The American Journal of Human Genetics, Volume *110*

Supplemental information

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

Brian L. Browning and Sharon R. Browning

Supplemental Information

Supplemental Subjects and Methods

UK Biobank data

The first release of UK Biobank whole genome sequence data includes 150,119 individuals.[1] The mean sequence coverage is 32.5, and the sequence coverage per individual is at least 23.5.[1] We analyzed version 12.1 of the UK Biobank data.

The UK Biobank has classified 125,363 of the sequenced individuals as White British based on self-report and principal component analysis.[2] The sequenced genomes include 41 parent-offspring trios. The White British subset contains 31 of these trios. The remaining 10 trios have at least one member who is not classified as White British. The 41 trios include 78 distinct parents.

Phasing pipeline

The UK Biobank Research Analysis Platform is hosted on the Amazon Web Services compute cloud. Users can interact with the Research Analysis Platform via the DNAnexus web interface or the DNAnexus command line utility. The phasing pipeline uses the DNAnexus command line utility.

The pipeline is invoked once for each chromosome. Chromosomes can be processed in parallel. The pipeline has four steps: marker filtering, file concatenation, genotype phasing, and file indexing.

The marker filtering step filters a chromosome's unphased genotype data with BCFtools 1.10.2.[3] The unfiltered, unphased sequence data for a chromosome are stored in hundreds of VCF files, each of which contains a 50 kb interval of genotype data for all sequenced individuals. We process batches of 100 input VCF files that contain data for consecutive genomic intervals, and we use one virtual machine for each batch. The marker filter excludes markers that are not SNVs, markers with ≥ 5% missing data, and markers with AAScore ≤ 0.95. For each batch, the marker filtering step creates one bgzip-compressed output file containing the filtered genotypes. The output file for the first batch on a chromosome includes the VCF

header lines so that concatenating the batch output files for a chromosome in genomic order will produce a valid VCF file. We use virtual machines that have 36 CPU cores and 72 GB of memory for the marker filtering step.

The file concatenation step concatenates the filtered files for a chromosome using the linux cat command. This step does not require any data compression or decompression because the concatenation of bgzip files is a bgzip file. The output is a single, bgzip-compressed VCF file that contains filtered, unphased genotype data for a chromosome. We use a virtual machine that has 2 CPU cores and 3.75 GB of memory for this step.

The genotype phasing step phases the chromosome's genotypes using Beagle 5.4 with default parameters on a virtual machine with 96 CPU cores and 786 GB of memory.[4]

The file indexing step indexes the phased VCF file for the chromosome with tabix 1.10.2-3 on a virtual machine with 2 CPU cores and 4 GB of memory.[5]

A separate orchestrator program runs the four steps and ensures that each step is completed before beginning the next step. The orchestrator program runs on a virtual machine with 2 CPU cores and 4 GB of memory.

Instructions and software for running the pipeline are available in a public GitHub repository (github.com/browning-lab/ukb-phasing). The repository contains a README file with instructions, genetic maps for each chromosome, two shell scripts, and five DNAnexus applets. One shell script copies the software to the DNAnexus platform, and one shell script executes the pipeline. The DNAnexus applets perform the work. There is one applet for task orchestration and one applet for each step of the workflow. Each applet creates a virtual machine, installs software on the virtual machine, copies input files to the virtual machine, runs the analysis, and copies output files to object storage.

The pipeline uses two types of virtual machines: spot instances and on-demand instances. Spot instances have a lower cost, but they may not be available when a job is submitted, and they can be terminated at any time by the cloud provider. Our pipeline uses spot instances for marker filtering, file concatenation, and VCF file indexing. If a spot instance running one of these jobs is terminated or if a spot instance is not available within 15 minutes after job submission, the job is automatically run on an on-demand instance.

Supplemental Tables

Table S1: Effect of AAScore filtering in non-White-British trio offspring

AAScore	Include non-SNVs	Markers	SER	MB / single switch error	MB / phase error
> 0.80	Yes	12,288,985	0.0063	1.5	0.48
> 0.90	Yes	11,307,099	0.0059	1.7	0.52
> 0.95	Yes	9,592,309	0.0060	1.9	0.52
> 0.95	No	8,833,023	0.0057	1.9	0.54

Table S1: Effect of AAScore filtering on chromosome 20 phase error rates in 10 non-White-British trio offspring. 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.

Table S2: Effect of allele frequency filtering in non-White-British trio offspring

Nonmajor allele threshold	Markers	SER	MB / single switch error	MB / phase error
≥ 1	8,833,023	0.0015	2.4	1.7
≥ 3	3,784,979	0.0014	2.6	1.8
≥ 30	855,024	0.0015	2.6	1.7
≥ 300	318,273	0.0015	2.4	1.7

Table S2: Effect of allele frequency filtering on chromosome 20 phase error rates in 10 non-White-British trio offspring. 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.

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. 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.
3. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience 10.
4. 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.
5. Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27, 718-719.