Tutorial

Test Report

Automatically generated by genipe

June 16th, 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
$[0.4 - 0.5]$	3,235	34.6
$[0.5 - 0.6]$	21,616	50.0
$[0.6 - 0.7]$	21,382	57.8
$[0.7 - 0.8]$	25,600	65.6
$[0.8 - 0.9]$	39,780	74.2
$[0.9 - 1.0]$	13,168,787	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
$[0.4 - 0.5]$	1,633	36.8
$[0.5 - 0.6]$	14,993	51.7
$[0.6 - 0.7]$	15,204	59.8
$[0.7 - 0.8]$	18,570	68.7
$[0.8 - 0.9]$	29,524	77.7
$[0.9 - 1.0]$	15,563,966	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
$[0.4 - 0.5]$	1,981	36.0
$[0.5 - 0.6]$	14,418	52.2
$[0.6 - 0.7]$	14,484	60.3
$[0.7 - 0.8]$	17,452	68.7
$[0.8 - 0.9]$	27,950	77.6
$[0.9 - 1.0]$	11,597,705	98.6

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.4
$[\geq 0.1]$	100.0	98.4
$[\geq 0.2]$	100.0	98.4
$[\geq 0.3]$	100.0	98.4
$[\geq 0.4]$	100.0	98.4
$[\geq 0.5]$	100.0	98.4
$[\geq 0.6]$	99.9	98.5
$[\geq 0.7]$	99.7	98.5
$[\geq 0.8]$	99.6	98.6
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,790	35.9
$[0.5 - 0.6]$	14,006	51.9
$[0.6 - 0.7]$	14,393	58.7
$[0.7 - 0.8]$	17,294	66.8
$[0.8 - 0.9]$	26,825	76.4
$[0.9 - 1.0]$	10,871,042	98.4

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.2
$[\geq 0.1]$	100.0	98.2
$[\geq 0.2]$	100.0	98.2
$[\geq 0.3]$	100.0	98.2
$[\geq 0.4]$	100.0	98.2
$[\geq 0.5]$	100.0	98.2
$[\geq 0.6]$	99.8	98.3
$[\geq 0.7]$	99.7	98.3
$[\geq 0.8]$	99.6	98.4
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,428	36.9
$[0.5 - 0.6]$	11,505	51.6
$[0.6 - 0.7]$	11,363	60.0
$[0.7 - 0.8]$	13,990	68.2
$[0.8 - 0.9]$	21,982	76.6
$[0.9 - 1.0]$	10,892,552	98.8

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.6
$[\geq 0.1]$	100.0	98.6
$[\geq 0.2]$	100.0	98.6
$[\geq 0.3]$	100.0	98.6
$[\geq 0.4]$	100.0	98.6
$[\geq 0.5]$	100.0	98.6
$[\geq 0.6]$	99.9	98.7
$[\geq 0.7]$	99.8	98.7
$[\geq 0.8]$	99.7	98.7
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,283	36.6
$[0.5 - 0.6]$	11,196	50.7
$[0.6 - 0.7]$	11,011	60.6
$[0.7 - 0.8]$	13,483	67.8
$[0.8 - 0.9]$	21,116	76.5
$[0.9 - 1.0]$	11,904,711	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	98.7
$[\geq 0.7]$	99.8	98.8
$[\geq 0.8]$	99.7	98.8
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,515	34.1
$[0.5 - 0.6]$	11,870	52.4
$[0.6 - 0.7]$	11,657	60.0
$[0.7 - 0.8]$	14,040	68.2
$[0.8 - 0.9]$	21,837	77.3
$[0.9 - 1.0]$	9,119,351	98.8

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.6
$[\geq 0.1]$	100.0	98.6
$[\geq 0.2]$	100.0	98.6
$[\geq 0.3]$	100.0	98.6
$[\geq 0.4]$	100.0	98.6
$[\geq 0.5]$	100.0	98.6
$[\geq 0.6]$	99.9	98.6
$[\geq 0.7]$	99.7	98.7
$[\geq 0.8]$	99.6	98.7
$[\geq 0.9]$	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
$[0.4 - 0.5]$	861	36.7
$[0.5 - 0.6]$	8,681	52.9
$[0.6 - 0.7]$	8,608	60.7
$[0.7 - 0.8]$	10,482	69.9
$[0.8 - 0.9]$	16,817	78.1
$[0.9 - 1.0]$	10,366,561	99.0

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.9
$[\geq 0.1]$	100.0	98.9
$[\geq 0.2]$	100.0	98.9
$[\geq 0.3]$	100.0	98.9
$[\geq 0.4]$	100.0	98.9
$[\geq 0.5]$	100.0	98.9
$[\geq 0.6]$	99.9	98.9
$[\geq 0.7]$	99.8	98.9
$[\geq 0.8]$	99.7	99.0
$[\geq 0.9]$	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
$[0.4 - 0.5]$	984	37.7
$[0.5 - 0.6]$	8,634	53.1
$[0.6 - 0.7]$	9,025	60.9
$[0.7 - 0.8]$	10,948	68.5
$[0.8 - 0.9]$	17,397	77.7
$[0.9 - 1.0]$	8,396,002	98.7

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.5
$[\geq 0.1]$	100.0	98.5
$[\geq 0.2]$	100.0	98.5
$[\geq 0.3]$	100.0	98.5
$[\geq 0.4]$	100.0	98.5
$[\geq 0.5]$	100.0	98.5
$[\geq 0.6]$	99.9	98.6
$[\geq 0.7]$	99.8	98.6
$[\geq 0.8]$	99.6	98.7
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,386	35.1
$[0.5 - 0.6]$	11,071	52.7
$[0.6 - 0.7]$	11,165	58.8
$[0.7 - 0.8]$	13,597	67.7
$[0.8 - 0.9]$	21,160	76.7
$[0.9 - 1.0]$	8,866,831	98.7

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.5
$[\geq 0.1]$	100.0	98.5
$[\geq 0.2]$	100.0	98.5
$[\geq 0.3]$	100.0	98.5
$[\geq 0.4]$	100.0	98.5
$[\geq 0.5]$	100.0	98.6
$[\geq 0.6]$	99.8	98.6
$[\geq 0.7]$	99.7	98.6
$[\geq 0.8]$	99.6	98.7
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,210	36.1
$[0.5 - 0.6]$	9,946	51.2
$[0.6 - 0.7]$	10,005	60.1
$[0.7 - 0.8]$	11,916	68.7
$[0.8 - 0.9]$	18,307	77.4
$[0.9 - 1.0]$	8,541,636	98.8

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.6
$[\geq 0.1]$	100.0	98.6
$[\geq 0.2]$	100.0	98.6
$[\geq 0.3]$	100.0	98.6
$[\geq 0.4]$	100.0	98.6
$[\geq 0.5]$	100.0	98.6
$[\geq 0.6]$	99.9	98.7
$[\geq 0.7]$	99.7	98.7
$[\geq 0.8]$	99.6	98.7
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,467	39.2
$[0.5 - 0.6]$	10,596	52.3
$[0.6 - 0.7]$	10,454	60.3
$[0.7 - 0.8]$	13,193	68.4
$[0.8 - 0.9]$	20,631	77.4
$[0.9 - 1.0]$	7,983,629	98.8

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.5
$[\geq 0.1]$	100.0	98.5
$[\geq 0.2]$	100.0	98.5
$[\geq 0.3]$	100.0	98.5
$[\geq 0.4]$	100.0	98.5
$[\geq 0.5]$	100.0	98.5
$[\geq 0.6]$	99.8	98.6
$[\geq 0.7]$	99.7	98.7
$[\geq 0.8]$	99.5	98.7
$[\geq 0.9]$	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
$[0.4 - 0.5]$	837	36.1
$[0.5 - 0.6]$	7,113	52.9
$[0.6 - 0.7]$	7,762	59.0
$[0.7 - 0.8]$	9,247	68.1
$[0.8 - 0.9]$	14,197	76.8
$[0.9 - 1.0]$	6,681,324	98.9

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.6
$[\geq 0.1]$	100.0	98.6
$[\geq 0.2]$	100.0	98.6
$[\geq 0.3]$	100.0	98.6
$[\geq 0.4]$	100.0	98.6
$[\geq 0.5]$	100.0	98.7
$[\geq 0.6]$	99.9	98.7
$[\geq 0.7]$	99.8	98.8
$[\geq 0.8]$	99.6	98.8
$[\geq 0.9]$	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
$[0.4 - 0.5]$	631	36.4
$[0.5 - 0.6]$	5,716	52.8
$[0.6 - 0.7]$	6,086	60.9
$[0.7 - 0.8]$	7,043	70.5
$[0.8 - 0.9]$	11,422	78.2
$[0.9 - 1.0]$	5,773,472	99.0

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.8
$[\geq 0.1]$	100.0	98.8
$[\geq 0.2]$	100.0	98.8
$[\geq 0.3]$	100.0	98.8
$[\geq 0.4]$	100.0	98.8
$[\geq 0.5]$	100.0	98.8
$[\geq 0.6]$	99.9	98.9
$[\geq 0.7]$	99.8	98.9
$[\geq 0.8]$	99.7	99.0
$[\geq 0.9]$	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
$[0.4 - 0.5]$	757	42.7
$[0.5 - 0.6]$	6,595	53.4
$[0.6 - 0.7]$	6,804	59.4
$[0.7 - 0.8]$	8,068	68.5
$[0.8 - 0.9]$	13,350	78.1
$[0.9 - 1.0]$	4,755,486	98.9

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.6
$[\geq 0.1]$	100.0	98.6
$[\geq 0.2]$	100.0	98.6
$[\geq 0.3]$	100.0	98.6
$[\geq 0.4]$	100.0	98.6
$[\geq 0.5]$	100.0	98.6
$[\geq 0.6]$	99.8	98.7
$[\geq 0.7]$	99.7	98.7
$[\geq 0.8]$	99.5	98.8
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,194	36.6
$[0.5 - 0.6]$	9,651	50.9
$[0.6 - 0.7]$	9,536	57.8
$[0.7 - 0.8]$	11,521	66.6
$[0.8 - 0.9]$	18,234	75.6
$[0.9 - 1.0]$	4,483,794	98.5

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.1
$[\geq 0.1]$	100.0	98.1
$[\geq 0.2]$	100.0	98.1
$[\geq 0.3]$	100.0	98.1
$[\geq 0.4]$	100.0	98.1
$[\geq 0.5]$	100.0	98.1
$[\geq 0.6]$	99.8	98.2
$[\geq 0.7]$	99.5	98.3
$[\geq 0.8]$	99.3	98.4
$[\geq 0.9]$	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
$[0.4 - 0.5]$	1,112	38.5
$[0.5 - 0.6]$	8,487	51.6
$[0.6 - 0.7]$	8,690	59.4
$[0.7 - 0.8]$	10,246	68.1
$[0.8 - 0.9]$	15,828	74.9
$[0.9 - 1.0]$	3,777,397	98.4

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.1
$[\geq 0.1]$	100.0	98.1
$[\geq 0.2]$	100.0	98.1
$[\geq 0.3]$	100.0	98.1
$[\geq 0.4]$	100.0	98.1
$[\geq 0.5]$	100.0	98.1
$[\geq 0.6]$	99.8	98.2
$[\geq 0.7]$	99.5	98.3
$[\geq 0.8]$	99.2	98.4
$[\geq 0.9]$	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
$[0.4 - 0.5]$	637	41.0
$[0.5 - 0.6]$	6,196	51.1
$[0.6 - 0.7]$	6,134	60.2
$[0.7 - 0.8]$	7,397	69.1
$[0.8 - 0.9]$	11,495	78.7
$[0.9 - 1.0]$	5,603,491	99.0

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.8
$[\geq 0.1]$	100.0	98.8
$[\geq 0.2]$	100.0	98.8
$[\geq 0.3]$	100.0	98.8
$[\geq 0.4]$	100.0	98.8
$[\geq 0.5]$	100.0	98.8
$[\geq 0.6]$	99.9	98.8
$[\geq 0.7]$	99.8	98.9
$[\geq 0.8]$	99.6	98.9
$[\geq 0.9]$	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
$[0.4 - 0.5]$	915	38.2
$[0.5 - 0.6]$	7,199	50.9
$[0.6 - 0.7]$	7,354	57.7
$[0.7 - 0.8]$	8,615	67.1
$[0.8 - 0.9]$	13,219	76.4
$[0.9 - 1.0]$	2,382,348	98.1

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	97.6
$[\geq 0.1]$	100.0	97.6
$[\geq 0.2]$	100.0	97.6
$[\geq 0.3]$	100.0	97.6
$[\geq 0.4]$	100.0	97.6
$[\geq 0.5]$	100.0	97.6
$[\geq 0.6]$	99.7	97.7
$[\geq 0.7]$	99.4	97.8
$[\geq 0.8]$	99.0	97.9
$[\geq 0.9]$	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
$[0.4 - 0.5]$	619	34.7
$[0.5 - 0.6]$	5,668	52.7
$[0.6 - 0.7]$	5,457	60.4
$[0.7 - 0.8]$	6,731	67.1
$[0.8 - 0.9]$	10,578	77.2
$[0.9 - 1.0]$	4,350,437	98.9

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.6
$[\geq 0.1]$	100.0	98.6
$[\geq 0.2]$	100.0	98.6
$[\geq 0.3]$	100.0	98.6
$[\geq 0.4]$	100.0	98.6
$[\geq 0.5]$	100.0	98.7
$[\geq 0.6]$	99.9	98.7
$[\geq 0.7]$	99.7	98.8
$[\geq 0.8]$	99.6	98.8
$[\geq 0.9]$	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
$[0.4 - 0.5]$	508	35.4
$[0.5 - 0.6]$	3,823	50.4
$[0.6 - 0.7]$	3,788	59.2
$[0.7 - 0.8]$	4,650	67.4
$[0.8 - 0.9]$	7,226	77.1
$[0.9 - 1.0]$	2,403,525	98.5

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.3
$[\geq 0.1]$	100.0	98.3
$[\geq 0.2]$	100.0	98.3
$[\geq 0.3]$	100.0	98.3
$[\geq 0.4]$	100.0	98.3
$[\geq 0.5]$	100.0	98.3
$[\geq 0.6]$	99.8	98.4
$[\geq 0.7]$	99.7	98.4
$[\geq 0.8]$	99.5	98.5
$[\geq 0.9]$	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
$[0.4 - 0.5]$	468	41.5
$[0.5 - 0.6]$	4,158	52.7
$[0.6 - 0.7]$	4,456	60.6
$[0.7 - 0.8]$	5,271	67.4
$[0.8 - 0.9]$	8,312	77.5
$[0.9 - 1.0]$	2,321,025	98.5

Interval	Called (%)	Concordance (%)
$[\geq 0.0]$	100.0	98.2
$[\geq 0.1]$	100.0	98.2
$[\geq 0.2]$	100.0	98.2
$[\geq 0.3]$	100.0	98.2
$[\geq 0.4]$	100.0	98.2
$[\geq 0.5]$	100.0	98.2
$[\geq 0.6]$	99.8	98.3
$[\geq 0.7]$	99.6	98.4
$[\geq 0.8]$	99.4	98.5
$[\geq 0.9]$	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
$[0.4 - 0.5]$	26,451	36.6
$[0.5 - 0.6]$	213,138	51.8
$[0.6 - 0.7]$	214,818	59.5
$[0.7 - 0.8]$	259,354	67.9
$[0.8 - 0.9]$	407,187	76.8
$[0.9 - 1.0]$	169,805,072	98.7

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

3.2 Completion rate

To evaluate the completion rate, we first used a probability threshold of $\geq 90.0\%$, 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 an information value under 0.00, and a completion rate under 98.0%, 12,287,066 (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 12,287,066 imputed variants with a completion rate $\geq 98.0\%$, there were 7,353,673 (59.8%) variants with a minor allele frequency (MAF) $\geq 1\%$, 5,834,834 (47.5%) variants with a MAF $\geq 5\%$, and 6,452,232 (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.

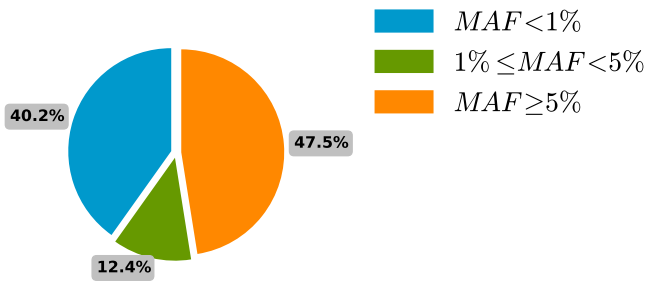

Figure 1: Proportions of minor allele frequencies for imputed sites with a completion rate of 98.0% or more at a probability of 90.0% or more.

4 Conclusions

Statistical analyses will be performed with the genome-wide imputed dataset, which include 12,287,066 imputed variants (done with an information threshold of ≥ 0.00, and a completion rate of $\geq 98.0\%$ at an imputation probability threshold of $\geq 90.0\%$). This total includes 1,928,081 previously genotyped variants.

All files were generated in the `genipe` directory and were separated by chromosomes (`genipe/chr*` directories). The final (merged) results (generated by IMPUTE2) are located in the `genipe/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 information threshold (≥ 0.00) and 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.impute2_info`: the IMPUTE2 marker-wise information file (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	00:00:12	12	00:00:11
2	00:00:12	13	00:00:10
3	00:00:12	14	00:00:10
4	00:00:12	15	00:00:10
5	00:00:12	16	00:00:10
6	00:00:11	17	00:00:09
7	00:00:11	18	00:00:09
8	00:00:11	19	00:00:09
9	00:00:11	20	00:00:09
10	00:00:11	21	00:00:09
11	00:00:10	22	00:00:09

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

Chrom	Time	Chrom	Time
1	00:00:29	12	00:00:15
2	00:00:26	13	00:00:10
3	00:00:32	14	00:00:09
4	00:00:27	15	00:00:10
5	00:00:22	16	00:00:12
6	00:00:23	17	00:00:09
7	00:00:24	18	00:00:07
8	00:00:25	19	00:00:06
9	00:00:16	20	00:00:06
10	00:00:19	21	00:00:06
11	00:00:18	22	00:00:03

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	00:00:01	14	00:00:01
4	00:00:01	15	00:00:01
5	00:00:01	16	00:00:00
6	00:00:01	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	00:00:00
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	Chrom	Time
1	00:00:19	12	00:00:12
2	00:00:22	13	00:00:10
3	00:00:17	14	00:00:08
4	00:00:17	15	00:00:07
5	00:00:16	16	00:00:09
6	00:00:16	17	00:00:07
7	00:00:14	18	00:00:07
8	00:00:14	19	00:00:05
9	00:00:11	20	00:00:05
10	00:00:14	21	00:00:03
11	00:00:13	22	00:00:03

Table XXVIII: Execution time for the 'plink_final_exclude_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	00:00:01	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	00:00:00
10	00:00:01	21	00:00:00
11	00:00:01	22	00:00:00

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

Chrom	Time	Chrom	Time
1	01:35:00	12	00:54:24
2	01:43:59	13	00:42:52
3	01:19:47	14	00:38:19
4	01:14:44	15	00:32:47
5	01:14:45	16	00:33:16
6	01:15:56	17	00:28:08
7	01:03:16	18	00:36:58
8	01:06:03	19	00:18:24
9	00:55:42	20	00:28:33
10	01:00:35	21	00:15:45
11	00:57:09	22	00:15:14

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

Chrom	Nb Seg.	Mean T.	Max T.	Chrom	Nb Seg.	Mean T.	Max T.
1	50	00:02:20	00:04:07	12	27	00:01:56	00:03:16
2	49	00:02:34	00:04:15	13	24	00:01:28	00:02:50
3	40	00:02:23	00:03:43	14	22	00:01:26	00:02:39
4	39	00:02:25	00:03:43	15	21	00:01:21	00:02:26
5	37	00:02:11	00:03:16	16	19	00:01:37	00:03:21
6	35	00:02:20	00:03:35	17	17	00:01:26	00:02:10
7	32	00:02:11	00:03:23	18	16	00:01:38	00:02:17
8	30	00:02:14	00:03:18	19	12	00:01:24	00:01:54
9	29	00:01:42	00:02:44	20	13	00:01:32	00:02:03
10	28	00:02:00	00:03:03	21	10	00:01:07	00:01:47
11	28	00:02:02	00:02:59	22	11	00:01:00	00:02:18

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

Chrom	Time	Chrom	Time
1	00:04:02	12	00:02:30
2	00:04:16	13	00:01:53
3	00:03:38	14	00:01:41
4	00:03:44	15	00:01:31
5	00:03:16	16	00:01:39
6	00:03:27	17	00:01:28
7	00:03:04	18	00:01:20
8	00:02:53	19	00:01:13
9	00:02:17	20	00:01:05
10	00:02:36	21	00:00:43
11	00:02:35	22	00:00:42

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

Chrom	Time	Chrom	Time
1	00:00:58	12	00:00:13
2	00:00:54	13	00:00:25
3	00:00:48	14	00:00:09
4	00:00:53	15	00:00:20
5	00:00:36	16	00:00:06
6	00:00:45	17	00:00:07
7	00:00:48	18	00:00:05
8	00:00:33	19	00:00:05
9	00:00:35	20	00:00:04
10	00:00:21	21	00:00:03
11	00:00:30	22	00:00:03