bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

[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).

Recent work has leveraged the unique genealogical structure and extensive genotyping (>30%) 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 outbred populations by harnessing long (>4cM) identical-by-descent (IBD) tracts shared among distantly related individuals. We applied Eagle to *N*=150K 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 exquisite phasing accuracy (switch error rate ≈0.3%, corresponding to perfect phase at the scale of >10Mb). Moreover, we observed that Eagle imputed masked genotypes with accuracy R^2>0.75 down to a minor allele frequency of 0.1%. Compared to computationally tractable alternatives, Eagle attained large improvements in phasing and imputation accuracy at *N*=150K and smaller improvements at smaller sample sizes, illustrating the advantages that LRP-based imputation will yield as very large reference panels become available.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (HMM) 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. "Long-range phasing" (LRP) [13] is an alternative approach that harnesses long IBD tracts shared among related individuals; in such IBD regions, phase inference is straightforward at sites for which at least one individual is homozygous. LRP has been successfully 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, easily identified IBD tracts (>10cM) in close relatives, LRP has previously only been successfully applied in isolated populations with a large fraction of individuals genotyped. In less ideal 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 based on long (>4cM) tracts of IBD sharing in closely or distantly related individuals, and concludes with two HMM decoding iterations to refine phase calls. We demonstrate the efficiency and accuracy of Eagle by phasing N=150K samples from the UK Biobank (see URLs); at large sample sizes, Eagle matches the accuracy of the best HMM-based methods and is far more computationally efficient (e.g., 14x faster than SHAPEIT2 [12]). We also show that when phasing N=150K UK samples, Eagle imputes missing genotypes (in-sample) with accuracy R^2>0.75 down to a minor allele frequency of 0.1%, and when used to pre-phase N=150K samples within a standard imputation pipeline, Eagle improves accuracy in downstream imputation (over existing options tractable for pre-phasing N=150K samples), with larger improvements expected as imputation reference panels grow. We have released Eagle as open source software (see URLs).

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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" [13] for the proband, which can then immediately be used to call phase. While this approach is simple in principle, two major challenges have precluded its application to outbred cohorts. First, identifying IBD is difficult both in terms of accuracy and computational cost; moreover, the most widely used IBD inference methods rely on first phasing the data [30–32]. Second, LRP by itself can phase only sites at which the proband has at least one homozygous relative; for cohorts representing a sizable fraction of a population, only 5% of sites may be left unphased [13], but for smaller cohorts, this fraction may exceed 25% even in isolated populations [28], limiting the utility 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. First, Eagle rapidly detects probable IBD tracts by identifying long regions of agreement at homozygous sites (i.e., identity by state, IBS≥1), scoring identified regions using allele frequency and linkage disequilibrium information, and checking overlapping regions for consistency; Eagle uses the detected IBD to perform accurate initial long-range phasing in high-IBD regions. Second, Eagle performs local phase refinement in overlapping ≈1cM windows by detecting complementary haplotype pairs (among haplotypes inferred in the previous step) using locality-sensitive hashing [33, 34]; specifically, Eagle searches the estimated haplotypes for long haplotypes consistent with each diploid individual and then searches for hash matches to the implied complementary haplotypes. Third, Eagle finalizes phase calls by running two fast HMM decoding iterations using up to 80 local reference haplotypes and aggressively pruning the search space to ≤200 states per position. All three steps are multithreaded and make use of bit operations to perform key computations in 64-SNP blocks. (For full details, see Online Methods and the Supplementary Note.)

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Computational cost

We benchmarked Eagle against state-of-the-art phasing methods—Beagle [8], HAPI-UR [11], and SHAPEIT2 [12] (see URLs)—on subsets of the UK Biobank data set containing N=15K, 50K, or 150K samples (Online Methods). For our first benchmark, we phased only the first 40cM of chromosome 10 ($\approx$1% of the data, 5,824 SNPs spanning 18Mb) to allow as many methods as possible to complete in $<$2 weeks (using up to 10 cores on a single compute node; all methods except HAPI-UR support multithreading over 10 cores). Throughout this paper we consider 200 node-days—corresponding to a 2-day limit for our analysis of 1% of the genome—to be a reasonable practical limit for an analysis to be genome-wide tractable (understanding that computations can be parallelized across multiple compute nodes), but we allowed jobs to run for up to 2 weeks (14 node-days) in this experiment for completeness. We observed that Eagle achieved a 1–2 order of magnitude speedup over other methods across the sample size range (Fig. 1a and Supplementary Table 1), attaining a 14x speedup over SHAPEIT2 and a 12x speedup over HAPI-UR at N=150K. (Beagle was unable to phase 1% of the genome in 2 weeks at N=150K.) Eagle analysis of all N=150K samples together was genome-wide tractable, whereas SHAPEIT2 and HAPI-UR were only genome-wide tractable for batched analyses involving 10 batches of N=15K samples at a time (Supplementary Fig. 1 and Supplementary Table 1). We note that (like other methods) Eagle has parameters that produce a trade-off in speed and accuracy (Online Methods); Eagle's `--fast` mode achieved a further $\approx$2x speedup over the default while incurring only a slight loss of accuracy (Supplementary Table 2). All methods exhibited superlinear but subquadratic scaling of running time with sample size, consistent with the presence of both linear and quadratic algorithmic components. We also observed that Eagle achieved modest (2–8x) savings in memory cost compared to other methods (Fig. 1b and Supplementary Table 2). 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 British ethnicity; see Online Methods). Specifically, we included all trio children and excluded all trio parents in each

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phasing run; we then assessed computational phase accuracy in trio children at all trio-phased sites (i.e., 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=150K samples over the same 1% of the genome as above, Eagle and SHAPEIT2 achieved near-identical, remarkably low (≈0.3%) mean switch error rates (Fig. 1c and Supplementary Table 1), though we note that SHAPEIT2 is not genome-wide tractable for N=150K samples (Fig. 1a). 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 methods except SHAPEIT2 at sample sizes of N=50K and N=15K, with only a 9% increase in switch error rate relative to SHAPEIT2 at N=50K (0.78% for Eagle vs. 0.71% for SHAPEIT2). To confirm these results, we performed a similar benchmark of Eagle and SHAPEIT2 on N=60K GERA samples of more diverse European ancestry [35,36] (Online Methods) and observed similar results (0.94% switch error rate for Eagle vs. 0.83% for SHAPEIT2; Supplementary Table 3).

We next performed a comprehensive, multiple-chromosome assessment of genome-wide tractable phasing options for N=150K UK Biobank samples. Based on our running time benchmarks (Supplementary Fig. 1a), three analysis options satisfied this requirement: Eagle analysis of all N=150K samples together (Eagle 1x150K), SHAPEIT2 analysis of the data in 10 batches of N=15K samples (SHAPEIT2 10x15K), and HAPI-UR analysis of the data in 10 batches (HAPI-UR 10x15K). We benchmarked each of these methods on 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 (Table 1, Supplementary Fig. 2, and Supplementary Table 4) confirmed our previous benchmarks (Fig. 1a) and were consistent across chromosomes (Supplementary Fig. 2). In particular, we observed that Eagle analysis of all N=150K samples together completed >3x faster than SHAPEIT2 10x15K analysis while achieving a 77% decrease in switch error rate. For comparison, the publicly available UK Biobank imputation documentation (see URLs) indicates that the current UK Biobank data release was phased using an unpublished method, SHAPEIT3, that achieved ≈2x speedup and ≈15% decreased error compared to SHAPEIT2 10x15K analysis; the decision by UK Biobank to use SHAPEIT3 instead of SHAPEIT2 to analyze N=150K samples (despite the much higher error rate of SHAPEIT3 compared to running SHAPEIT2 in a single 1x150K batch) is consistent with our (somewhat arbitrary)

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

200 node-day limit for genome-wide tractability. Finally, as noted above, Eagle has parameters that produce a trade-off in speed and accuracy (Online Methods); the `--fast` mode of Eagle completed these analyses twice as quickly as the default mode with slightly higher switch error (0.36% vs. 0.31% at N=150K; Supplementary Table 2).

In-sample imputation accuracy

We next investigated the utility of Eagle for genotype imputation. 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. 3).

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). (The last two analyses are not genome-wide tractable (Fig. 1a), but we ran them on the 1% of the genome analyzed above for completeness.) 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

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 across all MAF bins (Fig. 2a and Supplementary Table 5); 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 6).

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 (Fig. 2b and Supplementary Table 7). 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 analyses) in samples who 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. (The genome-wide tractable SHAPEIT2 10x15K analysis yielded much lower in-sample imputation R^2 than Eagle 1x150K analysis for all ethnicities; Supplementary Table 7.) 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 5 and 6).

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 [37–39]). Standard GWAS imputation differs from in-sample imputation in three ways (Supplementary Fig. 3). 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 [40]), 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 al-

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lele "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 ≈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 [41]; additionally, phasing methods may not be optimized for analysis of genotype data with a uniform 2% missing rate. 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).

GWAS imputation accuracy

Lastly, we investigated the benefits of using Eagle for pre-phasing [5] within an existing imputation pipeline: the Sanger Imputation Service, which supports imputation using up to N=32K 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 but also contains samples of other ancestries; see URLs.) We considered two genome-wide tractable pre-phasing procedures: Eagle pre-phasing of all N=150K UK Biobank samples and SHAPEIT2 10x15K pre-phasing of N=150K samples (similar in accuracy to the actual SHAPEIT3 pre-phasing performed by UK Biobank; see above and URLs). To benchmark imputation accuracy, we completely masked 700 SNPs (100 in each of seven MAF bins) in each of three chromosomes, pre-phased the remaining SNPs with Eagle and SHAPEIT2, imputed the same subset of N=15K pre-phased samples using the Sanger Imputation Service, and computed R^2 between 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 N=32K HRC), Eagle pre-phasing using all N=150K samples improved imputation R^2 by increasing amounts for increasingly rare SNPs, with a gain of 0.020

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(s.e.m. 0.002) in R^2 for MAF 0.1–0.2% SNPs (Table 2). When imputation was performed using only the N=4K UK10K reference panel (see URLs), gains were roughly half as large (Supplementary Table 8). Finally, to verify that similar improvements could be obtained at genome-wide SNPs (vs. the subsets of SNPs we masked), we ran the 1000 Genomes GBR samples through the same pipeline (after pre-phasing them together with the UK Biobank samples) and again observed a modest improvement using UK10K imputation (Supplementary Table 9). (We were unable to perform this experiment using HRC imputation because the HRC contains the 1000 Genomes data.) These results demonstrate that high-accuracy pre-phasing is already beneficial for GWAS imputation at current reference sizes (N=4K UK10K samples and N=32K diverse European HRC samples) and that gains will increase as reference panels grow, consistent with our in-sample imputation results projecting future performance with N=150K reference samples.

Discussion

We have developed a fast and accurate LRP-based phasing method, Eagle, and demonstrated that LRP can be effective in an outbred population. Ever since Kong et al. [13] 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"—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 utility for enhancing the accuracy of downstream imputation.

Eagle is a very different method from the "pure" LRP approach of Kong et al. [13]: in order to create an algorithm that could harness limited, 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" [1]. Indeed, these ideas have implicitly begun to converge within sophisticated HMM-based methods (e.g., 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 MCMC sampling iterations, iteratively copying phase information from progressively more accurate sets of best reference haplotypes.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 and using them to phase long stretches of genome, and then applying only two iterations of HMM decoding to correct errors and fill in unphased regions (Supplementary Fig. 4). (For LRP in the extensively genotyped Icelandic population, only the first step was necessary [13].) 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 speedup is essential at large sample sizes: due to computational constraints, the production phasing of UK Biobank samples (using SHAPEIT3, see URLs) incurred a switch error rate that we estimate was ≈4x higher than what would have been produced by SHAPEIT2 (were SHAPEIT2 N=150K analysis genome-wide tractable). Eagle eliminates the need to compromise on accuracy, allowing computational phasing to take full advantage of very large sample sizes (achieving perfect phase at the scale of >10Mb for N=150K samples).

Beyond our immediate goal of fast and accurate phasing, we envision that the primary downstream application of Eagle will be genotype imputation in the UK Biobank and future population cohorts of similar or larger size. We have demonstrated the utility of Eagle within current imputation pipelines and the promise of this approach for use in future data sets (e.g., imputation using N=150K reference samples). However, as we noted in Results, realizing this potential will require a few additional steps. First, as currently implemented, Eagle is optimized for phasing array data and will need to be modified to phase sequence data—or integrated with methods that make use of sequencing reads [40]. Second, an imputation algorithm capable of rapidly and accurately imputing pre-phased 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 the Positional Burrows-Wheeler Transformation (PBWT) [42]—which like Eagle applies fast string matching algorithms in favor of exact statistical modeling—and the Beagle v4.1 imputation software under development [43] 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

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

underway: e.g., Genomics England plans to sequence 100,000 genomes by 2017 (see URLs).

We also anticipate several direct applications for high-accuracy phasing in large outbred populations. First, Eagle-phased haplotypes could potentially improve haplotype-based estimates of heritability explained by haplotypes tagging rare SNPs [44]. Second, Eagle-phased haplotypes could be post-processed using existing IBD-calling methods [30–32] to perform population-based linkage analysis, which previous work has indicated may require very large sample sizes to achieve genome-wide significance [45]. Third, the long IBD tracts Eagle already identifies could be used to study recent fine-scale demography [46], e.g., urbanization in the UK in the past ten generations, complementing studies of deeper British population structure [47]. Fourth, Eagle-phased haplotypes could potentially be interrogated to make inferences about recombination rate [16, 23].

While Eagle provides new levels of efficiency (and accuracy compared to tractable alternatives) for phasing very large cohorts, we note a few limitations. First, Eagle relies on the IBD present within very large data sets to achieve high accuracy; on smaller data sets (e.g., N=15K), we recommend SHAPEIT2, which provides higher accuracy and is computationally tractable for such data sets. Second, along similar lines, we observed that when phasing all UK Biobank N=150K samples together, Eagle achieved lower accuracy than SHAPEIT2 1x150K (though much higher than SHAPEIT2 10x15K, which is genome-wide tractable) on the $<$10K samples of non-European ancestry (due to limited IBD). In practice, such samples are easily detected (e.g., by using Fast-PCA [36] or SNPweights [48]) 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 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 million-sample data sets. 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, e.g., limiting the set of haplotypes considered as potential surrogate parents via clustering methods (as in SHAPEIT3; see URLs). 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.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

URLs. Eagle v1.0 software and source code,
http://data.broadinstitute.org/alkesgroup/Eagle/.

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.

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 QC Documentation, http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf.

UK Biobank Imputation Documentation (including brief description of SHAPEIT3), http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf.

GERA data set, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1.

1000 Genomes data set, http://www.1000genomes.org/.

UK10K project, http://www.uk10k.org/.

Haplotype Reference Consortium, 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/.

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.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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, Supplementary Fig. 4 provides a snapshot of the progress of the algorithm after each step for our main N=150K phasing experiment.

Step 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 main N=150K experiment, this step required ≈10% of the total computation time and achieved near-perfect phasing within long swaths of genome covering most of each sample (corresponding to regions with IBD to several relatives) (Supplementary Fig. 4a). 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 (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 [49,50]. 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.

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 [51], 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 Par-

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ente2, we use only allele frequencies and LD Scores [52] 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.

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).

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 all surrogates are also 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 and surrogate paternal haplotypes.

Step 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

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

main N=150K experiment, this step required ≈20% of the total computation time and reduced the switch error rate to ≈1.5% (Supplementary Fig. 4b). 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 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 above).

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 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 approximate nearest neighbor search in Hamming space; to do so, we apply locality-sensitive hashing (LSH) [33, 34]. 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 lowest-error complementary haplotype pair in each window (i.e., 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 (i.e., proband hets for which both surrogate haplotypes have the same allele), we defer to the surrogate with higher confidence (from step 1), and 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.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [53]). 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. In our main N=150K experiment, this step required ≈70% of the total computation time and reduced the switch error rate to ≈0.4% after the first HMM iteration and ≈0.3% after the second (Supplementary Fig. 4c,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 $O(MN)$ 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 [53] 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 (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 ≈1cM haplotypes that match the Viterbi-inferred maternal and paternal haplotypes up to at most two errors. In rare cases, the

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Fast mode of Eagle algorithm. Many parameters of the Eagle algorithm can potentially be modified to trade off 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 0.3cM to 0.5cM and reducing the comprehensiveness of the second HMM iteration (by reducing its beam search width from P=200 to 100 and only re-phasing the samples processed in the first half of the first HMM iteration).

UK Biobank data set. We analyzed data from the UK Biobank, consisting of 152,729 samples typed at ≈800K SNPs. Using PLINK2 [54]) (see URLs), we removed 480 individuals marked for exclusion from genomic analyses based on missingness and heterozygosity filters, leaving 152,249 samples (see URLs, Genotyping and QC). We restricted the SNP set to autosomal, biallelic SNPs with MAF≥0.1% and missingness ≤5%, leaving 627K SNPs (26,695 on the short arm of chromosome 1, 31,090 on chromosome 10, and 16,367 on chromosome 20). We identified 72 trios based on IBS0<0.001, sex of parents, and age of trio members (see URLs, Genotyping and QC). Of the 72 trio children, 69 self-reported British ethnicity, one self-reported Indian ethnicity, and one self-reported Caribbean ethnicity. The remaining trio child did not self-report any ethnicity, but her parents self-reported Irish and "Any other white background" as their ethnicities. UK Biobank genotyping and QC analyses indicated that self-reported ethnicity aligned closely with genetic ancestry (see URLs); however, UK Biobank also curated a subset of 120,286 self-reported British samples recommended for GWAS. Aside from having homogeneous genetic ancestry, this subset did not contain close relatives (see URLs).

GERA data set. We analyzed GERA samples (see URLs; dbGaP study accession phs000674.v1.p1) typed on the GERA EUR chip [55]. The data contained 62,318 samples, of which we removed 961 with <90% European ancestry as determined by SNPweights v2.0 (ref. [48]). Among this subset

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of samples, we identified 197 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 (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 (10 computational threads). For HAPI-UR (v1.01), we set the maximum window size to 80 (as recommended based on genotyping density) and combined results from three parallel runs of the algorithm using different random seeds [11]. We note that a new minor version of SHAPEIT v2 (r837) has been released since we performed our benchmarks; however, the change log indicates that this update only affected a feature of the software (pertaining to sequencing reads) that we did not use.

Evaluation of phasing performance. For our benchmark analyses of N=150K UK Biobank samples, we removed 144 trio parents and phased the remaining 152,105 samples. For our benchmarks on N=50K or 15K samples, we phased all 72 trio children along with 1/3 or 1/10 of the remaining non-trio parent samples (50,752 or 15,270 samples in total). We evaluated phasing accuracy in trio children by comparing computational phase calls to trio phase calls (ignoring SNPs with Mendel 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]), i.e., ≈15% of all SNPs (varying slightly among samples). In our main results, we reported mean switch error rates over the 70 European-ancestry trio children (according to self-reported ethnicity; see above). 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

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

output. For each SNP, we computed adjusted R^2 between actual and imputed masked genotype values according to the formula

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

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 over MAF strata, treating R^2 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 (e.g., the 120K British samples curated by UK Biobank for GWAS), we computed R^2 using only masked genotypes from samples in the 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 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=15K samples (all 1000 Genomes samples plus 10% of the UK Biobank samples) using SHAPEIT2. Finally, we used the Sanger Imputation Service to impute the N=15K SHAPEIT2-phased samples and the same subset of Eagle-phased samples using both the UK10K panel (3,781 samples) and the Haplotype Reference Consortium (r1) panel (32,488 samples) with the PBWT imputation algorithm [42] (see URLs). We assessed imputation R^2 in N=12K curated British samples at the masked and imputed SNPs, computing means and standard errors across MAF strata as before (treating R^2 from different SNPs as approximately independent given that each MAF bin contained <1 SNP per cM). We further assessed imputation R^2 in UK10K-imputed 1000 Genomes GBR samples (N=89); since sequence data was available for these samples, we computed R^2 at all UK10K-imputed SNPs in the 1000 Genomes data set. We computed means of R^2 across MAF strata and estimated standard errors using a 100-block jackknife to account for linkage disequilibrium among SNPs.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

33. 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).

34. 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).

35. 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).

36. Galinsky, K. J. *et al.* Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. *bioRxiv* 018143 (2015).

37. 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).

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

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

40. 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).

41. 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).

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

43. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. Platform talk presented at the 65th Annual Meeting of The American Society of Human Genetics, October 8, 2015.

44. Bhatia, G. *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases. *bioRxiv* 022418 (2015).

45. Browning, S. R. & Thompson, E. A. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* **190**, 1521–1531 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics* **97**, 404–418 (2015).

47. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

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

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

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

51. 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).

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

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

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

55. 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).

56. 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).

57. 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).

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

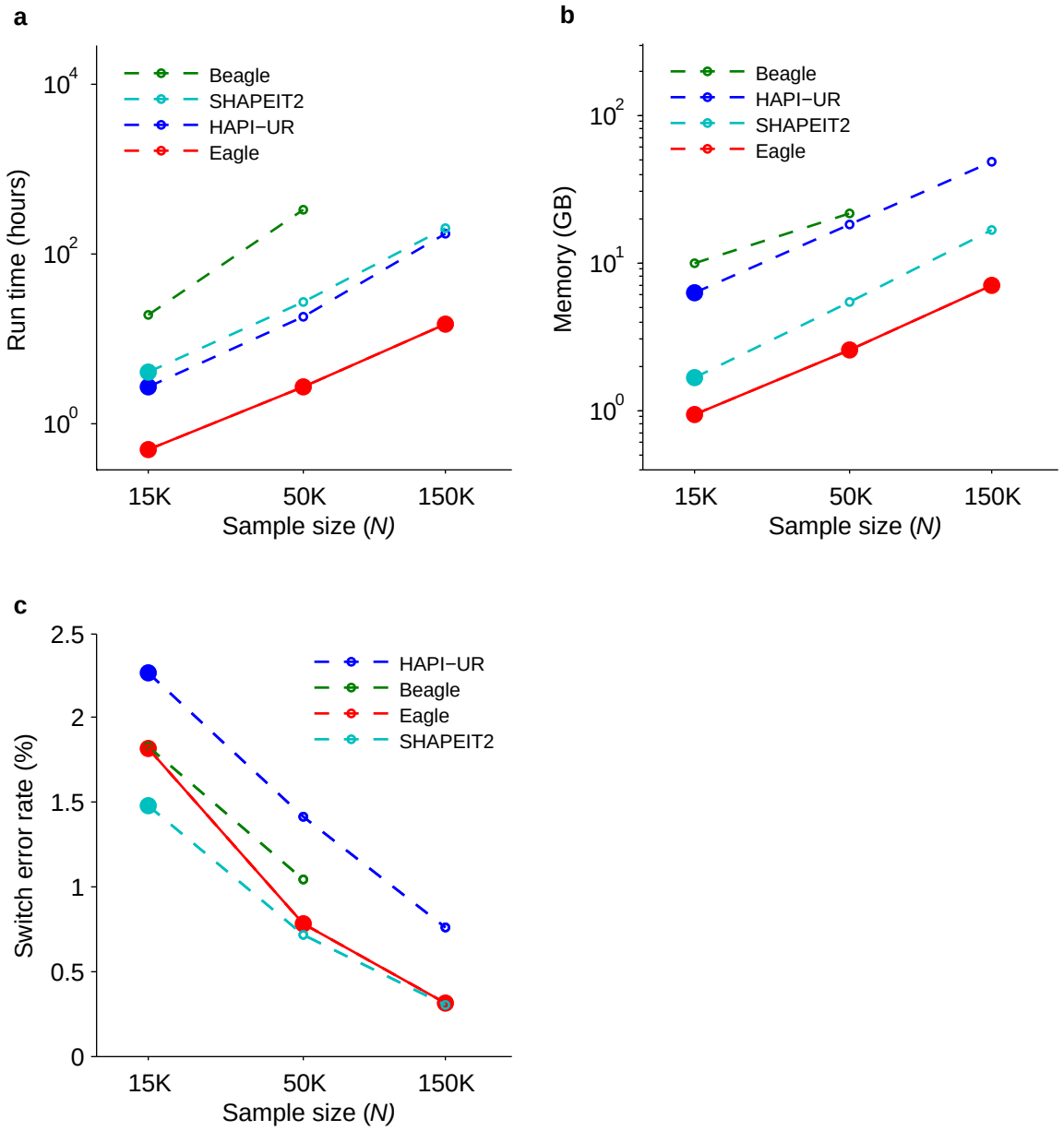

Figure 1. Computational cost and accuracy of phasing methods. Benchmarks of Eagle and existing phasing methods on N=15K, 50K, and 150K UK Biobank samples and M=5,824 SNPs on chromosome 10. (**a**) Run times and (**b**) memory using up to 10 cores of a 2.27 GHz Intel Xeon L5640 processor and up to two weeks of computation. (**c**) Mean switch error rate over 70 European-ancestry trios. Large, filled markers (connected by solid lines) indicate tractable computations for genome-wide phasing of 150K samples (i.e., projected run times of <200 days for phasing 150K/N batches of N samples genome-wide, corresponding to <2 days for M=5,824 SNPs; Supplementary Fig. 1). All methods except HAPI-UR supported multi-threading. 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 experiment, 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.) Numeric data are provided in Supplementary Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

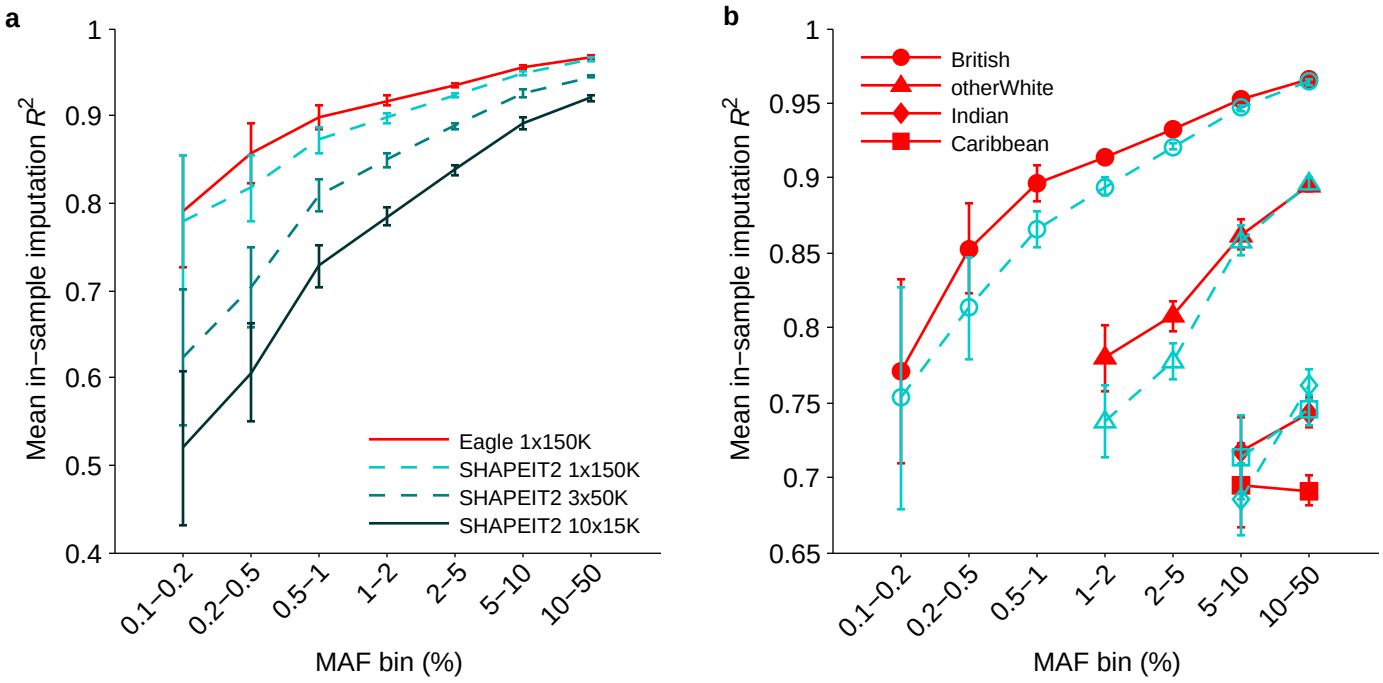

Figure 2. In-sample imputation accuracy of Eagle and SHAPEIT2. We randomly masked 2% of the genotypes in all N=150K UK Biobank samples and phased the first 40cM of chromosome 10 using Eagle (on the full cohort) and SHAPEIT2 (on all samples at once as well as in N=50K and N=15K batches), imputing all masked genotypes in the process. Solid lines indicate genome-wide tractable approaches for phasing 150K samples (i.e., methods requiring <200 node-days for phasing 150K/N batches of N samples genome-wide). **(a)** Accuracy of the imputed genotypes on the subset of 120K British 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 ethnicity, stratified by MAF in those samples. 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). The British and Irish results were near-identical (Supplementary Table 7), so we did not plot Irish results to improve readability. For the ethnicities with <5,000 samples, we plotted results only for MAF bins corresponding to an expected minor allele count ≥2 among masked samples. Error bars, s.e.m. Numeric data are provided in Supplementary Tables 5 and 7.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Computational cost and accuracy of genome-wide tractable methods in chromosome-scale analyses of *N*=150K samples.

Method	Run time	Switch error rate
Eagle 1x150K	7.9 days	0.31%
SHAPEIT2 10x15K	24.4 days	1.35%
HAPI-UR 10x15K	15.8 days	2.20%

Benchmarks of Eagle and existing phasing methods on N=150K UK Biobank samples. Reported run times are totals for phasing chromosomes 1 (short arm), 10, and 20 (using the same hardware and multithreading options as in Figure 1). Reported switch error rates are averages over the three chromosomes. Per-chromosome results are reported in Supplementary Fig. 2 and Supplementary Table 4. Run times are aggregated over the three chromosomes; we consider a method genome-wide tractable if 24 node-days are sufficient to analyze the 12% of the genome considered in this experiment. (Thus, SHAPEIT2 10x15K analysis is at our limit for genome-wide tractability.) For the SHAPEIT2 and HAPI-UR benchmarks, we phased only one 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 (Supplementary Fig. 2), exceeding our computational resources.

bioRxiv preprint doi: https://doi.org/10.1101/028282; this version posted October 4, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. 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.