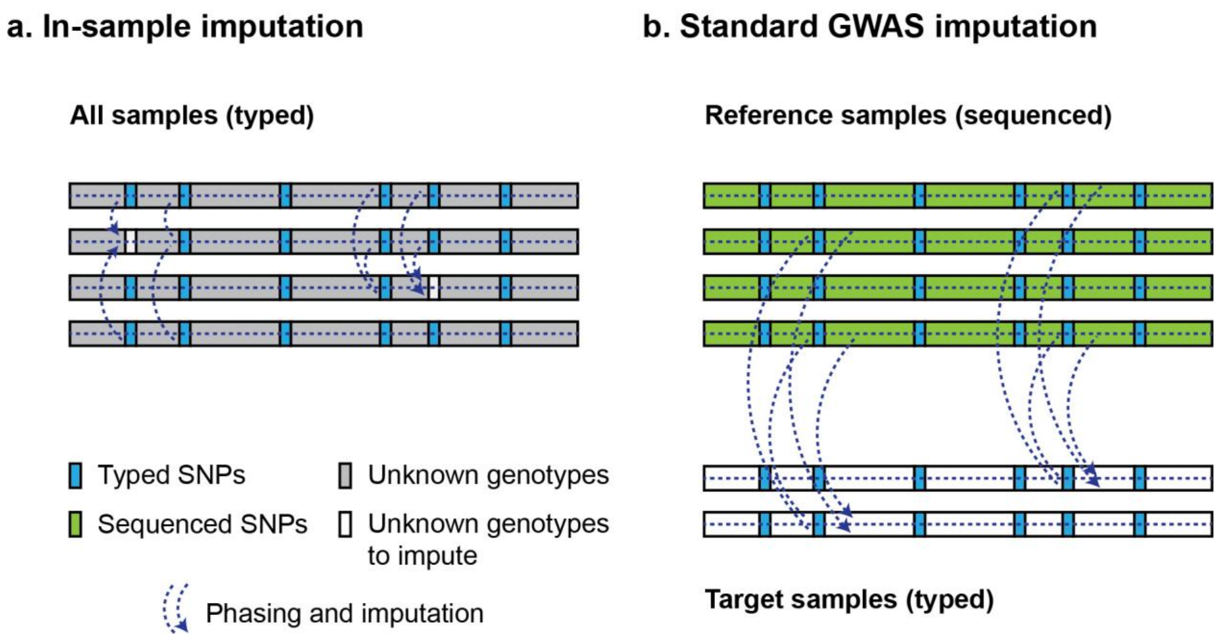
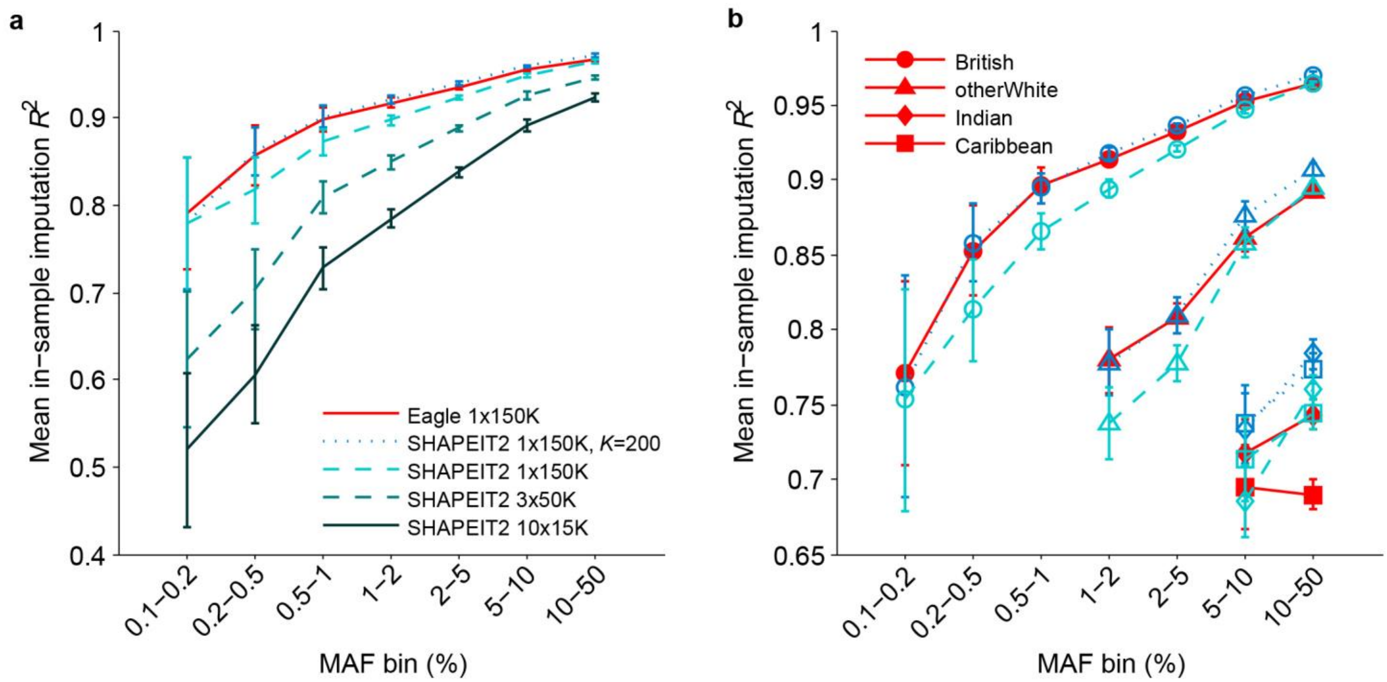


Supplementary Figure 1

Comparison of in-sample imputation and standard GWAS imputation.

Standard GWAS imputation differs from in-sample imputation in three ways. First, GWAS imputation usually involves imputing sequence data from a reference panel into a (genotyped but not sequenced) target sample, which typically requires phasing the sequenced reference (possibly using read information[36]), phasing the target sample (possibly using the phased reference), and imputing reference data into the target sample; in contrast, in-sample imputation involves only one sample, serving as both target and reference, that is simultaneously phased and imputed. Second, GWAS imputation pipelines produce probabilistic allele 'dosage' estimates, whereas phasing methods produce hard calls at missing genotypes (thus achieving suboptimal imputation R^2). Third, typical GWAS impute sequenced SNPs into target samples that are fully typed at a set of ascertained array SNPs, whereas phasing methods impute missing data at ascertained SNPs. (The latter task may be slightly harder than the former, as genotyping arrays are sometimes optimized to minimize redundancy among ascertained SNPs; thus, the LD between a typical ascertained SNP and its closest ascertained proxy may be lower than the LD between a typical sequenced SNP and its closest ascertained proxy. On the other hand, the fact that rare variants on genotyping arrays are typically enriched in densely typed fine-mapping regions may make in-sample imputation easier.) For all of these reasons, different algorithms are typically used for phasing versus GWAS imputation (e.g., SHAPEIT[10,12] versus IMPUTE[2,55] and MaCH[4] versus minimac[5,56]).

Nature Genetics: doi:10.1038/ng.3571



Supplementary Figure 2

In-sample imputation accuracy of Eagle and SHAPEIT2.

We randomly masked 2% of the genotypes in all N = 150,000 UK Biobank samples and phased the first 40 cM of chromosome 10 using Eagle (on the full cohort) and SHAPEIT2 (on all samples at once with either K = 100 (default) or 200 states as well as in N = 50,000 and 15,000 batches), imputing all masked genotypes in the process. (**a**) Accuracy of the imputed genotypes on the subset of 120,000 UK samples curated by UK Biobank for GWAS (~80% of all samples), stratified by MAF in those samples. (**b)** Accuracy of the imputed genotypes on subsets of samples defined by self-reported ancestry, stratified by MAF in those samples. The five largest ancestry groups in the data set were British (137,178 samples), Irish (3,977), "any other white background" (4,760), Indian (1,324), and Caribbean (1,028). The British and Irish results were nearly identical (**Supplementary Table 11**), so we did not plot Irish results to improve readability. For the ancestry groups with <5,000 samples, we plotted results only for MAF bins corresponding to an expected minor allele count of ≥2 among masked samples. Error bars, s.e.m. Numerical data are provided in **Supplementary Tables 9** and **11**.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Information for "Fast and accurate long-range phasing in a UK Biobank cohort"

Po-Ru Loh, Pier Francesco Palamara, Alkes L Price

Contents

Nature Genetics: doi:10.1038/ng.3571

Supplementary Note

The Eagle algorithm is overviewed in Online Methods. Here, we provide additional methodological details not fully described earlier. (To allow this note to be self-contained, we repeat some content provided in Online Methods, filling in details omitted earlier due to space limitations.)

Eagle proceeds in three main steps. The first and second step each iterate through all individuals in the data exactly once, updating each individual's phase in turn; the third step performs two such iterations. To help guide intuition, Figure 1 provides a snapshot of the progress of the algorithm after each step for our first N=150K phasing benchmark (Figure 2).

1 Direct IBD-based phasing using long IBD

For each proband in turn, Eagle scans all other (diploid) individuals for long genomic segments ($>$4cM) in which one (haploid) chromosome is likely to be shared IBD with the proband. Eagle then analyzes these probable IBD matches for consistency, identifies a consistent subset, and uses this subset to make phase calls. In our N=150K analyses, this step required $\approx$10% of the total computation time (Supplementary Table 2) and achieved near-perfect phasing within long swaths of genome covering most of each sample (corresponding to regions with IBD to several relatives) (Fig. 1a). In more detail, our algorithm applies the following four procedures to each proband in turn.

1.1 Detecting possible IBD: Scanning diploid genotypes for IBS$>$0 runs

First, we run a fast $O(MN)$-time scan against all other individuals for long runs of diploid genotypes containing no opposite homozygotes (i.e., IBS$>$0). This filtering procedure is expedient for analyses of very large data sets as it operates directly on diploid data and thus requires little computation; a few variations of the approach have previously been developed [41,42]. Our implementation achieves a very low constant factor in its running time by using bit operations to analyze blocks of 16–64 SNPs simultaneously and using dynamic programming to record the longest ten IBS$>$0 stretches starting at each SNP block. We partition SNPs into blocks as follows: moving sequentially across the genome, we initialize each new block to contain the next 16 SNPs. We then continue to add subsequent SNPs to the block until it either contains 64 SNPs or reaches a maximum span of 0.3cM; upon reaching either limit, we end the current block and begin the next block.

Nature Genetics: doi:10.1038/ng.3571

Pseudocode for IBS>0 scan.

```
INPUT:
- genoBits[][]: (# SNP blocks) x (# samples) matrix of bit mask pairs (is0, is2)
                each isX bit is set iff the genotype has allele count X
- proband: index of sample to use as proband
- numLong: number of longest IBS>0 runs to record for each start block
OUTPUT:
- topInds[][]: (# SNP blocks) x (numLong) matrix of sample indices
               records longest numLong IBS>0 runs starting at each block

WORK ARRAYS:
runStarts[] := (# samples) array: start of current IBS>0 run for each sample
                                  (algorithm iterates forward across the genome)
runStartFreqs[] := (# SNP blocks + 1) histogram (i.e., counts) of runStarts
runStartsNext[], runStartFreqsNext[]: storage arrays for updating the above

ALGORITHM:

N := (# samples)
B := (# SNP blocks)
runStarts[0..N-1] := 0           # initialize run starts and histogram
runStartFreqs[0] := N
runStartFreqs[1..B] := 0

for b = 0 to B-1                 # iterate forward across the genome

  runStartFreqsNext[0..B] := 0   # initialize histogram for next iteration

  for i = 0 to N-1               # iterate over samples
    if (genoBits[b][proband].is0 & genoBits[b][i].is2) |
         (genoBits[b][proband].is2 & genoBits[b][i].is0)   # bitwise opp-hom check
      runStartsNext[i] := b+1            # opposite homozygous sites => end run
    else
      runStartsNext[i] := runStarts[i]   # no opp-hom sites => continue run
    end if

    runStartFreqsNext[runStartsNext[i]]++
  end for

  for start = 0 to B
    if runStartFreqsNext[start] < numLong && runStartFreqs[start] >= numLong
      topInds[start][0..runStartFreqsNext[start]-1] :=
        all samples i with (runStartsNext[i] == start)   # runs continuing past b
      topInds[start][runStartFreqsNext[start]..numLong-1] :=
        subset of samples i with (runStarts[i] == start) # runs ending at b
```


Nature Genetics: doi:10.1038/ng.3571

```
    end if
  end for

  runStarts[0..N-1] := runStartsNext[0..N-1]       # prepare to advance 1 block
  runStartFreqs[0..B] := runStartFreqsNext[0..B]

end for
```

1.2 Detecting probable IBD: Estimating likelihoods of IBD or not IBD

Second, we compute an approximate likelihood ratio score for each potential IBD match identified by the above scan. This procedure is similar in spirit to Parente2 [43], which likewise computes approximate likelihood ratio scores to increase sensitivity and specificity of IBD calls. Our approach prioritizes speed over accuracy; instead of using a haplotype frequency model as in Parente2, we use only allele frequencies and LD Scores [44] to compute an approximate likelihood ratio for the observed match having occurred due to IBD versus by chance. We apply this procedure within a seed-and-extend framework in which we begin with long IBS>0 matches but consider extending them beyond IBS=0 sites (to tolerate genotyping errors). We record all extended matches with length >4cM and likelihood ratio $>10N$ (where N is the number of samples) as probable IBD matches.

In detail, for each long IBS>0 match between the proband and another sample identified by the scan (the "surrogate"), we first extend the match in each direction until we reach a SNP block containing ≥2 IBS=0 sites. As we extend the match in either direction, we keep track of the cumulative approximate log odds ratio for the match having arisen due to IBD (i.e., a shared haplotype) rather than by chance. We estimate the log odds at a given SNP m as

$$\text{approx log OR} = \text{crop}_{[\log P_{\text{err}}, -\log P_{\text{err}}]}\left(\frac{\log P(g_{\text{pro}} \mid g_{\text{sur}}, \text{IBD}) - \log P(g_{\text{pro}} \mid \text{no IBD})}{\text{LD Score}(m)}\right), \quad (2)$$

where:

- g_{pro} is the proband's genotype
- g_{sur} is the surrogate sample's genotype
- LD Score$(m) = \sum_{\text{SNPs } m' \text{ within 1cM of } m} r^2(m, m')$ (ref. [44]) roughly corrects for the redundant contributions of SNPs in LD [50, 51]
- $P(g_{\text{pro}} \mid \text{no IBD})$ is the probability of observing the proband's genotype by chance, i.e., the frequency of the (diploid) genotype g_{pro}

Nature Genetics: doi:10.1038/ng.3571

- $P(g_{\text{pro}} \mid g_{\text{sur}}, \text{IBD})$ is the probability of observing the proband's genotype conditional on sharing one haplotype with the surrogate sample

- $\text{crop}_{[\log P_{\text{err}}, -\log P_{\text{err}}]}$ denotes cropping the approximate odds ratio to be no more extreme in either direction than the chance of a genotype error (the constant P_{err}, 0.003 by default).

For a SNP in Hardy-Weinberg equilibrium with '1' allele frequency p and '0' allele frequency $1-p$, the probabilities $P(g_{\text{pro}} \mid \text{no IBD})$ and $P(g_{\text{pro}} \mid g_{\text{sur}}, \text{IBD})$ are as follows:

	$g_{\text{pro}} = 0$	$g_{\text{pro}} = 1$	$g_{\text{pro}} = 2$
$P(g_{\text{pro}} \mid \text{no IBD})$	$(1-p)^2$	$2p(1-p)$	p^2
$P(g_{\text{pro}} \mid g_{\text{sur}} = 0, \text{IBD})$	$1-p$	p	0
$P(g_{\text{pro}} \mid g_{\text{sur}} = 1, \text{IBD})$	$\frac{1-p}{2}$	$\frac{1}{2}$	$\frac{p}{2}$
$P(g_{\text{pro}} \mid g_{\text{sur}} = 2, \text{IBD})$	0	$1-p$	p

The approximate log odds ratio for a match is just the sum of per-SNP log odds ratios across SNPs in the match; thus, as we extend a match, we update its cumulative log odds ratio simply by adding the score of each successive SNP. We record the position in each direction at which the cumulative score is maximized, and we use these positions as the start and the end of the final match.

1.3 Identifying consistent IBD: Trimming and pruning

Third, we analyze the set of identified probable IBD matches for consistency, truncating or eliminating matches until we reach a consistent set. For any pair of overlapping probable IBD matches between the proband and potential surrogate parents 1 and 2, the implied shared haplotypes can be (a) consistent with the proband sharing the same haplotype with both surrogates 1 and 2, (b) consistent with the proband sharing one of its haploytpes with surrogate 1 and other with surrogate 2, or (c) inconsistent with both of these possibilities. We first identify pairs of overlapping probable IBD matches in which scenario (c) occurs; for these pairs, we assume the longer match is correct and trim the shorter match until consistency under either scenario (a) or (b) is achieved. If any match drops below 3cM after during this trimming procedure, we discard the match. At the end of the procedure, all remaining pairs of trimmed matches are consistent. We then perform a final check for global consistency of implied phase orientations among all matches, i.e., we reduce (if necessary) to a subset of matches that can each be assigned to either a surrogate maternal haplotype or a surrogate paternal haplotype in a manner that respects pairwise constraints (a) and (b).

Explicitly, for each pair of matches with nonempty intersection, we look for sites in their intersection at which the proband is heterozygous and both surrogates 1 and 2 are homozygous ("het-hom-hom sites"). If both surrogates are homozygous for the same allele, they map to the same haplotype (maternal or paternal) of the proband (situation (a) above); otherwise, they map to

Nature Genetics: doi:10.1038/ng.3571

opposite haplotypes (situation (b) above). In practice, we sometimes observe sites of both types (situation (c) above), indicating an error in at least one of the IBD calls (or a genotype error); typically, the reason is that one IBD call includes a true sub-region of IBD but extends beyond it. We deal with this situation by identifying the longest consistent sub-region in the intersection of the calls (i.e., the longest stretch of genome containing only het-hom-hom sites at which surrogates 1 and 2 are the same or only het-hom-hom sites at which surrogates 1 and 2 are opposite). We then trim the shorter of the two IBD calls until the intersection of the IBD calls contains only the consistent sub-region. (We trim the shorter IBD call because the longer IBD call is more likely to be correct.)

After trimming, we are left with a set of pairwise consistent IBD calls (between the proband and various surrogates), but there is still a chance that the set as a whole may not be consistent: each IBD call must ultimately map to either the proband's maternal haplotype or the proband's paternal haplotype, and this mapping must be simultaneously consistent for all pairs of calls. In graph theoretic language, if we let the IBD calls be vertices of a graph, then we wish to bicolor the graph while respecting same-color constraints (represented by one set of edges connecting pairs of calls in situation (a) above) and opposite-color constraints (represented by another set of edges connecting pairs of calls in situation (b) above). Checking for the existence of a valid coloring requires only a search through the graph. Thus, we prune our set of trimmed IBD calls to a globally consistent subset by starting with the empty subset and iteratively attempting to add each IBD call to the set; we iterate through the IBD calls from longest to shortest. At each iteration, we run a breadth-first search to check whether the augmented subset is still globally consistent; if so, we augment the subset, and if not, we discard the IBD call.

Pseudocode for pruning trimmed IBD calls to a consistent subset.

```
INPUT:
- IBDcalls[]: list of IBD calls (longest to shortest)
- constraints[][]: pairwise sign constraints among IBD calls (1=same, -1=opp)

OUTPUT:
- IBDpruned[]: pruned list of IBD calls consistent with constraints

ALGORITHM:

IBDpruned := []      # initialize list of consistent IBD calls
for u in IBDcalls    # iterate through IBD calls (longest to shortest)
  IBDpruned.insert(u)
  if checkSigns(IBDpruned,constraints) == false
    IBDpruned.erase(u)
  end if
end for
```


Nature Genetics: doi:10.1038/ng.3571

```
###

function checkSigns(IBDpruned,constraints)
  signs[..] := 0             # initialize signs to 0; signs will become 1 or -1
  q := []                    # initialize breadth-first search queue
  for u in IBDpruned
    if signs[u] == 0         # IBD call u has not been processed
      q.push(u)
      while !q.empty()       # breadth-first search
        v := q.pop()
        for w in constraints[v]
          if signs[w] != 0   # IBD call w has been processed: check consistency
            if constraints[v][w] != signs[v] * signs[w]
              return false   # failure: inconsistency found
            end if
          else               # IBD call w has not been processed
            signs[w] := signs[v] * constraints[v][w]
            q.push(w)
          end if
        end for
      end while
    end if
  end for
  return true   # success: no inconsistencies found
end
```

1.4 Making phase calls: Weighing the evidence

Fourth, we use the surrogate maternal and paternal haplotypic assignments of probable IBD regions to make phase calls. Whenever at least one surrogate is homozygous at a proband het, we use that surrogate to phase the site. (If homozygous surrogates disagree on the phasing of a site, we always defer to the longest surrogate with longest IBD to the proband.) If all surrogates are heterozygous, we make a probabilistic phase call based on the allele frequency of the SNP and the difference between the numbers of (heterozygous) surrogate maternal haplotypes (n_{mat}) and surrogate paternal haplotypes (n_{pat}). Specifically, let p be the minor allele frequency of the SNP. If we condition on the proband's maternal allele being the minor allele, then the probability of observing hets in all n_{mat} maternal surrogates is $(1-p)^{n_{\mathrm{mat}}}$ and the probability of observing hets in all n_{pat} paternal surrogates is $p^{n_{\mathrm{pat}}}$ (assuming only one haplotype is shared per IBD match and non-shared haplotypes are independent). If we condition on the proband's paternal allele being the minor allele, the probabilities are $p^{n_{\mathrm{mat}}}$ and $(1-p)^{n_{\mathrm{pat}}}$. Thus, the odds ratio of the minor allele being maternal vs. paternal is $((1-p)/p))^{n_{\mathrm{mat}}-n_{\mathrm{pat}}}$. We randomly hard-call step 1 phase according

Nature Genetics: doi:10.1038/ng.3571

to this odds ratio, and we also record the call probability as an estimate of phase confidence to use in step 2 along with the hard call.

Finally, we note one additional subtlety: occasionally, the proband may share both haplotypes IBD with a surrogate (e.g., a sibling). In such situations, hets in the surrogate provide no probabilistic information for phasing hets in the proband. Fortunately, regions of double-IBD are easy to identify with high sensitivity and specificity (as the diploid genotypes must exactly match); we use an approximate likelihood ratio score similar to our approach for calling single-IBD, and we exclude likely double-IBD regions from the calculation above.

2 Local phase refinement using long and short IBD

For each diploid proband in turn, Eagle analyzes overlapping $\approx$1cM windows of genome, searching for pairs of haplotypes (from the output of step 1) that approximately sum to the diploid proband within the window. Eagle then makes phase calls according to the haplotype pairs that most closely match the proband. In our N=150K analyses, this step required $\approx$20% of the total computation time (Supplementary Table 2) and reduced the switch error rate to $\approx$1.5% (Fig. 1b). In more detail, our algorithm applies the following three procedures to each proband in turn.

2.1 Detecting diploid-haploid long IBD: Scanning for IBS>0 runs

First, we run a fast $O(MN)$-time scan to find probable IBD with other haploid chromosomes (according to phase calls made in step 1). This procedure begins analogously to the first component of step 1; again, we look for long segments of IBS>0 (now between the diploid proband and haploid potential surrogates), now allowing a single mismatch site (IBS=0) within runs. We then attempt to extend the identified seed matches and record the ten longest matches covering each SNP block (as defined in step 1).

Explicitly, we run a fast $O(MN)$-time scan between the (diploid) proband and the $2N-2$ haploid chromosomes of the remaining $N-1$ samples in the cohort (according to the hard-called phase from step 1, with random phase calls in segments lacking IBD); we ignore the phase calls made for the proband in step 1. At each SNP block, we identify the 20 samples with the longest runs of IBS>0 to the proband (allowing $\leq$1 error) starting exactly at that block (i.e., with IBS=0 in the previous block). (Here, IBS>0 is equivalent to IBS=1 because we are comparing the proband to haploid surrogates.) We treat the identified matches as seeds, and we extend each seed forward and backward until we reach a block containing $\geq$4 IBS=0 sites (among the 16–64 sites in the block; most blocks have SNP counts in the upper end of this range). The idea behind this extension procedure is to retain sensitivity despite errors in the step 1 phase calls; even in well-phased regions (with IBD to many surrogates), step 1 phasing is error-prone at common SNPs for which

Nature Genetics: doi:10.1038/ng.3571

all surrogates are hets. Finally, among all extended IBD matches between the proband and haploid surrogates produced in this manner, we record for each block the longest 10 extended matches covering that block.

2.2 Finding complementary short IBD: Locality-sensitive hashing

Second, for each window of three consecutive blocks (containing a total of up to 192 SNPs spanning up to 0.9cM), and for each of the ten longest haplotype matches covering the center block in the window, we search for haplotypes approximately complementary (within the window) to the long haplotype. The idea is that often, only one of the proband's haplotypes belongs to a long IBD tract (several cM); however, in such cases, the other haplotype is often shared in a short IBD tract ($\approx$1 cM), allowing confident phase inference if the complementary haplotype can be found to exist. Looking for a complementary haplotype in an error-tolerant manner amounts to performing approximate nearest neighbor search in Hamming space; to do so, we apply locality-sensitive hashing (LSH) [45, 46]. In brief, LSH overcomes the "curse of dimensionality" by building multiple hash tables (here, ten per window) using different random subsets of SNPs (here, up to 32); then, when searching for a complementary haplotype, chances are high that at least one hash table will not include any SNPs with errors, allowing the approximate match to be found.

Explicitly, for each 3-block window, we build 10 hash tables containing B = 23, 24, ..., 32 SNPs selected independently at random from all MAF$\geq$2% SNPs in the window. For each hash table, we hash each of the $2N$ haplotypes called in step 1 as a B-bit string encoding major/minor allele status at the selected B SNPs. For memory efficiency, we store at most 99 haplotype indices per hash key; if $>$99 haplotypes hash to the same key (which occurs for common haplotypes), we store a random 99-element subset of these haplotype indices. Because the hash table is static once created, we further optimize memory by storing occurring keys in a sorted array, each of which contains a pointer to the list of haplotype index values corresponding to the key; to perform hash lookups, we run binary searches on the sorted key array. The total number of bytes required by this implementation is $12K + 4V$, where K is the number of keys and V the number of stored values.

2.3 Making phase calls: Settling disagreements and linking blocks

Third, for each 3-block window, we select the lowest-error complementary haplotype pair for that window (i.e., the pair of surrogate haploid parents—one found via long IBD and the other identified by LSH—with fewest conflicts between the sum of the haploid surrogates vs. the diploid proband over the 3-block window). We use this surrogate parental pair to phase the block in the center of the window. This procedure is fairly straightforward, with the only subtleties being that (i) to avoid simply copying phase from double-IBD matches, we require the surrogate haploid parents to be derived from distinct individuals; (ii) to phase error hets (i.e., proband hets for which both

Nature Genetics: doi:10.1038/ng.3571

surrogate haplotypes have the same allele), we defer to the surrogate with higher confidence (using the call probabilities saved from step 1); and (iii) when transitioning from one block to the next, we choose the orientation of the next complementary haplotype pair that best continues the current surrogate maternal and paternal haplotypes. More precisely, for (iii), we identify the five proband-het SNPs closest to the block transition and compare how these SNPs are phased by the surrogate parental pair for the previous block vs. the next block; we then decide which orientation to use for the next surrogate parental pair relative to the previous based on majority vote (over the five SNPs near the transition).

3 Approximate HMM decoding

For each diploid proband in turn, Eagle identifies candidate surrogate parental haplotypes (from the output of step 2) for use within an HMM (similar to the Li-Stephens model [47]). Eagle then computes an approximate maximum likelihood path through the HMM using a modified Viterbi algorithm (aggressively pruning the state space to increase speed) and calls phase according to the HMM decoding. Finally, Eagle post-processes the phase calls to correct sporadic errors by explicitly taking into account haplotype frequencies and long IBD. Eagle runs two iterations of this entire procedure, and Eagle performs each iteration in 10 batches of N/10 samples, updating hard-called haplotypes available as surrogates and all derivative data (e.g., hash tables) after each batch. In our N=150K analyses, this step required ≈70% of the total computation time (Supplementary Table 2) and reduced the switch error rate to ≈0.4% after the first HMM iteration and ≈0.3% after the second (Fig. 1c,d). In more detail, our algorithm applies the following three procedures to each proband in turn (in each HMM iteration).

3.1 Identifying surrogate parents: Scanning for long IBD and complements

First, we compile a set of reference haplotypes for the proband for each SNP block. This procedure begins analogously to the first component of step 2, identifying long haplotype matches using a fast $O(MN)$ search within a seed-and-extend framework. To ensure that both maternal and paternal surrogates are represented among the reference haplotypes, we augment the set of long haplotype matches with complementary haplotypes found using LSH. In total, we store $K \leq 80$ reference haplotypes per block.

In more detail, we begin by running the fast $O(MN)$ diploid-haploid IBS>0 search algorithm used in step 2 (on updated haploid chromosomes corresponding to current phase calls); in the first iteration of this step, we record the 100 longest 1-err IBS>0 runs starting at each block, and in the second iteration, we additionally record the 100 longest 0-err IBS>0 runs starting at each block. We then apply the seed extension algorithm described in step 2 with a more stringent extension

Nature Genetics: doi:10.1038/ng.3571

criterion: we extend only until we reach a block with $\geq$2 IBS=0 sites.

The matches identified above serve as a starting point for constructing a set of $K \leq 80$ reference haplotypes specific to each block (for subsequent use in HMM decoding on the proband); for each block b, we construct the final reference set as follows. First, we include a total of up to 20 haplotypes from among the longest 1-err (or in the second iteration, 1-err and 0-err) IBS>0 runs starting at block b or block $b+1$. Second, we augment the reference set with the longest extended matches covering block b until we reach a total of 40 references or we run out of extended matches covering block b. Third, for each of the $\leq$40 reference haplotypes selected thus far, we attempt to find another haplotype that is exactly complementary to it within the region starting from a random SNP within block b through the end of block $b+2$ (or as far as possible if no such haplotype exists). We do so by using LSH as in step 2 with slightly different parameters aimed at increasing sensitivity to find matches on shorter scales: we hash SNP sets spanning intervals ranging from 3 blocks (as in step 2) down to only 1 block, and we include four additional hash tables. The overall intuition is that the first two groups of references include haplotypes with the longest IBD possible to the proband, while the third group ensures that at least short-range surrogates are available for both the maternal and paternal chromosomes (even if one side lacks IBD).

3.2 Finding a parsimonious path: Approximate HMM decoding

Second, we compute an approximate Viterbi decoding of an HMM similar to the Li-Stephens model [47] using the sets of local reference haplotypes found above. A path through the HMM consists of a sequence of state pairs (one maternal reference haplotype and one paternal reference haplotype) at each location; we score a path according to the number of transitions on the maternal side, the number of transitions on the paternal side, and the number (and types) of Mendel errors between the proband and surrogate parents. An exact Viterbi decoding of this HMM using dynamic programming requires $O(MK^3)$ time, which is too expensive for us; instead, we perform the dynamic programming within a beam search, pruning the search space from K^2 state pairs to the top P=100–200 state pairs at each location (100 in the first decoding iteration, 200 in the second) and thus limiting the complexity to $O(MKP)$. We then phase the proband according to the approximate Viterbi path.

In more detail, we first consider the diploid analog of the original Li-Stephens model [47] in which HMM states are ordered pairs of haplotype indices (each selected from among the $2N-2$ non-proband haplotypes, for a total of $O(N^2)$ possible state pairs) and we wish to find the maximum likelihood path (i.e., sequence of state pairs) through the M SNPs being phased. This computation can be performed using the Viterbi algorithm in $O(MN^4)$ time if we naively allow all-to-all transitions between state pairs and $O(MN^3)$ time if we allow only transitions in either the maternal or the paternal references (but not both) from one SNP to the next. Either way, the full computation is far too expensive for large N and M, so we make several approximations. First,

Nature Genetics: doi:10.1038/ng.3571

instead of performing the full Viterbi dynamic programming search, we perform a beam search: at each position, we prune the search to the top (most likely) P state pairs. Second, at each position, instead of considering transitions to all $2N-2$ reference haplotypes, we only consider transitions to the K references selected above for that position. These approximations reduce the computational complexity to $O(MKP)$. Finally, for a further constant-factor reduction in cost, we perform computations in blocks of 16–64 SNPs as elsewhere, allowing only one state transition (for either the maternal or paternal reference but not both) within each block.

The details of the model—i.e., score penalties for transitions and Mendel errors, equivalent to HMM transition and emission probabilities—are as follows. We assess a score penalty of 3 for each transition between references, a penalty of 2 for Mendel errors at proband hets, and a penalty of 1 for Mendel errors at non-hets. Under this very basic model, we observed that the best-scoring path already yielded very accurate phase, but we noticed a tendency for occasional switch errors to occur near block boundaries when the best-scoring path included transitions in both the maternal and paternal references in rapid succession, one at the end of block b and the other at the beginning of block $b+1$. We therefore added a penalty for transitions near block boundaries. (The engineering details are a bit complex, but the penalty is roughly equivalent to an added penalty of 3 for transitions within 4 SNPs of a boundary, 2 for transitions within 8 SNPs of a boundary, and 1 for transitions within 12 SNPs of a boundary; for details, see the `computeSwitchScore()` function in the Eagle code.)

Overall, the Eagle HMM is very rudimentary compared to the HMMs used by advanced HMM-based methods such as Beagle [8], HAPI-UR [11], and SHAPEIT2 [12]. For phasing very large samples containing long IBD, our intuition is that precise probabilistic modeling is unnecessary: once long IBD has been identified, the right phasing should be fairly obvious even to a crude model, and the key is to rapidly identify and use such IBD. Our approach is optimized for this purpose; the approximations we use lend themselves to fast (approximate) Viterbi decoding rather than careful MCMC sampling.

3.3 Cleaning up errors: Using haplotype frequencies and respecting IBD

Third, we post-process the phase calls to correct sporadic errors. Within each window of three consecutive blocks, we use LSH to determine the frequencies of ≈1cM haplotypes that match the Viterbi-inferred maternal and paternal haplotypes up to at most two errors. In rare cases, the haplotype frequencies give strong evidence to flip the phase of one or two SNPs, in which case we override the Viterbi phase call. Finally, we also check the Viterbi-inferred maternal and paternal haplotypes for consistency with the longest previously-identified IBD segments; in rare cases when the Viterbi phasing requires a phase switch >1.5cM from either end of a probable IBD segment, we override the switch.

Explicitly, for each block, we run LSH queries (in the 3-block window centered at that block)

Nature Genetics: doi:10.1038/ng.3571

for the hash keys of the maternal and paternal haplotypes under the Viterbi phasing. The queries return haplotypes matching the Viterbi parental haplotypes at the 23–32 SNPs used in each hash table; however, these haplotypes may differ from the input Viterbi haplotypes at remaining SNPs. We record haplotypes differing from the input Viterbi haplotypes at ≤ 2 sites, and at each proband het, we use these haplotypes to generate frequency tables for the ref/alt allele according to the near-matches to each Viterbi haplotype. We then compute an odds ratio for keeping vs. flipping the Viterbi phasing of the proband het; if the odds ratio exceeds a threshold of 10, we flip the phasing. (We reduce the threshold to 2 at Mendel error SNPs.)

Nature Genetics: doi:10.1038/ng.3571

4 Appendix: In-sample imputation accuracy

To project the imputation accuracy that will be achievable in the UK population using LRP-based methods once a reference panel of N=150K sequenced UK samples becomes available, we performed in-sample imputation of masked genotypes in the UK Biobank data set. Explicitly, we randomly masked 2% of all genotypes, phased the modified data set (automatically obtaining imputed genotypes at masked SNPs), and assessed concordance between imputed and actual genotypes. This procedure is commonly used to assess accuracy of phasing methods [1, 9, 10], and for very large sample sizes, enough genotypes are masked per SNP (here, ≈3,000) that R^2 between imputed and actual genotypes can be assessed across the minor allele frequency (MAF) spectrum (e.g., a 0.1% variant is expected to have a minor allele count of 6 among 3,000 masked genotypes). We note that from an engineering perspective, in-sample imputation differs from standard GWAS imputation in a few important ways (detailed below); however, from a statistical perspective, in-sample imputation on N samples is similar to standard GWAS phasing and imputation on a target sample using a reference panel of size N: both tasks entail copying shared haplotypes (identified based on data at typed SNPs) from a set of N samples (Supplementary Fig. 1).

We benchmarked in-sample imputation using Eagle and SHAPEIT2 (the two most accurate phasing algorithms according to our previous benchmarks). For Eagle, we imputed all N=150K samples together (Eagle 1x150K), and for SHAPEIT2, we performed imputation in 10 batches of N=15K samples (SHAPEIT2 10x15K), 3 batches of N=50K samples (SHAPEIT2 3x50K), or in a single batch of all N=150K samples (SHAPEIT2 1x150K) using either default parameters (K=100) or twice the default number of conditioning states (K=200). We then assessed imputation R^2 stratified by MAF, first focusing on accuracy within N=120K genetically homogeneous samples curated by UK Biobank for GWAS (a subset of the 88% of samples who self-reported British ethnicity; see Online Methods and URLs). We observed that both Eagle and SHAPEIT2 1x150K analyses achieved mean in-sample imputation R^2>0.75 down to a MAF of 0.1%, with Eagle slightly more accurate than SHAPEIT2 K=100 across all MAF bins and of similar accuracy to SHAPEIT2 K=200 (Supplementary Fig. 2a and Supplementary Table 9); in contrast, SHAPEIT2 10x15K analysis achieved R^2<0.6 for MAF 0.1%-variants. We confirmed these results in chromosome-scale analyses as before (Supplementary Table 10).

We further investigated in-sample imputation performance of Eagle and SHAPEIT2 as a function of self-reported ethnicity. As UK Biobank genotyping and QC analyses indicated that self-reported ethnicity aligned closely with genetic ancestry (see URLs), we stratified our in-sample imputation assessment by self-reported ethnicity (Supplementary Fig. 2b and Supplementary Table 11). We observed that in-sample imputation R^2 for British and Irish samples (comprising 88% and 3% of the samples) closely matched our previous results, as expected, while accuracy was lower (but still slightly higher for Eagle vs. SHAPEIT2 1x150K, K=100 analyses) in samples who

Nature Genetics: doi:10.1038/ng.3571

reported "any other white background" (3%). Accuracy was lowest in non-white samples, and in these samples, SHAPEIT2 1x150K achieved slightly higher in-sample imputation accuracy than Eagle, as expected for low amounts of IBD. Consistent with these findings, we observed a modest decrease in in-sample imputation R^2 across all methods (with little relative change between methods) when evaluated on all N=150K UK Biobank samples versus the N=120K curated British samples in our main analyses (Supplementary Tables 9 and 10).

As noted above, some caution is warranted in interpreting these results, as in-sample imputation of missing data distributed across SNPs generally does not arise in GWAS (except in the context of low-coverage sequencing [52–54]). Standard GWAS imputation differs from in-sample imputation in three ways (Supplementary Fig. 1). First, GWAS imputation usually involves imputing sequence data from a reference panel into a (genotyped but not sequenced) target sample, which typically requires phasing the sequenced reference (possibly using read information [37]), phasing the target sample (possibly using the phased reference), and imputing reference data into the target sample; here, we have only one N=150K sample as both target and reference that we simultaneously phase and impute. Second, GWAS imputation pipelines produce probabilistic allele "dosage" estimates, whereas phasing methods produce hard calls at missing genotypes; thus, R^2 using imputed allele dosages is expected to be even higher. Third, typical GWAS impute sequenced SNPs into target samples that are fully typed at a set of ascertained array SNPs; here, we imputed masked data in $\approx$98%-typed array SNPs. (The latter task may be slightly harder than the former, as genotyping arrays are sometimes optimized to minimize redundancy among ascertained SNPs [55]; additionally, phasing methods may not be optimized for analysis of genotype data with a uniform 2% missing rate. On the other hand, the fact that rare variants on genotyping arrays are typically enriched in densely-typed fine-mapping regions may make in-sample imputation easier.) For all of these reasons, different algorithms are typically used for phasing vs. GWAS imputation (e.g., SHAPEIT [10, 12] vs. IMPUTE [2, 56], MaCH [4] vs. minimac [5, 57]). Despite these caveats, our results give reason for optimism that when sequenced ancestry-matched reference panels of size N=150K become available, high-accuracy imputation of rare variants will be possible using LRP-based approaches such as Eagle: we expect that efficient imputation of MAF>0.1% variants at R^2>0.75 will be possible using Eagle and appropriate extensions (see Discussion).

Nature Genetics: doi:10.1038/ng.3571

References

1. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**, 703–714 (2011).

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

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

4. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816–834 (2010).

5. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* **44**, 955–959 (2012).

6. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**, 449–462 (2005).

7. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**, 629–644 (2006).

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

9. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210–223 (2009).

10. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2012).

11. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* **91**, 238–251 (2012).

12. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).

13. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* **40**, 1068–1075 (2008).

Nature Genetics: doi:10.1038/ng.3571

14. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744–747 (2009).

15. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).

16. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).

17. Thorleifsson, G. *et al.* Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. *Nature Genetics* **42**, 906–909 (2010).

18. Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics* **43**, 316–320 (2011).

19. Rafnar, T. *et al.* Mutations in BRIP1 confer high risk of ovarian cancer. *Nature Genetics* **43**, 1104–1107 (2011).

20. Gudmundsson, J. *et al.* Discovery of common variants associated with low TSH levels and thyroid cancer risk. *Nature Genetics* **44**, 319–322 (2012).

21. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics* **44**, 1326–1329 (2012).

22. Helgason, H. *et al.* A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. *Nature Genetics* **45**, 1371–1374 (2013).

23. Kong, A. *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics* **46**, 11–16 (2014).

24. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature Genetics* **46**, 294–298 (2014).

25. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* **47**, 435–444 (2015).

26. Steinberg, S. *et al.* Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nature Genetics* (2015).

27. Helgason, H. *et al.* Loss-of-function variants in ATM confer risk of gastric cancer. *Nature Genetics* **47**, 906–910 (2015).

28. Palin, K., Campbell, H., Wright, A. F., Wilson, J. F. & Durbin, R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic Epidemiology* **35**, 853–860 (2011).

29. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLOS Genetics* **10**, e1004234 (2014).

Nature Genetics: doi:10.1038/ng.3571

30. Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* **12**, 1–10 (2015).

31. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19**, 318–326 (2009).

32. Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics* **88**, 173–182 (2011).

33. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

34. Banda, Y. *et al.* Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* **200**, 1285–1295 (2015).

35. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. *American Journal of Human Genetics* **98**, 456–472 (2016).

36. O'Connell, J. *et al.* Haplotype estimation for biobank scale datasets. *Nature Genetics* (in press).

37. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *American Journal of Human Genetics* **93**, 687–696 (2013).

38. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).

39. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics* **98**, 116–126 (2016).

40. Chen, C.-Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).

41. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS ONE* (2012).

42. Huang, L., Bercovici, S., Rodriguez, J. M. & Batzoglou, S. An effective filter for IBD detection in large datasets. *PLOS ONE* **9**, e92713 (2014).

43. Rodriguez, J. M., Bercovici, S., Huang, L., Frostig, R. & Batzoglou, S. Parente2: a fast and accurate method for detecting identity by descent. *Genome Research* **25**, 280–289 (2015).

44. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015).

45. Indyk, P. & Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM Symposium on Theory of Computing*, 604–613 (ACM, 1998).

Nature Genetics: doi:10.1038/ng.3571

46. Gionis, A., Indyk, P. & Motwani, R. Similarity search in high dimensions via hashing. In *Proceedings of the 25th VLDB Conference*, vol. 99, 518–529 (1999).

47. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

48. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16 (2015).

49. Kvale, M. N. *et al.* Genotyping informatics and quality control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051–1060 (2015).

50. Zou, F., Lee, S., Knowles, M. R. & Wright, F. A. Quantification of population structure using correlated SNPs by shrinkage principal components. *Human Heredity* **70**, 9–22 (2010).

51. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**, 1011–1021 (2012).

52. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research* (2011).

53. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics* **44**, 631–635 (2012).

54. Cai, N. *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).

55. Hoffmann, T. J. *et al.* Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89 (2011).

56. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics* **5**, e1000529 (2009).

57. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 1. Computational cost and accuracy of phasing methods.

(a) Running time and memory cost

N	Eagle	SHAPEIT2	HAPI-UR	Beagle
15K	0.5 hr / 0.9 GB	4.2 hr / 1.7 GB	2.8 hr / 6.3 GB	19.5 hr / 9.9 GB
50K	2.7 hr / 2.6 GB	28.2 hr / 5.5 GB	18.9 hr / 18.1 GB	335.3 hr / 21.4 GB
150K	15.0 hr / 7.0 GB	207.6 hr / 16.5 GB	181.0 hr / 48.1 GB	–

(b) Switch error rate

N	Eagle	SHAPEIT2	HAPI-UR	Beagle
15K	1.50% (0.094%)	1.15% (0.075%)	1.94% (0.09%)	1.50% (0.074%)
50K	0.628% (0.048%)	0.564% (0.05%)	1.26% (0.077%)	0.896% (0.059%)
150K	0.308% (0.034%)	0.303% (0.035%)	0.755% (0.063%)	–

(c) Running time for phasing N=150K samples in batches

Batches	Eagle	SHAPEIT2	HAPI-UR	Beagle
10x15K	0.2 days	1.8 days	1.2 days	8.1 days
3x50K	0.3 days	3.5 days	2.4 days	41.9 days
1x150K	0.6 days	8.7 days	7.5 days	–

(This table provides numeric data plotted in Figure 2.) We benchmarked Eagle and existing phasing methods on N=15K, 50K, and 150K UK Biobank samples and M=5,824 SNPs on chromosome 10. (**a**) Run times and memory are reported for runs using up to 10 cores of a 2.27 GHz Intel Xeon L5640 processor and up to two weeks of computation. (**b**) Mean switch error rates (s.e.m.) are over 70 European-ancestry trios. (**c**) Run times for phasing N=150K samples in 10 batches of 15K samples, 3 batches of 50K samples, and 1 batch of 150K samples (i.e., 10x, 3x, and 1x the run times reported in (a)). All methods except HAPI-UR supported multithreading. As the HAPI-UR documentation suggested merging results from three independent runs with different random seeds, we parallelized these runs across three cores. (For the N=150K analysis, HAPI-UR encountered a failed assertion bug for some random seeds, so we needed to try six random seeds to find three working seeds. We did not count this extra work against HAPI-UR.)

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 2. Detailed run time breakdown of Eagle at varying sample sizes.

(a) N=15K samples

	Step 1	Step 2	Step 3	Total
$O(MN^2)$ component	0.5 min	1.1 min	2.6 min	4.2 min
Other computation	0.3 min	0.9 min	23.9 min	25.1 min
Total	0.9 min	1.9 min	26.5 min	29.2 min

(b) N=50K samples

	Step 1	Step 2	Step 3	Total
$O(MN^2)$ component	6.7 min	13.3 min	30.3 min	50.3 min
Other computation	2.5 min	5.0 min	103.4 min	110.9 min
Total	9.2 min	18.3 min	133.8 min	161.2 min

(c) N=150K samples

	Step 1	Step 2	Step 3	Total
$O(MN^2)$ component	66.2 min	153.3 min	290.7 min	510.2 min
Other computation	23.5 min	21.8 min	336.4 min	381.6 min
Total	89.7 min	175.1 min	627.1 min	891.9 min

These table provides detailed run time breakdowns of Eagle's three algorithmic steps in runs on N=15K, 50K, and 150K UK Biobank samples and M=5,824 SNPs on chromosome 10. Run times are reported for runs using up to 10 cores of a 2.27 GHz Intel Xeon L5640 processor. Each of the three steps—(1) direct IBD-based phasing, (2) local phase refinement, and (3) approximate HMM decoding—involve an all-pairs $O(MN^2)$ computation followed by an additional computation that is inexpensive for step 1 and scales closer to linearly in sample size (N) for steps 2 and 3. (The tables above do not include time needed to write output, which increases total run times by $\approx 1\%$.)

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 3. Phasing performance on GERA data.

Method	Run time	Memory	Switch error rate
Eagle	15.8 hr	15.7 GB	0.820% (0.035%)
SHAPEIT2	229.7 hr	44.4 GB	0.704% (0.033%)

We phased chromosome 10 (32,741 SNPs) for N=60K European-ancestry GERA individuals using 10 cores of a 2.27 GHz Intel Xeon L5640 processor. We report mean switch error rate (s.e.m.) on 197 children from European-ancestry trios in independent pedigrees.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 4. List of 10,000-SNP regions analyzed in Eagle and SHAPEIT2 *N*=150K analyses.

Chromosome	Base pair range (hg19)	Physical span	Genetic span
1	157.1–204.1 Mb	47.0 Mb	51.5 cM
2	204.7–243.0 Mb	38.3 Mb	58.8 cM
4	53.2–106.8 Mb	53.7 Mb	49.1 cM
6	0.2–29.8 Mb	29.6 Mb	50.0 cM
7	72.0–127.1 Mb	55.1 Mb	48.1 cM
9	78.7–121.3 Mb	42.7 Mb	54.7 cM
11	69.6–117.2 Mb	47.6 Mb	49.3 cM
14	19.3–65.2 Mb	46.0 Mb	59.0 cM
16	58.3–90.2 Mb	31.9 Mb	55.2 cM
19	28.3–59.1 Mb	30.8 Mb	57.6 cM

We defined these ten regions by (i) listing all SNPs in order from chromosome 1–22, (ii) splitting this list into 10 chunks, and (iii) selecting the 10,000-SNP region in the middle of each chunk (shifting the region if necessary to avoid crossing chromosome boundaries or centromeres).

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 5. Distributions of discrepancy counts in 10Mb segments phased using Eagle and SHAPEIT2 on *N*=150K samples.

	Percentage of 10Mb segments with specified number of discrepancies					
Method	0	1	2	3	4	≥5
Eagle `--fast`	61.2% (1.9%)	14.6% (0.7%)	3.6% (0.4%)	0.9% (0.2%)	0.6% (0.2%)	19.2% (1.5%)
Eagle	63.3% (1.9%)	15.6% (0.6%)	2.7% (0.4%)	0.8% (0.1%)	0.5% (0.1%)	17.1% (1.5%)
SHAPEIT2 K=100 (3 blocks)	56.9% (1.5%)	12.3% (0.7%)	1.9% (0.3%)	0.8% (0.2%)	0.9% (0.2%)	27.1% (1.2%)
SHAPEIT2 K=200 (4 blocks)	63.5% (1.6%)	12.4% (0.8%)	1.8% (0.3%)	1.0% (0.2%)	0.4% (0.1%)	20.8% (1.0%)
SHAPEIT2 K=400 (5 blocks)	64.9% (1.5%)	13.3% (0.8%)	2.2% (0.3%)	0.6% (0.1%)	0.5% (0.2%)	18.4% (1.0%)

(This table provides detailed discrepancy distributions for the analyses presented in Table 1.) We benchmarked various parameter settings of Eagle and SHAPEIT2 in ten analyses of 10,000-SNP regions (Supplementary Table 4), phasing all N=150K UK Biobank samples in each analysis. We partitioned SHAPEIT2 analyses into 3, 4, or 5 blocks (with an overlap of 500 SNPs) as necessitated by computational constraints; we ligated SHAPEIT2 output using hapfuse v1.6.2. The number of discrepancies within a 10Mb segment is defined as the minimum number of SNPs with incorrect phase when comparing a phased haplotype to either trio-phased haplotype [13]. Percentages of 10Mb segments with a given number of discrepancies are means (s.e.m.) over the ten 10,000-SNP regions. Within each 10,000-SNP region, we computed the distribution of discrepancies across the 70 European-ancestry trios on as many disjoint 10Mb segments as could fit in the region while leaving a 1Mb buffer on each end. That is, for a 10,000-SNP region of length L, we considered $70\lfloor(L-2)/10\rfloor$ segments.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 6. Computational cost and accuracy of Eagle on *N*=150K samples using additional parameter settings.

Method	Run time	Switch error rate	Switch error rate
Eagle `--fast`	2.8 days	0.317% (0.012%)	0.152% (0.012%)
Eagle	5.0 days	0.272% (0.009%)	0.118% (0.007%)
Eagle 2x HMM beam width	6.2 days	0.285% (0.012%)	0.123% (0.009%)
Eagle w=0.2cM	7.3 days	0.276% (0.011%)	0.118% (0.008%)
Eagle w=0.2cM, 2x HMM beam width	9.2 days	0.266% (0.011%)	0.108% (0.008%)

(This table provides benchmark results analogous to Table 1 for a few additional parameter settings of Eagle; the first two rows of the two tables are shared.) We benchmarked various parameter settings of Eagle in ten analyses of 10,000-SNP regions (Supplementary Table 4), phasing all N=150K UK Biobank samples in each analysis. Switch error rates are means (s.e.m.) over the ten regions, assessed on 70 European-ancestry trios. Switch error rates without blips ignore switches arising when 1–2 SNPs are oppositely phased relative to $\geq$10 consistently phased SNPs on both sides. The parameter settings tested are as follows:

- Eagle `--fast`: larger limit on SNP block span (w=0.5cM); reduced approximate HMM search (see Online Methods for details)
- Eagle: default w=0.3cM limit on SNP block span; default approximate HMM search
- Eagle 2x HMM beam width: default w=0.3cM; twice as many states in approximate HMM beam search width
- Eagle w=0.2cM: w=0.2cM; default approximate HMM search
- Eagle w=0.2cM, 2x HMM beam width: w=0.2cM; twice as many states in approximate HMM beam search width.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 7. Computational cost and accuracy of Eagle and SHAPEIT2 on *N*=150K samples using various parameters.

Method	Run time	Switch error rate	Switch error rate without blips
Eagle `--fast`	1.4 days	0.342% (0.016%)	0.180% (0.013%)
Eagle	2.6 days	0.284% (0.015%)	0.130% (0.010%)
SHAPEIT2 W=2Mb, K=100 (3 blocks)	50.1 days	0.306% (0.022%)	0.167% (0.015%)
SHAPEIT2 W=2Mb, K=200 (4 blocks)	56.7 days	0.272% (0.025%)	0.133% (0.014%)
SHAPEIT2 W=2Mb, K=400 (5 blocks)	69.7 days	0.248% (0.018%)	0.111% (0.008%)
SHAPEIT2 W=4Mb, K=100 (3 blocks)	33.5 days	0.398% (0.051%)	0.244% (0.038%)
SHAPEIT2 W=4Mb, K=200 (4 blocks)	40.4 days	0.350% (0.036%)	0.200% (0.023%)
SHAPEIT2 W=4Mb, K=400 (5 blocks)	54.4 days	0.292% (0.019%)	0.151% (0.014%)

(This table is analogous to Table 1 and includes benchmark data for SHAPEIT2 run with a window size of 4Mb on a pilot subset of five 10,000-SNP regions.) We benchmarked various parameter settings of Eagle and SHAPEIT2 in five analyses of 10,000-SNP regions (every other line of Supplementary Table 4, i.e., the regions from chromosomes 2, 6, 9, 14, and 19), phasing all N=150K UK Biobank samples in each analysis. We partitioned SHAPEIT2 analyses into 3, 4, or 5 blocks (with an overlap of 500 SNPs) as necessitated by computational constraints; we ligated SHAPEIT2 output using hapfuse v1.6.2. Run times are totals across all ten regions (using 16 cores of a 2.60 GHz Intel Xeon E5-2650 v2 processor). Switch error rates are means (s.e.m.) over the ten regions, assessed on 70 European-ancestry trios. Switch error rates without blips ignore switches arising when 1–2 SNPs are oppositely phased relative to $\geq$10 consistently phased SNPs on both sides.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 8. Computational cost and accuracy of efficient methods for chromosome-scale analyses of *N*=150K samples.

(a) Running time for phasing N=150K samples in batches

Chromosome	Eagle 1x150K	SHAPEIT2 10x15K	HAPI-UR 10x15K
chr1p	2.7 days / 27.8 GB	8.8 days / 7.0 GB	5.6 days / 25.5 GB
chr10	3.3 days / 32.8 GB	10.4 days / 8.2 GB	6.7 days / 30.7 GB
chr20	1.8 days / 19.1 GB	5.2 days / 4.5 GB	3.5 days / 17.0 GB

(b) Switch error rate

Chromosome	Eagle 1x150K	SHAPEIT2 10x15K	HAPI-UR 10x15K
chr1p	0.29% (0.03%)	0.82% (0.05%)	1.68% (0.07%)
chr10	0.30% (0.02%)	0.87% (0.04%)	1.67% (0.05%)
chr20	0.32% (0.03%)	1.02% (0.06%)	1.96% (0.06%)

We phased the short arm of chromosome 1 (26,695 SNPs), chromosome 10 (31,090 SNPs), and chromosome 20 (16,367 SNPs) using up to 10 cores of a 2.27 GHz Intel Xeon L5640 processor. We report mean switch error rate (s.e.m.) over 70 children from European-ancestry trios. For the SHAPEIT2 and HAPI-UR benchmarks, we phased only one N=15K batch of the data (containing all trio children and 10% of the remaining samples) and scaled running times up by 10. We note that the HAPI-UR runs only used 3 cores, whereas Eagle and SHAPEIT2 performed multithreaded computations on 10 cores; however, parallelizing HAPI-UR jobs to fully use all cores would require $>$100GB memory, exceeding our computational resources.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 9. In-sample imputation accuracy of Eagle and SHAPEIT2.

(a) Mean in-sample imputation R^2 in curated British samples (s.e.m.)

MAF bin	Eagle 1x150K	SHAPEIT2 1x150K K=200	SHAPEIT2 1x150K K=100	SHAPEIT2 3x50K K=100	SHAPEIT2 10x15K K=100
0.1–0.2%	0.790 (0.032)	0.779 (0.038)	0.778 (0.038)	0.624 (0.039)	0.520 (0.044)
0.2–0.5%	0.857 (0.017)	0.862 (0.014)	0.817 (0.019)	0.704 (0.023)	0.606 (0.028)
0.5–1%	0.898 (0.007)	0.901 (0.006)	0.872 (0.007)	0.808 (0.009)	0.728 (0.012)
1–2%	0.917 (0.003)	0.921 (0.002)	0.897 (0.003)	0.849 (0.004)	0.784 (0.005)
2–5%	0.935 (0.001)	0.939 (0.001)	0.923 (0.001)	0.888 (0.002)	0.838 (0.003)
5–10%	0.955 (0.001)	0.959 (0.001)	0.949 (0.001)	0.926 (0.002)	0.891 (0.003)
10–50%	0.966 (0.001)	0.971 (0.001)	0.965 (0.001)	0.947 (0.001)	0.923 (0.002)

(b) Mean in-sample imputation R^2 in all samples (s.e.m.)

MAF bin	Eagle 1x150K	SHAPEIT2 1x150K K=200	SHAPEIT2 1x150K K=100	SHAPEIT2 3x50K K=100	SHAPEIT2 10x15K K=100
0.1–0.2%	0.694 (0.029)	0.697 (0.032)	0.681 (0.034)	0.583 (0.039)	0.482 (0.035)
0.2–0.5%	0.803 (0.017)	0.825 (0.015)	0.788 (0.019)	0.697 (0.023)	0.593 (0.027)
0.5–1%	0.855 (0.008)	0.863 (0.008)	0.834 (0.008)	0.767 (0.011)	0.700 (0.013)
1–2%	0.892 (0.003)	0.900 (0.002)	0.873 (0.003)	0.826 (0.004)	0.760 (0.005)
2–5%	0.914 (0.001)	0.922 (0.001)	0.905 (0.002)	0.869 (0.002)	0.819 (0.003)
5–10%	0.938 (0.002)	0.946 (0.001)	0.935 (0.002)	0.911 (0.002)	0.877 (0.003)
10–50%	0.952 (0.001)	0.960 (0.001)	0.953 (0.001)	0.935 (0.001)	0.910 (0.002)

(The first table provides numeric data plotted in Supplementary Fig. 2a.) We randomly masked 2% of the genotypes in all N=150K UK Biobank samples and phased the first 40cM of chromosome 10 using Eagle (on the full cohort) and SHAPEIT2 (on all samples at once with either K=100 (default) or 200 states as well as in N=50K and N=15K batches), imputing all masked genotypes in the process. We then evaluated the accuracy of the imputed genotypes on **(a)** the subset of British samples curated by UK Biobank for GWAS (≈80% of all samples) or **(b)** all samples, stratifying by minor allele frequency in the selected samples.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 10. In-sample imputation accuracy of Eagle and SHAPEIT2 in chromosome-scale analyses.

(a) Mean in-sample imputation R^2 in curated British samples (s.e.m.)

MAF bin	Eagle 1x150K	SHAPEIT2 10x15K
0.1–0.2%	0.790 (0.007)	0.599 (0.009)
0.2–0.5%	0.864 (0.003)	0.704 (0.005)
0.5–1%	0.906 (0.001)	0.779 (0.003)
1–2%	0.923 (0.001)	0.823 (0.001)
2–5%	0.939 (0.000)	0.863 (0.001)
5–10%	0.959 (0.000)	0.908 (0.001)
10–50%	0.971 (0.000)	0.937 (0.000)

(b) Mean in-sample imputation R^2 in all samples (s.e.m.)

MAF bin	Eagle 1x150K	SHAPEIT2 10x15K
0.1–0.2%	0.734 (0.006)	0.585 (0.007)
0.2–0.5%	0.822 (0.003)	0.679 (0.005)
0.5–1%	0.874 (0.002)	0.751 (0.003)
1–2%	0.900 (0.001)	0.800 (0.001)
2–5%	0.922 (0.000)	0.846 (0.001)
5–10%	0.944 (0.000)	0.894 (0.001)
10–50%	0.958 (0.000)	0.925 (0.000)

We randomly masked 2% of the genotypes in all N=150K UK Biobank samples and phased the short arm of chromosome 1 (26,695 SNPs), chromosome 10 (31,090 SNPs), and chromosome 20 (16,367 SNPs) using Eagle (on the full cohort) and SHAPEIT2 (in 10 batches of N=15K samples), imputing all masked genotypes in the process. We then evaluated the accuracy of the imputed genotypes on **(a)** the subset of British samples curated by UK Biobank for GWAS ($\approx$80% of all samples) or **(b)** all samples, stratifying by minor allele frequency in the selected samples.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 11. In-sample imputation accuracy of Eagle and SHAPEIT2 stratified by ethnicity.

(a) Mean in-sample imputation R^2 (s.e.m.), Eagle 1x150K

MAF bin	British	Irish	otherWhite	Indian	Caribbean
0.1–0.2%	0.771 (0.031)	NA	NA	NA	NA
0.2–0.5%	0.853 (0.015)	NA	NA	NA	NA
0.5–1%	0.896 (0.006)	NA	NA	NA	NA
1–2%	0.914 (0.002)	0.925 (0.008)	0.780 (0.011)	NA	NA
2–5%	0.932 (0.001)	0.936 (0.004)	0.808 (0.005)	NA	NA
5–10%	0.953 (0.001)	0.952 (0.003)	0.862 (0.005)	0.718 (0.011)	0.695 (0.014)
10–50%	0.965 (0.001)	0.965 (0.001)	0.893 (0.002)	0.743 (0.005)	0.690 (0.005)

(b) Mean in-sample imputation R^2 (s.e.m.), SHAPEIT2 1x150K, K=200

MAF bin	British	Irish	otherWhite	Indian	Caribbean
0.1–0.2%	0.762 (0.037)	NA	NA	NA	NA
0.2–0.5%	0.858 (0.013)	NA	NA	NA	NA
0.5–1%	0.895 (0.005)	NA	NA	NA	NA
1–2%	0.918 (0.002)	0.929 (0.008)	0.778 (0.011)	NA	NA
2–5%	0.936 (0.001)	0.933 (0.004)	0.810 (0.006)	NA	NA
5–10%	0.957 (0.001)	0.958 (0.003)	0.876 (0.005)	0.734 (0.012)	0.737 (0.013)
10–50%	0.970 (0.001)	0.969 (0.001)	0.907 (0.002)	0.784 (0.005)	0.774 (0.005)

(c) Mean in-sample imputation R^2 (s.e.m.), SHAPEIT2 1x150K, K=100

MAF bin	British	Irish	otherWhite	Indian	Caribbean
0.1–0.2%	0.753 (0.037)	NA	NA	NA	NA
0.2–0.5%	0.813 (0.017)	NA	NA	NA	NA
0.5–1%	0.866 (0.006)	NA	NA	NA	NA
1–2%	0.894 (0.003)	0.894 (0.010)	0.738 (0.012)	NA	NA
2–5%	0.921 (0.001)	0.915 (0.004)	0.778 (0.006)	NA	NA
5–10%	0.947 (0.001)	0.950 (0.003)	0.858 (0.005)	0.686 (0.012)	0.713 (0.014)
10–50%	0.964 (0.001)	0.962 (0.001)	0.895 (0.002)	0.760 (0.005)	0.744 (0.005)

(d) Mean in-sample imputation R^2 (s.e.m.), SHAPEIT2 10x15K, K=100

MAF bin	British	Irish	otherWhite	Indian	Caribbean
0.1–0.2%	0.502 (0.040)	NA	NA	NA	NA
0.2–0.5%	0.598 (0.027)	NA	NA	NA	NA
0.5–1%	0.724 (0.012)	NA	NA	NA	NA
1–2%	0.781 (0.005)	0.824 (0.012)	0.686 (0.013)	NA	NA
2–5%	0.835 (0.003)	0.831 (0.006)	0.731 (0.007)	NA	NA
5–10%	0.889 (0.003)	0.897 (0.005)	0.820 (0.006)	0.612 (0.013)	0.601 (0.016)
10–50%	0.921 (0.002)	0.922 (0.002)	0.868 (0.003)	0.710 (0.006)	0.627 (0.006)

(These tables provide numeric data plotted in Supplementary Fig. 2b along with data for SHAPEIT2 using 10 batches of N=15K samples.) The caption is on the next page.

Nature Genetics: doi:10.1038/ng.3571

Extended caption for Supplementary Table 11. We randomly masked 2% of the genotypes in all N=150K UK Biobank samples and phased the first 40cM of chromosome 10 using Eagle (on the full cohort) and SHAPEIT2 (on all samples at once with either K=100 (default) or 200 states as well as in N=15K batches), imputing all masked genotypes in the process. We then evaluated the accuracy of the imputed genotypes on subsets of the UK Biobank cohort defined by self-reported ethnicity. The five largest ethnicities in the data set were British (137,178 samples), Irish (3,977), "Any other white background" (4,760), Indian (1,324), and Caribbean (1,028). For the ethnicities with $<$5,000 samples, we report results only for minor allele frequency bins corresponding to an expected minor allele count $\geq$2 in 2% of the samples.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 12. HRC imputation accuracy after pre-phasing using SHAPEIT2 or Eagle.

MAF bin	SHAPEIT2 10x15K	Eagle 1x150K	Difference
0.1–0.2%	0.574 (0.012)	0.594 (0.012)	0.020 (0.002)
0.2–0.5%	0.665 (0.010)	0.679 (0.010)	0.013 (0.002)
0.5–1%	0.753 (0.009)	0.765 (0.009)	0.012 (0.001)
1–2%	0.786 (0.008)	0.798 (0.008)	0.012 (0.001)
2–5%	0.812 (0.007)	0.822 (0.007)	0.010 (0.001)
5–10%	0.881 (0.007)	0.888 (0.006)	0.007 (0.000)
10–50%	0.924 (0.004)	0.928 (0.004)	0.004 (0.000)

We pre-phased N=15K samples using SHAPEIT2 and pre-phased all N=150K samples using Eagle; we then imputed the same subset of N=15K pre-phased samples using the Haplotype Reference Consortium (r1) imputation panel. Each row reports mean imputation R^2 (s.e.m.) assessed in curated British samples over 300 masked SNPs, 100 each in chromosomes 1 (short arm), 10, and 20.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 13. UK10K imputation accuracy after pre-phasing using SHAPEIT2 or Eagle.

MAF bin	SHAPEIT2 10x15K	Eagle 1x150K	Difference
0.1–0.2%	0.457 (0.015)	0.468 (0.014)	0.010 (0.002)
0.2–0.5%	0.563 (0.013)	0.571 (0.013)	0.008 (0.001)
0.5–1%	0.673 (0.012)	0.680 (0.012)	0.007 (0.001)
1–2%	0.719 (0.010)	0.726 (0.010)	0.008 (0.001)
2–5%	0.754 (0.009)	0.760 (0.009)	0.006 (0.001)
5–10%	0.840 (0.008)	0.845 (0.008)	0.004 (0.000)
10–50%	0.892 (0.006)	0.894 (0.006)	0.002 (0.000)

We pre-phased N=15K samples using SHAPEIT2 and pre-phased all N=150K samples using Eagle; we then imputed the same subset of N=15K pre-phased samples using the UK10K imputation panel. Each row reports mean imputation R^2 (s.e.m.) assessed in curated British samples over 300 masked SNPs, 100 each in chromosomes 1 (short arm), 10, and 20.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 14. UK10K imputation accuracy on sequenced SNPs after pre-phasing using SHAPEIT2 or Eagle.

MAF bin	SHAPEIT2 10x15K	Eagle 1x150K	Difference
1–2%	0.797 (0.003)	0.802 (0.003)	0.004 (0.001)
2–5%	0.895 (0.002)	0.897 (0.002)	0.001 (0.000)
5–10%	0.943 (0.001)	0.943 (0.001)	0.000 (0.000)
10–50%	0.965 (0.001)	0.965 (0.001)	0.000 (0.000)

We pre-phased 89 GBR samples from the 1000 Genomes data set together with N=15K samples using SHAPEIT2 or all N=150K samples using Eagle; we then imputed the pre-phased samples using the UK10K imputation panel. Each row reports mean imputation R^2 (s.e.m.) assessed in curated British samples over all SNPs in the corresponding MAF range.

Nature Genetics: doi:10.1038/ng.3571

Supplementary Table 15. Sensitivity of Eagle to increased genotyping error.

Added error	Switch error rate	Switch error rate without blips
0%	0.304% (0.020%)	0.137% (0.013%)
0.5%	0.858% (0.029%)	0.226% (0.024%)
1%	1.528% (0.038%)	0.432% (0.032%)
2%	2.974% (0.048%)	1.164% (0.040%)

We assessed Eagle's robustness to genotyping error by adding random errors at 0.5%, 1% or 2% of genotypes on chromosome 10. Specifically, with probability 0.5%, 1%, or 2%, we modified each non-missing genotype by 1 (i.e., homozygous genotypes became hets and heterozygous genotypes became homozygous for one or the other allele with uniform probability 1/2). We then phased the modified data set using Eagle. Switch error rates are means (s.e.m.) over 70 European-ancestry trios. Switch error rates without blips ignore switches arising when 1–2 SNPs are oppositely phased relative to $\geq$10 consistently phased SNPs on both sides.

Nature Genetics: doi:10.1038/ng.3571