Supplementary Information for "Rapid genotype imputation from sequence with reference panels"

Davies *et al.*

April 12, 2021

Contents

1 Supplementary Note

1.1 QUILT

1.1.1 Notation

We note that when possible notation follows that of the STITCH model [1] and the SEW model [2], though some changes have been made. In particular, for the random variables under the model, of O, R, H and Q, to use upper case for random variables and lower case for observations of those variables. Further, we try to use lower case for indices, upper case for constants, and Greek characters for parameters of the model.

Summary of notation related to counts and constants

Symbol	Definition
N_s	Number of SNPs
N_g	Number of grids, with $N_g = \lceil \frac{N_s}{32} \rceil$
N_r	Number of sequencing reads
N_{ek}	Defined as $\frac{4N_e}{K}$, where N_e is the effective population size
K	Number of haplotypes in haplotype reference panel
L	Physical position of SNPs (*i.e.* L_t is the physical position for SNP t)
G	Grid for SNPs, with G_t being the grid location for SNP t, with $G_t = \lceil \frac{t}{32} \rceil$
P	Recombination rate between grids, with P_g being between grids g and $g+1$
D	Physical distance between grids, with D_g being between grids g and $g+1$
I	Vector of reference haplotypes to consider in Gibbs sampling

Summary of notation related to indices

Symbol	Definition
t	Choice of SNP, $t \in \{1, ..., N_s\}$
g	Choice of grid, $g \in \{1, ..., N_g\}$
i	Choice of maternal ($i = 1$) or paternal ($i = 2$) haplotype
k	Choice of haplotype in haplotype reference panel, $k \in \{1, ..., K\}$
v	Choice of sequencing read, $v \in \{1, ..., N_r\}$
j	Choice of SNP in read, with $j \in \{1, ..., a_v\}$, where a_v is the number of SNPs intersected by read r_v

Summary of notation related to reads

Symbol	Definition
R and r	Random and observed variables denoting reads, consisting of u, s, b, and derived a and c
R_v and r_v	Read with index v which spans a_v SNPs, has central grid c_v, with SNP indices u_v, sequenced bases s_v and base qualities b_v, or $r_v = \{u_v, s_v, b_v\}$
a_v	Number of SNPs spanned by read r_v, with $a_v = \lvert u_v \rvert$
c_v	Central grid for read r_v, with $c_v \in \{1, ..., N_g\}$, and c_v being defined deterministically by u_v
$u_{v,j}$	For SNP j in read r_v, the indices of SNPs intersected by that read, *i.e.* $u_{v,j} \in \{1, ..., N_s\}$, and for example it has physical position $L_{u_{v,j}}$
$s_{v,j}$	Sequencing base for SNP j in read r_v, with $s_{v,j} = 1$ for the alternate base and 0 for the reference base
$b_{v,j}$	Base quality for SNP j in read r_v
$e^l_{v,j}$	Probability of SNP j from read r_v coming from an underlying genotype l, or $P(S_{v,j} = s_{v,j} \lvert Gen = l)$
O and o	Random and observed variables denoting all observed data from the sequencing reads $o = \{o_g \lvert g = 1, ..., N_g\}$, or alternatively, $o = \{r_v \lvert v = 1, ..., N_r\}$
o_g	Observations comprised of reads with central grid g, $o_g = \{r_v \lvert c_v = g\}$
o^i	Observations on either maternal ($i = 1$) or paternal ($i = 2$) chromosome, $o^i = \{r_v \lvert h_v = i\}$
o^i_g	Observations from a particular chromosome at grid g, $o^i_g = \{r_v \lvert c_v = g, h_v = i\}$

Summary of notation related to hidden variables

Symbol	Definition
H and h	Random and observed variables, vector of haplotype membership labels, with $h_v = 1$ denoting read v is from the maternal haplotype and $h_v = 2$ for read v coming from the paternal haplotype
H_{-v} and h_{-v}	Vector H or h with read v removed
Q and q	Random and observed variables, vector of hidden states, with q_g denoting the hidden copying state from the haplotype reference panel in grid g, with $q_g \in \{1, ..., K\}$

Summary of notation related to parameters of the model

Symbol	Definition
σ_g	Recombination distance between grids g and $g + 1$
$\theta_{t,k}$	Probability reference haplotype k emits the alternate base at SNP t, defined as $\theta_{t,k} = w$ when reference haplotype k at SNP t has the reference and $1 - w$ when it has the alternate, for a fixed error rate w, default $w = 0.001$
λ	Parameters of the model $\lambda = \{\sigma, \theta\}$

Summary of notation related to hidden Markov model probabilities

Symbol	Definition
$\alpha_g^i(k)$	Traditional HMM forward probabilities, here specifically for observations for haplotype i up to grid g for hidden state k, where $\alpha_g^i(k) = P(O_1^i = o_g^i, ..., O_g^i = o_g^i, Q_g^i = k\|\lambda)$
$\beta_g^i(k)$	Traditional HMM backward probabilities, $P(O_{g+1}^i = o_{g+1}^i, ..., O_{N_g}^i = o_{N_g}^i \| Q_g^i = k, \lambda)$
$\gamma_g^i(k)$	Traditional HMM posterior state probabilities, $P(Q_g^i = k\|O^i, \lambda) = \frac{\alpha_g^i(k)\beta_g^i(k)}{P(O^i\|\lambda)}$

1.1.2 Model setup, and simulating under the model

We consider a Li and Stephens model[3] in which a sample haplotype can be modelled as a mosaic of reference haplotypes. We assume that we are considering the region to be imputed has a high enough marker density that it is sufficient to consider discretized regions of length 32 SNPs, where we allow recombinations to occur between but not within such regions. We use the term "grid" to refer to these fixed windows. Let indexes of t refer to SNPs and g to grid, and let N_s and N_g be constants representing the number of SNPs and grids, respectively. Let L_t be the physical position of SNP t, and G_t be the grid of SNP t, where G_t is the ceiling of t divided by 32. Note here the difference between the special use of N and other constants like L, where notation N_x refers to a count where the subscript is convenient to remember the constant, rather than an index, i.e. N_s is the number of SNPs, rather than the s^{th} entry of vector N, but L_t and other terms are instead indexed by t. Let D_g and P_g be the physical distance and recombination rate between grid points g and $g+1$, respectively, where D_g is taken as the difference between the average of L for SNPs with grid points g and $g+1$ respectively, and similarly for P for recombination rate. Let K be the number of haplotypes in the haplotype reference panel. Then the recombination distance between grids g and $g+1$ is $\sigma_g = D_g P_g$, and the probability of recombination between grid points g and $g+1$ is $1-e^{-N_{ek}\sigma_g}$, where $N_{ek} = \frac{4N_e}{K}$, although in practice we allow N_{ek} to be user defined and find that higher values of N_{ek} can generate more accurate imputation, potentially due to non-random relatedness among haplotype reference panel members. Conditional on recombination locations, we sample a reference haplotype that will be copied continuously between recombination break points, and we sample this haplotype uniformly at random. We define q_g as the observed reference haplotype copied in grid g, which will be constant between recombination break points.

Finally, we simulate sequencing reads. Let N_r be a constant referring to the number of reads. Let r_v refer to the v^{th} observed sequencing read. We first sample a central grid c_v the read is sampled from, and start and end positions yielding a set of SNPs with indexes $u_{v,j}$ that the read overlaps, a set of observed bases $s_{v,j}$ at these SNPs, and a set of Phred base qualities $b_{v,j}$ of these observed

bases. There are a_v of these SNPs in total defining the number of j values. Then $r_v = \{a_v, u_v, s_v, b_v\}$ captures the information from this read. We assume that for all sequencing reads, the underlying sequenced molecule differs from the reference haplotype according to some fixed mutation/error rate. When simulating data, this then allows us to independently sample $b_{v,j}$, for each j, and using both the sequencing and reference haplotype error rate, sample $s_{v,j}$.

We make two different assumptions that govern the underlying reference haplotype sequence over the SNPs $u_{v,j}$, and use these differently in two parts of the model. The first, more realistic assumption, is that we expect the true underlying sequence of the sequenced read at SNP position $u_{v,j} = t$ and with copied reference haplotype $q_{G_t} = k$ to be that of reference haplotype k and SNP t, i.e. the true underlying molecule will reflect the copied reference haplotype haplotype at that position. This assumption applies when we perform imputation using the per-haplotype full reference panel imputation. The second assumption, is that, instead, we expect the true underlying molecule will reflect the copied haplotype at only the central grid of the read. Thus the sequenced read at SNP position $u_{v,j} = t$ and with copied reference haplotype at central read grid position c_v with $q_{c_v} = k$ will reflect the sequence of reference haplotype k and SNP t. This approximation is used for computational convenience: it allows, in a diploid model, us to apply a hidden Markov model because it implies the probability of observations at a given grid point reflect will only reflect the hidden copying state at that grid point, and not on the hidden state at any other grid point. We use this second assumption in the Gibbs sampler and what follows next.

1.1.3 Probabilities under the model

Let λ be the parameters of the model $\lambda = (\theta, \sigma)$, where σ is recombination distance, and θ depends on the haplotype reference panel. We use a hidden Markov model where Q_g is the random variable of hidden state and q_g is its realisation for a sample haplotype at grid point g, that is q_g is in $\{1, ..., K\}$, and both Q and q have length N_g. Initial hidden states are taken to be equally probable with $P(Q_1 = k) = \frac{1}{K}$ for all k. For state transitions, with probability $e^{-N_{ek}\sigma_g}$, no recombination occurs between grids g and $g+1$, and with probability $1 - e^{-N_{ek}\sigma_g}$ a recombination occurs and a new state is chosen randomly from the K options, giving

$$P(Q_{g+1} = q_{g+1} | Q_g = q_g, \lambda) = \begin{cases} e^{-N_{ek}\sigma_g} + (1 - e^{-N_{ek}\sigma_g})/K & \text{if } q_{g+1} = q_g \\ (1 - e^{-N_{ek}\sigma_g})/K & \text{if } q_{g+1} \neq q_g \end{cases} \quad (1)$$

Let Q and q refer to results for an arbitrary chromosome, and Q_i and q_i refer to haplotype i, where $i = 1$ arbitrarily refers to maternal and $i = 2$ to paternal origin.

For the emission of reads, consider the SNP j in the read indexed by v. Given the Phred scaled base quality of $b_{v,j}$, we have that the probability that this base is called erroneously is $\epsilon_{v,j} = 10^{-b_{v,j}/10}$, and hence, compared to the

true underlying genotype $Gen = l$, for observed sequence $s_{v,j}$, that

$$P(S_{v,j} = s_{v,j}|Gen = l) = \begin{cases} 1 - \epsilon_{v,j} & \text{if } s_{v,j} = l \\ \frac{1}{3}\epsilon_{v,j} & \text{if } s_{v,j} \neq l \end{cases} \tag{2}$$

For convenience, set $e^l_{v,j} = P(S_{v,j} = s_{r,j}|Gen = l)$. For the reference haplotypes, for some fixed error w (default 0.001), define $\theta_{t,k} = w$ if reference haplotype k has a reference base at SNP t and $1 - w$ if reference haplotype k has the alternate base at SNP t. We therefore have that the probability of read v, given it was emitted in central grid c_v on haplotype i with hidden copying state $q^i_{c_v}$, is

$$P(R_v = r_v|Q^i_{c_v} = q^i_{c_v}, \lambda) = \prod_{j=1}^{a_v} \left(\theta_{u_{v,j},q^i_{c_v}} e^1_{v,j} + (1 - \theta_{u_{v,j},q^i_{c_v}})e^0_{v,j}\right) \tag{3}$$

Let H be a random variable which is a vector of read labels of length N_r, with observation h with h_v in $\{1, 2\}$, where 1 arbitrarily refers to maternal and 2 to paternal origin. Let O be a random variable representing the entire set of sequencing reads, with observation o. We can consider o by indexing by reads $o = \{r_v|v = 1, ..., N_r\}$, and specific observations at g as the reads that intersect that grid $o_g = \{r_v|c_v = g\}$. Let o^i be the subset of reads drawn from parental haplotype i, with i in $\{1, 2\}$, with $o^i = \{r_v|h_v = i\}$. We can therefore calculate for sequencing reads o_i, hidden reference panel states q_i, and parameters λ that

$$P(O^i = o^i, Q^i = q^i|\lambda) = \frac{1}{K} \prod_{g=2}^{N_g} P(Q^i_g = q^i_g|Q^i_{g-1} = q^i_{g-1}, \lambda) \prod_{r_v \in o^i} P(R_v = r_v|Q^i_{c_v} = q^i_{c_v}, \lambda) \tag{4}$$

Finally we can generate the complete data probability for all reads given the observed data o, a hidden vectors of copying states q^1 and q^2, read labels h, and parameters of the model λ as

$$P(O, Q^1, Q^2|\lambda) = P(O^1, Q^1|\lambda) \times P(O^2, Q^2|\lambda) \tag{5}$$

1.1.4 Gibbs sampling setup

Consider Gen as the diploid genotype or some arbitrary SNP that we are interested in estimating, and that we want to posterior genotype probabilities $P(Gen|O, \lambda)$, which can be used to give us the diploid genotype dosage $E[Gen|O, \lambda]$ in the normal way as

$$E[Gen|O, \lambda] = \sum_{g=0}^{2} g \times P(Gen = g|O, \lambda) \tag{6}$$

We can calculate $P(Gen = g|O, \lambda)$ using Monte Carlo as follows (with $\approx$ becoming $=$ in the limit)

$$P(Gen = g|O, \lambda) = \sum_{H \in \{1,2\}^{N_r}} P(Gen = g|H, O, \lambda) P(H|O, \lambda) \tag{7}$$

$$\approx \sum_{H \sim P(H|O,\lambda)} P(Gen = g|H, O, \lambda) \tag{8}$$

We can rapidly calculate $P(Gen = g|H, O, \lambda)$ under the model as

$$P(Gen = g|H, O, \lambda) = \begin{cases} P(Hap^1 = 0|H, O, \lambda) \times P(Hap^2 = 0|H, O, \lambda) & \text{if } g = 0 \\ P(Hap^1 = 0|H, O, \lambda) \times P(Hap^2 = 1|H, O, \lambda) + & \\ P(Hap^1 = 1|H, O, \lambda) \times P(Hap^2 = 0|H, O, \lambda) & \text{if } g = 1 \\ P(Hap^1 = 1|H, O, \lambda) \times P(Hap^2 = 1|H, O, \lambda) & \text{if } g = 2 \end{cases}$$

and further we can rapidly calculate $P(Hap^i = 1|H, O, \lambda)$ for some SNP t as

$$P(Hap^i = 1|H, O, \lambda) = \sum_{k=1}^{K} \theta_{t,k} P(Q^i_{G_t} = k|O, \lambda) \tag{9}$$

recalling that $\theta_{t,k}$ gives the probability that reference haplotype k carries the alternate allele, and $P(Q^i_{G_t} = k|O, \lambda)$ is the probability the the sample copies from haplotype k at grid point G_t.

We therefore need rapid draws of $H \sim P(H|O, \lambda)$, which can be done using Gibbs sampling in way described in the following subsection.

1.1.5 Gibbs sampling practical

We first suppose that we are working not with the full set of haplotypes but with some subset, call it I (for example, I could be $I = \{3, 7, ...\}$. Suppose we therefore consider a recast version of θ under this new I, such that $\theta^*_{t,k^*} = \theta_{t,I_{k^*}}$, and that $K^* = |I|$. In this section, it is arbitrary to use either θ^* or K^* versus θ and K, as the math does does not change. Therefore for simplicity of notation in what is already crowded mathematics, we omit the $*$, but note that in practice, we used the reduced set that depends on I.

Now, for drawing a new h from H, suppose we start with some initial realization for H, call it h, at random, *i.e.* $P(H_v) = \frac{1}{2} \forall v \in 1, ..., N_r$. Let h_v be the current label for read v and h^o_v be the alternate read label *i.e.* $h^o_v = 3 - h_v$. Now, with Gibbs sampling, we want to efficiently calculate

$$P(H_v = h^o_v|H_{-v}, O, \lambda) = \frac{P(O, H_v = h^o_v, H_{-v} = h_{-v}|\lambda)}{\sum_{j=1}^{2} P(O, H_v = j, H_{-v} = h_{-v}|\lambda)} \tag{10}$$

This can be done if we can efficiently calculate $P(O, H_v = h^o_v, H_{-v} = h_{-v}|\lambda)$, noting that we already have $P(O, H_v = h_v, H_{-v} = h_{-v}|\lambda)$ under the HMM. Now suppose we have, for a particular realization of h, run an HMM including

forward backward pass, separately for each of $i = 1$ and $i = 2$, generating forward-backward variables in the normal way, with

$$\alpha_g^i(k) = P(O_1^i = o_1^i, ..., O_g^i = o_g^i, Q_g^i = k|\lambda) \tag{11}$$

$$\beta_g^i(k) = P(O_{g+1}^i = o_{g+1}^i, ..., O_{N_g}^i = o_{N_g}^i|Q_g^i = k, \lambda) \tag{12}$$

$$\gamma_g^i(k) = P(Q_g^i = k|O^i = o^i, \lambda) = \frac{\alpha_g^i(k)\beta_g^i(k)}{P(O^i = o^i|\lambda)} \tag{13}$$

Then we have already calculated

$$P(O, H_v = h_v, H_{-v} = h_{-v}|\lambda) = P(O^{h_v}|\lambda)P(O^{h_v^o}|\lambda)\left(\frac{1}{2}\right)^{N_r} \tag{14}$$

and we want to efficiently calculate

$$P(O, H_v = h_v^o, H_{-v} = h_{-v}|\lambda) = P(O^{h_v,-v}|\lambda)P(O^{h_v^o,+v}|\lambda)\left(\frac{1}{2}\right)^{N_r} \tag{15}$$

where $O^{h_v,-v}$ means we take read v out of the observations for haplotype h_v *i.e.* $O^{h_v,-v} = \{r_{v*}|h_{v*} = h_v, v* \neq v\}$, and similarly $O^{h_v^o,+v}$ means we add observation from read v to haplotype h_v^o.

Now, recalling the definition of the standard HMM variables given above. For α, we have that

$$\alpha_g^i(k) = \left[\sum_{j=1}^{K} \alpha_{g-1}^i(k)P(Q_g^i = k|Q_{g-1}^i = j, \lambda)\right] P(O_g^i = o_g^i|Q_g = k, \lambda) \tag{16}$$

Now, considering $\alpha_g^{h_v,-v}(k)$ and $\alpha_g^{h_v^o,+v}(k)$ as

$$\alpha_g^{h_v,-v}(k) = \frac{\alpha_g^{h_v}(k)}{P(R_v = r_v|Q_g = k, \lambda)} \tag{17}$$

$$\alpha_g^{h_v^o,+v}(k) = \alpha_g^{h_v^o}(k)P(R_v = r_v|Q_g = k, \lambda) \tag{18}$$

Recall that in HMMs we often work with scaled versions of the forward and backward variables, for example $\hat{\alpha}_g^i(k) = \left(\prod_{\tau=1}^{g} c_\tau^i\right)\alpha_g^i(k)$, where c^i here in this paragraph is the traditional scaling variable, of length N_g here, with for example first entry $c_1^i = \frac{1}{\sum_{k=1}^{K}\alpha_1^i(k)}$. Now for readability further consider the temporary variable $D = \frac{1}{c_g^{h_v}C^{h_v}}\frac{1}{c_g^{h_v^o}C^{h_v^o}}$ where $C^i = \prod_{\tau=1}^{N_g}\left(c_\tau^i\right)$, as well as the temporary

variable $E = \left(\frac{1}{2}\right)^{N_r}$. Then we have that

$$\begin{aligned}
&P(O = o, H_{-v} = h_{-v}, H_v = h_v^o, |\lambda) \\
=&P(O^{h_v,-v} = o^{h_v,-v}|\lambda)P(O^{h_v^o,+v} = o^{h_v^o,+v}|\lambda)E \\
=&\left(\sum_{k=1}^{K} \alpha_g^{h_v,-v}(k)\beta_g^{h_v}(k)\right)\left(\sum_{k=1}^{K} \alpha_g^{h_v^o,+v}(k)\beta_g^{h_v^o}(k)\right)E \\
=&\left(\sum_{k=1}^{K} \hat{\alpha}_g^{h_v,-v}(k)\hat{\beta}_g^{h_v}(k)\right)\left(\sum_{k=1}^{K} \hat{\alpha}_g^{h_v^o,+v}(k)\hat{\beta}_g^{h_v^o}(k)\right)DE \\
=&\left(\sum_{k=1}^{K} \frac{\hat{\alpha}_g^{h_v}(k)}{P(R_v = r_v|Q_g = k, \lambda)}\hat{\beta}_g^{h_v}(k)\right) \times \\
&\left(\sum_{k=1}^{K} \hat{\alpha}_g^{h_v^o}(k)P(R_v = r_v|Q_g = k, \lambda)\hat{\beta}_g^{h_v^o}(k)\right)DE
\end{aligned} \tag{19}$$

If we do not select the new label in Gibbs sampling, nothing changes to the forward and backward variables. If we do, we update the forward variable and scaling variable, continuing through all such reads at grid g, before we move to the next grid point and any reads inside it. Once we have completed an entire forward pass, we run a backwards pass to update the β variable, and can continue the Gibbs sampling anew.

1.1.6 Full haploid imputation

As mentionned above, with H, we can relax an assumption necessary for the HMM, while additionally switching to using the full reference panel. As described above, consider instead a model in which observations, denoted here with stars $*$ to denote the differences in the assumptions of the underlying model, reflect that the underlying sequenced molecule has bases defined by the reference haplotypes at that particular location, i.e. instead of $\mathrm{E}[s_{v,j}|u_{v,j} = t, Q_{c_v} = k] = \theta_{t,k}$we have that $\mathrm{E}[s_{v,j}|u_{v,j} = t, Q_{G_t} = k] = \theta_{t,k}$. Note that under this assumption the information from sequenced bases belonging to reads from a haplotype no longer depends on the specific read they are in, so we re-consider $o^i = \{r_v|h_v = i\}$ as $o^{*i} = \{(u_{v,j}, s_{v,j}, b_{v,j})|h_v = i\}$, and in particular, at a SNP t, this means that

$$o_t^{*i} = \{(s_{v,j}, b_{v,j})|u_{v,j} = t\} \tag{20}$$

We can merge results for a given SNP into haplotype likelihoods for SNP t and genotype $l \in \{0, 1\}$ as $e_{t,l}^{*i}$ with

$$e_{t,l}^{*i} = P(Gen_t = l|O_t^{*i} = o_t^{*i}) = \prod_{(v,j)|u_{v,j}=t} e_{v,j}^l \tag{21}$$

Where recall that $e_{v,j}^l = P(S_{v,j} = s_{r,j}|Gen = l)$. We therefore get that

$$P(O_g^{*i} = o_g^{*i}|Q_g^i = q_g^i, \lambda) = \prod_{t|G_t=g} \left(\theta_{t,q_g^i} e_{t,1}^{*i} + (1 - \theta_{t,q_g^i})e_{t,0}^{*i}\right) \tag{22}$$

With $e^{*i}_{t,l} = 1$ if there are no entries with $u_{v,j} = t$ for SNPs for haplotype i (*i.e.* no part of any read intersect SNP t). The full probability for a vector of hidden copying states q^i for haplotype i is therefore

$$P(O^{*i} = o^{*i}, Q^i = q^i|\lambda) = \frac{1}{K}\prod_{g=2}^{N_g} P(Q^i_g = q^i_g|Q^i_{g-1} = q^i_{g-1}, \lambda)\prod_{g=1}^{N_g} P(O^{*i}_g = o^{*i}_g|Q^i_g = q^i_g, \lambda) \tag{23}$$

where the transition probabilities are as defined above for the Gibbs sampler, using the appropriate recombination distance based on the full reference panel size. Posterior probabilities under this model can be calculated in the usual way for HMMs, and genotype dosages obtained from posterior probabilities in the obvious way *i.e.* for a diploid genotype dosage at SNP t of Gen_t, that

$$E[Gen_t|O, \lambda] = \sum_{i=1}^{2}\sum_{k=1}^{K}\theta_{t,k}P(Q^i_{G_t} = k|O^{*i} = o^{*i}, \lambda) \tag{24}$$

1.1.7 Full haploid imputation selection of new subset of the reference panel

To begin, we use H and the per-chromosome full reference panel imputation described above. This per-haplotype full reference panel K can be used to generate $\alpha^i_g(k)$, $\beta^i_g(k)$, and $\gamma^i_g(k)$, though in practice, we only record $\alpha^i_g(k)$, as these matrices are very large, and the posterior state and hence genotype probabilities can be calculated without recording β and γ.

Here, we first thin the list of grid points to be considered as a user defined parameter, here 10%, and then do calculate and store $\gamma^i_g(k)$ for those positions. Without loss of generality, consider that one of these thinned grids in indeed grid g, then we take $\gamma^i_g(k)$ separately for each haplotype i and do a partial sort of those values, which, for some user defined value of top matches to return, with default value 5, orders and stores indices for all reference haplotypes at that grid g with a posterior probability greater than or equal to the value of the 5^{th} top value.

Once this is complete for each haplotype i and each thinned grid point, we take the vector of indices of the previous reference panel subset (default size 400) and randomly select some subset to replace (default 100, call it A). We then start with the previously retained haplotypes (here, default 300, call it B).

We then run a while loop, where we increment some counter c by one per pass through of the while loop, starting from $c = 1$. We also start with A as empty. Suppose we have run through the while loop a few times, so that A is non-empty. Then we begin this pass through of the while loop by selecting a set of putative haplotypes to add, A^*, by taking the haplotypes across each thinned grid points that have the c^{th} best posterior probability γ^i_g for their i. We then make A^* unique by removing duplicates, and removing values already in A and B. If A^* fits into A such that the size of A is less than its maximum, we do this, and continue the while loop, incrementing c. Otherwise, we sample amongst A^*

at random to fill the rest of A. If we've run out of retained γ_g^i before we fill A, we fill the remainder of A at random among haplotypes not in B or A.

1.1.8 Block Gibbs sampling

Recall with normal Gibbs sampling, we need to efficiently calculate

$$P(H_v = h_v^o | H_{-v}, O, \lambda) = \frac{P(O, H_v = h_v^o, H_{-v} = h_{-v}|\lambda)}{\sum_{j=0}^{1} P(O, H_v = j, H_{-v} = h_{-v}|\lambda)} \tag{25}$$

with diploid block Gibbs sampling, we want to consider resampling some entire continuous block of reads $V \subseteq \{1, ..., N_r\}$. Consider that in that block we want to consider two sets of reads, first, the current set, and second, re-assigning each read to the opposite haplotype, *i.e.* we have two options, h_V and h_V^o, where each read is re-assigned in the obvious way, for example $h_{V_1}^o = (3 - h_{V_1})$. We therefore need to calculate

$$P(H_V = h_V^o | H_{-V} = h_{-V}, O, \lambda) = \frac{P(O = o, H_V = h_V^o, H_{-V} = h_{-V}|\lambda)}{\sum_{h_V^* \in \{h_V, h_V^o\}} P(O, H_V = h_V^*, H_{-V} = h_{-V}|\lambda)} \tag{26}$$

This is easy enough to calculate for random V but inefficient when considering multiple V. Opterationally, we therefore proceed as follows. First, we identify discrete blocks that we want to re-sample in, where discretization is chosen so that reads from the same grid cannot fall into different blocks, in a manner described below. This discretizes $v = 1, ..., N_r$ reads into $\{V_1, ..., \}$, such that $V_i \cap V_j = \emptyset \forall i, j$ and further $\forall v_a \in V_i, v_b \in V_j, c_{v_a} \neq c_{v_b}$. We can therefore proceed with updating in a single forward-backward pass, where over a particular V, we update the forward algorithm from the minimum $\min(c_v | v \in V)$ to the maximum grid $\max(c_v | v \in V)$ for the flipped option, and if we accept it in the Gibbs sampling, we update the forward variable α for the two chromosomes, and flip all read labels from $\min(V)$ to N_r.

As for the discretization process, we use code similar to calculating haploid updates from the STITCH model, and calculate, given the posterior probabilities for each haplotype, the expected number of switches between grids g and $g + 1$, over each pair of grids. We then sum this for the two haplotypes, and further, multiply it by the recombination rate, effectively focusing it around real potential recombination hotspots, and remove noise from this procedure. We then smooth this rate over a range, default 5000bp, and identify peaks using a peak finding algorithm, again, from the STITCH model heuristic for identifying ancestral haplotype breakpoints. This returns a discrete list of peaks being a set of grid points, from which we generate a discretized grid, and hence generate V_i as $V = \{V_1, ..., \}$ being the reads with central grid in a respective V_j.

1.1.9 Phasing

For any given Gibbs sample, considering all samplings as burn-in except for the single final sampling, it is possible to get haplotype dosage using the logic laid

out in Equation 9. However this does not work across Gibbs samplings due to the arbitrariness of the maternal and paternal haplotype. Therefore, briefly, for phasing, we attempt to define a consensus set of read labels across the Gibbs sampled iterations, and then we use this to initialize a single final sampling, whereupon completion we report the haploid dosages.

In more detail, suppose we have run the Gibbs samples, with final read labellings h^l for Gibbs samplings $l = 1, ..., N_{gs}$, where N_{gs} is the number of Gibbs samples, and h^l_v is the read label for read v in Gibbs sample l. We additionaly generate a binary 0/1 vector for each Gibbs sample, with length the number of reads, called x, where x^l_v is a binary 0/1 value indicating whether it is likely (1) or unlikely (0) that that read specifically came from one or the other imputed haplotype. In more detail towards x^l_v, first consider $p^l_{v,i} = P(R_v = r_v|D = d^{l,i}, \lambda)$ is the probability read v came from imputed haplotype i with imputed dosage $d^{l,i}$, i.e. for dosage d_t at SNP t that $P(R_v = r_v|D = d, \lambda) = \prod_{j=1}^{a_v}(d_{u_{v,j}}e^1_{v,j} + (1 - d_{u_{v,j}})e^0_{v,j})$, and then suppose we define

$$x^l_v = \begin{cases} 1 & \text{if } \frac{\max\left(p^l_{v,1}, p^l_{v,2}\right)}{p^l_{v,1}+p^l_{v,2}} > 0.95 \\ 0 & \text{otherwise} \end{cases}$$

In other words, x^l_v is a 1 if, after imputation, there is substantial evidence it came from either haplotypes 1 or 2 specifically, rather than being likely to come from either haplotype.

We next only considered reads for which x^l_v was 1 for all l, as reads confidently assigned in every run, consider this subset as h^l_{v*}. Consider an exemplar read labelling, which can be arbitrary, for convenience here choose 1, *i.e.* begin with the set of read labels that will initialize phasing, h^p as h^1, where here p just indicates *phasing* rather than a label. Then we determined all pairs of sequential reads, $v*$ and $v*+1$, where there existed a difference in label between the read pairs at those respective positions, *i.e.* where $\sum_{a=1}^{N_{gs}} |h^a_{v*} - h^1_{v*}| \neq \sum_{a=1}^{N_{gs}} |h^a_{v*+1} - h^1_{v*+1}|$. In choosing which read label to keep between those, we took the majority vote, *i.e.* with respect to the canonical read labelling, if more than half of the read labels changed between their previous and new positions, we changed the canonical read label from that read through to the ends of reads, *i.e.* if $\frac{\sum_{a=1}^{N_{gs}} |h^a_{v*+1} - h^a_{v*}|}{N_{gs}} > 0.5$, we flipped h^p_{v*} for $v*$ to the end of the confidentally chosen reads, as well as for the canonical haplotype.

In this way, we used the read labels h^l from the Gibbs samples, to arrive at a new set of read labels that attempted to model the Gibbs sampled read labels with as few switches as possible, and used this as the starting value for h for another round of the model.

1.1.10 Other parameters

We used default parameters of QUILT including 400 haplotypes in the Gibbs sampler, 7 Gibbs samples, 3 iterations of Gibbs sampling and full per-haplotype imputations. When updating read labels given a reference panel subset, per

round of Gibbs sampling, we did 20 full iterations across all read labels, with block Gibbs on iterations 3, 6 and 9. When updating reference panel subset given read labels, we thinned to 10% of SNPs, sought to re-select 100 haplotypes out of 400, and kept the top 5 matches at each thinned grid point for later consideration. Other parameters are as defined in the help pages of the QUILT software.

1.2 HLA imputation

1.2.1 Reference panel construction

We downloaded full-length HLA alignments for annotated HLA genes and pseudogenes HLA-A to HLA-Y, from the HLA database IPD-IMGT/HLA [4]. This provides a set of (aligned) sequenced alleles for each region, hereafter the "HLA database sequences". Because we utilized these alleles directly in calling, and due to extensive complications in genotyping and mapping reads in these regions within standard pipelines, we excluded SNPs in these regions themselves from the QUILT component of HLA typing, but retained SNPs in flanking regions, and used reads intersecting these regions to regain information from the excluded SNPs.

For each HLA region in the set HLA-A, -B, -C, -DQB1 and -DRB1 we constructed a panel of reference haplotypes using the HRC reference data for a subset of individuals drawn from the 1000 Genomes Project[5], and a set of HLA types previously inferred using high-coverage exome sequence data for 1000 genomes samples, and the PolyPheMe software [6]. These were previously demonstrated to have high accuracy [6]. These reference panels excluded all members of the 5 tests populations (ASW, CEU, CHB, PJL, PUR), and had between 3674 (DQB1) and 4082 (C) haplotypes in them.

We fixed the identified HRC haplotypes for each individual, and phased the pair of HLA types against these as follows, to identify a single HLA allele for each haplotype. First, we used the HLA database sequences to construct SNP haplotypes at HRC sites by identifying which allele was carried by each haplotype at variant HRC sites they overlapped. For each 4-digit allele, this yielded a set of predicted probabilities of carrying the non-reference allele (averaging over all 4, 6 or 8 digit sequences carrying that allele) at each site. Now for 1000G HLA-typed individuals' HRC data, we calculated the difference between both possible phasings and their predicted types based on their HLA alleles, counting mismatches as cases with a difference between observed type and predicted type probabilities of >0.9. We then took the phasing producing the smaller number of mismatches, provided this was < 4 mismatches while the alternative showed > 4 mismatches, or if only one HLA type was obtained and overlapped the database, provided this had at least 2 fewer mismatches than the alternative phasing. Although this approach phased the majority of individuals, there were remaining cases with 1000 genomes derived HLA-types not matching those in the reference database and unclear phasing (e.g. because both possible phasings gave similar numbers of mismatches within the HLA region itself). To phase

these, we used other phased HRC individuals directly, now extending the HLA region to include sites outside the HLA gene itself: first identifying haplotypes for each HLA-allele at each site (probabilistically averaging over sequences carrying that allele), and then using these new probabilities as above i.e. counting mismatching sites with a difference of > 0.9. We identified a phasing if one phasing over the extended region produced at least two fewer mismatches than the other. This was applied successively, fixing already-phased individuals and extending the HLA gene surrounds by successively adding in sets of 50 SNPs upstream and downstream, until all individuals were phased (or else an upper limit of 1000 SNPs was reached). Data for any remaining unphased individuals, those with no identified type at a particular HLA gene or those with types not identified in our HLA database sequences, was removed from the panel for that region.

1.2.2 QUILT-HLA state inference

For each of the HLA genes listed above, we generated a reference panel using the phased HLA data as above, the HRC reference haplotypes for the corresponding haplotypes, and a window of 500kb on either side, removing SNPs falling within HLA genes as described above. We then ran a modified version of QUILT. As the number of reference haplotypes was comparatively small, we used a single stage Gibbs sampler including all reference haplotypes (approximately 4000), using the final parental haplotype read labels for that sample in a final full panel per-haplotype imputation to get per-haplotype posterior state probabilities. We generated 20 Gibbs samples per sample, retaining posterior state probabilities from phasing at each sample, so that per sampling, we did the following. Let A^i be the underlying allele for some arbitrary HLA gene of interest for haplotype i (e.g. the truth for some sample could be $A^1 = 2$ and $A^2 = 4$ where the 2nd indexed HLA-allele is HLA-A*01:02 and the 4th indexed HLA allele could be HLA-A*02:01), and let B_j be the set of reference haplotypes that contain allele j (e.g. $B_j = \{3, 6\}$ means the 3rd and 6th reference haplotypes have allele indexed by j). Then per Gibbs sampling, for the SNP indexed by t that is physically closest to the physical center of the HLA gene, and at grid position G_t, for haplotype i, we calculated the prediction using posterior state probability as

$$P(A^i = j|O^i, \lambda) = \sum_{k \in B_j} P(Q^i_{G_t} = k|O, \lambda) \tag{27}$$

We then calculated joint probabilities as the product of the per-haplotype probabilities

$$P((A^1, A^2) = (j_1, j_2)|O, \lambda) = P(A^1 = j_1|O^1, \lambda) \times P(A^2 = j_2|O^2, \lambda) \tag{28}$$

and could therefore sum this across Gibbs samplings, to yield per-sample phased HLA allele probabilities. We then summed across phase identical probabilities, and took the most likely pair as the inferred allele, and assigned a confidence as the posterior probability of that allele pairing.

1.2.3 QUILT read based HLA typing

The above approach specifically avoids using information on reads falling within HLA genes themselves, but instead captures long-range LD among carriers of particular HLA-allelic types. To capture read-based information, we constructed a likelihood for each HLA database sequence and each of the five HLA genes, as follows.

First, we identified reads potentially mapping within the corresponding gene by using all those whose mapped positions fell within the appropriate region of the chromosome 6 reference sequence (extended 1000bp upstream to conservatively capture reads beginning upstream). We also identified reads mapped using bwa-mem to alternative HLA contigs, e.g. HLA-A*69:01 is a contig of length 2917. We used both types of read going forward. A challenge of HLA-typing is mismapped (or unmapped) reads among a highly homologous set of genes and pseudogenes, particularly in certain regions of these genes, and we utilized filters to remove suspicious reads, rather than simply using for example a minimum mapping quality. A second challenge is the highly polymorphic nature of these regions, which have an extremely high density of SNPs and indels distinguishing HLA database sequences, and this makes remapping of reads essential.

As an initial set of filters we removed reads (i) whose mate pair mapped on another chromosome, or $>$ 1000bp away for reads mapping on chromosome 6, or where (ii) the read itself has an alternative mapping to another chromosome, or elsewhere on chromosome 6 than the HLA-gene under consideration. A second set of filters tested directly if each read (or its complementary sequence) could be specifically mapped to a single HLA sequence, using the full HLA database sequence set. We took bases 11-20, 21-30, 121-130 and 131-140 of each 150-bp read (for shorter or longer reads, adapting in the obvious manner), to define four 10-bp sequences. For each of these, we identified all possible HLA sequence matches (e.g. an exact match to some 10-bp region of HLA-A*01:01:01:01), in not just the HLA gene being considered, but across the entire database. Considering these as possible matches, we found all HLA haplotypes with at least two matches and the expected base-pair spacing between these matches. We retained all read pairs where at least one of the pair could map to the expected HLA region, only to this region, and where the mate pair did not map uniquely to some other region. We discarded all other reads, i.e. those with no mapping, with non-unique mapping, or with inconsistent mate-pair mapping. The 10-bp matches also immediately identified a strand for each read pair, and a set of possible aligned read start positions within the appropriate HLA region (with potentially several start positions in the same region among indel-containing HLA alleles in the database), which we used going forward. Note that these filters will remove, potentially, reads containing indels, or with sufficient errors that they do not possess exact matching over at least two of the four 10-bp regions considered here, which is potentially problematic for less high-quality sequencing. However they have the advantage of considering the large space of possible allelic types across a broad set of HLA loci, and rapidly remapping reads

over this space. Having filtered reads, we then constructed a likelihood for the remaining set of curated reads as follows, using the HLA database sequences for the region. For each read, we calculated, for each potential start point of that read in the region-wide alignment, a likelihood for all possible HLA sequences. Given a read of length l and a start point a for an alignment, against potential HLA sequence i, this likelihood is calculated by ignoring the possibility of novel indels as follows (indels are permitted among database HLA sequences). (We adjust our notation here compared to other sections discussing reads, because in this analysis we do not merely consider SNPs, but all bases – because, for example, indels and multi-base variants occur frequently among HLA alleles, and variant positions can overlap within the highly polymorphic HLA.)

Let the sequenced bases s_p of a read indexed by p be

$$s_{p1}, s_{p2}, s_{p3}, ..., s_{pl} \tag{29}$$

and the corresponding region of the reference sequence be

$$r_{ia}, r_{i(a+1)}, r_{i(a+2)}, ..., r_{i(a+l-1)} \tag{30}$$

Then if we have a Phred quality score q_{pk} for base k, we attach a likelihood of observing base s_{pk} of $1-10^{-q_{pk}/10}$ if $s_{pk} = r_{i(a+k-1)}$ and for a mismatching base, assuming all sequencing errors are equally likely, a likelihood of $10^{-q_{pk}/10}/3$ if $s_{pk} \neq r_{i(a+k-1)}$. Assuming all bases are independent, the overall likelihood of read p coming from a true allele i is then maximized over possible a:

$$L(i; s_p) = \max_a \prod_{k=1}^{l} \left(1 - 10^{-q_{pk}/10}\right)^{I_{s_{pk}=r_{i(a+k-1)}}} \left(10^{-q_{pk}/10}/3\right)^{1-I_{s_{pk}=r_{i(a+k-1)}}} \tag{31}$$

Paired-end reads (which come from the same haplotype *i.e.* HLA allele) are dealt with by multiplying the likelihoods from each member of the read pair together, to make a single likelihood for the read pair which we also denote by $L(i; s_p)$. If this maximized likelihood is below the equivalent of that for a read with 5 mismatching bases of Phred-score 30 ($p = 0.001$ of a mismatch), the read was considered as mismatching all alleles and excluded from the overall likelihood calculation below.

Given a total of n reads from this individual, for a possible pair of alleles i, j this individual carries, we may then construct the overall likelihood of that individual as

$$L(i, j; s) = \prod_{p=1}^{n} \frac{1}{2} \left(L(i; s_p) + L(j; s_p)\right) \tag{32}$$

We calculate this for all pairs of alleles, and multiply the corresponding likelihoods by those from QUILT, with a few necessary modifications detailed below, to obtain a joint likelihood. Because the QUILT approach uses only information outside the HLA target gene while this approach uses only reads from within this gene, we may regard the information as complementary. We note the above

is a full-likelihood approach, at least for those reads we manage to uniquely and correctly map in the target gene, and subject to (in)completeness of the database.

1.2.4 Combining QUILT state inference with sequence-based information

Because the 1000 genomes HLA types are all four-digit types, and the above HLA database sequences have varying accuracies (up to 8-digit accuracy), and there is incomplete overlap between the two, it is not completely straightforward to combine their information. To do this, we only use likelihoods for four-digit alleles in the intersection of the two sets, mainly removing alleles in the HLA database that were not observed in the 1000 genomes samples. Because the HLA database can have many alleles with the same 4-digit resolution, we assign a uniform prior on all alleles with a particular 4-digit code, so the read-based likelihood is the average over the high-resolution terms of all pairs of alleles with a given code. Then, we simply multiply the resulting likelihoods together to yield a combined score for each possible pair of HLA types at each gene.

1.2.5 SNP2HLA

We imputed HLA alleles in a manner emulating the approach of SNP2HLA in the following manner. First, similar to the approach for imputation generally described in the main text for arrays, we generated genotypes in short non-overlapping windows at array sites for the UK Biobank Axiom array and for the Illumina Global Screening Array using the GATK UnifiedGenotyper module, using high coverage whole genome sequences filtered to array sites. We then combined these into a single input file, one for each array type and HLA-gene, imputing using SNPs from 500kbp upstream to 500kbp downstream of the centre of that HLA gene. We then performed imputation from genotypes using the specified version of Beagle from SNP2HLA (Beagle 4.1) and using the HRC reference panel, removing from the HRC panel data for NA12878 or each of the 5 populations being tested. Next, using the same reference panel we used for QUILT based HLA typing, we built a SNP2HLA reference panel for each gene. This takes the form of a phased VCF, with each HLA allele encoded as a distinct SNP in the reference panel. In other words, in the SNP2HLA style phased VCF, the summed dosage of each reference haplotype over the SNPs representing the different HLA alleles would be 1, and the sum of the dosages for the two haplotypes, representing one sample, would be 2. Next, using the output of the HRC based imputation as input genotypes, and using the above-described custom SNP2HLA style reference panel, we imputed HLA alleles using Beagle 4.1 using options impute=true, grobs=true, niterations=5. We then generated per-sample per-HLA-gene HLA alleles and a measure of confidence from the output in the following way. For a single sample, for a single HLA-gene, the above approach yields dosages for all alleles. We extracted those with non-zero values, collapsed them to be unique to 4 digit accuracy if not already done, and

then re-normalized them to have sum 1. If there was only a single imputed allele for a given sample for that gene, we took this to imply the inferred type was homozygous for that allele (with confidence 1). If there was more than one imputed allele and the dosage for the most likely allele was more than twice as likely as the second most likely allele, we took this to mean the sample was homozygous for the most likely allele, with the normalized dosage for that allele as the confidence. Otherwise, we took as the prediction the two most likely alleles, and the confidence the sum of their normalized dosages. In this way we would necessarily for each sample get a diploid genotype of HLA alleles and a confidence score between 0 and 1.

1.3 Haplotagging

Haplotagging uses barcoded and bead-bound Tn5 transposomes to fragment ("tagment") and barcode genomic DNA. Effectively, multiple reads belonging to the same input DNA molecule, i.e., a haplotype, will be tagged with the same 4-segment DNA barcode under Illumina sequencing (split into i7 and i5 indexes, each 13 nucleotides long). Under haplotagging, as presently configured up to 96 individuals can be multiplexed in a single Illumina sequencing lane by means of one of the four barcode segments (designed "C" segment here). We performed haplotagging essentially as described in Meier *et al.*[7]. Briefly, for each DNA sample (NA12878 and the 5-Family samples), we extracted high molecular weight DNA (typically ¿ 100 kbp) using Nanobind (Circulomics, Baltimore MA, USA) or in-house equivalent. For the 1000G-GBR samples (MGP00003) we directly obtained extracted DNA from the Coriell Institute for Medical Research. For each sample, 3 ng of DNA were tagmented using 5 μl of haplotagging beads (approximately 3.5 million beads and thus 3.5 million barcodes). Next, either the entire batch of beads for NA12878, or each sample were subsampled to obtain a pool of 3.5 million beads that were used for a subsequent thermocycling reaction to amplify the haplotag libraries as described in Meier *et al.*[7]. These libraries were pooled and sequenced on a HiSeq3000 instrument as a 2x150 paired-end run with extra index cycles of 13 and 12nt at the Genome Core of the Max Planck Institute for Developmental Biology, Tübingen, Germany. The resulting data were demultiplexed as described in Meier *et al.*[7] prior to standard read alignment and further analysis. The barcode information is retained in the BX tag in BAM files.

References

[1] Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nature Genetics. 2016 Jul;advance online publication. Available from: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3594.html.

[2] Bowden R, Davies RW, Heger A, Pagnamenta AT, Cesare Md, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. Nature Communications. 2019 Apr;10(1):1–9. Available from: https://www.nature.com/articles/s41467-019-09637-5.

[3] Li N, Stephens M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics. 2003 Jan;165(4):2213–2233. Available from: http://www.genetics.org/content/165/4/2213.

[4] Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Research. 2020;48(D1):D948–D955.

[5] "The 1000 Genomes Project Consortium". A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74. Available from: http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html.

[6] Abi-Rached L, Gouret P, Yeh JH, Cristofaro JD, Pontarotti P, Picard C, et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. PLOS ONE. 2018 Oct;13(10):e0206512. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0206512.

[7] Meier JI, Salazar PA, Kučka M, Davies RW, Dréau A, Aldás I, et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. bioRxiv. 2020 Jun;p. 2020.05.25.113688. Publisher: Cold Spring Harbor Laboratory Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2020.05.25.113688v2.

[8] Karnes JH, Shaffer CM, Bastarache L, Gaudieri S, Glazer AM, Steiner HE, et al. Comparison of HLA allelic imputation programs. PLoS ONE. 2017 Feb;12(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5312875/.

2 Supplementary Tables

	Method	Input	pse	pse2	disc
1	Optimal	Illumina ht	0.04	4.38	5.82
2	Optimal	Illumina	0.03	4.33	5.78
3	Optimal	ONT	0.12	13.03	17.79
4	QUILT	Illumina ht	0.23	8.9	13.61
5	QUILT	Illumina	0.28	10.94	17.07
6	QUILT	ONT	0.36	18.35	27.92
7	GLIMPSE	Illumina	0.63	13.74	21.47
8	GLIMPSE	ONT	1.72	29.52	50.5

(a) 0.1X coverage

	Method	Input	pse	pse2	disc
1	Optimal	Illumina ht	0.01	2.04	2.57
2	Optimal	Illumina	0.01	1.95	2.46
3	Optimal	ONT	0.03	4.75	6.7
4	QUILT	Illumina ht	0.11	3.72	5.26
5	QUILT	Illumina	0.18	5.13	7.74
6	QUILT	ONT	0.18	7.99	12.1
7	GLIMPSE	Illumina	0.23	5.4	7.96
8	GLIMPSE	ONT	0.7	20.32	35.09

(b) 0.25X coverage

	Method	Input	pse	pse2	disc
1	Optimal	Illumina ht	0	1.36	1.53
2	Optimal	Illumina	0	1.31	1.5
3	Optimal	ONT	0.01	2.38	3.12
4	QUILT	Illumina ht	0.09	2.07	2.66
5	QUILT	Illumina	0.11	2.62	3.52
6	QUILT	ONT	0.14	4.98	7.51
7	GLIMPSE	Illumina	0.14	2.51	3.3
8	GLIMPSE	ONT	0.59	16.05	27.98

(c) 0.5X coverage

	Method	Input	pse	pse2	disc
1	Optimal	Illumina ht	0	0.94	1.08
2	Optimal	Illumina	0	0.96	1.07
3	Optimal	ONT	0.01	1.3	1.58
4	QUILT	Illumina ht	0.08	1.48	1.75
5	QUILT	Illumina	0.13	1.73	2.12
6	QUILT	ONT	0.09	2.28	3.24
7	GLIMPSE	Illumina	0.13	1.66	2.04
8	GLIMPSE	ONT	0.61	11.18	19.12

(d) 1.0X coverage

	Method	Input	pse	pse2	disc
1	Optimal	Illumina ht	0	0.72	0.81
2	Optimal	Illumina	0	0.74	0.82
3	Optimal	ONT	0	0.98	1.12
4	QUILT	Illumina ht	0.1	1.13	1.27
5	QUILT	Illumina	0.15	1.23	1.32
6	QUILT	ONT	0.08	1.52	2.01
7	GLIMPSE	Illumina	0.15	1.25	1.34
8	GLIMPSE	ONT	0.7	6.95	11.03

(e) 2.0X coverage

Supplementary Table 1: Effect of coverage on phasing performance

Method	Optimal	Optimal	Optimal	QUILT	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	HT	Illumina	ONT	HT	Illumina	ONT	Illumina	ONT
Cov	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
(0,0.0001]	0.102	0.089	0.079	0.062	0.058	0.048	0.054	0.019
(0.0001,0.0002]	0.275	0.221	0.176	0.171	0.149	0.079	0.142	0.02
(0.0002,0.0005]	0.336	0.308	0.187	0.192	0.193	0.097	0.146	0.032
(0.0005,0.001]	0.531	0.522	0.329	0.448	0.402	0.222	0.304	0.105
(0.001,0.002]	0.606	0.593	0.457	0.46	0.416	0.343	0.335	0.148
(0.002,0.005]	0.687	0.677	0.504	0.552	0.502	0.384	0.423	0.207
(0.005,0.01]	0.724	0.742	0.567	0.602	0.563	0.441	0.479	0.265
(0.01,0.02]	0.819	0.816	0.627	0.718	0.679	0.507	0.609	0.345
(0.02,0.05]	0.88	0.876	0.682	0.807	0.772	0.582	0.707	0.424
(0.05,0.1]	0.923	0.918	0.764	0.857	0.825	0.678	0.774	0.5
(0.1,0.2]	0.933	0.929	0.764	0.861	0.829	0.686	0.794	0.519
(0.2,0.5]	0.934	0.934	0.769	0.869	0.837	0.68	0.806	0.526
(0.5,0.95]	0.937	0.936	0.789	0.874	0.849	0.711	0.82	0.574
(0.95,1]	0.869	0.858	0.748	0.827	0.822	0.659	0.748	0.488

(a) 0.1X

Method	Optimal	Optimal	Optimal	QUILT	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	HT	Illumina	ONT	HT	Illumina	ONT	Illumina	ONT
Cov	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
(0,0.0001]	0.148	0.153	0.093	0.103	0.117	0.07	0.131	0.033
(0.0001,0.0002]	0.355	0.339	0.198	0.289	0.276	0.156	0.285	0.06
(0.0002,0.0005]	0.478	0.462	0.397	0.367	0.38	0.346	0.371	0.137
(0.0005,0.001]	0.64	0.636	0.565	0.553	0.534	0.475	0.522	0.283
(0.001,0.002]	0.68	0.685	0.629	0.635	0.605	0.581	0.562	0.322
(0.002,0.005]	0.775	0.77	0.703	0.721	0.69	0.618	0.661	0.369
(0.005,0.01]	0.844	0.842	0.778	0.792	0.753	0.698	0.734	0.465
(0.01,0.02]	0.893	0.893	0.842	0.852	0.817	0.776	0.801	0.474
(0.02,0.05]	0.934	0.937	0.883	0.91	0.891	0.832	0.875	0.659
(0.05,0.1]	0.965	0.966	0.918	0.941	0.928	0.881	0.917	0.71
(0.1,0.2]	0.972	0.973	0.914	0.951	0.935	0.871	0.927	0.719
(0.2,0.5]	0.977	0.979	0.916	0.954	0.94	0.87	0.931	0.685
(0.5,0.95]	0.978	0.982	0.924	0.962	0.948	0.883	0.942	0.742
(0.95,1]	0.922	0.914	0.848	0.884	0.852	0.826	0.85	0.555

(b) 0.25X

Method	Optimal	Optimal	Optimal	QUILT	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	HT	Illumina	ONT	HT	Illumina	ONT	Illumina	ONT
Cov	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
(0,0.0001]	0.164	0.183	0.121	0.148	0.14	0.122	0.18	0.077
(0.0001,0.0002]	0.38	0.374	0.322	0.319	0.29	0.263	0.352	0.131
(0.0002,0.0005]	0.491	0.515	0.453	0.493	0.469	0.381	0.477	0.228
(0.0005,0.001]	0.667	0.684	0.633	0.638	0.606	0.543	0.609	0.304
(0.001,0.002]	0.744	0.76	0.7	0.709	0.678	0.669	0.672	0.428
(0.002,0.005]	0.807	0.812	0.761	0.784	0.752	0.686	0.738	0.458
(0.005,0.01]	0.869	0.871	0.849	0.851	0.831	0.797	0.822	0.594
(0.01,0.02]	0.928	0.927	0.89	0.905	0.892	0.845	0.892	0.668
(0.02,0.05]	0.955	0.957	0.931	0.945	0.935	0.908	0.934	0.725
(0.05,0.1]	0.98	0.981	0.96	0.973	0.967	0.936	0.966	0.782
(0.1,0.2]	0.984	0.986	0.963	0.978	0.972	0.934	0.97	0.782
(0.2,0.5]	0.987	0.988	0.965	0.98	0.975	0.937	0.974	0.771
(0.5,0.95]	0.989	0.99	0.971	0.983	0.98	0.941	0.978	0.798
(0.95,1]	0.925	0.931	0.926	0.918	0.9	0.898	0.914	0.679

(c) 0.5X

Method	Optimal	Optimal	Optimal	QUILT	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	HT	Illumina	ONT	HT	Illumina	ONT	Illumina	ONT
Cov	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
(0,0.0001]	0.172	0.213	0.153	0.146	0.134	0.143	0.209	0.104
(0.0001,0.0002]	0.387	0.429	0.282	0.334	0.351	0.268	0.372	0.098
(0.0002,0.0005]	0.555	0.58	0.494	0.531	0.532	0.466	0.509	0.242
(0.0005,0.001]	0.763	0.762	0.697	0.729	0.713	0.655	0.72	0.389
(0.001,0.002]	0.809	0.83	0.746	0.776	0.754	0.726	0.765	0.46
(0.002,0.005]	0.849	0.853	0.823	0.824	0.81	0.799	0.801	0.558
(0.005,0.01]	0.913	0.918	0.886	0.897	0.892	0.862	0.876	0.653
(0.01,0.02]	0.948	0.954	0.935	0.938	0.93	0.911	0.92	0.674
(0.02,0.05]	0.971	0.971	0.965	0.962	0.958	0.954	0.954	0.802
(0.05,0.1]	0.985	0.985	0.981	0.981	0.978	0.972	0.977	0.839
(0.1,0.2]	0.989	0.989	0.983	0.986	0.983	0.972	0.983	0.847
(0.2,0.5]	0.991	0.992	0.985	0.988	0.986	0.975	0.986	0.839
(0.5,0.95]	0.993	0.993	0.987	0.989	0.988	0.979	0.987	0.865
(0.95,1]	0.963	0.957	0.942	0.955	0.943	0.927	0.955	0.709

(d) 1.0X

Method	Optimal	Optimal	Optimal	QUILT	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	HT	Illumina	ONT	HT	Illumina	ONT	Illumina	ONT
Cov	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
(0,0.0001]	0.272	0.276	0.194	0.252	0.209	0.188	0.274	0.108
(0.0001,0.0002]	0.521	0.537	0.347	0.434	0.422	0.34	0.456	0.139
(0.0002,0.0005]	0.645	0.648	0.574	0.631	0.625	0.572	0.551	0.329
(0.0005,0.001]	0.778	0.794	0.748	0.774	0.768	0.734	0.714	0.419
(0.001,0.002]	0.854	0.854	0.82	0.811	0.806	0.784	0.785	0.511
(0.002,0.005]	0.877	0.888	0.862	0.864	0.853	0.841	0.832	0.611
(0.005,0.01]	0.934	0.938	0.922	0.93	0.925	0.907	0.908	0.744
(0.01,0.02]	0.966	0.967	0.952	0.959	0.953	0.946	0.939	0.81
(0.02,0.05]	0.978	0.978	0.97	0.974	0.972	0.964	0.967	0.875
(0.05,0.1]	0.988	0.988	0.987	0.986	0.986	0.984	0.985	0.901
(0.1,0.2]	0.992	0.992	0.988	0.99	0.99	0.983	0.988	0.913
(0.2,0.5]	0.994	0.994	0.991	0.992	0.992	0.987	0.991	0.906
(0.5,0.95]	0.994	0.995	0.992	0.993	0.993	0.988	0.991	0.912
(0.95,1]	0.955	0.95	0.963	0.958	0.954	0.949	0.96	0.72

(e) 2.0X

Supplementary Table 2: Imputation accuracy for NA12878 across methods, data and coverage

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle	Beagle
Data	HT	HT	HT	HT	HT	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.062	0.103	0.148	0.146	0.252	0.188	0.121
(0.0001,0.0002]	0.171	0.289	0.319	0.334	0.434	0.276	0.259
(0.0002,0.0005]	0.192	0.367	0.493	0.531	0.631	0.43	0.43
(0.0005,0.001]	0.448	0.553	0.638	0.729	0.774	0.606	0.566
(0.001,0.002]	0.46	0.635	0.709	0.776	0.811	0.694	0.683
(0.002,0.005]	0.552	0.721	0.784	0.824	0.864	0.745	0.72
(0.005,0.01]	0.602	0.792	0.851	0.897	0.93	0.819	0.824
(0.01,0.02]	0.718	0.852	0.905	0.938	0.959	0.885	0.905
(0.02,0.05]	0.807	0.91	0.945	0.962	0.974	0.95	0.944
(0.05,0.1]	0.857	0.941	0.973	0.981	0.986	0.979	0.971
(0.1,0.2]	0.861	0.951	0.978	0.986	0.99	0.986	0.978
(0.2,0.5]	0.869	0.954	0.98	0.988	0.992	0.989	0.984
(0.5,0.95]	0.874	0.962	0.983	0.989	0.993	0.989	0.981
(0.95,1]	0.827	0.884	0.918	0.955	0.958	0.931	0.933

(a) Haplotagged Illumina

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle	Beagle
Data	Illumina	Illumina	Illumina	Illumina	Illumina	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.058	0.117	0.14	0.134	0.209	0.188	0.121
(0.0001,0.0002]	0.149	0.276	0.29	0.351	0.422	0.276	0.259
(0.0002,0.0005]	0.193	0.38	0.469	0.532	0.625	0.43	0.43
(0.0005,0.001]	0.402	0.534	0.606	0.713	0.768	0.606	0.566
(0.001,0.002]	0.416	0.605	0.678	0.754	0.806	0.694	0.683
(0.002,0.005]	0.502	0.69	0.752	0.81	0.853	0.745	0.72
(0.005,0.01]	0.563	0.753	0.831	0.892	0.925	0.819	0.824
(0.01,0.02]	0.679	0.817	0.892	0.93	0.953	0.885	0.905
(0.02,0.05]	0.772	0.891	0.935	0.958	0.972	0.95	0.944
(0.05,0.1]	0.825	0.928	0.967	0.978	0.986	0.979	0.971
(0.1,0.2]	0.829	0.935	0.972	0.983	0.99	0.986	0.978
(0.2,0.5]	0.837	0.94	0.975	0.986	0.992	0.989	0.984
(0.5,0.95]	0.849	0.948	0.98	0.988	0.993	0.989	0.981
(0.95,1]	0.822	0.852	0.9	0.943	0.954	0.931	0.933

(b) Illumina

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle	Beagle
Data	ONT	ONT	ONT	ONT	ONT	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.049	0.072	0.117	0.147	0.188	0.188	0.121
(0.0001,0.0002]	0.091	0.158	0.284	0.279	0.357	0.276	0.259
(0.0002,0.0005]	0.117	0.337	0.403	0.488	0.576	0.43	0.43
(0.0005,0.001]	0.226	0.48	0.56	0.664	0.733	0.606	0.566
(0.001,0.002]	0.354	0.58	0.676	0.741	0.784	0.694	0.683
(0.002,0.005]	0.382	0.624	0.694	0.802	0.844	0.745	0.72
(0.005,0.01]	0.466	0.7	0.793	0.864	0.911	0.819	0.824
(0.01,0.02]	0.522	0.78	0.85	0.913	0.946	0.885	0.905
(0.02,0.05]	0.588	0.836	0.912	0.955	0.964	0.95	0.944
(0.05,0.1]	0.681	0.884	0.938	0.973	0.984	0.979	0.971
(0.1,0.2]	0.694	0.868	0.934	0.974	0.984	0.986	0.978
(0.2,0.5]	0.685	0.87	0.938	0.976	0.987	0.989	0.984
(0.5,0.95]	0.716	0.881	0.944	0.98	0.988	0.989	0.981
(0.95,1]	0.66	0.836	0.89	0.928	0.952	0.931	0.933

(c) ONT

Supplementary Table 3: Imputation accuracy for NA12878 across methods, data and coverage

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	HT	Illumina	Illumina
Cov	0.1	0.1	0.1	0.1	0.1
(0,0.0001]	0.024	0.025	0.021	0.019	0.015
(0.0001,0.0002]	0.161	0.161	0.165	0.145	0.116
(0.0002,0.0005]	0.283	0.286	0.281	0.25	0.184
(0.0005,0.001]	0.443	0.446	0.442	0.405	0.327
(0.001,0.002]	0.458	0.463	0.45	0.407	0.328
(0.002,0.005]	0.521	0.527	0.518	0.461	0.364
(0.005,0.01]	0.635	0.636	0.625	0.569	0.476
(0.01,0.02]	0.714	0.709	0.71	0.655	0.564
(0.02,0.05]	0.835	0.837	0.822	0.774	0.705
(0.05,0.1]	0.882	0.885	0.867	0.823	0.761
(0.1,0.2]	0.894	0.898	0.881	0.837	0.782
(0.2,0.5]	0.908	0.911	0.891	0.85	0.804
(0.5,0.95]	0.914	0.917	0.899	0.861	0.818
(0.95,1]	0.736	0.74	0.722	0.678	0.597

(a) 0.1X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	HT	Illumina	Illumina
Cov	0.25	0.25	0.25	0.25	0.25
(0,0.0001]	0.032	0.031	0.032	0.029	0.029
(0.0001,0.0002]	0.197	0.197	0.208	0.199	0.187
(0.0002,0.0005]	0.339	0.339	0.343	0.329	0.308
(0.0005,0.001]	0.502	0.498	0.505	0.492	0.465
(0.001,0.002]	0.524	0.525	0.531	0.514	0.478
(0.002,0.005]	0.601	0.596	0.609	0.582	0.538
(0.005,0.01]	0.708	0.709	0.721	0.692	0.655
(0.01,0.02]	0.795	0.795	0.801	0.778	0.742
(0.02,0.05]	0.895	0.896	0.895	0.878	0.85
(0.05,0.1]	0.938	0.941	0.937	0.922	0.906
(0.1,0.2]	0.947	0.952	0.948	0.933	0.916
(0.2,0.5]	0.957	0.96	0.956	0.941	0.926
(0.5,0.95]	0.957	0.96	0.956	0.942	0.927
(0.95,1]	0.795	0.794	0.804	0.785	0.768

(b) 0.25X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	HT	Illumina	Illumina
Cov	0.5	0.5	0.5	0.5	0.5
(0,0.0001]	0.04	0.04	0.038	0.038	0.053
(0.0001,0.0002]	0.222	0.225	0.228	0.228	0.235
(0.0002,0.0005]	0.364	0.368	0.372	0.367	0.366
(0.0005,0.001]	0.536	0.538	0.537	0.533	0.527
(0.001,0.002]	0.563	0.564	0.568	0.559	0.551
(0.002,0.005]	0.632	0.634	0.64	0.628	0.615
(0.005,0.01]	0.745	0.75	0.758	0.746	0.73
(0.01,0.02]	0.825	0.83	0.831	0.822	0.808
(0.02,0.05]	0.913	0.917	0.917	0.911	0.904
(0.05,0.1]	0.95	0.954	0.954	0.949	0.944
(0.1,0.2]	0.961	0.965	0.964	0.961	0.954
(0.2,0.5]	0.968	0.971	0.97	0.967	0.961
(0.5,0.95]	0.968	0.971	0.97	0.966	0.961
(0.95,1]	0.818	0.824	0.823	0.814	0.814

(c) 0.5X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	HT	Illumina	Illumina
Cov	1.0	1.0	1.0	1.0	1.0
(0,0.0001]	0.05	0.05	0.047	0.048	0.078
(0.0001,0.0002]	0.235	0.234	0.247	0.244	0.264
(0.0002,0.0005]	0.398	0.4	0.405	0.404	0.417
(0.0005,0.001]	0.561	0.564	0.563	0.563	0.575
(0.001,0.002]	0.588	0.591	0.594	0.588	0.593
(0.002,0.005]	0.658	0.66	0.662	0.658	0.656
(0.005,0.01]	0.767	0.776	0.778	0.773	0.77
(0.01,0.02]	0.845	0.849	0.849	0.846	0.844
(0.02,0.05]	0.925	0.928	0.927	0.926	0.924
(0.05,0.1]	0.958	0.961	0.96	0.96	0.959
(0.1,0.2]	0.966	0.971	0.969	0.969	0.968
(0.2,0.5]	0.973	0.976	0.975	0.974	0.973
(0.5,0.95]	0.972	0.975	0.974	0.973	0.973
(0.95,1]	0.834	0.836	0.838	0.837	0.852

(d) 1.0X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	HT	Illumina	Illumina
Cov	2.0	2.0	2.0	2.0	2.0
(0,0.0001]	0.059	0.064	0.056	0.06	0.113
(0.0001,0.0002]	0.269	0.269	0.274	0.275	0.307
(0.0002,0.0005]	0.431	0.431	0.437	0.437	0.471
(0.0005,0.001]	0.586	0.587	0.593	0.592	0.608
(0.001,0.002]	0.616	0.618	0.615	0.616	0.633
(0.002,0.005]	0.678	0.682	0.681	0.68	0.693
(0.005,0.01]	0.793	0.797	0.797	0.796	0.803
(0.01,0.02]	0.862	0.866	0.863	0.862	0.866
(0.02,0.05]	0.933	0.937	0.934	0.936	0.938
(0.05,0.1]	0.962	0.965	0.963	0.964	0.966
(0.1,0.2]	0.97	0.973	0.971	0.972	0.973
(0.2,0.5]	0.976	0.978	0.977	0.977	0.978
(0.5,0.95]	0.975	0.977	0.976	0.976	0.977
(0.95,1]	0.841	0.845	0.845	0.846	0.871

(e) 2.0X

Supplementary Table 4: Imputation accuracy for 5-Family samples methods, data and coverage

Method	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	Illumina
Cov	0.1	0.1	0.1
(0,0.0001]	0.161	0.136	0.088
(0.0001,0.0002]	0.242	0.206	0.125
(0.0002,0.0005]	0.304	0.26	0.16
(0.0005,0.001]	0.413	0.36	0.238
(0.001,0.002]	0.481	0.425	0.31
(0.002,0.005]	0.575	0.516	0.401
(0.005,0.01]	0.658	0.602	0.498
(0.01,0.02]	0.728	0.679	0.589
(0.02,0.05]	0.811	0.767	0.704
(0.05,0.1]	0.85	0.81	0.761
(0.1,0.2]	0.858	0.819	0.776
(0.2,0.5]	0.868	0.834	0.796
(0.5,0.95]	0.879	0.848	0.813
(0.95,1]	0.79	0.748	0.671

(a) 0.1X

Method	QUILT	QUILT	GLIMPSE
Data	HT	Illumina	Illumina
Cov	0.25	0.25	0.25
(0,0.0001]	0.223	0.212	0.185
(0.0001,0.0002]	0.337	0.312	0.259
(0.0002,0.0005]	0.425	0.393	0.322
(0.0005,0.001]	0.55	0.513	0.432
(0.001,0.002]	0.629	0.593	0.519
(0.002,0.005]	0.723	0.687	0.615
(0.005,0.01]	0.799	0.765	0.704
(0.01,0.02]	0.852	0.823	0.774
(0.02,0.05]	0.907	0.886	0.856
(0.05,0.1]	0.938	0.919	0.9
(0.1,0.2]	0.943	0.926	0.909
(0.2,0.5]	0.947	0.931	0.917
(0.5,0.95]	0.952	0.938	0.925
(0.95,1]	0.895	0.874	0.843

(b) 0.25X

Supplementary Table 5: Imputation accuracy for GBR samples with haplotagging data across methods, data and coverage

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	Illumina	ONT	Illumina	ONT	Illumina	ONT
Cov	0.1	0.1	0.1	0.1	0.1	0.1
(0,0.0001]	0.124	0.062	0.067	0.055	0.054	0.034
(0.0001,0.0002]	0.134	0.067	0.074	0.065	0.057	0.034
(0.0002,0.0005]	0.181	0.105	0.117	0.098	0.093	0.046
(0.0005,0.001]	0.263	0.147	0.178	0.142	0.16	0.104
(0.001,0.002]	0.347	0.166	0.233	0.158	0.207	0.119
(0.002,0.005]	0.465	0.266	0.307	0.249	0.29	0.187
(0.005,0.01]	0.541	0.298	0.361	0.281	0.344	0.224
(0.01,0.02]	0.596	0.322	0.412	0.309	0.402	0.247
(0.02,0.05]	0.643	0.371	0.468	0.36	0.46	0.29
(0.05,0.1]	0.662	0.39	0.489	0.382	0.481	0.318
(0.1,0.2]	0.68	0.401	0.507	0.397	0.499	0.326
(0.2,0.5]	0.715	0.454	0.551	0.443	0.548	0.372
(0.5,0.95]	0.737	0.488	0.58	0.477	0.577	0.403
(0.95,1]	0.626	0.39	0.448	0.366	0.458	0.305

(a) 0.1X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	Illumina	ONT	Illumina	ONT	Illumina	ONT
Cov	0.25	0.25	0.25	0.25	0.25	0.25
(0,0.0001]	0.201	0.126	0.16	0.126	0.129	0.066
(0.0001,0.0002]	0.219	0.154	0.155	0.149	0.12	0.075
(0.0002,0.0005]	0.318	0.226	0.238	0.218	0.179	0.111
(0.0005,0.001]	0.43	0.292	0.314	0.284	0.282	0.166
(0.001,0.002]	0.536	0.351	0.401	0.33	0.367	0.216
(0.002,0.005]	0.672	0.454	0.502	0.443	0.482	0.313
(0.005,0.01]	0.757	0.522	0.565	0.503	0.548	0.363
(0.01,0.02]	0.8	0.559	0.624	0.539	0.607	0.394
(0.02,0.05]	0.83	0.587	0.66	0.567	0.641	0.428
(0.05,0.1]	0.844	0.613	0.675	0.591	0.659	0.453
(0.1,0.2]	0.857	0.629	0.691	0.611	0.671	0.474
(0.2,0.5]	0.881	0.664	0.728	0.643	0.711	0.511
(0.5,0.95]	0.889	0.683	0.747	0.662	0.728	0.53
(0.95,1]	0.817	0.605	0.686	0.585	0.662	0.384

(b) 0.25X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	Illumina	ONT	Illumina	ONT	Illumina	ONT
Cov	0.5	0.5	0.5	0.5	0.5	0.5
(0,0.0001]	0.236	0.177	0.19	0.173	0.164	0.089
(0.0001,0.0002]	0.305	0.236	0.246	0.227	0.2	0.107
(0.0002,0.0005]	0.408	0.329	0.34	0.329	0.269	0.155
(0.0005,0.001]	0.516	0.425	0.42	0.426	0.37	0.255
(0.001,0.002]	0.636	0.516	0.529	0.504	0.487	0.342
(0.002,0.005]	0.767	0.62	0.648	0.602	0.617	0.441
(0.005,0.01]	0.849	0.699	0.717	0.678	0.698	0.509
(0.01,0.02]	0.885	0.729	0.756	0.705	0.734	0.527
(0.02,0.05]	0.912	0.759	0.787	0.737	0.768	0.557
(0.05,0.1]	0.924	0.773	0.807	0.748	0.785	0.576
(0.1,0.2]	0.932	0.79	0.82	0.765	0.797	0.584
(0.2,0.5]	0.945	0.809	0.839	0.783	0.822	0.616
(0.5,0.95]	0.948	0.823	0.853	0.802	0.837	0.628
(0.95,1]	0.892	0.744	0.774	0.718	0.761	0.545

(c) 0.5X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	Illumina	ONT	Illumina	ONT	Illumina	ONT
Cov	1.0	1.0	1.0	1.0	1.0	1.0
(0,0.0001]	0.297	0.216	0.272	0.216	0.224	0.109
(0.0001,0.0002]	0.364	0.294	0.337	0.312	0.262	0.125
(0.0002,0.0005]	0.474	0.414	0.452	0.434	0.357	0.209
(0.0005,0.001]	0.601	0.538	0.546	0.539	0.477	0.342
(0.001,0.002]	0.7	0.638	0.64	0.629	0.598	0.439
(0.002,0.005]	0.825	0.749	0.758	0.731	0.735	0.554
(0.005,0.01]	0.9	0.818	0.83	0.797	0.812	0.625
(0.01,0.02]	0.927	0.848	0.861	0.823	0.847	0.648
(0.02,0.05]	0.947	0.869	0.89	0.845	0.874	0.671
(0.05,0.1]	0.955	0.882	0.902	0.859	0.887	0.686
(0.1,0.2]	0.962	0.892	0.91	0.868	0.893	0.693
(0.2,0.5]	0.97	0.905	0.925	0.884	0.908	0.718
(0.5,0.95]	0.971	0.909	0.929	0.888	0.913	0.732
(0.95,1]	0.926	0.838	0.86	0.818	0.849	0.615

(d) 1.0X

Method	Optimal	Optimal	QUILT	QUILT	GLIMPSE	GLIMPSE
Data	Illumina	ONT	Illumina	ONT	Illumina	ONT
Cov	2.0	2.0	2.0	2.0	2.0	2.0
(0,0.0001]	0.336	0.249	0.322	0.254	0.268	0.14
(0.0001,0.0002]	0.436	0.358	0.417	0.366	0.319	0.173
(0.0002,0.0005]	0.56	0.512	0.537	0.528	0.443	0.298
(0.0005,0.001]	0.664	0.624	0.64	0.628	0.58	0.445
(0.001,0.002]	0.754	0.719	0.732	0.712	0.699	0.562
(0.002,0.005]	0.862	0.823	0.835	0.809	0.815	0.681
(0.005,0.01]	0.924	0.894	0.896	0.881	0.884	0.765
(0.01,0.02]	0.946	0.918	0.923	0.899	0.916	0.788
(0.02,0.05]	0.961	0.932	0.941	0.916	0.933	0.804
(0.05,0.1]	0.967	0.938	0.95	0.921	0.943	0.811
(0.1,0.2]	0.972	0.944	0.955	0.927	0.947	0.814
(0.2,0.5]	0.979	0.954	0.964	0.938	0.956	0.825
(0.5,0.95]	0.979	0.956	0.964	0.94	0.958	0.826
(0.95,1]	0.939	0.909	0.92	0.89	0.91	0.721

(e) 2.0X

Supplementary Table 6: Imputation accuracy for Shafin *et al* samples methods, data and coverage

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle 5.1	Beagle 5.1
Data	Illumina	Illumina	Illumina	Illumina	Illumina	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.16	0.203	0.223	0.259	0.288	0.203	0.183
(0.0001,0.0002]	0.127	0.172	0.204	0.244	0.277	0.161	0.13
(0.0002,0.0005]	0.199	0.267	0.341	0.374	0.44	0.204	0.173
(0.0005,0.001]	0.342	0.465	0.544	0.607	0.668	0.356	0.302
(0.001,0.002]	0.492	0.616	0.686	0.744	0.797	0.506	0.422
(0.002,0.005]	0.65	0.78	0.837	0.875	0.903	0.681	0.58
(0.005,0.01]	0.722	0.846	0.899	0.928	0.946	0.784	0.679
(0.01,0.02]	0.771	0.886	0.93	0.952	0.966	0.855	0.743
(0.02,0.05]	0.812	0.912	0.949	0.965	0.975	0.905	0.794
(0.05,0.1]	0.83	0.924	0.955	0.971	0.979	0.929	0.85
(0.1,0.2]	0.843	0.933	0.963	0.976	0.983	0.946	0.89
(0.2,0.5]	0.867	0.947	0.971	0.983	0.988	0.959	0.919
(0.5,0.95]	0.876	0.95	0.974	0.984	0.989	0.962	0.921
(0.95,1]	0.819	0.914	0.947	0.965	0.974	0.901	0.794

(a) ASW

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle 5.1	Beagle 5.1
Data	Illumina	Illumina	Illumina	Illumina	Illumina	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.108	0.169	0.221	0.276	0.373	0.179	0.145
(0.0001,0.0002]	0.21	0.307	0.389	0.467	0.548	0.313	0.275
(0.0002,0.0005]	0.284	0.418	0.481	0.565	0.648	0.383	0.363
(0.0005,0.001]	0.426	0.562	0.64	0.699	0.771	0.543	0.52
(0.001,0.002]	0.517	0.656	0.734	0.782	0.839	0.641	0.631
(0.002,0.005]	0.606	0.745	0.805	0.854	0.891	0.735	0.723
(0.005,0.01]	0.707	0.825	0.877	0.913	0.939	0.819	0.832
(0.01,0.02]	0.771	0.875	0.917	0.943	0.959	0.881	0.898
(0.02,0.05]	0.85	0.928	0.956	0.97	0.979	0.943	0.941
(0.05,0.1]	0.896	0.96	0.977	0.984	0.988	0.976	0.963
(0.1,0.2]	0.91	0.97	0.984	0.989	0.992	0.985	0.975
(0.2,0.5]	0.92	0.974	0.987	0.992	0.994	0.988	0.982
(0.5,0.95]	0.928	0.976	0.988	0.992	0.995	0.988	0.982
(0.95,1]	0.827	0.905	0.937	0.956	0.965	0.928	0.911

(b) CEU

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle 5.1	Beagle 5.1
Data	Illumina	Illumina	Illumina	Illumina	Illumina	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.093	0.164	0.224	0.307	0.402	0.111	0.119
(0.0001,0.0002]	0.158	0.273	0.378	0.49	0.601	0.163	0.171
(0.0002,0.0005]	0.211	0.361	0.481	0.6	0.707	0.23	0.271
(0.0005,0.001]	0.3	0.472	0.583	0.693	0.789	0.318	0.39
(0.001,0.002]	0.421	0.581	0.678	0.768	0.84	0.43	0.485
(0.002,0.005]	0.532	0.673	0.752	0.825	0.876	0.551	0.561
(0.005,0.01]	0.646	0.771	0.833	0.884	0.92	0.671	0.669
(0.01,0.02]	0.741	0.852	0.899	0.93	0.95	0.793	0.803
(0.02,0.05]	0.794	0.898	0.938	0.959	0.972	0.888	0.887
(0.05,0.1]	0.816	0.918	0.955	0.972	0.981	0.935	0.91
(0.1,0.2]	0.838	0.93	0.963	0.978	0.986	0.954	0.929
(0.2,0.5]	0.857	0.942	0.97	0.983	0.989	0.965	0.944
(0.5,0.95]	0.872	0.947	0.972	0.984	0.99	0.967	0.949
(0.95,1]	0.74	0.861	0.911	0.937	0.955	0.856	0.816

(c) CHB

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle 5.1	Beagle 5.1
Data	Illumina	Illumina	Illumina	Illumina	Illumina	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.287	0.395	0.475	0.545	0.622	0.353	0.333
(0.0001,0.0002]	0.3	0.433	0.519	0.61	0.69	0.358	0.331
(0.0002,0.0005]	0.357	0.502	0.594	0.682	0.764	0.397	0.367
(0.0005,0.001]	0.437	0.59	0.685	0.759	0.828	0.482	0.456
(0.001,0.002]	0.487	0.629	0.709	0.783	0.837	0.533	0.498
(0.002,0.005]	0.534	0.674	0.753	0.813	0.861	0.573	0.547
(0.005,0.01]	0.636	0.764	0.833	0.881	0.917	0.689	0.671
(0.01,0.02]	0.692	0.817	0.878	0.917	0.941	0.768	0.775
(0.02,0.05]	0.771	0.887	0.929	0.954	0.969	0.878	0.871
(0.05,0.1]	0.816	0.92	0.955	0.973	0.982	0.94	0.905
(0.1,0.2]	0.833	0.933	0.965	0.979	0.987	0.958	0.928
(0.2,0.5]	0.858	0.945	0.973	0.984	0.99	0.969	0.948
(0.5,0.95]	0.869	0.949	0.974	0.985	0.99	0.971	0.95
(0.95,1]	0.685	0.837	0.891	0.923	0.948	0.827	0.785

(d) PJL

Method	QUILT	QUILT	QUILT	QUILT	QUILT	Beagle 5.1	Beagle 5.1
Data	Illumina	Illumina	Illumina	Illumina	Illumina	UKBB	GSA
Cov	0.1	0.25	0.5	1.0	2.0		
(0,0.0001]	0.079	0.11	0.142	0.177	0.231	0.13	0.088
(0.0001,0.0002]	0.181	0.266	0.324	0.369	0.451	0.268	0.223
(0.0002,0.0005]	0.34	0.45	0.511	0.579	0.651	0.41	0.384
(0.0005,0.001]	0.415	0.545	0.627	0.689	0.746	0.48	0.456
(0.001,0.002]	0.5	0.629	0.707	0.766	0.814	0.565	0.52
(0.002,0.005]	0.613	0.741	0.808	0.854	0.888	0.67	0.619
(0.005,0.01]	0.709	0.826	0.878	0.913	0.937	0.781	0.735
(0.01,0.02]	0.766	0.874	0.919	0.943	0.96	0.851	0.81
(0.02,0.05]	0.822	0.915	0.947	0.964	0.975	0.912	0.87
(0.05,0.1]	0.859	0.942	0.966	0.977	0.983	0.954	0.921
(0.1,0.2]	0.88	0.955	0.975	0.984	0.989	0.971	0.948
(0.2,0.5]	0.897	0.963	0.981	0.988	0.992	0.979	0.963
(0.5,0.95]	0.906	0.966	0.983	0.989	0.993	0.98	0.965
(0.95,1]	0.792	0.9	0.931	0.953	0.964	0.897	0.829

(e) PUR

Supplementary Table 7: Imputation accuracy for 1000 Genomes populations across methods, data and coverage

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	rare	States	45	80	62.2	85.7	91.2	68.4	94.9	91.2	61.4	100
ASW	rare	Joint	45	80	62.2	85.7	91.2	70.2	95	94.7	89.5	94.1
ASW	common	States	75	82.7	64	93.8	96.8	55.6	97.1	90.5	58.7	97.3
ASW	common	Joint	75	82.7	64	93.8	98.4	63.5	97.5	95.2	93.7	96.6
ASW	all	States	120	81.7	63.3	90.8	94.2	61.7	95.9	90.8	60	98.6
ASW	all	Joint	120	81.7	63.3	90.8	95	66.7	96.2	95	91.7	95.5
CEU	rare	States	38	100	89.5	100	97.4	68.4	100	92.1	68.4	100
CEU	rare	Joint	38	100	89.5	100	97.4	68.4	100	97.4	84.2	100
CEU	common	States	160	97.5	82.5	100	98.8	76.2	99.2	99.4	67.5	100
CEU	common	Joint	160	97.5	82.5	100	100	73.8	100	100	83.8	100
CEU	all	States	198	98	83.8	100	98.5	74.7	99.3	98	67.7	100
CEU	all	Joint	198	98	83.8	100	99.5	72.7	100	99.5	83.8	100
CHB	rare	States	69	84.1	79.7	85.5	92.8	44.9	96.8	91.3	33.3	100
CHB	rare	Joint	69	84.1	79.7	85.5	92.8	40.6	96.4	95.7	66.7	97.8
CHB	common	States	137	89.8	86.9	92.4	95.6	38.7	100	97.1	48.9	100
CHB	common	Joint	137	89.8	86.9	92.4	96.4	43.8	100	98.5	73	100
CHB	all	States	206	87.9	84.5	90.2	94.7	40.8	98.8	95.1	43.7	100
CHB	all	Joint	206	87.9	84.5	90.2	95.1	42.7	98.9	97.6	70.9	99.3
PJL	rare	States	33	69.7	84.8	78.6	78.8	42.4	92.9	84.8	48.5	93.8
PJL	rare	Joint	33	69.7	84.8	78.6	78.8	48.5	93.8	87.9	84.8	92.9
PJL	common	States	159	93.1	86.8	97.8	98.1	70.4	99.1	98.7	70.4	99.1
PJL	common	Joint	159	93.1	86.8	97.8	98.1	71.7	99.1	99.4	93.1	99.3
PJL	all	States	192	89.1	86.5	94.6	94.8	65.6	98.4	96.4	66.7	98.4
PJL	all	Joint	192	89.1	86.5	94.6	94.8	67.7	98.5	97.4	91.7	98.3
PUR	rare	States	91	81.3	81.3	83.8	81.3	63.7	86.2	79.1	61.5	85.7
PUR	rare	Joint	91	81.3	81.3	83.8	83.5	68.1	88.7	91.2	91.2	91.6
PUR	common	States	117	91.5	65	96.1	98.3	70.1	100	97.4	63.2	100
PUR	common	Joint	117	91.5	65	96.1	99.1	71.8	100	97.4	84.6	100
PUR	all	States	208	87	72.1	90	90.9	67.3	94.3	89.4	62.5	93.8
PUR	all	Joint	208	87	72.1	90	92.3	70.2	95.2	94.7	87.5	96.2
ALL	rare	States	412	83.7	78.9	89.2	90.8	59.2	94.7	88.6	56.1	95.7
ALL	rare	Joint	412	83.7	78.9	89.2	91.7	60.9	95.2	94.7	84.7	95.7
ALL	common	States	512	93.8	79.5	96.8	97.7	64.1	99.4	98.6	63.1	99.7
ALL	common	Joint	512	93.8	79.5	96.8	98.2	65.8	99.7	98.8	84.2	99.8
ALL	all	States	924	89.3	79.2	93.4	94.6	61.9	97.4	94.2	60	98
ALL	all	Joint	924	89.3	79.2	93.4	95.3	63.6	97.8	97	84.4	97.9

(a) HLA-A

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	rare	States	62	87.1	69.4	97.7	82.4	35.3	100	91.2	58.8	100
ASW	rare	Joint	62	87.1	69.4	97.7	82.4	30.9	100	92.6	88.2	95
ASW	common	States	58	86.2	60.3	97.1	90.4	30.8	100	92.3	53.8	100
ASW	common	Joint	58	86.2	60.3	97.1	88.5	28.8	100	96.2	92.3	95.8
ASW	all	States	120	86.7	65	97.4	85.8	33.3	100	91.7	56.7	100
ASW	all	Joint	120	86.7	65	97.4	85	30	100	94.2	90	95.4
CEU	rare	States	61	85.2	77	100	90.2	54.1	100	88.5	45.9	96.4
CEU	rare	Joint	61	85.2	77	100	88.5	63.9	100	95.1	72.1	97.7
CEU	common	States	137	97.8	89.8	100	99.3	72.3	100	99.3	73	100
CEU	common	Joint	137	97.8	89.8	100	99.3	79.6	100	100	86.1	100
CEU	all	States	198	93.9	85.9	100	96.5	66.7	100	96	64.6	99.2
CEU	all	Joint	198	93.9	85.9	100	96	74.7	100	98.5	81.8	99.4
CHB	rare	States	119	84.9	72.3	91.9	86.6	31.1	94.6	81.5	37	95.5
CHB	rare	Joint	119	84.9	72.3	91.9	88.2	34.5	95.1	92.4	68.1	96.3
CHB	common	States	87	88.5	78.2	97.1	90.8	40.2	97.1	92	43.7	100
CHB	common	Joint	87	88.5	78.2	97.1	93.1	42.5	97.3	96.6	70.1	96.7
CHB	all	States	206	86.4	74.8	94.2	88.3	35	95.8	85.9	39.8	97.6
CHB	all	Joint	206	86.4	74.8	94.2	90.3	37.9	96.2	94.2	68.9	96.5
PJL	rare	States	80	78.8	58.8	87.2	80	47.5	94.7	78.8	32.5	92.3
PJL	rare	Joint	80	78.8	58.8	87.2	82.5	50	95	88.8	68.8	96.4
PJL	common	States	112	91.1	68.8	94.8	96.4	57.1	98.4	92	39.3	97.7
PJL	common	Joint	112	91.1	68.8	94.8	95.5	60.7	98.5	95.5	70.5	98.7
PJL	all	States	192	85.9	64.6	91.9	89.6	53.1	97.1	86.5	36.5	95.7
PJL	all	Joint	192	85.9	64.6	91.9	90.1	56.2	97.2	92.7	69.8	97.8
PUR	rare	States	130	73.8	63.1	85.4	76.9	30.8	100	80	40	92.3
PUR	rare	Joint	130	73.8	63.1	85.4	80	33.1	100	88.5	63.8	94
PUR	common	States	78	85.9	66.7	92.3	94.9	43.6	100	91	51.3	100
PUR	common	Joint	78	85.9	66.7	92.3	94.9	44.9	100	96.2	65.4	100
PUR	all	States	208	78.4	64.4	88.1	83.7	35.6	100	84.1	44.2	95.7
PUR	all	Joint	208	78.4	64.4	88.1	85.6	37.5	100	91.3	64.4	96.3
ALL	rare	States	639	82.3	66.5	92.2	84.7	39.1	97.6	84.4	41.5	96.2
ALL	rare	Joint	639	82.3	66.5	92.2	85.9	42.3	97.8	92	69.6	96
ALL	common	States	285	94.7	82.5	98.3	98.6	59.6	100	97.9	61.4	100
ALL	common	Joint	285	94.7	82.5	98.3	98.2	62.5	100	98.9	82.5	99.6
ALL	all	States	924	86.1	71.4	94.4	89	45.5	98.6	88.5	47.6	97.7
ALL	all	Joint	924	86.1	71.4	94.4	89.7	48.5	98.7	94.2	73.6	97.2

(b) HLA-B

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	rare	States	35	94.3	94.3	97	88.6	71.4	92	88.6	74.3	96.2
ASW	rare	Joint	35	94.3	94.3	97	91.4	74.3	92.3	97.1	100	97.1
ASW	common	States	87	95.4	97.7	95.3	97.7	70.1	98.4	97.7	89.7	98.7
ASW	common	Joint	87	95.4	97.7	95.3	96.6	69	98.3	98.9	97.7	98.8
ASW	all	States	122	95.1	96.7	95.8	95.1	70.5	96.5	95.1	85.2	98.1
ASW	all	Joint	122	95.1	96.7	95.8	95.1	70.5	96.5	98.4	98.4	98.3
CEU	rare	States	56	91.1	98.2	90.9	92.9	82.1	91.3	92.9	94.6	92.5
CEU	rare	Joint	56	91.1	98.2	90.9	92.9	80.4	91.1	94.6	96.4	94.4
CEU	common	States	142	99.3	99.3	99.3	98.6	84.5	100	99.3	97.9	100
CEU	common	Joint	142	99.3	99.3	99.3	99.3	85.2	100	99.3	100	99.3
CEU	all	States	198	97	99	96.9	97	83.8	97.6	97.5	97	97.9
CEU	all	Joint	198	97	99	96.9	97.5	83.8	97.6	98	99	98
CHB	rare	States	36	86.1	100	86.1	77.8	61.1	95.5	86.1	91.7	87.9
CHB	rare	Joint	36	86.1	100	86.1	80.6	63.9	95.7	88.9	94.4	91.2
CHB	common	States	170	88.2	84.7	95.8	97.6	69.4	100	98.2	78.2	99.2
CHB	common	Joint	170	88.2	84.7	95.8	98.2	72.4	100	99.4	92.9	100
CHB	all	States	206	87.9	87.4	93.9	94.2	68	99.3	96.1	80.6	97
CHB	all	Joint	206	87.9	87.4	93.9	95.1	70.9	99.3	97.6	93.2	98.4
PJL	rare	States	56	91.1	89.3	98	98.2	78.6	97.7	98.2	94.6	98.1
PJL	rare	Joint	56	91.1	89.3	98	98.2	82.1	97.8	98.2	92.9	98.1
PJL	common	States	136	99.3	98.5	99.3	99.3	89.7	100	99.3	94.9	99.2
PJL	common	Joint	136	99.3	98.5	99.3	99.3	89.7	100	99.3	97.1	100
PJL	all	States	192	96.9	95.8	98.9	99	86.5	99.4	99	94.8	98.9
PJL	all	Joint	192	96.9	95.8	98.9	99	87.5	99.4	99	95.8	99.5
PUR	rare	States	68	88.2	95.6	89.2	92.6	67.6	95.7	94.1	79.4	98.1
PUR	rare	Joint	68	88.2	95.6	89.2	94.1	75	94.1	97.1	91.2	96.8
PUR	common	States	140	99.3	97.9	100	98.6	75.7	100	97.9	88.6	99.2
PUR	common	Joint	140	99.3	97.9	100	99.3	79.3	100	99.3	97.1	99.3
PUR	all	States	208	95.7	97.1	96.5	96.6	73.1	98.7	96.6	85.6	98.9
PUR	all	Joint	208	95.7	97.1	96.5	97.6	77.9	98.1	98.6	95.2	98.5
ALL	rare	States	384	89.1	93.5	93.3	93.5	68.8	96.6	94.5	83.9	96.9
ALL	rare	Joint	384	89.1	93.5	93.3	94.3	71.9	96.4	96.9	93.8	97.2
ALL	common	States	542	98.2	96.1	98.7	98.5	82.3	99.6	98.7	92.3	99
ALL	common	Joint	542	98.2	96.1	98.7	98.9	83.4	99.6	99.3	97.8	99.4
ALL	all	States	926	94.4	95	96.5	96.4	76.7	98.5	97	88.8	98.2
ALL	all	Joint	926	94.4	95	96.5	97	78.6	98.4	98.3	96.1	98.5

(c) HLA-C

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	rare	States	22	81.8	68.2	86.7	77.3	18.2	100	86.4	54.5	100
ASW	rare	Joint	22	81.8	68.2	86.7	86.4	45.5	100	90.9	95.5	95.2
ASW	common	States	100	86	73	93.2	82	14	85.7	84	48	91.7
ASW	common	Joint	100	86	73	93.2	89	34	97.1	93	83	95.2
ASW	all	States	122	85.2	72.1	92	81.1	14.8	88.9	84.4	49.2	93.3
ASW	all	Joint	122	85.2	72.1	92	88.5	36.1	97.7	92.6	85.2	95.2
CEU	rare	States	17	100	94.1	100	94.1	29.4	100	100	64.7	100
CEU	rare	Joint	17	100	94.1	100	100	29.4	100	100	100	100
CEU	common	States	169	87.6	67.5	96.5	92.9	31.4	98.1	94.1	45.6	98.7
CEU	common	Joint	169	87.6	67.5	96.5	95.3	49.1	100	99.4	100	99.4
CEU	all	States	186	88.7	69.9	96.9	93	31.2	98.3	94.6	47.3	98.9
CEU	all	Joint	186	88.7	69.9	96.9	95.7	47.3	100	99.5	100	99.5
CHB	rare	States	9	55.6	44.4	75	87.5	25	100	92.9	55.4	100
CHB	rare	Joint	9	55.6	44.4	75	92.9	48.2	100	96.4	94.6	98.1
CHB	common	States	81	97.5	71.6	98.3	97.1	17.6	100	100	38.2	100
CHB	common	Joint	81	97.5	71.6	98.3	97.1	32.4	100	100	97.1	100
CHB	all	States	90	93.3	68.9	96.8	91.1	22.2	100	95.6	48.9	100
CHB	all	Joint	90	93.3	68.9	96.8	94.4	42.2	100	97.8	95.6	98.8
PJL	rare	States	35	94.3	85.7	100	88.6	31.4	100	94.3	74.3	100
PJL	rare	Joint	35	94.3	85.7	100	94.3	62.9	95.5	97.1	97.1	100
PJL	common	States	157	86.6	58.6	93.5	93	23.6	94.6	95.5	61.1	99
PJL	common	Joint	157	86.6	58.6	93.5	95.5	54.8	98.8	98.7	98.1	98.7
PJL	all	States	192	88	63.5	95.1	92.2	25	95.8	95.3	63.5	99.2
PJL	all	Joint	192	88	63.5	95.1	95.3	56.2	98.1	98.4	97.9	98.9
PUR	rare	States	21	85.7	57.1	91.7	85.4	19.5	87.5	87.8	51.2	85.7
PUR	rare	Joint	21	85.7	57.1	91.7	95.1	39	93.8	90.2	95.1	94.9
PUR	common	States	147	86.4	58.5	95.3	92.1	15.7	95	89.8	40.2	100
PUR	common	Joint	147	86.4	58.5	95.3	92.1	29.9	97.4	99.2	93.7	99.2
PUR	all	States	168	86.3	58.3	94.9	90.5	16.7	92.9	89.3	42.9	95.8
PUR	all	Joint	168	86.3	58.3	94.9	92.9	32.1	96.3	97	94	98.1
ALL	rare	States	201	81.6	63.7	92.2	85.1	17.4	97.1	86.1	53.2	93.5
ALL	rare	Joint	201	81.6	63.7	92.2	89.6	42.8	97.7	93	90.5	96.7
ALL	common	States	557	90.3	66.8	96.2	91.9	24.6	95.6	94.3	50.1	99.3
ALL	common	Joint	557	90.3	66.8	96.2	95.2	44.2	98.8	98.9	96.9	98.9
ALL	all	States	758	88	66	95.2	90.1	22.7	95.9	92.1	50.9	97.7
ALL	all	Joint	758	88	66	95.2	93.7	43.8	98.5	97.4	95.3	98.3

(d) HLA-DQB1

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	rare	States	76	65.8	57.9	72.7	64.5	5.3	100	51.3	0	NaN
ASW	rare	Joint	76	65.8	57.9	72.7	64.5	7.9	100	75	56.6	88.4
ASW	common	States	46	93.5	65.2	96.7	71.7	0	NaN	69.6	4.3	100
ASW	common	Joint	46	93.5	65.2	96.7	67.4	8.7	50	69.6	63	69
ASW	all	States	122	76.2	60.7	82.4	67.2	3.3	100	58.2	1.6	100
ASW	all	Joint	122	76.2	60.7	82.4	65.6	8.2	80	73	59	80.6
CEU	rare	States	43	67.4	32.6	78.6	55.8	0	NaN	34.9	7	33.3
CEU	rare	Joint	43	67.4	32.6	78.6	48.8	9.3	0	55.8	51.2	68.2
CEU	common	States	155	83.9	46.5	97.2	81.9	16.8	100	67.7	12.3	100
CEU	common	Joint	155	83.9	46.5	97.2	82.6	21.9	100	83.2	55.5	96.5
CEU	all	States	198	80.3	43.4	94.2	76.3	13.1	100	60.6	11.1	90.9
CEU	all	Joint	198	80.3	43.4	94.2	75.3	19.2	89.5	77.3	54.5	90.7
CHB	rare	States	76	63.2	46.1	82.9	68.4	9.2	100	56.6	9.2	100
CHB	rare	Joint	76	63.2	46.1	82.9	65.8	11.8	88.9	73.7	47.4	83.3
CHB	common	States	130	85.4	54.6	94.4	86.9	11.5	100	63.8	3.8	100
CHB	common	Joint	130	85.4	54.6	94.4	91.5	17.7	100	97.7	73.8	99
CHB	all	States	206	77.2	51.5	90.6	80.1	10.7	100	61.2	5.8	100
CHB	all	Joint	206	77.2	51.5	90.6	82	15.5	96.9	88.8	64.1	94.7
PJL	rare	States	45	68.9	55.6	72	68.9	4.4	100	60	2.2	100
PJL	rare	Joint	45	68.9	55.6	72	64.4	8.9	100	68.9	44.4	80
PJL	common	States	147	91.2	70.1	96.1	91.2	13.6	100	76.2	11.6	100
PJL	common	Joint	147	91.2	70.1	96.1	95.2	24.5	100	94.6	73.5	99.1
PJL	all	States	192	85.9	66.7	91.4	85.9	11.5	100	72.4	9.4	100
PJL	all	Joint	192	85.9	66.7	91.4	88	20.8	100	88.5	66.7	96.1
PUR	rare	States	107	65.4	29.9	75	60.7	12.1	100	44.9	1.9	100
PUR	rare	Joint	107	65.4	29.9	75	64.5	15	93.8	70.1	51.4	90.9
PUR	common	States	101	76.2	31.7	90.6	73.3	3	66.7	73.3	0	NaN
PUR	common	Joint	101	76.2	31.7	90.6	78.2	7.9	100	87.1	62.4	93.7
PUR	all	States	208	70.7	30.8	82.8	66.8	7.7	93.8	58.7	1	100
PUR	all	Joint	208	70.7	30.8	82.8	71.2	11.5	95.8	78.4	56.7	92.4
ALL	rare	States	631	70.2	44.5	83.6	67.4	6.7	97.6	53.1	4	92
ALL	rare	Joint	631	70.2	44.5	83.6	68.3	11.7	89.2	74.6	54	87.4
ALL	common	States	295	94.9	60	97.7	93.9	16.3	100	82.4	10.5	100
ALL	common	Joint	295	94.9	60	97.7	96.3	23.7	100	97.3	73.6	99.1
ALL	all	States	926	78.1	49.5	89.1	75.8	9.7	98.9	62.4	6	96.4
ALL	all	Joint	926	78.1	49.5	89.1	77.2	15.6	94.4	81.9	60.3	91.9

(e) HLA-DRB1

Supplementary Table 8: HLA imputation accuracy for 1000 Genomes populations across populations, methods, and coverage Qmethod refers to whether QUILT results are just using results from a labelled haplotype reference panel (States), or jointly using the labelled reference panel and read mapping (Joint). Array results are unaffected by this setting and are listed twice for simplicity. Population ALL refers to running all individuals as if they were from a single population. freq referring to frequency refers to whether analyses are restricted to alleles that are rare (less than or equal to 5 % frequency in the truth samples for that population) or common. Across the 9 accuracy columns, the prefix A, L and H refers to A for the array based method (SNP2HLA style using Beagle), and otherwise L for low (0.1X) and H for high (2.0X) coverage sequence using QUILT-HLA. For the suffix, accA is accuracy across all samples regardless of confidence (i.e. always taking the most likely called HLA alleles), while perT is percent of samples that meet the threshold (0.90), and accT is accuracy only on the high confidence HLA alleles.

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	27:05	States	2	100	100	100	100	100	100	100	100	100
ASW	27:05	Joint	2	100	100	100	100	100	100	100	100	100
CEU	27:05	States	6	83.3	66.7	100	100	66.7	100	100	50	100
CEU	27:05	Joint	6	83.3	66.7	100	100	66.7	100	100	83.3	100
CHB	27:05	States	1	100	100	100	100	100	100	100	0	NaN
CHB	27:05	Joint	1	100	100	100	100	100	100	100	100	100
PJL	27:05	States	3	100	33.3	100	100	66.7	100	100	33.3	100
PJL	27:05	Joint	3	100	33.3	100	100	66.7	100	100	66.7	100
PUR	27:05	States	1	100	100	100	100	100	100	100	100	100
PUR	27:05	Joint	1	100	100	100	100	100	100	100	100	100
ALL	27:05	States	13	92.3	69.2	100	100	76.9	100	100	53.8	100
ALL	27:05	Joint	13	92.3	69.2	100	100	76.9	100	100	84.6	100

(a) HLA-B, 27:05

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	57:01	States	0									
ASW	57:01	Joint	0									
CEU	57:01	States	13	100	84.6	100	100	69.2	100	100	61.5	100
CEU	57:01	Joint	13	100	84.6	100	100	92.3	100	100	92.3	100
CHB	57:01	States	1	100	100	100	100	0	NaN	100	100	100
CHB	57:01	Joint	1	100	100	100	100	0	NaN	100	100	100
PJL	57:01	States	8	100	75	100	100	75	100	100	50	100
PJL	57:01	Joint	8	100	75	100	100	87.5	100	100	87.5	100
PUR	57:01	States	5	100	80	100	100	40	100	100	40	100
PUR	57:01	Joint	5	100	80	100	100	40	100	100	100	100
ALL	57:01	States	27	100	81.5	100	100	63	100	100	55.6	100
ALL	57:01	Joint	27	100	81.5	100	100	77.8	100	100	92.6	100

(b) HLA-B, 57:01

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	58:01	States	8	87.5	62.5	100	75	12.5	100	100	50	100
ASW	58:01	Joint	8	87.5	62.5	100	75	12.5	100	100	100	100
CEU	58:01	States	0									
CEU	58:01	Joint	0									
CHB	58:01	States	16	93.8	81.2	100	93.8	62.5	100	87.5	62.5	100
CHB	58:01	Joint	16	93.8	81.2	100	100	62.5	100	100	68.8	100
PJL	58:01	States	5	100	40	100	100	60	100	100	0	NaN
PJL	58:01	Joint	5	100	40	100	100	60	100	100	80	100
PUR	58:01	States	2	100	100	100	50	50	100	100	50	100
PUR	58:01	Joint	2	100	100	100	100	50	100	100	100	100
ALL	58:01	States	31	93.5	71	100	87.1	48.4	100	93.5	48.4	100
ALL	58:01	Joint	31	93.5	71	100	93.5	48.4	100	100	80.6	100

(c) HLA-B, 58:01

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	02:01	States	6	83.3	50	100	83.3	50	100	100	50	100
ASW	02:01	Joint	6	83.3	50	100	100	83.3	100	100	100	100
CEU	02:01	States	17	94.1	76.5	100	100	23.5	100	100	35.3	100
CEU	02:01	Joint	17	94.1	76.5	100	94.1	52.9	100	100	100	100
CHB	02:01	States	5	100	60	100	100	0	NaN	100	40	100
CHB	02:01	Joint	5	100	60	100	100	40	100	100	100	100
PJL	02:01	States	36	86.1	69.4	92	94.4	13.9	100	97.2	63.9	100
PJL	02:01	Joint	36	86.1	69.4	92	97.2	30.6	100	100	97.2	100
PUR	02:01	States	11	81.8	45.5	100	100	9.1	100	90.9	45.5	100
PUR	02:01	Joint	11	81.8	45.5	100	100	36.4	100	100	100	100
ALL	02:01	States	75	88	65.3	95.9	96	17.3	100	97.3	52	100
ALL	02:01	Joint	75	88	65.3	95.9	97.3	41.3	100	100	98.7	100

(d) HLA-DQB1, 02:01

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	03:02	States	4	75	25	100	100	0	NaN	75	0	NaN
ASW	03:02	Joint	4	75	25	100	100	0	NaN	100	25	100
CEU	03:02	States	0									
CEU	03:02	Joint	0									
CHB	03:02	States	0									
CHB	03:02	Joint	0									
PJL	03:02	States	0									
PJL	03:02	Joint	0									
PUR	03:02	States	8	50	25	100	87.5	25	100	50	0	NaN
PUR	03:02	Joint	8	50	25	100	87.5	25	100	87.5	62.5	100
ALL	03:02	States	12	58.3	25	100	91.7	16.7	100	58.3	0	NaN
ALL	03:02	Joint	12	58.3	25	100	91.7	16.7	100	91.7	50	100

(e) HLA-DRB1, 03:02

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	04:01	States	5	80	40	100	40	0	NaN	0	0	NaN
ASW	04:01	Joint	5	80	40	100	60	0	NaN	60	40	100
CEU	04:01	States	25	88	52	100	52	0	NaN	20	0	NaN
CEU	04:01	Joint	25	88	52	100	48	0	NaN	52	16	100
CHB	04:01	States	3	33.3	0	NaN	0	0	NaN	33.3	0	NaN
CHB	04:01	Joint	3	33.3	0	NaN	0	0	NaN	66.7	0	NaN
PJL	04:01	States	1	0	100	0	0	0	NaN	0	0	NaN
PJL	04:01	Joint	1	0	100	0	0	0	NaN	0	0	NaN
PUR	04:01	States	4	75	50	100	25	0	NaN	0	0	NaN
PUR	04:01	Joint	4	75	50	100	50	0	NaN	50	25	100
ALL	04:01	States	38	78.9	47.4	94.4	42.1	0	NaN	15.8	0	NaN
ALL	04:01	Joint	38	78.9	47.4	94.4	44.7	0	NaN	52.6	18.4	100

(f) HLA-DRB1, 04:01

pop	freq	Qmethod	N	AaccA	AperT	AaccT	LaccA	LperT	LaccT	HaccA	HperT	HaccT
ASW	08:01	States	0									
ASW	08:01	Joint	0									
CEU	08:01	States	5	100	40	100	80	0	NaN	80	0	NaN
CEU	08:01	Joint	5	100	40	100	20	20	0	0	80	0
CHB	08:01	States	1	0	0	NaN	100	0	NaN	100	0	NaN
CHB	08:01	Joint	1	0	0	NaN	100	0	NaN	0	0	NaN
PJL	08:01	States	1	100	100	100	100	0	NaN	0	0	NaN
PJL	08:01	Joint	1	100	100	100	0	0	NaN	0	0	NaN
PUR	08:01	States	2	100	100	100	50	0	NaN	100	0	NaN
PUR	08:01	Joint	2	100	100	100	50	50	0	0	100	0
ALL	08:01	States	9	88.9	55.6	100	77.8	0	NaN	77.8	0	NaN
ALL	08:01	Joint	9	88.9	55.6	100	33.3	22.2	0	0	66.7	0

(g) HLA-DRB1, 08:01

Supplementary Table 9: HLA imputation accuracy for 1000 Genomes populations for specific alleles with medical importance Format is the same as the previous Supplementary Table, except that freq here is just the specific allele under interrogation, and rows for which there are no truth values are listed as blank. Alleles are selected as per Karnes *et al.* [8]

	Pop	Cov	Per-X-cost	Pheno-cost	GR	Gimp	Ggsa	r2_imp	r2_gsa
1	CHB	0.1	26.458	5	3.527	33311.535	9444.559	0.857	0.944
2	CHB	0.1	26.458	50	1.344	12696.680	9444.559	0.857	0.944
3	CHB	0.1	13.229	5	4.134	39047.455	9444.559	0.857	0.944
4	CHB	0.1	13.229	50	1.378	13015.533	9444.559	0.857	0.944
5	CHB	0.1	6.615	5	4.524	42725.404	9444.559	0.857	0.944
6	CHB	0.1	6.615	50	1.396	13180.959	9444.559	0.857	0.944
7	CHB	0.25	26.458	5	2.690	25410.385	9444.559	0.942	0.944
8	CHB	0.25	26.458	50	1.376	12998.143	9444.559	0.942	0.944
9	CHB	0.25	13.229	5	3.611	34103.758	9444.559	0.942	0.944
10	CHB	0.25	13.229	50	1.460	13784.682	9444.559	0.942	0.944
11	CHB	0.25	6.615	5	4.356	41141.161	9444.559	0.942	0.944
12	CHB	0.25	6.615	50	1.505	14214.217	9444.559	0.942	0.944
13	CHB	0.5	26.458	5	1.835	17329.334	9444.559	0.970	0.944
14	CHB	0.5	26.458	50	1.272	12013.272	9444.559	0.970	0.944
15	CHB	0.5	13.229	5	2.770	26164.219	9444.559	0.970	0.944
16	CHB	0.5	13.229	50	1.417	13383.751	9444.559	0.970	0.944
17	CHB	0.5	6.615	5	3.718	35115.494	9444.559	0.970	0.944
18	CHB	0.5	6.615	50	1.503	14193.623	9444.559	0.970	0.944
19	CHB	1.0	26.458	5	1.110	10486.151	9444.559	0.983	0.944
20	CHB	1.0	26.458	50	1.070	10107.608	9444.559	0.983	0.944
21	CHB	1.0	13.229	5	1.860	17567.376	9444.559	0.983	0.944
22	CHB	1.0	13.229	50	1.289	12178.291	9444.559	0.983	0.944
23	CHB	1.0	6.615	5	2.808	26523.621	9444.559	0.983	0.944
24	CHB	1.0	6.615	50	1.437	13567.595	9444.559	0.983	0.944
25	CHB	2.0	26.458	5	0.619	5842.710	9444.559	0.989	0.944
26	CHB	2.0	26.458	50	0.804	7591.070	9444.559	0.989	0.944
27	CHB	2.0	13.229	5	1.117	10552.498	9444.559	0.989	0.944
28	CHB	2.0	13.229	50	1.077	10171.559	9444.559	0.989	0.944
29	CHB	2.0	6.615	5	1.872	17678.527	9444.559	0.989	0.944
30	CHB	2.0	6.615	50	1.298	12255.344	9444.559	0.989	0.944

(a) Common variant GWAS-type analysis

	Pop	Cov	Per-X-cost	Pheno-cost	GR	Gimp	Ggsa	r2_imp	r2_gsa
1	CHB	0.1	26.458	5	3.377	16374.772	4848.488	0.421	0.485
2	CHB	0.1	26.458	50	1.287	6241.239	4848.488	0.421	0.485
3	CHB	0.1	13.229	5	3.959	19194.348	4848.488	0.421	0.485
4	CHB	0.1	13.229	50	1.320	6397.976	4848.488	0.421	0.485
5	CHB	0.1	6.615	5	4.332	21002.297	4848.488	0.421	0.485
6	CHB	0.1	6.615	50	1.336	6479.293	4848.488	0.421	0.485
7	CHB	0.25	26.458	5	3.232	15671.161	4848.488	0.581	0.485
8	CHB	0.25	26.458	50	1.653	8016.249	4848.488	0.581	0.485
9	CHB	0.25	13.229	5	4.338	21032.561	4848.488	0.581	0.485
10	CHB	0.25	13.229	50	1.753	8501.326	4848.488	0.581	0.485
11	CHB	0.25	6.615	5	5.233	25372.687	4848.488	0.581	0.485
12	CHB	0.25	6.615	50	1.808	8766.230	4848.488	0.581	0.485
13	CHB	0.5	26.458	5	2.500	12120.865	4848.488	0.678	0.485
14	CHB	0.5	26.458	50	1.733	8402.588	4848.488	0.678	0.485
15	CHB	0.5	13.229	5	3.774	18300.355	4848.488	0.678	0.485
16	CHB	0.5	13.229	50	1.931	9361.158	4848.488	0.678	0.485
17	CHB	0.5	6.615	5	5.066	24561.253	4848.488	0.678	0.485
18	CHB	0.5	6.615	50	2.048	9927.617	4848.488	0.678	0.485
19	CHB	1.0	26.458	5	1.689	8186.860	4848.488	0.768	0.485
20	CHB	1.0	26.458	50	1.628	7891.320	4848.488	0.768	0.485
21	CHB	1.0	13.229	5	2.829	13715.390	4848.488	0.768	0.485
22	CHB	1.0	13.229	50	1.961	9507.965	4848.488	0.768	0.485
23	CHB	1.0	6.615	5	4.271	20707.805	4848.488	0.768	0.485
24	CHB	1.0	6.615	50	2.185	10592.638	4848.488	0.768	0.485
25	CHB	2.0	26.458	5	1.022	4957.452	4848.488	0.840	0.485
26	CHB	2.0	26.458	50	1.328	6440.910	4848.488	0.840	0.485
27	CHB	2.0	13.229	5	1.847	8953.637	4848.488	0.840	0.485
28	CHB	2.0	13.229	50	1.780	8630.416	4848.488	0.840	0.485
29	CHB	2.0	6.615	5	3.094	14999.966	4848.488	0.840	0.485
30	CHB	2.0	6.615	50	2.145	10398.476	4848.488	0.840	0.485

(b) Rare variant GWAS-type analysis

	Pop	Cov	Per-X-cost	Pheno-cost	BR	Bimp	Bgsa	r2_imp	r2_gsa
1	CHB	0.1	26.458	5	1.071	1.000	0.933	0.857	0.944
2	CHB	0.1	26.458	50	1.059	0.990	0.935	0.857	0.944
3	CHB	0.1	13.229	5	1.069	1.000	0.936	0.857	0.944
4	CHB	0.1	13.229	50	1.063	0.994	0.935	0.857	0.944
5	CHB	0.1	6.615	5	1.066	1.000	0.938	0.857	0.944
6	CHB	0.1	6.615	50	1.061	0.994	0.937	0.857	0.944
7	CHB	0.25	26.458	5	1.070	1.000	0.934	0.942	0.944
8	CHB	0.25	26.458	50	1.068	0.997	0.933	0.942	0.944
9	CHB	0.25	13.229	5	1.069	1.000	0.936	0.942	0.944
10	CHB	0.25	13.229	50	1.065	0.998	0.937	0.942	0.944
11	CHB	0.25	6.615	5	1.065	1.000	0.939	0.942	0.944
12	CHB	0.25	6.615	50	1.066	0.999	0.937	0.942	0.944
13	CHB	0.5	26.458	5	1.067	1.000	0.938	0.970	0.944
14	CHB	0.5	26.458	50	1.065	0.993	0.933	0.970	0.944
15	CHB	0.5	13.229	5	1.067	1.000	0.937	0.970	0.944
16	CHB	0.5	13.229	50	1.065	0.998	0.936	0.970	0.944
17	CHB	0.5	6.615	5	1.067	1.000	0.937	0.970	0.944
18	CHB	0.5	6.615	50	1.070	1.000	0.934	0.970	0.944
19	CHB	1.0	26.458	5	1.046	0.978	0.935	0.983	0.944
20	CHB	1.0	26.458	50	1.034	0.968	0.936	0.983	0.944
21	CHB	1.0	13.229	5	1.075	1.000	0.930	0.983	0.944
22	CHB	1.0	13.229	50	1.060	0.995	0.939	0.983	0.944
23	CHB	1.0	6.615	5	1.072	1.000	0.933	0.983	0.944
24	CHB	1.0	6.615	50	1.063	0.999	0.940	0.983	0.944
25	CHB	2.0	26.458	5	0.575	0.537	0.934	0.989	0.944
26	CHB	2.0	26.458	50	0.872	0.817	0.936	0.989	0.944
27	CHB	2.0	13.229	5	1.044	0.981	0.940	0.989	0.944
28	CHB	2.0	13.229	50	1.041	0.973	0.935	0.989	0.944
29	CHB	2.0	6.615	5	1.068	1.000	0.936	0.989	0.944
30	CHB	2.0	6.615	50	1.070	0.996	0.931	0.989	0.944

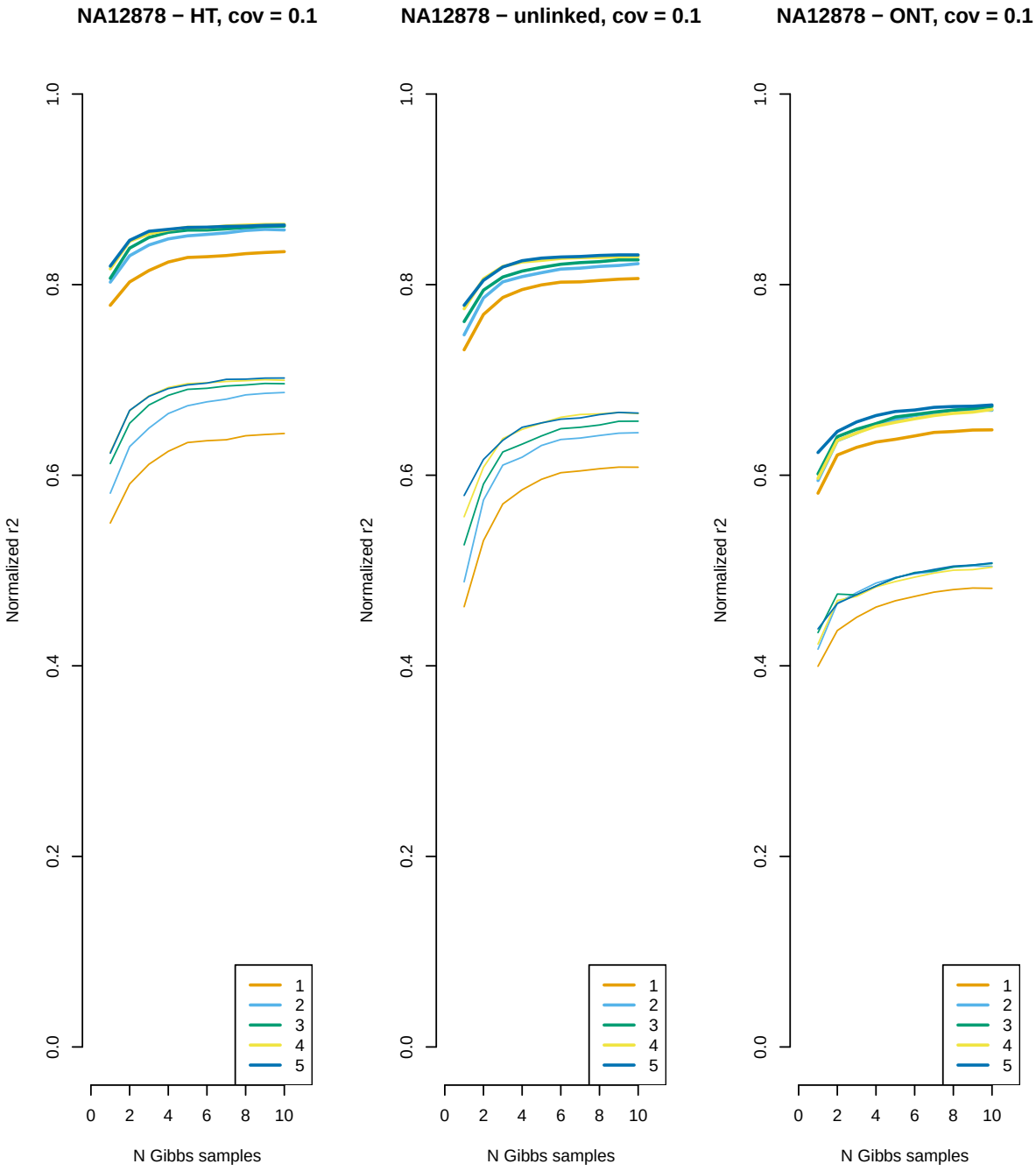
(c) Common variant burden-type analysis

	Pop	Cov	Per-X-cost	Pheno-cost	BR	Bimp	Bgsa	r2_imp	r2_gsa
1	CHB	0.1	26.458	5	6.307	0.990	0.157	0.421	0.485
2	CHB	0.1	26.458	50	1.885	0.318	0.169	0.421	0.485
3	CHB	0.1	13.229	5	6.060	0.998	0.165	0.421	0.485
4	CHB	0.1	13.229	50	1.987	0.328	0.165	0.421	0.485
5	CHB	0.1	6.615	5	6.053	1.000	0.165	0.421	0.485
6	CHB	0.1	6.615	50	2.065	0.343	0.166	0.421	0.485
7	CHB	0.25	26.458	5	6.179	0.990	0.160	0.581	0.485
8	CHB	0.25	26.458	50	3.928	0.631	0.161	0.581	0.485
9	CHB	0.25	13.229	5	6.015	1.000	0.166	0.581	0.485
10	CHB	0.25	13.229	50	4.294	0.692	0.161	0.581	0.485
11	CHB	0.25	6.615	5	6.321	1.000	0.158	0.581	0.485
12	CHB	0.25	6.615	50	4.361	0.716	0.164	0.581	0.485
13	CHB	0.5	26.458	5	6.105	0.964	0.158	0.678	0.485
14	CHB	0.5	26.458	50	4.269	0.726	0.170	0.678	0.485
15	CHB	0.5	13.229	5	5.823	1.000	0.172	0.678	0.485
16	CHB	0.5	13.229	50	5.164	0.829	0.160	0.678	0.485
17	CHB	0.5	6.615	5	5.914	1.000	0.169	0.678	0.485
18	CHB	0.5	6.615	50	5.156	0.874	0.170	0.678	0.485
19	CHB	1.0	26.458	5	4.713	0.757	0.161	0.768	0.485
20	CHB	1.0	26.458	50	4.400	0.719	0.163	0.768	0.485
21	CHB	1.0	13.229	5	6.298	0.993	0.158	0.768	0.485
22	CHB	1.0	13.229	50	5.333	0.870	0.163	0.768	0.485
23	CHB	1.0	6.615	5	6.325	1.000	0.158	0.768	0.485
24	CHB	1.0	6.615	50	5.556	0.931	0.168	0.768	0.485
25	CHB	2.0	26.458	5	1.658	0.279	0.168	0.840	0.485
26	CHB	2.0	26.458	50	3.348	0.546	0.163	0.840	0.485
27	CHB	2.0	13.229	5	5.458	0.863	0.158	0.840	0.485
28	CHB	2.0	13.229	50	5.083	0.829	0.163	0.840	0.485
29	CHB	2.0	6.615	5	6.248	0.998	0.160	0.840	0.485
30	CHB	2.0	6.615	50	5.769	0.946	0.164	0.840	0.485

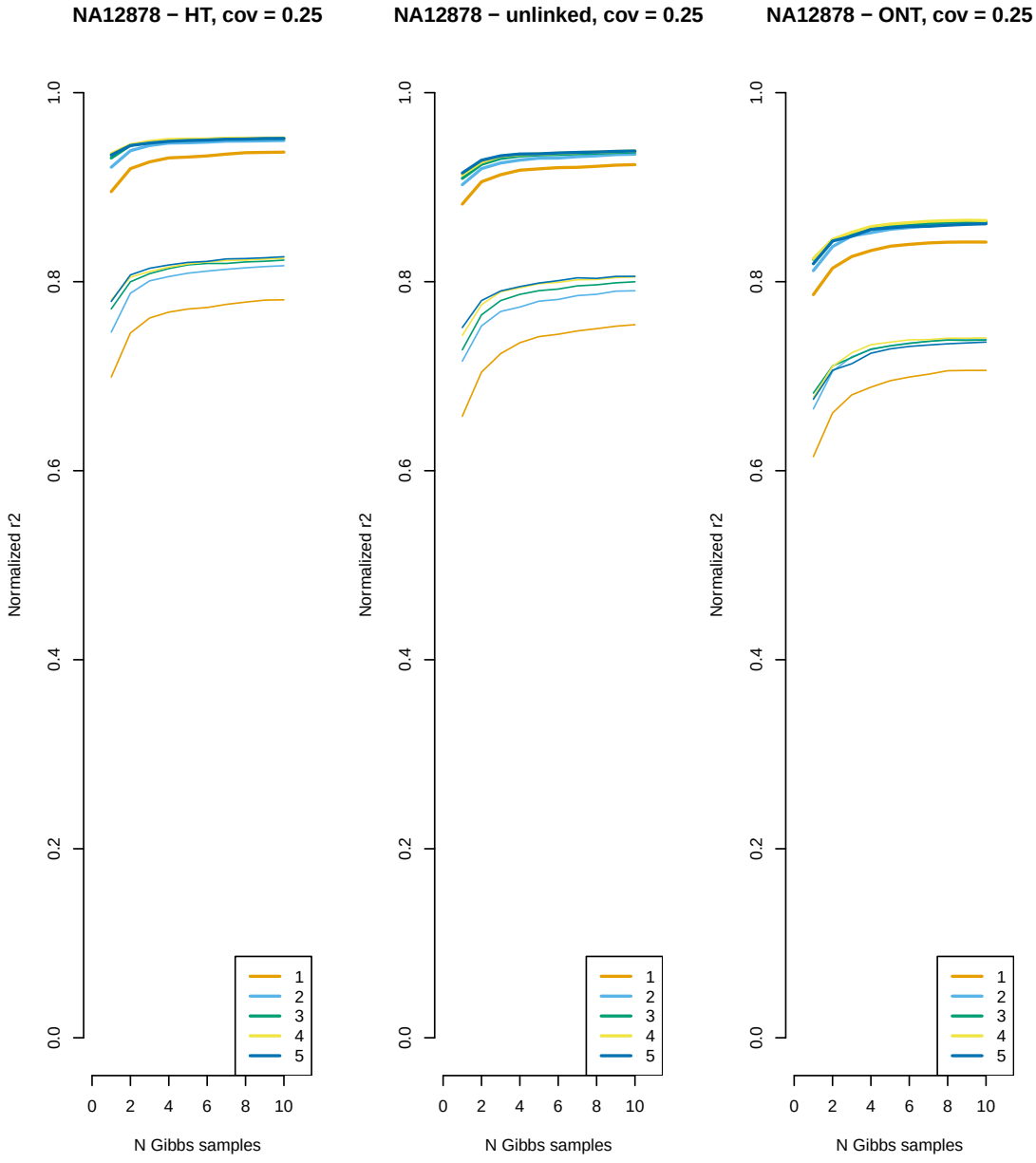
(d) Rare variant burden-type analysis

Supplementary Table 10: Relative effective sample size and power of lc-WGS and QUILT versus genotyping microarrays For the CHB population, for imputation accuracy for either rare (0.1-0.2 percent) or common (20-50 percent) SNPs, ratio of effective sample sizes and powers for GWAS-type and burden-type analyses are given. Per-X costs are converted from USD to GBP assuming whole genome costs of 1000/30X, 500/30X and 250/30X. Columns Gimp and Ggsa give the GWAS effective sample size for a given imputation accuracy and cost for lc-WGS and QUILT (imp) and array (gsa), and GR gives their ratio. Similarly the columns Bimp and Bgsa give the powers for burden test analyses, also for imputation and array (GSA), and BR gives their ratio.

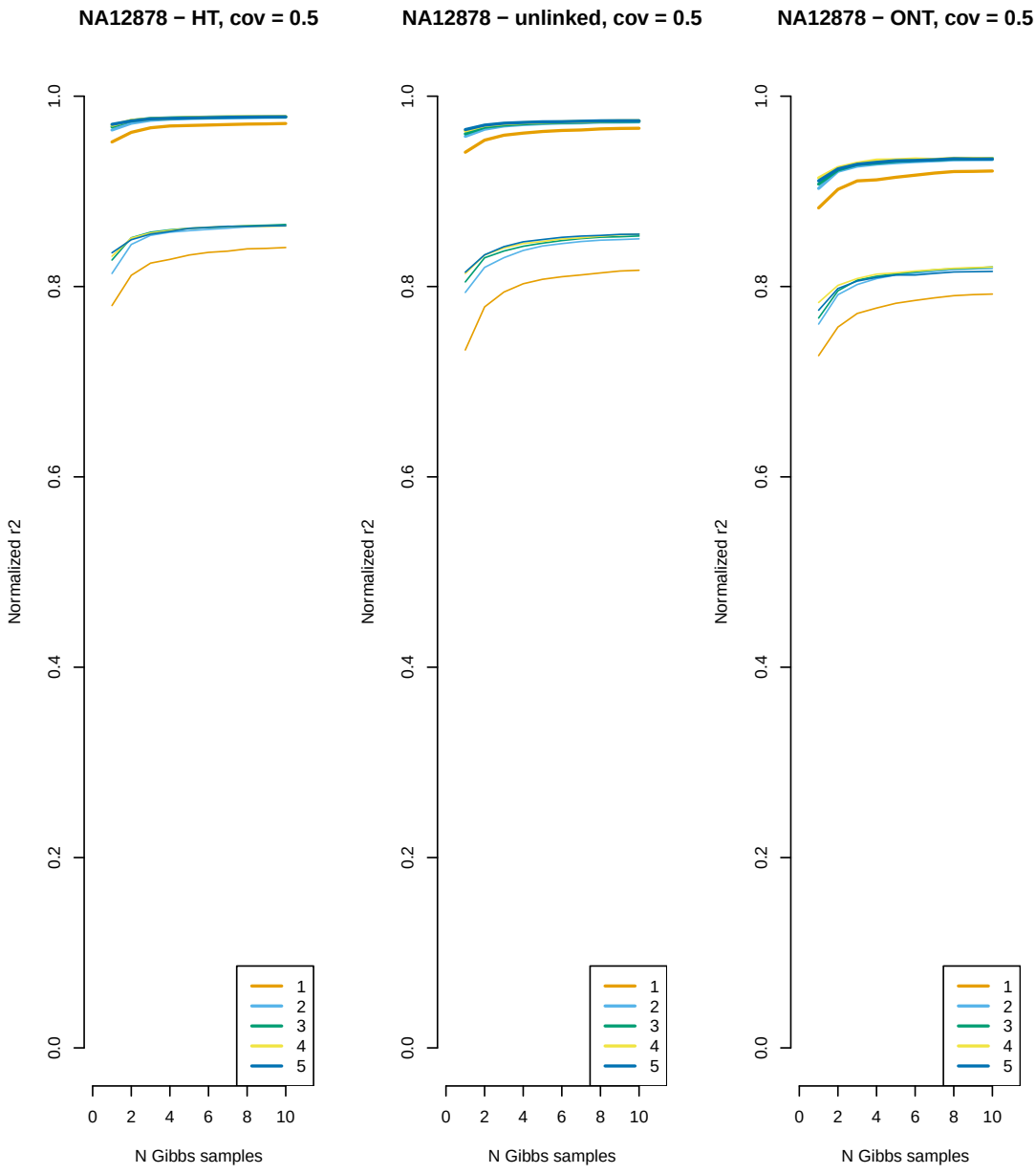
3 Supplementary Figures



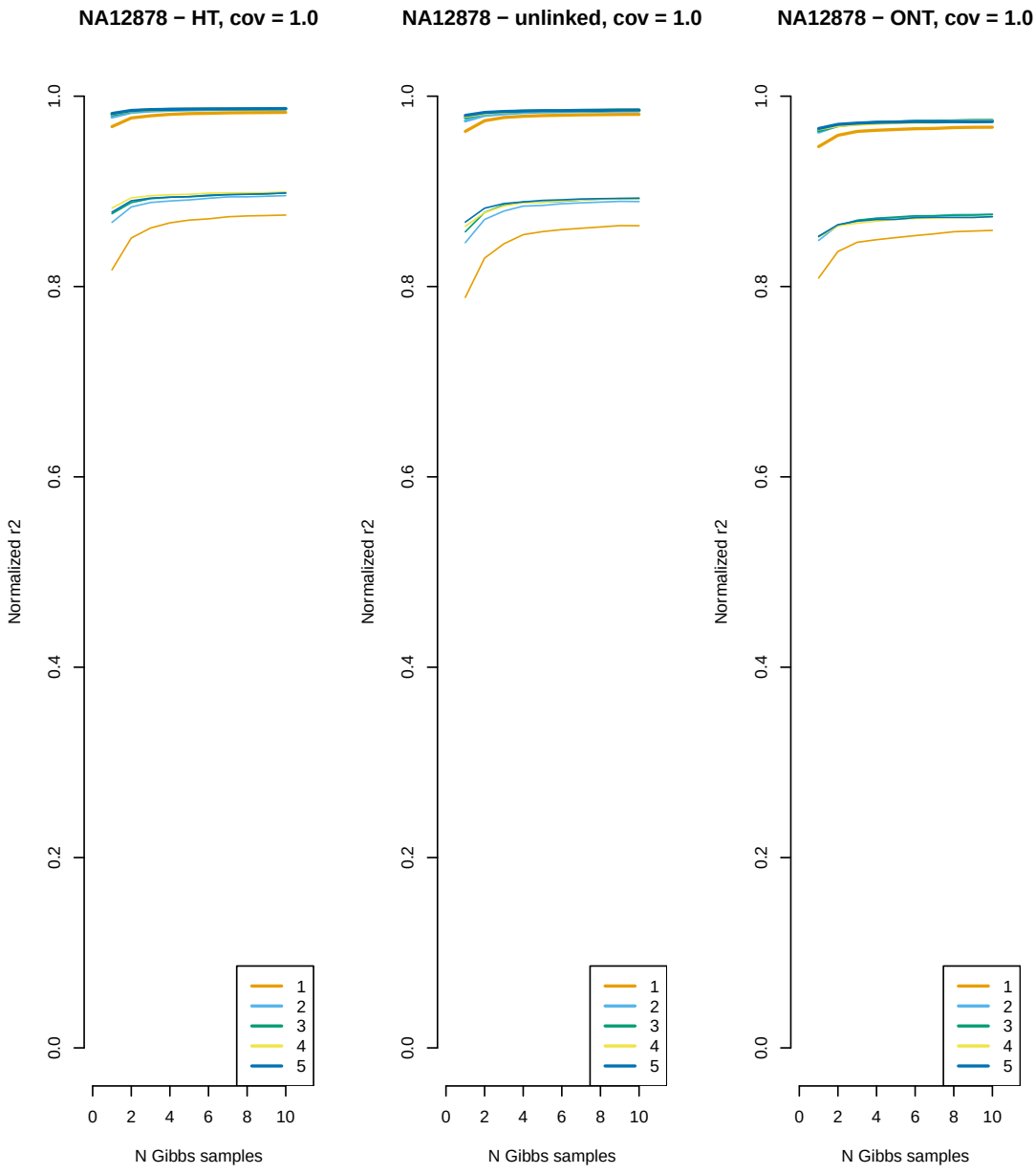
(a) 0.1X coverage



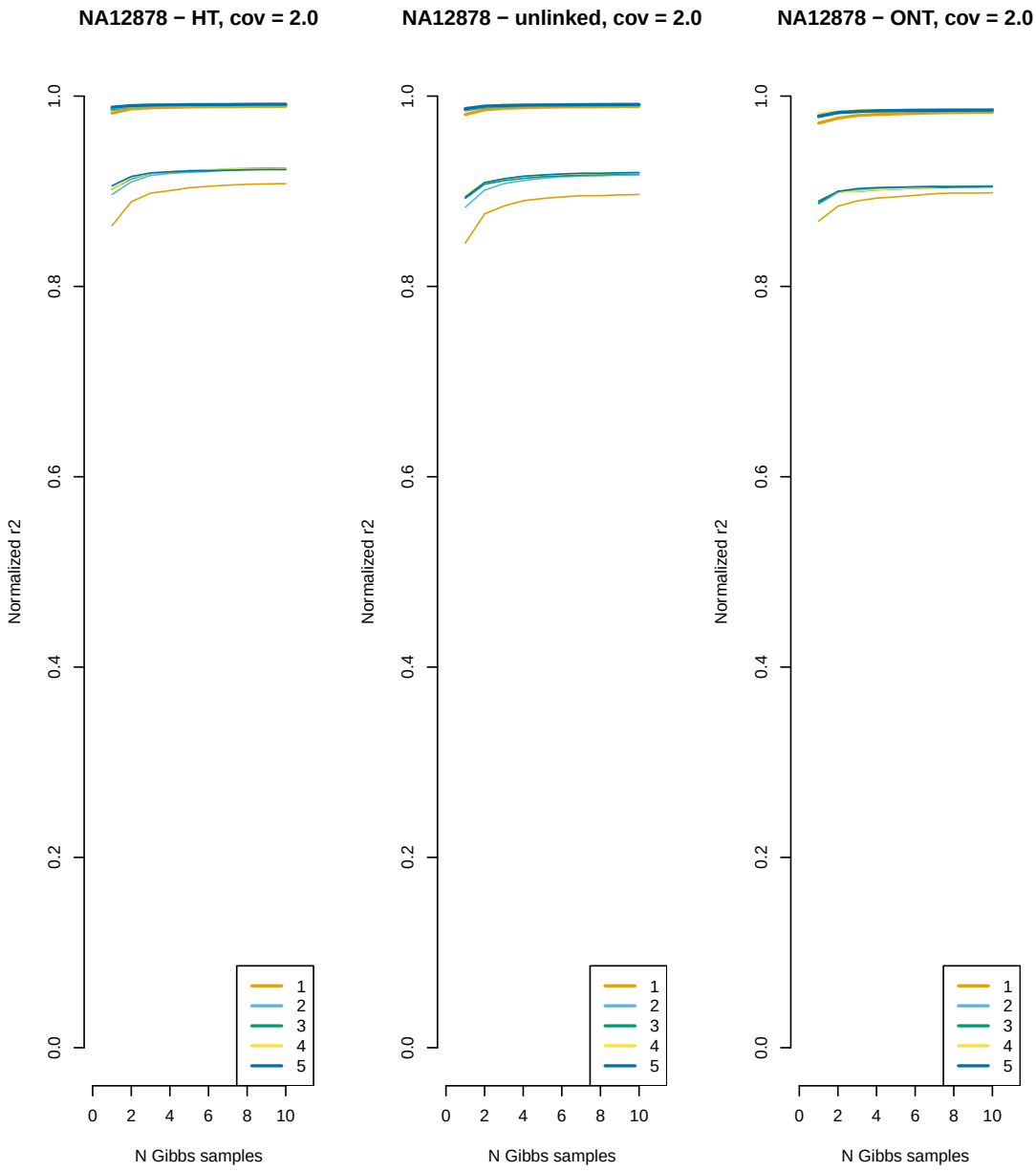
(b) 0.25X coverage



(c) 0.5X coverage

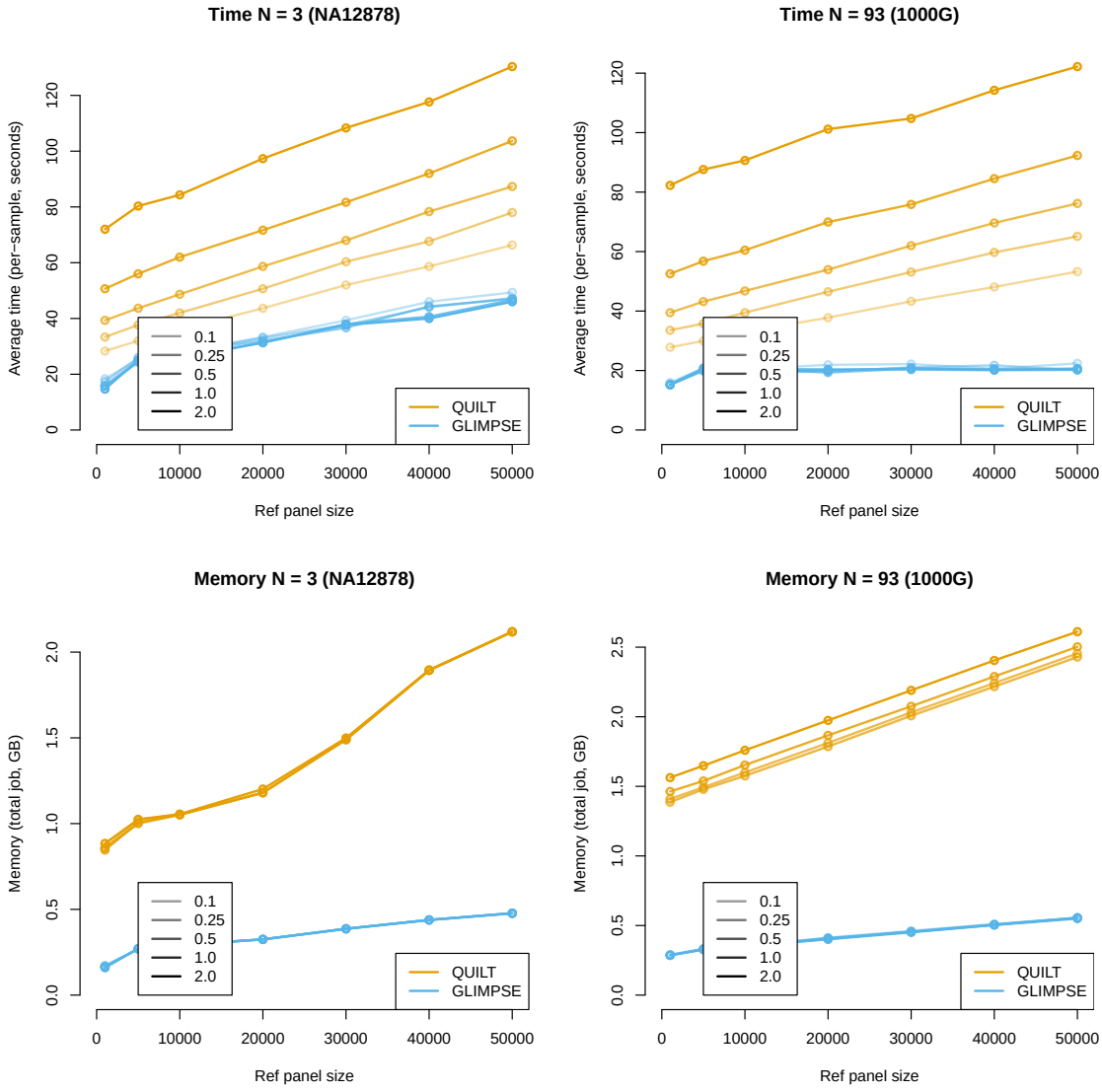


(d) 1.0X coverage

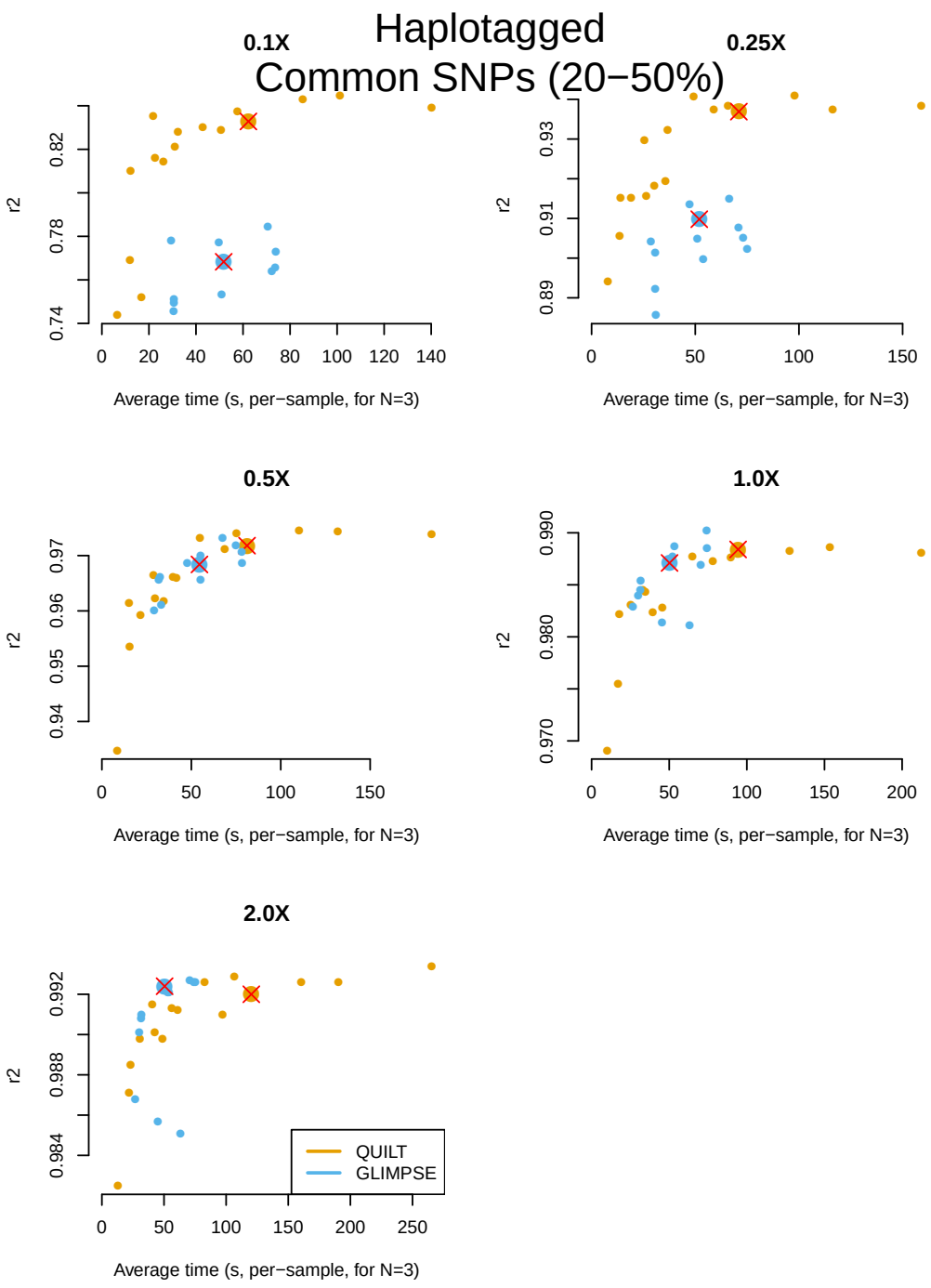


(e) 2.0X coverage

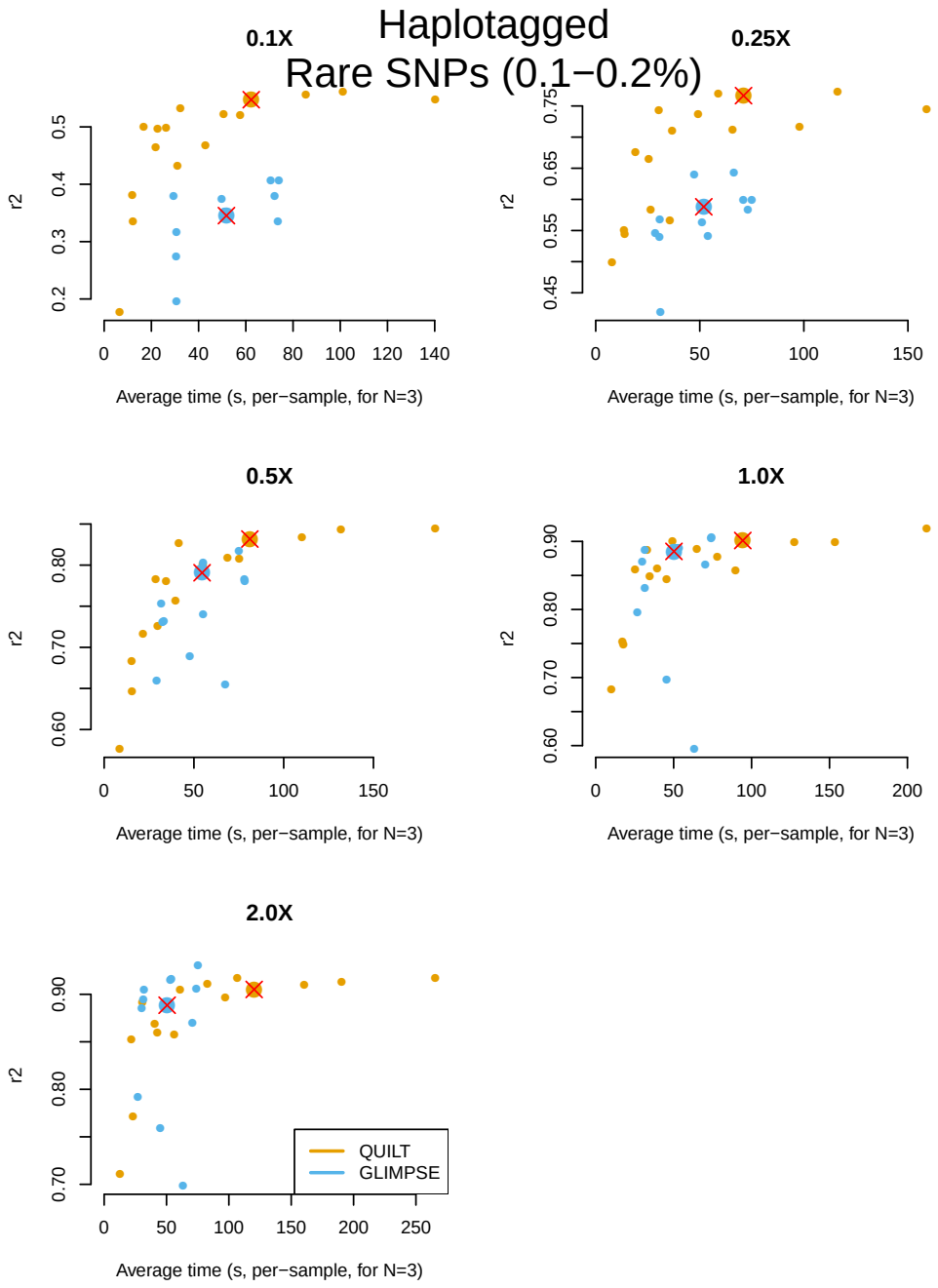
Supplementary Figure 1: Effect of parameters on QUILT imputation performance Effect of varying number of Gibbs samples and number of full haploid iterations on imputation performance for NA12878



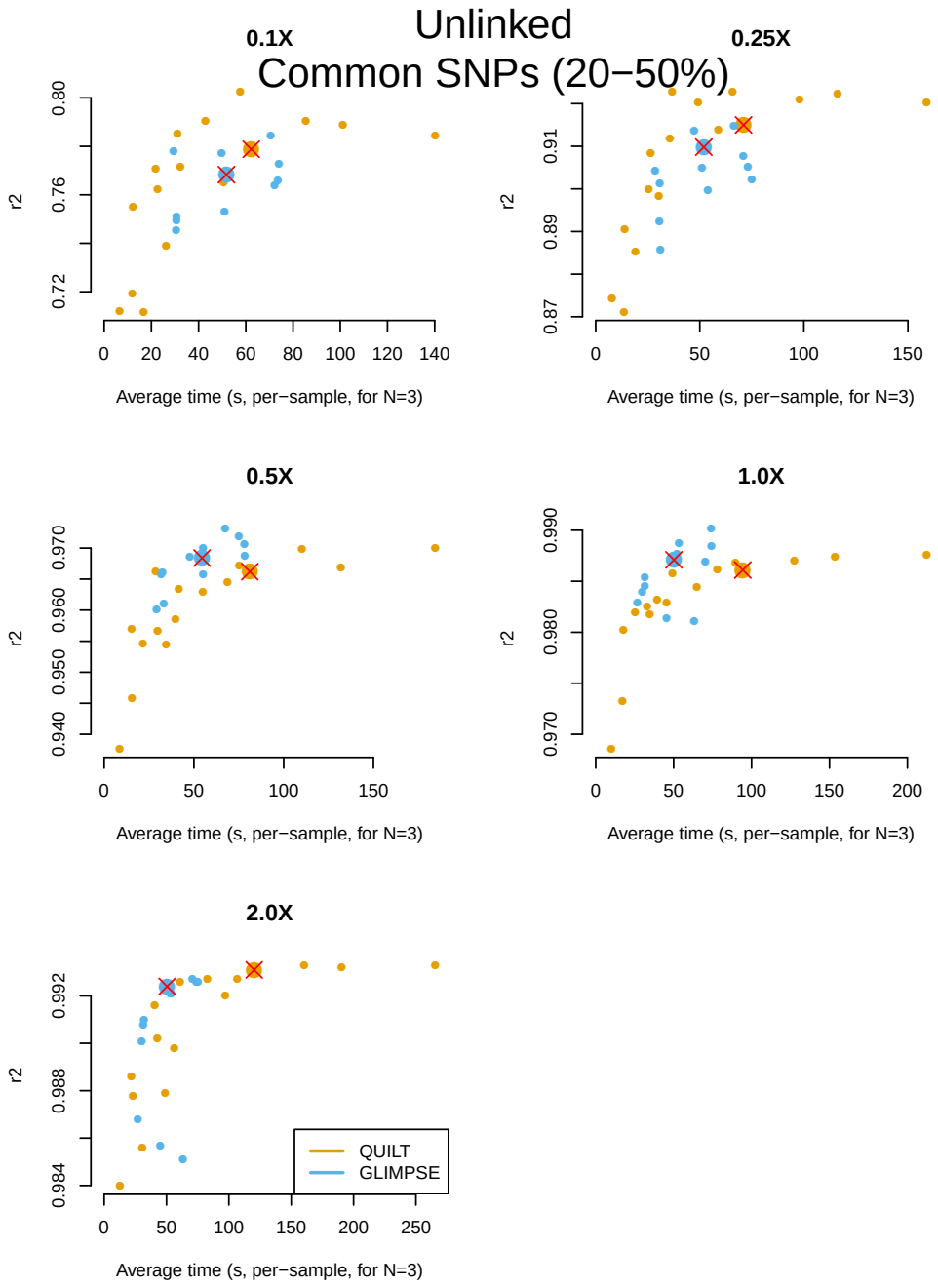
Supplementary Figure 2: Run time and memory usage Top row gives results for N=3 samples (NA12878), bottom row gives results for N=93 samples (1000G). Left column gives speed, which is given as per-sample time (*i.e.* the time plotted is the total time divided divided by the number of samples). Right column gives memory per-job using `/usr/bin/time -v` for "Maximum resident set size" (*i.e.* the memory plotted is per job and not divided by the number of samples). Results shown are average per region, where each region was 2 Mbp long with 500 kbp buffer (a total of 200 Mbp was imputed). Results are shown for 1X coverage. x-axis values are 1000, 5000, 10000, 20000, 30000, 40000, 50000



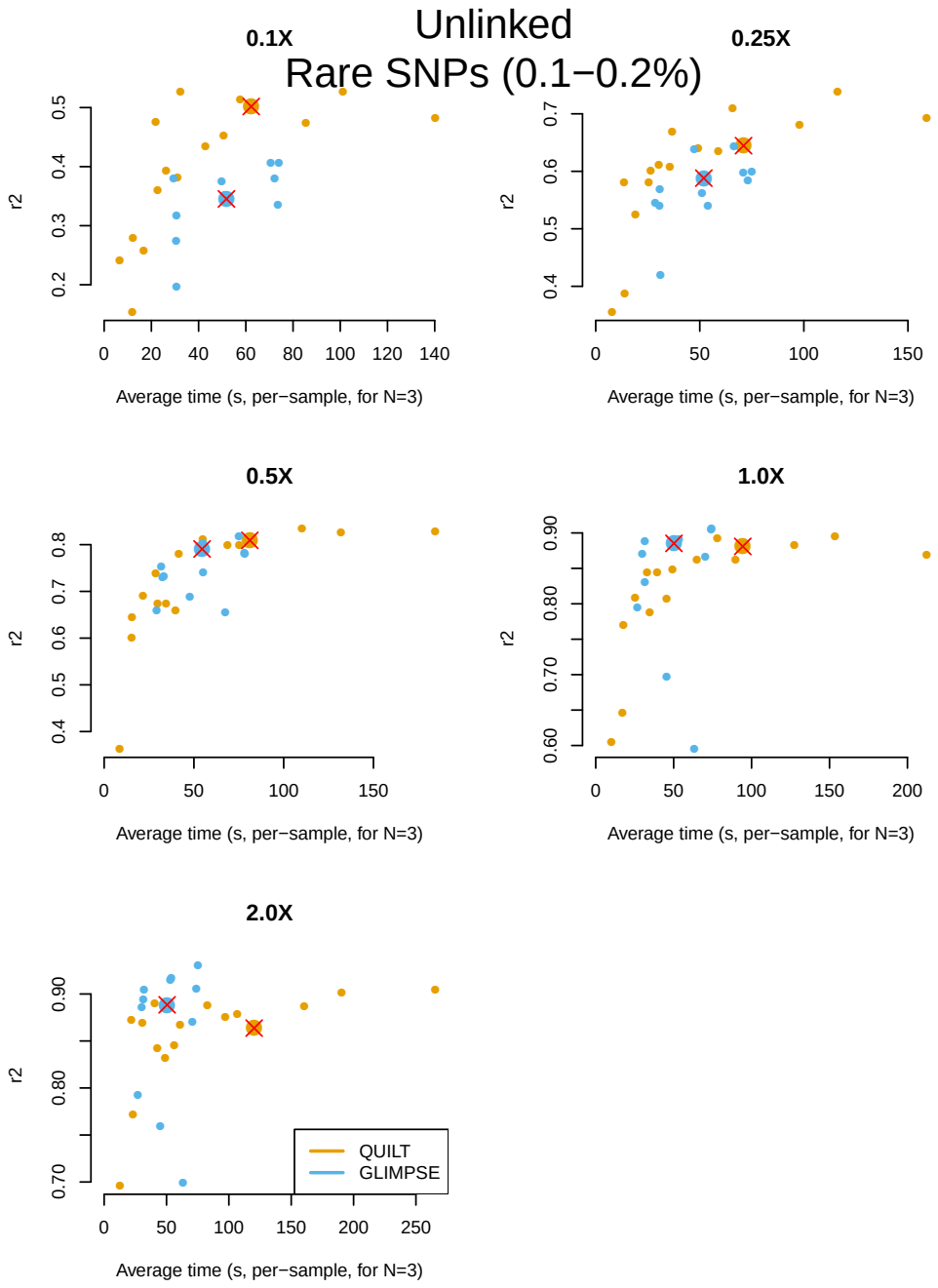
(a) Haplotagged common SNPs (20-50% allele frequency)



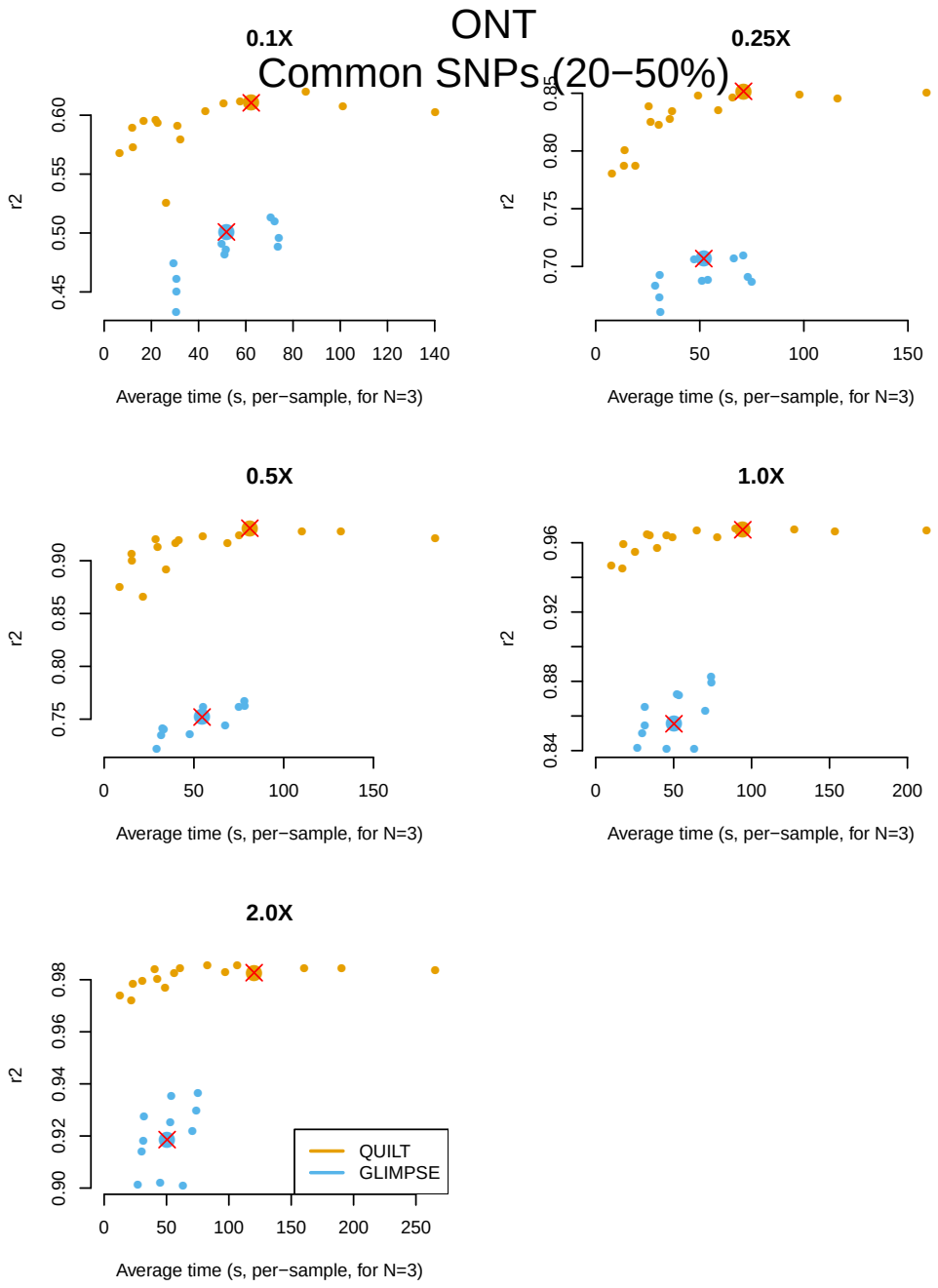
(b) Haplotagged rare SNPs (0.1-0.2% allele frequency)



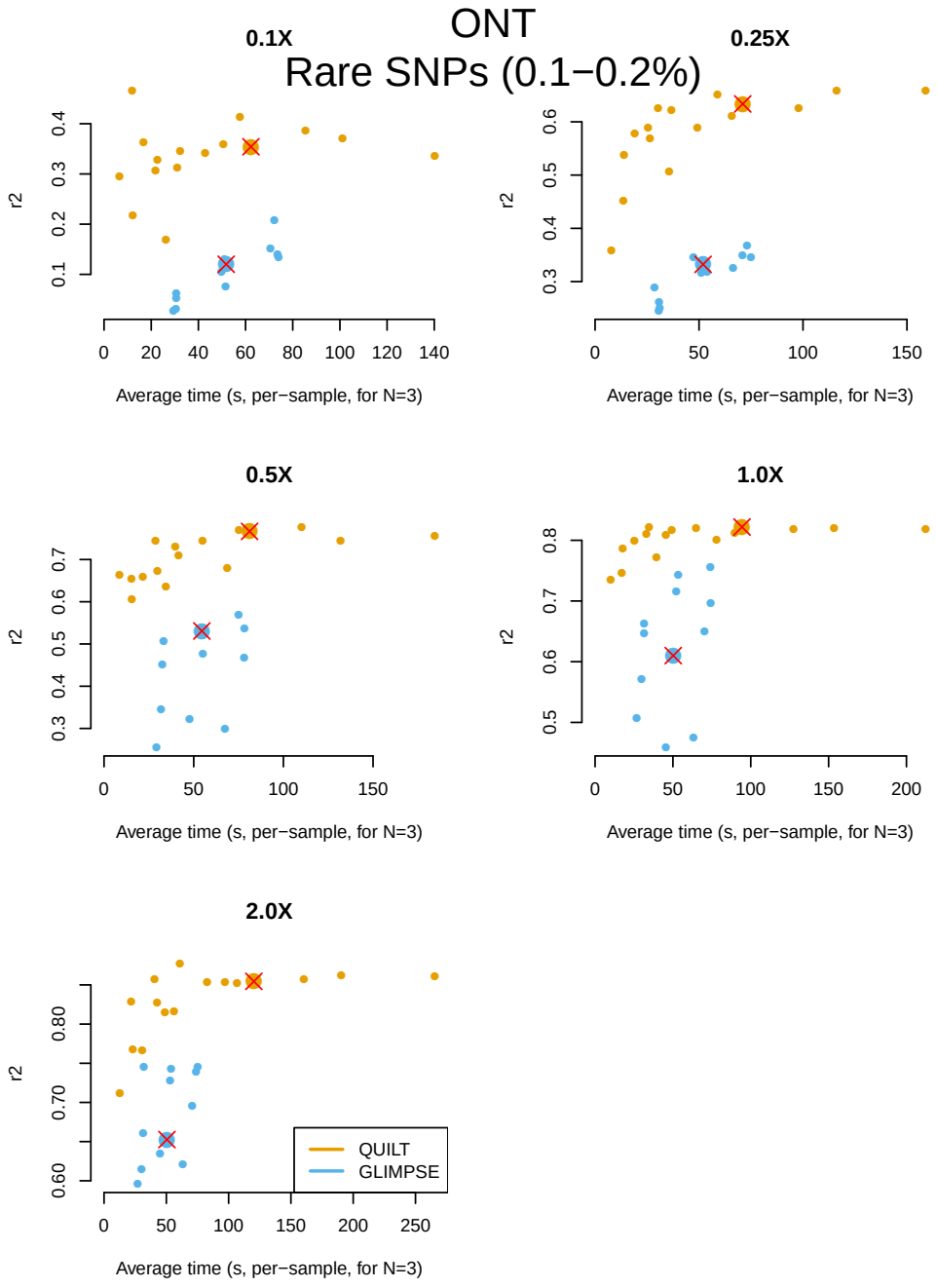
(c) Illumina common SNPs (20-50% allele frequency)



(d) Illumina rare SNPs (0.1-0.2% allele frequency)

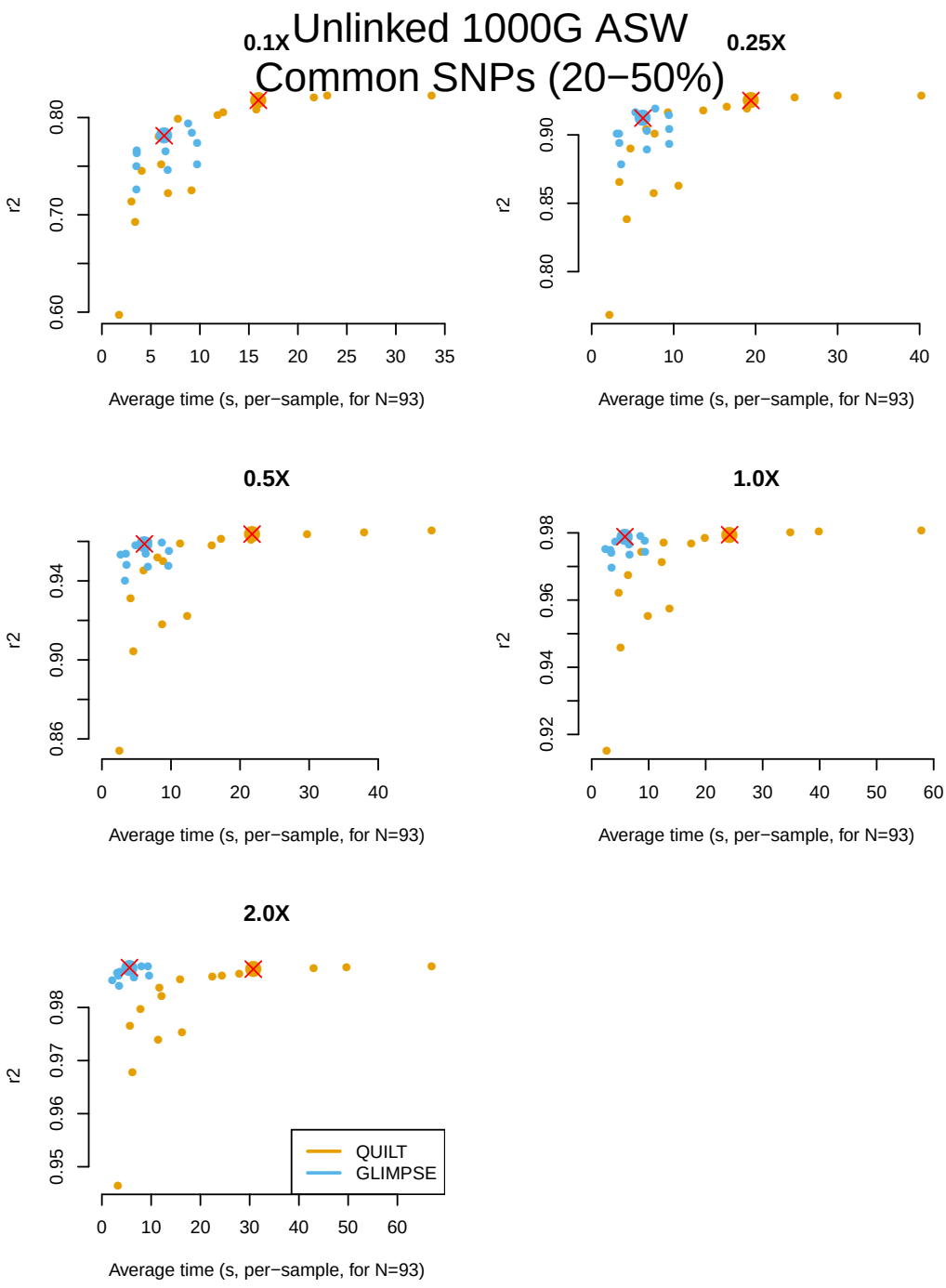


(e) ONT common SNPs (20-50% allele frequency)

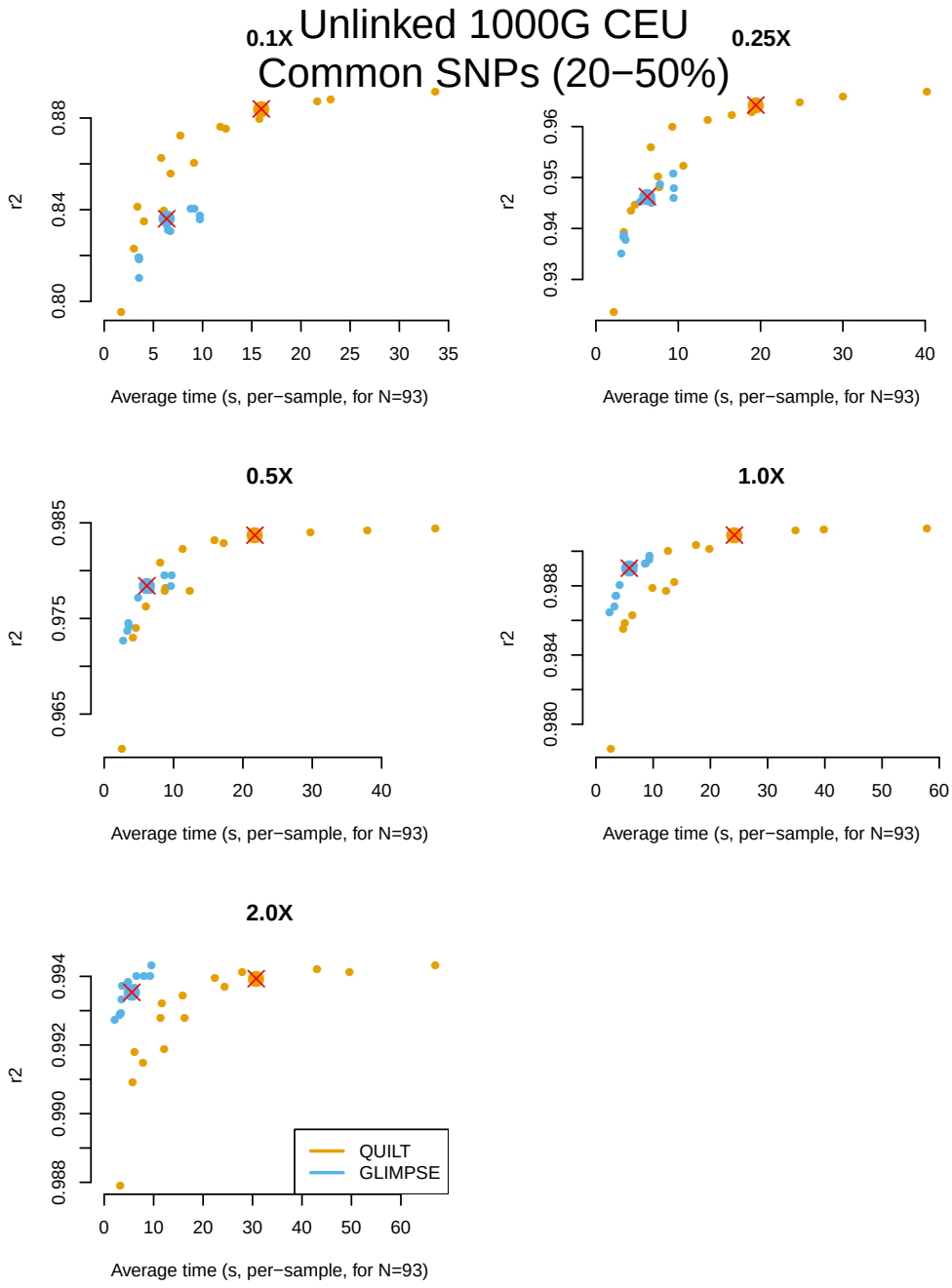


(f) ONT rare SNPs (0.1-0.2% allele frequency)

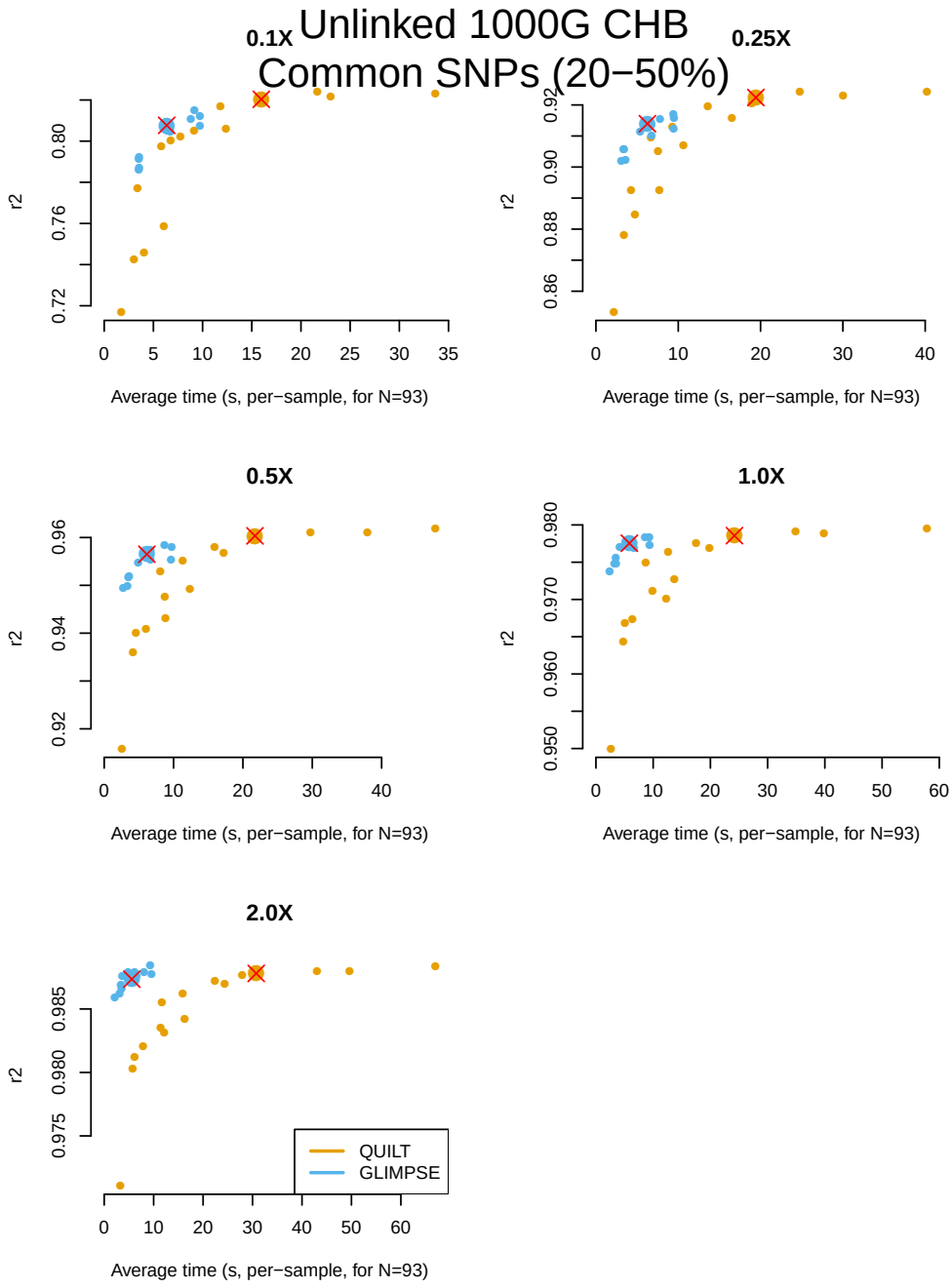
Supplementary Figure 3: QUILT and GLIMPSE accuracy over parameter values, for NA12878 data Shown are QUILT and GLIMPSE performance for NA12878 for different data types as a function of run time given different parameter settings. QUILT varies number of Gibbs sample iterations (1, 3, 7 (default), 10) and iterative Gibbs sampling iterations (1, 2, 3 (default), 5), while GLIMPSE varies number of burn in and Gibbs sampling iterations (5+5, 10+10 (default), 15+15) and pbwt-depth (1, 2 (default), 4, 8). Analysis was performed using chromosome 20 0-26 Mbp to reduce computational burden. We note that QUILT is computationally linear in the number of samples to impute while GLIMPSE decreases, and so for moderate sample sizes, the GLIMPSE values should be shifted to approximately 1/2 their current values for lower coverages to 1/6 of their original values for higher coverage (*i.e.* leftwards).



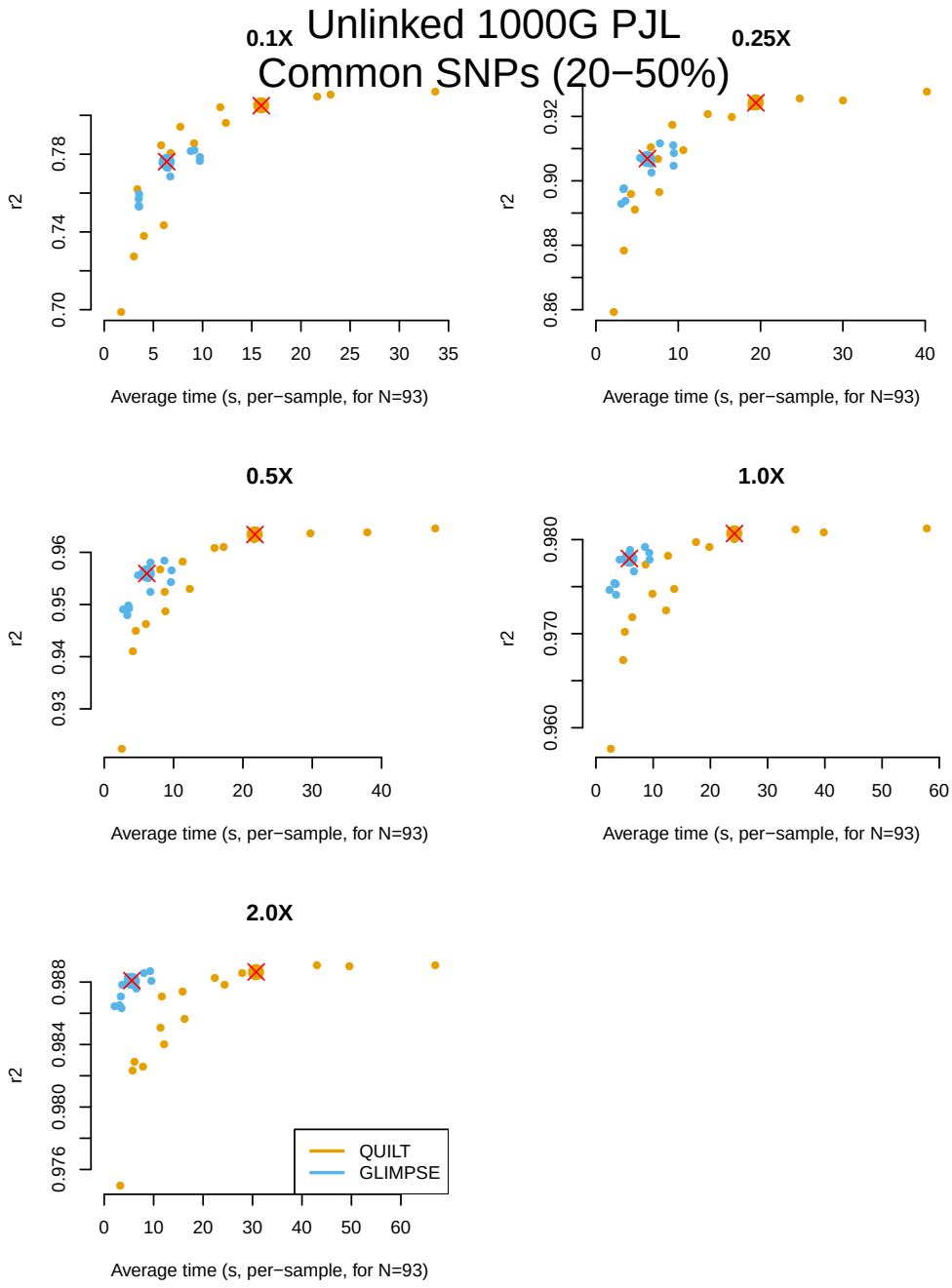
(a) Illumina 1000G common SNPs (20-50% allele frequency)



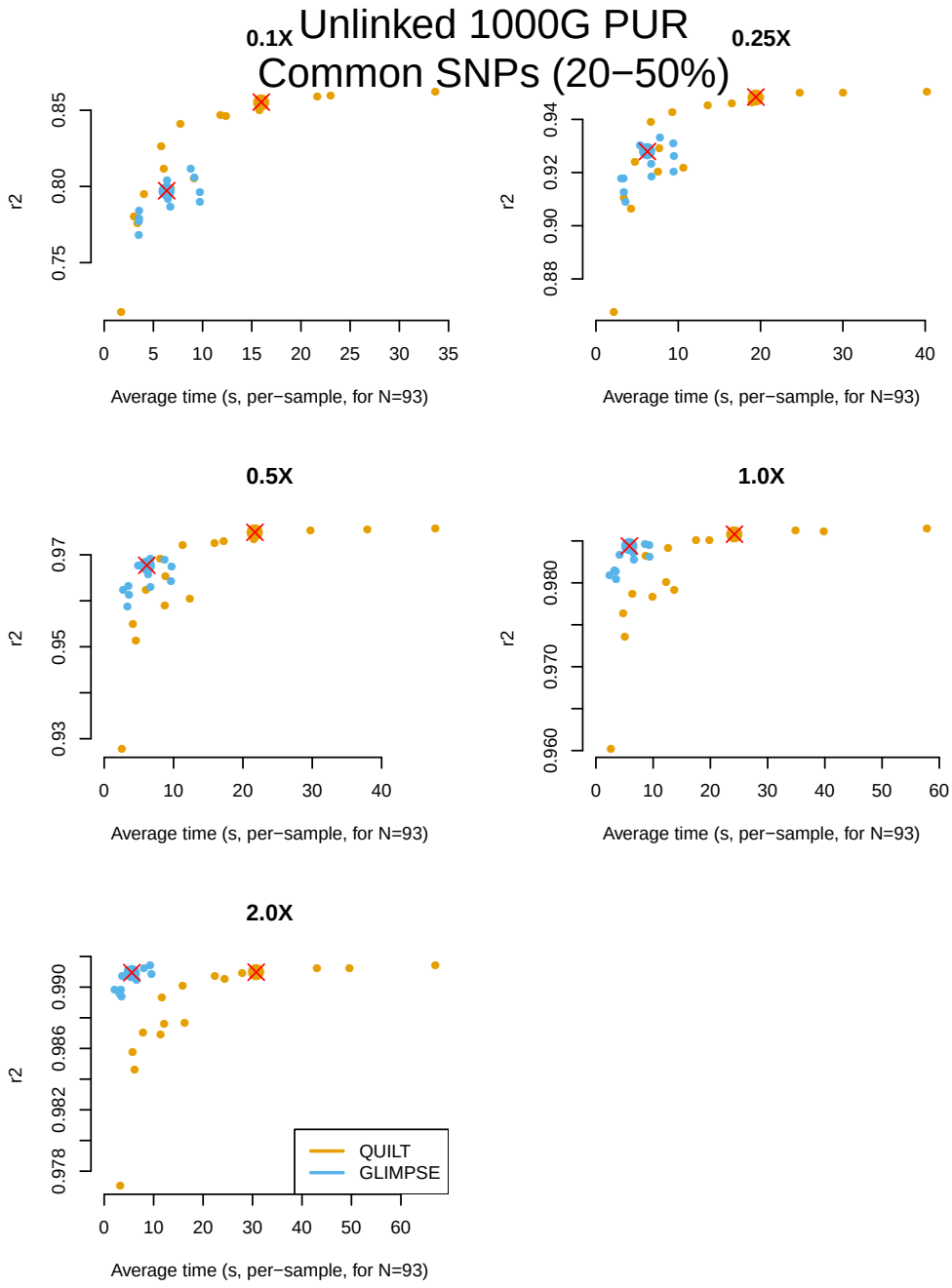
(b) Illumina 1000G common SNPs (20-50% allele frequency)



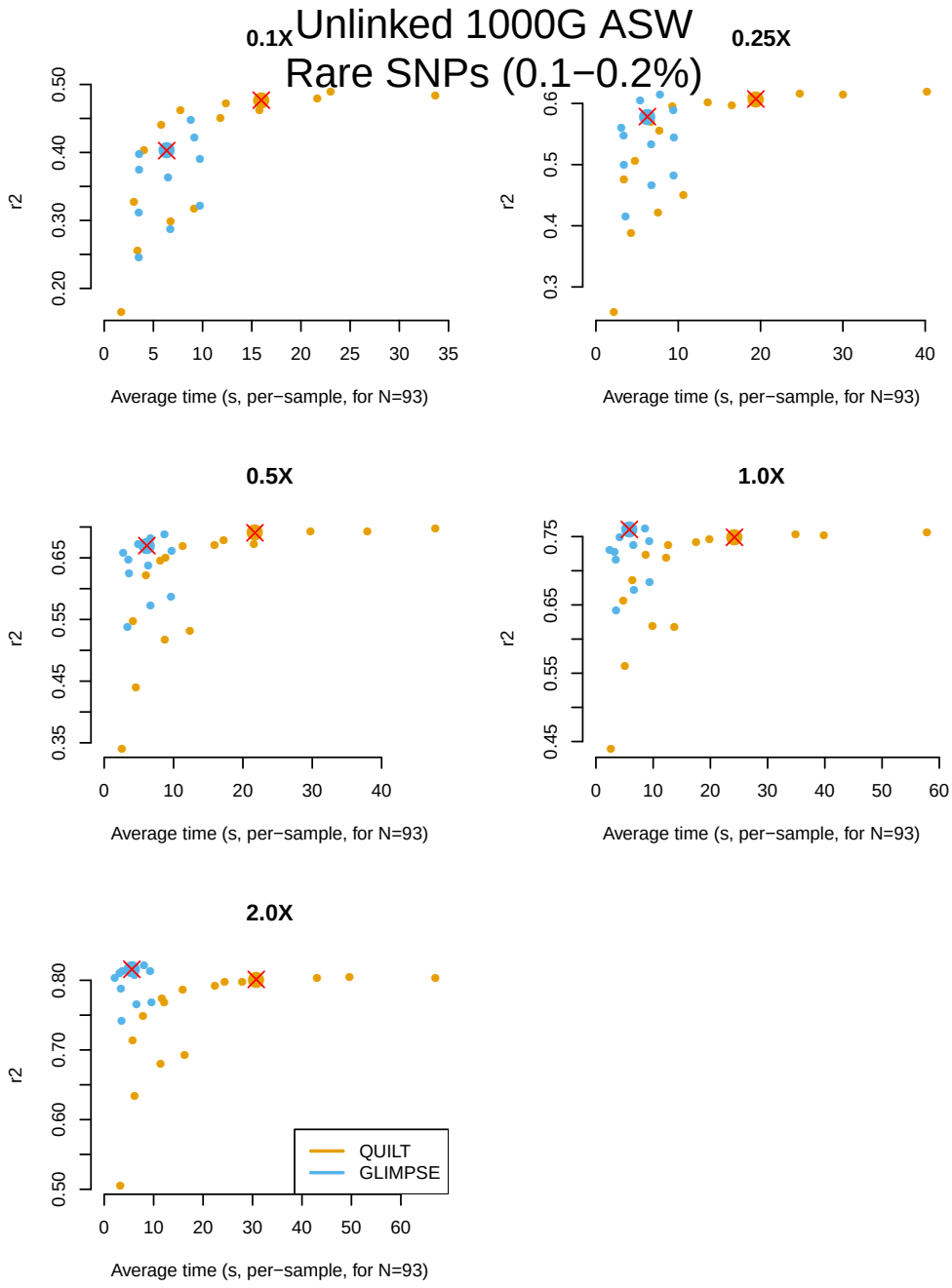
(c) Illumina 1000G common SNPs (20-50% allele frequency)



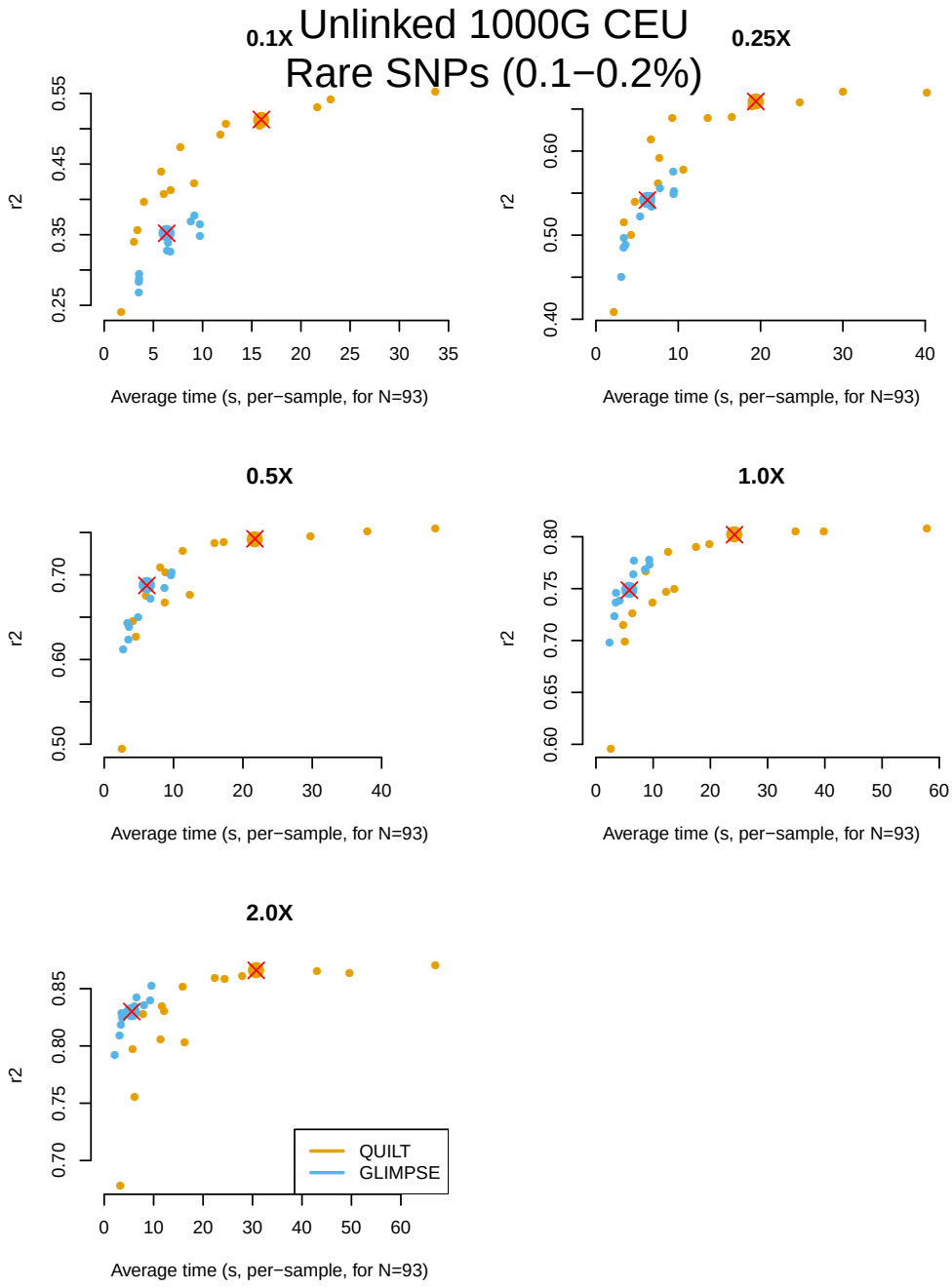
(d) Illumina 1000G common SNPs (20-50% allele frequency)



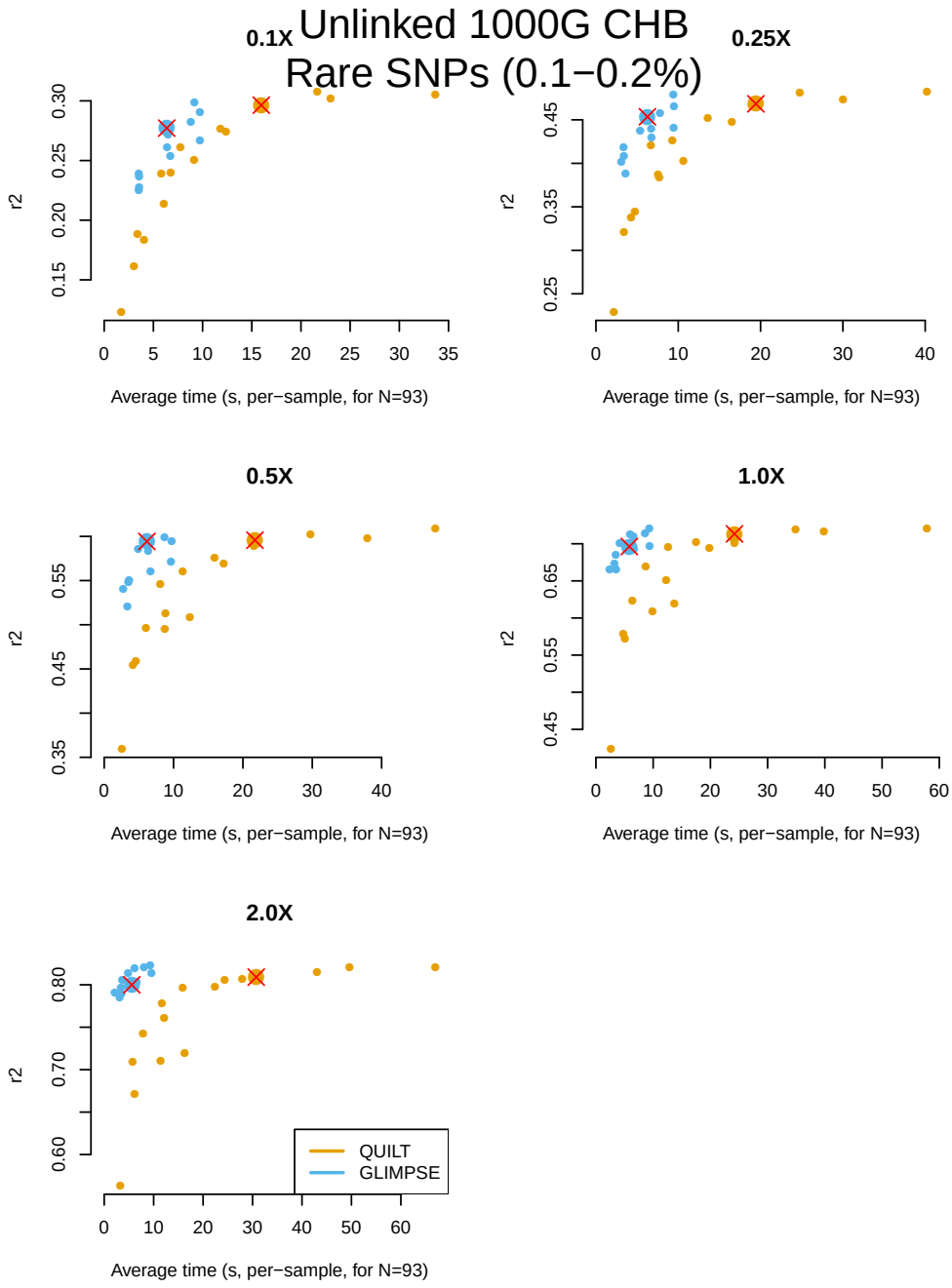
(e) Illumina 1000G common SNPs (20-50% allele frequency)



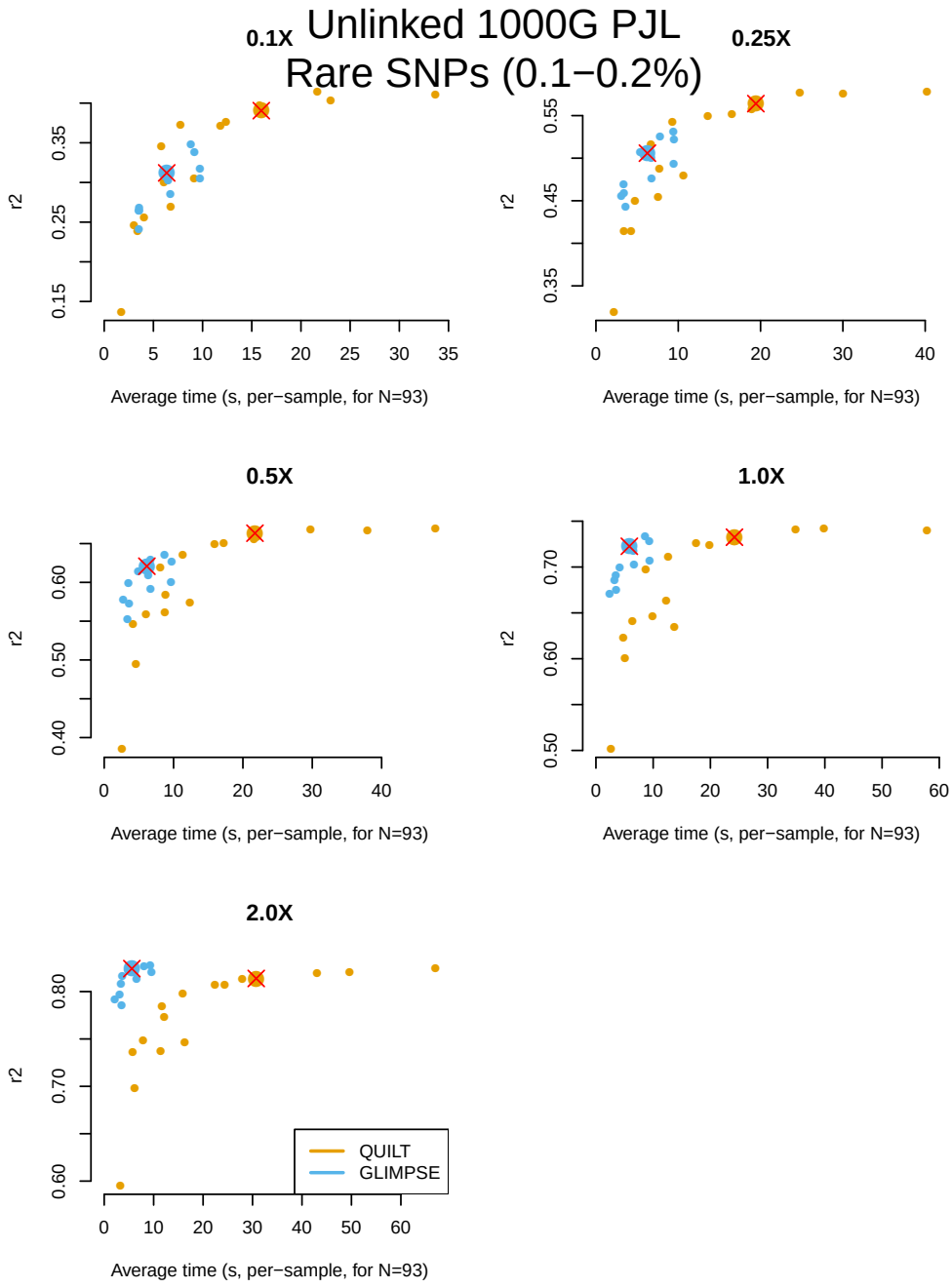
(f) Illumina 1000G rare SNPs (0.1-0.2% allele frequency)



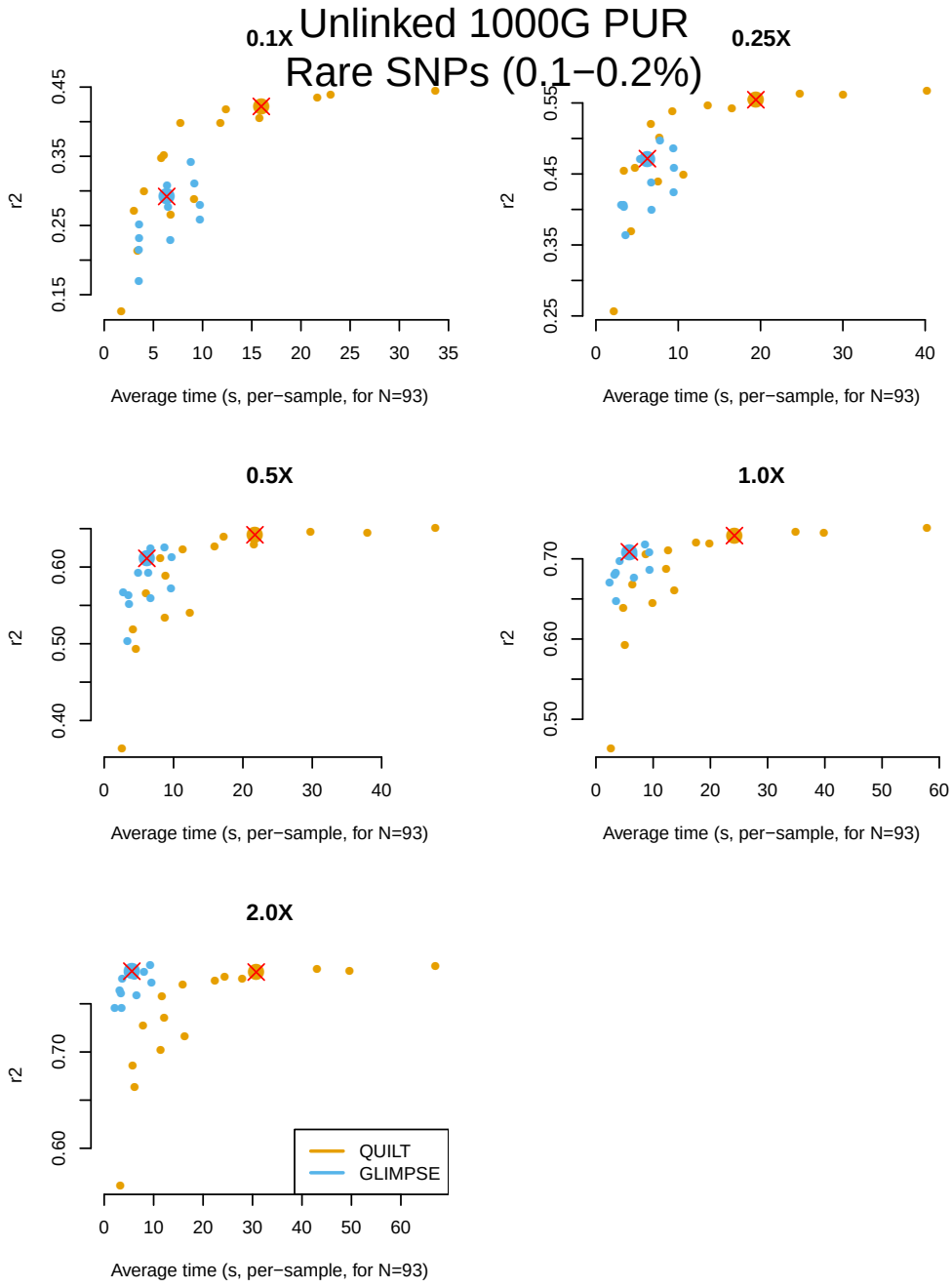
(g) Illumina 1000G rare SNPs (0.1-0.2% allele frequency)



(h) Illumina 1000G rare SNPs (0.1-0.2% allele frequency)

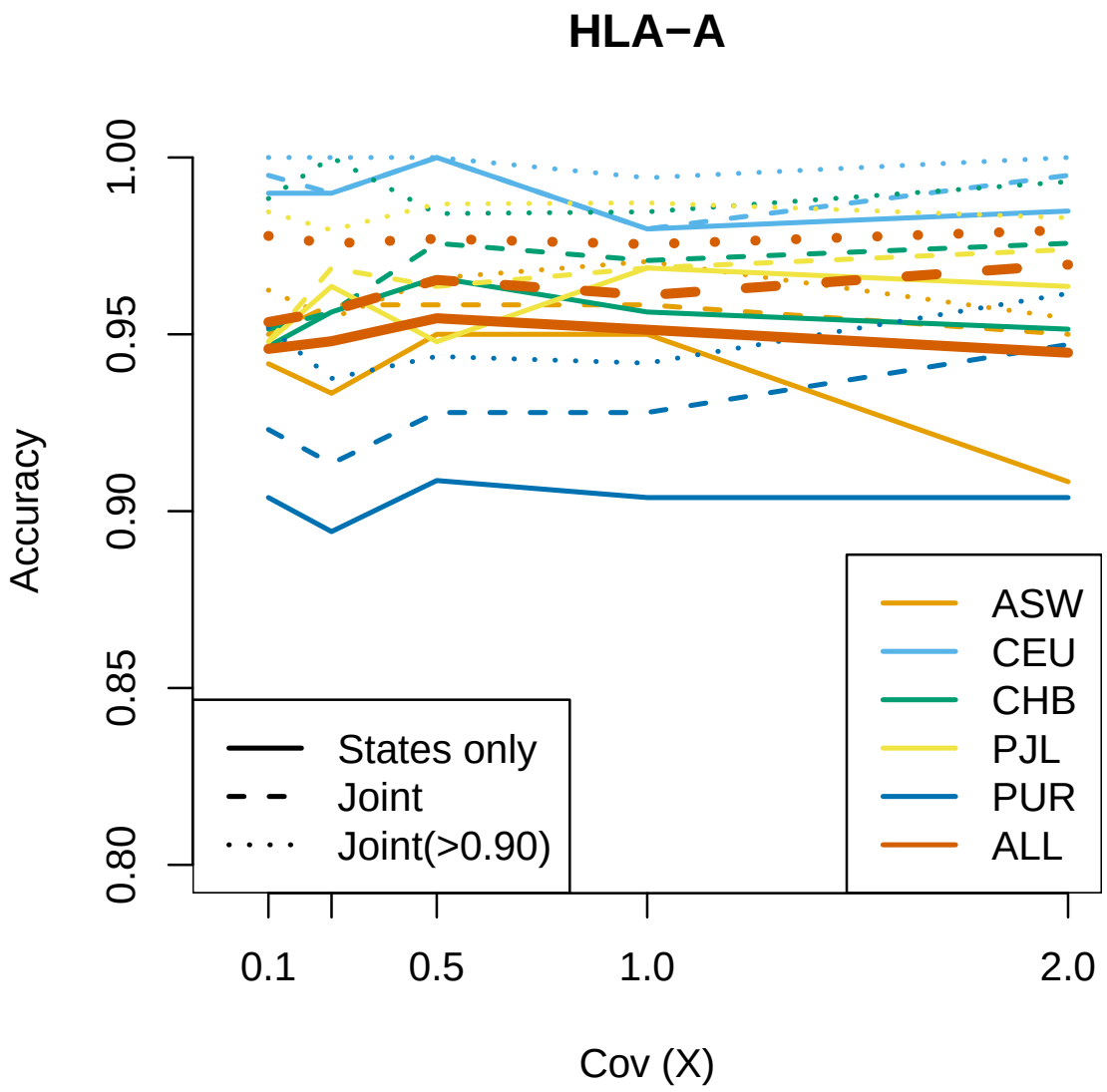


(i) Illumina 1000G rare SNPs (0.1-0.2% allele frequency)

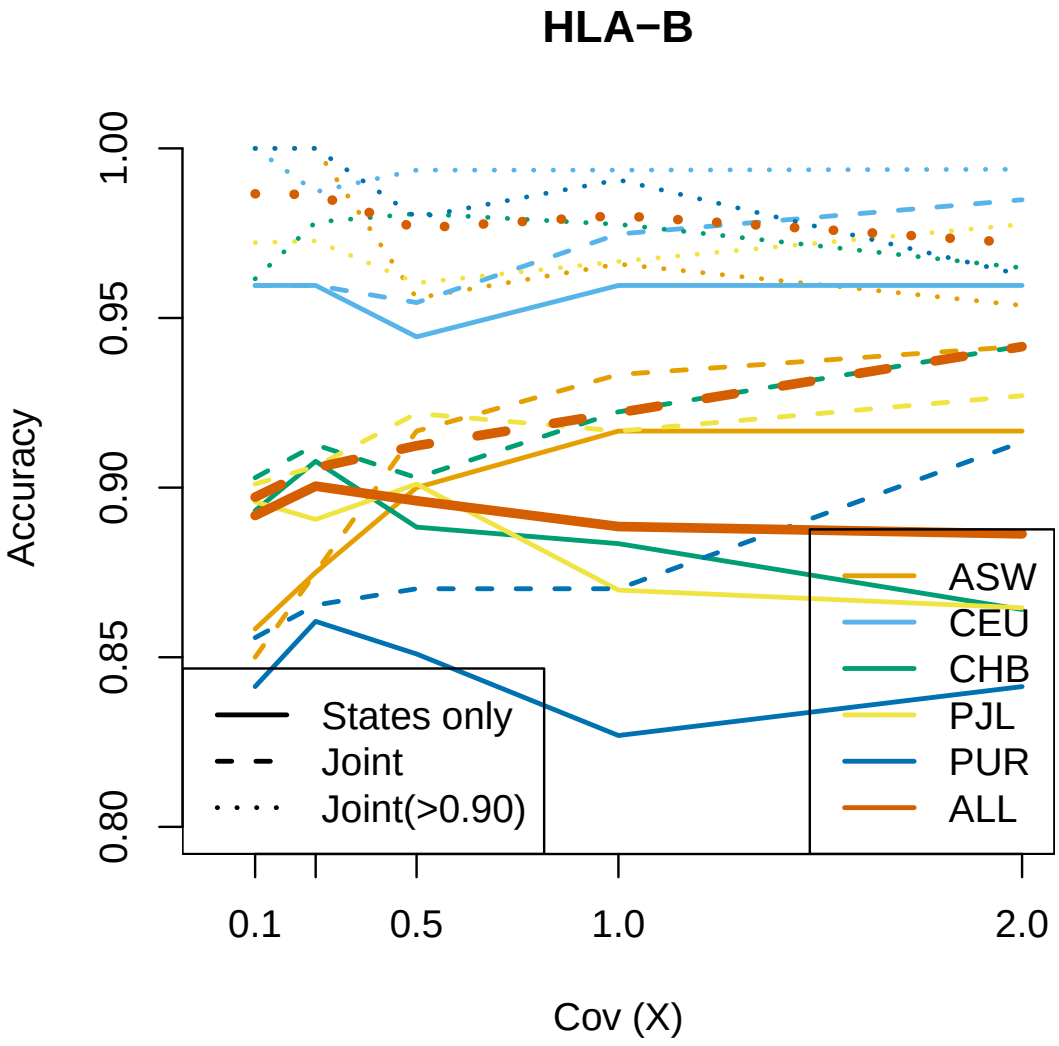


(j) Illumina 1000G rare SNPs (0.1-0.2% allele frequency)

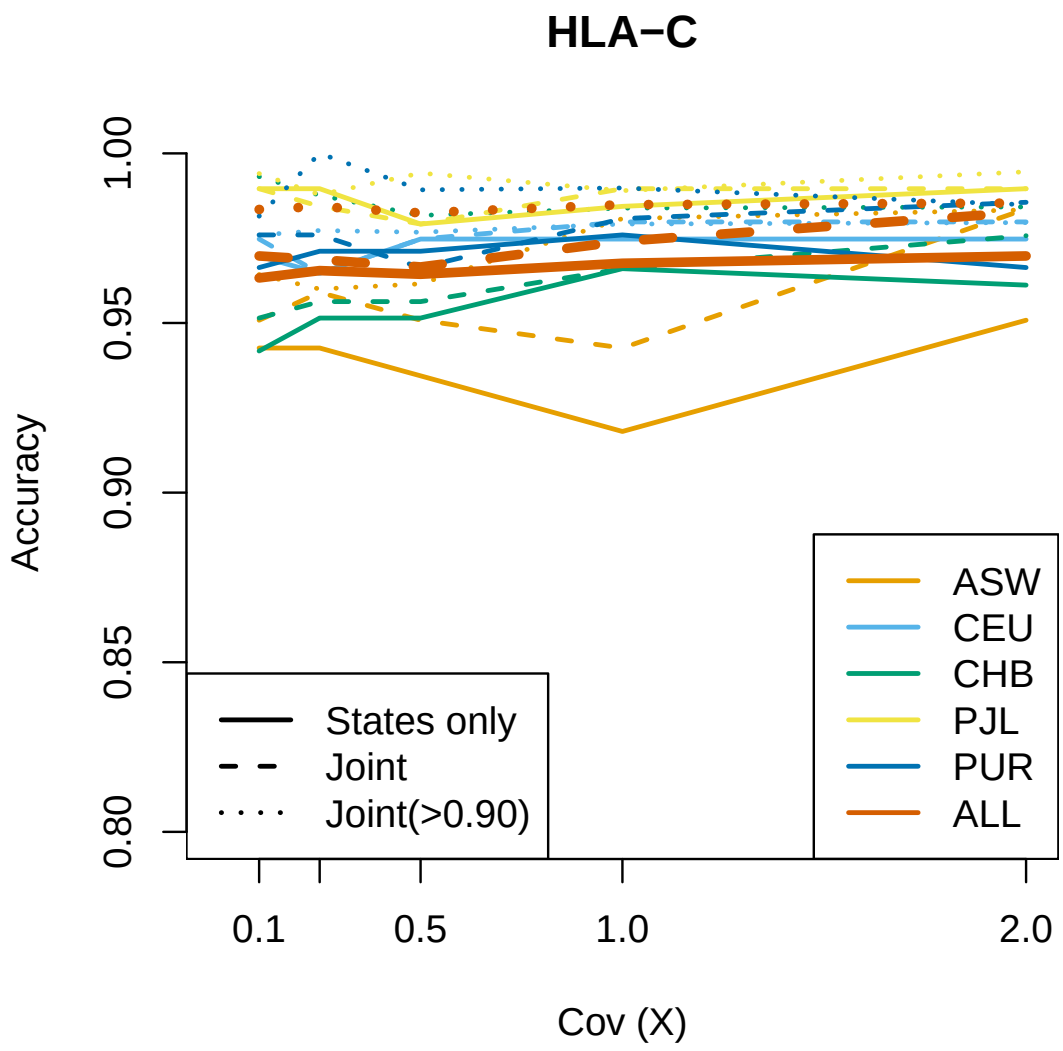
Supplementary Figure 4: QUILT and GLIMPSE accuracy over parameter values, for 1000 Genomes data Shown are QUILT and GLIMPSE performance for 1000 Genomes for Illumina data as a function of run time given different parameter settings. QUILT varies number of Gibbs sample iterations (1, 3, 7 (default), 10) and iterative Gibbs sampling iterations (1, 2, 3 (default), 5), while GLIMPSE varies number of burn in and Gibbs sampling iterations (5+5, 10+10 (default), 15+15) and pbwt-depth (1, 2 (default), 4, 8). Analysis was performed using chromosome 20 0-26 Mbp to reduce computational burden. This analysis was performed with $N = 93$ samples, and given the relationship between sample size and run time for both QUILT and GLIMPSE, results are indicative of performance expected at larger sample sizes.


HLA−A
Accuracy
1.00
0.95
0.90
0.85
0.80
0.1
0.5
1.0
2.0
Cov (X)
States only
Joint
Joint(>0.90)
ASW
CEU
CHB
PJL
PUR
ALL

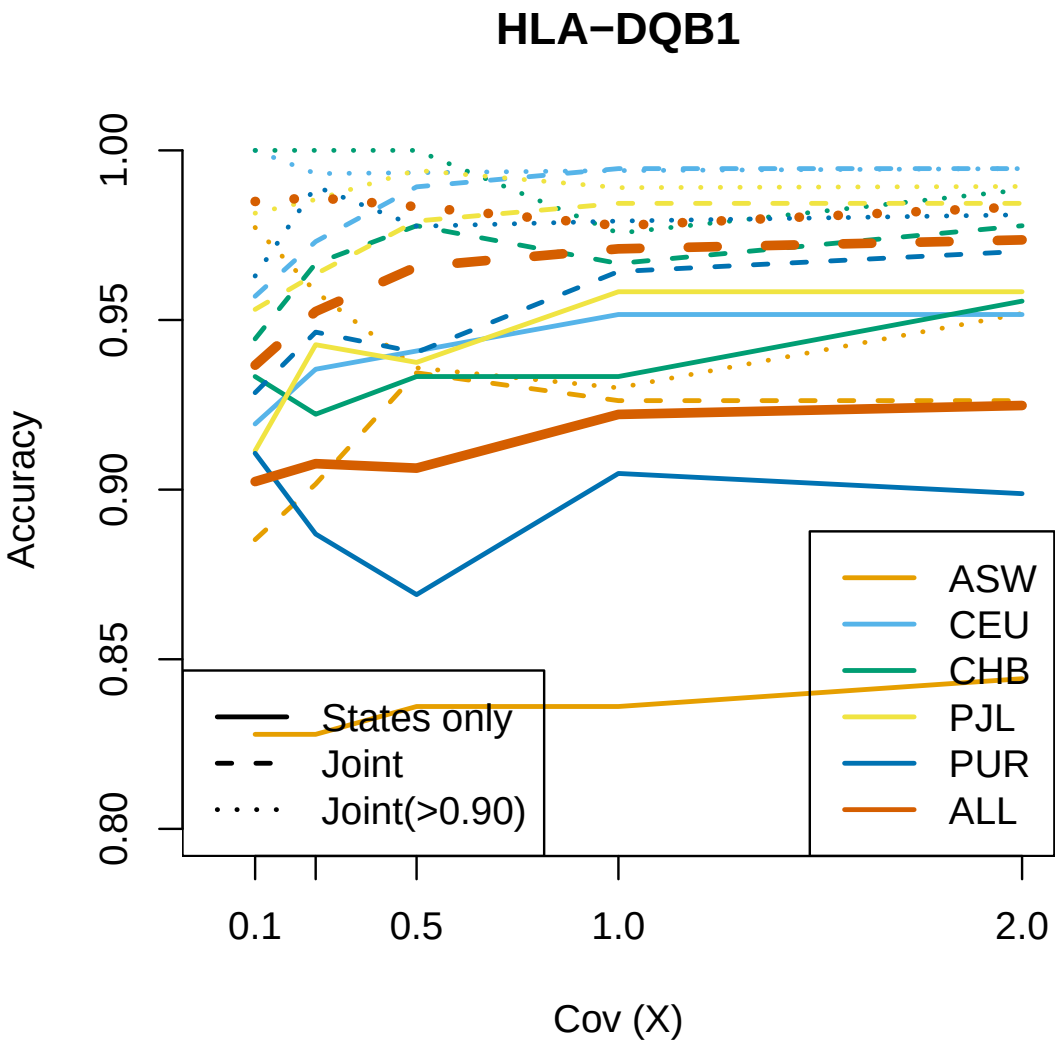
(a) HLA-A


HLA−B
Accuracy
1.00
0.95
0.90
0.85
0.80
0.1
0.5
1.0
2.0
Cov (X)
States only
Joint
Joint(>0.90)
ASW
CEU
CHB
PJL
PUR
ALL

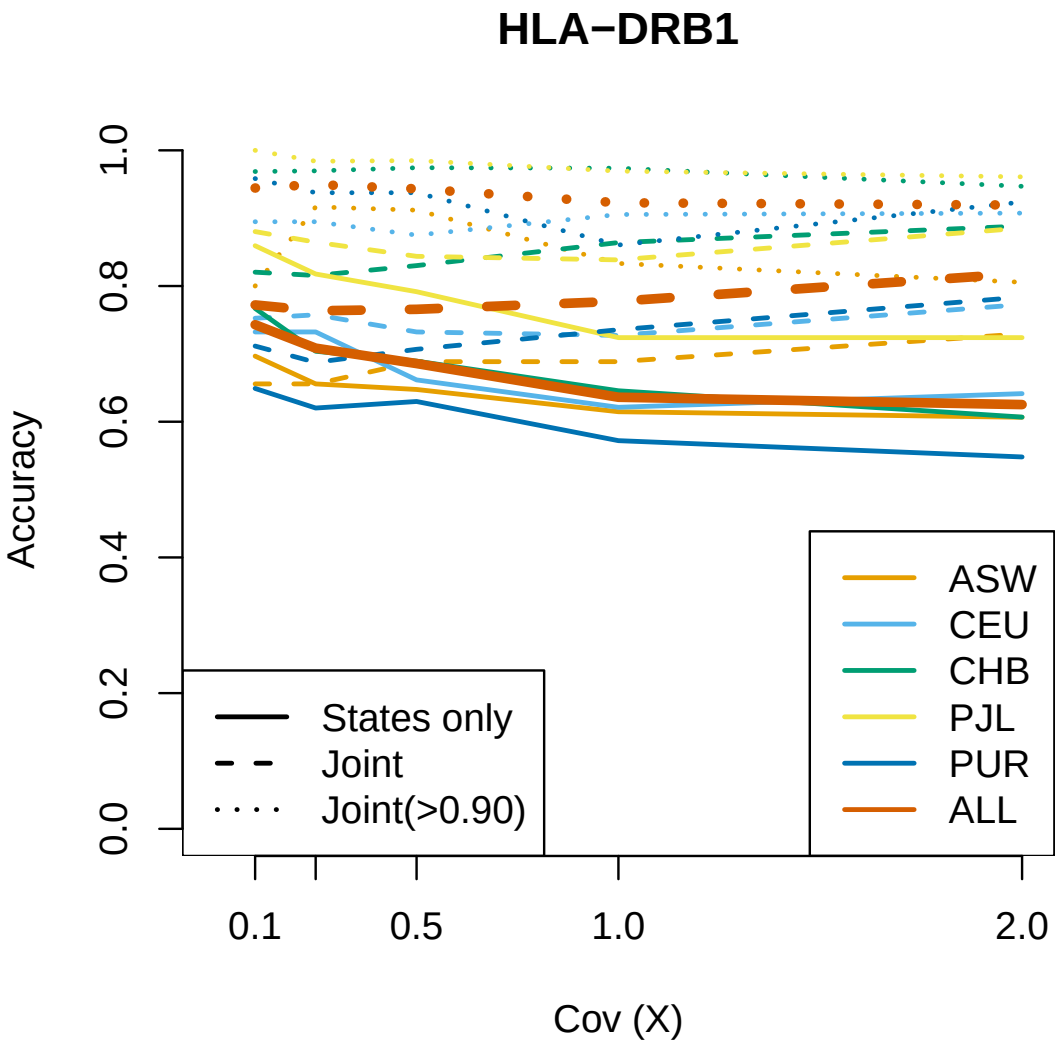
(b) HLA-B


HLA−C
Accuracy
1.00
0.95
0.90
0.85
0.80
0.1
0.5
1.0
2.0
Cov (X)
States only
Joint
Joint(>0.90)
ASW
CEU
CHB
PJL
PUR
ALL

(c) HLA-C

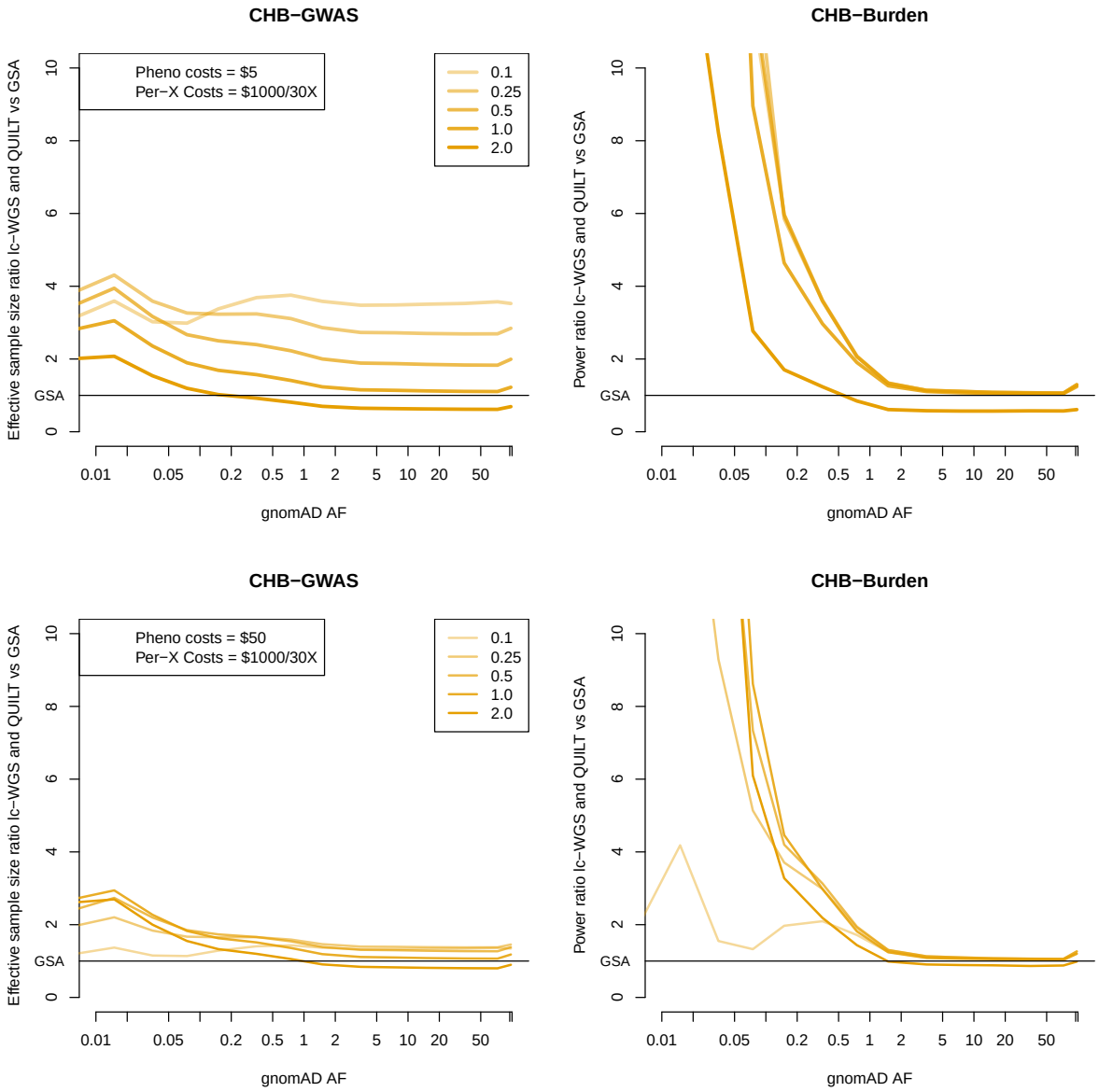

HLA−DQB1
Accuracy
1.00
0.95
0.90
0.85
0.80
0.1
0.5
1.0
2.0
Cov (X)
States only
Joint
Joint(>0.90)
ASW
CEU
CHB
PJL
PUR
ALL

(d) HLA-DQB1

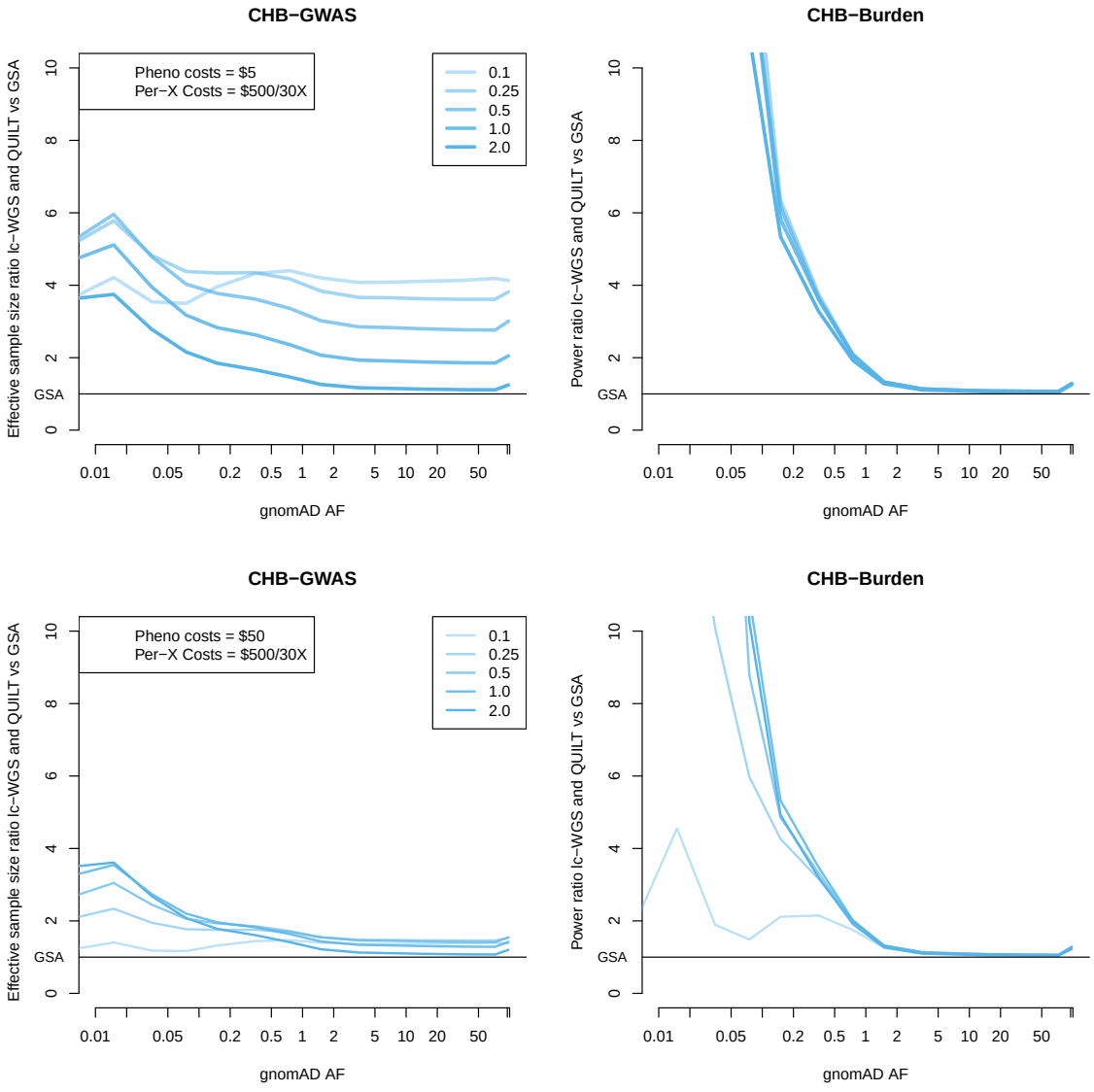


(e) HLA-DRB1

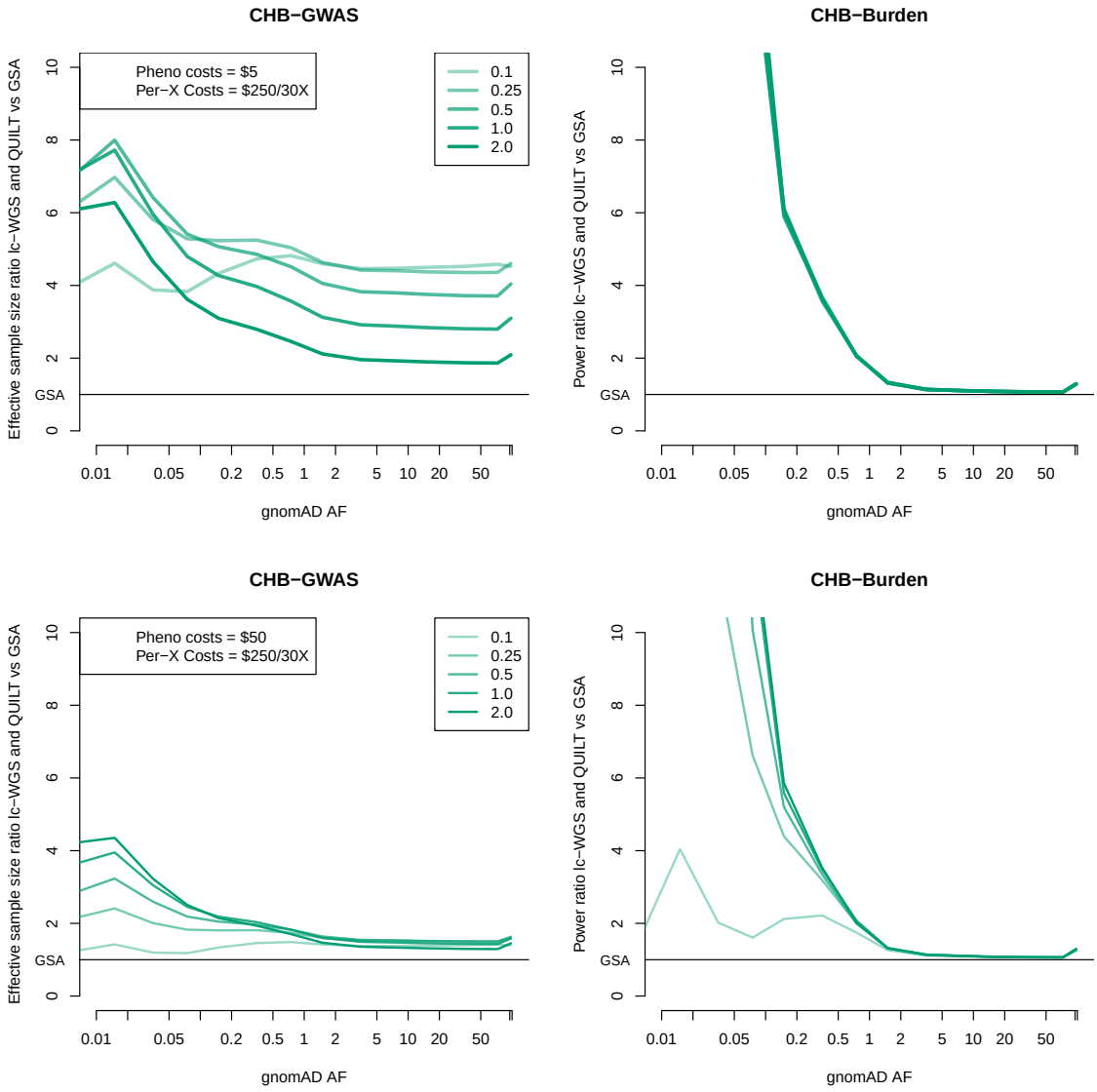
Supplementary Figure 5: HLA imputation accuracy as function of coverage States only refers to performing HLA typing using only posterior probabilities, while joint uses reads, and joint (> 0.90) is confidently called loci only



(a) Per-X cost of $1000/30X



(b) Per-X cost of $500/30X



(c) Per-X cost of $250/30X

Supplementary Figure 6: Relative effective sample size and power of lc-WGS and QUILT versus genotyping microarrays, as a function of allele frequency Using imputation accuracy from the CHB population for variable allele frequencies, the ratio of effective sample size or power is shown for either GWAS-style or burden-style analyses. Results vary per-X sequencing costs and phenotyping costs. Per-X costs are done assuming $1000/30X, $500/30X and $250/30X, and phenotyping costs of $5 or $50 per sample.