Statistical phasing of 150,119 sequenced genomes in the UK Biobank

Authors

Brian L. Browning, Sharon R. Browning

Correspondence

browning@uw.edu

We present an open-source pipeline for filtering, phasing, and indexing 150,119 UK Biobank genomes. This pipeline makes it possible to apply haplotype-based methods to these data. We use the pipeline to phase 406 million single-nucleotide variants on chromosomes 1–22 and X at a cost of £2,309.

Browning & Browning, 2023, The American Journal of Human Genetics *110*, 161–165
January 5, 2023 © 2022 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2022.11.008

Statistical phasing of 150,119 sequenced genomes in the UK Biobank

Brian L. Browning[1,2,*] and Sharon R. Browning[2]

Summary

The first release of UK Biobank whole-genome sequence data contains 150,119 genomes. We present an open-source pipeline for filtering, phasing, and indexing these genomes on the cloud-based UK Biobank Research Analysis Platform. This pipeline makes it possible to apply haplotype-based methods to UK Biobank whole-genome sequence data. The pipeline uses BCFtools for marker filtering, Beagle for genotype phasing, and Tabix for VCF indexing. We used the pipeline to phase 406 million single-nucleotide variants on chromosomes 1–22 and X at a cost of £2,309. The maximum time required to process a chromosome was 2.6 days. In order to assess phase accuracy, we modified the pipeline to exclude trio parents. We observed a switch error rate of 0.0016 on chromosome 20 in the White British trio offspring. If we exclude markers with nonmajor allele frequency $< 0.1\%$ after phasing, this switch error rate decreases by 80% to 0.00032.

Genotype phasing is the inference of the two allele sequences that are inherited from an individual's parents. The UK Biobank has released whole-genome sequence data for 150,119 genomes.[1] Phasing large samples of genomes is computationally demanding, but it is desirable because phased genotypes are the input data for many powerful analyses.[2–9]

Analysis of UK Biobank sequence data is restricted to the UK Biobank Research Analysis Platform that is hosted on the Amazon Web Services (AWS) cloud.[1] Analyses performed in a compute cloud are more complex than analyses performed on a local compute cluster. Cloud-based analysis pipelines must create virtual machines, install software on the virtual machines, copy input files to the virtual machines, and copy output files to persistent storage.

Researchers must also determine an appropriate level of quality control (QC) filtering for sequence data. Phase accuracy decreases if there is inadequate QC filtering, but it can also decrease if aggressive QC filtering discards too many accurately genotyped markers. Choosing an appropriate QC filter often requires performing tests that measure phase accuracy for different levels of filtering.

In this paper we present an open-source pipeline for phasing the 150,119 genomes in the first release of UK Biobank whole-genome sequence data. The pipeline filters the sequence data with BCFtools,[10] phases the filtered sequence data with Beagle,[11] and indexes the phased sequence data with Tabix[12] (see supplemental subjects and methods).

The pipeline is simple to use and produces reproduceable results. One linux command downloads the pipeline software and genetic maps. A second command uploads the software and genetic maps to the UK Biobank Research Analysis Platform. A third command filters, phases, and indexes the sequence data for a chromosome.

The pipeline paves the way for researchers to apply haplotype-based methods to UK Biobank sequence data. The pipeline can also serve as an exemplar for phasing other large sequenced cohorts, such as the NIH All of Us Research Program.[13]

We used the pipeline to phase 406 million single-nucleotide variants on chromosomes 1–22 and X at a cost of £2,309, which is £0.0154 per genome (Table 1). The phased SNVs have a mean density of 1 SNV per 7.5 base pairs. The time for processing a chromosome ranged from 0.5 to 2.6 days. The total size of the bgzip-compressed output files was 691 GB. The cost of storing the phased output files on the UK Biobank Research Analysis platform is less than £10 per month.

The pipeline uses two types of virtual machine instances: spot instances and on-demand instances. Spot instances are less expensive, but they can be terminated at any time by the cloud provider. The pipeline uses spot instances for short-running compute jobs. If a job running on a spot instance fails due to spot instance termination, the job is rerun on an on-demand instance. When we applied our pipeline to chromosomes 1–22 and the X chromosome, approximately 5% of the jobs running on spot instances had to be rerun on on-demand instances. For the worst-case scenario in which all spot instances are terminated immediately before job completion, we estimate that the cost of applying the pipeline to chromosomes 1–22 and X would increase by approximately £1,200.

The pipeline's QC filter excludes markers with AAScore ≤ 0.95, markers with $\geq 5\%$ missing data, and non-SNV markers. This QC filter produced the highest genotype phase accuracy in the filter tests described below. If one wishes to include structural variants, the pipeline documentation

[1] Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA 98195, USA; [2] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
*Correspondence: browning@uw.edu

https://doi.org/10.1016/j.ajhg.2022.11.008.

Table 1. Cost of phasing 406,184,991 SNVs on chromosome 1–22 and X in 150,119 individuals

Task	DNAnexus compute node	Instance type	Cost (GBP)
Orchestrate analysis	mem2_ssd1_v2_x2	on-demand	39.18
Filter markers	mem1_ssd1_v2_x72	spot	322.05
Catenate files	mem1_hdd1_v2_x2	spot	1.00
Phase genotypes	mem3_ssd1_v2_x96	on-demand	1,941.03
Index phased VCF	mem1_ssd2_v2_x2	spot	5.67
Total			2,308.93

Each row describes a task, the type of DNAnexus compute node that was used, the instance type (spot or on-demand), and the cost in British pounds. See supplemental subjects and methods for more information. Spot instance can be terminated at any time by the cloud provider. Compute jobs that failed due to spot instance termination were automatically rerun on on-demand instances.

explains how to change the BCFtools[10] filter expression to include these variants. We present results below that show the differences in number of markers and phase accuracy when statistical phasing includes and excludes structural variants.

The UK Biobank sequence data contain 41 parent-offspring trios. We used the trio offspring to estimate statistical phase accuracy for different QC filters. When estimating statistical phase accuracy, we excluded trio parents before statistical phasing, and we assumed that the Mendelian phase in the offspring is the true genotype phase. The Mendelian phase of a heterozygous genotype is the phase determined by the parents' genotypes and Mendelian inheritance rules. If both parents are heterozygous or if a parent has a missing genotype, Mendelian phasing leaves the offspring heterozygote unphased.

A switch error occurs when a heterozygote is incorrectly phased with respect to the preceding phased heterozygote. A switch error can be a single switch error or a paired switch error.[14,15] A single switch error is not immediately preceded or followed by another switch error. A paired switch is immediately preceded or followed by another switch error (Figure 1). We refer to two consecutive paired switch errors as a double switch error. An isolated double switch error can be considered to be one phase error because a single heterozygote is incorrectly phased with respect to the surrounding heterozygotes (Figure 1).[14,15] A double switch error can arise when there is no information available to infer a heterozygous genotype's phase, such as when gene conversion or mutation has introduced an allele onto a new haplotypic background and only one individual in the sample has inherited the introduced allele.

We used chromosome 20 data to estimate phase accuracy, and we measured the statistical phase error rate using three metrics: switch error rate, mean megabase (Mb) distance per single switch error, and mean Mb distance per phase error, where a phase error is defined to be either a single switch error or two consecutive paired switch errors (Figure 1). The first metric (switch error rate) is the proportion of pairs of consecutive phased heterozygotes that are incorrectly phased. The second metric is the mean distance between single switch errors. Single switch errors are particularly detrimental because any two heterozygotes with an intervening single switch error will be incorrectly phased in the absence of other phasing errors. The third metric (mean Mb distance per phase errors) estimates the mean length of genomic segments with perfect haplotype phase. We exclude non-SNVs when measuring phase accuracy. This ensures that the markers used to estimate phase accuracy do not have overlapping positions on a chromosome.

We performed four analyses that investigated the impact of quality-control (QC) filters on phase accuracy. We used these analyses to select the QC filter for our pipeline. For these four analyses, we excluded markers with $\geq 5\%$ missing genotypes and we varied the minimum required AAScore. The AAScore is an estimate of the probability that a variant is a true positive.[1] At least one non-reference allele had to have an AAScore above a given threshold to pass the AAScore filter.

More than 80% of the participants in the UK Biobank are classified as White British based on principal component analysis and self-report.[16] The White British subset contains 31 sequenced parent-offspring trios. Table 2 shows chromosome 20 phase accuracy in the 31 White British trio offspring. Phase accuracy increases as the AAScore increases from 0.8 to 0.95, and there is a further increase in phase accuracy if non-SNV markers are excluded. The last two lines of Table 2 show that inclusion of structural variants increases the number of markers by 8.6% and reduces phase accuracy by approximately 10%. Phase accuracy in each analysis is estimated at the set of markers common to all four analyses (SNVs with AAScore > 0.95). The results in Table 2 show that QC filters should be applied before statistical phasing because markers with lower-quality genotypes can decrease phase accuracy at markers with higher-quality genotypes. Our phasing pipeline applies marker QC filters before statistical phasing.

When the input data for phasing are SNVs with AAScore > 0.95, the mean distance between single switch errors in the 31 White British trio offspring is 20.8 Mb, and the mean distance between phase errors is 2.3 Mb (Table 2). There are 10 additional sequenced trios that contain at least one non-White British member. In these additional trio offspring, the mean distance between single

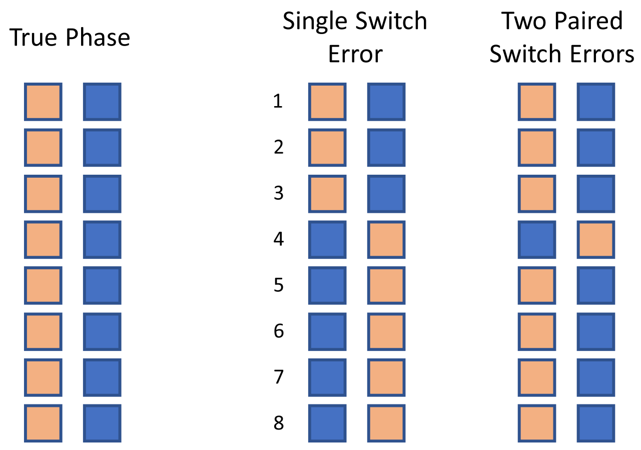

Figure 1. Single and paired switch errors
Each column of squares represents a true or estimated haplotype at eight heterozygous genotypes. Tan and blue squares represent alleles inherited from the mother and father, respectively. The left pair of columns shows the true haplotypes. A switch error is a heterozygote that is phased incorrectly with respect to the preceding heterozygote. A single switch error is a switch error that is not immediately preceded or followed by another switch error. A paired switch error is immediately preceded or followed by another (paired) switch error. The two haplotypes in the middle pair of columns have one single switch error since the fourth heterozygote is incorrectly phased with respect to the preceding heterozygote. The two haplotypes in the right pair of columns have two consecutive paired switch errors since both the fourth and fifth heterozygotes are incorrectly phased with respect to the preceding heterozygote. This figure is based on a figure in Browning and Browning.[15]

switch errors is 1.9 Mb, and the mean distance between phase errors of 0.54 Mb (Table S1). The phase accuracy in the White British samples is much higher than in the non-White British samples.

We performed another four analyses that assessed the impact of allele-frequency filters on chromosome 20 phase accuracy. The allele-frequency filters excluded markers with nonmajor allele count less than 1, 3, 30, or 300. All markers in the unphased data are polymorphic. Thus, the filter that excludes markers with <1 nonmajor allele does not exclude any markers. In each analysis, we applied the pipeline's QC filter and one of the allele-frequency filters, then we excluded trio parents, and then we phased the filtered data. We assessed phase accuracy at the set of markers common to all analyses (SNVs having a nonmajor allele count ≥ 300). Assessing accuracy at these markers is equivalent to applying an allele-frequency filter after phasing that excludes markers with nonmajor allele count less than 300.

Table 3 shows chromosome 20 phase accuracy for each allele-frequency filter in the White British trio offspring, and Table S2 shows chromosome 20 phase accuracy for each allele-frequency filter in the remaining 10 trio offspring. These results show that filtering on allele frequency before phasing with Beagle does not improve phase accuracy at SNVs with nonmajor allele count ≥300. For these data, allele frequency filters can be applied after phasing to good effect.

A comparison of the last row of Table 2 and the first row of Table 3 shows that excluding markers with nonmajor allele count less than 300 after phasing substantially improves all three measures of phase accuracy in the White British trio offspring. Excluding these markers decreases the switch error rate by 80% (from 0.0016 to 0.00032), increases the mean Mb distance per single switch error by 6.9 Mb (from 20.8 to 27.7 Mb), and increases the mean Mb distance per phase error by a factor of 4.3 (from 2.3 to 9.8 Mb). Allele-frequency filtering trades a reduction in markers for improved phase accuracy. Applying allele-frequency filters after phasing, rather than before, allows analysts to choose an allele-frequency filter that is most appropriate for a specific downstream analysis. You could use a higher-frequency threshold if you want to maximize the distance between phase errors (e.g., when detecting identity by descent segments). Alternatively, you could use a lower-frequency threshold if the downstream analysis can tolerate a shorter distance between phase errors and you are interested in lower-frequency variants (e.g., when detecting variants influencing allele-specific expression).

Most of the improvement in phase accuracy from allele-frequency filtering is due to a decrease in paired switch errors. Excluding markers with nonmajor allele count less than 300 reduced the mean number of paired switch errors per White British trio offspring on chromosome 20 from 49.7 to 8.4. There is a smaller reduction in single switch errors per offspring (from 3.1 to 2.3) because exclusion of low-frequency markers after phasing eliminates a single switch error only if all phased heterozygotes between a single switch error and the preceding or succeeding single switch error are excluded.

These measures of phase accuracy assume that the Mendelian phase in trio offspring is the true phase. However, genotype error creates both Mendelian phase errors and spurious heterozygotes.[15] The true switch error rate for the White British samples at SNVs with nonmajor allele count ≥300 could be significantly lower than the observed switch error rate reported in Table 3.[15]

We anticipate that we will be able to phase much larger data sets with Beagle. Beagle's computation time and memory requirements scale linearly with sample size. Larger data sets can be phased if Beagle's memory requirements are less than the computer's available memory. Beagle's memory requirements are approximately proportional to its sliding window length. By default, Beagle uses a 40 cM sliding window with 2 cM overlap. The window length can be reduced with little loss of phase accuracy because the 2 cM overlap ensures that the phase of each heterozygote is informed by all markers within a 1 cM radius. Reducing the window length can result in longer run times because more windows are required to cover a chromosome and overlap regions are phased twice. However, if the reduced window length is at least twice the overlap length (i.e., >4 cM if using a 2 cM overlap), the overlap regions are phased at most twice, and the increase in phasing time is less than a factor of 2.

Table 2. **Effect of AAScore filtering on phase error rates in 31 White British trio offspring**

AAScore	Exclude non-SNVs	Markers	SER	Mb / single switch error	Mb / phase error
> 0.80	no	12,288,985	0.0021	8.8	1.7
> 0.90	no	11,307,099	0.0019	11.2	1.9
> 0.95	no	9,592,309	0.0017	18.4	2.1
> 0.95	yes	8,833,023	0.0016	20.8	2.3

After marker filtering, statistical phase was inferred in 150,041 UK Biobank participants who are not trio parents. Statistical phase accuracy was then calculated in trio offspring for 8,833,023 chromosome 20 SNVs with AAScore > 0.95 under the assumption that the Mendelian phase is the true phase. For each analysis, the table reports the AAScore threshold, the inclusion status of non-SNVs, the number of filtered markers, the switch error rate (SER), the mean Mb distance per single switch error, and the mean Mb distance per phase error. A switch error is a heterozygote that is phased incorrectly with respect to the preceding heterozygote. A single switch error is a switch error that is not immediately preceded or followed by another switch error. A phase error is a single switch error or two consecutive switch errors.

The number of polymorphic sites increases with sample size for sequence data, but the additional polymorphic sites do not consume much additional memory because the additional sites have low nonmajor allele frequency. For low-frequency markers, Beagle stores in memory only the indices of samples and haplotypes that carry each nonmajor allele.

Beagle can phase the UK Biobank sequence data on a virtual machine having 50% less memory than the virtual machines used in our pipeline (384 GB instead of 768 GB). To confirm this, we used Beagle to phase all markers on chromosome 1, including structural variants, that had AAScore > 0.95 and < 5% missing genotypes. We ran Beagle on a virtual machine with 384 GB of memory and 96 CPU cores, and we used a 30 cM sliding window length. Based on this analysis, we estimate that Beagle could phase more than 1 million individuals if we reduce the window length by a factor of 4 (from 30 cM to 7.5 cM) and increase the memory by a factor of 4 (from 384 GB to 1,536 GB). Virtual machines with up to 1,952 GB of memory and 128 CPU cores are currently available on the Research Analysis Platform.

Increasing the sample size from 150,000 individuals to 1 million individuals would increase pipeline run times. One way to reduce these run times would be to divide each chromosome into overlapping segments, phase the chromosome segments in parallel, and then merge the haplotype segments for each individual to obtain the chromosome-wide phasing.

Our estimate that data sets with 1 million genomes can be phased is based on current methods and technology. In the past, the arrival of larger data sets has spurred the development of more efficient phasing methods.[11,14,17] In addition, cloud providers have periodically introduced new classes of virtual machines that have more memory and CPU cores. We anticipate that future improvements in methods and computational resources will allow researchers to phase data sets with millions of genomes. Larger sample sizes will make it possible to achieve even lower phase error rates.[15]

Data and code availability

The pipeline and software for phasing the 150,019 sequenced UK Biobank samples are available at https://github.com/browning-lab/ukb-phasing/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.11.008.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 19934. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number HG008359. The content is solely the responsibility of the authors

Table 3. **Effect of allele frequency filtering on phase error rates in 31 White British trio offspring**

Nonmajor alleles	Markers	SER	Mb / single switch error	Mb / phase error
≥ 1	8,833,023	0.00032	27.7	9.8
≥ 3	3,784,979	0.00034	22.4	9.2
≥ 30	855,024	0.00034	22.4	9.0
≥ 300	318,273	0.00035	20.8	8.6

After marker QC and allele frequency filtering, statistical phase was inferred for chromosome 20 markers in 150,041 UK Biobank participants who are not trio parents. Statistical phase accuracy was then calculated in trio offspring for 318,273 chromosome 20 SNVs with nonmajor allele count ≥300 under the assumption that the Mendelian phase is the true phase. For each analysis, the table reports the nonmajor allele count threshold before phasing, the number of filtered markers, the switch error rate (SER), the mean Mb distance per single switch error, and the mean Mb distance per phase error. A switch error is a heterozygote that is phased incorrectly with respect to the preceding heterozygote. A single switch error is a switch error that is not immediately preceded or followed by another switch error. A phase error is a single switch error or two consecutive switch errors.

and does not necessarily represent the official views of the National Institutes of Health.

Declaration of interests

The authors declare no competing interests.

Received: October 3, 2022
Accepted: November 8, 2022
Published: November 29, 2022

Web resources

UK Biobank Research Analysis Platform, https://ukbiobank.dnanexus.com

References

1. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150, 119 genomes in the UK Biobank. Nature *607*, 732–740.
2. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.
3. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
4. Ramstetter, M.D., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Mezey, J.G., and Williams, A.L. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. Genetics *207*, 75–82.
5. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A fast and simple method for detecting identity-by-descent segments in large-scale data. Am. J. Hum. Genet. *106*, 426–437.
6. Zhou, Y., Browning, B.L., and Browning, S.R. (2020). Population-specific recombination maps from segments of identity by descent. Am. J. Hum. Genet. *107*, 137–148.
7. Browning, S.R., and Browning, B.L. (2020). Probabilistic estimation of identity by descent segment endpoints and detection of recent selection. Am. J. Hum. Genet. *107*, 895–910.
8. Browning, S.R., Waples, R.K., and Browning, B.L. (2022). Fast, accurate local ancestry inference with FLARE. Preprint at bioRxiv.
9. Wohns, A.W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2022). A unified genealogy of modern and ancient genomes. Science *375*, eabi8264.
10. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.
11. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890.
12. Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718–719.
13. The All of Us Research Program Investigators, Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The "All of Us" research program. N. Engl. J. Med. *381*, 668–676.
14. Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.
15. Browning, B.L., and Browning, S.R. (2022). Genotype error biases trio-based estimates of haplotype phase accuracy. Am. J. Hum. Genet. *109*, 1016–1025.
16. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
17. Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436.