Genotype inference analysis

Version 2

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

Note about individuals' populations: As explained in previous reports and in the wiki, we had 20 individuals from different populations. Some individuals had confusing or unespecific ancestry information, so I did a PCA to confirm their origin. The final distribution was 12 EUR, 3 AFR, 3 EAS, 1 SAS, 1 Admixed.

IMPUTE2 infers unknown genotypes for variants in a sample using a set of reference haplotypes from genotyped individuals. As a general rule, the more reference individuals, the better is the imputation precision. On the other hand, we must be careful when imputing recurrent inversions: haplotypes from opposite orientations could be more similar than normal and differences in recurrence rates between populations can be misleading for the program.

We have a list of 109 inversions that were classified according to their imputability (i.e. IMPUTE2 success rate in inferring the same inversion orientations as in experimental genotype validations in a set of individuals). The imputability information was available for AFR and EUR populations and we had a global (GLB) estimation as well. We selected for the analysis only those that were autosomal and imputable in at least one population (82 inversions).

We imputed each inversion in those individuals classified as "Tagged", "Imputable" or "No_Polymorphic" in their respective population (EAS, SAS and the admixed individual were assessed according to the GLB column). The reference sample size was a maximum of 500 haplotypes. When we had more than 500 reference haplotypes available, IMPUTE2 automatically selected them sorted by similarity to the individual to genotype. The origin of references varied depending on the inversions' recurrence: from the global dataset for unique inversions and from a same-population dataset for recurrent inversions.

The result table (example in Table 1, shows for each inversion and individual (inversion-individual pair), the Genotype predicted by IMPUTE2 and a Probability score indicating how reliable is the result. 28 inversion-individual pairs, mostly from SAS and Admixed individuals, did not have enough reference samples to be imputed.

Table 1: Sample rows from the imputation results

Inversion	Individual	Population	Imputability	Origin	Probability	Genotype
HsInv0003	NC6	AFR	Tagged	Unique	0.951	HET
HsInv0003	NC3	EUR	Tagged	Unique	1.000	INV
HsInv0003	NC27	SAS	Tagged	Unique	1.000	INV
HsInv0003	NC14	EUR	Tagged	Unique	1.000	HET

We established a minimum threshold of 90% Probability for a result to be accepted, which is the highest we can go without losing too much information (Figure 1).

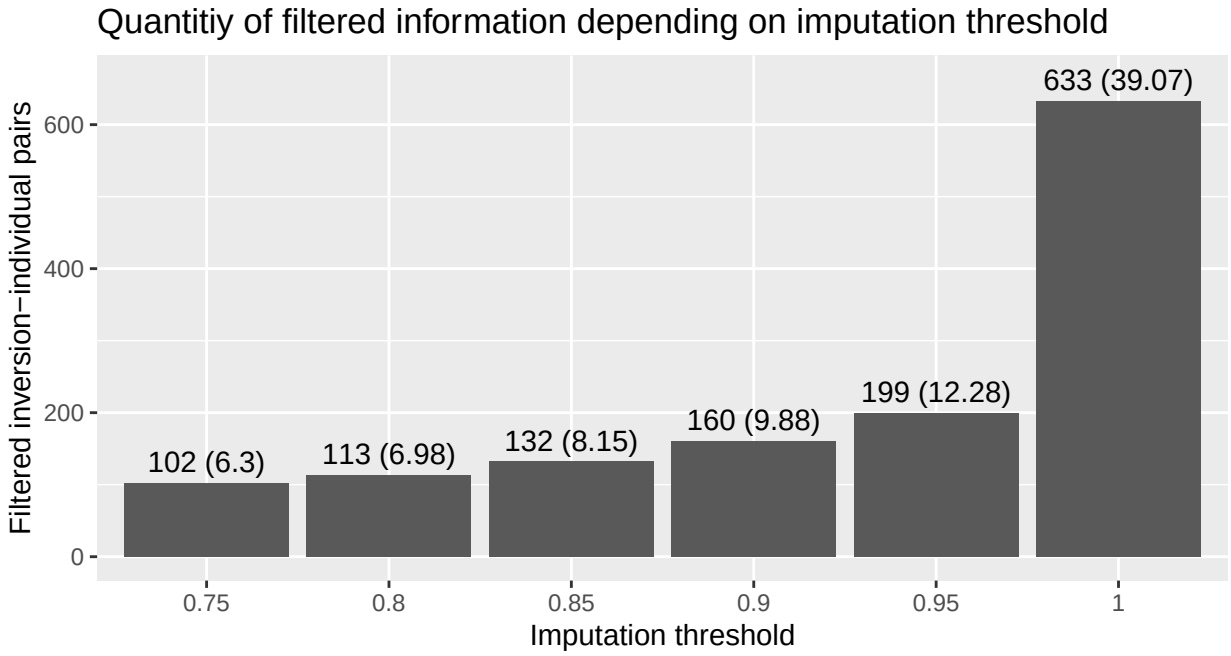


Figure 1: This plot shows how many inversion-individual pairs will be filtered depending on the threshold applied to imputation results. From the 6 thresholds here proposed, 0.9 is the highest one that leaves out less than 10% of the data

2 Tag SNP genotyping results

We had 22 inversions with known perfect tag SNPs (LD = 1) in suitable populations. I made a script 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 2) 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 pair 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 2: 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 2). Some inversions have mean relative coverages below 30% (Figure 3, and I confirmed that it is a general tendency of those regions, and not a specific individual having a generalized low relative coverage (Figure 4). Despite having significantly low relative coverages, some of these inversions have >500 SNPs sequenced, which could be enough to impute the inversion orientation.

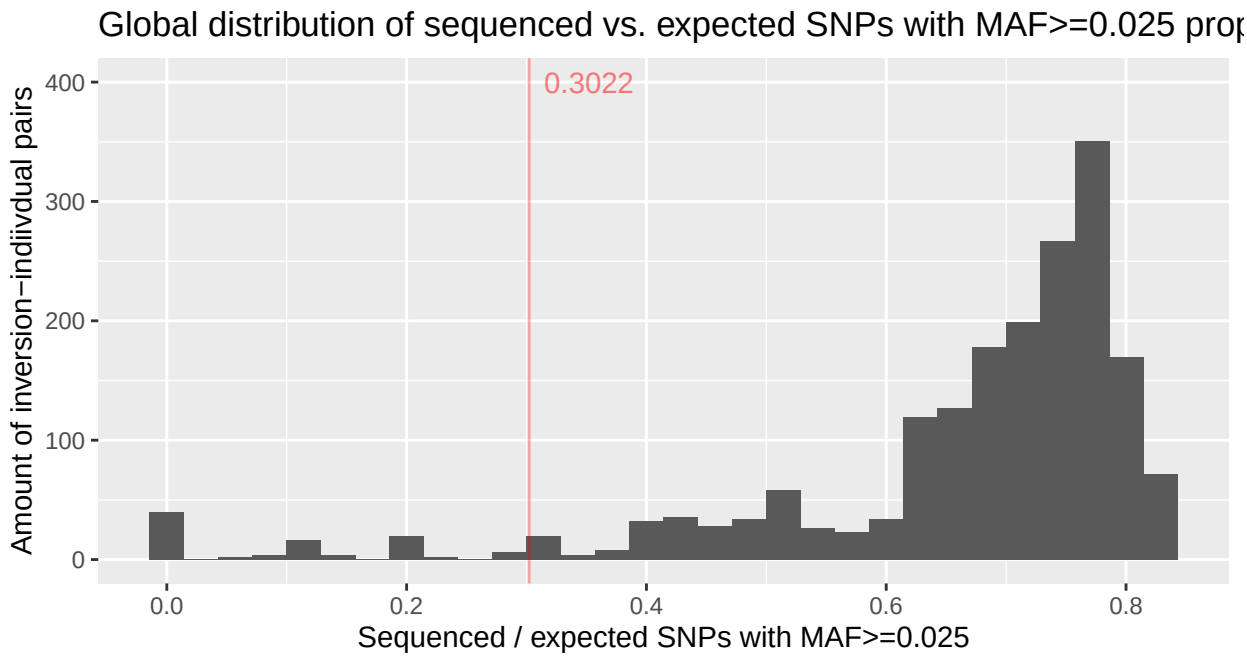


Figure 2: Histogram showing the distribution of realtive coverages in the whole dataset. The red line corresponds to the 95% quantile limit. Most inversion-individual pairs have sequenced vs. expected SNPs with MAF >=0.025 ratios between 0.6 and 0.8.

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 that are imputable according to Jon's previous studies (i.e. "Tagged", "Imputable" or "No_Polymorphic") and the probability value of the predicted genotype is >= 0.9.
- **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.Tag.Imput**: Good imputation and tagSNP results that match with imputation predictions. Enough coverage

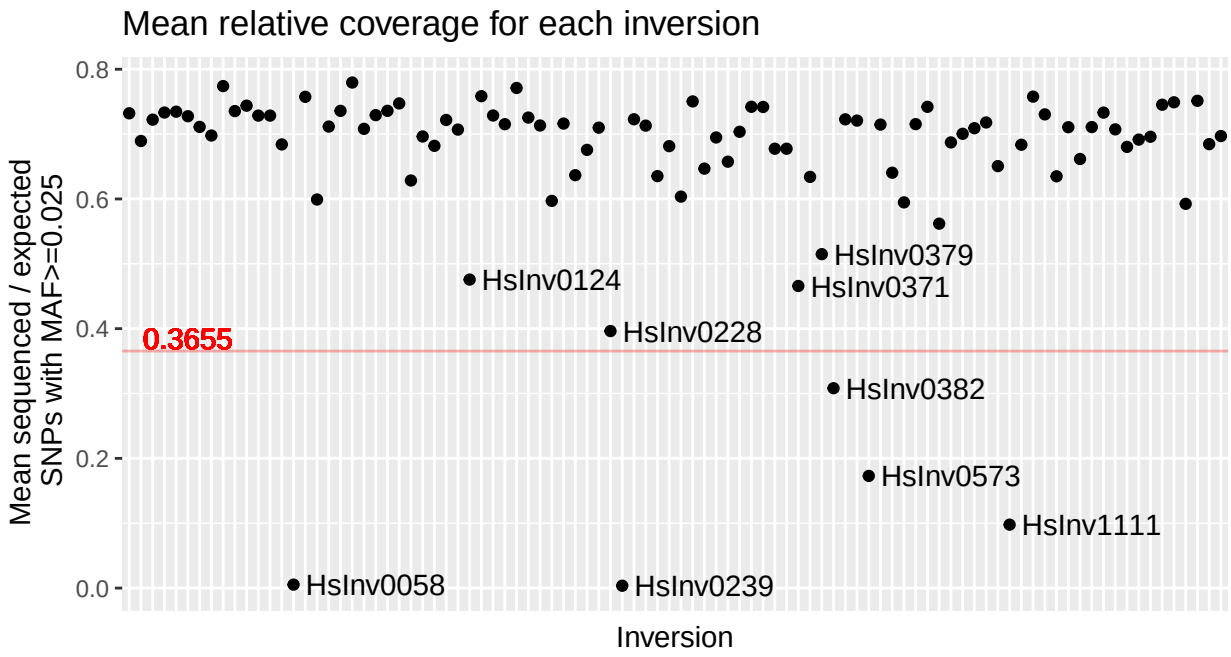


Figure 3: Each point is the mean sequenced / expected value for SNPs with MAF>=0.025 in a specific inversion region. Labels are shown for inversions with low relative coverages. Only those below the red line are significant.

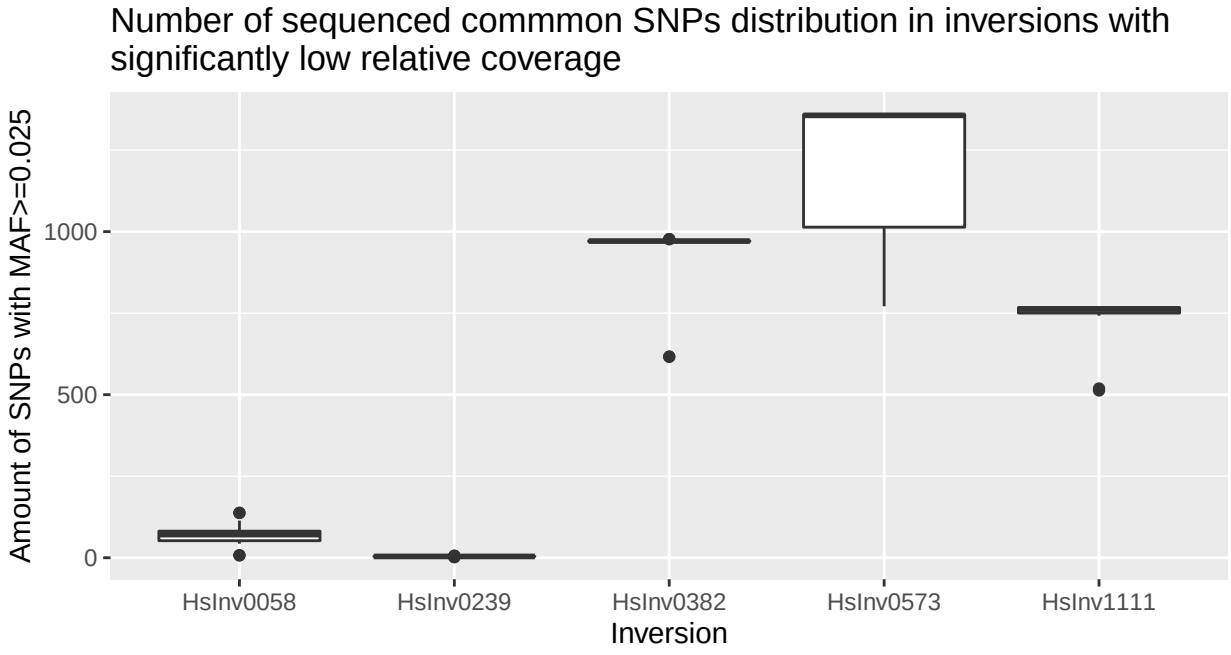


Figure 4: Box plots show the distribution of sequneced vs. expected ratio of SNPs with MAF>=0.025 in the 20 individuals for a given inversion. Only inversions with significantly low mean relative coverage are included. The significantly low relative coverage is not caused by a specific individual, it is rather a generalized tendency of each region.

  - **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.
- **To check (C)**
  - **C.Tag.Imput**: Good imputation and tag SNP results, not matching. Enough coverage.
- **Rejected (R)**
  - **R.badImputation**: Bad imputation results without tagSNP information.
  - **R.badCoverage**: Bad coverage.
  - **R.badImputability**: Not imputable.

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 3) is stored in allgenotypes_classified.csv.

Table 3: Sample summary table for the results

Inversion	Individual	Population	Result	Imputability	Origin	Imp.probability	Imp.genotype	TagSNP.sequenced	TagSNP.probability
HsInv0003	NC15	EUR	A.Tag.Imput	Tagged	Unique	1	INV	22	1.0000000
HsInv0003	NC25	AFR	A.Tag.Imput	Tagged	Unique	1	HET	22	1.0000000
HsInv0003	NC16	EUR	A.Tag.Imput	Tagged	Unique	1	INV	22	1.0000000
HsInv0003	NC26	ALL	A.Tag.Imput	Tagged	Unique	1	INV	22	0.9090909
HsInv0003	NC22	AFR	A.Tag.Imput	Tagged	Unique	1	INV	22	0.9545455
HsInv0003	NC2	EUR	A.Tag.Imput	Tagged	Unique	1	HET	22	1.0000000

TagSNP.population	TagSNP.genotype	Cov.seq.vs.exp_maf	Cov.sequenced_maf	HET.genotype	HOMO.genotype	all.genotype
GLB	INV	0.7726433	2926	7	13	20
GLB	HET	0.6926327	2623	7	13	20
GLB	INV	0.7721151	2924	7	13	20
GLB	INV	0.6902561	2614	7	13	20
GLB	INV	0.6910483	2617	7	13	20
GLB	HET	0.7713229	2921	7	13	20

In addition, a per-inversion summary table stored in inversions_summary.csv show how are the 20 individual results distributed regarding this inversion before applying the "R.lowSample" tag, the amount of correctly heterozygous, homozygous and total genotyped individuals, and a column indicating whether the inversion was "Accepted", "lowSample" or "unableToGenotype".

Figure 5 shows the incidence of each classification category. **1136 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.

```
invSummary<-melt(invSummary, id.vars =  c("Inversion", "HET.genotype", "HOMO.genotype", "all.genotype",

rejectList<-invSummary[invSummary$variable == "R.badImputability" & invSummary$value == 20,"Inversion"]

invSummaryPlot<-invSummary[!(invSummary$Inversion %in% rejectList),]

sTab<-merge(aggregate(value    ~ Inversion, invSummaryPlot, max), invSummaryPlot )

invSummaryPlot$Inversion<-factor(invSummaryPlot$Inversion, levels =  sTab[order(sTab$variable, sTab$valu
```

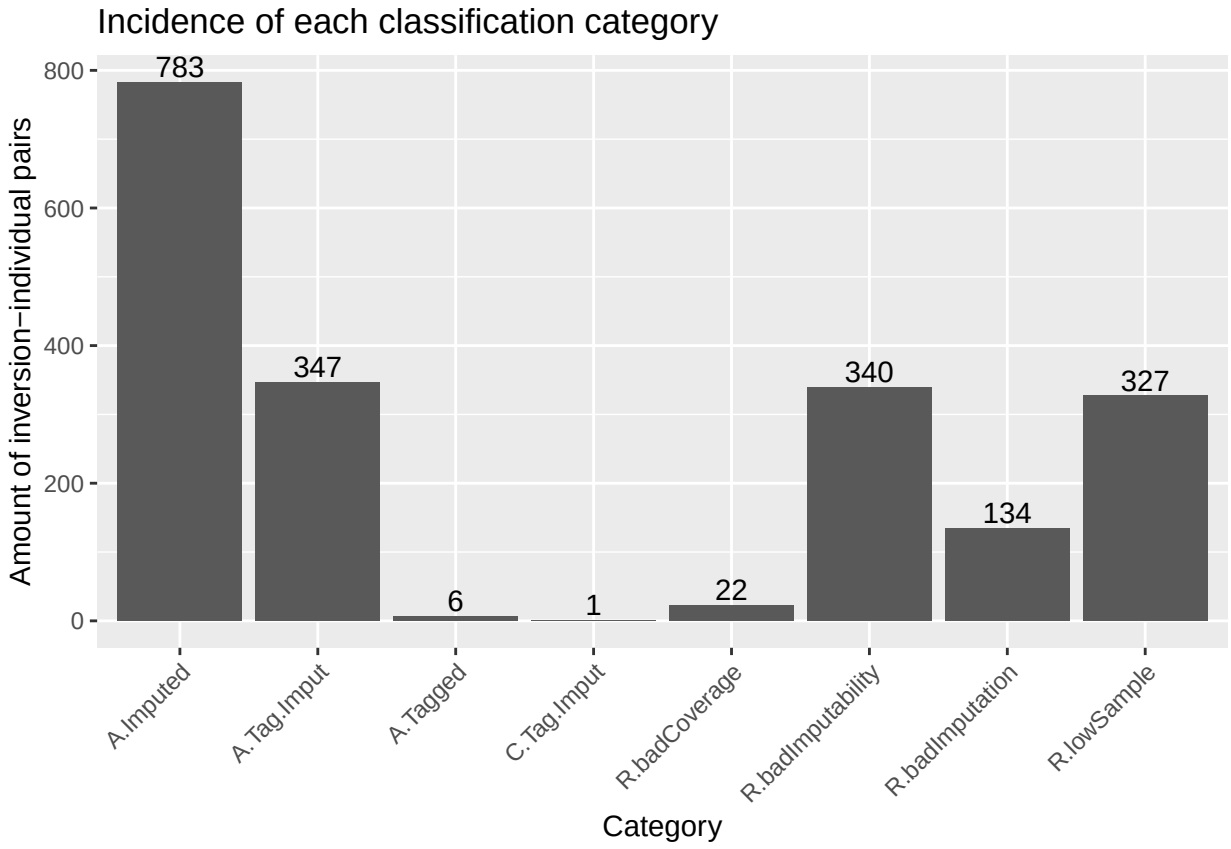
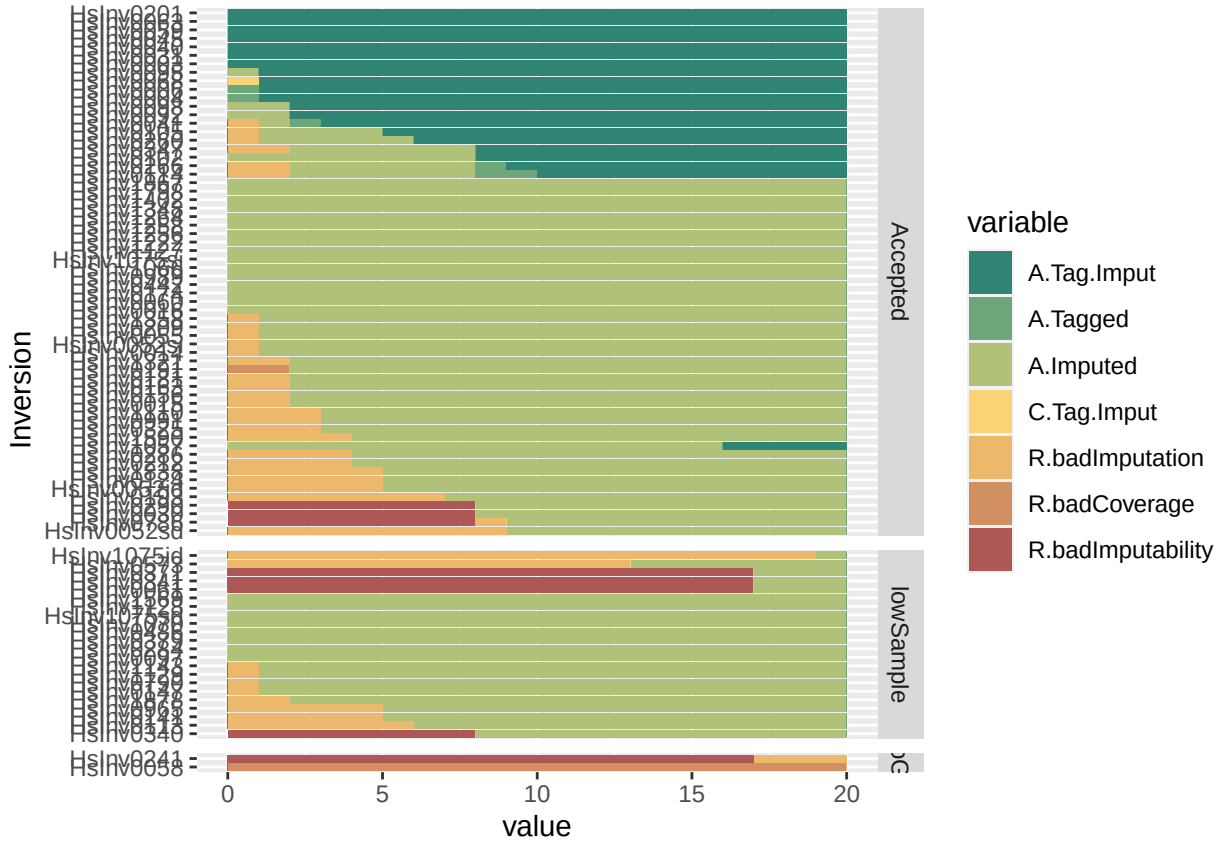


Figure 5: Final status of each inversion-individual pair. In most cases the imputation worked well, and almost half of the imputation results are supported by a tag SNP check. 5 pairs were recovered thanks to the tag SNP check and 18 cases have to yet be revised and are potentially useful. The main reasons 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.

```r
# blues<-colorRampPalette(c("#0B9885", "#A3EA99"))
# oranges<-colorRampPalette(c("#F3EB90",  "#dd6e42" ))
colors<-colorRampPalette(c("#308474", "#8cb87d","#f9d476",  "#e5ab68","#ad5757"))

invSummaryPlot$variable<-factor(invSummaryPlot$variable, levels = c("A.Tag.Imput", "A.Tagged", "A.Imputed", "C.Tag.Imput", "R.badImputation", "R.badCoverage", "R.badImputability"))

ggplot(invSummaryPlot)+geom_bar(aes(x = Inversion, y =value, fill = variable), stat="identity")+coord_flip()+
  # scale_fill_brewer(palette = "Spectral", direction = -1)
  scale_fill_manual(values = c(colors(7)))
```



5 Future directions

- An imputability assessment in EAS population will allow us to process the correspongind individuals as we did with EUR and AFR.