Tutorial

Test Report

Automatically generated by GWIP

April 01, 2015

Contents

1 Background

The aim of this project is to perform genome-wide imputation using the study cohort.

2 Methods

The following (cleaned) files provided information about the study cohort dataset for 90 samples and 2,278,357 markers (including 0 markers located on sexual or mitochondrial chromosomes):

- `data/hapmap_CEU_r23a_hg19.bed`
- `data/hapmap_CEU_r23a_hg19.bim`
- `data/hapmap_CEU_r23a_hg19.fam`

IMPUTE2's pre-phasing approach can work with phased haplotypes from SHAPEIT, a highly accurate phasing algorithm that can handle mixtures of unrelated samples, duos or trios. The usage of SHAPEIT is highly recommended to infer haplotypes underlying the study genotypes. The phased haplotypes are then passed to IMPUTE2 for imputation. Although pre-phasing allows for very fast imputation, it leads to a small loss in accuracy since the estimation uncertainty in the study haplotypes is ignored. SHAPEIT version v2.r790 [1] and IMPUTE2 version 2.3.2 [2, 3, 4] were used for this analysis. Binary pedfiles were processed using Plink version v1.07 [5].

To speed up the pre-phasing and imputation steps, the dataset was split by chromosome. The following quality steps were then performed on each chromosome:

1. Ambiguous markers with alleles `A/T` and `C/G`, duplicated markers (same position), and markers located on special chromosomes (sexual or mitochondrial chromosomes) were excluded from the imputation. An initial strand check was also performed using the human reference genome. **In total, 349,533 ambiguous, 0 duplicated and 0 special markers were excluded. Also, 338 markers were flipped because of strand issue.**
2. Markers' strand was checked using the SHAPEIT algorithm and IMPUTE2's reference files. **In total, 743 markers had an incorrect strand and were flipped using Plink.**
3. The strand of each marker was checked again using SHAPEIT against IMPUTE2's reference files. **In total, 743 markers were found to still be on the wrong strand, and were hence excluded from the final dataset using Plink.**

In total, 1,928,081 were used for phasing using SHAPEIT. IMPUTE2 was then used to impute markers genome-wide using its reference file (filtering out sites where `ALL<0.01` or `ALL>0.99`).

3 Results

3.1 Cross-validation

According to IMPUTE2's documentation, the cross-validation tables are "based on an internal cross-validation that is performed during each IMPUTE2 run. For this analysis, the program masks the genotypes of one variant at a time in the study data and imputes the masked genotypes by using the remaining study and reference data. The imputed genotypes are then compared with the original genotypes to produce the concordance statistics."

Tables I to XXII show the cross-validation results for the autosomes (chromosomes 1 to 22). Table XXIII shows the cross-validation results across the autosomes.

Table I: IMPUTE2's internal cross-validation for chromosome 1. Tables show the percentage of concordance between genotyped calls and imputed calls for 13,280,400 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	3,239	34.6
[0.5 – 0.6]	21,599	49.9
[0.6 – 0.7]	21,348	57.7
[0.7 – 0.8]	25,640	65.5
[0.8 – 0.9]	39,754	74.3
[0.9 – 1.0]	13,168,820	98.2

Interval	Called (%)	Concordance (%)
[$\geq$ 0.0]	100.0	97.9
[$\geq$ 0.1]	100.0	97.9
[$\geq$ 0.2]	100.0	97.9
[$\geq$ 0.3]	100.0	97.9
[$\geq$ 0.4]	100.0	97.9
[$\geq$ 0.5]	100.0	97.9
[$\geq$ 0.6]	99.8	98.0
[$\geq$ 0.7]	99.6	98.0
[$\geq$ 0.8]	99.5	98.1
[$\geq$ 0.9]	99.2	98.2

Table II: IMPUTE2's internal cross-validation for chromosome 2. Tables show the percentage of concordance between genotyped calls and imputed calls for 15,643,890 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,646	37.4
[0.5 – 0.6]	14,921	51.6
[0.6 – 0.7]	15,196	59.6
[0.7 – 0.8]	18,565	68.8
[0.8 – 0.9]	29,513	77.6
[0.9 – 1.0]	15,564,049	98.9

Interval	Called (%)	Concordance (%)
[$\geq$ 0.0]	100.0	98.7
[$\geq$ 0.1]	100.0	98.7
[$\geq$ 0.2]	100.0	98.7
[$\geq$ 0.3]	100.0	98.7
[$\geq$ 0.4]	100.0	98.7
[$\geq$ 0.5]	100.0	98.7
[$\geq$ 0.6]	99.9	98.8
[$\geq$ 0.7]	99.8	98.8
[$\geq$ 0.8]	99.7	98.9
[$\geq$ 0.9]	99.5	98.9

Table III: IMPUTE2's internal cross-validation for chromosome 3. Tables show the percentage of concordance between genotyped calls and imputed calls for 11,673,990 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,969	36.0
[0.5 – 0.6]	14,403	52.2
[0.6 – 0.7]	14,511	60.2
[0.7 – 0.8]	17,419	68.8
[0.8 – 0.9]	27,956	77.7
[0.9 – 1.0]	11,597,732	98.6

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.4
[≥ 0.1]	100.0	98.4
[≥ 0.2]	100.0	98.4
[≥ 0.3]	100.0	98.4
[≥ 0.4]	100.0	98.4
[≥ 0.5]	100.0	98.4
[≥ 0.6]	99.9	98.5
[≥ 0.7]	99.7	98.5
[≥ 0.8]	99.6	98.6
[≥ 0.9]	99.3	98.6

Table IV: IMPUTE2's internal cross-validation for chromosome 4. Tables show the percentage of concordance between genotyped calls and imputed calls for 10,945,350 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,784	36.1
[0.5 – 0.6]	13,995	51.8
[0.6 – 0.7]	14,370	58.7
[0.7 – 0.8]	17,284	66.8
[0.8 – 0.9]	26,793	76.3
[0.9 – 1.0]	10,871,124	98.4

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.2
[≥ 0.1]	100.0	98.2
[≥ 0.2]	100.0	98.2
[≥ 0.3]	100.0	98.2
[≥ 0.4]	100.0	98.2
[≥ 0.5]	100.0	98.2
[≥ 0.6]	99.8	98.3
[≥ 0.7]	99.7	98.3
[≥ 0.8]	99.6	98.4
[≥ 0.9]	99.3	98.4

Table V: IMPUTE2's internal cross-validation for chromosome 5. Tables show the percentage of concordance between genotyped calls and imputed calls for 10,952,820 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,420	36.9
[0.5 – 0.6]	11,497	51.7
[0.6 – 0.7]	11,356	60.0
[0.7 – 0.8]	13,978	68.3
[0.8 – 0.9]	21,975	76.5
[0.9 – 1.0]	10,892,594	98.8

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.6
[≥ 0.1]	100.0	98.6
[≥ 0.2]	100.0	98.6
[≥ 0.3]	100.0	98.6
[≥ 0.4]	100.0	98.6
[≥ 0.5]	100.0	98.6
[≥ 0.6]	99.9	98.7
[≥ 0.7]	99.8	98.7
[≥ 0.8]	99.7	98.7
[≥ 0.9]	99.5	98.8

Table VI: IMPUTE2's internal cross-validation for chromosome 6. Tables show the percentage of concordance between genotyped calls and imputed calls for 11,962,800 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 − 0.1]	0	0.0
[0.1 − 0.2]	0	0.0
[0.2 − 0.3]	0	0.0
[0.3 − 0.4]	0	0.0
[0.4 − 0.5]	1,284	36.1
[0.5 − 0.6]	11,223	50.8
[0.6 − 0.7]	10,988	60.6
[0.7 − 0.8]	13,497	67.7
[0.8 − 0.9]	21,092	76.6
[0.9 − 1.0]	11,904,716	98.9

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.7
[≥ 0.1]	100.0	98.7
[≥ 0.2]	100.0	98.7
[≥ 0.3]	100.0	98.7
[≥ 0.4]	100.0	98.7
[≥ 0.5]	100.0	98.7
[≥ 0.6]	99.9	98.7
[≥ 0.7]	99.8	98.8
[≥ 0.8]	99.7	98.8
[≥ 0.9]	99.5	98.9

Table VII: IMPUTE2's internal cross-validation for chromosome 7. Tables show the percentage of concordance between genotyped calls and imputed calls for 9,180,270 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 − 0.1]	0	0.0
[0.1 − 0.2]	0	0.0
[0.2 − 0.3]	0	0.0
[0.3 − 0.4]	0	0.0
[0.4 − 0.5]	1,518	33.8
[0.5 − 0.6]	11,889	52.5
[0.6 − 0.7]	11,684	60.1
[0.7 − 0.8]	14,097	68.0
[0.8 − 0.9]	21,851	77.2
[0.9 − 1.0]	9,119,231	98.8

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.6
[≥ 0.1]	100.0	98.6
[≥ 0.2]	100.0	98.6
[≥ 0.3]	100.0	98.6
[≥ 0.4]	100.0	98.6
[≥ 0.5]	100.0	98.6
[≥ 0.6]	99.9	98.6
[≥ 0.7]	99.7	98.7
[≥ 0.8]	99.6	98.7
[≥ 0.9]	99.3	98.8

Table VIII: IMPUTE2's internal cross-validation for chromosome 8. Tables show the percentage of concordance between genotyped calls and imputed calls for 10,412,010 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 − 0.1]	0	0.0
[0.1 − 0.2]	0	0.0
[0.2 − 0.3]	0	0.0
[0.3 − 0.4]	0	0.0
[0.4 − 0.5]	868	36.6
[0.5 − 0.6]	8,653	53.0
[0.6 − 0.7]	8,594	60.8
[0.7 − 0.8]	10,524	69.8
[0.8 − 0.9]	16,817	78.1
[0.9 − 1.0]	10,366,554	99.0

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.9
[≥ 0.1]	100.0	98.9
[≥ 0.2]	100.0	98.9
[≥ 0.3]	100.0	98.9
[≥ 0.4]	100.0	98.9
[≥ 0.5]	100.0	98.9
[≥ 0.6]	99.9	98.9
[≥ 0.7]	99.8	98.9
[≥ 0.8]	99.7	99.0
[≥ 0.9]	99.6	99.0

Table IX: IMPUTE2's internal cross-validation for chromosome 9. Tables show the percentage of concordance between genotyped calls and imputed calls for 8,442,990 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	983	37.0
[0.5 – 0.6]	8,639	52.9
[0.6 – 0.7]	9,017	60.7
[0.7 – 0.8]	10,952	68.7
[0.8 – 0.9]	17,362	77.7
[0.9 – 1.0]	8,396,037	98.7

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.5
[≥ 0.1]	100.0	98.5
[≥ 0.2]	100.0	98.5
[≥ 0.3]	100.0	98.5
[≥ 0.4]	100.0	98.5
[≥ 0.5]	100.0	98.5
[≥ 0.6]	99.9	98.6
[≥ 0.7]	99.8	98.6
[≥ 0.8]	99.6	98.7
[≥ 0.9]	99.4	98.7

Table X: IMPUTE2's internal cross-validation for chromosome 10. Tables show the percentage of concordance between genotyped calls and imputed calls for 8,925,210 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,387	35.1
[0.5 – 0.6]	11,082	52.5
[0.6 – 0.7]	11,175	58.9
[0.7 – 0.8]	13,576	67.8
[0.8 – 0.9]	21,170	76.6
[0.9 – 1.0]	8,866,820	98.7

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.5
[≥ 0.1]	100.0	98.5
[≥ 0.2]	100.0	98.5
[≥ 0.3]	100.0	98.5
[≥ 0.4]	100.0	98.5
[≥ 0.5]	100.0	98.6
[≥ 0.6]	99.8	98.6
[≥ 0.7]	99.7	98.6
[≥ 0.8]	99.6	98.7
[≥ 0.9]	99.3	98.7

Table XI: IMPUTE2's internal cross-validation for chromosome 11. Tables show the percentage of concordance between genotyped calls and imputed calls for 8,593,020 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,205	36.0
[0.5 – 0.6]	9,931	51.3
[0.6 – 0.7]	10,022	60.1
[0.7 – 0.8]	11,865	68.9
[0.8 – 0.9]	18,376	77.5
[0.9 – 1.0]	8,541,621	98.8

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.6
[≥ 0.1]	100.0	98.6
[≥ 0.2]	100.0	98.6
[≥ 0.3]	100.0	98.6
[≥ 0.4]	100.0	98.6
[≥ 0.5]	100.0	98.6
[≥ 0.6]	99.9	98.6
[≥ 0.7]	99.8	98.7
[≥ 0.8]	99.6	98.7
[≥ 0.9]	99.4	98.8

Table XII: IMPUTE2's internal cross-validation for chromosome 12. Tables show the percentage of concordance between genotyped calls and imputed calls for 8,039,970 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,472	38.9
[0.5 – 0.6]	10,571	52.4
[0.6 – 0.7]	10,450	60.3
[0.7 – 0.8]	13,188	68.4
[0.8 – 0.9]	20,625	77.3
[0.9 – 1.0]	7,983,664	98.8

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.5
[≥ 0.1]	100.0	98.5
[≥ 0.2]	100.0	98.5
[≥ 0.3]	100.0	98.5
[≥ 0.4]	100.0	98.5
[≥ 0.5]	100.0	98.5
[≥ 0.6]	99.8	98.6
[≥ 0.7]	99.7	98.7
[≥ 0.8]	99.5	98.7
[≥ 0.9]	99.3	98.8

Table XIII: IMPUTE2's internal cross-validation for chromosome 13. Tables show the percentage of concordance between genotyped calls and imputed calls for 6,720,480 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	826	37.2
[0.5 – 0.6]	7,137	52.9
[0.6 – 0.7]	7,761	59.5
[0.7 – 0.8]	9,241	68.0
[0.8 – 0.9]	14,217	76.8
[0.9 – 1.0]	6,681,298	98.9

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.7
[≥ 0.1]	100.0	98.7
[≥ 0.2]	100.0	98.7
[≥ 0.3]	100.0	98.7
[≥ 0.4]	100.0	98.7
[≥ 0.5]	100.0	98.7
[≥ 0.6]	99.9	98.7
[≥ 0.7]	99.8	98.8
[≥ 0.8]	99.6	98.8
[≥ 0.9]	99.4	98.9

Table XIV: IMPUTE2's internal cross-validation for chromosome 14. Tables show the percentage of concordance between genotyped calls and imputed calls for 5,804,370 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	635	36.4
[0.5 – 0.6]	5,697	52.9
[0.6 – 0.7]	6,023	60.9
[0.7 – 0.8]	7,049	70.5
[0.8 – 0.9]	11,414	78.3
[0.9 – 1.0]	5,773,552	99.0

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.8
[≥ 0.1]	100.0	98.8
[≥ 0.2]	100.0	98.8
[≥ 0.3]	100.0	98.8
[≥ 0.4]	100.0	98.8
[≥ 0.5]	100.0	98.8
[≥ 0.6]	99.9	98.9
[≥ 0.7]	99.8	98.9
[≥ 0.8]	99.7	99.0
[≥ 0.9]	99.5	99.0

Table XV: IMPUTE2's internal cross-validation for chromosome 15. Tables show the percentage of concordance between genotyped calls and imputed calls for 4,791,060 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	754	41.9
[0.5 – 0.6]	6,589	53.0
[0.6 – 0.7]	6,795	59.6
[0.7 – 0.8]	8,076	68.5
[0.8 – 0.9]	13,349	78.2
[0.9 – 1.0]	4,755,497	98.9

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.6
[≥ 0.1]	100.0	98.6
[≥ 0.2]	100.0	98.6
[≥ 0.3]	100.0	98.6
[≥ 0.4]	100.0	98.6
[≥ 0.5]	100.0	98.6
[≥ 0.6]	99.8	98.7
[≥ 0.7]	99.7	98.7
[≥ 0.8]	99.5	98.8
[≥ 0.9]	99.3	98.9

Table XVI: IMPUTE2's internal cross-validation for chromosome 16. Tables show the percentage of concordance between genotyped calls and imputed calls for 4,533,930 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,184	36.9
[0.5 – 0.6]	9,655	50.9
[0.6 – 0.7]	9,464	57.9
[0.7 – 0.8]	11,548	66.7
[0.8 – 0.9]	18,250	75.5
[0.9 – 1.0]	4,483,829	98.5

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.1
[≥ 0.1]	100.0	98.1
[≥ 0.2]	100.0	98.1
[≥ 0.3]	100.0	98.1
[≥ 0.4]	100.0	98.1
[≥ 0.5]	100.0	98.1
[≥ 0.6]	99.8	98.2
[≥ 0.7]	99.5	98.3
[≥ 0.8]	99.3	98.4
[≥ 0.9]	98.9	98.5

Table XVII: IMPUTE2's internal cross-validation for chromosome 17. Tables show the percentage of concordance between genotyped calls and imputed calls for 3,821,760 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	1,100	37.1
[0.5 – 0.6]	8,470	51.6
[0.6 – 0.7]	8,731	59.5
[0.7 – 0.8]	10,219	68.1
[0.8 – 0.9]	15,853	75.2
[0.9 – 1.0]	3,777,387	98.4

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.1
[≥ 0.1]	100.0	98.1
[≥ 0.2]	100.0	98.1
[≥ 0.3]	100.0	98.1
[≥ 0.4]	100.0	98.1
[≥ 0.5]	100.0	98.1
[≥ 0.6]	99.8	98.2
[≥ 0.7]	99.5	98.3
[≥ 0.8]	99.2	98.4
[≥ 0.9]	98.8	98.4

Table XVIII: IMPUTE2's internal cross-validation for chromosome 18. Tables show the percentage of concordance between genotyped calls and imputed calls for 5,635,350 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	651	41.5
[0.5 – 0.6]	6,222	51.1
[0.6 – 0.7]	6,161	60.3
[0.7 – 0.8]	7,403	69.1
[0.8 – 0.9]	11,506	78.7
[0.9 – 1.0]	5,603,407	99.0

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.8
[≥ 0.1]	100.0	98.8
[≥ 0.2]	100.0	98.8
[≥ 0.3]	100.0	98.8
[≥ 0.4]	100.0	98.8
[≥ 0.5]	100.0	98.8
[≥ 0.6]	99.9	98.8
[≥ 0.7]	99.8	98.9
[≥ 0.8]	99.6	98.9
[≥ 0.9]	99.4	99.0

Table XIX: IMPUTE2's internal cross-validation for chromosome 19. Tables show the percentage of concordance between genotyped calls and imputed calls for 2,419,650 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	942	38.9
[0.5 – 0.6]	7,187	51.3
[0.6 – 0.7]	7,397	57.4
[0.7 – 0.8]	8,600	67.1
[0.8 – 0.9]	13,276	76.3
[0.9 – 1.0]	2,382,248	98.1

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	97.5
[≥ 0.1]	100.0	97.5
[≥ 0.2]	100.0	97.5
[≥ 0.3]	100.0	97.5
[≥ 0.4]	100.0	97.5
[≥ 0.5]	100.0	97.6
[≥ 0.6]	99.7	97.7
[≥ 0.7]	99.4	97.8
[≥ 0.8]	99.0	97.9
[≥ 0.9]	98.5	98.1

Table XX: IMPUTE2's internal cross-validation for chromosome 20. Tables show the percentage of concordance between genotyped calls and imputed calls for 4,379,490 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 – 0.1]	0	0.0
[0.1 – 0.2]	0	0.0
[0.2 – 0.3]	0	0.0
[0.3 – 0.4]	0	0.0
[0.4 – 0.5]	620	32.7
[0.5 – 0.6]	5,657	52.6
[0.6 – 0.7]	5,462	60.4
[0.7 – 0.8]	6,761	67.1
[0.8 – 0.9]	10,550	77.1
[0.9 – 1.0]	4,350,440	98.9

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.6
[≥ 0.1]	100.0	98.6
[≥ 0.2]	100.0	98.6
[≥ 0.3]	100.0	98.6
[≥ 0.4]	100.0	98.6
[≥ 0.5]	100.0	98.7
[≥ 0.6]	99.9	98.7
[≥ 0.7]	99.7	98.8
[≥ 0.8]	99.6	98.8
[≥ 0.9]	99.3	98.9

Table XXI: IMPUTE2's internal cross-validation for chromosome 21. Tables show the percentage of concordance between genotyped calls and imputed calls for 2,423,520 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 − 0.1]	0	0.0
[0.1 − 0.2]	0	0.0
[0.2 − 0.3]	0	0.0
[0.3 − 0.4]	0	0.0
[0.4 − 0.5]	507	37.5
[0.5 − 0.6]	3,843	50.4
[0.6 − 0.7]	3,825	59.1
[0.7 − 0.8]	4,652	67.6
[0.8 − 0.9]	7,234	77.1
[0.9 − 1.0]	2,403,459	98.5

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.2
[≥ 0.1]	100.0	98.2
[≥ 0.2]	100.0	98.2
[≥ 0.3]	100.0	98.2
[≥ 0.4]	100.0	98.2
[≥ 0.5]	100.0	98.3
[≥ 0.6]	99.8	98.4
[≥ 0.7]	99.7	98.4
[≥ 0.8]	99.5	98.5
[≥ 0.9]	99.2	98.5

Table XXII: IMPUTE2's internal cross-validation for chromosome 22. Tables show the percentage of concordance between genotyped calls and imputed calls for 2,343,690 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 − 0.1]	0	0.0
[0.1 − 0.2]	0	0.0
[0.2 − 0.3]	0	0.0
[0.3 − 0.4]	0	0.0
[0.4 − 0.5]	459	40.8
[0.5 − 0.6]	4,173	52.8
[0.6 − 0.7]	4,434	60.7
[0.7 − 0.8]	5,250	67.4
[0.8 − 0.9]	8,322	77.5
[0.9 − 1.0]	2,321,052	98.5

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.2
[≥ 0.1]	100.0	98.2
[≥ 0.2]	100.0	98.2
[≥ 0.3]	100.0	98.2
[≥ 0.4]	100.0	98.2
[≥ 0.5]	100.0	98.2
[≥ 0.6]	99.8	98.3
[≥ 0.7]	99.6	98.4
[≥ 0.8]	99.4	98.5
[≥ 0.9]	99.0	98.5

Table XXIII: IMPUTE2's internal cross-validation across the genome. Tables show the percentage of concordance between genotyped calls and imputed calls for 170,926,020 genotypes.

Interval	Nb Geno	Concordance (%)
[0.0 − 0.1]	0	0.0
[0.1 − 0.2]	0	0.0
[0.2 − 0.3]	0	0.0
[0.3 − 0.4]	0	0.0
[0.4 − 0.5]	26,453	36.6
[0.5 − 0.6]	213,033	51.8
[0.6 − 0.7]	214,764	59.6
[0.7 − 0.8]	259,384	67.9
[0.8 − 0.9]	407,255	76.8
[0.9 − 1.0]	169,805,131	98.7

Interval	Called (%)	Concordance (%)
[≥ 0.0]	100.0	98.5
[≥ 0.1]	100.0	98.5
[≥ 0.2]	100.0	98.5
[≥ 0.3]	100.0	98.5
[≥ 0.4]	100.0	98.5
[≥ 0.5]	100.0	98.5
[≥ 0.6]	99.9	98.6
[≥ 0.7]	99.7	98.6
[≥ 0.8]	99.6	98.7
[≥ 0.9]	99.3	98.7

3.2 Completion rate

To evaluate the completion rate, we first used a probability threshold of ≥ 90.0%, which means that a genotype must have one of the three allele combination (`AA`, `AB` or `BB`) probabilities higher or equal to 90.0% to be considered as a *good call*.

For the 13,771,150 imputed variants, an average completion rate of 98.9% was obtained. When removing variants with a completion rate under 98.0%, 12,287,509 (89.2%) markers were left, with an average completion rate of 100.0%, meaning that there is a mean of 0.0 missing genotypes (for 90 samples) for each markers.

A total of 1,928,081 variants were previously genotyped, 406,033 (21.1%) of which had a call rate lower than 100% (*i.e.* 406,033 missing genotypes). A total of 406,033 (100.0%) missing genotypes were imputed with high quality (*i.e.* 1,928,081 markers now have a call rate of 100%).

3.3 Minor allele frequencies

Out of the 12,287,509 imputed variants with a completion rate $\geq$ 98.0%, there were 7,354,048 (59.8%) variants with a minor allele frequency (MAF) $\geq$ 1%, 5,835,100 (47.5%) variants with a MAF $\geq$ 5%, and 6,452,409 (52.5%) variants with a MAF $<$ 5%. Figure 1 shows the proportions of ultra rare ($MAF < 1\%$), rare ($1\% \leq MAF < 5\%$) and common ($MAF \geq 5\%$) variants.

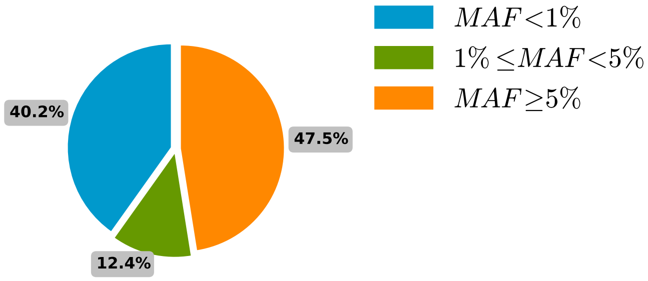

Figure 1: Proportions of minor allele frequencies for imputed sites with a completion rate of 98.0% or more at a probability of 90.0% or more.

4 Conclusions

Statistical analyses will be performed with the genome-wide imputed dataset, which include 12,287,509 imputed variants (done with an imputation probability threshold of $\geq$ 90.0% and a completion rate of $\geq$ 98.0%, including 1,928,081 previously genotyped variants.

All files were generated in the `gwip` directory and were separated by chromosomes (`gwip/chr*` directories). The final (merged) results (generated by IMPUTE2) are located in the `gwip/chr*/final_impute2` directories. All the output files are described below.

- `chr*.imputed.alleles`: description of the reference and alternative allele at each site.
- `chr*.imputed.completion_rates`: number of missing values and completion rate for all site (using a probability threshold $\geq$ 90.0%).
- `chr*.imputed.good_sites`: list of sites which pass the completion rate threshold ($\geq$ 98.0%) using the probability threshold $\geq$ 90.0%.
- `chr*.imputed.impute2`: imputation results (merged from all segments.
- `chr*.imputed.imputed_sites`: list of imputed sites (excluding sites that were previously genotyped in the study cohort).
- `chr*.imputed.log`: log file of the merging procedure.
- `chr*.imputed.maf`: minor allele frequency (along with minor allele identification) for all sites using the probability threshold $\geq$ 90.0%.
- `chr*.imputed.map`: a map file describing the genomic location of all sites.
- `chr*.imputed.sample`: the sample file generated by the phasing step.

References

[1] Delaneau O, Zagury JF, Marchini J: **Improved whole-chromosome phasing for disease and population genetic studies**. *Nature methods* 2013, **10**:5–6. [DOI:10.1038/nmeth.2307].

[2] Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS genetics* 2009, **5**(6):e1000529. [DOI:10.1371/journal.pgen.1000529].

[3] Howie B, Marchini J, Stephens M: **Genotype imputation with thousands of genomes**. *G3: Genes, Genomes, Genetics* 2011, **1**(6):457–470. [DOI:10.1534/g3.111.001198].

[4] Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: **Fast and accurate genotype imputation in genome-wide association studies through pre-phasing**. *Nature Genetics* 2012, **44**(8):955–959. [DOI:10.1038/ng.2354].

[5] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al.: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *The American Journal of Human Genetics* 2007, **81**(3):559–575. [DOI:10.1086/519795].

Annex I: Execution times

The following tables show the execution time required by all the different tasks. All tasks are split by chromosomes. Execution times for imputation for each chromosome are means of individual segment times. Computing all genotyped markers' missing rate took 22 seconds.

Table XXIV: Execution time for the '`plink_exclude_chr*`' tasks.

Chrom	Time	Chrom	Time
1	00:00:13	12	00:00:11
2	00:00:13	13	00:00:10
3	00:00:13	14	00:00:11
4	00:00:11	15	00:00:10
5	00:00:13	16	00:00:10
6	00:00:12	17	00:00:08
7	00:00:13	18	00:00:09
8	00:00:12	19	00:00:09
9	00:00:10	20	00:00:09
10	00:00:11	21	00:00:09
11	00:00:11	22	00:00:09

Table XXV: Execution time for the '`shapeit_check_chr*_1`' tasks.

Chrom	Time	Chrom	Time
1	00:00:27	12	00:00:16
2	00:00:35	13	00:00:14
3	00:00:22	14	00:00:09
4	00:00:28	15	00:00:11
5	00:00:26	16	00:00:10
6	00:00:22	17	00:00:09
7	00:00:18	18	00:00:08
8	00:00:20	19	00:00:06
9	00:00:14	20	00:00:06
10	00:00:20	21	00:00:04
11	00:00:15	22	00:00:04

Table XXVI: Execution time for the '`plink_flip_chr*`' tasks.

Chrom	Time	Chrom	Time
1	00:00:02	12	00:00:01
2	00:00:02	13	00:00:01
3	00:00:02	14	00:00:01
4	00:00:01	15	00:00:01
5	00:00:01	16	00:00:01
6	00:00:02	17	00:00:00
7	00:00:01	18	00:00:01
8	00:00:01	19	00:00:00
9	00:00:01	20	00:00:01
10	00:00:01	21	00:00:00
11	00:00:01	22	00:00:00

Table XXVII: Execution time for the 'shapeit_check_chr*_2' tasks.

Chrom	Time
1	00:00:23
2	00:00:24
3	00:00:19
4	00:00:20
5	00:00:22
6	00:00:18
7	00:00:15
8	00:00:16
9	00:00:11
10	00:00:13
11	00:00:13
12	00:00:13
13	00:00:09
14	00:00:08
15	00:00:08
16	00:00:08
17	00:00:07
18	00:00:08
19	00:00:05
20	00:00:06
21	00:00:04
22	00:00:04

Table XXVIII: Execution time for the 'plink_final_exclude_chr*' tasks.

Chrom	Time
1	00:00:02
2	00:00:02
3	00:00:02
4	00:00:01
5	00:00:02
6	00:00:01
7	00:00:01
8	00:00:01
9	00:00:01
10	00:00:01
11	00:00:01
12	00:00:01
13	00:00:01
14	00:00:01
15	00:00:01
16	00:00:01
17	00:00:00
18	00:00:01
19	00:00:00
20	00:00:01
21	00:00:00
22	00:00:00

Table XXIX: Execution time for the 'shapeit_phase_chr*' tasks.

Chrom	Time
1	01:33:58
2	01:43:31
3	01:19:04
4	01:13:50
5	01:14:30
6	01:15:38
7	01:02:36
8	01:05:48
9	00:56:08
10	00:57:56
11	00:54:33
12	00:52:53
13	00:42:25
14	00:36:59
15	00:32:05
16	00:31:57
17	00:27:22
18	00:34:26
19	00:17:57
20	00:27:27
21	00:15:12
22	00:14:27

Table XXX: Execution time for the 'impute2_chr*' tasks.

Chrom	Nb Seg.	Mean T.	Max T.
1	50	00:02:12	00:03:41
2	49	00:02:26	00:03:59
3	40	00:02:12	00:03:36
4	39	00:02:13	00:03:35
5	37	00:02:02	00:03:06
6	35	00:02:06	00:03:20
7	32	00:02:00	00:03:06
8	30	00:02:04	00:03:03
9	29	00:01:36	00:02:37
10	28	00:01:58	00:03:12
11	28	00:01:51	00:02:39
12	27	00:01:51	00:02:54
13	24	00:01:29	00:02:38
14	22	00:01:20	00:02:30
15	21	00:01:18	00:02:29
16	19	00:01:32	00:02:55
17	17	00:01:23	00:02:07
18	16	00:01:33	00:02:16
19	12	00:01:32	00:02:30
20	13	00:01:23	00:01:52
21	10	00:01:04	00:01:50
22	11	00:00:53	00:02:02

Table XXXI: Execution time for the 'merge_impute2_chr*' tasks.

Chrom	Time
1	00:05:09
2	00:05:38
3	00:04:31
4	00:04:46
5	00:04:06
6	00:04:29
7	00:03:42
8	00:03:41
9	00:03:02
10	00:03:23
11	00:03:19
12	00:03:15
13	00:02:29
14	00:02:33
15	00:02:15
16	00:02:23
17	00:02:07
18	00:02:10
19	00:01:47
20	00:01:15
21	00:00:46
22	00:01:05

Table XXXII: Execution time for the 'bgzip_chr*' tasks.

Chrom	Time
1	00:00:21
2	00:00:20
3	00:00:21
4	00:00:18
5	00:00:16
6	00:00:20
7	00:00:25
8	00:00:13
9	00:00:08
10	00:00:17
11	00:00:12
12	00:00:08
13	00:00:06
14	00:00:05
15	00:00:05
16	00:00:06
17	00:00:05
18	00:00:05
19	00:00:04
20	00:00:04
21	00:00:03
22	00:00:03