H_r = 1|O, \lambda)P(R_r|q_1, \lambda) + P(H_r = 2|O, \lambda)P(R_r|\lambda)$ instead, so this equation was used when calculating the state probabilities.

To use this, we need an estimate of the probability of a label given the data. To do this, consider a read $R_r$, with lead SNP $c_r$, and label $H_r$. Then we can calculate the following

$$P(H_r = 1|O, \lambda) = \sum_{q_1, q_2} P(H_r|q_1, q_2, O, \lambda)P(q_1, q_2|O, \lambda)$$ \hspace{1cm} (10)

$$= \sum_{q_1, q_2} P(H_r|q_1, q_2, R_r, \lambda)P(q_1, q_2|O, \lambda)$$ \hspace{1cm} (11)

$$= \sum_{q_1, q_2} \frac{P(R_r|q_1, \lambda)}{P(R_r|\lambda)} P(q_1, q_2|O, \lambda)$$ \hspace{1cm} (12)

$$= \mathbb{E}_{q_1, q_2} \left[ \frac{P(R_r|q_1, \lambda)}{P(R_r|\lambda)} \right]|O, \lambda \right]$$ \hspace{1cm} (13)

$$\approx \frac{\mathbb{E}_{q_1} \left[ P(R_r|q_1, \lambda)|O, \lambda \right]}{\sum_{h=1}^2 \mathbb{E}_{q_h} \left[ P(R_r|q_h, \lambda)|O, \lambda \right]}$$ \hspace{1cm} (14)

This uses a prior probability on labels of $P(H_r = 1) = P(H_r = 2) = \frac{1}{2}$. We also use the approximation that the expectation of ratios is equivalent to the ratio of expectations, to avoid a calculation with computational complexity of order $K^2$. To perform this calculation we use

$$P(R_r|haph, \lambda) = \mathbb{E}_{q_h} \left[ P(R_r|q_h, \lambda)|O, \lambda \right] \approx \sum_{k=1}^K P(R_r|q_h = k, \lambda')P(q_k|O, \lambda')$$ \hspace{1cm} (15)

where $\lambda'$ are the parameters from the previous iteration.

Therefore, in calculating the complete data probability for the pseudo-haploid model for haplotype $H = 1$, we use the probability of the observation at SNP $t$
given state $q_t = k_t$ and parameters $\lambda$ as

$$
P_{H = h} (O_t | q_t = k_t, \lambda) = \prod_{j=1}^{J_r} P_{H = h} (R_r | q_t = k_t, \lambda)
$$

$$
= \prod_{j=1}^{J_r} P(H_r = h | O, \lambda) P(R_r | q_t = k_t, \lambda) + P(H_r \neq h | O, \lambda) P(R_r | \text{hap}, \lambda)
$$

(16)

where $P(H_r = h | O, \lambda)$ is from Equation 14, $P(R_r | q_t = k_t, \lambda)$ is as defined in the main text, $P(H_r \neq h | O, \lambda) = 1 - P(H_r = h | O, \lambda)$, and $P(R_r | \text{hap}, \lambda)$ is from Equation 15.

1.2 Maximization and parameter updating

In the EM algorithm, one defines a “complete dataset” $D$ including the observed data ($O$, the reads), as well as the hidden parameters ($Q$, the hidden states). Given a set of parameters $\lambda$, one defines the log-likelihood of the complete data as $L(\lambda) = \log(l(\lambda|D)) = \log(l(\lambda|D = (O,Q)))$. Given a current set of parameters $\lambda_i$, we generate a new set of parameters $\lambda_{i+1}$ to maximize the expectation of $l(\lambda_{i+1})$ with respect to the distribution of hidden parameters obtained by $\lambda_i$

$$
U(\lambda_{i+1}, \lambda_i) = \mathbb{E}[l(\lambda_{i+1})|O, \lambda_i]
$$

$$
= \sum_Q P(Q|O, \lambda_i) \log(P(O, Q|\lambda_{i+1}))
$$

(17)

Standard theory implies that by choosing $\lambda_{i+1}$ to maximize $U(\lambda_{i+1}, \lambda_i)$, we also increase the likelihood of the observed data, $l(\lambda_{i+1}|O) > l(\lambda_i|O)$.

In applying the EM algorithm, we first initialize with a set of parameters $\lambda^0$. For each subsequent iteration $i = 1, 2, ..., \lambda^i$, we then iteratively alternate between the “Expectation” phase, where we calculate $U(\lambda^{i+1}, \lambda^i)$, and the “Maximization” phase, where we calculate $\lambda^{i+1}$ to maximize $U(\lambda^{i+1}, \lambda^i)$. In the Expectation phase, the crucial component is calculating the state probabilities $P(Q|O, \lambda^i)$ - these are calculated using the forward and backward algorithms. To calculate the updates in the Maximization stage, we must further augment the latent space to model how many recombinations occur between SNPs, whether emissions were due to occurrences of an alternate base or a reference base, and whether observed reads were from the maternal or paternal haplotype. To calculate the updates in the Maximization stage, we must further augment the latent space to model whether transitions occur due to recombinations or not, and whether emissions were due to occurrences of an alternate base or a reference base. In this new augmented latent space, for some fixed set of hidden parameters for the $N$ samples, consider some sums that can be calculated. Let $n^t_k$ be the number of sample haplotypes in state $k$ at the first SNP, $n^t_{\text{st}}$ be the number of sample haplotypes which do not recombine between SNPs $t$ and $t+1$, $n^t_{\text{switch}}$ be the number of sample haplotypes which switch into ancestry $k$
between SNPs $t$ and $t+1$, and $n_{k,s}^t$ be the number of reads that have a reference $s = 0$ or alternate $s = 1$ base for SNP $t$ that are in state $k$ for their central SNP.

Then the complete data log likelihood is

$$l(\lambda) = \log(P(O, Q|\lambda))$$

$$= \sum_{k=1}^{K} n_{k}^1 \log(\pi_k)$$

$$+ \sum_{t=1}^{T-1} n_{\text{stay}}^{t} \log(e^{-G \sigma_t}) + \sum_{t=1}^{T-1} \sum_{k=1}^{K} n_{\text{switch},k}^{t} \log((1 - e^{-G \sigma_t}) \alpha_{t,k})$$

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} n_{k,1}^{t} \log(\theta_{t,k}) + \sum_{t=1}^{T} \sum_{k=1}^{K} n_{k,0}^{t} \log((1 - \theta_{t,k})) \quad (18)$$

Calculating updates for a parameter is done by taking the derivative of $U(\lambda^{i+1}, \lambda^i)$ with respect to that parameter, setting it equal to 0 and solving. Employing the notation $E[x|O, \lambda] = E_{\lambda}[x]$, it is easy to calculate the following updates for $\lambda^{i+1} = (\pi^{i+1}, \theta^{i+1}, \alpha^{i+1}, \sigma^{i+1})$

$$\pi^{i+1}_k = \frac{E_{\lambda^i}[n_{k}^1]}{\sum_{j=1}^{K} E_{\lambda^i}[n_{j}^1]} \quad (19)$$

$$\theta^{i+1}_{t,k} = \frac{E_{\lambda^i}[n_{k,1}^{t}]}{E_{\lambda^i}[n_{k,0}^{t}] + E_{\lambda^i}[n_{k,1}^{t}]} \quad (20)$$

$$\alpha^{i+1}_{t,k} = \frac{E_{\lambda^i}[n_{\text{switch},k}^{t}]}{\sum_{j=1}^{K} E_{\lambda^i}[n_{\text{switch},j}^{t}]} \quad (21)$$

$$\sigma^{i+1}_t = \frac{1}{-G} \log \left( \frac{\sum_{k=1}^{K} E_{\lambda^i}[n_{\text{switch},k}^{t}]}{\sum_{k=1}^{K} E_{\lambda^i}[n_{\text{switch},k}^{t}] + E_{\lambda^i}[n_{\text{stay}}^{t}]} \right) \quad (22)$$

### 1.2.1 Useful Variables

We use a standard forward backward HMM implementation with a set of parameters $\lambda$. Recall that $q_t$ is the hidden state at SNP $t$. We use the following notations for states $k_t$ at SNP $t$ and $k_{t+1}$ at SNP $t+1$

$$\alpha_t(k_t) = P(O_1O_2...O_t, q_t = k_t|\lambda)$$

$$\beta_t(k_t) = P(O_{t+1}O_{t+2}...O_T|q_t = k_t, \lambda)$$

$$\gamma_t(k_t) = P(q_t = k_t|O, \lambda) = \frac{\alpha_t(k_t) \beta_t(k_t)}{P(O|\lambda)}$$

$$\xi_t(k_t, k_{t+1}) = P(q_t = k_t, q_{t+1} = k_{t+1}|O, \lambda)$$

$$= \frac{\alpha_t(k_t)P(q_{t+1} = k_{t+1}|q_t = k_t, \lambda)\beta_{t+1}(k_{t+1})P(O_{t+1}|q_{t+1} = k_{t+1}, \lambda)}{P(O|\lambda)}$$
Recall that \(\xi\) switch into state \(k\) the main text as probability that the sample is in the first state at SNP \(t\)

To update the prior parameters, we need the expectation of

\[ \text{Initial probabilities} \]

1.2.2 Haploid model

Define a variable \(I\) as the count of the number of sample haplotypes in state \(k\) at the first SNP. Denote the probability that the sample is in the first state at SNP \(t\) by \(\gamma_t(k)\). Let \(\gamma_{n,t}(k)\) be \(\gamma_t(k)\) for sample \(n\). We can therefore calculate the required expectation from the main text as

\[
E_{\lambda}[n_{k}^{1}] = \sum_{n=1}^{N} \gamma_{n,1}(k)
\]

(23)

1.2.2 Haploid model

Initial probabilities

To update the prior parameters, we need the expectation of \(n_{k}^{1}\), which we define as the number of sample haplotypes in state \(k\) at the first SNP. Denote the probability that the sample is in the first state at SNP \(t\) by \(\gamma_t(k)\). Let \(\gamma_{n,t}(k)\) be \(\gamma_t(k)\) for sample \(n\). We can therefore calculate the required expectation from the main text as

\[
E_{\lambda}[n_{k}^{1}] = \sum_{n=1}^{N} \gamma_{n,1}(k)
\]

(23)

Transition matrix probabilities

To update the transition parameters, we use an augmented state space where we have knowledge of how many recombinations occurred between two SNPs. Define a variable \(I_t\) as the count of the number of recombinations between SNPs \(t\) and \(t+1\); in the haploid model, this takes value 0 or 1. This will allow us to calculate the expectation of \(n_{\text{stay}}^{t}\), the number of sample haplotypes that do not recombine between SNPs \(t\) and \(t+1\), and \(n_{\text{switch},k}^{t}\), the number that switch into state \(k\) between SNPs \(t\) and \(t+1\).

We extend our transition probability to include \(I_t\) as follows

\[
P(q_{t+1} = k_{t+1}, I_t | q_t = k_t, \lambda) = \begin{cases} 
  e^{-G\sigma_t} & \text{if } k_t = k_{t+1} \text{ and } I_t = 0 \\
  0 & \text{if } k_t \neq k_{t+1} \text{ and } I_t = 0 \\
  (1 - e^{-G\sigma_t})\alpha_{t,k_{t+1}} & \text{if } I_t = 1 
\end{cases}
\]

Recall that \(\xi_t(k_t, k_{t+1})\) is

\[
\xi_t(k_t, k_{t+1}) = \frac{\alpha_t(k_t)P(q_{t+1} = k_{t+1}|q_t = k_t, \lambda)\beta_{t+1}(k_{t+1})P(O_{t+1}|q_{t+1} = k_{t+1}, \lambda)}{P(O|\lambda)}
\]

(24)
Denote the probability given the observed data $O$ that across SNP $t$, the sample has states $k_t$, $k_{t+1}$ and indicator $I_t$ by $\xi_t(k_t, k_{t+1}, I_t)$. Then

$$
\xi_t(k_t, k_{t+1}, I_t) = \frac{\alpha_t(k_t)P(q_{t+1} = k_{t+1}, I_t|q_t = k_t, \lambda)\beta_{t+1}(k_{t+1})P(O_{t+1}|q_{t+1} = k_{t+1}, \lambda)}{P(O|\lambda)} 
$$

Let $\xi_t(k_t, k_{t+1}, I_t)$ be $\xi_{n,t}(k_t, k_{t+1}, I_t)$ for sample $n$. We can therefore calculate expectations as

$$
E_\lambda[n_{\text{stay}}^t] = \sum_{n=1}^{N} \sum_{k=1}^{K} \xi_{n,t}(k, k, I_t) = 0
$$

$$
E_\lambda[n_{\text{switch}}^t] = \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{k=1}^{K} \xi_{n,t}(i, k, I_t) = 1
$$

and since

$$
E_\lambda[n_{\text{stay}}^t] = 1 - \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{k=1}^{K} \xi_{n,t}(i, k, I_t) = 1 - \sum_{k=1}^{K} E_\lambda[n_{\text{switch}}^t]
$$

it is therefore sufficient to calculate $E_\lambda[n_{\text{switch}}^t]$ to perform the EM updating from the main text.

**Emission matrix probabilities**

To update the emission parameters, we use an augmented state space where we have knowledge of whether emissions were due to the alternate or reference base. Recall that $\phi_{r,j}^i$ is the probability SNP $j$ in read $R_r$ came from a read with underlying genotype $i$. Denote by $H_{j}^r$ a variable which takes value 1 if the underlying base is the alternate base and 0 if it is the reference base. We will use this to calculate the expectation of $n_{s,a}^t$, the number of reads with a base at SNP $t$ that contain the alternate ($s=1$) or reference ($s=0$) base where the sample was in state $k$ at the central SNP of the read.

Recall that the original definition of the probability of read $R_r$ given hidden state $k$ at SNP $t$ and parameters $\lambda$ is

$$
P(R_r|q_t = k, \lambda) = \prod_{j=1}^{J_r} P(R_{r,j}|q_t = k, \lambda) = \prod_{j=1}^{J_r} (\phi_{r,j}^1 \theta_{u_{r,j},k} + \phi_{r,j}^0 (1 - \theta_{u_{r,j},k}))
$$

We extend our emission probability to include $H_{j}^r$ as follows

$$
P(R_r, H_{j}^r|q_t = k_t, \lambda) = \left\{ \begin{array}{ll} 
\prod_{i \neq j} P(R_{r,i}|q_t = k_t, \lambda) \phi_{r,j}^1 \theta_{u_{r,j},k} & \text{if } H_{j}^r = 1 \\
\prod_{i \neq j} P(R_{r,i}|q_t = k_t, \lambda) \phi_{r,j}^0 (1 - \theta_{u_{r,j},k}) & \text{if } H_{j}^r = 0
\end{array} \right.
$$

For read $R_r$ with central SNP $c_r$, the probability of the observation (set of reads)
at SNP \( t = c_r \) and \( H_f^r \) becomes

\[
P(O_t, H_f^r | q_t = k_t, \lambda) = \begin{cases} 
P(O_t | q_t = k_t, \lambda) \frac{\phi^l_{r,j} \theta_{u_{r,j}, k}}{\phi^l_{r,j} \theta_{u_{r,j}, k} + \phi^r_{r,j} (1 - \theta_{u_{r,j}, k})} & \text{if } H_f^r = 1 \\
P(O_t | q_t = k_t, \lambda) \frac{\phi^r_{r,j} (1 - \theta_{u_{r,j}, k})}{\phi^l_{r,j} \theta_{u_{r,j}, k} + \phi^r_{r,j} (1 - \theta_{u_{r,j}, k})} & \text{if } H_f^r = 0 
\end{cases}
\]

We expand \( \gamma_t(k_t) \) as

\[
\gamma_t(k_t) = \frac{\alpha_t(k_t) \beta_t(k_t)}{P(O | \lambda)} = \frac{\sum_{l=1}^{K} \alpha_{t-1}(l) P(q_t = k_t | q_{t-1} = l, \lambda) P(O_t | q_t = k, \lambda) \beta_t(k_t)}{P(O | \lambda)}
\]

where we note that for \( t = 1 \), we substitute \( \pi_k \) for \( \sum_{l=1}^{K} \alpha_{t-1}(l) P(q_t = k_t | q_{t-1} = l, \lambda) \). Denote the probability that for SNP \( j \) in read \( R_r \) with central SNP \( t = c_r \), the sample has a hidden state \( k_t \) and has indicator \( H_f^r \) given observed data \( O \) and parameters \( \lambda \) by \( \gamma_t(k_t, H_f^r) \). Then

\[
\gamma_t(k_t, H_f^r) = \frac{\sum_{l=1}^{K} \alpha_{t-1}(l) P(q_t = k_t | q_{t-1} = l, \lambda) P(O_t, H_f^r | q_t = k, \lambda)}{P(O | \lambda)}
\]

\[
\gamma_t(k_t, H_f^r) = \begin{cases} 
\gamma_t(k_t) \frac{\phi^l_{r,j} \theta_{u_{r,j}, k}}{\phi^l_{r,j} \theta_{u_{r,j}, k} + \phi^r_{r,j} (1 - \theta_{u_{r,j}, k})} & \text{if } H_f^r = 1 \\
\gamma_t(k_t) \frac{\phi^r_{r,j} (1 - \theta_{u_{r,j}, k})}{\phi^l_{r,j} \theta_{u_{r,j}, k} + \phi^r_{r,j} (1 - \theta_{u_{r,j}, k})} & \text{if } H_f^r = 0 
\end{cases}
\]

Let \( \gamma_n(k_t, H_f^r) \) be \( \gamma_t(k_t, H_f^r) \) for sample \( n \), and let \( A_n \) be the complete set of SNPs \( j \) from reads \( R_r \) for sample \( n \) such that \( u_{r,j} = t \). We can therefore calculate the required expectations from the main text as

\[
E_{\lambda}[n^t_{k,r} |] = \sum_{n=1}^{N} \sum_{(r,j) \in A_n} \gamma_n(k, H_f^r = 1) \tag{34}
\]

\[
E_{\lambda}[n^t_{k,r} |] = \sum_{n=1}^{N} \sum_{(r,j) \in A_n} \gamma_n(k, H_f^r = 0) \tag{35}
\]

### 1.2.3 Pseudo-haploid model

In the pseudo-haploid model, the only changes to the likelihood occur through the emissions; as such, we need to re-calculate Equations [34] and [35]. To update the emission parameters for the pseudo-haploid model, we use an augmented state space where we have knowledge of whether emissions were due to the alternate or reference base, and further have knowledge of whether the read came from the maternal or paternal haplotype. Recall that \( \phi^l_{r,j} \) is the probability that observed base \( j \) in read \( R_r \) came from a read with underlying genotype \( i \).
Recall that $H^j_t$ is an indicator variable which takes value 1 if the underlying base is the alternate base and 0 if it is the reference base. Let $H_r$ take value 1 if the read came from the maternal haplotype and 2 if it came from the paternal haplotype. We will use these to calculate the expectation of $n^j_{k,s}$, the number of reads that emit the alternate base ($s = 1$) or reference base ($s = 0$) given they are in state $k$ at the central SNP of the read.

Recall that for each individual, we make two forward backward passes of the algorithm, once for the maternal haplotype ($h = 1$), and a second time for the paternal haplotype ($h = 2$). We also attempt to probabilistically infer for each read which haplotype it came from. Let $H$ refer to the haplotype we are currently modelling (maternal or paternal).

First, recall that the original definition of the probability while modelling haplotype $h$ of read $R_r$ given hidden state $k$ at SNP $t$ and parameters $\lambda$ is

$$P_{H=h}(R_r | q_t = k_t, \lambda) = P(R_r | q_t = k_t, \lambda)P(H_r = h(O, \lambda)$$

$$+ P(R_r | H_r \neq h, \lambda)P(H_r \neq h(O, \lambda))$$  \hspace{1cm} (36)$$

For notational convenience, set $F_{r,j,h} = P(H_r = h | O, \lambda)\left[ \prod_{i \neq j} P(R_{r,i} | q_t = k, \lambda) \right]$. We therefore expand the emission probability to include $H^j_t$ and $H_r$ as follows

$$P_{H=h}(R_r, H^j_r, H_r | q_t = k, \lambda) = \begin{cases} F_{r,j,h} \theta_{ur,j,k} \phi^1_{r,j} & \text{if } H^j_t = 1, H_r = h \\ F_{r,j,h}(1 - \theta_{ur,j,k}) \phi^0_{r,j} & \text{if } H^j_t = 0, H_r = h \\ P(H_r \neq h | O, \lambda)P(R_r | H_r \neq h, \lambda) & \text{if } H_r \neq h \end{cases}$$

Denote the probability that haplotype $h$ of the sample is in state $k_t$ at SNP $t$ with $H^j_r$ and $H_r$ given observed data $O$ and parameters $\lambda$ by $\gamma_{t,h}(k_t, H^j_r, H_r)$. Then, we get that

$$\gamma_{t,h}(k_t, H^j_r, H_r) = \begin{cases} \gamma_{t,h}(k_t) \frac{F_{r,j,h} \theta_{ur,j,k} \phi^1_{r,j}}{P_{H=h}(R_r | q_t = k_t, \lambda)} & \text{if } H^j_t = 1, H_r = h \\ \gamma_{t,h}(k_t) \frac{F_{r,j,h}(1 - \theta_{ur,j,k}) \phi^0_{r,j}}{P_{H=h}(R_r | q_t = k_t, \lambda)} & \text{if } H^j_t = 0, H_r = h \end{cases}$$

Let $\gamma_{n,t,h}(k_t, H^j_r, H_r)$ be $\gamma_{t,h}(k_t, H^j_r, H_r)$ for sample $n$, and let $A_n$ be the complete set of SNPs $j$ from reads $R_r$ for sample $n$ such that $ur,j = t$. We can therefore calculate the required expectations from the main text as

$$\mathbb{E}_\lambda[n^1_{k,s}] = \sum_{n=1}^N \sum_{(r,j) \in A_n} \sum_{h=1}^2 \gamma_{n,c,r,h}(k, H^j_r = 1, H_r = h)$$  \hspace{1cm} (37)$$

$$\mathbb{E}_\lambda[n^0_{k,s}] = \sum_{n=1}^N \sum_{(r,j) \in A_n} \sum_{h=1}^2 \gamma_{n,c,r,h}(k, H^j_r = 0, H_r = h)$$  \hspace{1cm} (38)$$

1.2.4 Diploid model

Initial probabilities
To update the prior parameters, we need the expectation of \( n_t^k \), which we define as the number of sample haplotypes in state \( k \) at the first SNP. Denote the probability that sample \( n \) is in pairs of states \((k_{t,1}, k_{t,2})\) at SNP \( t \) given observed data \( O \) by \( \gamma_{n,t}(k_{t,1}, k_{t,2}) \). We can therefore calculate the required expectation from the main text as

\[
E_\lambda [n_t^k] = \sum_{n=1}^{N} \sum_{j=1}^{K} (\gamma_{n,1}(k,j) + \gamma_{n,1}(j,k))
\]

**Transition probabilities**

To update the transition parameters for the diploid model, we use an augmented state space where we have knowledge of how many recombinations occurred between two SNPs. Here we define a variable \( I_t \) which counts the number of recombinations that occur between SNPs \( t \) and \( t+1 \) for the two haplotypes of the diploid sample, and takes values 0, 1 or 2. This will allow us to calculate the expectation of \( n_{t,stay}^k \), the number of sample haplotypes that do not recombine between SNPs \( t \) and \( t+1 \), and \( n_{t,switch,k}^t \), the number of sample haplotypes that switch into state \( k \) between SNPs \( t \) and \( t+1 \).

We can therefore extend the diploid transition probability to include \( I_t \) by multiplying the haploid transition probabilities as follows

\[
P(q_{t+1} = (k_{t+1,1}, k_{t+1,2}), I_t | q_t = (k_{t,1}, k_{t,2}), \lambda) =
\]

\[
\begin{cases}
  e^{-2G \sigma_t} & \text{if } I_t = 0 \text{ and } k_{t+1,1} = k_{t,1} \text{ and } k_{t+1,2} = k_{t,2} \\
  e^{-G \sigma_t} (1 - e^{-G \sigma_t}) \alpha_{t,k_{t+1,1}} & \text{if } I_t = 1 \text{ and } k_{t+1,1} \neq k_{t,1} \text{ and } k_{t+1,2} = k_{t,2} \\
  e^{-G \sigma_t} (1 - e^{-G \sigma_t}) \alpha_{t,k_{t+1,2}} & \text{if } I_t = 1 \text{ and } k_{t+1,1} = k_{t,1} \text{ and } k_{t+1,2} \neq k_{t,2} \\
  e^{-G \sigma_t} (1 - e^{-G \sigma_t}) (\alpha_{t,k_{t+1,1}} + \alpha_{t,k_{t+1,2}}) & \text{if } I_t = 1 \text{ and } k_{t+1,1} = k_{t,1} \text{ and } k_{t+1,2} = k_{t,2} \\
  (1 - e^{-G \sigma_t})^2 \alpha_{t,k_{t+1,1}} \alpha_{t,k_{t+1,2}} & \text{if } I_t = 2 \\
  0 & \text{otherwise}
\end{cases}
\]

Denote the probability under the diploid model that the sample is in states \((k_{t,1}, k_{t,2})\) at SNP \( t \) and states \((k_{t+1,1}, k_{t+1,2})\) at SNP \( t+1 \) and has indicator variable \( I_t \) given observed data \( O \) and parameters \( \lambda \) by \( \xi_t((k_{t,1}, k_{t,2}), (k_{t+1,1}, k_{t+1,2}), I_t) \). Then

\[
\xi_t((k_{t,1}, k_{t,2}), (k_{t+1,1}, k_{t+1,2}), I_t) = \frac{1}{P(O | \lambda)} \alpha_{t,k_{t,1}} \beta_{t+1,k_{t+1,1},k_{t+1,2}} P(O_{t+1} | q_t = (k_{t,1}, k_{t,2}), \lambda) \\
P(q_{t+1} = (k_{t+1,1}, k_{t+1,2}), I_t | q_t = (k_{t,1}, k_{t,2}), \lambda)
\]

Let \( m_{switch,k}^t \) be the number of haplotypes of the sample that switch into state \( k \) between SNPs \( t \) and \( t+1 \). We can calculate \( E_\lambda [m_{switch,k}^t] \), and from this \( E_\lambda [n_{t,stay}^k] \), by summing across \( E_\lambda [m_{switch,k}^t] \) for all \( N \) samples, and so we can calculate the required expectations from the main text by performing the calculations below. Note that we simplify the summation to give a formulation that enables
quadratic versus linear computational complexity in $K$. A similar approach is done for the haploid model to achieve linear versus quadratic computational complexity (not shown).

$$
\mathbb{E}_\lambda [m_{\text{switch,k}}] = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{k_3=1}^{K} \sum_{j=0}^{2} j \times \left( \xi_t ((k_1, k_2), (k, k_3), I_t = j) + \xi_t ((k_1, k_2), (k, k_3), I_t = j) \right)
= \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} 1 \times \left( \xi_t ((k_1, k_2), (k, k_2), I_t = 1) + \xi_t ((k_1, k_2), (k, k_1), I_t = 1) \right)
+ \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{k_3=1}^{K} 2 \times \left( \frac{1}{2} \xi_t ((k_1, k_2), (k, k_3), I_t = 2) + \frac{1}{2} \xi_t ((k_1, k_2), (k, k_3), I_t = 2) \right)
$$

$$
= \sum_{k_1=1}^{K} \sum_{k_3=1}^{K} 2 \times \xi_t ((k_1, k_3), (k, k_3), I_t = 1)
+ \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{k_3=1}^{K} 2 \times \xi_t ((k_1, k_2), (k, k_3), I_t = 2)
\tag{42}
$$

$$
= 2 \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{k_3=1}^{K} \alpha_t (k_1, k_3) \beta_{t+1} (k, k_3) P(O_{t+1} | q_t = (k, k_3), \lambda) \alpha_{t,k} (1 - e^{-G_{\sigma_t}}) e^{-G_{\sigma_t}} P(O | \lambda)
+ 2 \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{k_3=1}^{K} \alpha_t (k_1, k_2) \beta_{t+1} (k, k_3) P(O_{t+1} | q_t = (k, k_3), \lambda) \alpha_{t,k} \alpha_{t,k_3} (1 - e^{-T_{\sigma_t}})^2 P(O | \lambda)
$$

$$
= \frac{2 \alpha_{t,k}}{P(O | \lambda)} \sum_{k_3=1}^{K} \left( 1 - e^{-G_{\sigma_t}} e^{-G_{\sigma_t}} \sum_{k_1=1}^{K} \alpha_t (k_1, k_3) \right)
+ \alpha_{k_3}^t (1 - e^{-G_{\sigma_t}})^2 \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \alpha_t (k_1, k_2) \beta_{t+1} (k, k_3) P(O_{t+1} | q_t = (k, k_3), \lambda)
\tag{44}
$$

**Emission probabilities**

To update the emission parameters for the diploid model, we use an augmented state space as in for the pseudo-haploid model where we have knowledge of whether emissions were due to the alternate or reference base, and further have knowledge of whether the read came from the maternal or paternal haplotype. Recall that: $\phi_{t,j}^k$ is the probability that observed base $j$ in read $R_j$ came from a read with underlying genotype $i$; $H^j_r$ is an variable which takes value 1 if the underlying base is the alternate base and 0 if it is the reference base; and $H^r$ is a variable that takes value 1 if the read came from the maternal haplotype and 2 if from the paternal haplotype. We will use these to calculate the expectation of $n_{k,s}^t$, the number of reads that emit the alternate base ($s = 1$) or reference base ($s = 0$) given they are in state $k$ at their central SNP.
Recall from the main text that the probability of an observation (set of reads) at SNP \( t \) in the diploid model is

\[
P(O_t | q_t = (k_{t,1}, k_{t,2}), \lambda) = \frac{1}{2} P(R_r | q_t = k_{t,1}, \lambda) + \frac{1}{2} P(R_r | q_t = k_{t,2}, \lambda) \quad (45)
\]

For notational convenience set

\[
F_{r,j,H_r} = \frac{1}{2} P(R_r | q_t = k_{H_r}, \lambda) \left( \frac{1}{2 \theta_{t,k_{t,H_r}} \phi_{r,j}^1} + (1 - \theta_{t,k_{t,H_r}}) \phi_{r,j}^0 \right)
\]

We can therefore calculate the probability that SNP \( j \) in read \( R_r \) with central SNP \( c_r \) has indicator variable \( H^j_r \) and \( H_r \) and observation for SNP \( t = c_r \) of \( O_t \) given the pair of hidden states \( (k_{t,1}, k_{t,2}) \) and parameters \( \lambda \) as

\[
P(O_t, H^j_r, H_r | q_t = (k_{t,1}, k_{t,2}), \lambda) = \begin{cases} 
P(O_t, | q_t = (k_{t,1}, k_{t,2}), \lambda) F_{r,j,H_r} \theta_{t,k_{t,H_r}} \phi_{r,j}^1 & \text{if } H^j_r = 1 \\ P(O_t, | q_t = (k_{t,1}, k_{t,2}), \lambda) F_{r,j,H_r} (1 - \theta_{t,k_{t,H_r}}) \phi_{r,j}^0 & \text{if } H^j_r = 0 \end{cases}
\]

Denote the probability for SNP \( j \) in read \( R_r \) that at the central SNP of the read \( t = c_r \) is in the pair of states \( (k_{t,1}, k_{t,2}) \) given the observed data \( O \) and parameters \( \lambda \) by \( \gamma_t(k_{t,1}, k_{t,2}, H^j_r, H_r) \). Then

\[
\gamma_t(k_{t,1}, k_{t,2}, H^j_r, H_r) = \begin{cases} 
\gamma_t(k_{t,1}, k_{t,2}) F_{r,j,H_r} \theta_{t,k_{t,H_r}} \phi_{r,j}^1 & \text{if } H^j_r = 1 \\
\gamma_t(k_{t,1}, k_{t,2}) F_{r,j,H_r} (1 - \theta_{t,k_{t,H_r}}) \phi_{r,j}^0 & \text{if } H^j_r = 0 
\end{cases}
\]

Let \( \gamma_n(k_{t,1}, k_{t,2}, H^j_r, H_r) \) be \( \gamma_t(k_{t,1}, k_{t,2}, H^j_r, H_r) \) for sample \( n \), and let \( A_n \) be the complete set of SNPs \( j \) and reads \( R_r \) for sample \( n \) such that \( u_{r,j} = t \). We can calculate the required expectations from the main text as

\[
E_{\lambda}[n^k_{k,s}] = \sum_{n=1}^{N} \sum_{(r,j) \in A_n} \sum_{i=1}^{K} \left( \gamma_{n,c_r}(k, i, H^j_r = s, H_r = 1) + \gamma_{n,c_r}(i, k, H^j_r = s, H_r = 2) \right) (47)
\]

### 1.3 Efficient calculation of forward backward variables

We take the time here to write out the forward backwards calculations that we used for the diploid case, as symmetries in the transition matrix allow us to make the calculation in quadratic, rather than quartic time with respect to \( K \). Similar calculations (not shown) are used for the haploid model to ensure linear versus quadratic computational complexity in \( K \). We note that these calculations are not original and are given in very similar form in the original fastPHASE paper [2], but we reproduce them here as they represent important simplifications for computational reasons.
\[
\alpha_{t+1}(k_3, k_4) = \left[ \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \alpha_t(k_1, k_2) P(q_{t+1} = (k_3, k_4)|q_t = (k_1, k_2), \lambda) \right] P(O_{t+1}|q_{t+1} = (k_3, k_4), \lambda) \\
= \left[ \alpha_t(k_3, k_4)(e^{-G_{\sigma t}})^2 + \sum_{k=1}^{K} e^{-G_{\sigma t}}(1 - e^{-G_{\sigma t}})\alpha_{t,k_3}\alpha_t(k, k) + \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} (1 - e^{-G_{\sigma t}})^2\alpha_{t,k_3}\alpha_{t,k_4}\alpha_t(k_1, k_2) \right] P(O_{t+1}|q_{t+1} = (k_3, k_4)) \\
= \left[ \alpha_t(k_3, k_4)(e^{-G_{\sigma t}})^2 + \alpha_{t,k_3}A_{t,1}(k_4) + \alpha_{t,k_4}A_{t,2}(k_3) + \alpha_{t,k_3}\alpha_{t,k_4}B_t \right] \times P(O_{t+1}|q_{t+1} = (k_3, k_4), \lambda)
\]

where

\[
A_{t,1}(k_4) = e^{-G_{\sigma t}}(1 - e^{-G_{\sigma t}})\sum_{k=1}^{K} \alpha_t(k, k_4) \\
A_{t,2}(k_3) = e^{-G_{\sigma t}}(1 - e^{-G_{\sigma t}})\sum_{k=1}^{K} \alpha_t(k_3, k) \\
B_t = (1 - e^{-G_{\sigma t}})^2 \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \alpha_t(k_1, k_2)
\]

As such, the forward calculation can be done in quadratic time with respect to the number of ancestral haplotypes \( K \).

Similarly, for the backward calculation we get that

\[
\beta_t(k_1, k_2) = \sum_{k_3=1}^{K} \sum_{k_4=1}^{K} P(q_{t+1} = (k_3, k_4)|q_t = (k_1, k_2), \lambda)P(O_{t+1}|q_{t+1} = (k_3, k_4), \lambda)\beta_{t+1}(k_3, k_4) \\
= (e^{-G_{\sigma t}})^2 P(O_{t+1}|q_{t+1} = (k_1, k_2), \lambda)\beta_{t+1}(k_3, k_4) + (e^{-G_{\sigma t}})(1 - e^{-G_{\sigma t}}) \left( \sum_{k=1}^{K} \alpha_{t,k}P(O_{t+1}|q_{t+1} = (k, k_2), \lambda)\beta_{t+1}(k, k_2) + \sum_{k=1}^{K} \alpha_{t,k}P(O_{t+1}|q_{t+1} = (k_1, k), \lambda)\beta_{t+1}(k_1, k) \right) + (1 - e^{-G_{\sigma t}})^2 \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \alpha_{t,k_3}\alpha_{t,k_4}P(O_{t+1}|q_{t+1} = (k_3, k_4), \lambda)\beta_{t+1}(k_3, k_4) \\
= (e^{-G_{\sigma t}})^2 P(O_{t+1}|q_{t+1} = (k_1, k_2), \lambda)\beta_{t+1}(k, k_2) + E_{t,1}(k_2) + E_{t,2}(k_1) + F_t
\]
where

\[ E_{t,1}(k_2) = (e^{-G\sigma_t})(1 - e^{-G\sigma_t}) K \sum_{k=1}^{K} \alpha_{t,k} P(O_{t+1}|q_{t+1} = (k, k_2), \lambda) \beta_{t+1}(k, k_2) \]

(51)

\[ E_{t,2}(k_1) = (e^{-G\sigma_t})(1 - e^{-G\sigma_t}) K \sum_{k=1}^{K} \alpha_{t,k} P(O_{t+1}|q_{t+1} = (k_1, k), \lambda) \beta_{t+1}(k_1, k) \]

(52)

\[ F_t = (1 - e^{-G\sigma_t})^2 K \sum_{k_3=1}^{K} \sum_{k_4=1}^{K} \alpha_{t,k_3} \alpha_{t,k_4} P(O_{t+1}|q_{t+1} = (k_3, k_4), \lambda) \beta_{t+1}(k_3, k_4) \]

(53)

1.4 Initialization

Haploid probabilities \( \pi_k \) are initialized with equal weights \( \pi_k = \frac{1}{K} \), as are diploid priors \( \pi_{k_1,k_2} = \frac{1}{K^2} \). The state probabilities \( \alpha_{t,k} \) are also initialized with equal weights \( \alpha_{t,k} = \frac{1}{K} \). The recombination distance is initialized assuming a constant recombination rate multiplied by the physical distance between SNPs, for example assuming \( \sigma_t = d_t \times 0.5\text{cM/Mb} \) where \( d_t \) is the physical distance between SNPs \( t \) and \( t+1 \). Finally, given a lower bound \( \delta \) on emission probabilities, for example \( \delta = 0.0001 \), \( \theta_{t,k} \) are sampled from a uniform distribution with minimum value \( \delta \) and maximum value \( 1 - \delta \). Note that \( G \) is left as a user set parameter, which can be approximated for outbred populations using external estimates of \( N_e \) with \( G = \frac{4N_e}{K} \).

1.5 Parameter bounding

After parameter updating, newly calculated parameters are bounded with default but user tunable parameters. Prior probabilities \( \pi_k \), new state parameters \( \alpha_{t,k} \), and emission probabilities \( \theta_{t,k} \) (and \( 1 - \theta_{t,k} \)) whose values are less than a threshold are set equal to that threshold, and then probabilities re-normalized as appropriate to have sum 1. Under default conditions this bound is \( 1 \times 10^{-4} \). For the recombination distance, values of \( \sigma_t \) that exceed implied upper (default 100 cM/Mb) and lower (default 0.1 cM/Mb) bounds are reset to the bound value.

1.6 Heuristics

Since emission probabilities \( \theta \) are initialized at random, STITCH can get stuck in local minima, for which two heuristics are employed at various (default) iterations. First, to help overcome unnecessary switches between ancestral haplotype backgrounds, at iterations 4, 8, 12 and 16, pairs of haplotype states are calculated for each sample between pairs of nearby SNPs (starting at SNP 51, then every further 100th SNP) by multiplying their marginal ancestral haplotype
probabilities. If, across all samples, for each pair of nearby SNPs, there exists a re-ordering of ancestral haplotype states that minimizes switching, then that switch, or switches, is performed, and local SNPs (plus or minus 20 from the break) are reset with $\theta$ from a $U(0,1)$ distribution. Second, to help fill unused ancestral haplotypes, and to overcome superimposed ancestral haplotypes, at iterations 6, 10, 14 and 18, ancestral haplotype usage in the most recent iteration is discretized by averaging over 100 SNP intervals, and every continuous interval of infrequently used ancestral haplotype ($<0.5\%$) is identified. Values of $\theta$ over each interval are then refilled for that ancestral haplotype by copying from another sampled ancestral haplotype chosen with sampling probability proportional to ancestral haplotype usage over that interval. $\theta$ is then reset using 80% of these filled values and 20% noise from a $U(0,1)$ distribution.

1.7 Guidance behind parameter options

STITCH contains many parameter options that can be modified by the user, for example upper and lower bounds on recombination rate. However, most of these are reasonable for the majority of anticipated applications of STITCH. For the analyses presented here for the CFW and CONVERGE populations, we varied: $K$ (option K), the number of ancestral haplotypes; whether the diploid or pseudo-haploid method was used (option method); the number of pseudo-haploid iterations (option switchModelIteration): the number of generations when the population was founded (or can be so approximated) $G$ (option nGen) (which we set as 100 for the CFW analyses and $4\times20000K$ for the CONVERGE studies). We also, for model evaluation purposes only, invoked a flag on whether reads were split into new reads containing one SNP each (option readAware), the number of computer cores available to the process (option nCores), and whether the process is running in a server or cluster environment (option environment).

We anticipate that in using STITCH, the majority of users will achieve desired results, both in terms of accuracy and computational speed, through varying $K$, $G$, the method (diploid or pseudo-haploid), and the number of pseudo-haploid iterations.

In terms of selecting $K$, the diploid or pseudo-haploid method, and the number of pseudo-haploid iterations, we recommend imputing a small region of the genome, such as a chromosome, using the diploid mode with a range of $K$, and then evaluate performance. We recommend that to evaluate imputation performance, users obtain validation data, using either genotyping microarrays or higher coverage sequencing (like 10X). In the absence of external validation data, we recommend the info score distribution or its average. If, for the diploid method and a choice of $K$, results start to deteriorate, then choose the diploid mode and $K$ that gave optimal performance. If results do not deteriorate but become computationally impractical, we recommend applying the pseudo-haploid method for a range of pseudo-haploid and diploid iterations (as was done here for CONVERGE), and choosing the combination that gives optimal results under the given computational constraints.

For $G$ (or nGen), we recommend setting this to a reasonable $a\ priori$ esti-
mate, like was available for the CFW mice, or to use $4 \times \frac{N_e}{K}$, when the population is wild or has not been through a strong bottleneck. We note that STITCH should be fairly robust to this parameter choice. Users may also increase the minimum and maximum allowed recombination rates if they are less certain about this parameter.

Finally, while we do not give specific guidance on study design strategies and sequencing depths, we note that in designing low coverage sequencing only studies, users should try to ensure adequate population sequencing coverage to ensure the ancestral haplotypes are well reconstructed, particularly in the case when the founding structure is well known. For example, if a population was founded with $K = 8$ haplotypes, then to achieve a given level of per-ancestral haplotype coverage (e.g. 30X), while sequencing each sample at a given level (e.g. 0.2X), one should consider sequencing in excess of $\frac{30 \times K}{0.2} = 1200$ samples. Drift in the population (i.e. non-equal ancestral haplotype usage in the population) would require additional samples or depth for reconstruction of rare haplotypes in the population.
2 Supplementary Tables
### Supplementary Table 1A: Genotype concordance for CFW using STITCH (K=4, diploid) at all SNPs

Results give genotype concordance stratified by genotype class and allele frequency. Discrete genotype calls are generated for imputation as the genotype with the maximum genotype posterior probability. Results are given genome-wide (autosomal and chromosome X). Allele freq = allele frequencies are the frequency of the minor allele. Type is either High Cov = high coverage (10X) sequencing (4 samples) or Array = MegaMuga (44 samples). Columns contain either Num = Number of non-missing genotypes considered (samples times SNPs for sequencing or array), or Per = Percent of imputed best guess genotypes that match sequencing or array genotypes. Hom Major = homozygous for the major allele, Het = heterozygous, Hom Minor = homozygous for the minor allele. Note that truth (sequencing or array) genotypes contain some missing data.

| Allele freqs | Type  | Num  | Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|-------|------|-----------|---------------|---------|---------|--------------|--------------|
| [0.0,0.01)   | High Cov | 1,139,724 | 99.98   | 3.958         | 17.08   | 14      | 0            |              |
| [0.01,0.02)  | High Cov | 1,016,641 | 99.84   | 20,516        | 64.72   | 124     | 4.84         |              |
| [0.02,0.05)  | High Cov | 3,756,581 | 99.71   | 213,186       | 81.62   | 3,407   | 52.51        |              |
| [0.05,0.1]   | High Cov | 4,072,071 | 99.57   | 552,747       | 89.85   | 22,958  | 75.8         |              |
| [0.1,0.2)    | High Cov | 3,781,946 | 99.23   | 1,164,122     | 92.84   | 117,126 | 90.83        |              |
| [0.2,0.3)    | High Cov | 1,973,117 | 98.69   | 1,274,057     | 94.86   | 204,474 | 93.69        |              |
| [0.3,0.4]    | High Cov | 1,201,299 | 98.08   | 1,296,792     | 95.83   | 328,621 | 95.31        |              |
| [0.4,0.5]    | High Cov | 730,043  | 97.29   | 1,417,056     | 96.59   | 452,854 | 96.13        |              |
| [0.0,0.01)   | Array  | 3,101  | 99.97   | 106           | 56.6    | 3       | 33.33        |              |
| [0.01,0.02)  | Array  | 19,788 | 99.93   | 803           | 84.43   | 20      | 50           |              |
| [0.02,0.05)  | Array  | 135,504 | 99.9    | 9,386         | 93.51   | 312     | 73.08        |              |
| [0.05,0.1)   | Array  | 161,620 | 99.86   | 24,850        | 95.99   | 1,238   | 81.18        |              |
| [0.1,0.2)    | Array  | 163,416 | 99.76   | 55,438        | 97.59   | 5,965   | 94.25        |              |
| [0.2,0.3)    | Array  | 82,595  | 99.57   | 54,880        | 98.27   | 9,586   | 97.83        |              |
| [0.3,0.4]    | Array  | 45,709  | 99.33   | 49,416        | 98.36   | 14,094  | 98.24        |              |
| [0.4,0.5]    | Array  | 33,823  | 99.21   | 53,605        | 98.79   | 22,887  | 98.93        |              |
### Supplementary Table 1B: Genotype concordance for CFW using Beagle (default) at all SNPs

| Allele freqs | Type   | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|--------|---------------|---------------|---------|---------|---------------|---------------|
| [0.0,0.01)   | High Cov | 1,139,728 | 95            | 3,958   | 19.45   | 10            | 0             |
| [0.01,0.02)  | High Cov | 1,016,641 | 91.05         | 20,516  | 14.9    | 124           | 0             |
| [0.02,0.05)  | High Cov | 3,756,615 | 90.2          | 213,186 | 16.04   | 3,373         | 0.18          |
| [0.05,0.1)   | High Cov | 4,072,282 | 84.46         | 552,747 | 23.3    | 22,747        | 0.32          |
| [0.1,0.2)    | High Cov | 3,782,741 | 65.12         | 1,164,122 | 45.7   | 116,331       | 0.87          |
| [0.2,0.3)    | High Cov | 1,973,436 | 23.14         | 1,274,072 | 84.16  | 204,155       | 2.16          |
| [0.3,0.4)    | High Cov | 1,196,914 | 14.13         | 1,296,792 | 90.73  | 333,006       | 3.85          |
| [0.4,0.5]    | High Cov | 718,518  | 13.02         | 1,417,056 | 90.02  | 464,379       | 6.1           |
| [0,0.01)     | Array   | 3,101       | 91.36         | 106     | 17.92   | 3             | 0             |
| [0.01,0.02)  | Array   | 19,788     | 93.68         | 803     | 15.44   | 20            | 0             |
| [0.02,0.05)  | Array   | 135,504    | 90.06         | 9,386   | 17.45   | 312           | 0             |
| [0.05,0.1)   | Array   | 161,581    | 84.58         | 24,850  | 23.57   | 1,277         | 0.23          |
| [0.1,0.2)    | Array   | 163,393    | 63.19         | 55,438  | 48.4    | 5,988         | 0.78          |
| [0.2,0.3)    | Array   | 82,553     | 21.1          | 54,880  | 88.32   | 9,628         | 1.65          |
| [0.3,0.4)    | Array   | 45,562     | 15.71         | 49,416  | 92.11   | 14,241        | 1.94          |
| [0.4,0.5]    | Array   | 33,002     | 15.13         | 53,605  | 91.53   | 23,708        | 3.28          |
Supplementary Table 1C: Genotype concordance for CFW using findhap (maxlen=10000, minlen=100, steps=3, iters=4) at all SNPs

| Allele freqs | Type  | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|-------|---------------|---------------|---------|---------|---------------|---------------|
| [0,0.01)     | High Cov | 1,138,593    | 96.5          | 3,958   | 30.19   | 1,145         | 12.4          |
| [0.01,0.02)  | High Cov | 1,012,728    | 92.28         | 20,516  | 52.67   | 4,037         | 9.14          |
| [0.02,0.05)  | High Cov | 3,739,212    | 89.72         | 213,186 | 81.45   | 20,776        | 13.59         |
| [0.05,0.1)   | High Cov | 4,040,317    | 89.29         | 552,747 | 86.98   | 54,712        | 15.61         |
| [0.1,0.2)    | High Cov | 3,730,302    | 91.25         | 1,164,122 | 79.28   | 168,770       | 34.79         |
| [0.2,0.3)    | High Cov | 1,966,686    | 93.17         | 1,274,072 | 71.31   | 210,905       | 51.8          |
| [0.3,0.4)    | High Cov | 1,198,042    | 90.45         | 1,296,792 | 62.92   | 331,878       | 64.81         |
| [0.4,0.5]    | High Cov | 722,465      | 87.28         | 1,417,056 | 57.79   | 460,432       | 74.34         |
| [0.001)      | Array   | 3,101        | 97.84         | 106     | 37.74   | 3             | 0             |
| [0.01,0.02)  | Array   | 19,745       | 93.78         | 803     | 69.12   | 63            | 7.94          |
| [0.02,0.05)  | Array   | 135,174      | 90.06         | 9,386   | 86.74   | 642           | 19.47         |
| [0.05,0.1)   | Array   | 160,288      | 87.06         | 24,850  | 89.34   | 2,570         | 18.6          |
| [0.1,0.2)    | Array   | 161,093      | 89.2          | 55,438  | 82.78   | 8,288         | 33.69         |
| [0.2,0.3)    | Array   | 82,595       | 91.56         | 54,880  | 73.4    | 9,586         | 52.22         |
| [0.3,0.4)    | Array   | 45,673       | 87.74         | 49,416  | 68      | 14,130        | 61.78         |
| [0.4,0.5]    | Array   | 33,145       | 83.72         | 53,605  | 63.99   | 23,565        | 69.89         |
Supplementary Table 1D: Genotype concordance for CFW using STITCH (K=4, diploid) at all SNPs (post QC)

| Allele freqs | Type   | Num Hom Major | Per Hom Major | Num Het  | Per Het  | Num Hom Minor | Per Hom Minor |
|--------------|--------|---------------|---------------|----------|----------|---------------|---------------|
| [0.001)      | High Cov | 1,139,724     | 99.98         | 3,958    | 17.08    | 14            | 0             |
| [0.001,0.02) | High Cov | 1,016,641     | 99.84         | 20,516   | 64.72    | 124           | 4.84          |
| [0.02,0.05)  | High Cov | 3,756,581     | 99.71         | 213,186  | 81.62    | 3,407         | 52.51         |
| [0.05,0.1)   | High Cov | 4,072,071     | 99.57         | 552,747  | 89.85    | 22,958        | 75.8          |
| [0.1,0.2)    | High Cov | 3,781,946     | 99.23         | 1,164,122| 92.84    | 117,126       | 90.83         |
| [0.2,0.3)    | High Cov | 1,973,117     | 98.69         | 1,274,072| 94.86    | 204,474       | 93.69         |
| [0.3,0.4)    | High Cov | 1,201,299     | 98.08         | 1,296,792| 95.83    | 328,621       | 95.31         |
| [0.4,0.5]    | High Cov | 730,043       | 97.29         | 1,417,056| 96.59    | 452,854       | 96.13         |
| [0.001)      | Array   | 3,101         | 99.97         | 106      | 56.6     | 3             | 33.33         |
| [0.01,0.02)  | Array   | 19,788        | 99.93         | 803      | 84.43    | 20            | 50            |
| [0.02,0.05)  | Array   | 135,504       | 99.9          | 9,386    | 93.51    | 312           | 73.08         |
| [0.05,0.1)   | Array   | 161,620       | 98.86         | 24,850   | 95.99    | 1,238         | 81.18         |
| [0.1,0.2)    | Array   | 163,416       | 99.76         | 55,438   | 97.59    | 5,965         | 94.25         |
| [0.2,0.3]    | Array   | 82,595        | 99.57         | 54,880   | 98.27    | 9,586         | 97.83         |
| [0.3,0.4]    | Array   | 45,709        | 99.33         | 49,416   | 98.36    | 14,094        | 98.24         |
| [0.4,0.5]    | Array   | 33,823        | 99.21         | 53,605   | 98.79    | 22,887        | 98.93         |
Supplementary Table 2: Performance of CFW study under different programs and options

Results are given for chromosomes 18 and 19. All STITCH results are for the diploid model with 40 iterations. Program options are as follows. For STITCH, RU refers to read unaware (i.e. split each read spanning multiple SNPs into sub-reads spanning one read each). For Beagle, shown are the number of iterations (i.e. burnin-its, phase-its, and impute-its to this value), window is the window size, and msf is the (singlescale) model scale factor. For findhap, options correspond directly to parameter options. Note that times for STITCH do not include the generation of input data from BAMs, which took about 1-1.5 hours per chromosome for chromosomes 18 and 19, irrespective of other program options. Similarly, times for findhap do not include conversion time from VCF to the findhap input format. Av r^2 is the average r^2 for SNPs on the Illumina MegaMUGA array, with no filtration for QC for any method. Time is the average time in hours for chromosomes 18 and 19, where all programs were run on 1 core on 2.60 GHz Intel E5-2650 chips.

| Program | Options                      | Time | Av r^2 |
|---------|-------------------------------|------|--------|
| STITCH  | K=2                           | 7.2  | 0.622  |
| STITCH  | K=3                           | 11.3 | 0.957  |
| STITCH  | K=4                           | 18.6 | 0.972  |
| STITCH  | K=5                           | 25.3 | 0.97   |
| STITCH  | K=6                           | 37.5 | 0.966  |
| STITCH  | K=7                           | 49   | 0.964  |
| STITCH  | K=8                           | 59   | 0.967  |
| STITCH  | K=4, RU                       | 18.8 | 0.873  |
| Beagle  | its=5, window=50000, msf=1    | 6.1  | 0.074  |
| Beagle  | its=5, window=100000, msf=1   | 4.7  | 0.073  |
| Beagle  | its=10, window=50000, msf=1   | 17.2 | 0.085  |
| Beagle  | its=20, window=50000, msf=1   | 34.1 | 0.109  |
| Beagle  | its=5, window=50000, msf=0.4  | 72.4 | 0.088  |
| Beagle  | its=5, window=50000, msf=0.6  | 7.7  | 0.079  |
| Beagle  | its=5, window=50000, msf=0.8  | 6.6  | 0.073  |
| Beagle  | its=5, window=50000, msf=1.0  | 5.3  | 0.072  |
| Beagle  | its=5, window=50000, msf=1.2  | 4.9  | 0.071  |
| Beagle  | its=5, window=50000, msf=1.4  | 5.7  | 0.071  |
| Beagle  | its=5, window=50000, msf=1.6  | 5.2  | 0.071  |
| Beagle  | its=5, window=50000, msf=1.8  | 5.2  | 0.071  |
| Beagle  | its=5, window=50000, msf=2.0  | 5.1  | 0.071  |
| findhap | maxlen=100000, minlen=1000, steps=3, iters=4 | 0.6  | 0.225  |
| findhap | maxlen=100000, minlen=1000, steps=2, iters=6 | 0.7  | 0.226  |
| findhap | maxlen=100000, minlen=1000, steps=5, iters=10 | 2.2  | 0.15   |
| findhap | maxlen=10000, minlen=100, steps=3, iters=4 | 0.5  | 0.523  |
| findhap | maxlen=50000, minlen=500, steps=3, iters=4 | 0.6  | 0.281  |
| findhap | maxlen=200000, minlen=2000, steps=3, iters=4 | 0.5  | 0.169  |
Supplementary Table 3: Performance of CONVERGE study under different programs and options with no reference panel

Results are given for the first 10 Mbp of chromosome 20, run in 0.5 Mbp regions with 0.1 Mbp buffers. Program options are as follows. For STITCH, all options were run using 40 EM iterations, split into either diploid (D) or pseudo-haploid (PH) iterations, while RU refers to read unaware (i.e. split each read spanning multiple SNPs into sub-reads spanning one read each). For Beagle, shown are the number of iterations (i.e. burnin-its, phase-its, and impute-its to this value). For findhap, options correspond directly to parameter options. Note that times for STITCH do not include the generation of input data from BAMs, which took about 30 minutes per region, irrespective of other program options. Similarly, times for findhap do not include conversion time from VCF to the findhap input format. Av r² is the average r² for SNPs on the Illumina HumanOmniZhongHua-8 array for common (MAF 5% to 95%) variants, with no filtration for QC for any method. Time is the average in hours for each 0.5Mbp region, where all programs were run on 4 cores on 2.60 GHz Intel E5-2650 chips.

| Program     | Options      | Time  | Av r² |
|-------------|--------------|-------|-------|
| STITCH      | K=20, its=40D | 24.5  | 0.922 |
| STITCH      | K=20, its=40PH| 8.0   | 0.875 |
| STITCH      | K=20, its=34PH;6D | 10.6  | 0.920 |
| STITCH      | K=20, its=35PH;5D | 9.9   | 0.919 |
| STITCH      | K=20, its=36PH;4D | 9.6   | 0.918 |
| STITCH      | K=20, its=37PH;3D | 9.3   | 0.917 |
| STITCH      | K=20, its=38PH;2D | 8.8   | 0.911 |
| STITCH      | K=20, its=39PH;1D | 8.4   | 0.898 |
| STITCH      | K=20, its=38PH;2D, RU | 9.4   | 0.910 |
| STITCH      | K=30, its=40D   | 52.2  | 0.927 |
| STITCH      | K=30, its=38PH;2D | 12.4  | 0.917 |
| STITCH      | K=40, its=38PH;2D | 16.5  | 0.920 |
| STITCH      | K=60, its=38PH;2D | 27.7  | 0.923 |
| STITCH      | K=80, its=38PH;2D | 42.2  | 0.925 |
| STITCH      | K=100, its=38PH;2D | 61.1  | 0.927 |
| Beagle      | its=5         | 12.5  | 0.874 |
| findhap     | maxlen=100000, minlen=1000, steps=3, iters=4 | 0.4   | 0.437 |
| findhap     | maxlen=100000, minlen=1000, steps=2, iters=6 | 0.4   | 0.437 |
| findhap     | maxlen=100000, minlen=1000, steps=5, iters=10 | 1.4   | 0.426 |
| findhap     | maxlen=100000, minlen=100, steps=3, iters=4 | 0.3   | 0.434 |
| findhap     | maxlen=500000, minlen=500, steps=3, iters=4 | 0.4   | 0.448 |
| findhap     | maxlen=200000, minlen=2000, steps=3, iters=4 | 0.5   | 0.414 |
Supplementary Table 4A: Genotype concordance for CONVERGE using STITCH (K=40, 38 PH iterations, 2 D iterations) (without a reference panel) at all SNPs

Results give genotype concordance stratified by genotype class and allele frequency. Discrete genotype calls are generated for imputation as the genotype with the maximum genotype posterior probability. Results are given for the first 10 Mbp region of chromosome 20, run in 20 0.5 Mbp regions with 0.1 Mbp buffers. Allele freqs = allele frequencies are the frequency of the minor allele. Type is either High Cov = high coverage (10X) sequencing (9 samples) or Array = HumanOmniZhongHua-8 (72 samples). Columns contain either Num = Number of non-missing genotypes considered (samples times SNPs for sequencing or array), or Per = Percent of imputed best guess genotypes that match sequencing or array genotypes. Hom major = homozygous for the major allele, Het = heterozygous, Hom Minor = homozygous for the minor allele. Note that truth (sequencing or array) genotypes contain some missing data.

| Allele freqs | Type    | Num  | Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|---------|------|-----------|---------------|---------|---------|---------------|---------------|
| [0,0.01)     | High Cov| 16234| 99.98     | 879           | 27.19   | 14      | 0             |               |
| [0.01,0.02)  | High Cov| 3973 | 99.72     | 483           | 62.11   | 6       | 16.67         |               |
| [0.02,0.05)  | High Cov| 11325| 99.59     | 1725          | 83.65   | 31      | 19.35         |               |
| [0.05,0.1)   | High Cov| 14634| 99.33     | 2931          | 91.23   | 116     | 65.52         |               |
| [0.1,0.2]    | High Cov| 28004| 99.13     | 10059         | 96.01   | 1183    | 84.53         |               |
| [0.2,0.3]    | High Cov| 18880| 98.59     | 12285         | 96.81   | 2134    | 90.63         |               |
| [0.3,0.4]    | High Cov| 15750| 97.69     | 15356         | 97.49   | 4300    | 94.12         |               |
| [0.4,0.5]    | High Cov| 15750| 97.69     | 15356         | 97.49   | 4300    | 94.12         |               |
| [0.01,0.02)  | Array   | 9691 | 99.95     | 100           | 70      | 0       | NA            |               |
| [0.02,0.05]  | Array   | 4519 | 99.91     | 156           | 73.08   | 0       | NA            |               |
| [0.05,0.1]   | Array   | 11883| 99.82     | 834           | 88.13   | 12      | 83.33         |               |
| [0.1,0.2]    | Array   | 18317| 99.42     | 3003          | 91.44   | 114     | 73.68         |               |
| [0.2,0.3]    | Array   | 29193| 98.92     | 10165         | 93.38   | 866     | 85.68         |               |
| [0.3,0.4]    | Array   | 20139| 97.89     | 12835         | 94.39   | 2201    | 89.41         |               |
| [0.4,0.5]    | Array   | 14833| 97.2      | 15879         | 95.44   | 4171    | 93.02         |               |
| [0.01,0.02)  | Array   | 10172| 96.29     | 16290         | 95.67   | 6766    | 94.8          |               |
Supplementary Table 4B: Genotype concordance for CONVERGE using Beagle (default) (without a reference panel) at all SNPs

| Allele freqs | Type    | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|---------|---------------|---------------|---------|---------|---------------|---------------|
| [0,0.01)     | High Cov| 16,234        | 100           | 879     | 56.09   | 14            | 0             |
| [0.01,0.02)  | High Cov| 3,973         | 100           | 483     | 64.6    | 6             | 0             |
| [0.02,0.05)  | High Cov| 11,325        | 99.97         | 1,725   | 73.74   | 31            | 6.45          |
| [0.05,0.1)   | High Cov| 14,634        | 99.64         | 2,931   | 84.61   | 116           | 50.86         |
| [0.1,0.2)    | High Cov| 28,004        | 99.49         | 10,059  | 90.55   | 1,183         | 82.25         |
| [0.2,0.3)    | High Cov| 18,880        | 98.98         | 12,285  | 93.09   | 2,134         | 88.71         |
| [0.3,0.4]    | High Cov| 15,750        | 97.96         | 15,356  | 94.27   | 4,300         | 93.37         |
| [0.4,0.5]    | High Cov| 10,469        | 97.05         | 16,280  | 95.12   | 7,445         | 96.55         |
| [0.0,0.01)   | Array   | 9,691         | 100           | 100     | 54      | 0             | NA            |
| [0.01,0.02)  | Array   | 4,519         | 100           | 156     | 57.69   | 0             | NA            |
| [0.02,0.05)  | Array   | 11,883        | 99.91         | 834     | 76.5    | 12            | 75            |
| [0.05,0.1]   | Array   | 18,317        | 99.77         | 3,003   | 81.22   | 114           | 69.3          |
| [0.1,0.2]    | Array   | 29,193        | 99.47         | 10,165  | 86.54   | 866           | 83.03         |
| [0.2,0.3]    | Array   | 20,139        | 98.46         | 12,835  | 89.44   | 2,201         | 86.37         |
| [0.3,0.4]    | Array   | 14,833        | 97.5          | 15,879  | 91.91   | 4,171         | 91.63         |
| [0.4,0.5]    | Array   | 10,172        | 96.23         | 16,290  | 92.98   | 6,766         | 94.66         |
Supplementary Table 4C: Genotype concordance for CONVERGE using findhap (maxlen=50000, minlen=500, steps=3, iters=4) (without a reference panel) at all SNPs

| Allele freqs | Type   | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|--------|---------------|---------------|---------|---------|---------------|---------------|
| [0,0.01]     | High Cov | 13,485        | 99.69         | 562     | 48.22   | 10            | 0             |
| [0.01,0.02)  | High Cov | 3,701         | 99.08         | 453     | 60.71   | 5             | 20            |
| [0.02,0.05)  | High Cov | 10,771        | 97.96         | 1,623   | 61.06   | 27            | 3.7           |
| [0.05,0.1]   | High Cov | 14,328        | 94.45         | 2,875   | 68.49   | 115           | 18.26         |
| [0.1,0.2)    | High Cov | 26,632        | 93.43         | 9,443   | 69      | 1,083         | 42.84         |
| [0.2,0.3)    | High Cov | 17,884        | 87.82         | 11,566  | 66.72   | 1,964         | 50.61         |
| [0.3,0.4)    | High Cov | 15,229        | 84.5          | 14,665  | 67.23   | 4,212         | 60.73         |
| [0.4,0.5]    | High Cov | 9,766         | 77.63         | 15,546  | 67.34   | 7,050         | 67.48         |
| [0,0.01]     | Array   | 8,690         | 99.57         | 94      | 46.81   | 0             | NA            |
| [0.01,0.02)  | Array   | 3,894         | 98.02         | 133     | 46.62   | 0             | NA            |
| [0.02,0.05)  | Array   | 11,302        | 97.34         | 773     | 56.4    | 8             | 37.5          |
| [0.05,0.1]   | Array   | 17,884        | 93.63         | 2,934   | 63.53   | 113           | 20.35         |
| [0.1,0.2)    | Array   | 27,294        | 91.11         | 9,529   | 62.09   | 813           | 31.12         |
| [0.2,0.3)    | Array   | 18,845        | 85.16         | 12,052  | 63.57   | 2,046         | 40.27         |
| [0.3,0.4]    | Array   | 14,260        | 78.72         | 15,331  | 65.34   | 4,068         | 49.68         |
| [0.4,0.5]    | Array   | 9,903         | 72.2          | 15,787  | 66.08   | 6,003         | 58.47         |
Supplementary Table 4D: Genotype concordance for CONVERGE using STITCH (K=40, 38 PH iterations, 2 D iterations) (without a reference panel) at all SNPs (that pass QC)

| Allele freqs | Type       | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|------------|---------------|---------------|---------|---------|---------------|---------------|
| [0.0,0.01)   | High Cov   | 6,266         | 99.97         | 241     | 83.82   | 1             | 0             |
| [0.01,0.02)  | High Cov   | 2,725         | 99.67         | 323     | 85.45   | 2             | 50            |
| [0.02,0.05)  | High Cov   | 10,001        | 99.59         | 1,541   | 90.53   | 29            | 17.24         |
| [0.05,0.1)   | High Cov   | 13,916        | 99.35         | 2,760   | 94.28   | 109           | 68.81         |
| [0.1,0.2)    | High Cov   | 27,327        | 99.19         | 9,773   | 97.24   | 1,155         | 85.97         |
| [0.2,0.3)    | High Cov   | 18,291        | 98.9          | 11,938  | 97.76   | 2,064         | 92.34         |
| [0.3,0.4)    | High Cov   | 15,294        | 98.12         | 14,869  | 98.14   | 4,170         | 95.4          |
| [0.4,0.5]    | High Cov   | 10,174        | 97.67         | 15,802  | 98.15   | 7,260         | 96.85         |
| [0.0,0.01)   | Array      | 6,763         | 99.94         | 76      | 90.79   | 0             | NA            |
| [0.01,0.02)  | Array      | 3,898         | 99.9          | 132     | 84.09   | 0             | NA            |
| [0.02,0.05)  | Array      | 11,068        | 99.83         | 787     | 92.12   | 11            | 90.91         |
| [0.05,0.1)   | Array      | 17,968        | 99.49         | 2,925   | 92.89   | 110           | 76.36         |
| [0.1,0.2)    | Array      | 27,902        | 99.06         | 9,643   | 95.49   | 809           | 90.36         |
| [0.2,0.3)    | Array      | 18,709        | 98.36         | 11,832  | 96.49   | 2,047         | 92.82         |
| [0.3,0.4)    | Array      | 14,263        | 97.69         | 15,252  | 96.48   | 4,000         | 95.15         |
| [0.4,0.5]    | Array      | 9,573         | 97.49         | 15,336  | 96.77   | 6,376         | 96.86         |
Supplementary Table 5: Performance of CONVERGE study under different programs and options with a reference panel

Results are given for the first 10 Mbp of chromosome 20, run in 0.5 Mbp regions with 0.1 Mbp buffers. Program options are as follows. For STITCH, all options were run using 40 EM iterations, split into either diploid (D) or pseudo-haploid (PH) iterations. For Beagle, shown are the number of iterations (i.e. burnin-its, phase-its, and impute-its to this value). Note that times for STITCH do not include the generation of input data from BAMs, which took about 30 minutes per region, irrespective of other program options. Av r2 is the average r^2 for SNPs on the Illumina HumanOmniZhongHua-8 array for common (MAF 5% to 95%) variants, with no filtration for QC for any method. Time is the average in hours for each 0.5Mbp region, where all programs were run on 4 cores on 2.60 GHz Intel E5-2650 chips.

| Program | Options                          | Time | Av r2 |
|---------|----------------------------------|------|-------|
| STITCH  | K=20, its=38PH:2D                | 5.4  | 0.911 |
| STITCH  | K=40, its=38PH:2D                | 10.2 | 0.922 |
| STITCH  | K=60, its=38PH:2D                | 16.6 | 0.925 |
| Beagle  | its=5, no ref panel              | 7.8  | 0.886 |
| Beagle  | its=4                            | 114.4| 0.946 |
| Beagle  | its=3                            | 74.5 | 0.943 |
| Beagle  | its=2                            | 39.7 | 0.939 |
| Beagle  | its=1                            | 12.0 | 0.930 |
Supplementary Table 6A: Genotype concordance for CONVERGE using STITCH (K=40, 38 PH iterations, 2D iterations) (without a reference panel) at reference panel SNPs (1000G ASN). Results give genotype concordance stratified by genotype class and allele frequency. Discrete genotype calls are generated for imputation as the the genotype with the maximum genotype posterior probability. Results are given for the first 10 Mbp region of chromosome 20, run in 20 0.5 Mbp regions with 0.1 Mbp buffers. Allele freqs = allele frequencies are the frequency of the minor allele. Type is either High Cov = high coverage (10X) sequencing (9 samples) or Array = HumanOmniZhongHua-8 (72 samples). Columns contain either Num = Number of non-missing genotypes considered (samples times SNPs for sequencing or array), or Per = Percent of imputed best guess genotypes that match sequencing or array genotypes. Hom major = homozygous for the major allele, Het = heterozygous, Hom Minor = homozygous for the minor allele. Note that truth (sequencing or array) genotypes contain some missing data.

| Allele freqs | Type    | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|---------|---------------|---------------|---------|---------|---------------|---------------|
| [0.0,0.01)   | High Cov| 13,968        | 99.96         | 596     | 35.91   | 10            | 0             |
| [0.01,0.02)  | High Cov| 3,798         | 99.74         | 460     | 63.04   | 6             | 0             |
| [0.02,0.05)  | High Cov| 11,259        | 99.72         | 1,714   | 85.3    | 31            | 16.13         |
| [0.05,0.1)   | High Cov| 14,634        | 99.33         | 2,931   | 91.88   | 116           | 63.79         |
| [0.1,0.2)    | High Cov| 27,980        | 99.1          | 10,054  | 95.91   | 1,183         | 85.88         |
| [0.2,0.3)    | High Cov| 18,875        | 98.56         | 12,282  | 96.88   | 2,133         | 90.53         |
| [0.3,0.4)    | High Cov| 15,737        | 97.59         | 15,315  | 97.43   | 4,298         | 94.04         |
| [0.4,0.5]    | High Cov| 10,463        | 96.82         | 16,261  | 97.72   | 7,434         | 95.96         |
| [0.0,0.01)   | Array   | 8,975         | 99.96         | 97      | 64.95   | 0             | NA            |
| [0.01,0.02)  | Array   | 4,519         | 99.82         | 156     | 80.13   | 0             | NA            |
| [0.02,0.05)  | Array   | 11,740        | 99.74         | 833     | 88.36   | 12            | 83.33         |
| [0.05,0.1)   | Array   | 18,317        | 99.45         | 3,003   | 91.81   | 114           | 76.32         |
| [0.1,0.2)    | Array   | 29,144        | 98.93         | 10,142  | 93.7    | 866           | 86.95         |
| [0.2,0.3)    | Array   | 20,139        | 97.85         | 12,835  | 94.66   | 2,201         | 90.37         |
| [0.3,0.4)    | Array   | 14,833        | 97.01         | 15,879  | 95.54   | 4,171         | 92.78         |
| [0.4,0.5]    | Array   | 10,172        | 96.06         | 16,290  | 95.75   | 6,766         | 94.77         |
### Supplementary Table 6B: Genotype concordance for CONVERGE using Beagle (default) (without a reference panel) at reference panel SNPs (1000G ASN)

| Allele freqs | Type      | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|-----------|---------------|---------------|---------|---------|---------------|---------------|
| [0.01,0.02)  | High Cov  | 13,968        | 100           | 596     | 55.54   | 10            | 0             |
| [0.02,0.05)  | High Cov  | 3,798         | 100           | 460     | 66.09   | 6             | 0             |
| [0.05,0.1)   | High Cov  | 11,259        | 99.95         | 1,714   | 75.61   | 31            | 12.9          |
| [0.1,0.2)    | High Cov  | 27,980        | 99.41         | 10,054  | 91.87   | 1,183         | 83.94         |
| [0.2,0.3)    | High Cov  | 18,875        | 98.95         | 12,282  | 93.93   | 2,133         | 89.45         |
| [0.3,0.4)    | High Cov  | 15,737        | 97.99         | 15,315  | 95.09   | 4,298         | 93.9          |
| [0.4,0.5]    | High Cov  | 10,463        | 97.24         | 16,261  | 95.79   | 7,434         | 96.7          |
| [0.01,0.02]  | Array     | 8,975         | 100           | 97      | 56.7    | 0             | NA            |
| [0.02,0.05]  | Array     | 4,519         | 100           | 156     | 58.33   | 0             | NA            |
| [0.05,0.1]   | Array     | 11,740        | 99.88         | 833     | 78.63   | 12            | 75            |
| [0.1,0.2]    | Array     | 18,317        | 99.72         | 3,003   | 83.25   | 114           | 71.05         |
| [0.2,0.3]    | Array     | 29,144        | 99.32         | 10,142  | 88.55   | 866           | 84.06         |
| [0.3,0.4]    | Array     | 20,139        | 98.38         | 12,835  | 90.7    | 2,201         | 87.6          |
| [0.4,0.5]    | Array     | 14,833        | 97.57         | 15,879  | 92.88   | 4,171         | 92.14         |
| [0.01,0.02]  | Array     | 10,172        | 96.26         | 16,290  | 93.49   | 6,766         | 95.26         |
Supplementary Table 6C: Genotype concordance for CONVERGE using Beagle (its=3) (with a reference panel) at reference panel SNPs (1000G ASN)

| Allele freqs | Type  | Num Hom Major | Per Hom Major | Num Het | Per Het | Num Hom Minor | Per Hom Minor |
|--------------|-------|---------------|---------------|---------|---------|--------------|--------------|
| [0.0,0.01)   | High Cov | 13,968        | 99.96         | 596     | 66.95   | 10           | 0            |
| [0.01,0.02]  | High Cov | 3,798         | 99.63         | 460     | 81.09   | 6            | 33.33        |
| [0.02,0.05)  | High Cov | 11,259        | 99.86         | 1,714   | 91.54   | 31           | 19.35        |
| [0.05,0.1]   | High Cov | 14,634        | 99.55         | 2,931   | 95.19   | 116          | 67.24        |
| [0.1,0.2)    | High Cov | 27,980        | 99.32         | 10,054  | 97.13   | 1,183        | 88.33        |
| [0.2,0.3)    | High Cov | 18,875        | 98.95         | 12,282  | 97.44   | 2,133        | 92.45        |
| [0.3,0.4]    | High Cov | 15,737        | 98.2          | 15,315  | 97.54   | 4,298        | 95.23        |
| [0.4,0.5]    | High Cov | 10,463        | 97.69         | 16,261  | 97.82   | 7,434        | 97.19        |
| [0.0,0.01)   | Array   | 8,975         | 99.98         | 97      | 81.44   | 0            | NA           |
| [0.01,0.02]  | Array   | 4,519         | 99.96         | 156     | 83.33   | 0            | NA           |
| [0.02,0.05]  | Array   | 11,740        | 99.8          | 833     | 93.28   | 12           | 83.33        |
| [0.05,0.1]   | Array   | 18,317        | 99.6          | 3,003   | 94.21   | 114          | 84.21        |
| [0.1,0.2]    | Array   | 29,144        | 99.22         | 10,142  | 95.5    | 866          | 91.11        |
| [0.2,0.3]    | Array   | 20,139        | 98.72         | 12,835  | 95.96   | 2,201        | 94           |
| [0.3,0.4]    | Array   | 14,833        | 97.9          | 15,879  | 96.52   | 4,171        | 95.35        |
| [0.4,0.5]    | Array   | 10,172        | 97.53         | 16,290  | 96.62   | 6,766        | 97.04        |
**Supplementary Table 7: Performance of STITCH on CONVERGE study original imputation**

Results are over the first 10 Mbp of chromosome 20. Beagle methodology was the same as done in the original CONVERGE paper and as explained in the text. STITCH results are for $K = 40, 38$ pseudo-haploid iterations, 2 diploid iterations. All sites with removal of SNPs failing QC also removed SNPs with Hardy-Weinberg p-value less than $10^{-6}$. Av $r^2$ is the average $r^2$ for SNPs on the Illumina HumanOmniZhongHua-8 array for high frequency (MAF 5% to 95%) variants.

| Method | SNP set | % SNPs | Av $r^2$ |
|--------|---------|--------|----------|
| Beagle | All     | 100    | 0.933    |
| STITCH | All     | 100    | 0.92     |
| Beagle | info>0.4| 90     | 0.939    |
| STITCH | info>0.4| 90     | 0.939    |
| Beagle | info>0.9| 78     | 0.968    |
| STITCH | info>0.9| 75     | 0.972    |
Supplementary Table 8: Effect of filtering on imputation performance

Results are given for chromosome 19. QC is defined per-run and reflects info > 0.4 and HWE p-value > $1 \times 10^{-6}$. $r^2$ values are against the 4 10X mice.

| Set | Description                  | SNPs   | Number of SNPs | T1/Tv | VQSR r² | No VQSR r² |
|-----|-------------------------------|--------|----------------|-------|---------|------------|
| 1   | VQSR                          | All    | 152,486        | 2.07  | 0.937   |            |
| 2   | VQSR                          | Post-QC| 122,878        | 2.21  | 0.968   |            |
| 3   | No VQSR, Round 1              | All    | 355,123        | 1.48  | 0.745   | 0.945      |
| 4   | No VQSR, Round 1              | Post-QC| 136,164        | 2.08  | 0.938   |            |
| 5   | No VQSR, Round 2              | All    | 136,164        | 2.08  | 0.938   |            |
| 6   | No VQSR, Round 2              | Post-QC| 128,054        | 2.14  | 0.952   |            |
| 7   | Intersect Set 2 and Set 6     |        | 115,567        | 2.22  | 0.967   | 0.969      |
| 8   | Present Set 2, absent Set 6    |        | 7,311          | 2.13  | 0.915   |            |
| 9   | Present Set 6, absent Set 2    |        | 12,487         | 1.55  | 0.930   |            |
References

[1] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1977;39(1):1–38. Available from: http://www.jstor.org/stable/2984875

[2] Scheet P, Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics. 2006 Apr;78(4):629–644. Available from: http://www.sciencedirect.com/science/article/pii/S000292970763701X