# Imputation report

## Context

Our goal is to quantify the effect of inversions in meiotic recombination, both in heterozygosis and homozygosis. We hypothesize that inverted regions may fail to synapse in heterozygous individuals and that homozygous individuals may have different distribution of crossovers depending on the orientation they carry. We have data from Bell et al. 2019, which sequence the genomes of 31,228 gametes from 20 sperm donors, identifying 813,122 crossover events. The first step in this project is to infer these individuals' genotypes of as many inversions as possible.

## Methods

Inversion genotypes were imputed with IMPUTE2, a genotype imputation and haplotype phasing program. Using a set of reference haplotypes genotyped for the inversions of interest and a genetic map, it infers the inversion status for our set of study individuals.

### Study panel preparation

We were provided with the VCFs for the 20 individuals in Bell et al. (2019) study. They were sequenced with a method called Sperm-seq: many individual sperm cells (974-2,274 gametes per donor) are sequenced to low coverage (median of ~1% of the haploid genome for each cell) simultaneously. Then, allelic haplotypes for the full length of every chromosome were inferred with ~40x coverage per donor.

The population of origin for each individual was provided by the sperm bank, but there were cases with conflicting ancestry information so I checked their ancestry with a Principal Component Analysis. The individuals from the 1000 Genomes Project and our study individuals were clustered using chromosome 1 with `qctools pca`. Then, our study individuals were classified according to the 1000 Genomes Project population cluster they fit into. **Figure 1** shows an example of the PCA result with some individuals. The final ancestry classifications are in **Table 1**.

A main condition in IMPUTE2 is that variants must be mapped against the same genome build in the reference and test panels. In addition, it is important that all shared SNPs between reference and test panels are aligned to the same allele coding. Our reference panel is in hg39 and all variants are aligned to the '+' strand, while the provided VCFs were in hg38 and the strand alignment changed in some regions. To be sure that shared variants had exactly the same coordinates and strand alignment in both panels, I copied the 1000 Genomes Project Phase 3 coordinates into the test VCFs using the variant ID as a reference and created a file showing their strand orientation, which can be used with the IMPUTE2 option `-strand_g` to account for strand alignment changes along the sequence.

### Reference panel preparation

Experimental genotypes for 111 inversions were available, 92 of which were autosomic. For each imputed inversion, a reference panel was created by merging these genotypes with the variants from the 1000 Genomes Project Phase 3 at 500 kb at each side of the inversion for the individuals in common.

Due to local differences in inversion recurrence I was advised to use as a reference panel only those individuals with the same ancestry as the individuals in the test panel. However, some of my individuals were admixed. With them, an alternative methodology can be used in which the reference panel icludes all the available individuals regardless of their ancestry, but IMPUTE2 uses only the closest n individuals to the test panel.

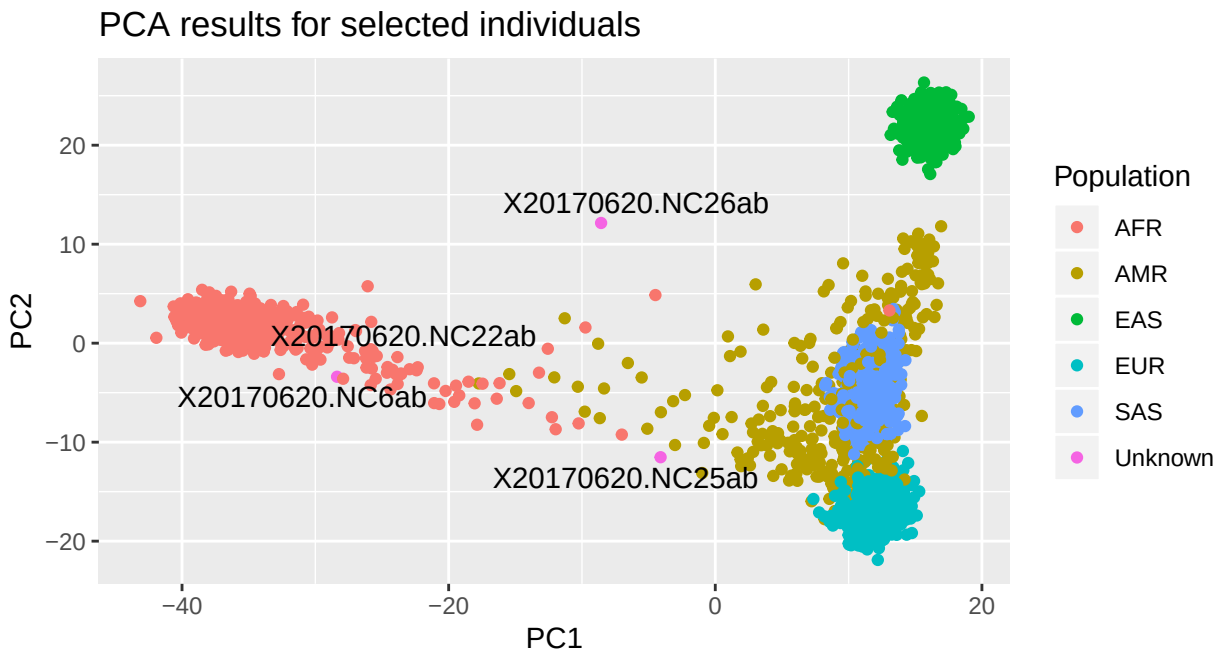


Figure 1: Principal Component Analysis for individuals NC6ab, NC22ab, NC25ab and NC26ab. Individuals NC6ab, NC22ab and NC25ab were reported by the sperm bank to be African American, while individual NC26ab was reported to be African American Asian. According to this analysis, individuals NC6ab and NC22ab were classified as AFR and individuals NC25ab and NC26ab as admixed.

Table 1: Principal Component Analysis results.

| Individual | Reported in original paper | PCA result |
|---|---|---|
| nc1 | European | EUR |
| nc2 | European | EUR |
| nc3 | European | EUR |
| nc4 | European | EUR |
| NC6ab | African American | AFR |
| nc8 | Asian | EAS |
| nc9 | Asian | EAS |
| nc10 | Asian | EAS |
| nc11 | European | EUR |
| nc12 | European | EUR |
| nc13 | European | EUR |
| nc14 | European | EUR |
| nc15 | European | EUR |
| nc16 | European | EUR |
| nc17 | European | EUR |
| nc18 | European | EUR |
| NC22ab | African American | AFR |
| NC25ab | African American | Admixed |
| NC26ab | African American Asian | Admixed |
| nc27 | Asian (conflicting ancestry information) | SAS |

Table 2: Genotype file modifications

| Original genotype | HsInv0052si and HsInv0370si | HsInv0052nd and HsInv0370nd |
|---|---|---|
| STD/STD | STD | STD |
| STD/INV | HET | STD |
| INV/INV | INV | STD |
| STD/DEL | | HET |
| INV/DEL | | HET |
| DEL/DEL | | DEL |

To test which is the most suitable threshold, I imputed all the inversions in all the individuals in four ways: using all individuals from the same population as those in test panel (the first methodology, which can be considered a control) and using all individuals regardless of their ancestry but making IMPUTE2 to select the 500, 250 or 100 closest individuals to those in test panel (the second methodology with relaxed, medium and strict conditions, respectively).

**Inversions with deletion allele**

Inversions HsInv0052 and HsInv0370 have three alleles: standard (STD), inverted (INV) and a deletion (DEL) that spans the inversion region. In the IMPUTE2 documentation, they explain that two or more variants can have the same position, for example a SNP and an INDEL. Following this example, I have divided these inversions into two variants. HsInv0052si and HsInv0370si refer to the standard and inverted alleles and HsInv0052nd and HsInv0370nd refer to not-deleted and deleted alleles. The genotype files were modified accordingly (**Table 2**).

# Results

The output of IMPUTE2 is the probability for each genotype to be true. For each inversion and individual, the result with the highest probability has been included in a table and processed with R. However, only genotypes with a probability higher than 80% can be considered reliable.

The results of imputation by individual and inversion are in the attached spreadsheet (sheet "Imputation results"). For each individual, inversions are tagged and sorted by imputability (with the categories "No_Polymorphic", "Tagged", "Imputable", "No_Imputable"). In addition, two columns summarize the reliability of the result by comparing the four methodological conditions that were tested:

- Uniformity: "good" when all four conditions report the same genotype, "alert" when all the conditions fot the second methodology predict the same genotype, which is different from that predicted by the first methodology and "bad" when conditions within the second methodology do not agree. Admixed individuals can only have "good" and "bad" Uniformity tags.
- Quality: "good" when all four conditions report probabilities >80% for their predicted genotypes, "alert" when some of the conditions report probabilities <80%, and "bad" when all conditions report <80% probabilites.

For HsInv0052 and HsInv0370, the results for both parts of the imputation were interpreted (column Notes). In short, the "nd" part is used to know how many DEL alleles are there and the "si" part to know the content of the not-deleted alleles, i.e "nd" = HET + "si" = STD is interpreted as STD/DEL. In the case of DEL/DEL genotype, I expect "nd" to be DEL with a good Uniformity and Quality and "si" to have not so good Uniformity or Qualities, unless the DEL haplotype is similar enough to STD or INV haplotypes.

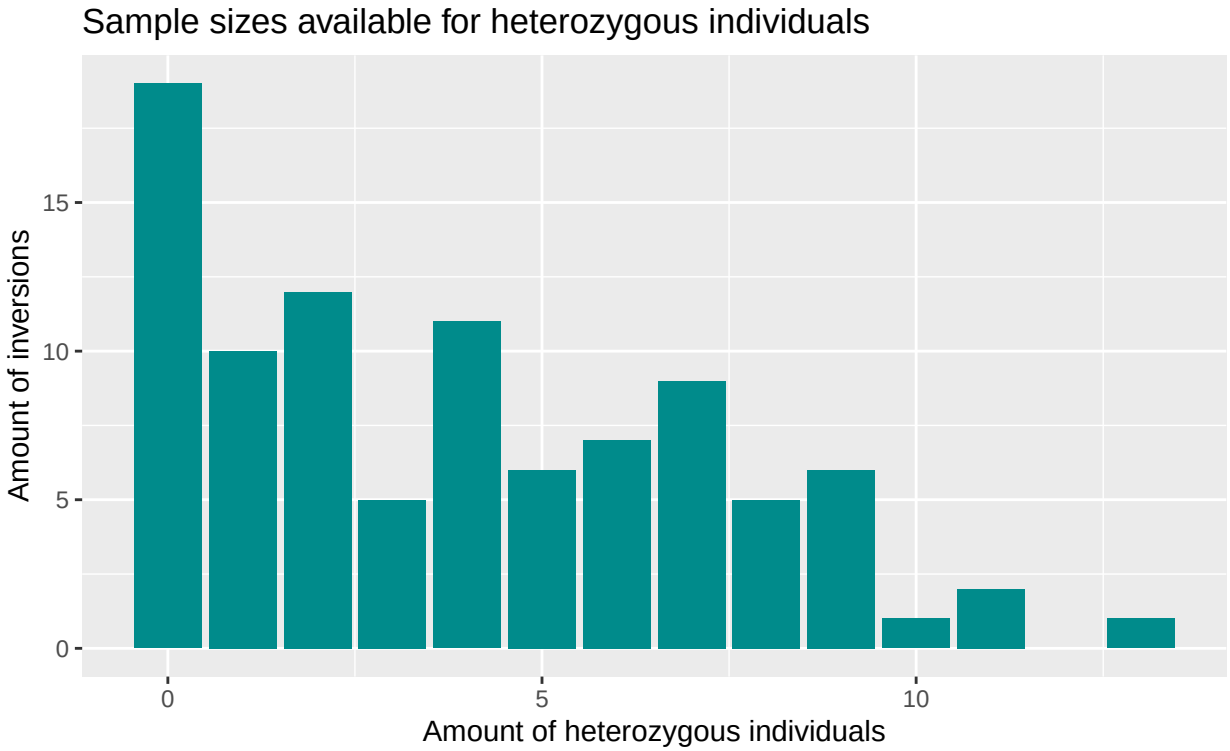


Figure 2: Distribution of sample sizes for valid heterozygous genotypes. ~20% of inversions do not have any heterozygous individuals. Most of the inversions have 9 or less valid heterozygous samples.

Finally, the Uniformity and Quality of the imputation results was summarized by inversion and population and can be found in the attached spreadsheet (sheet "Summary by inversion").

## Discussion

According to these preliminary results, and taking into account only genotypes that can be considered correctly predicted (good Uniformity and Quality), 74 inversions (~80%) have heterozygous individuals. In addition, for 33 inversions (~35%) heterozygous and homozygous individuals for both orientations can be compared. The sample size for heterozygous individuals for each inversion is maximum 13 (**Figure 2**).

In general, imputation results are considered valid when they have good Uniformity and Quality. There are several ways in which the amount of valid results can be increased:

- Less strict filters: results with good Uniformity and probability >80% in their corresponding population, with poor performance when imputed using all individuals but good performance in their corresponding population or those with probability <80% with their own population, but excellent performance when using all the individuals as reference could also be accepted, depending on the inversion imputability.
- In many cases, there are not enough SAS individuals genotyped to be used as a reference panel for individual nc27. This one could be filtered and interpreted as if it was admixed.
- Inversion HsInv0058 showed very poor performance, which was surprising considering it has tag SNPs. The problem was that none of the SNPs around the region in the test VCFs is a tag SNP for HsInv0058. This problem can be addressed using breakseq to genotype it. I'm in process of downloading part of the data available in dbGAP to check the format and how raw is the data, which will determine how time-consuming this solution would be.