nature genetics

Fast and accurate long-range phasing in a UK Biobank cohort

Po-Ru Loh[1,2], Pier Francesco Palamara[1,2] & Alkes L Price[1–3]

Recent work has leveraged the extensive genotyping of the Icelandic population to perform long-range phasing (LRP), enabling accurate imputation and association analysis of rare variants in target samples typed on genotyping arrays. Here we develop a fast and accurate LRP method, Eagle, that extends this paradigm to populations with much smaller proportions of genotyped samples by harnessing long (>4-cM) identical-by-descent (IBD) tracts shared among distantly related individuals. We applied Eagle to $N \approx 150{,}000$ samples (0.2% of the British population) from the UK Biobank, and we determined that it is 1–2 orders of magnitude faster than existing methods while achieving similar or better phasing accuracy (switch error rate ≈ 0.3%, corresponding to perfect phase in a majority of 10-Mb segments). We also observed that, when used within an imputation pipeline, Eagle prephasing improved downstream imputation accuracy in comparison to prephasing in batches using existing methods, as necessary to achieve comparable computational cost.

Haplotype phasing is a fundamental question in human genetics[1] and a key step in genotype imputation[2–5]. Most existing methods for statistical phasing apply hidden Markov models (HMMs) to iteratively refine haplotype frequency models and improve phase calls[6–12]. This approach produces accurate phase inference at large sample sizes but is computationally challenging. LRP[13] is an alternative approach that harnesses long IBD tracts shared among related individuals; in such IBD regions, phase inference is straightforward and extremely accurate at sites for which at least one individual is homozygous. LRP has successfully been used in the Icelandic population to rapidly determine highly accurate phase and impute rare variants, producing insights into fine-scale recombination and enabling dozens of discoveries regarding numerous diseases[14–27]. However, because existing implementations of LRP rely on very long (>10-cM), easily identified IBD tracts in close relatives, LRP has previously only been successfully applied in isolated populations or populations with large fractions of individuals genotyped. In more general settings, existing LRP approaches are unable to phase a sizable fraction of sites[28] and have been observed to achieve worse performance (both in terms of accuracy and run time) than conventional HMM-based approaches[29].

Here we develop a new algorithm, Eagle, that surmounts these challenges by combining the key ideas of LRP and conventional methods: Eagle begins with an LRP approach, making initial phase calls on the basis of long (>4-cM) tracts of IBD sharing in closely or distantly related individuals, and concludes with two approximate HMM decoding iterations to refine phase calls. We demonstrate the efficiency and accuracy of Eagle by phasing $N \approx 150{,}000$ samples from the UK Biobank[30] (see URLs); at large sample sizes, Eagle matches the accuracy of the best HMM-based methods and is far more computationally efficient (for example, it is 14 times faster than SHAPEIT2; ref. 12). We also show that, when phasing $N \approx 150{,}000$ UK samples, Eagle imputes missing genotypes (within the data set) with accuracy (R^2) >0.75 down to a minor allele frequency (MAF) of 0.1%. When used to prephase $N \approx 150{,}000$ samples within a standard imputation pipeline, Eagle improves accuracy in downstream imputation (as compared to prephasing using existing methods on batches of $N \approx 15{,}000$ samples at comparable cost), with larger improvements expected as imputation reference panels grow. We have released Eagle as open source software (see URLs).

RESULTS

Overview of methods

The basic idea of our approach is to harness IBD from distant relatedness (up to ~12 generations from a common ancestor) that is pervasive within very large cohorts. IBD between a proband and other individuals provides a 'surrogate family' (ref. 13) for the proband, which can then immediately be used to call phase. Although this approach is simple in principle, two major challenges have precluded its application to cohorts representing small fractions of large outbred populations. First, identifying IBD is difficult, in terms of both accuracy and computational cost; moreover, the most widely used IBD inference methods rely on first phasing the data[31–33]. Second, LRP by itself can phase only sites at which the proband has at least one relative who is a homozygote; for cohorts representing a sizable fraction of a population, 2–5% of sites may be left unphased[13,15], but for smaller cohorts this fraction may exceed 25%, even in isolated populations[28], limiting the usefulness of LRP as a general purpose method. Our algorithm, Eagle, overcomes the first challenge by employing a new, fast IBD-scanning strategy and overcomes the second challenge by introducing an approximate HMM computation that rapidly refines LRP phase calls.

The Eagle algorithm has three main steps (**Fig. 1**). First, Eagle rapidly detects probable IBD tracts by identifying long regions of

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to P.-R.L. (loh@hsph.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

Received 3 September 2015; accepted 22 April 2016; published online 6 June 2016; doi:10.1038/ng.3571

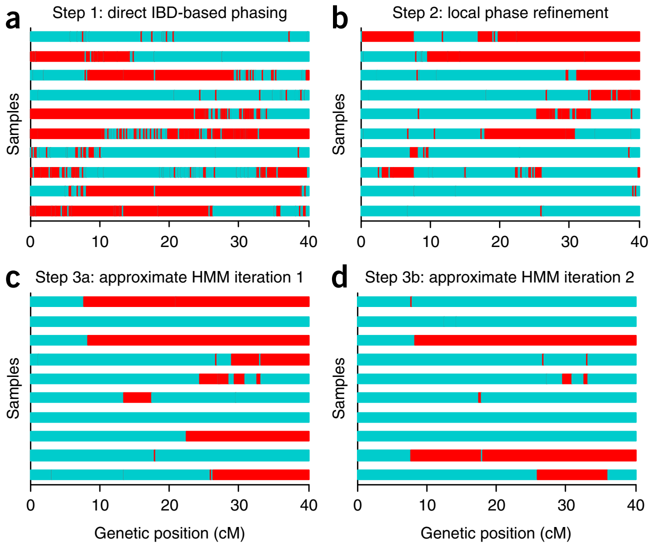

Figure 1 Eagle algorithm and example phase calls after each step. We show phase calls for ten children from family trios after each successive step of the Eagle algorithm (applied to phase the first 40 cM of chromosome 10 in all $N \approx 150{,}000$ UK Biobank samples except the parents from family trios). At all trio-phased sites, red and blue indicate whether the first Eagle-phased haplotype for each child matches the maternal or paternal haplotype, respectively. (**a**) After the first step, a sizable proportion of each genome is covered by long segments of nearly perfect phase; these segments are the regions in which long IBD is available from several relatives. (**b**) The second step, which uses both long and short IBD segments, fixes most of the phase switch errors in the first step. (**c**,**d**) The subsequent approximate HMM iterations further reduce the error rate.

agreement at homozygous sites (at which alleles for each haplotype are known without phasing), scoring identified regions using allele frequency and linkage disequilibrium (LD) information and checking overlapping regions for consistency; Eagle uses the detected IBD to perform accurate initial LRP in high-IBD regions (**Fig. 1a**). Second, Eagle performs local phase refinement in overlapping ~1-cM windows by detecting complementary haplotype pairs (among the haplotypes inferred in the previous step); specifically, for each diploid individual, Eagle rephases the individual by searching for long stretches of IBD with haplotypes from step 1 and then checking for the existence of haplotypes complementary to the IBD hits (**Fig. 1b**). Third, Eagle finalizes phase calls by running two approximate HMM decoding iterations using up to 80 local reference haplotypes and aggressively pruning the search space to ≤200 states per position using a fast path-scoring scheme (**Fig. 1c**,**d**). All three steps are multithreaded and make use of bit operations to perform key computations in 64-SNP blocks. (For full details, see the Online Methods and **Supplementary Note**.)

Computational cost

We benchmarked Eagle against state-of-the-art phasing methods—Beagle[8], HAPI-UR[11], and SHAPEIT2 (ref. 12) (see URLs)—on subsets of the UK Biobank data set containing $N \approx 15{,}000$, 50,000, or 150,000 samples (Online Methods). After quality control, this data set contained 626,636 autosomal markers with average heterozygosity of 0.189 and a MAF distribution typical of genotyping arrays: 42,612 variants with MAF 0.1–1%, 234,616 variants with MAF 1–5%, and 349,408 variants with MAF 5–50%. (Our quality control procedure excluded very rare variants with MAF <0.1%; see the Online Methods.) For our first benchmark, we phased only the first 40 cM of chromosome 10 (~1% of the data; 5,824 SNPs spanning 18 Mb) to allow as many methods as possible to complete in <2 weeks (using up to ten cores on a single compute node; all methods except HAPI-UR support multithreading over ten cores). We observed that Eagle achieved an increase in speed of 1–2 orders of magnitude over the other methods across the sample size range (**Fig. 2a** and **Supplementary Table 1**), attaining a 14-fold increase in speed over SHAPEIT2 and a 12-fold increase in speed over HAPI-UR at $N \approx 150{,}000$. (Beagle was unable to phase 1% of the genome in 2 weeks at $N \approx 150{,}000$.) We note that, like other methods, Eagle has parameters that produce a tradeoff between speed and accuracy (Online Methods); Eagle's --fast mode achieved a further ~2-fold increase in speed over the default while incurring only a slight loss of accuracy (**Table 1**). All methods exhibited superlinear but subquadratic scaling of run time with sample size, consistent with the presence of both linear and quadratic algorithmic components. (For a detailed discussion of the run time scaling for each of Eagle's algorithmic steps, see the Online Methods and **Supplementary Table 2**.) We also observed that Eagle achieved modest (two- to eightfold) savings in memory cost in comparison to other methods (**Fig. 2b** and **Supplementary Table 1**). All methods exhibited memory cost scaling roughly linearly with sample size.

Phasing accuracy

We assessed the accuracy of each phasing method using gold standard data from the 70 European-ancestry trios in the UK Biobank data set (all but one of which self-reported as having British ancestry; Online Methods). Specifically, we included all trio children and excluded all trio parents in each phasing run; we then assessed computational

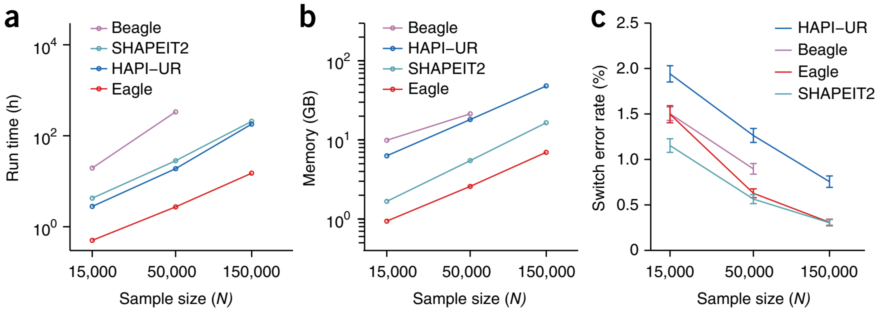

Figure 2 Computational cost and accuracy of phasing methods. Benchmarks for Eagle and existing phasing methods (all run with default options) on $N \approx 15{,}000$, 50,000, and 150,000 UK Biobank samples and $M = 5{,}824$ SNPs on chromosome 10. (**a**,**b**) Log–log plots of run times (**a**) and memory consumption (**b**) using up to ten cores of a 2.27-GHz Intel Xeon L5640 processor and up to 2 weeks of computation. (**c**) Mean switch error rate over 70 European-ancestry trios; error bars, s.e.m. 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 analysis with $N \approx 150{,}000$ samples, 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.) Numerical data are provided in **Supplementary Table 1**.

Table 1 Computational cost and accuracy of Eagle and SHAPEIT2 on $N \approx 150{,}000$ samples using various parameters

Method	Run time (d)	Switch error rate (%)	Switch error rate without blips (%)	0-discrepancy 10-Mb segments (%)	≤2-discrepancy 10-Mb segments (%)
Eagle --fast	2.8	0.317 (0.012)	0.152 (0.012)	61.2 (1.9)	79.3 (1.7)
Eagle	5.0	0.272 (0.009)	0.118 (0.007)	63.3 (1.9)	81.6 (1.6)
SHAPEIT2 K = 100 (3 blocks)	106.8	0.302 (0.015)	0.159 (0.010)	56.9 (1.5)	71.2 (1.3)
SHAPEIT2 K = 200 (4 blocks)	118.8	0.261 (0.015)	0.124 (0.009)	63.5 (1.6)	77.8 (1.1)
SHAPEIT2 K = 400 (5 blocks)	152.8	0.239 (0.012)	0.101 (0.005)	64.9 (1.5)	80.4 (1.1)

We benchmarked various parameter settings for Eagle and SHAPEIT2 in analyses of ten 10,000-SNP regions comprising 16% of the genome (listed in **Supplementary Table 4**), phasing all $N \approx 150{,}000$ UK Biobank samples in each analysis. We split the SHAPEIT2 analyses into three, four, or five 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 shown as means (s.e.m.) over the ten regions, assessed for 70 European-ancestry family trios. Switch error rates without blips ignore switches arising when 1 or 2 SNPs are oppositely phased relative to ≥10 consistently phased SNPs on both sides of the blip. The number of discrepancies within a 10-Mb segment is defined as the minimum number of SNPs with incorrect phase when comparing a phased haplotype to either trio-phased haplotype[1]; the percentages of 10-Mb segments with 0 or ≤2 discrepancies are shown as means (s.e.m.) over the ten 10,000-SNP regions. Detailed discrepancy distributions are provided in **Supplementary Table 5**.

phase accuracy in trio children at all trio-phased sites (that is, SNPs heterozygous in the child and homozygous in at least one parent; comprising ~80% of heterozygous SNPs per trio child). We observed that, when phasing $N \approx 150{,}000$ samples over the same 1% of the genome as above, Eagle and SHAPEIT2 achieved nearly identical, very low (~0.3%) mean switch error rates (**Fig. 2c** and **Supplementary Table 1**). The accuracy of Eagle relative to SHAPEIT2 degraded slightly with decreasing sample size (as expected with limited IBD in an outbred population); interestingly, however, Eagle still achieved better accuracy than all other methods except SHAPEIT2 at sample sizes of $N \approx 50{,}000$ and 15,000, with only an 11% (s.e.m. = 9%) increase in switch error rate relative to SHAPEIT2 at $N \approx 50{,}000$: switch error rate of 0.63% (0.05%) for Eagle versus 0.56% (0.05%) for SHAPEIT2. To confirm these results, we performed a similar benchmark of Eagle and SHAPEIT2 on $N \approx 60{,}000$ samples from the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort of more diverse European ancestry[34,35] (Online Methods) and observed similar results: switch error rate of 0.82% (0.03%) for Eagle versus 0.70% (0.03%) for SHAPEIT2, a 16% (2%) increase (**Supplementary Table 3**).

We next undertook a detailed comparison of the phasing accuracy achieved by the two most accurate methods, Eagle and SHAPEIT2, when run on $N \approx 150{,}000$ samples. For this comparison, we analyzed ten 10,000-SNP regions (of median length 44 Mb) comprising 16% of the genome (**Supplementary Table 4**). To overcome the high computational cost of SHAPEIT2 analyses with $N \approx 150{,}000$ (**Fig. 2a**), we performed these benchmarks on the Lisa Genetic Cluster Computer (see URLs), which offered high-throughput parallel computation in batches of 16-core, 5-d jobs. Because SHAPEIT2 was unable to complete 10,000-SNP analyses within a single job, we split each 10,000-SNP region into three overlapping blocks of 3,667 SNPs each (with an overlap of 500 SNPs); we ligated the results from analyses of these blocks using hapfuse v1.6.2 (see URLs). In these benchmarks, we observed that Eagle achieved a slightly lower switch error rate than SHAPEIT2 run with default parameters (K = 100 conditioning states[12]): switch error rate of 0.272% (0.009%) for Eagle versus 0.302% (0.015%) for SHAPEIT2 (P = 0.03, one-sided paired t test) (**Table 1**). One caveat regarding these comparisons is that splitting and ligating the SHAPEIT2 analyses may have incurred a slight loss of accuracy[12]; although computational limitations prevented us from running SHAPEIT2 on full 10,000-SNP regions, performing the full analyses in single computations could improve accuracy.

In light of the very low switch error rates achieved by both Eagle and SHAPEIT2 in analyses with $N \approx 150{,}000$ samples, we further investigated the nature of the errors made by each method. We observed that many of the switch errors accrued by Eagle were the result of 'blips' involving one or occasionally two adjacent SNPs oppositely phased relative to surrounding SNPs (**Fig. 1d**); such errors can arise from genotyping error or from recent mutation or gene conversion. We therefore computed an alternative metric, 'switch error rate without blips', in which we ignored errors in which 1 or 2 SNPs were oppositely phased relative to ≥10 consistently phased SNPs on both sides of the 1- or 2-SNP blip. Assessment with this metric showed that one- or two-SNP blips (previously counted as two switches) accounted for the majority (~60%) of Eagle's switch errors; similarly, such errors accounted for roughly half of the switch errors for SHAPEIT2 (**Table 1**). We further considered the metric 'discrepancies within a 10-Mb segment' (ref. 13) and observed that both Eagle and SHAPEIT2 achieved perfect phase in the majority of the 10-Mb segments phased (**Table 1** and **Supplementary Table 5**).

As both Eagle and SHAPEIT2 have parameters that affect the tradeoff between speed and accuracy, we also investigated the effect of running each method with non-default parameter settings. For Eagle, we benchmarked its --fast mode, which speeds up analysis by increasing the size of SNP blocks and limiting the approximate HMM computation (Online Methods). We observed that the --fast mode of Eagle completed analyses roughly twice as quickly as the default mode with a slightly higher switch error rate (0.317%, s.e.m. = 0.012%) (**Table 1**). We also benchmarked slower parameter settings that decrease the size of SNP blocks or expand the approximate HMM computation; these modifications had little effect on accuracy (**Supplementary Table 6**). For SHAPEIT2, we increased the number of conditioning states from K = 100 (the default) to K = 200 or 400, simultaneously increasing the number of ligated blocks to four or five per 10,000-SNP region to keep per-job run times within the 5-d limit (**Table 1**). We observed that using K = 200 conditioning states achieved accuracy similar to Eagle, whereas using K = 400 states achieved the lowest switch error rate of all methods tested (0.239%, s.e.m. = 0.012%; P = 0.007 versus Eagle, one-sided paired t test) (**Table 1**). (We note that SHAPEIT2 analyses with K = 400 required ~40% more computational time than default SHAPEIT2 analyses with K = 100; although the run time scaling of SHAPEIT2 is asymptotically linear with respect to K (ref. 12), the quadratic component of the computation, which is independent of K, dominates at very large N values and typical K values.) We also considered increasing the window size parameter for SHAPEIT2 from 2 Mb to 4 Mb, but the results of a pilot analysis indicated that doing so substantially decreased accuracy (**Supplementary Table 7**).

Finally, we assessed the accuracy of the analysis options for efficiently phasing $N \approx 150{,}000$ samples. In addition to Eagle (run on all $N \approx 150{,}000$ samples together; 1 × 150,000), we considered batching approaches requiring up to three times the run time of Eagle 1 × 150,000 analysis. According to our run time benchmarks (**Fig. 2a** and **Supplementary Table 1**), SHAPEIT2 or HAPI-UR analysis of the data in ten batches of $N \approx 15{,}000$ samples (10 × 15,000) satisfied this constraint. We benchmarked each method in three chromosome-scale tests: the short arm of chromosome 1 (26,695 SNPs), chromosome 10 (31,090 SNPs), and chromosome 20 (16,367 SNPs), amounting to 12% of the genome. Our results (**Supplementary Table 8**) confirmed our previous benchmarks (**Fig. 2**) and were consistent across chromosomes. In particular, we observed that Eagle analysis of all $N \approx 150,000$ samples together completed three times faster than SHAPEIT2 10 × 15,000 analysis while achieving a 67% (1%) decrease in switch error rate: switch error rate of 0.30% (0.01%) for Eagle 1 × 150,000 versus 0.90% (0.06%) for SHAPEIT2 10 × 15,000 (**Supplementary Table 8**).

In-sample imputation and GWAS imputation accuracy

We next investigated the use of Eagle for genotype imputation. First, to project the imputation accuracy that will be achievable in the UK population using LRP-based methods once a reference panel of $N \approx 150,000$ sequenced UK samples becomes available (**Supplementary Fig. 1**), we performed in-sample imputation of masked genotypes in the UK Biobank data set (Online Methods and **Supplementary Note**). In these benchmarks (**Supplementary Fig. 2** and **Supplementary Tables 9–11**), Eagle and SHAPEIT2 both achieved mean in-sample imputation R^2 >0.75 down to a MAF of 0.1%. As in our benchmarks of switch error rate (**Table 1**), Eagle was slightly more accurate than SHAPEIT2 run with default parameters and achieved accuracy similar to that from SHAPEIT2 run with $K = 200$ states; when compared to SHAPEIT2 10 × 15,000, Eagle 1 × 150,000 was much more accurate (**Supplementary Fig. 2** and **Supplementary Table 9**). In-sample imputation on N samples bears some similarities to standard genome-wide association study (GWAS) phasing and imputation on a target sample using a reference panel of size N (as both tasks entail copying shared haplotypes—identified on the basis of data at typed SNPs—from a set of N samples); however, the two tasks also have several important differences and require different algorithms and software (**Supplementary Fig. 1**). We therefore caution that our in-sample imputation results may not be representative of GWAS imputation performance using a reference panel with $N \approx 150,000$ samples.

We also investigated the benefits of using Eagle for prephasing[5] within an existing imputation pipeline—the Sanger Imputation Service, which currently supports imputation using up to $N \approx 32,000$ sequenced reference individuals from the Haplotype Reference Consortium (HRC; see URLs). We note that the HRC is predominantly European and contains a substantial fraction of UK samples, although it also contains samples of other ancestries (see URLs). We considered two fast prephasing procedures: Eagle prephasing of all $N \approx 150,000$ UK Biobank samples and SHAPEIT2 10 × 15,000 prephasing of $N \approx 150,000$ samples. To benchmark imputation accuracy, we completely masked 700 SNPs (100 in each of seven MAF bins) on each of three chromosomes, prephased the remaining SNPs with Eagle and SHAPEIT2, imputed the same subset of $N \approx 15,000$ prephased samples using the Sanger Imputation Service, and computed R^2 values for comparisons of the true genotypes of the masked SNPs and their imputed genotype dosages across curated British samples (Online Methods; see URLs). This benchmarking procedure is commonly used to assess the accuracy of phasing and imputation pipelines[5,9]. We observed that, when imputation was performed using the largest reference panel available (the HRC with $N \approx 32,000$ samples), Eagle prephasing using all $N \approx 150,000$ samples improved imputation R^2 values by increasing amounts for increasingly rare SNPs, with a gain of 0.020 (0.002) in R^2 values for SNPs with MAF = 0.1–0.2% ($R^2 = 0.594$ (0.012) for Eagle 1 × 150,000 prephasing versus $R^2 = 0.574$ (0.012) for SHAPEIT2 10 × 15,000 prephasing; **Supplementary Table 12**). We caution that these results are based on the preliminary release (r1) of the HRC; development of the HRC is still underway, and performance may improve with future releases. When imputation was performed using only the UK10K reference panel with $N \approx 4,000$ samples (see URLs), the gains in R^2 values were roughly half as large (**Supplementary Table 13**). Finally, to verify that similar improvements could be obtained at genome-wide SNPs (in comparison to the subsets of SNPs we masked), we ran the 1000 Genomes Project British in England and Scotland (GBR) samples through the same pipeline after prephasing them together with the UK Biobank samples and again observed a modest improvement using UK10K imputation (**Supplementary Table 14**). We were unable to perform this analysis using HRC imputation because the HRC contains the 1000 Genomes Project data. These results demonstrate that high-accuracy prephasing is already beneficial for GWAS imputation at current reference sizes ($N \approx 4,000$ UK10K samples and $N \approx 32,000$ diverse European HRC samples) and that gains will increase as reference panels grow, corroborating our in-sample imputation results projecting high future accuracy with $N \approx 150,000$ reference samples.

DISCUSSION

We have developed a fast and accurate LRP-based phasing method, Eagle, and demonstrated that LRP can be effective in a cohort representing a small fraction of a large outbred population. Ever since Kong *et al.* established the efficacy of LRP in the Icelandic population—speculating that "having as little as 1% of a population genotyped may be adequate for the method to yield useful results" (ref. 13)—the extension of LRP to more general settings has been eagerly anticipated but up to now unrealized[1]. We have successfully applied Eagle to phase 0.2% of the UK population and demonstrated its usefulness in enhancing the accuracy of downstream imputation. We note that LRP in 0.2% of the UK population cannot match the accuracy that was achieved in 11% of the Icelandic population[13,15] (which further improved with genotyping of >30% of the Icelandic population[25]); however, these results give reason for optimism that LRP-based phasing accuracy in the UK population (and other large outbred populations) will continue to improve as more individuals are genotyped.

Eagle is a very different method from the 'pure' LRP approach of Kong *et al.*[13]: to create an algorithm that could harness limited and often distant relatedness, we needed to combine aspects of LRP and conventional HMM-based phasing, confirming the hypothesis that "IBD-based phasing can be extended [...] by using more sensitive methods for detecting IBD and combining IBD-based phasing with population haplotype frequency models" (ref. 1). Indeed, these ideas have implicitly begun to converge within sophisticated HMM-based methods (for example, SHAPEIT2), as has recently been observed[29]. SHAPEIT2 takes a 'bottom–up' approach in which it steadily improves phase accuracy over the course of a few dozen Markov chain Monte Carlo (MCMC) sampling iterations, iteratively copying phase information from progressively more accurate sets of best reference haplotypes. This procedure eventually achieves high-accuracy phase for a proband's (distant) relatives, selects them as reference haplotypes, and uses them to phase the proband[29]. In contrast, Eagle takes a 'top–down' approach, first scanning all pairs of individuals for long IBD tracts, using these tracts to phase long stretches of the genome, and then applying only two iterations of approximate HMM decoding to correct errors and fill in unphased regions (**Fig. 1**). Thus, at a high level, the key methodological contribution of Eagle's top–down approach is its use of LRP to greatly improve speed (by over an order of magnitude) by eliminating the need to slowly build phase accuracy over many HMM sampling iterations. This increase in speed is essential at large sample sizes: because of computational constraints, the production phasing of UK Biobank samples was not performed using the most accurate method available, SHAPEIT2; instead, a new method was developed, SHAPEIT3, which was reported to achieve a higher switch error rate of ~0.4% (see URLs)[36]. At the very large sample sizes to come, experience from Iceland indicates that HMM iterations may not be necessary at all[13,15,25]; instead, optimizing accuracy will require solving problems of a different nature, for example, resolving conflicts in IBD information.

Beyond our immediate goal of fast and accurate phasing, we envision that the primary downstream application of Eagle will be genotype imputation (via prephasing with Eagle followed by imputation with other software) in the UK Biobank and future population cohorts of similar or larger size. We have demonstrated the use of Eagle in current imputation pipelines and the promise of this approach for use in future data sets (for example, imputation using $N \approx 150{,}000$ reference samples). However, realizing this potential will require additional work. First, as currently implemented, Eagle is optimized for phasing array data and will need to be modified to phase sequence data. In particular, the method will need to be modified to incorporate additional information available from paired-end reads[37] and from rare variants—which can greatly aid IBD calling—while accounting for increased error rates. Simulations with increased rates of genotyping error suggest that the Eagle algorithm is in principle quite robust to error (**Supplementary Table 15**), but additional tuning will undoubtedly be necessary. Second, an imputation algorithm capable of rapidly and accurately imputing prephased target samples using very large imputation reference panels will be needed. Several efforts to develop such methods are currently underway: the Sanger Imputation Service (see URLs) is already using a new (unpublished) imputation algorithm based on positional Burrows–Wheeler transformation (PBWT)[38], which like Eagle applies fast string matching algorithms in favor of exact statistical modeling; the Beagle v4.1 imputation software[39] and the Minimac3 imputation software (unpublished but in use by the Michigan Imputation Server; see URLs) likewise aim to satisfy these requirements. Finally, the sequence data itself will need to be generated. However, very large-scale sequencing projects are already underway: for example, Genomics England plans to sequence 100,000 genomes by 2017 (see URLs).

Eagle provides new levels of efficiency and accuracy in comparison to fast alternatives for phasing very large cohorts, although we note a few limitations. First, Eagle relies on the IBD present in very large data sets to achieve high accuracy; on smaller data sets (for example, $N \approx 15{,}000$), we recommend alternative methods. Second, along similar lines, we observed that, when phasing all $N \approx 150{,}000$ UK Biobank samples together, Eagle achieved lower accuracy than SHAPEIT2 on the <10,000 samples of non-European ancestry (owing to limited IBD). In practice, such samples are easily detected (for example, by using FastPCA[35] or SNPweights[40]) and could be phased separately with SHAPEIT2. Alternatively, a hybrid algorithm that uses the Eagle approach for most of the phasing computation but switches to the SHAPEIT2 model in segments of the genome lacking IBD would be ideal; developing such an algorithm is a direction for future work. Finally, despite Eagle's speed, its computational complexity contains a quadratic term, like all other published methods, and will become daunting for data sets with millions of samples. Most simply, this issue could be sidestepped by phasing very large samples in batches of a few hundred thousand samples at a time, but we expect that further algorithmic improvements will be possible, for example, limiting the set of haplotypes considered as potential surrogate parents via clustering methods (as in SHAPEIT3; ref. 36). Despite these limitations, we expect that Eagle in its current form—already much faster than existing methods with equal or better accuracy—will be a useful tool for large-sample phasing, and we believe further innovations will amplify the advantages of LRP-based phasing and imputation.

URLs. Eagle v1.0 software and source code, http://www.hsph.harvard.edu/alkes-price/software/; SHAPEIT v2 software, http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html; HAPI-UR v1.01 software, http://code.google.com/p/hapi-ur/; Beagle v4.0 software, http://faculty.washington.edu/browning/beagle/beagle.html; hapfuse v1.6.2 software, http://bitbucket.org/wkretzsch/hapfuse/src; PLINK2 software, http://www.cog-genomics.org/plink2; SNPweights v2.0 software, http://www.hsph.harvard.edu/alkes-price/software/; UK Biobank, http://www.ukbiobank.ac.uk/; UK Biobank genotyping and quality control documentation, http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf; UK Biobank phasing and imputation documentation (including brief description of SHAPEIT3), http://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf; 1000 Genomes Project data set, http://www.1000genomes.org/; UK10K project, http://www.uk10k.org/; Haplotype Reference Consortium (HRC), http://www.haplotype-reference-consortium.org/; Sanger Imputation Service, http://imputation.sanger.ac.uk/; Michigan Imputation Server, http://imputationserver.sph.umich.edu/; 100,000 Genomes Project, http://www.genomicsengland.co.uk/the-100000-genomes-project/; Lisa Genetic Cluster Computer, http://geneticcluster.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to G. Bhatia, S. Gusev, M. Lipson, B. Pasaniuc, N. Patterson, and N. Zaitlen for helpful discussions. This research was conducted using the UK Biobank Resource and was supported by US National Institutes of Health grants R01 HG006399 and R01 MH101244 and US National Institutes of Health fellowship F32 HG007805. Computational analyses were performed on the Orchestra High-Performance Compute Cluster at Harvard Medical School, which is partially supported by grant NCRR 1S10RR028832-01, and on the Lisa Genetic Cluster Computer hosted by SURFsara and financially supported by the Netherlands Scientific Organization (NWO 480-05-003, principal investigator D. Posthuma) along with a supplement from the Dutch Brain Foundation and VU University Amsterdam.

AUTHOR CONTRIBUTIONS

P.-R.L. and P.F.P. designed the algorithm. P.-R.L. implemented the algorithm and performed experiments. P.-R.L. and A.L.P. analyzed data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **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. *Nat. Genet.* **39**, 906–913 (2007).
3. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **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. *Genet. Epidemiol.* **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. *Nat. Genet.* **44**, 955–959 (2012).

6. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **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. *Am. J. Hum. Genet.* **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. *Am. J. Hum. Genet.* **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. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
10. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
11. Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91**, 238–251 (2012).
12. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
13. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
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. *Nat. Genet.* **42**, 906–909 (2010).
18. Holm, H. *et al.* A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
19. Rafnar, T. *et al.* Mutations in *BRIP1* confer high risk of ovarian cancer. *Nat. Genet.* **43**, 1104–1107 (2011).
20. Gudmundsson, J. *et al.* Discovery of common variants associated with low TSH levels and thyroid cancer risk. *Nat. Genet.* **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. *Nat. Genet.* **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. *Nat. Genet.* **45**, 1371–1374 (2013).
23. Kong, A. *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nat. Genet.* **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. *Nat. Genet.* **46**, 294–298 (2014).
25. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
26. Steinberg, S. *et al.* Loss-of-function variants in *ABCA7* confer risk of Alzheimer's disease. *Nat. Genet.* **47**, 445–447 (2015).
27. Helgason, H. *et al.* Loss-of-function variants in *ATM* confer risk of gastric cancer. *Nat. Genet.* **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. *Genet. Epidemiol.* **35**, 853–860 (2011).
29. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
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 Med.* **12**, e1001779 (2015).
31. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
32. Browning, B.L. & Browning, S.R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **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. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
36. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* http://dx.doi.org/10.1038/ng.3583 (2016).
37. Delaneau, O., Howie, B., Cox, A.J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **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. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
40. Chen, C.-Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).

ONLINE METHODS

Eagle algorithm. We outline the three main steps of the Eagle algorithm here; full details are provided in the **Supplementary Note**. The first and second steps each iterate through all individuals in the data set 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 phasing benchmark with $N \approx 150{,}000$ samples (**Fig. 2**).

Step 1: direct IBD-based phasing using long tracts of IBD. For each proband in turn, Eagle scans all other (diploid) individuals for long genomic segments (>4 cM in length) 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 analyses with $N \approx 150{,}000$ samples, this step required ~10% of the total computational time (**Supplementary Table 2**) and achieved nearly perfect phasing over long swaths of the genome covering most of each sample (corresponding to regions having IBD with several relatives) (**Fig. 1a**). In more detail, our algorithm applies the following four procedures to each proband in turn.

First, we run a fast $O(MN)$-time scan against all other individuals for long runs of diploid genotypes containing no opposite homozygotes (with identity by state (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 run time by using bit operations to analyze blocks of 16–64 SNPs simultaneously and using dynamic programming to record the longest ten stretches with IBS >0 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 0.3 cM; upon reaching either limit, we end the current block and begin the next block.

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 (ref. 43), which likewise computes approximate likelihood-ratio scores to increase the 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 as a result of IBD rather than by chance. We apply this procedure within a seed-and-extend framework in which we begin with long matches with IBS >0 but consider extending them beyond sites with IBS = 0 (to tolerate genotyping errors). We record all extended matches with length >4 cM and likelihood ratio >10N (where N is the number of samples) as probable IBD matches.

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 for a proband and potential surrogate parents 1 and 2, the implied shared haplotypes can be (i) consistent with the proband sharing the same haplotype with both surrogates 1 and 2, (ii) consistent with the proband sharing one of its haplotypes with surrogate 1 and the other with surrogate 2, or (iii) inconsistent with both of these possibilities. We first identify pairs of overlapping probable IBD matches in which scenario (iii) occurs; for these pairs, we assume that the longer match is correct and trim the shorter match until consistency under either scenario (i) or (ii) is achieved. If any match drops below 3 cM in length 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 the global consistency of implied phase orientations among all matches: that is, 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 (i) and (ii).

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 site heterozygous in the proband, we use that surrogate to phase the site. If all surrogates are also heterozygous, we make a probabilistic phase call based on the allele frequency of the SNP and the difference between the number of (heterozygous) surrogate maternal haplotypes and surrogate paternal haplotypes.

Step 2: local phase refinement using long and short tracts of IBD. For each diploid proband in turn, Eagle analyzes overlapping ~1-cM windows of the 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 analyses with $N \approx 150{,}000$ samples, this step required ~20% of the total computational time (**Supplementary Table 2**) and reduced the switch error rate to ~1.5% (**Fig. 1b**). In more detail, our algorithm applies the following three procedures to each proband in turn.

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 with IBS >0 (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 above).

Second, for each window of three consecutive blocks (containing a total of up to 192 SNPs spanning up to 0.9 cM) and for each of the ten longest haplotype matches covering that 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; however, in such cases, the other haplotype is often shared in a short IBD tract, 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 an 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.

Third, we select the complementary haplotype pair with the lowest error in each window (block triplet) and use it to phase the block in the center of the window. This procedure is fairly straightforward, with the only subtleties being that, at error SNPs (which are heterozygous in the proband but for which both surrogate haplotypes have the same allele), we defer to the surrogate with higher confidence (from step 1) and that, 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.

Step 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 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. In our analyses with $N \approx 150{,}000$ samples, this step required ~70% of the total computational time (**Supplementary Table 2**) and reduced the switch error rate to ~0.4% after the first HMM iteration and ~0.3% after the second HMM iteration (**Fig. 1c,d**). In more detail, our algorithm applies the following three procedures to each proband in turn (in each HMM iteration).

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 $O(MN)$-time 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.

Second, we compute an approximate Viterbi decoding of a diploid 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

doi:10.1038/ng.3571

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 Mendelian errors between the proband and surrogate parents. Exact Viterbi decoding of this HMM using dynamic programming requires $O(MK^3)$ time (for K^2 state pairs and $O(K)$ possible transitions per position), 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 and thus limiting the complexity to $O(MKP)$. We then phase the proband according to the approximate Viterbi path.

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 ~1-cM haplotypes that match the Viterbi-inferred maternal and paternal haplotypes with 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 >1.5 cM from either end of a probable IBD segment, we override the switch.

Fast mode of the Eagle algorithm. Many parameters of the Eagle algorithm can potentially be modified to adjust the tradeoff between accuracy and speed. For simplicity, we created a single --fast mode that roughly doubles Eagle's speed by increasing the maximum SNP block span from 0.3 cM to 0.5 cM and reducing the comprehensiveness of the second HMM iteration (by reducing its beam search width from $P = 200$ to $P = 100$ and only rephasing the samples processed in the first half of the first HMM iteration).

Scaling of Eagle run time. Each of the three steps of the Eagle algorithm involves an all-pairs $O(MN^2)$ computation (where M is the number of SNPs and N is the number of samples) followed by an additional computation; the latter computation is inexpensive for step 1 and scales close to linearly with N for steps 2 and 3 (**Supplementary Table 2**). Thus, the distribution of time spent per step changes slightly with sample size, but no specific step is asymptotically a bottleneck. Summing across the three steps, the all-pairs $O(MN^2)$ computation constitutes slightly over half of the total computational cost at $N \approx 150{,}000$ (**Supplementary Table 2**).

All components of the Eagle algorithm have run time linear with the number of SNPs (with a small constant factor via bit operations). For genotype array data consisting mostly of common SNPs, linear scaling is optimal; however, for rare-variant-dense data (for example, sequence data), sublinear scaling should be possible, as rare variants have much lower information content than common variants. We note that this scaling could be achieved with some additional engineering, for example, by applying Eagle to only a subset of common and low-frequency variants and incorporating compressed rare variants *post hoc* (in a manner similar to imputation).

UK Biobank data set. We analyzed data from the UK Biobank, consisting of 152,729 samples typed at ~800,000 SNPs. Using PLINK2 (ref. 48 and see URLs), we removed 480 individuals marked for exclusion from genomic analyses on the basis of missingness and heterozygosity filters, leaving 152,249 samples (see URLs, genotyping and quality control documentation). We restricted the SNP set to autosomal, biallelic SNPs with MAF ≥0.1% and missingness ≤5%, leaving 626,636 SNPs (26,695 on the short arm of chromosome 1, 31,090 on chromosome 10, and 16,367 on chromosome 20). We identified 72 family trios on the basis of IBS0 <0.001, sex of the parents, and age of the trio members (see URLs, genotyping and quality control documentation). Of the 72 trio children, 69 self-reported British ancestry, 1 self-reported Indian ancestry, and 1 self-reported Caribbean ancestry. The remaining trio child did not self-report any ancestry, but her parents self-reported Irish and "any other white background" as their ancestries. UK Biobank genotyping and quality control analyses indicated that self-reported ancestry aligned closely with genetic ancestry (see URLs); however, UK Biobank also curated a subset of 120,286 self-reported British samples recommended for GWAS.

GERA data set. We analyzed GERA samples (database of Genotypes and Phenotypes (dbGaP) study accession phs000674.v1.p1) typed on the GERA EUR chip[49]. The data contained 62,318 samples, of which we removed 961 with <90% European ancestry as determined by SNPweights v2.0 (ref. 40). Among this subset of samples, we identified 197 family trios from independent pedigrees according to relationships provided with the data release. We analyzed chromosome 10, which contained 32,741 SNPs.

Phasing software versions and parameter settings. We tested the latest version of each method available as of August 2015, using its recommended parameter settings. For Eagle (v1.0), SHAPEIT v2 (r790), and Beagle (v4.0 r1399), no command line arguments were required beyond file paths and threading settings (ten computational threads). For HAPI-UR (v1.01), we set the maximum window size to 80 SNPs (as recommended on the basis of genotyping density) and combined results from three parallel runs of the algorithm using different random seeds[11].

Evaluation of phasing performance. For our benchmark analyses of $N \approx 150{,}000$ UK Biobank samples, we removed 144 trio parents and phased the remaining 152,105 samples. For our benchmarks on $N \approx 50{,}000$ or 15,000 samples, we phased all 72 trio children along with one-third or one-tenth of the remaining non-trio-parent samples (50,752 and 15,270 samples in total, respectively). We evaluated phasing accuracy in trio children by comparing computational phase calls to trio phase calls (ignoring SNPs with Mendelian errors); trio phase was available at ~80% of heterozygous SNPs. For each child, we computed switch error rate by dividing the number of phase mismatches at consecutive trio-phased SNPs by the total number of trio-phased heterozygous SNPs minus 1 (ref. 1), that is, ~15% of all SNPs (varying slightly among samples). In our results, we report mean switch error rates over the 70 European-ancestry trio children (according to self-reported ancestry). We applied an analogous procedure for our GERA benchmarks (differing only in that we removed all known relatives of the trio children, as the data contained a few extended pedigrees, leaving 60,929 samples).

Evaluation of in-sample imputation accuracy. In our in-sample imputation benchmarks, we used the same SNP and sample subsets described above, but we modified the genotype data by randomly masking 2% of all genotypes, increasing the missingness of each SNP by ~0.02. We then phased the masked data, obtaining imputed genotypes at all masked SNPs in the phased output. For each SNP, we computed the adjusted R^2 value between the actual and imputed masked genotype values according to the formula

$$\text{adjusted } R^2 := R^2 - \frac{1-R^2}{n-2} \qquad (1)$$

where R^2 on the right is the usual coefficient of determination and n is the number of data points. (This adjustment corrects for upward bias due to finite sample size; for simplicity, we always use R^2 to refer to adjusted R^2 elsewhere in this manuscript.) We computed means and standard errors of R^2 values over MAF strata, treating R^2 values from different SNPs as approximately independent given that the ~2% subset of masked individuals varied from SNP to SNP. To assess in-sample imputation accuracy on a subset of samples (for example, the 120,000 British samples curated by UK Biobank for GWAS), we computed R^2 values using only masked genotypes from samples in that subset.

Evaluation of GWAS imputation accuracy. For computational efficiency, we performed all benchmarks of downstream imputation starting from a single data set, created as follows. First, we merged the 379 European-ancestry individuals from the 1000 Genomes Project Phase 1 integrated v3 release (see URLs) into the UK Biobank data set. Second, we entirely masked 700 random SNPs per chromosome, 100 in each of seven MAF bins (with MAF computed in the curated British samples). We phased all samples together using Eagle, and we phased a subset of $N \approx 15{,}000$ samples (all 1000 Genomes Project samples plus 10% of the UK Biobank samples) using SHAPEIT2. Finally, we used the Sanger Imputation Service to impute the $N \approx 15{,}000$ SHAPEIT2-phased samples and the same subset of Eagle-phased samples using both the UK10K panel (3,781 samples) and the HRC (r1) panel (32,488 samples) with the PBWT imputation algorithm[38] (see URLs). We assessed imputation R^2

values in $N \approx 12,000$ curated British samples at the masked and imputed SNPs, computing means and standard errors across MAF strata as before (treating R^2 values from different SNPs as approximately independent given that each MAF bin contained <1 SNP/cM). We further assessed imputation R^2 values in UK10K-imputed 1000 Genomes Project GBR samples ($N = 89$); as sequence data were available for these samples, we computed R^2 values at all UK10K-imputed SNPs in the 1000 Genomes Project data set. We computed means of R^2 values across MAF strata and estimated standard errors using a 100-block jackknife to account for LD among SNPs.

41. Henn, B.M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* **7**, e34267 (2012).
42. Huang, L., Bercovici, S., Rodriguez, J.M. & Batzoglou, S. An effective filter for IBD detection in large data sets. *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 Res.* **25**, 280–289 (2015).
44. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
45. Indyk, P. & Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. in *Proc. 30th Ann. ACM Symposium Theory Computing* 604–613 (ACM, 1998).
46. Gionis, A., Indyk, P. & Motwani, R. Similarity search in high dimensions via hashing. in *Proc. 25th VLDB Conf.* vol. **99**, 518–529 (Morgan Kaufmann Publishers, 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**, 7 (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).

doi:10.1038/ng.3571