# Phasing of the UK Biobank whole genome sequencing data interim release of 200,031 samples

Version 1.1

25 July 2023

Diogo Ribeiro*, Robin Hofmeister*, Simone Rubinacci*, Olivier Delaneau

Department of Computational biology, University of Lausanne, Switzerland

* Equal contribution

Correspondence should be addressed to olivier.delaneau@gmail.com

Latest version of this report: LINK

Bullet point summary:

- Data undergoes thorough quality control (QC)
- Statistical phasing of entire chromosomes performed using SHAPEIT5
- Phasing process considers available family information
- Phasing process accounts for haploidy of males on chromosome X
- Confidence levels provided for rare variants during phasing
- Haplotypes validated using family data and imputation
- Two imputation pipelines available: one for SNP array data and another for low coverage whole genome sequencing (WGS) data.

## Table of contents

## 0. Introduction

In this document we describe the phasing of 200,031 whole genome sequences (WGS) of the UK Biobank using the SHAPEIT5 software (Hofmeister, Ribeiro, Rubinacci, Delaneau, et al., 2022) (https://odelaneau.github.io/shapeit5/). This dataset is developed for use by any researcher with access approval to the UKB.

Statistical haplotype phasing is a procedure applied to distinguish the two inherited chromosome copies into haplotypes, as this information is not readily available through sequencing, but can be accurately predicted in large cohorts by leveraging identity-by-descent tracks. Phasing information unlocks several analyses, such as detecting combinations of heterozygous variants present in both copies of a gene (compound heterozygous events) and determining parent-of-origin effects (Hofmeister, et al., 2022). Moreover, accurate phasing of reference panels shows to improve genotype imputation (Hofmeister, Ribeiro, Rubinacci, Delaneau, et al., 2022; Rubinacci et al., 2022). SHAPEIT5 was specifically developed to phase large whole genome sequencing datasets, such as the UK biobank cohort. One of its main features is its high accuracy in phasing rare variants, allowing users to increase the pool of genetic variation in downstream analyses. broadening of the repertoire of genetic variation accessible for downstream analyses, thereby enhancing the potential for uncovering a wider range of genetic insights from the data.

In the following sections we describe the protocol used to phase chromosomes 1 to 22 and X on the set of 200,031 UKB genomes, leading to more than 684 million phased variants (SNPs and indels). We describe the quality control measures adopted prior to phasing, the phasing pipeline utilised, the validation and performance assessment of the haplotypes. Additionally, we provide a description of the imputation pipelines released to utilise the phased data as a haplotype reference panel. These pipelines enable the imputation of SNP array (Rubinacci et al., 2020) and low-coverage WGS (Rubinacci et al., 2021, 2022).

## 1. Dataset quality-control

The starting point is the interim WGS data release of the UK Biobank with approximately 200k samples, pVCF files (Population level WGS variants, pVCF

format - interim 200k release, field 24304) (Halldorsson et al., 2022). These consist of genotype calls from GraphTyper (Eggertsson et al., 2017) which include a AAscore field per variant, based on a logistic regression model that predicts the probability of being a true positive. This metric varies from 0 to 1, with values towards 1 indicating high quality. As the accurate phasing is dependent on the genotype call quality, we examined the quality of variants depending on the AAscore in several orthogonal ways, with the aim of defining an optimal AAscore filter threshold for the phasing procedure.

We started by exploring the AAscores of all 22,027,524 variants (SNPs and indels) present on chr17 across the 200k samples (before any filter). We observed that most variants have high AAscores (mean AAscore = 0.915, median = 0.968). We then used a baseline filter of AAscore >= 0.5, as used by the data producers (Halldorsson et al., 2022) and Hardy-Weinberg equilibrium (HWE) p-value < $1e^{-30}$ (calculated only on Caucasian individuals), which excluded 2,172,344 variants (9.86% of the variants). Of the remaining 19,855,180 variants, 18,088,253 were SNPs and 1,766,927 were indels, presenting markedly different distributions of AAscores (**Figure 1**, mean AAscore for SNPs = 0.950, for indels = 0.854). An AAscore threshold of >=0.8 excludes 10.9% of all variants, whereas a threshold of >0.95 excludes up to 37.1% of all variants.

### 1.1 AAscore and mappability *(D. Ribeiro)*

We observed that AAscores are largely correlated with read mappability, when visually comparing mean AAscores across 10kb bins and mappability tracks from Single-read and multi-read mappability after bisulfite conversion (Bismap, k24, GRCh38) (Karimzadeh et al., 2018). Lower AAscore stretches clearly matched drops in mappability across all chromosomes (**Figure 2a**). These often occur in and around centromeres, or in the short arms of acrocentric chromosomes. Of note, drops in AAscore are largely reduced when filtering for AAscore > 0.8 across chromosomes (**Figure 2b, Supplementary Figure 1**).

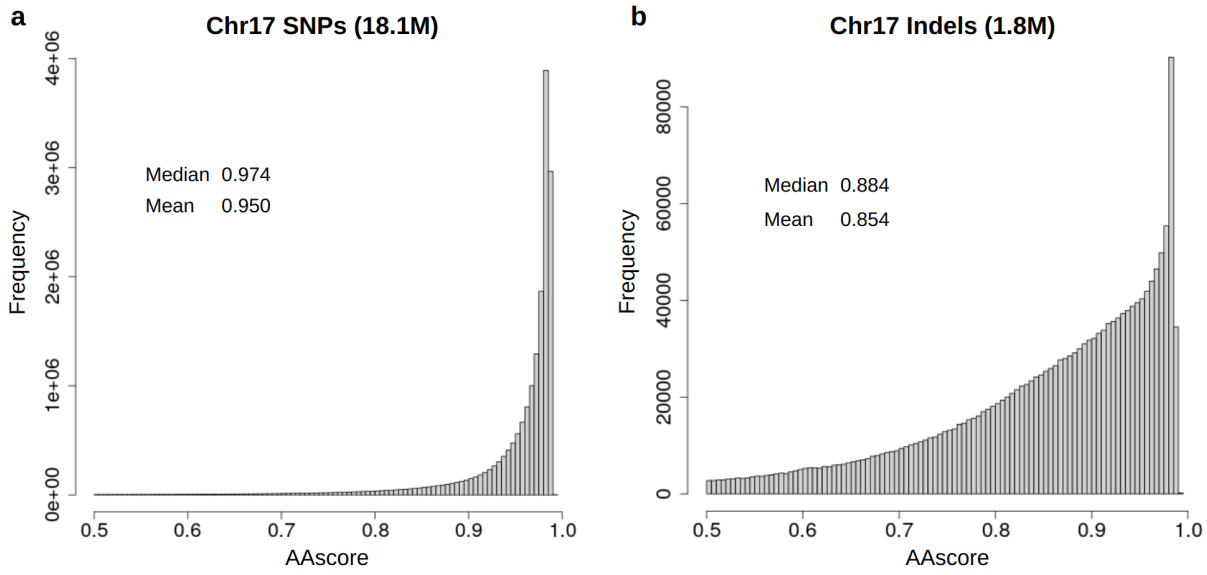



**Figure 1. Distribution of variant AAscores. (a)** for SNPs (N = 18,088,253)**, (b)** for indels (N = 1,766,927). Only variants with AAscore > 0.5 filter were considered.

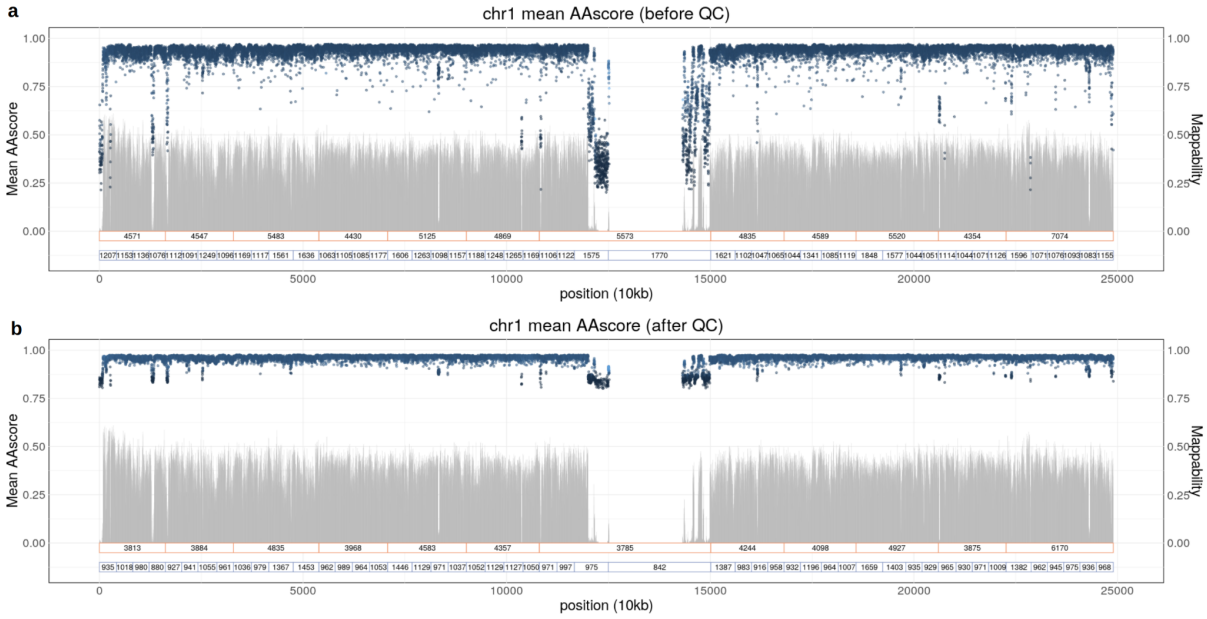



**Figure 2. Mappability and mean AAscore per 10kb bins in chromosome 1. (a)** before AAscore and HWE filters, **(b)** after AAscore > 0.8 and HWE filters. Mappability based on Bismap k24 shown in grey. AAscore shown in black (low variant density in 10kb bin) and blue (high variant density). Red rectangles show the chunks for SHAPEIT5 phase_common step, and blue rectangles the chunks for phase_rare, along with the total number of variants in the chunk (in thousands).

### 1.2 AAscore and Mendel error rates *(D. Ribeiro)*

To evaluate genotype calls and the AAscore metric, we measured the number of Mendel inconsistencies (disagreement between genotype and pedigree data) in genotype calls between 93 trios (mother-father-child) present in the UK Biobank 200k WGS dataset. We focused on 201,283 common variants (159,018 SNPs, 42,265 indels) with minor allele frequency (MAF) > 5% in chromosome 17. By counting the number of genotype inconsistencies (e.g. child presenting an allele not present in the parents), we observed that variants with low AAscores (e.g. 0.5 to 0.6, ~1.5% error rate), had approximately 6-fold higher error rates than variants with high AAscores (e.g. 0.8 to 0.9, 0.25% error rate, **Figure 3**). Variants with very high AAscores (>0.95) show error rates below 0.1%. Mendel inconsistency was estimated with an in-house tool (otools_2023-03-09_006c7e1) on BCF files filtered by AAscore > 0.5 and HWE p-value < $1e^{-30}$ (calculated only on Caucasian individuals).

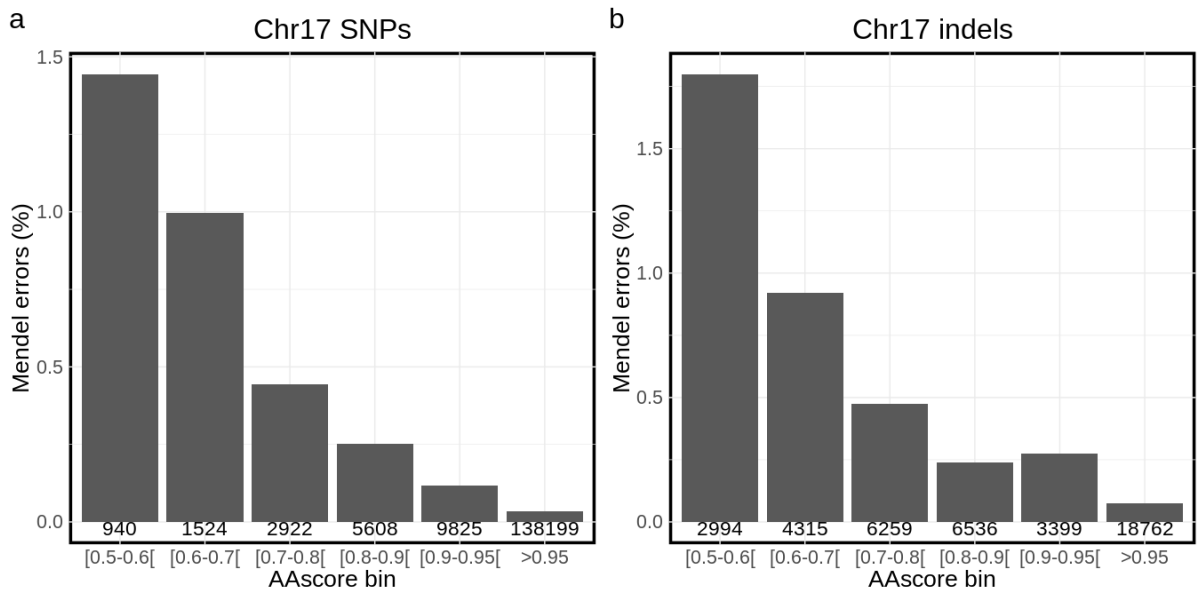



**Figure 3. Mendel inconsistency errors in UKB trios depending on AAscore. (a)** in 159,036 chr17 SNPs, **(b)** in 42,265 chr17 indels. Mendel Error percentage is calculated as the sum of inconsistencies divided by the number of variants/trios with alternative alleles.

### 1.3 AAscore and SNP array concordance *(D. Ribeiro)*

A key question when filtering variants is how many commonly assessed SNP array variants are being filtered out. We thus evaluated the overlap between the 200k WGS variants with the UK Biobank Axiom SNP array (field 22418, liftover to b38), depending on the AAscore filter. Overall, out of 22,669 SNP array variants in chr17, 22,615 (99.8%) are present in the WGS dataset, prior to any filtering (matching by chromosome, position, reference and alternative alleles). When filtering for variants with AAscore >=0.8, the vast majority of SNP array variants (22,317, 99.5%) are kept. We then compared the non-discordance rate (NRD, using bcftools stats -S) between WGS and SNP array genotypes (**Table 1**). As expected, we observed a higher NRD for variants with low AAscore (e.g. 2.09 for AAscore >=0.6 and <0.7), whereas NRD show to be several times lower in variants with high AAscore (e.g. 0.51 for AAscore >=0.9 and <0.95).

| AAscore bin | # variants | NRD | R/R Discord | R/A Discord | A/A Discord |
|---|---|---|---|---|---|
| >=0.5 & <0.6 | 7 | 7.54 | 2.97 | 0.45 | 0.30 |
| >=0.6 & <0.7 | 19 | 2.09 | 0.55 | 1.71 | 0.38 |
| >=0.7 & <0.8 | 68 | 1.35 | 0.04 | 1.21 | 0.08 |
| >=0.8 & <0.9 | 186 | 3.22 | 1.90 | 1.28 | 0.17 |
| >=0.9 & <0.95 | 419 | 0.51 | 0.09 | 0.33 | 0.23 |
| >=0.95 | 21498 | 0.27 | 0.06 | 0.10 | 0.11 |

**Table 1. Discordance rates between WGS and SNP array genotypes in chr17, depending on AAscore bin.** Non-reference (NRD), reference-reference (R/R), reference-alternative (R/R) and alternative-alternative (A/A) are shown.

### 1.4 AAscore and gnomAD variant overlap *(D. Ribeiro)*

Similarly to SNP arrays, we compared the overlap between the UKB 200k WGS variants and gnomAD database (Chen et al., 2022). For this, we gathered chr17 data from gnomAD v3.1.2 (WGS sites in 76,156 samples), retaining only 230,864 common sites in non-Finnish European samples (sites with MAF_nfe>0.05, and

AC_nfe>1000, with ‘PASS’ flag). We compared the overlap between these 230,864 gnomAD sites and the 201,283 chr17 UKB common sites (MAF 5%, no sample filter). Overall, 90.5% of the UKB variants with AAscore >=0.5 are present in gnomAD. This proportion increases to 93.3% and 95.3% of variants when considering AAscore >=0.8 and >=0.95, respectively. The likelihood of a variant being present in gnomAD increases with the AAscore (**Figure 4**), but remains relatively low for indels (<75%), possibly due to differences in indel calling between the softwares used.

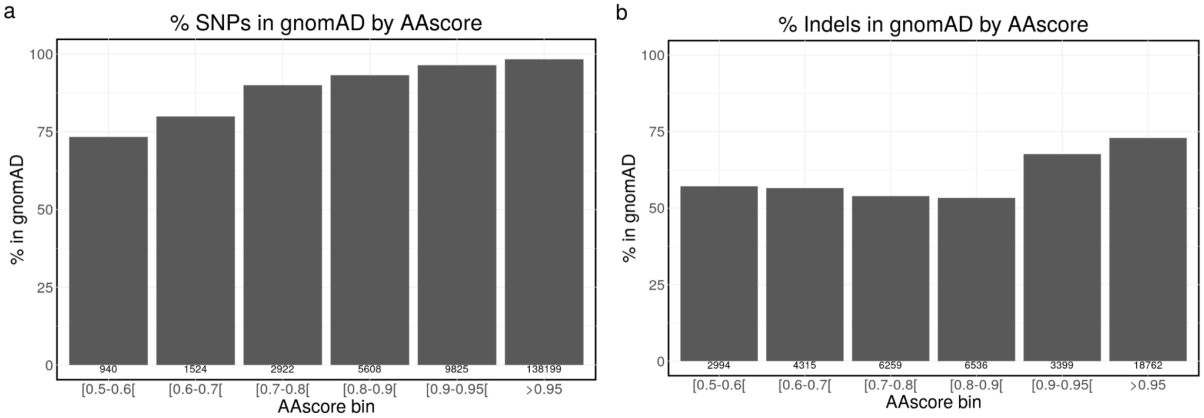



**Figure 4. Percentage of UKB WGS chr17 common sites in gnomAD depending on AAscore. (a)** in 159,018 chr17 SNPs, **(b)** in 42,265 chr17 indels.

We then compared the presence of gnomAD SNPs in the UKB WGS dataset. When focusing on a subset of 115,213 common SNPs (i.e. no indels) in gnomAD (AC_nfe > 10000, MAF_nfe>0.1), the UKB WGS dataset contains 99.7% of these sites before any filtering. When applying AAscore filtering of >=0.5 (and no other filtering) this is reduced to 97.9% and to 93.3% with AAscore >=0.8 (**Table 2**). The reason for this decrease may be due to different strigency in filter: the AAscore filtering in UKB may be more stringent than the “PASS” filtering of gnomAD, even for common variants.

### 1.5 AAscore and chrX heterozygous rate *(D. Ribeiro)*

As the chromosome X in males should be haploid (except in the PAR1 chrX:10,001-2,781,479 and PAR2 chrX:155,701,383-156,030,895 regions), and diploid in females, we evaluated heterozygous rates depending on AAscore. We expected that erroneous genotype calls inflate the number of heterozygous genotypes in males (males defined with UKB Field 22001). Focusing on 183,205

chrX common SNPs (MAF 5%, no indels, excluding PAR regions), for each individual we calculated the number of heterozygous genotypes by the total number of genotypes (with bcftools stats -s -). We calculated this heterozygous rate for different sets of SNPs depending on AAscore (**Table 3**). Reassuringly, high AAscores have very low heterozygous rates in males (0.05% for AAscore >=0.95). However, heterozygous rates in males varied per AAscore bin, with as much as 4.17% genotypes being heterozygous in SNPs with AAscore between 0.5 and 0.6. As a control, we measured heterozygous rates in females, which show an opposite trend, with increasing heterozygosity with higher AAscore (from 10.1% to 28.7%).

| AAscore filter | % gnomAD SNPs in UKB WGS |
|---|---|
| No filter | 99.7% (N = 114885) |
| >=0.5 | 97.8% (N = 112640) |
| >=0.8 | 93.2% (N = 107400) |
| >=0.95 | 83.4% (N = 96117) |

**Table 2. presence of gnomAD common SNPs in the UKB WGS dataset, depending on AAscore filter.** 115,213 gnomAD variants with MAF > 10% were used for this analysis.

| AAscore bin | # SNPs | Male heterozygous rate | Female heterozygous rate |
|---|---|---|---|
| <0.5 | 19527 | **7.57%** | 10.08% |
| >=0.5 & <0.6 | 3052 | **4.17%** | 12.07% |
| >=0.6 & <0.7 | 3599 | **2.39%** | 13.74% |
| >=0.7 & <0.8 | 5631 | **1.41%** | 16.60% |
| >=0.8 & <0.9 | 9601 | **0.86%** | 19.79% |
| >=0.9 & <0.95 | 12909 | **0.38%** | 24.36% |
| >=0.95 | 148413 | **0.05%** | 28.65% |

**Table 3. Male and female heterozygous rate on chrX common SNPs depending on AAscore.** Heterozygous rate is calculated per individual as the number of heterozygous genotypes among all genotypes for a certain set of variants (depending on AAscore).

We next compared the mean number of heterozygous genotypes (in chrX non-PAR regions) between males and females for SNPs with AAscore >=0.95 (**Figure 5**). We find a clear discrepancy, with an average 0.05% of heterozygous rate in males, compared to 28.7% in females. This metric allowed us to pinpoint samples that may have misassigned genetic sex or aneuploidies in the dataset (e.g. 898 males with heterozygous rates > 0.11%, 1101 females with heterozygous rates < 0.5%). This aided our decision to include only the 99% most chrX heterozygous females for the calculation of HWE p-values in chrX sites. In addition, when phasing chrX, we considered the 99% least chrX heterozygous male individuals as being haploid, as explained in below.

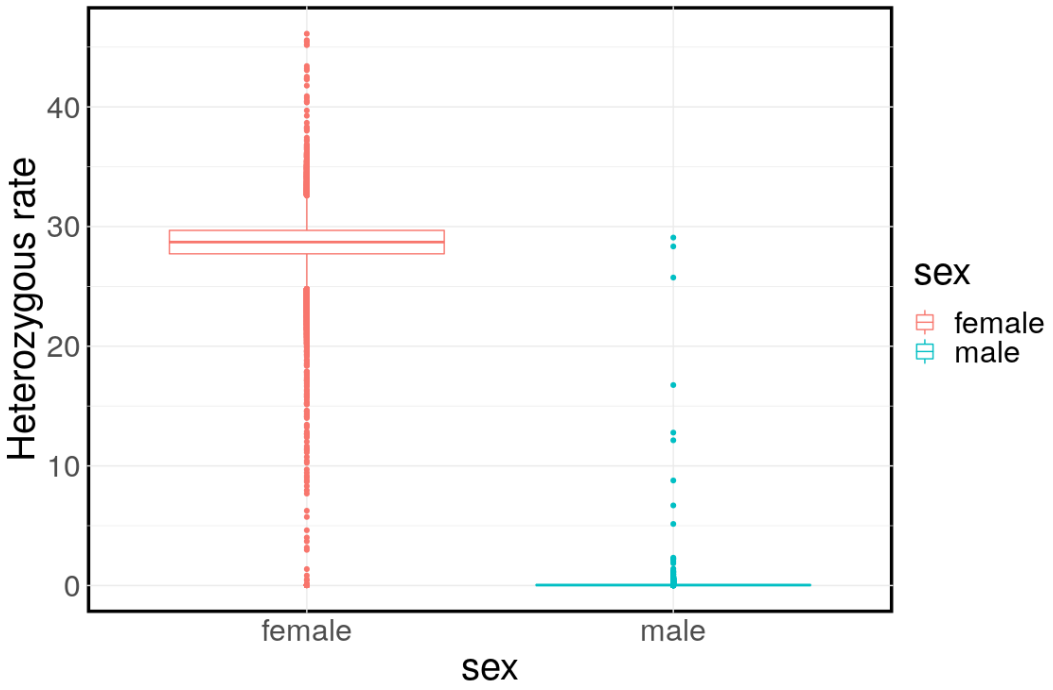



**Figure 5. Heterozygous rates in males and females.** Calculated on 148,413 SNPs with AAscore > 0.95. Females N = 110,024, Males N = 89,839.

### 1.6 AAscore and imputation accuracy *(R. Hofmeister)*

We next evaluated the impact of AAscore filtering on imputation accuracy, as imputation is a common application of phased reference panels. For this, we performed phasing and imputation experiments across three AAscore cutoffs (0.5, 0.8 and 0.95) on chr17 SNP array data from 995 British unrelated individuals using IMPUTE5 v1.1.5 (Rubinacci et al., 2020) (see section 3.2). We found that

considering variants with lower AAscores results in a slightly decreased imputation accuracy (**Figure 6**). Since the phasing procedure considers all variants in a chunk, a key question is whether including variants with low AAscores impacts the phasing of variants with higher AAscore, and thus their imputation accuracy. To assess this, we performed phasing with different AAscore cutoffs, as before, but measured imputation accuracy only for variants with AAscore > 0.95. We found that considering variants with lower AAscores does not affect imputation accuracy of sites with AAscore > 0.95 (**Figure 7**). As a result, the phased dataset can be filtered for AAscore a posteriori.

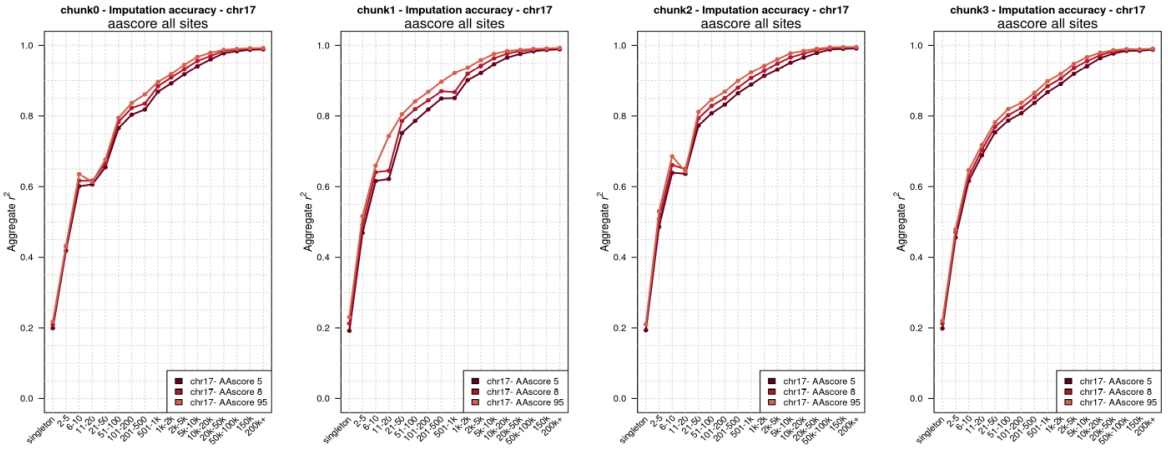



**Figure 6. Imputation accuracy depending on AAscore.** Imputation accuracy (r-squared, y axis) stratified by Minor Allele Count (MAC) bins (x axis) for chr17 split in 4 chunks. A panel per phasing chunk, each line representing a different AAscore filter (0.5, 0.8 and 0.95).

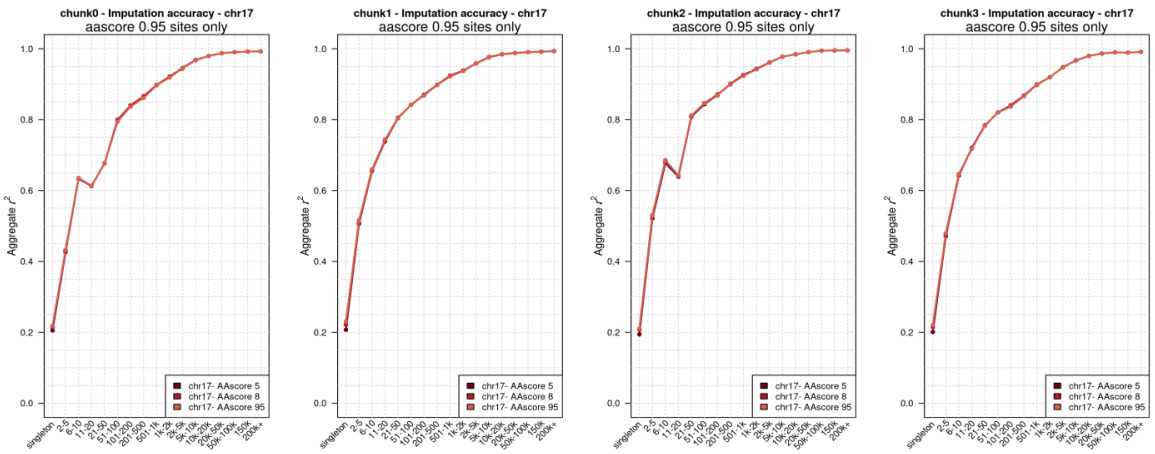



**Figure 7. Imputation accuracy and AAscore filters using sites with AAscore >=0.95.** Imputation accuracy (r-squared, y axis) stratified by Minor Allele Count (MAC) bins (x axis) for chr17 split in 4

chunks. A panel per phasing chunk, each line represents a different AAscore filter. In this figure, we filtered the phased data after the phasing procedure to keep only variant sites with AAscore>=0.95.

### 1.7 Data processing and filtering *(R. Hofmeister, S. Rubinacci, D. Ribeiro)*

Below we describe the filtering settings used for selecting which variants to phase. Overall, we retained more than 684 million SNPs and indels for phasing (**Table 4**), ~90% of the variants per chromosome. Most of the approximately 10% variants excluded were filtered out by AAscore, with a smaller amount excluded due to Hardy-Weinberg disequilibrium or other filters. The number of samples used are described in **Table 5**.

- *AAscore:* Through the above orthogonal evaluation of AAscores (mappability, Mendel inconsistencies, overlap with SNP array and gnomAD datasets, chrX heterozygous rates and imputation accuracy), we can conclude that AAscore is well calibrated with variant quality. We considered that a AAscore threshold of 0.5 is too lenient and may comprise our phasing procedure. While variant quality is highest with stringent thresholds (e.g. >=0.95), this threshold excludes 37.1% of all variants, which limits downstream analysis. Given this, we decided on a AAscore threshold of >=0.8, as a compromise between variant site quality (low error rates on most metrics) and number of sites excluded (7.85% of all variants).
- *Hardy-Weinberg equilibrium (HWE):* We excluded variants with HWE p-value $<1e^{-30}$, calculated on 113,637 unrelated Caucasian samples. For chrX, we calculated HWE only on the 99% (N=61,387) female unrelated Caucasian samples with the highest heterozygous rates (see section 1.5 for details).
- *Other filters:* Besides the previous filters, we also excluded variants with genotype missingness above 10% (F_MISSING > 0.1, using bcftools), i.e. >10% individuals with missing data for that variant. In addition, we excluded variants without a “PASS” in the VCF FILTER field and excluded variants where the alternative allele was missing (i.e. set as “*”). Of note, a higher proportion of variants are excluded with the “PASS” filter on chrX compared to autosomes. In addition, chrX PAR1 and PAR2 regions were not phased, which partially explains the higher rate of unphased variants in chrX.

- *Conversion from multi-allelic to bi-allelic:* To deal with multi-allelic sites in the input files, we converted multi-allelic to bi-allelic sites using the bcftools *norm* function (with -m -any parameters). This was executed before all above filters. This ensures compatibility of the data with all tools commonly used in the field.

| chr | Phased SNPs | Phased indels | Unphased SNPs | Unphased indels | Unphased % | Unphased Filtered by AAscore % | Unphased Filtered by HWE % |
|---|---|---|---|---|---|---|---|
| 1 | 49,028,590 | 3,508,622 | 3,624,714 | 1,938,101 | 10.59% | 92.21% | 5.18% |
| 2 | 54,348,858 | 3,839,699 | 3,062,538 | 1,969,373 | 8.65% | 91.05% | 4.14% |
| 3 | 45,302,488 | 3,220,355 | 2,042,704 | 1,565,866 | 7.44% | 89.34% | 5.33% |
| 4 | 43,546,711 | 3,139,277 | 2,164,061 | 1,482,616 | 7.81% | 88.80% | 5.62% |
| 5 | 40,657,162 | 2,885,161 | 2,010,525 | 1,412,059 | 7.86% | 90.60% | 4.92% |
| 6 | 38,165,066 | 2,802,268 | 1,661,946 | 1,374,375 | 7.41% | 89.24% | 5.36% |
| 7 | 35,671,487 | 2,541,459 | 2,346,964 | 1,382,369 | 9.76% | 91.63% | 5.11% |
| 8 | 35,062,502 | 2,342,549 | 1,922,519 | 1,141,357 | 8.19% | 88.48% | 5.34% |
| 9 | 27,058,701 | 1,833,852 | 2,826,239 | 1,042,437 | 13.39% | 92.85% | 5.18% |
| 10 | 29,898,454 | 2,104,663 | 1,790,582 | 1,128,444 | 9.12% | 90.99% | 5.25% |
| 11 | 30,622,290 | 2,119,160 | 1,566,212 | 1,068,945 | 8.05% | 89.84% | 5.70% |
| 12 | 29,414,835 | 2,156,598 | 1,505,020 | 1,131,797 | 8.35% | 90.03% | 5.95% |
| 13 | 21,733,330 | 1,621,803 | 994,358 | 779,048 | 7.59% | 89.18% | 5.66% |
| 14 | 19,846,065 | 1,446,113 | 1,242,319 | 757,879 | 9.39% | 90.78% | 4.20% |
| 15 | 18,075,979 | 1,290,116 | 1,566,423 | 745,600 | 11.94% | 92.77% | 6.14% |
| 16 | 20,178,833 | 1,263,556 | 1,864,945 | 782,296 | 12.35% | 92.21% | 4.04% |
| 17 | 17,372,974 | 1,286,557 | 1,361,870 | 814,046 | 11.66% | 90.76% | 6.01% |
| 18 | 17,047,126 | 1,237,219 | 984,130 | 610,784 | 8.72% | 91.28% | 8.16% |
| 19 | 13,101,931 | 981,149 | 1,035,254 | 662,906 | 12.06% | 90.78% | 4.70% |
| 20 | 14,075,021 | 972,197 | 869,612 | 539,884 | 9.37% | 91.12% | 7.59% |
| 21 | 7,885,781 | 578,110 | 1,044,333 | 344,096 | 16.40% | 92.56% | 6.24% |
| 22 | 8,055,945 | 566,895 | 1,243,662 | 392,371 | 18.97% | 93.02% | 6.23% |
| X | 23,248,080 | 1,551,508 | 4,040,602 | 1,131,225 | 20.85% | 55.27% | 2.54% |
| **Total** | **639,398,209** | **45,288,886** | **42,771,532** | **24,197,874** | **9.78%** | **88.11%** | **5.14%** |

**Table 4. Summary of phased and unphased variants per chromosome.** The percentage of unphased variants filtered by AAscore 0.8 and HWE are reported as a percentage of the total of unphased variants (SNPs and indels). Note that a variant can be excluded by several filters (e.g. a variant excluded by HWE and AAscore, PASS filter or F_MISSING filters).

| Dataset | Female | Male | Total | Total Caucasian | Total Other |
|---|---|---|---|---|---|
| All samples | 110,024 | 89,839 | 200,031 | 167,117 | 32,914 |
| Unrelated | 63,762 | 76,245 | 140,175 | 113,637 | 26,538 |
| Duos | 542 | 373 | 915 | 789 | 126 |
| Trios | 59 | 34 | 93 | 85 | 8 |

**Table 5. Summary of samples used from the UKB 200K WGS release.** Note that 168 samples have no genetic sex attributed.

## 2. Phasing procedure

We phased all the sequencing data in 4 steps as follows using SHAPEIT5. A full documentation of SHAPEIT5 can be found here.

1. We first phased the common variants (MAF >= 0.1%) using the **phase_common** program. For the UK Biobank WGS data, it is recommended to split the chromosome into large chunks (e.g. 20 cM) to ensure good accuracy while speeding up the computations.
2. We ligated the phased chunks at common variants (MAF >= 0.1%) of a chromosome using the **ligate** program. The ligation step is computationally efficient and uses variants in the intersection of the chunks to provide chromosome-wide haplotypes. The result of this step is used as a haplotype scaffold for the next step.
3. We phased rare variants (MAF < 0.1%) using the **phase_rare** program. To do this, we use the haplotype scaffold generated in step 2 and we proceed in relatively small chunks (e.g. 5Mb) to run many small jobs in parallel. At the end of this step, we have several phased chunks.
4. We concatenated all the phased chunks generated in step 3 using bcftools concat -n. As in the previous step haplotypes have been phased onto a haplotype scaffold, there is no need to ligate the chunks, and the files can be directly concatenated without decompression and recompression. This makes this step almost instantaneous, even for large cohorts.

For the phasing of the UK Biobank WGS 200k, we used SHAPEIT v5.1.0. Details of the parameters and the command lines we run can be found here (LINK). We utilized parent-offspring information to inform the phasing of common and rare variants when

possible (#trios=93 and #duos=915) using the SHAPEIT5 option *--pedigree* (see 2.2.5). This ensures that the first and second offspring haplotypes come from parent 1 and 2, respectively. We also enforced haploidy for male on chromosome X using the SHAPEIT5 option *--haploid* (see 2.2.6).

### 2.1 Chunking *(S. Rubinacci)*

To phase the UK Biobank WGS 200k, we used (i) large chunks (~20 cM) to phase common variants (step 1) and (ii) small chunks (~4Mb) to phase rare variants (step 3). To produce the small chunks for the step 3 (phasing rare variants), we use the GLIMPSE2_chunk tool as follows:

```
GLIMPSE2_chunk      --input ${bcf_input}
                    --output sites.txt
                    --window-cm 4
                    --window-mb 4
                    --window-count 30000
                    --buffer-cm 0.5
                    --buffer-mb 0.5
                    --buffer-count 3000
                    --sequential
                    --map chr${CHR}.b38.gmap.gz
                    --region chr${CHR}
```

This command ensures that the chunks are at least 4cM long and 5Mb long (when including the buffer region) and includes at least 30,000 common variants (MAF >0.1%). These criteria ensure that chunks overlapping centromeres contain enough variants to retain a good phasing and imputation accuracy. To produce the large chunks for the step 1 (phasing common variants), we manually merged the small chunks produced with the above command four by four. Chunks used for the UK Biobank WGS 200k phasing are available here. These can be used for any other analysis requiring chunking on the GRCh38 assembly.

### 2.2 Phasing *(R. Hofmeister)*

#### 2.2.1 Phasing common variants

To phase common variants we used a total of 145 chunks of on average 25 cM. We ran each phasing chunk as a single job on a *mem1_ssd1_v2_x72* instance. On average, each job took 6.5 hours and cost £4 on spot. A relatively small number of

jobs (n=4) ran on-demand and cost on average £16. The total cost for this step was about £630.

### 2.2.2 Ligate common variants

To ligate phased chunks from the previous step, we used the *SHAPEIT5 ligate* tool. We ran one job per chromosome on the *mem1_ssd1_v2_x8* instance. Jobs cost between £0.017 and £0.1 and lasted between 20 and 104 minutes (for chromosome 22 and chromosome 1, respectively).

### 2.2.3 Phasing rare variants

To phase rare variants on the autosomes, we used a total of 558 chunks of on average 6.5 cM. To optimise the running time and cost of this process, we grouped those chunks into 70 sets of 8 chunks and ran each set on the *mem3_ssd1_v2_x32* instance with 4 jobs in parallel using 8 threads each (with *xargs* -P 4). This allowed us to have the 70 jobs running in parallel on the UK Biobank RAP. Each of the 70 jobs took on average less than 9 hours and cost less than £2 when running on spot. (11 ran on-demand <£14 and <11 hours). The total cost for this step was about £272. For each rare heterozygous genotype, we reported the phasing confidence as the probability that the minor allele sits on the haplotypes reported in the FORMAT/GT field. This probability is given in the FORMAT/PP field and ranges from 0.5 (full uncertainty) to 1.0 (best phasing quality). Note that singletons were systematically given a score of 0.5.

### 2.2.4 Concatenating

We finally concatenated all phased chunks to obtain chromosome-wide phased files. For this, we used the *bcftools concat -n* command. We ran one job per chromosome on the *mem1_ssd1_v2_x16* machine. Jobs cost between £0.078 and £0.48 and lasted between 45 and 277 minutes (for chromosome 22 and chromosome 1, respectively).

### 2.2.5 Family phasing

When available, we leveraged knowledge of family relationships (parent-offspring relationships, duos/trios) to improve phasing accuracy. In addition, this also allows us to perform inter-chromosomal phasing, meaning that the first and second haplotypes

correspond to the same parents across the 22 autosomes (in our case, the first haplotype is the paternal one and the second haplotype is the maternal one). To perform this family phasing, we used the *--pedigree* option in SHAPEIT5. This takes as input a pedigree file that contains one line per sample having parent(s) in the dataset and three columns (offspringID fatherID and motherID), separated by TABs. Use NAs is allowed for unknown parents (in the case of duos). Family data is used to fix the phase of offspring heterozygous genotypes when possible, that is when (i) there is no Mendel inconsistency, and (ii) at least one parent is homozygous. In other words, it builds a scaffold of haplotypes for offspring from the parental genomes.

#### 2.2.6 Chromosome X phasing

For chromosome X phasing we first remove the PAR regions. We used the option *--haploid* to specify the list of male individuals to consider as haploid. We first identified putative males as genetically determined (UK Biobank field 22001). We additionally excluded individuals having sex chromosome aneuploidy (UK Blobank field 22019). Finally, we excluded from the list 898 males with heterozygosity > 0.109% (see section 1.5). As a result, we obtain a list of 88,877 males that we used as input for the *--haploid* option. Individuals that are not on the list are phased as diploids, in the same way as autosomes.

## 3. Phasing validation

To assess the accuracy of our phased haplotypes, we used different strategies. The first one leverages available family information. The second one is agnostic of family and leverages the fact that an accurate imputation relies on an accurately phased reference panel.

### 3.1 Phasing validation using family information *(R. Hofmeister)*

A standard approach is to leverage parent-offspring trios and duos in the data to measure the switch error rate (SER) in the offspring (phased without pedigree information). The SER measures how close estimated and true haplotypes are and The SER is defined as the fraction of successive pairs of heterozygous genotypes being correctly phased.

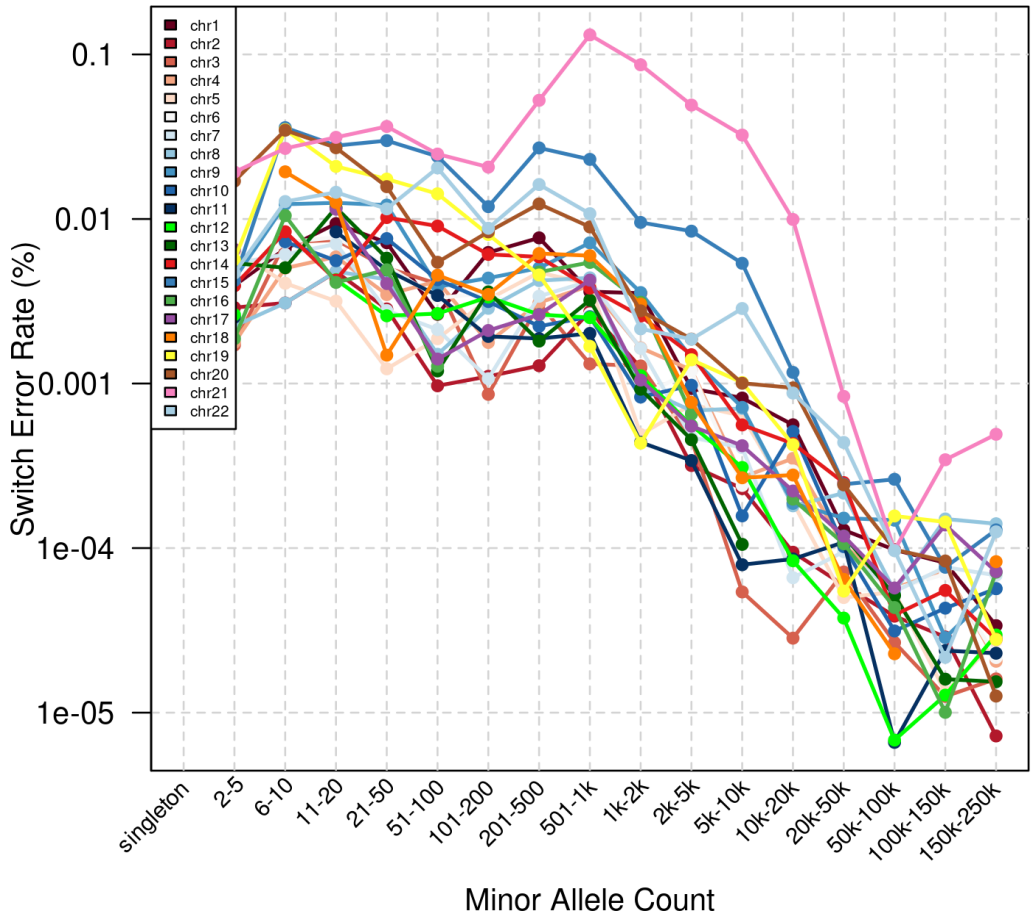



**Figure 8. Phasing Switch Error Rates (SER).** Phasing switch error rate (y-axis) stratified by MAC (x-axis) across the 22 autosomes. Computed using 93 trios and 915 duos.

In the context of this work, offspring have been phased together with their parents, so that many of their heterozygous genotypes are phased using mendel logic. We can therefore not use the approach described before. Instead, we looked at the switch error rates in the parents, under the assumption of no-recombination (which is true in most of the cases). This does not provide an unbiased estimate of the SER but instead a validation of the statistical phasing we performed. We should expect indeed extremely low switch errors if the statistical model is able to phase the parents conditioning on the offsprings haplotypes. We therefore measured SER stratified by bins of minor allele count (MAC). We assigned each heterozygous genotype to a given MAC bin and counted the fraction of heterozygous genotypes being correctly phased per MAC bin. This definition of SER has the advantage of showing how well statistical phasing performs depending on the frequency of the variants it phases (either common or rare). On average across the 22 autosomes in

93 trios and 915 duos, the phasing switch error rate (SER) is below 0.01% (**Figure 8; Supplementary Figure 2**). Notably, the highest SER are found in acrocentric chromosomes (e.g. chr15 and chr21), for which the phasing of the short arm is more challenging. The computation of SER was performed with the SHAPEIT5 ***switch*** program.

### 3.2 Phasing validation by imputation *(R. Hofmeister, S. Rubinacci)*

As genotype imputation accuracy directly improves with appropriate reference panel phasing, we evaluated imputation performance across autosomes, chrX and several difficult genomic regions.

#### 3.2.1 Autosome-wide

It is well established that genotype imputation performs better with an accurate phasing of the reference panel. We therefore leverage this feature to assess the accuracy of our phasing through genotype imputation. For this, we randomly selected 1,000 individuals of white British ancestry that are unrelated to any other UK Biobank participant. We built a reference panel for genotype imputation by removing these individuals from our haplotype callsets. We then used the available UK Biobank Axiom array data as input for genotype imputation, from which we imputed only our selected 1,000 individuals using IMPUTE5 v1.1.5 (Rubinacci et al., 2020). We ran the imputation in chunks of ~25cM (i.e, we used the same chunking as for the phasing of common variants; see section 2.1). Finally, we computed the phasing accuracy using the concordance tool of GLIMPSE v2.0.0 (Rubinacci et al., 2022). This imputation accuracy (aggregate $r^2$) is represented chromosome-wide and stratified by Minor Allele Count (MAC) in **Figure 9**. We observed that imputation accuracy increases with MAC as expected and is highly consistent across chromosomes.

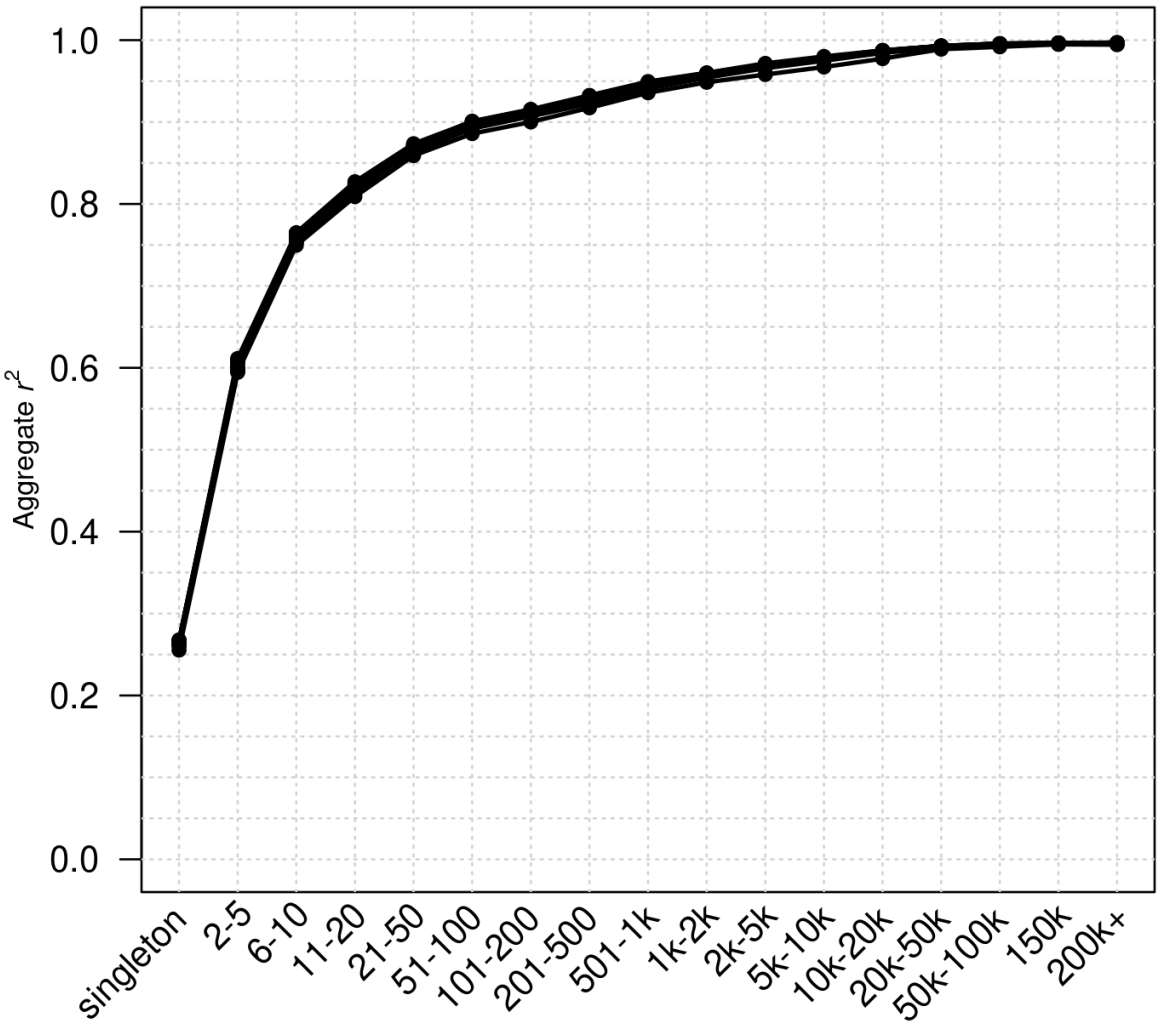



**Figure 9. Imputation accuracy.** Imputation accuracy ($r^2$, y-axis) stratified by MAC (x-axis) across the 22 autosomes. Each black line represents an autosome.

### 3.2.2 Region specific

We next assessed the accuracy of imputation per chunk, both for the large and the small chunks used to phase common and rare variants, respectively (see section 2.1). We observed that some chunks have lower imputation accuracy than others (**Supplementary Figure 3**). To investigate further, we modified the chunking to highlight the centromeric regions, as well as the regions before and after the centromere (**Supplementary Figure 4**). We found that drops in imputation accuracy occur for the first chunks of acrocentric chromosomes (e.g, chromosomes 21 and 22) as well as for chunks overlapping large centromeres (e.g, chromosome 9,

**Supplementary figure 3**). These chunks usually contain only a few variants and do not really impact the overall accuracy across the chromosome, as seen in **Figure 9**.

### 3.2.3 Chromosome X

We finally assessed imputation accuracy on chromosome X, which used a specific phasing parameter to account for variable ploidy. However, we observed a genotype imputation accuracy on chromosome X similar to that of autosomes (**Figure 10**).

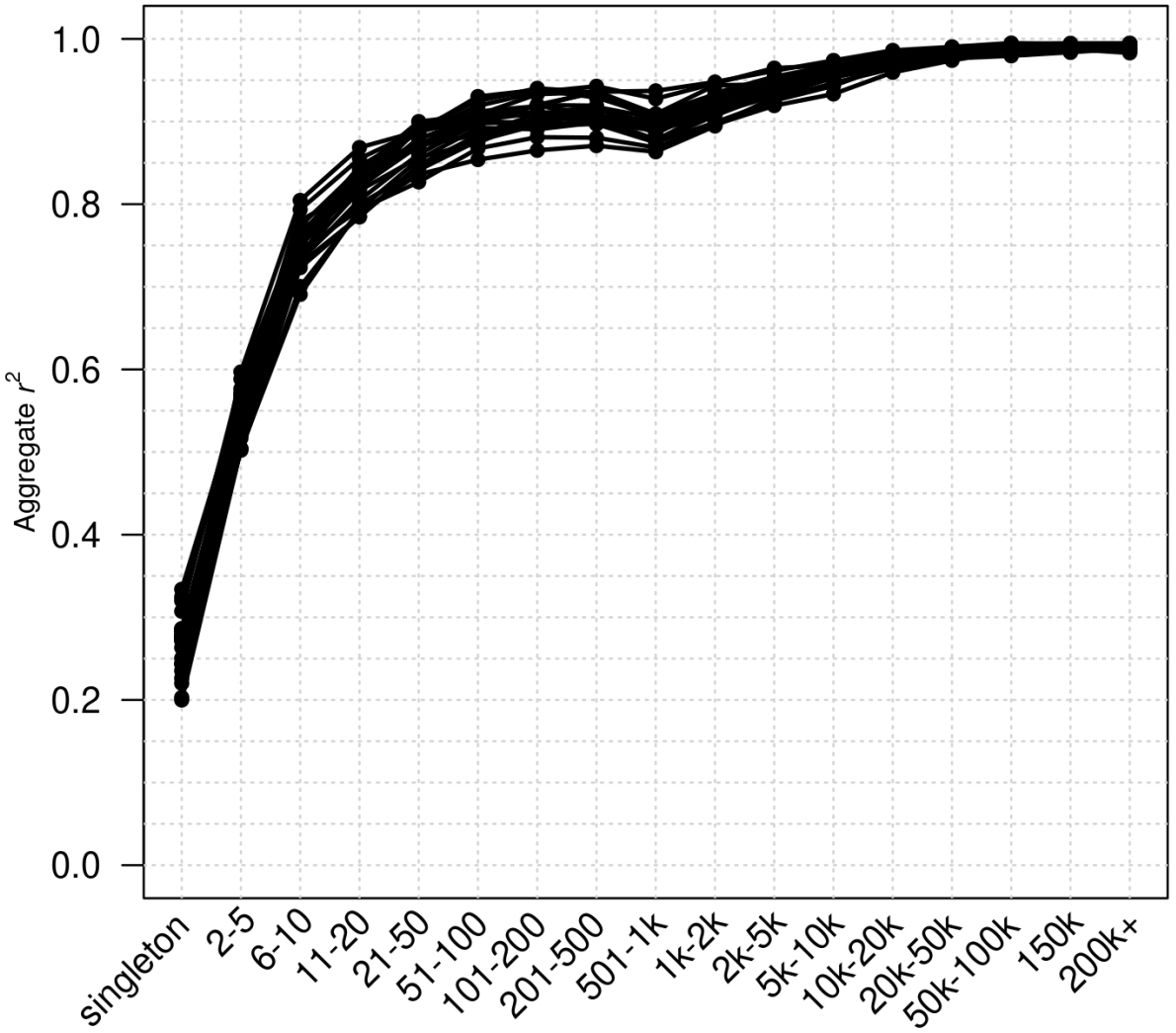



**Figure 10. Imputation accuracy on chromosome X.** Imputation accuracy ($r^2$, y-axis) stratified by Minor Allele Count (x-axis) across the small chunks of ~6cM on chromosome X. Each black line represents a chunk.

## 4. Data availability *(D. Ribeiro)*

We provided phased pVCF files for 200,011 individuals and about 684 million SNPs and indels, one file per chromosome (1-22 and X), together with index files. We have simplified the INFO field and only provide a phased GT field as well as the phasing probability (PP) of each genotype. We excluded 20 samples that withdrew from the UK Biobank study at the time the phasing data was returned to the UK Biobank (samples were excluded after filtering and phasing procedures). Besides this, we provide indexed BCF and TSV files comprising all input WGS variants and their original INFO field (from GraphType), with an added PHASED tag, representing whether this variant has been phased or not (which is dependent on our quality-control).

## 5. Released pipelines *(S. Rubinacci)*

### 5.1 Introduction

Genotype imputation is a computational technique used to estimate missing genotypes in SNP array data. It involves using a reference panel of haplotypes to predict the missing genotypes. This process can also be applied to low-coverage whole genome sequencing data, where it helps to fill-in missing genotypes or improve uncertain genotype calls obtained from sequencing reads.

We have developed pipelines that employ the UK Biobank reference panel to perform genotype imputation for both SNP array and low-coverage whole genome sequencing data. To accomplish this, we utilise efficient state-of-the-art tools such as IMPUTE5 (Rubinacci et al., 2020) for SNP array imputation and GLIMPSE2 (Rubinacci et al., 2021, 2022) for low-coverage WGS imputation. Our pipeline takes input from either a multi-sample VCF/BCF file containing genotypes from SNP arrays or a set of low-coverage BAM/CRAM files. Using the UK Biobank reference panel generated as described in the previous sections, the pipeline performs imputation by running applets and dx command jobs, specifically designed for the UKB RAP.

At the end of each of the imputation pipelines, a single multi-sample BCF file per chromosome is produced, containing genotype posteriors, dosages and phased

best-guess genotypes. Additional output (e.g. haploid dosages) can be obtained by specifying additional options to the imputation software using the appropriate option.

## 5.2 Low-coverage WGS imputation pipeline

The inputs of the low-coverage imputation pipeline are described in **Supplementary Table 1**. The workflow begins with a set of BAM/CRAM files low-coverage WGS reads uploaded on the UKB RAP. The pipeline's default parameters are optimised for cost-efficiency with few hundreds samples, but adjustments are necessary for larger sample sizes (e.g., using larger computing instances). It is recommended to run at least 100 samples in a single batch.

The low-coverage WGS imputation pipeline consists of two modules:

- **Convert reference**: this module is the first step when setting up the pipeline and it is run only once. Its purpose is to convert the phased pVCF files from the RAP into the binary representation of GLIMPSE2 using GLIMPSE_split_reference.
- **Imputation**: This module is responsible for performing genotype imputation using GLIMPSE2. It is the most computationally intensive task in the pipeline, taking in input the binary reference panel files and a set of BAM/CRAM files. At the end of the imputation step, a single ligation step is performed to provide chromosome-level phased genotypes.

## 5.3 SNP array imputation pipeline

The SNP array imputation pipeline requires specific inputs, which are described in **Supplementary Table 2**. The workflow begins with a multi-sample VCF/BCF file containing chromosome-wide SNP array data, which can be phased or unphased. It is important that the samples in the file are genotyped using the same SNP array and have a broad European origin matching the UK Biobank reference panel. The pipeline's default parameters are optimised for cost-efficiency with a few hundred samples, but adjustments are necessary for larger sample sizes (e.g., using larger computing instances). It is recommended to run at least 100 samples in a single batch.

The pipeline utilizes IMPUTE5, a licensed software owned by the University of Oxford. Its usage is permissible for academic purposes in accordance with the software's license.

The SNP array imputation pipeline consists of four modules that perform subsequent tasks. Here is a short description of each module:

- **Convert reference**: this module is the first step when setting up the pipeline and it is run only once. Its purpose is to convert the phased pVCF files from the RAP into the XCF sparse format used by SHAPEIT5 and IMPUTE5.
- **Prephasing**: this module is used for prephasing the data using SHAPEIT5 phase_common. It is necessary when working with unphased SNP array data. The output of this module is a phased target file in the XCF binary format.
- **Convert target**: this module is employed when the SNP array data has already been phased and uploaded to the RAP. In such cases, this module performs the conversion to the XCF binary format.
- **Imputation**: This module is responsible for performing genotype imputation using IMPUTE5 v1.2.0. It is the most computationally intensive task in the pipeline, taking in input XCF files for the reference and target panels. At the end of the imputation step, a single ligation step is performed to provide chromosome-level phased genotypes.

## 5.4 Software

The SNP array and low-coverage pipelines utilise the following software tools:

- XCFTOOLS v1.0.0 [LINK]: Used in the SNP array pipeline.
- SHAPEIT v5.1.0 [LINK]: Used in the SNP array pipeline.
- GLIMPSE v2.0.0 [LINK]: Used in the low-coverage pipeline.
- IMPUTE5 v1.2.0 [LINK]: Used in the SNP array pipeline. IMPUTE5 is a licensed software owned by the University of Oxford. Its usage is permissible for academic purposes in accordance with the software's licence.

## 5.5 Price estimates

**Table 5** provides price estimates for spot instances. It is well-documented in the literature that SNP array imputation is more efficient compared to low-coverage

WGS imputation. However, utilising our GLIMPSE2-based pipeline, we are able to impute a sample from the 200k reference panel for approximately 0.08 £ per sample. For SNP array imputation, leveraging the latest updates from IMPUTE5 (v.1.2.0) with the XCF file format, the cost is reported to be less than 0.001 £ per sample for batches larger than 500 samples.

It is important to consider that additional factors can influence the running time. In the case of low-coverage WGS, higher coverage levels result in longer running times and increased storage requirements. Similarly, with SNP array data, denser SNP arrays lead to higher memory usage and longer running times. The costs presented in **Table 5** should therefore be regarded as guidelines based on an average case scenario.

| **Number of target samples** | **Low-coverage pipeline. Estimated cost per sample (whole genome - 1.0x coverage)** | **SNP array pipeline. Estimated cost per sample (whole genome - Axiom array)** |
|---|---|---|
| 1 | 0.92 £ | 0.21 £ |
| 10 | 0.152 £ | 0.019 £ |
| 50 | 0.091 £ | 0.0049 £ |
| 100 | 0.082 £ | 0.00265 £ |
| 500 | 0.079 £ | 0.00096 £ |
| 1000 | 0.077 £ | 0.00084 £ |

**Table 5. Price estimates of the pipelines on the UK Biobank RAP (spot instances).**

## 6. Code availability

We processed the data, performed haplotype phasing and validation using the following software:

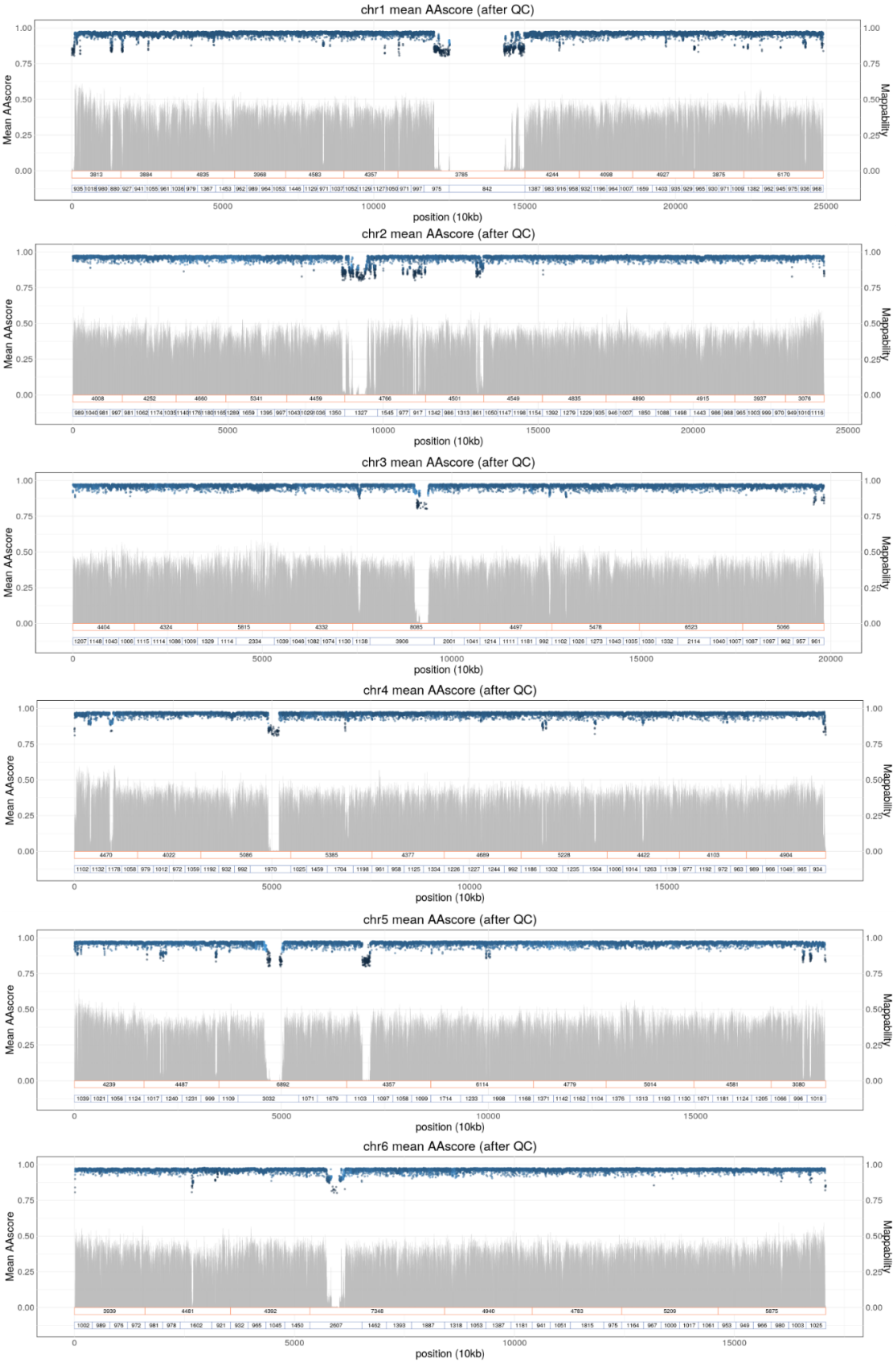
- SHAPEIT v5.1.0 [LINK]
- BCFtools v1.15.1 [LINK]
- GLIMPSE v2.0.0 [LINK]

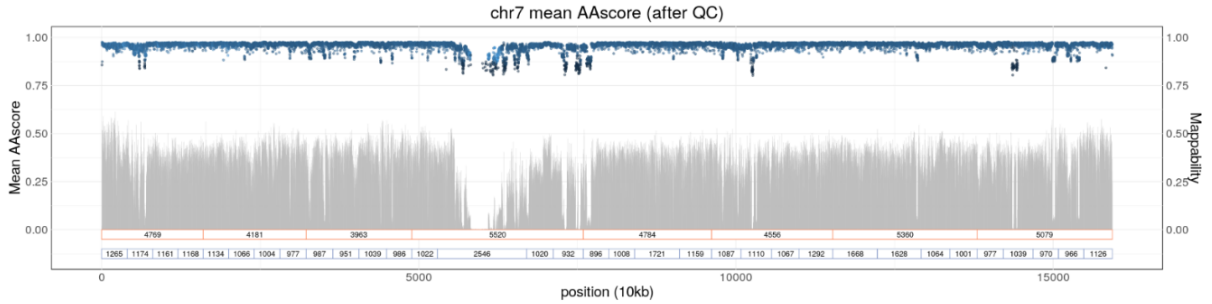
The pipelines to perform genotype imputation are released under MIT licence. However, please note that the programs used may be subject to different licences.
The pipelines are available at https://github.com/srubinacci/imputation-ukb-ref-panel.

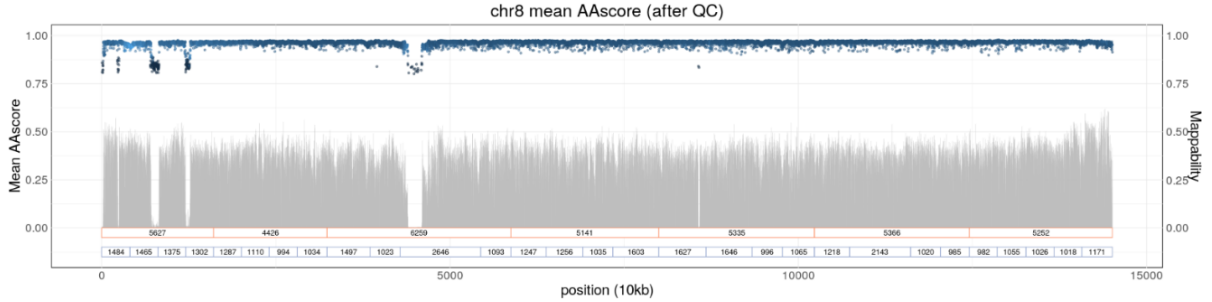
## 7. Funding and Data Access

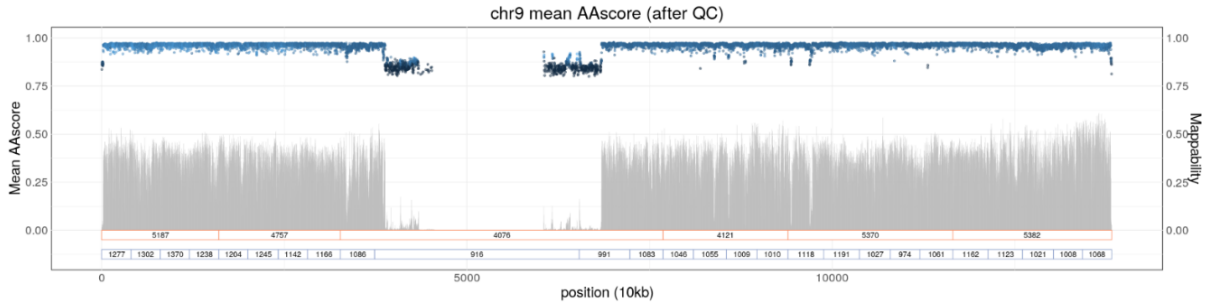
The UK Biobank data was obtained via approved application number 66995. All processing and analyses were carried out in the UKB Research Analysis Platform (RAP). All this work has been funded by the Swiss National Science Foundation project grant PP00P3_176977.

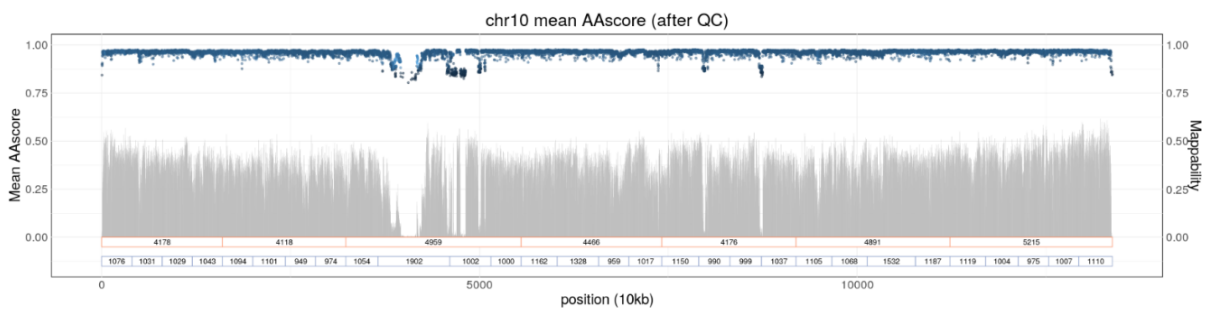
# 8. Supplementary figures

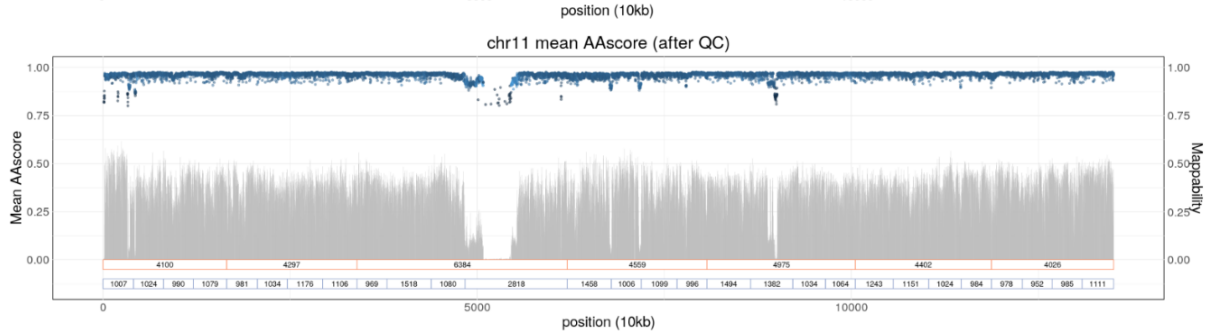


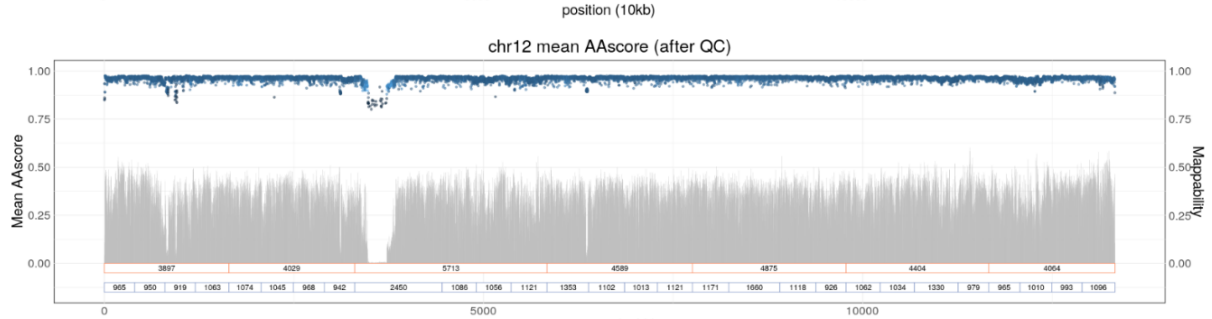

chr7 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

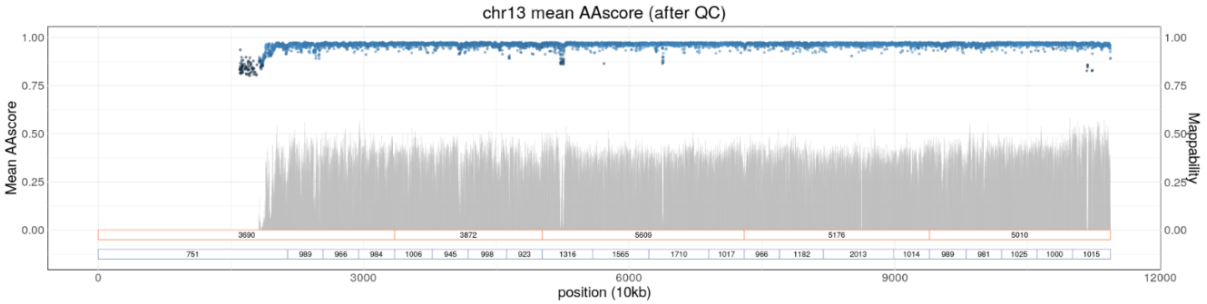

chr8 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

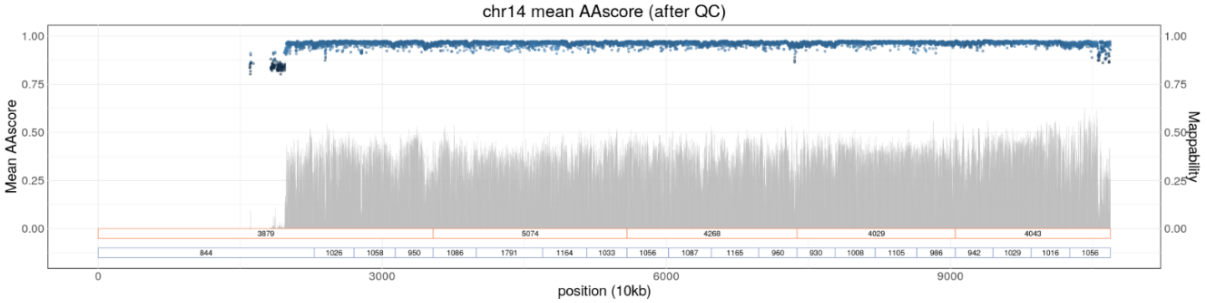

chr9 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

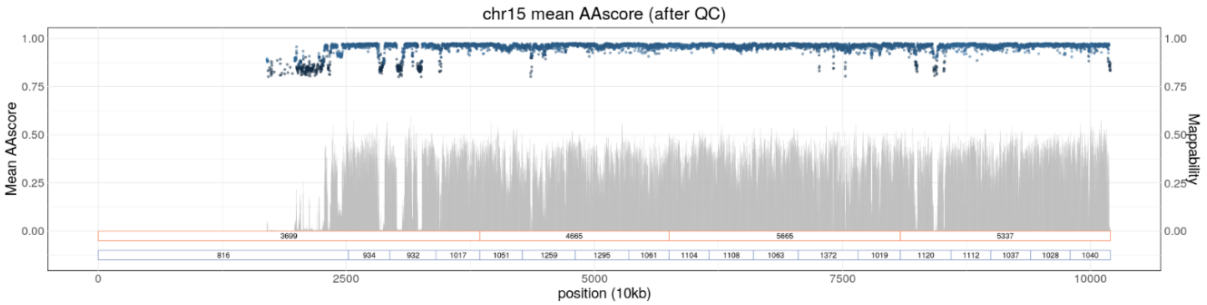

chr10 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

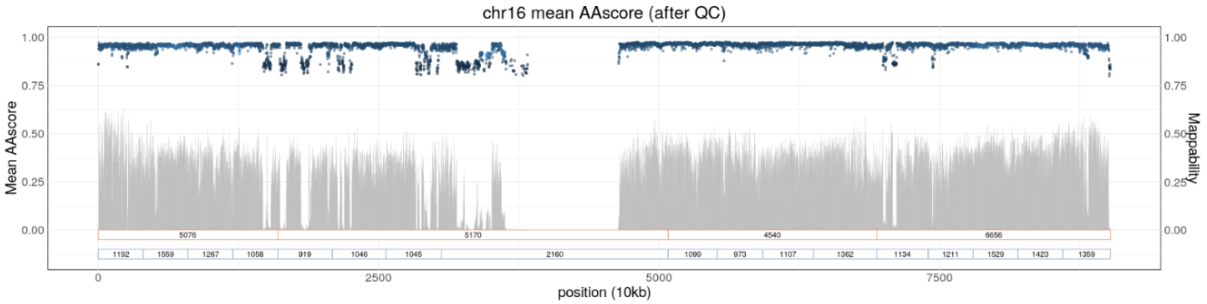

chr11 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

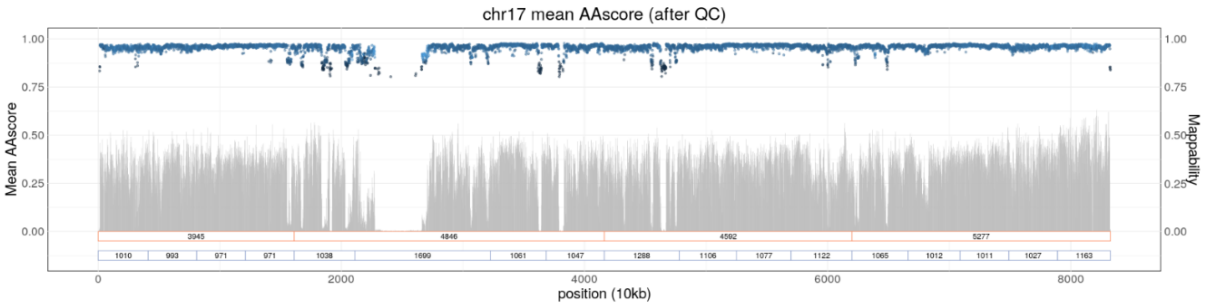

chr12 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

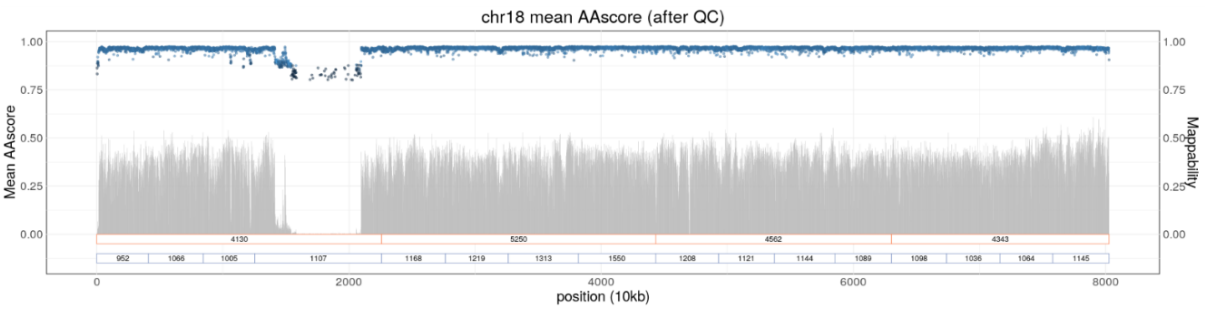

chr13 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

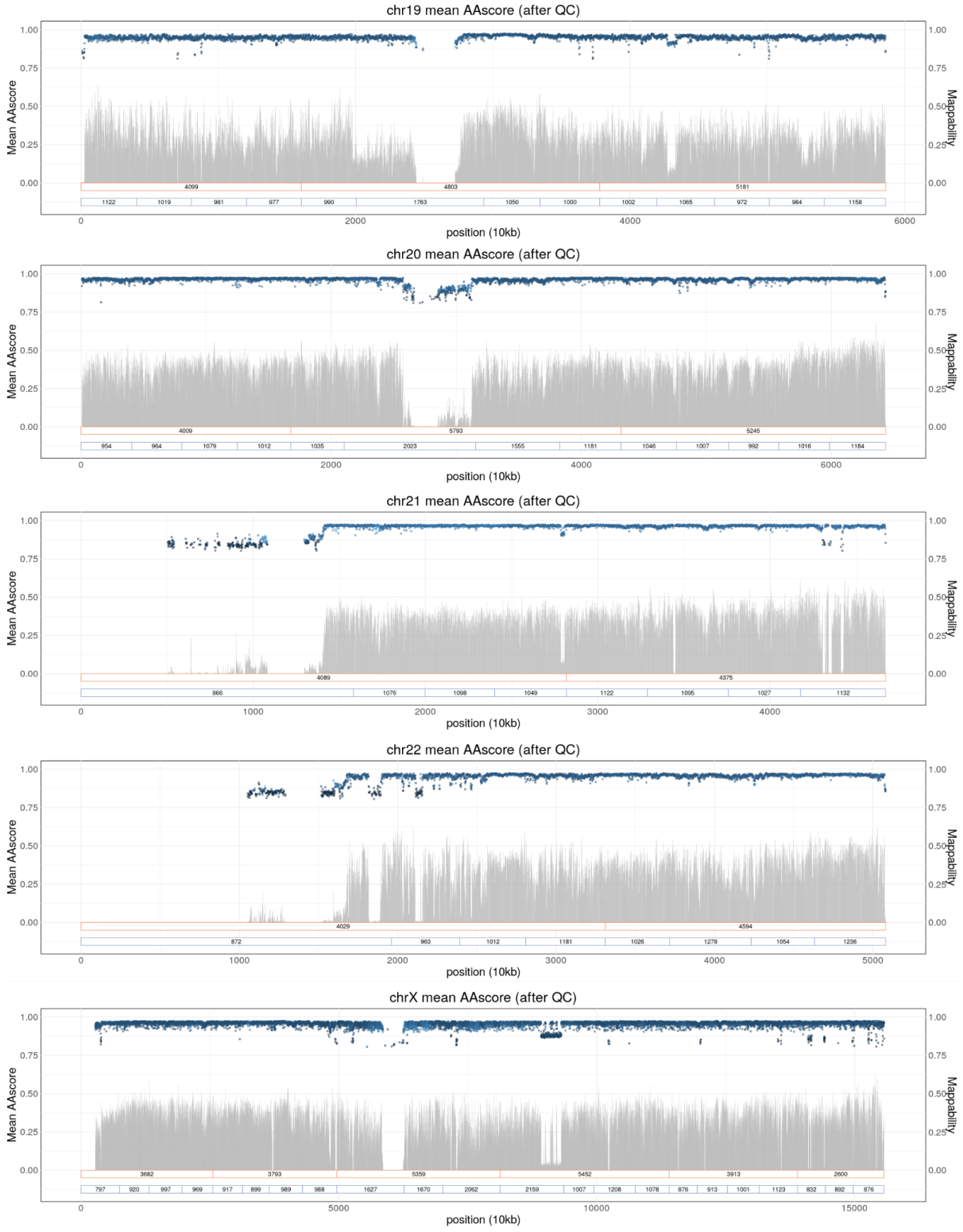

chr14 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)


chr15 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

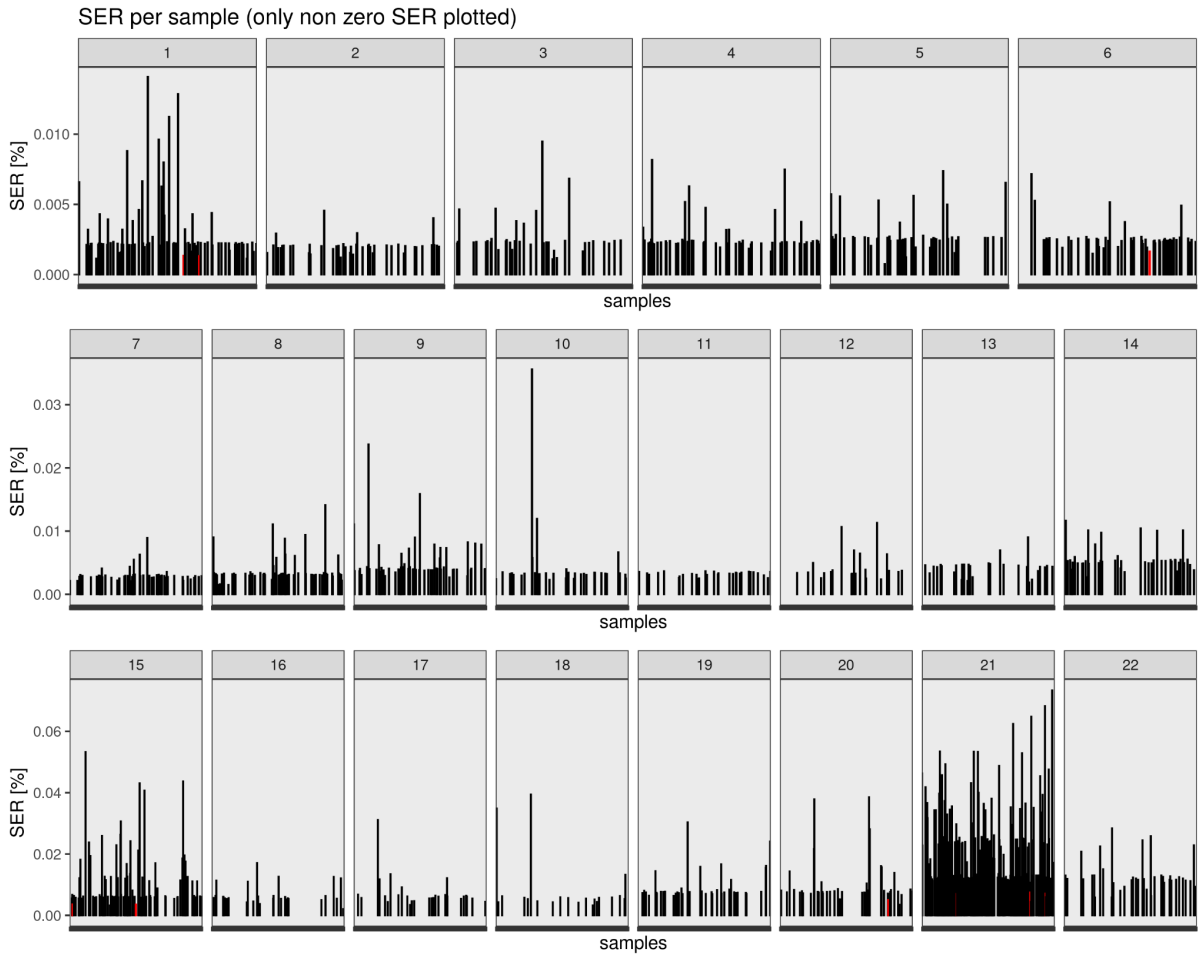

chr16 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)


chr17 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

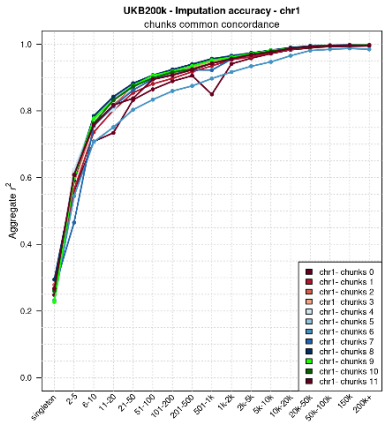

chr18 mean AAscore (after QC)
Mean AAscore
Mappability
position (10kb)

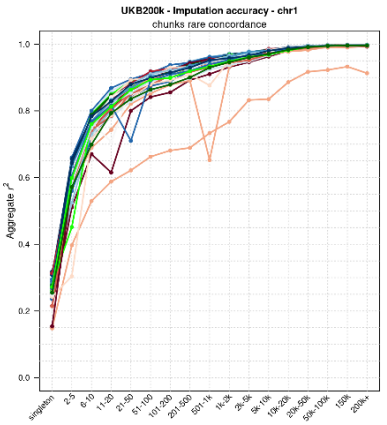


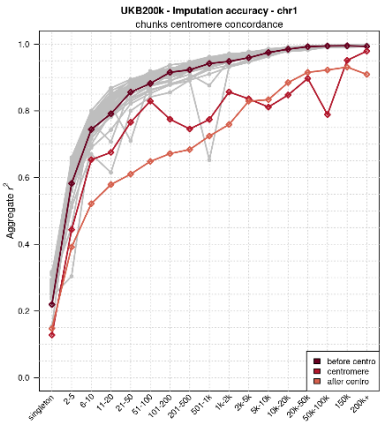

**Supplementary Figure 1. Mappability and mean AAscore per 10kb bins across chromosomes.** A plot is shown for each chromosome. Values after AAscore > 0.8 and HWE filters. Mappability based on Bismap k24 shown in grey. AAscore shown in black (low variant density in 10kb bin) and blue (high variant density). Red rectangles show the chunks for SHAPEIT5 phase_common step, and blue rectangles the chunks for phase_rare, along with the total number of variants in the chunk (in thousands).

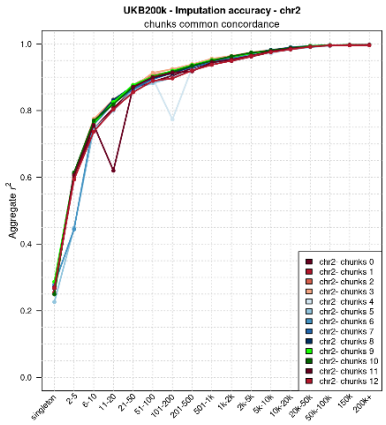


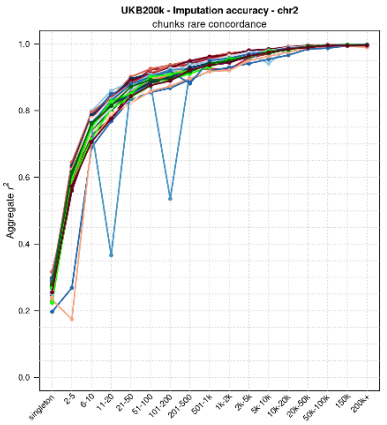

**Supplementary Figure 2. Phasing Switch Error Rates (SER).** Phasing switch error rate (y-axis) stratified by duo and trio offspring (x-axis) across the 22 autosomes. Each bar represents an individual with non-zero SER. Black bars for duos; Red bars for trios.

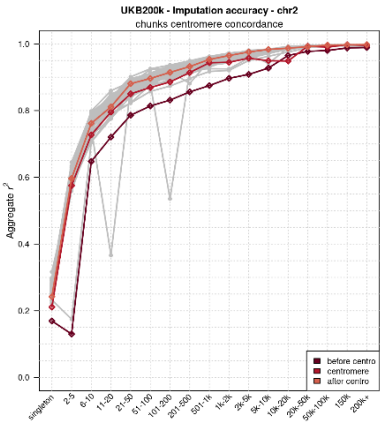

UKB200k - Imputation accuracy - chr1
chunks common concordance
Aggregate $r^2$
chr1- chunks 0
chr1- chunks 1
chr1- chunks 2
chr1- chunks 3
chr1- chunks 4
chr1- chunks 5
chr1- chunks 6
chr1- chunks 7
chr1- chunks 8
chr1- chunks 9
chr1- chunks 10
chr1- chunks 11

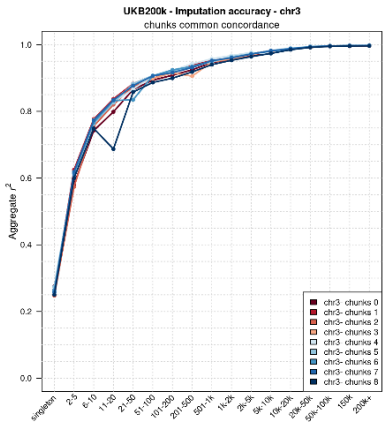

UKB200k - Imputation accuracy - chr1
chunks rare concordance
Aggregate $r^2$

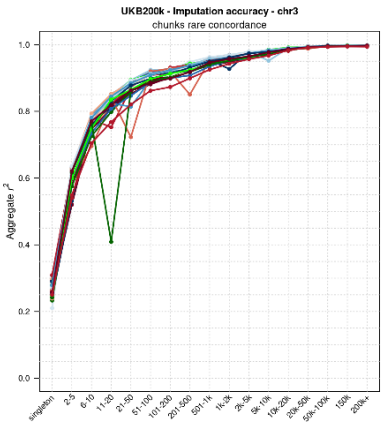

UKB200k - Imputation accuracy - chr1
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

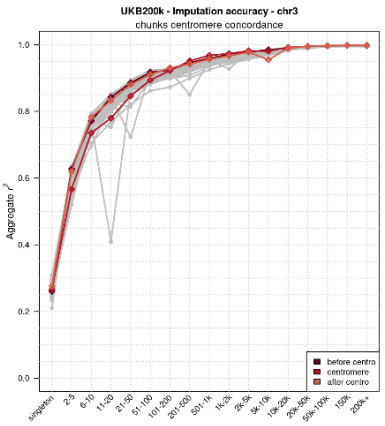

UKB200k - Imputation accuracy - chr2
chunks common concordance
Aggregate $r^2$
chr2- chunks 0
chr2- chunks 1
chr2- chunks 2
chr2- chunks 3
chr2- chunks 4
chr2- chunks 5
chr2- chunks 6
chr2- chunks 7
chr2- chunks 8
chr2- chunks 9
chr2- chunks 10
chr2- chunks 11
chr2- chunks 12

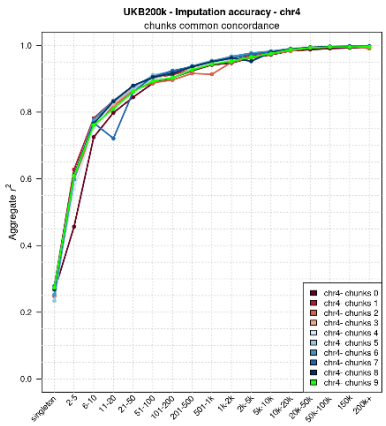

UKB200k - Imputation accuracy - chr2
chunks rare concordance
Aggregate $r^2$

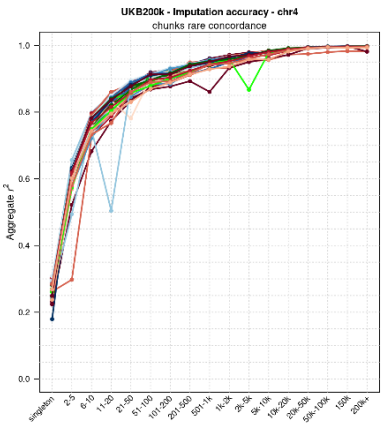

UKB200k - Imputation accuracy - chr2
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

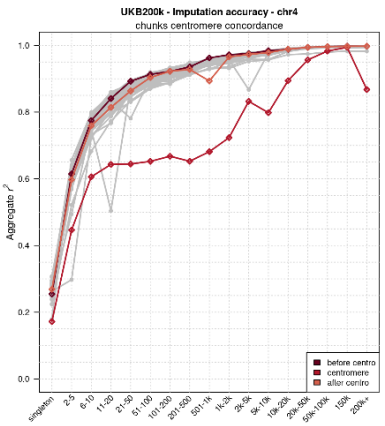

UKB200k - Imputation accuracy - chr3
chunks common concordance
Aggregate $r^2$
chr3- chunks 0
chr3- chunks 1
chr3- chunks 2
chr3- chunks 3
chr3- chunks 4
chr3- chunks 5
chr3- chunks 6
chr3- chunks 7
chr3- chunks 8

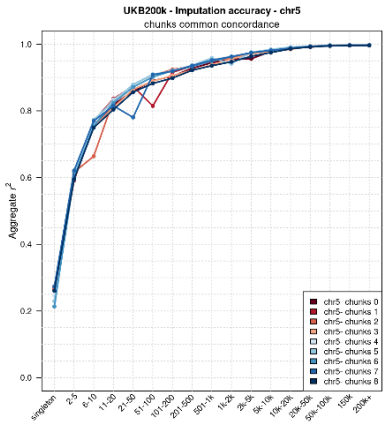

UKB200k - Imputation accuracy - chr3
chunks rare concordance
Aggregate $r^2$

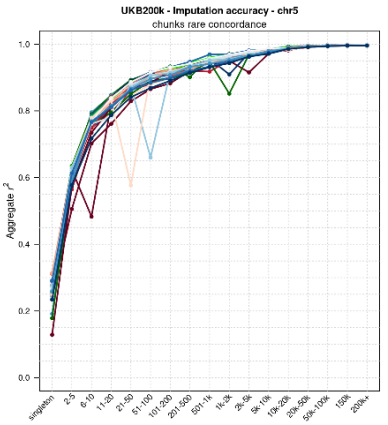

UKB200k - Imputation accuracy - chr3
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

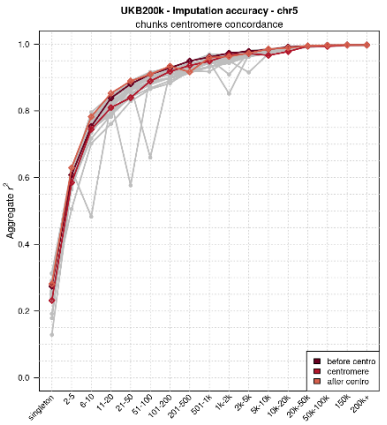

UKB200k - Imputation accuracy - chr4
chunks common concordance
Aggregate $r^2$
chr4- chunks 0
chr4- chunks 1
chr4- chunks 2
chr4- chunks 3
chr4- chunks 4
chr4- chunks 5
chr4- chunks 6
chr4- chunks 7
chr4- chunks 8
chr4- chunks 9

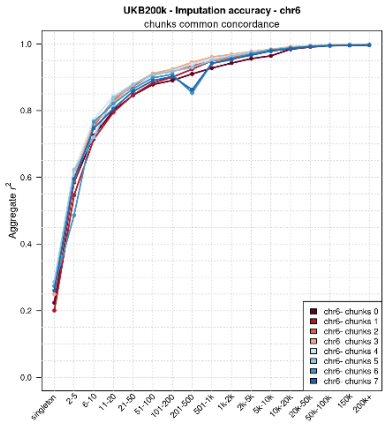

UKB200k - Imputation accuracy - chr4
chunks rare concordance
Aggregate $r^2$

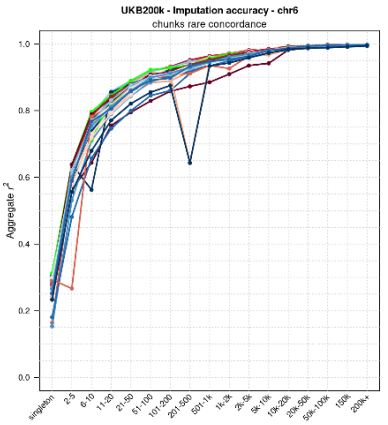

UKB200k - Imputation accuracy - chr4
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

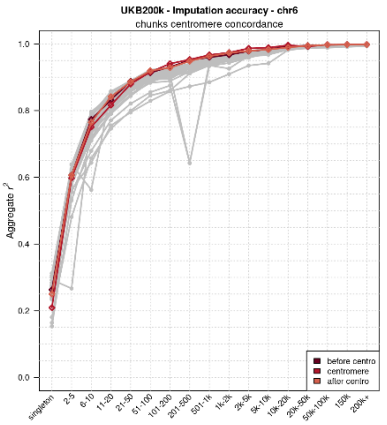

UKB200k - Imputation accuracy - chr5
chunks common concordance
Aggregate $r^2$
chr5- chunks 0
chr5- chunks 1
chr5- chunks 2
chr5- chunks 3
chr5- chunks 4
chr5- chunks 5
chr5- chunks 6
chr5- chunks 7
chr5- chunks 8

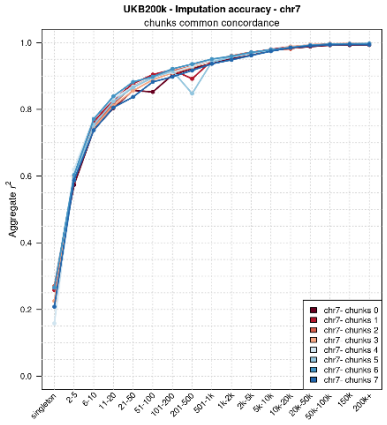

UKB200k - Imputation accuracy - chr5
chunks rare concordance
Aggregate $r^2$

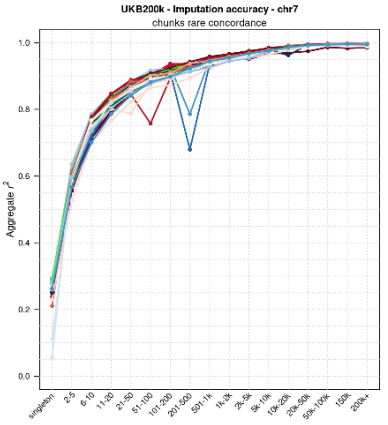

UKB200k - Imputation accuracy - chr5
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

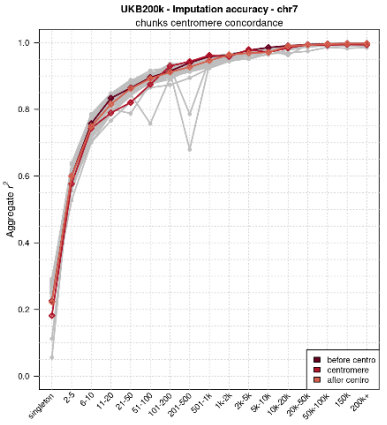

UKB200k - Imputation accuracy - chr6
chunks common concordance
Aggregate $r^2$
chr6- chunks 0
chr6- chunks 1
chr6- chunks 2
chr6- chunks 3
chr6- chunks 4
chr6- chunks 5
chr6- chunks 6
chr6- chunks 7

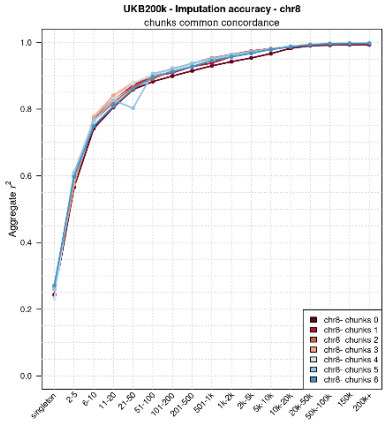

UKB200k - Imputation accuracy - chr6
chunks rare concordance
Aggregate $r^2$

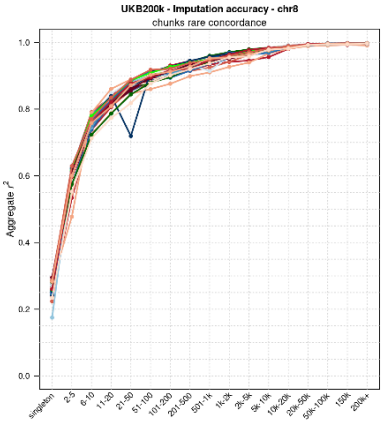

UKB200k - Imputation accuracy - chr6
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

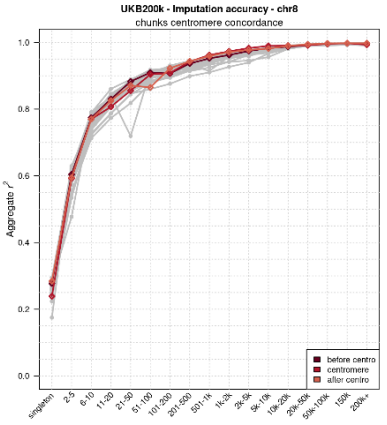

UKB200k - Imputation accuracy - chr7
chunks common concordance
Aggregate $r^2$
chr7- chunks 0
chr7- chunks 1
chr7- chunks 2
chr7- chunks 3
chr7- chunks 4
chr7- chunks 5
chr7- chunks 6
chr7- chunks 7

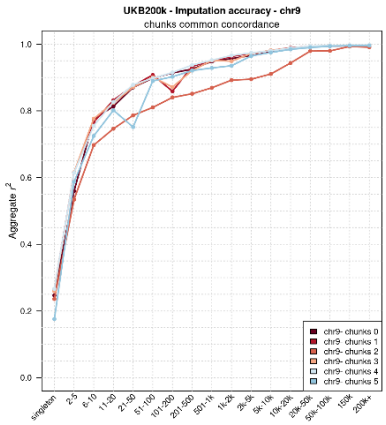

UKB200k - Imputation accuracy - chr7
chunks rare concordance
Aggregate $r^2$

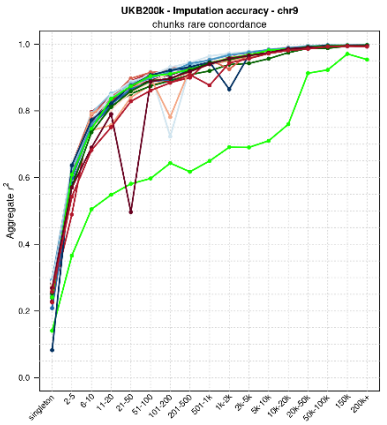

UKB200k - Imputation accuracy - chr7
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

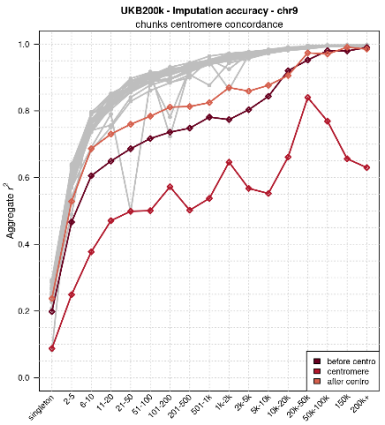

UKB200k - Imputation accuracy - chr8
chunks common concordance
Aggregate $r^2$
chr8- chunks 0
chr8- chunks 1
chr8- chunks 2
chr8- chunks 3
chr8- chunks 4
chr8- chunks 5
chr8- chunks 6

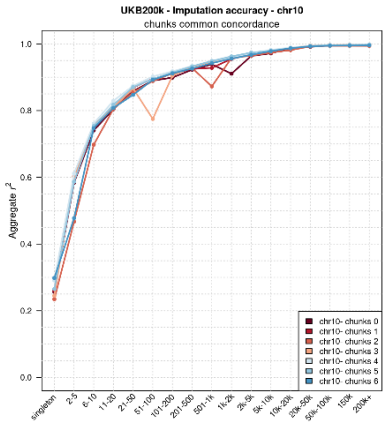

UKB200k - Imputation accuracy - chr8
chunks rare concordance
Aggregate $r^2$

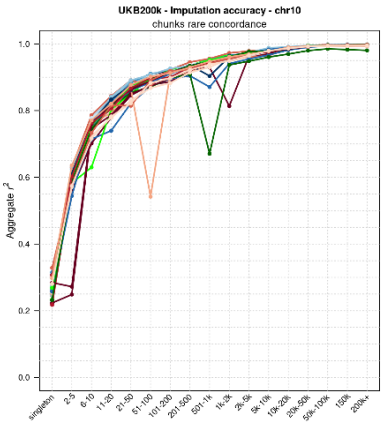

UKB200k - Imputation accuracy - chr8
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

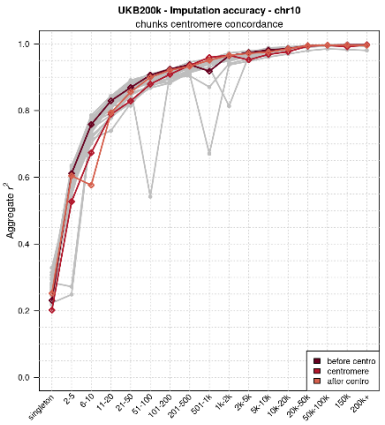

UKB200k - Imputation accuracy - chr9
chunks common concordance
Aggregate $r^2$
chr9- chunks 0
chr9- chunks 1
chr9- chunks 2
chr9- chunks 3
chr9- chunks 4
chr9- chunks 5

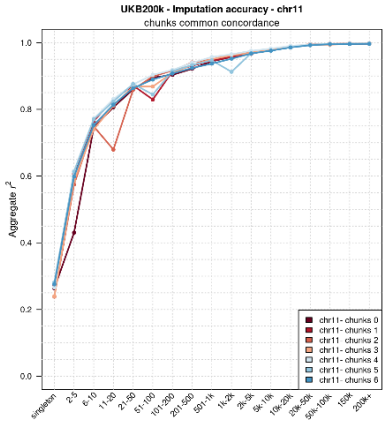

UKB200k - Imputation accuracy - chr9
chunks rare concordance
Aggregate $r^2$

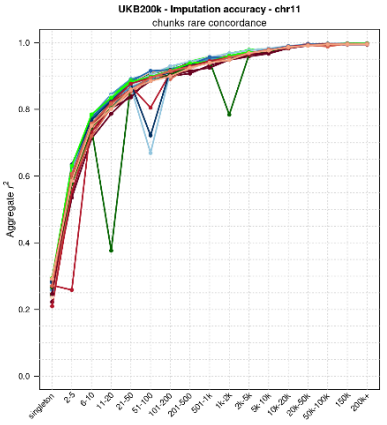

UKB200k - Imputation accuracy - chr9
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

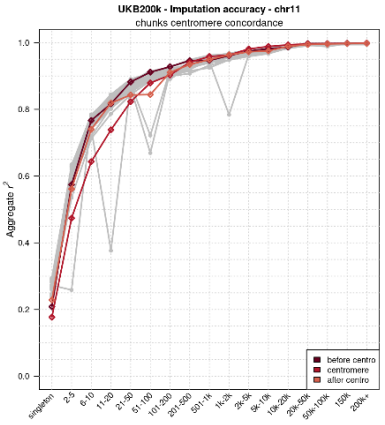

UKB200k - Imputation accuracy - chr10
chunks common concordance
Aggregate $r^2$
chr10- chunks 0
chr10- chunks 1
chr10- chunks 2
chr10- chunks 3
chr10- chunks 4
chr10- chunks 5
chr10- chunks 6

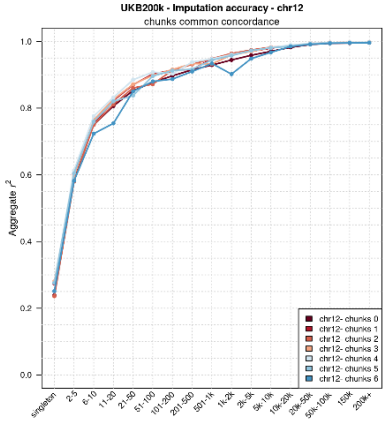

UKB200k - Imputation accuracy - chr10
chunks rare concordance
Aggregate $r^2$

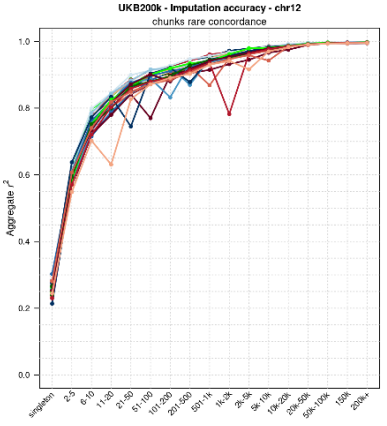

UKB200k - Imputation accuracy - chr10
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

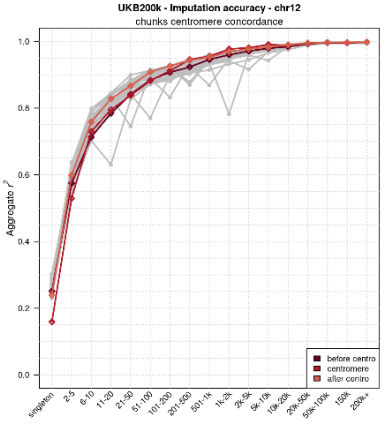

UKB200k - Imputation accuracy - chr11
chunks common concordance
Aggregate $r^2$
chr11- chunks 0
chr11- chunks 1
chr11- chunks 2
chr11- chunks 3
chr11- chunks 4
chr11- chunks 5
chr11- chunks 6

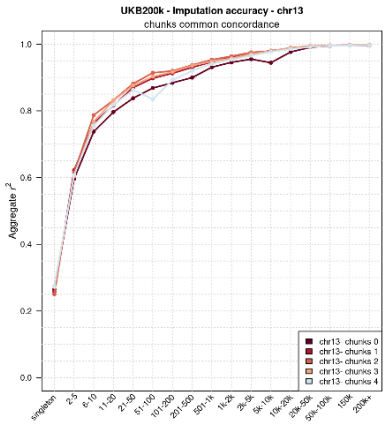

UKB200k - Imputation accuracy - chr11
chunks rare concordance
Aggregate $r^2$

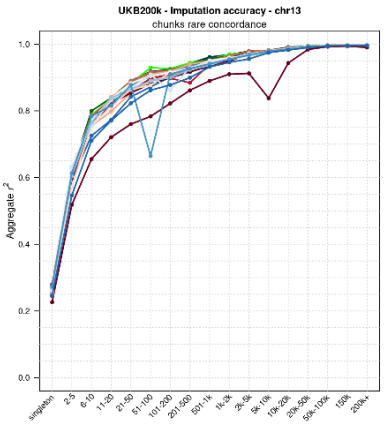

UKB200k - Imputation accuracy - chr11
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

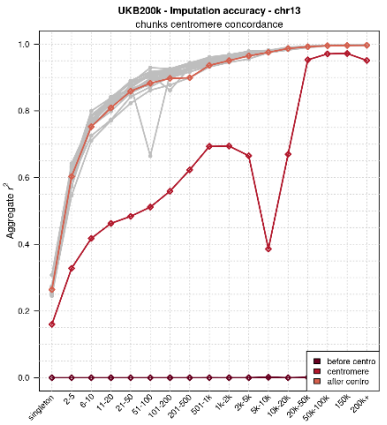

UKB200k - Imputation accuracy - chr12
chunks common concordance
Aggregate $r^2$
chr12- chunks 0
chr12- chunks 1
chr12- chunks 2
chr12- chunks 3
chr12- chunks 4
chr12- chunks 5
chr12- chunks 6

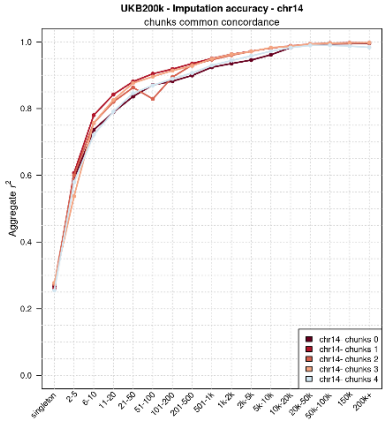

UKB200k - Imputation accuracy - chr12
chunks rare concordance
Aggregate $r^2$

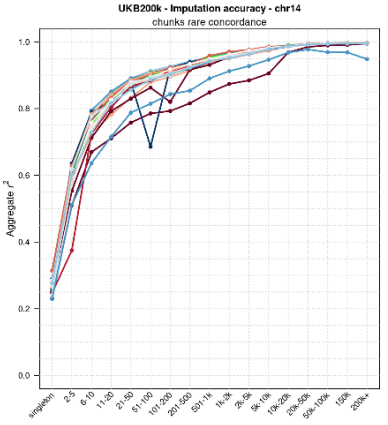

UKB200k - Imputation accuracy - chr12
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

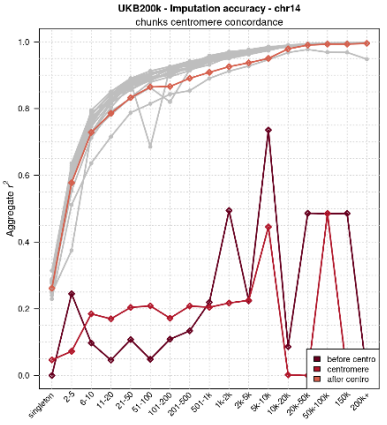

UKB200k - Imputation accuracy - chr13
chunks common concordance
Aggregate $r^2$
chr13- chunks 0
chr13- chunks 1
chr13- chunks 2
chr13- chunks 3
chr13- chunks 4

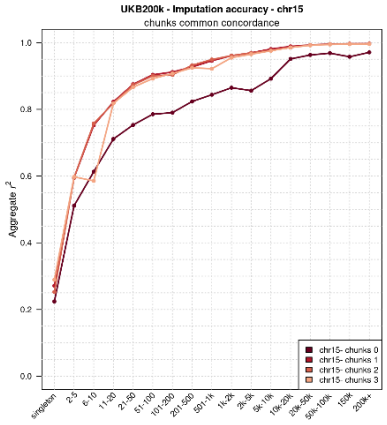

UKB200k - Imputation accuracy - chr13
chunks rare concordance
Aggregate $r^2$

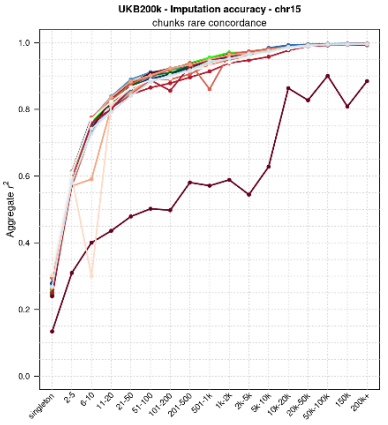

UKB200k - Imputation accuracy - chr13
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

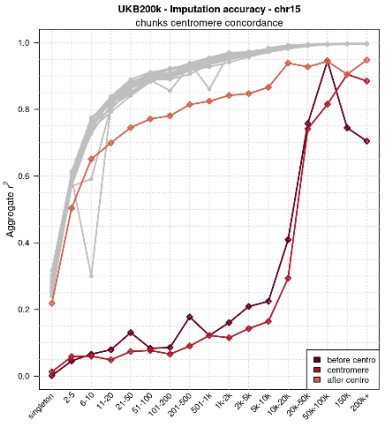

UKB200k - Imputation accuracy - chr14
chunks common concordance
Aggregate $r^2$
chr14- chunks 0
chr14- chunks 1
chr14- chunks 2
chr14- chunks 3
chr14- chunks 4

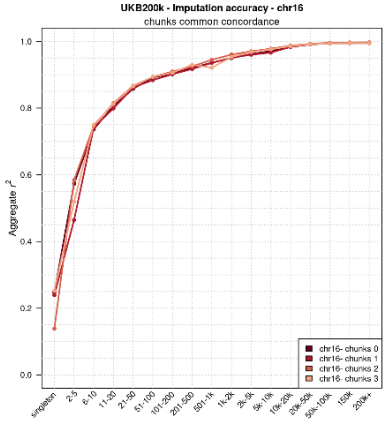

UKB200k - Imputation accuracy - chr14
chunks rare concordance
Aggregate $r^2$

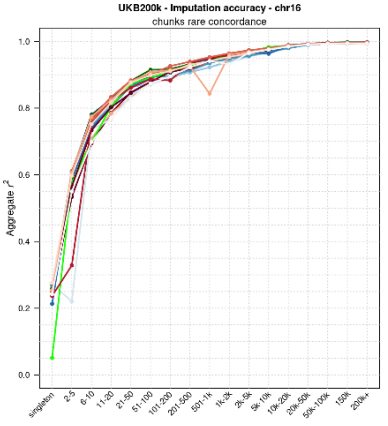

UKB200k - Imputation accuracy - chr14
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

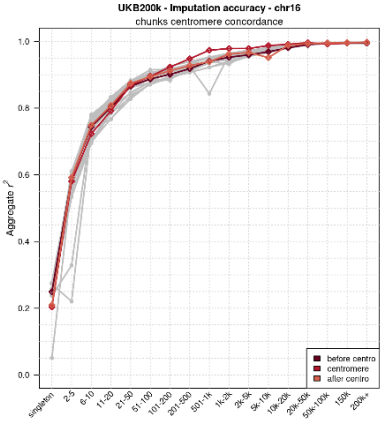

UKB200k - Imputation accuracy - chr15
chunks common concordance
Aggregate $r^2$
chr15- chunks 0
chr15- chunks 1
chr15- chunks 2
chr15- chunks 3

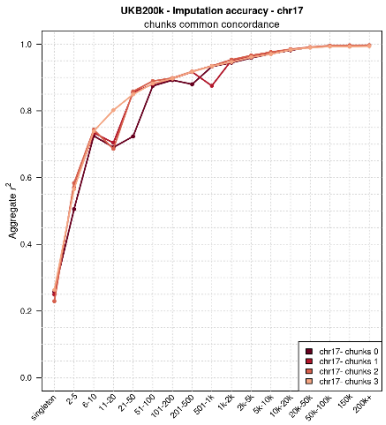

UKB200k - Imputation accuracy - chr15
chunks rare concordance
Aggregate $r^2$

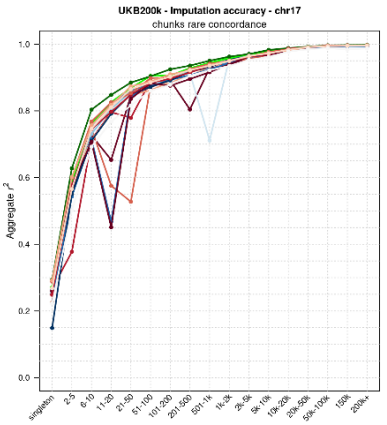

UKB200k - Imputation accuracy - chr15
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

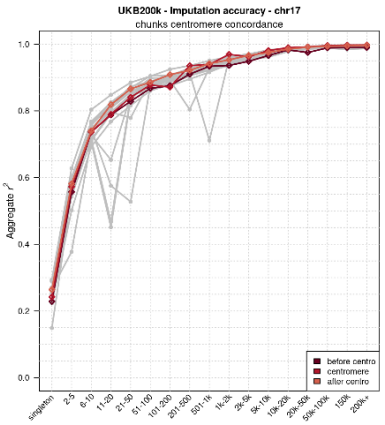

UKB200k - Imputation accuracy - chr16
chunks common concordance
Aggregate $r^2$
chr16- chunks 0
chr16- chunks 1
chr16- chunks 2
chr16- chunks 3

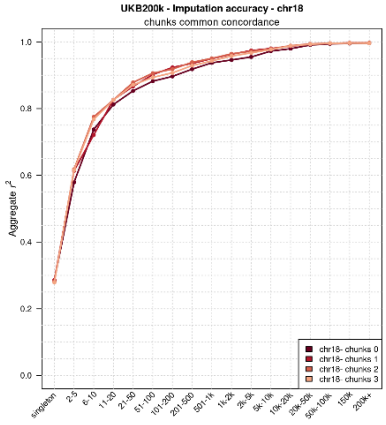

UKB200k - Imputation accuracy - chr16
chunks rare concordance
Aggregate $r^2$

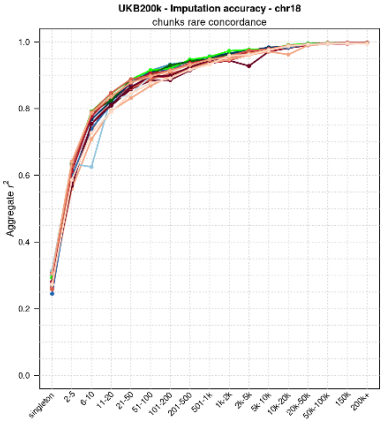

UKB200k - Imputation accuracy - chr16
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro

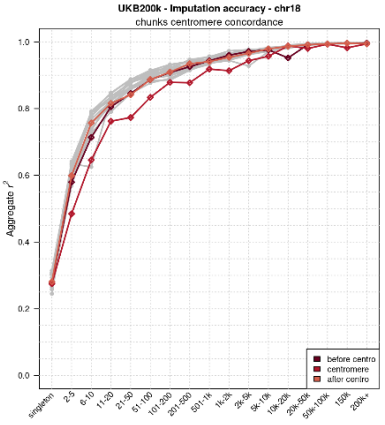

UKB200k - Imputation accuracy - chr17
chunks common concordance
Aggregate $r^2$
chr17- chunks 0
chr17- chunks 1
chr17- chunks 2
chr17- chunks 3
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

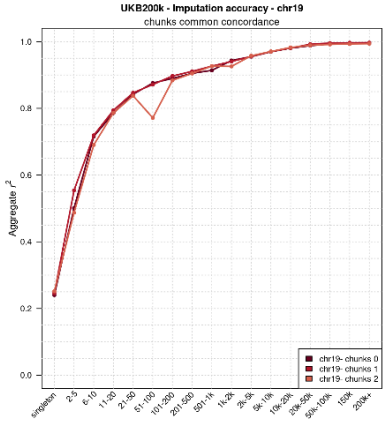

UKB200k - Imputation accuracy - chr17
chunks rare concordance
Aggregate $r^2$
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

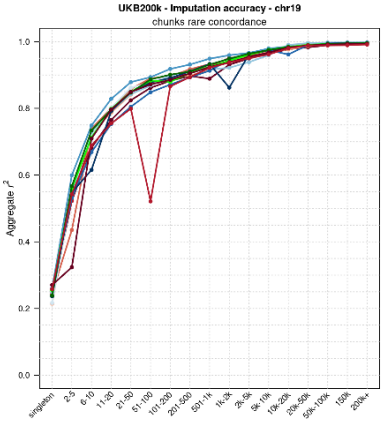

UKB200k - Imputation accuracy - chr17
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

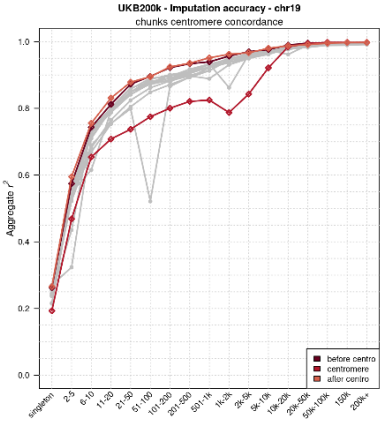

UKB200k - Imputation accuracy - chr18
chunks common concordance
Aggregate $r^2$
chr18- chunks 0
chr18- chunks 1
chr18- chunks 2
chr18- chunks 3
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

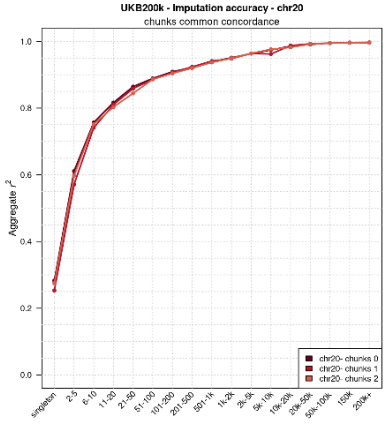

UKB200k - Imputation accuracy - chr18
chunks rare concordance
Aggregate $r^2$
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

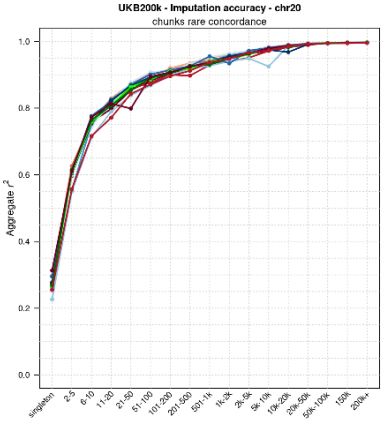

UKB200k - Imputation accuracy - chr18
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

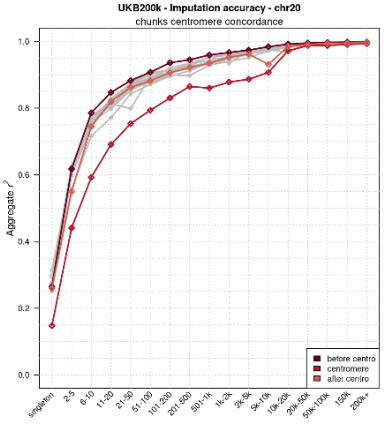

UKB200k - Imputation accuracy - chr19
chunks common concordance
Aggregate $r^2$
chr19- chunks 0
chr19- chunks 1
chr19- chunks 2
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

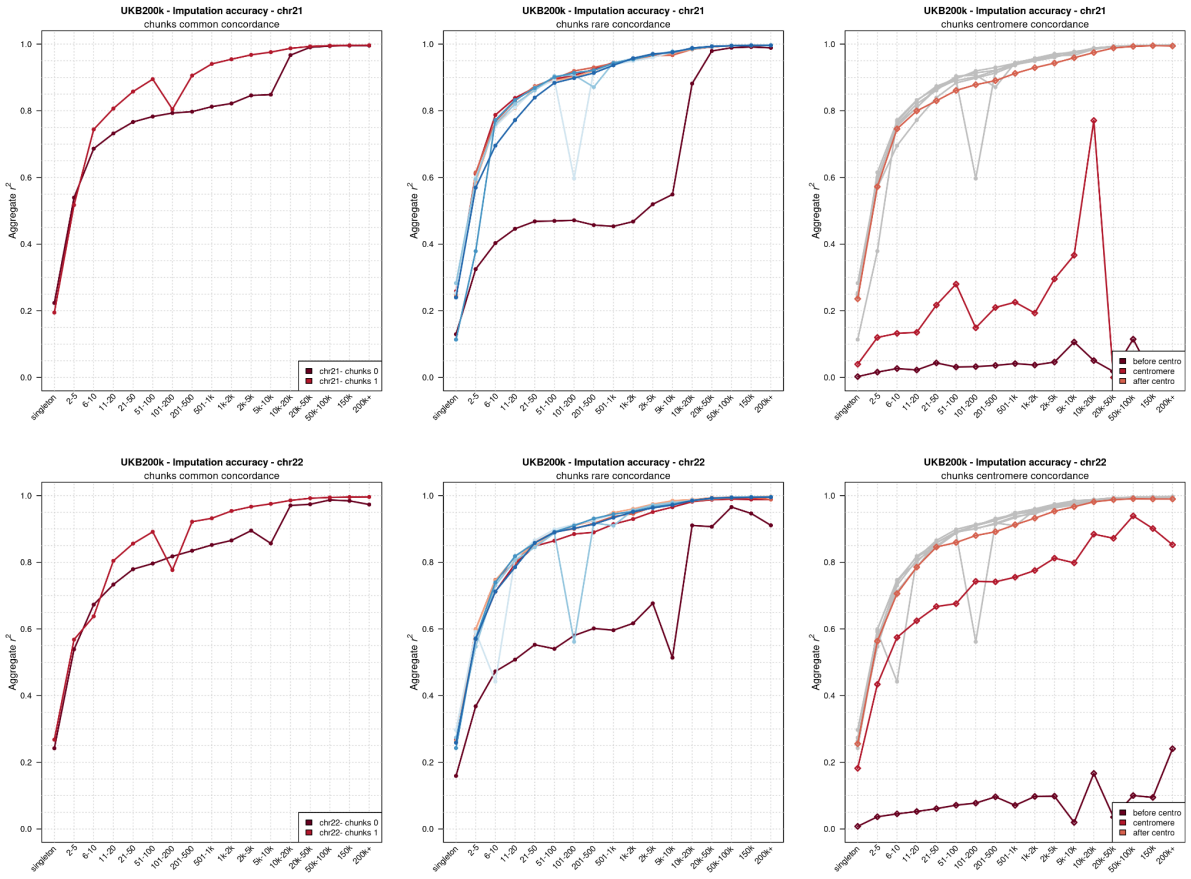

UKB200k - Imputation accuracy - chr19
chunks rare concordance
Aggregate $r^2$
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+


UKB200k - Imputation accuracy - chr19
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+

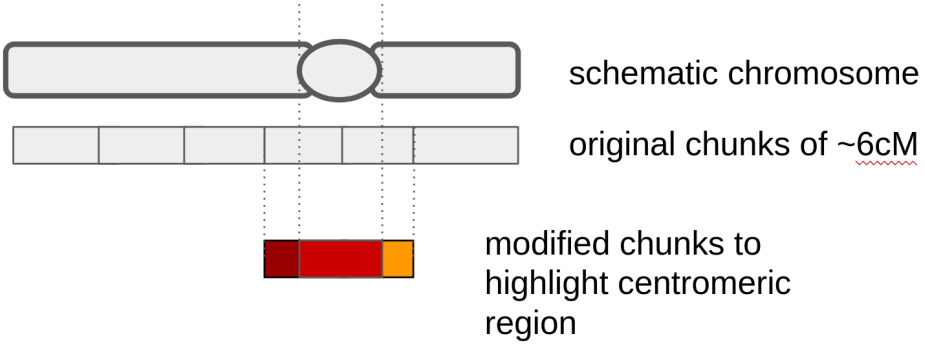

UKB200k - Imputation accuracy - chr20
chunks common concordance
Aggregate $r^2$
chr20- chunks 0
chr20- chunks 1
chr20- chunks 2
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+


UKB200k - Imputation accuracy - chr20
chunks rare concordance
Aggregate $r^2$
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+


UKB200k - Imputation accuracy - chr20
chunks centromere concordance
Aggregate $r^2$
before centro
centromere
after centro
singleton
2-5
6-10
11-20
21-50
51-100
101-200
201-500
501-1k
1k-2k
2k-5k
5k-10k
10k-20k
20k-50k
50k-100k
150k
200k+




**Supplementary Figure 3. Imputation accuracy by chromosome chunks.** Imputation accuracy ($r^2$, y-axis) stratified by Minor Allele Count (x-axis) across the 22 autosomes. Each row represents an autosome. Left panels: accuracy aggregated by chunks of ~25cM ("phase common" chunks). Middle panel: accuracy aggregated by chunks of ~6cM ("phase rare" chunks). Right panels: accuracy aggregated by chunks of ~6cM with modified chunking highlighting centromere regions (see Supplementary Figure 4).




**Supplementary Figure 4. Schematic representation of the modified chunking in centromere regions.** For each chromosome, the two chunks overlapping the centromere are modified into three chunks so that one chunk includes the region before the centromere, one chunk overlaps the

centromere, and the third chunk includes the region after the centromere. This allows us to understand the drop in accuracy around centromeric regions.

## 9. Supplementary Tables

| Low-coverage WGS imputation pipeline | | | | |
|---|---|---|---|---|
| **Name** | **Description** | **Type** | **Optional** | **Default** |
| app_pth | apps path | string | true | apps/ |
| batch_id | string used to identify batch of data (set of target samples) | string | true | batch_00000 |
| chr | chromosome name | string | false | |
| cnk_pth | imputation chunk file path | string | true | data/glimpse2/chunks/ |
| conversion_instance_type | instance type used for reference panel binary format conversions | string | true | mem2_ssd1_v2_x4 |
| imp_arg | glimpse2 arguments | string | true | |
| imputation_instance_type | instance type used for glimpse2 imputation | string | true | mem2_ssd1_v2_x4 |
| map_pth | genetic map path | string | true | data/glimpse2/maps/ |
| mount_inputs | whether to mount all files that were supplied as inputs to the app instead of downloading them to the local storage of the execution worker | boolean | true | true |
| out_pth | output directory path | string | true | /data/glimpse2/out/ |
| project | DNAnexus project name | string | false | |
| ref_bcf_pth | reference panel path in bcf file format | string | false | |
| ref_pfx | reference panel prefix name | string | true | |
| ref_sfx | reference panel suffix name | string | true | |
| ref_bin_pth | reference panel path in binary file format | string | true | data/glimpse2/ref_bin_phased/ |
| run_convert_reference_module | create reference panel in binary format | boolean | true | false |

| | | | | |
|---|---|---|---|---|
| run_impute_module | perform genotype imputation | boolean | true | true |
| tar_cram_pth | target panel bam/cram path | string | true | data/glimpse2/target_data/crams |

**Supplementary Table 1. Parameters of the low-coverage imputation pipeline**.

| SNP array imputation pipeline | | | | |
|---|---|---|---|---|
| **Name** | **Description** | **Type** | **Optional** | **Default** |
| app_pth | apps path | string | true | apps/ |
| batch_id | string used to identify batch of data (set of target samples) | string | true | batch_00000 |
| chr | chromosome name | string | false | |
| cnk_pth | imputation chunk file path | string | true | data/impute5/chunks/ |
| conversion_instance_type | instance type used for xcf conversions | string | true | mem1_ssd1_v2_x4 |
| imp_arg | impute5 arguments | string | true | |
| imputation_instance_type | instance type used for impute5 imputation | string | true | mem2_ssd1_v2_x4 |
| map_pth | genetic map path | string | true | data/impute5/maps/ |
| mount_inputs | whether to mount all files that were supplied as inputs to the app instead of downloading them to the local storage of the execution worker | boolean | true | false |
| out_pth | output directory path | string | true | /data/impute5/out/ |
| phasing_instance_type | instance type used for pre-phasing | string | true | mem2_ssd1_v2_x4 |
| phs_arg | shapeit5 arguments | string | true | |
| project | DNAnexus project name | string | false | |
| ref_bcf_pth | reference panel path in bcf file format | string | true | |
| ref_pfx | reference panel prefix name | string | true | |
| ref_sfx | reference panel suffix name | string | true | |

| | | | | |
|---|---|---|---|---|
| ref_xcf_pth | reference panel path in xcf file format | string | true | data/impute5/ref_xcf_phased/ |
| run_convert_reference_module | create reference panel in xcf format | boolean | true | false |
| run_convert_target_module | perform conversion of the already phased bcf files in xcf format, without running prephasing | boolean | true | false |
| run_impute_module | perform genotype imputation | boolean | true | true |
| run_phase_module | perform pre-phasing on the SNP array data using shapeit5_phase_common | boolean | true | false |
| tar_bcf_pth | target panel path (unphased data) | string | true | data/impute5/target_data |
| tar_pfx | target panel prefix name | string | true | |
| tar_sfx | target panel suffix name | string | true | |
| tar_xcf_pth | target panel path (pre-phased data) | string | true | data/impute5/tar_xcf_phased/ |

**Supplementary Table 2. Parameters of the SNP array imputation pipeline**.

## 10. References


Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., … gnomAD Project Consortium. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. In bioRxiv. https://doi.org/10.1101/2022.03.20.485034

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. Nature Genetics, 49(11), 1654–1660.

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., Kristmundsdottir, S., Sigurpalsdottir, B. D., Stefansson, O. A., Beyter, D., Holley, G., Tragante, V., Gylfason, A., Olason, P. I., Zink, F.,

… Stefansson, K. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature, 607(7920), 732–740.

Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., & Delaneau, O. (2022). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. In bioRxiv. https://doi.org/10.1101/2022.10.19.512867

Hofmeister, R. J., Rubinacci, S., Ribeiro, D. M., Buil, A., Kutalik, Z., & Delaneau, O. (2022). Parent-of-Origin inference for biobanks. Nature Communications, 13(1), 6668.

Karimzadeh, M., Ernst, C., Kundaje, A., & Hoffman, M. M. (2018). Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Research, 46(20), e120.

Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. PLoS Genetics, 16(11), e1009049.

Rubinacci, S., Hofmeister, R., da Mota, B. S., & Delaneau, O. (2022). Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. In bioRxiv. https://doi.org/10.1101/2022.11.28.518213

Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., & Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nature Genetics, 53(1), 120–126.