Genotype inference analysis

Ruth Gómez Graciani

In this report, I put together IMPUTE2 results, genotypes inferred from tag SNPs and a coverage analysis to elaborte a definitive dataset of inversions reliably genotyped in as much individuals as possible from the 20 originally available. Those inversions genotyped in more than 3 individuals with at least 1 heterozygous and 1 homozygous sample will be used in later analyses.

1 IMPUTE2 results

We had 111 inversions with imputability information. The imputability information was available for GLB, AFR and EUR populations, and each inversion was marked as "No_Imputable", "Tagged", "Imputable" or "No_Polymorphic" (example in Table 1). From this dataset, **83 inversions were autosomal and imputable** (i.e. "Tagged", "Imputable" or "No_Polymorphic" in at least 1 population).

Table 1: Example rows from the imputability information table

Inversion	GLB	AFR	EUR
HsInv0003	Tagged	Tagged	Tagged
HsInv0004	Tagged	Tagged	Tagged
HsInv0006	Tagged	Tagged	Tagged

Since we don't have have measurements about how well does IMPUTE2 predict inversion orientations in EAS, SAS and ALL (admixed) individuals, the imputation process was repeated for each inversion and individual multiple times using different reference panels: same population haplotypes when possible (as control), and the 500, 250 and 100 closest haplotypes from the general population. This way, general population results can be used to (1) double check control results in EAS and SAS populations, where we don't know how reliable is IMPUTE2, and (2) predict orientation when same-population control results are not available. SNPs within the inversion region + 500kb to each side were used.

The resulting table shows, for each inversion and individual, one Genotype and one Probability column for each reference panel that was used, marked as _con, _500, _250 and _100 to tell them apart (example in Table 2).

Table 2: Some example rows from IMPUTE2 results table. Only some of the Probability and Genotype columns are present in this example.

Inversion	Individual	Population	Genotype_con	Probability_con	Genotype_100	Probability_100	Genotype_250
HsInv0379	NC6	AFR	STD	1.000	STD	1.00	STD
HsInv0486	NC6	AFR	STD	0.998	STD	1.00	STD
HsInv0379	NC22	AFR	STD	1.000	STD	0.99	STD
HsInv0486	NC22	AFR	STD	0.998	STD	1.00	STD
HsInv0379	NC25	AFR	STD	1.000	STD	1.00	STD
HsInv0486	NC25	AFR	STD	0.998	STD	1.00	STD

Three columns were incorporated to this table, that help us decide how reliable each imputation result is.

- **Imputability:** This column contains the population-specific imputability information relevant to this inversion and individual, obtained from the imputability information table (example in Table 1). EUR and AFR individuals will have the information from their corresponding columns and SAS, ALL and EAS individuals from the GLB column.
- **Imp.min.probability:** For a predicted genotype to be considered we'll require a minimum probability value of 0.8. When there are Probability_con measurements, we'll only take those into account, because we trust same-population measurements more than whole population measurements. If this value is NULL due to a lack of samples in the reference panel, which happens in most ALL and SAS individuals but with some EAS individuals as well, all the remaining results (Probability_500, _250 and _100) will be required to meet the criteria. Thus, the Imp.min.quality (imputation minimum probability) column contains either the Probability_con value or the minimum value from Probability_500, _250 and _100 columns.
- **Imp.all.equal:** This column contains TRUE/FALSE values indicating the consistency of imputation results. All genotypes for EUR and AFR individuals contain automatically a TRUE value, because we will consider the same population control only. Other populations' genotypes have a TRUE value when all their predicted genotypes are equal. All 4 measurements are available for most EAS individuals, and many of the inversions did not have a Genotype_con result in SAS and ALL individuals.

As I mentioned before, one of the uses for imputation results that used a general-population reference is to predict inversion orientations when same-population results are not available. This way of genotyping the inversions is assuming that when the three general-population results have probabilities > 0.8 and predict the same genotype, they will be showing the same result a same-population reference would. Thus, if we observe any inversion to not follow this tendency, we should be careful with the prediction. I searched for inversions in AFR and EUR populations where all IMPUTE2 results had >0.8 Probability but where general-population results were different from the same-population result. Those inversions were: HsInv0052si, HsInv0191, HsInv0991, HsInv1075si, HsInv1222, HsInv1264, HsInv1402. In consequence, those results that in column "Imp.all.equal" were marked as TRUE (consistent results) **changed to NA in 17 inversion-individual pairs**.

Finally, **Imp.genotype** has the resulting genotype only for those inversions and individuals that are imputable, have imputation minimum probaility >= 0.8 and have TRUE or NA values in the Imp.all.equal column. The table with the new columns is stored in impute.genotypes_filtered.csv

2 Tag SNP genotyping results

We had 22 inversions with known perfect tag SNPs (LD = 1) in suitable populations. I made a sript that summarizes which are the tag SNP genotypes associated to each inversion orientation in the reference panel, and then uses this as a template to infer the inversion orientations in the sample individuals. The resulting table (example in Table 3) shows, for each inversion and individual:

- **TagSNP.existing:** how many tag SNPs we know
- **TagSNP.sequenced:** how many tag SNPs were actually sequenced in the individual VCFs
- **TagSNP.genotype:** the predominant predicted orientation
- **TagSNP.probability:** the percentage of sequenced SNPs that agreed with the orientation prediction.

Due to the looping strategy used by the program, some inversion-individual pairs can be repeated because they were compared against more than one population (e.g. GLB and EUR). Taking into account only results with TagSNP.probability >= 0.8, for each inversion and individual the predicted orientation was selected in order of priority: GLB prediction and then population-specific prediction if GLB not available. **366 out of 367 inversion-individual pairs had at least one valid result available**.

Table 3: Sample rows from the tagSNP check results

Inversion	Individual	TagSNP.existing	TagSNP.sequenced	TagSNP.genotype	TagSNP.probability
HsInv0061	NC1	1	1	INV	1
HsInv0061	NC2	1	1	INV	1
HsInv0061	NC3	1	1	INV	1
HsInv0061	NC11	1	1	INV	1

The table with each inversion-individual pair and population is stored in tag.genotypes_filtered_detail.csv. The global result for each inversion and individual is in tag.genotypes_filtered.csv.

3 Coverage check

For each region, I counted how many SNPs with a global MAF>=0.025 are there in the 1KGP VCFs and how many of them were sequenced in the sample individuals. The ratio between 1KGP SNPs (expected) and sequenced SNPs was used to detect regions and/or individuals with too low coverages, that we expect to give less reliable imputation results.

Most individuals and inversions have an acceptable proportion of the expected SNPs within the imputation region (between 60% and 80%) and 95% of inversion-individual pairs have values above 30% (Figure 1). Some inversions have mean relative coverages below 30% (Figure 2, and I confirmed that it is a general tendency of those regions, and not a specific individual having a generalized low relative coverage (Figure 3). Despite having significantly low relative coverages, some of these inversions have >500 SNPs sequenced, which could be enough to impute the inversion orientation.

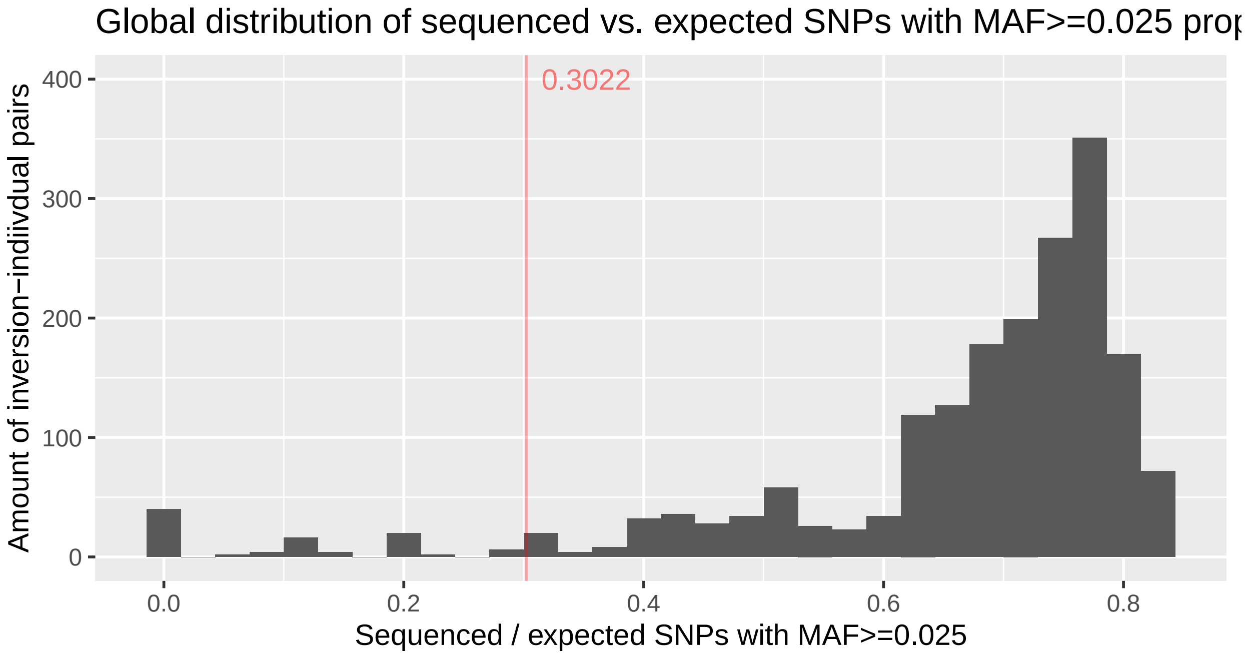

Figure 1: Histogram showing the distribution of realtive coverages in the whole dataset. The red line corresponds to the 95% quantile limit. Most inversion-individual pairs have sequenced vs. expected SNPs with MAF >=0.025 ratios between 0.6 and 0.8.

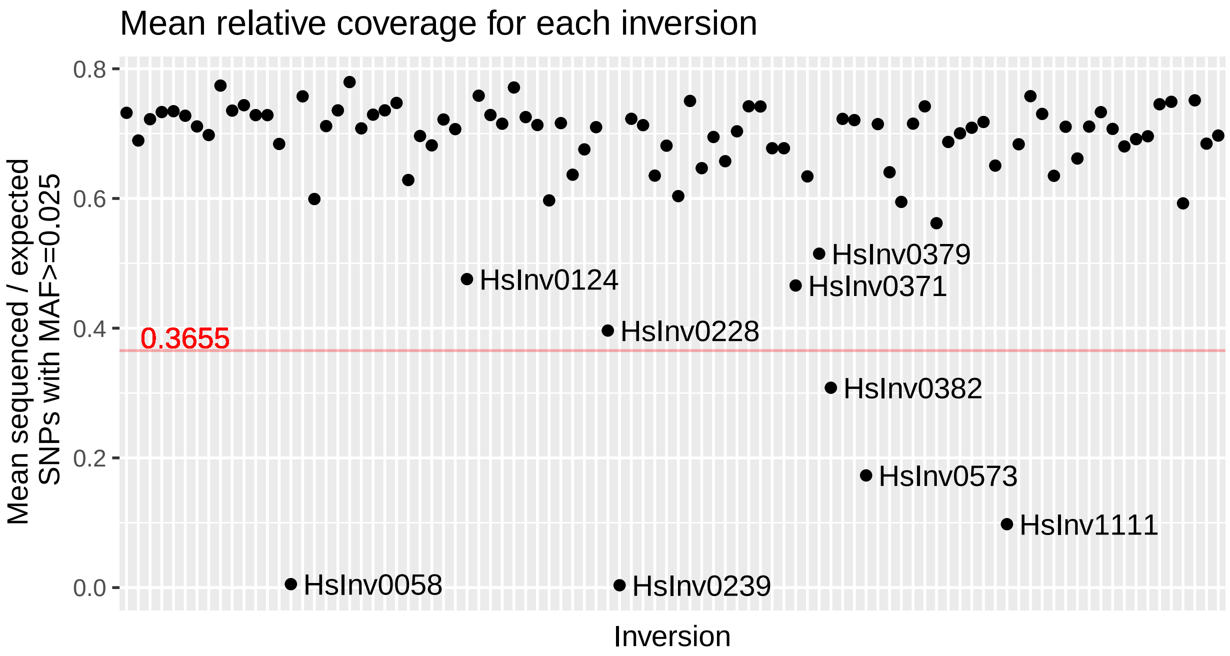

Figure 2: Each point is the mean sequenced / expected value for SNPs with MAF>=0.025 in a specific inversion region. Labels are shown for inversions with low relative coverages. Only those below the red line are significant.

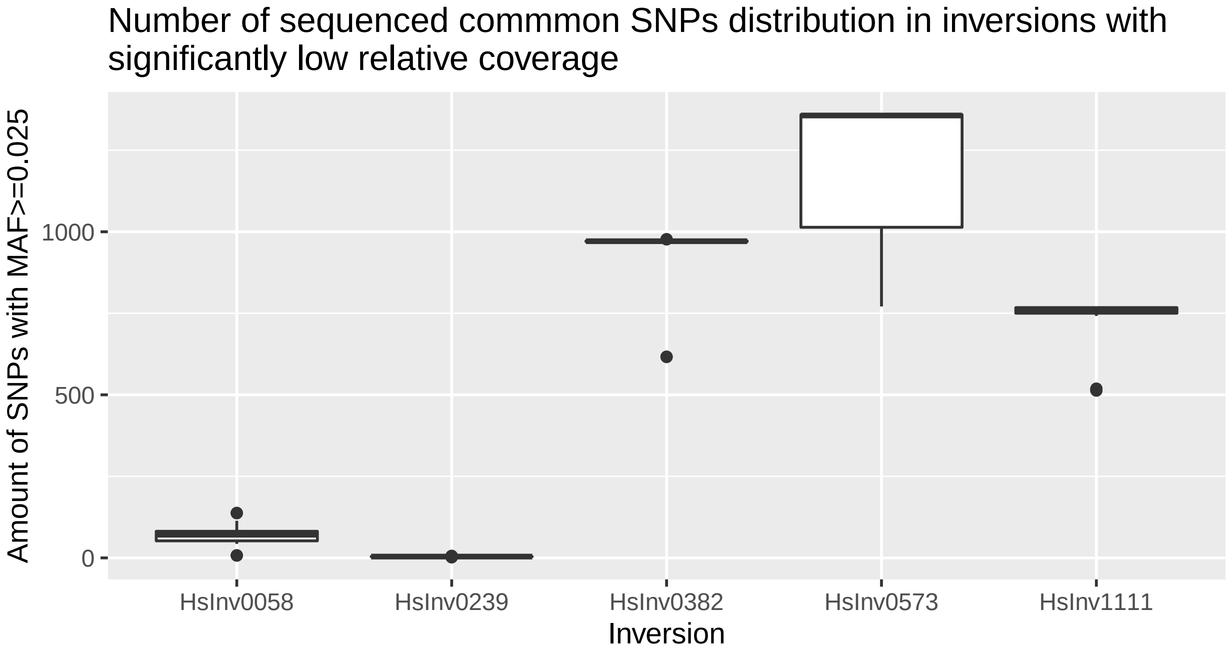

Figure 3: Box plots show the distribution of sequneced vs. expected ratio of SNPs with MAF>=0.025 in the 20 individuals for a given inversion. Only inversions with significantly low mean relative coverage are included. The significantly low relative coverage is not caused by a specific individual, it is rather a generalized tendency of each region.

4 Putting it all together

In this last part I create a table where we can see what happened with each inversion in each individual at a glance. Before explaining how inversions were classified, I need to define or revisit some key concepts:

- **Good imputation results:** those cases where Imp.min.quality >= 0.8 and Imp.all.equal is TRUE, that are imputable (i.e. “Tagged”, “Imputable” or “No_Polymorphic”) in this individual according to their population.
- **Good tagSNP results:** those cases where a tag SNP check result with TagSNP.probability >= 0.8 was found using either in GLB or same-population reference panels.
- **Enough coverage:** sequenced vs. existing SNPs with MAF >=0.025 proportion (Cov.seq.vs.exp_maf) is >=0.3 or more than 500 SNPs sequenced (Cov.sequenced_maf).

Knowing these conditions, we can find each inversion-individual pair classified as one of these categories:

- **Accepted (A)**
  - **A.Tagged**: Good tag SNP results with bad imputation results. Enough coverage.
  - **A.Imputed**:Good imputation results with bad or no tag SNP results. Enough coverage.
  - **A.Tag.Imput**: Good imputation and tagSNP results that match with imputation predictions. Enough coverage
- **To check (C)**
  - **C.Tag.Imput**: Good imputation and tag SNP results, not matching. Enough coverage.
  - **C.check.Imput**: Imputation results with NA in the Imp.all.equal column.
- **Rejected (R)**
  - **R.badImputability**: Not imputable.
  - **R.badCoverage**: Bad coverage.
  - **R.badImputation**: Bad imputation results without tagSNP information.

In some cases, a record could fit in more than one category (e.g. Not imputable and Bad Coverage), but I tried to apply them in an order that allows us to know the root reason for rejecting genotype prediction (in the example, Not imputable, because we wouldn’t be able to predict the orientation even with a Good Coverage).

Once the genotypes are filtered depending on the imputation, tagSNP and coverage results, they are aggregated and counted to filter inversions by the amount of different quality haplotypes available, introducing a new Rejected (R) category: **R.lowSample**, which will be assigned to previously Accepted (A) results whose inversions don’t have more than 3 individuals genotyped and both heterozygous and homozygous individuals available. The final table (example in Table 4) is stored in allgenotypes_classified.csv.

Table 4: Sample summary table for the results

Inversion	Individual	Population	Result	Imputability	Imp.min.quality	Imp.all.equal	Imp.genotype	TagSNP.sequenced	TagSNP.probability	TagSNP.population	TagSNP.genotype	Cov.seq.vs.exp_maf	Cov.sequenced_maf	HET.genotype	HOMO.genotype	all.genotype
HsInv0003	NC15	EUR	A.Tag.Imput	Tagged	1.000	TRUE	INV	22	1.0000000	GLB	INV	0.7726433	2926	7	13	20
HsInv0003	NC25	AFR	A.Tag.Imput	Tagged	1.000	TRUE	HET	22	1.0000000	GLB	HET	0.6926327	2623	7	13	20
HsInv0003	NC16	EUR	A.Tag.Imput	Tagged	1.000	TRUE	INV	22	1.0000000	GLB	INV	0.7721151	2924	7	13	20
HsInv0003	NC26	ALL	A.Tag.Imput	Tagged	1.000	TRUE	INV	22	0.9090909	GLB	INV	0.6902561	2614	7	13	20
HsInv0003	NC22	AFR	A.Tag.Imput	Tagged	1.000	TRUE	INV	22	0.9545455	GLB	INV	0.6910483	2617	7	13	20
HsInv0003	NC2	EUR	A.Tag.Imput	Tagged	1.000	TRUE	HET	22	1.0000000	GLB	HET	0.7713229	2921	7	13	20

In addition, a per-inversion summary table with the amount of correctly genotyped individuals and a column indicating whether the inversion was “Accepted”, “lowSample” or “unableToGenotype” is stored in inversions_summary.csv,

Figure 4 shows the incidence of each classification category. **1154 inversion-individual pairs from 62 inversions were finally accepted for analysis**. 22 inversions were genotyped but not in enough heterozygous or total individuals to be analyzed, and 14 couldn’t be genotyped at all.

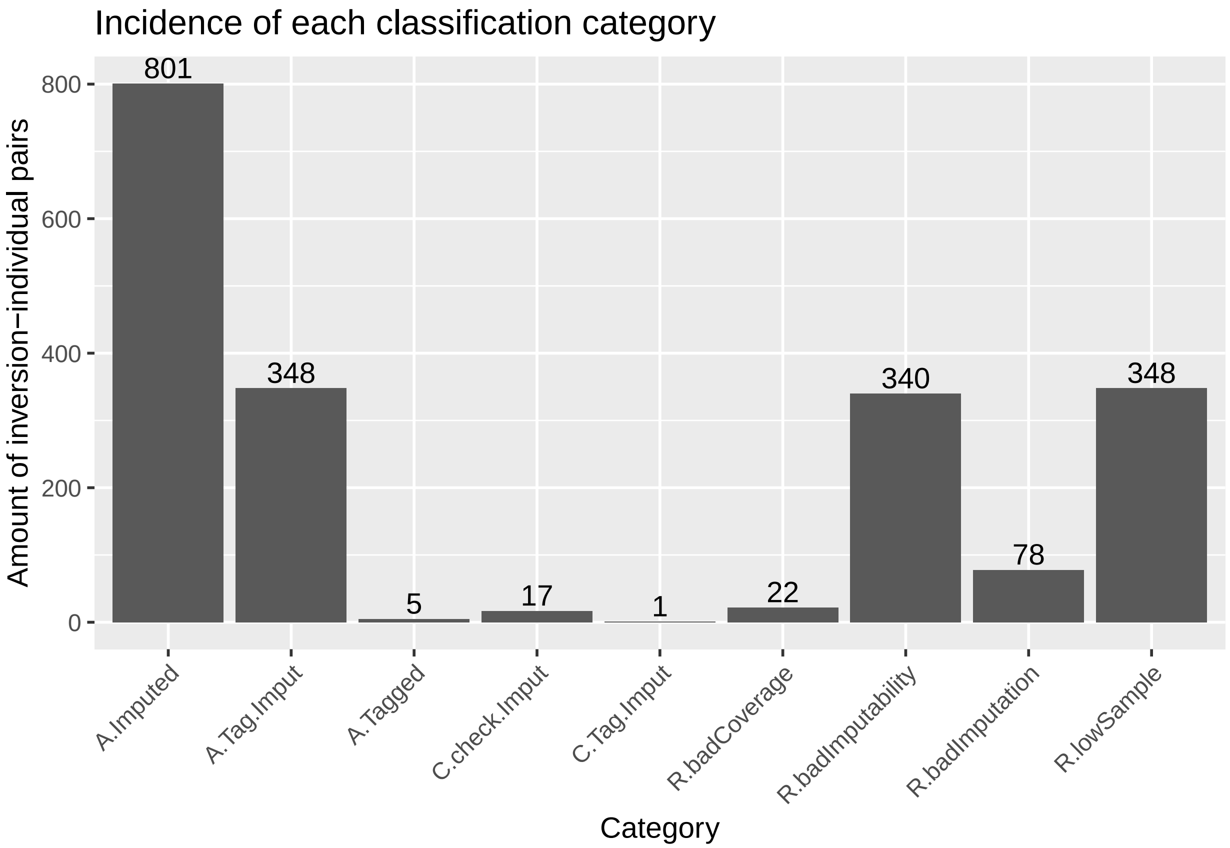

Figure 4: Final status of each inversion-individual pair. In most cases the imputation worked well, and almost half of the imputation results are supported by a tag SNP check. 5 pairs were recovered thanks to the tag SNP check and 18 cases have to yet be revised and are potentially useful. The main reasons for sample loss are the inversion not being imputable in the individual’s population and inversions not having enough samples to compare heterozygous vs homozygous tendencies.