Testing the performance of the imputation of MHC region in large datasets when using different reference panels

Locatelli E.[1], Treccani M.[1], Patuzzo C. [1], Veschetti L. [1], De Tomi E.[1], Dagnogo D.[1], Gallinaro M. [1], Stefani C.[1], Zipeto D. [1], Tamburin S.[1], Malerba G.[1]

[1] Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy

https://www.sites.google.com/view/gmlab

elena.locatelli@univr.it

INTRODUCTION

The major histocompatibility complex (MHC) contains a group of genes (~260 genes in ~4Mb) involved in several inflammatory disorders and immune response including the HLA-C gene. So far, the IPD-IMGT/HLA database reports more than 4000 different HLA-C alleles. Given the highly polymorphic nature of the gene, GWAS generally don't study or study only a small subset of polymorphic sites of the region. Imputation procedures may help in gaining additional information on this region. However, the successful imputation of the MHC region would require a reference panel with detailed information.

AIM OF THE PROJECT

The main goal of this study is to investigate whether imputation procedures using appropriate reference panels may effectively increase the number of polymorphic sites of the MHC region for association with complex traits.

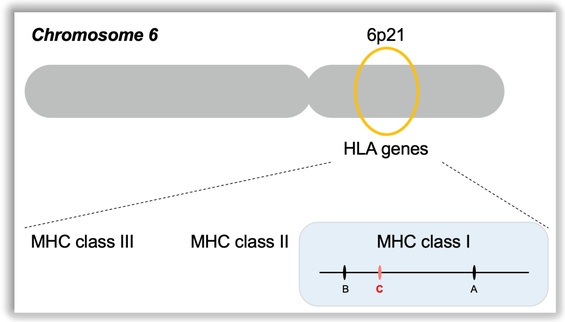


Figure 1: MHC complex scheme

METHODS

We studied the MHC region imputation performances using 3 different reference panels (Michigan and TOPMed imputation servers): TOPMed-r2, 1000 Genomes (Phase3, v5), and the novel four-digit multi-ethnic HLA panel (v1, 2021). Here, 5 datasets with more than 1000 individuals each genotyped by either Illumina or Affymetrix chip-arrays underwent imputation.
We then focused on the imputation results of the MHC region that surround the HLA-C gene (hg19: 31234948-31241032).

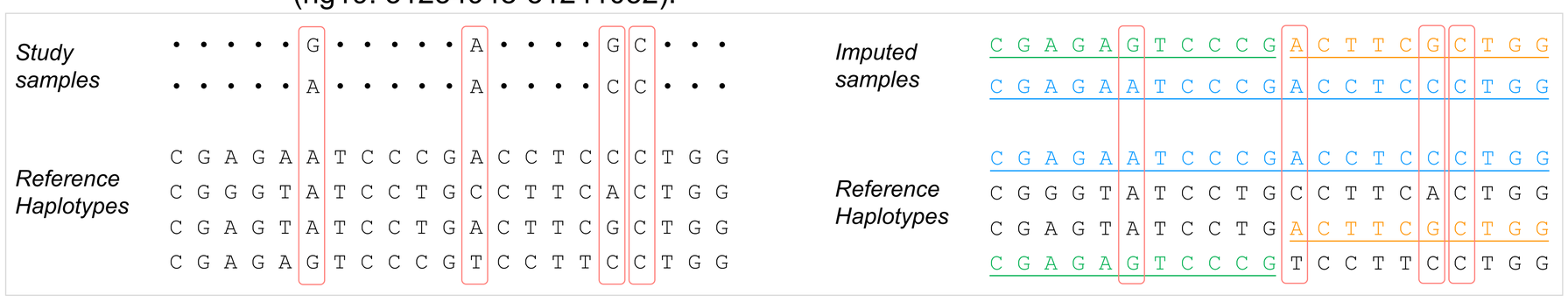


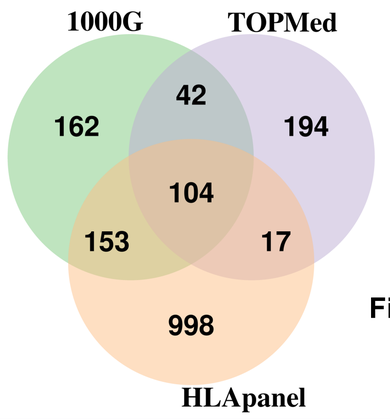
Figure 2: Imputation mechanism

RESULTS

Imputation reported a different number of markers for the different reference panels. HLA panels gave a higher number of imputed markers than the others (table 1).

Imputation panels	Number of imputed markers
1000G	482
TOPMed	365
HLA panel	1272

Table 1: Number of imputed markers for each reference panel.



104 markers were found to be in common between all three reference panels.

162 markers were found only by 1000G panel, 194 by TOPMed, and 998 by the HLA-panel (figure 3).

Figure 3: Reference panels comparison.

The 104 common markers showed high R2 values (>0.96). As expected, in the other marker groups, the R2 mean values were lower for markers with MAF<0.1 (>0.65 in 1000G, 0.15-0.20 in TOPMed, >0.40 in HLA panel).

Common markers					
1000G	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**
MAF < 0.1	R2 = 0.97 49 markers	R2 = 0.97 42 markers	R2 = 0.94 45 markers	R2 = 0.97 42 markers	R2 = 0.95 42 markers
0.1 < MAF ≤ 0.25	R2 = 0.99 38 markers	R2 = 0.98 45 markers	R2 = 0.98 41 markers	R2 = 0.98 44 markers	R2 = 0.96 44 markers
MAF > 0.25	R2 = 0.99 17 markers	R2 = 0.99 17 markers	R2 = 0.99 18 makers	R2 = 0.99 18 markers	R2 = 0.96 18 markers
TOPMed	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**
MAF < 0.1	R2 = 0.95 50 markers	R2 = 0.95 43 markers	R2 = 0.95 45 markers	R2 = 0.95 43 markers	R2 = 0.95 43 markers
0.1 < MAF ≤ 0.25	R2 = 0.99 35 markers	R2 = 0.99 42 markers	R2 = 0.99 39 markers	R2 = 0.99 42 markers	R2 = 0.99 42 markers
MAF > 0.25	R2 = 0.99 19 markers	R2 = 0.99 19 markers	R2 = 0.99 20 markers	R2 = 0.99 19 markers	R2 = 0.99 19 markers
HLA panel	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**
MAF < 0.1	R2 = 0.95 50 markers	R2 = 0.96 42 markers	R2 = 0.97 45 markers	R2 = 0.97 42 markers	R2 = 0.94 42 markers
0.1 < MAF ≤ 0.25	R2 = 0.99 35 markers	R2 = 0.99 43 markers	R2 = 0.99 39 markers	R2 = 0.99 43 markers	R2 = 0.99 43 markers
MAF > 0.25	R2 = 0.99 19 markers	R2 = 0.99 19 markers	R2 = 0.99 20 markers	R2 = 0.99 19 markers	R2 = 0.99 19 markers

Table 2: Number of markers and averaged R-squared values according to different MAF ranges among the 5 datasets.

The efficiency of the imputation was measured by the R-squared (R2) values stratifying the markers into 3 groups according to the minor allele frequency (MAF), table2.
The figure 4 does not show any strong difference in the distribution of the imputed markers when using each of the three different imputation panels.

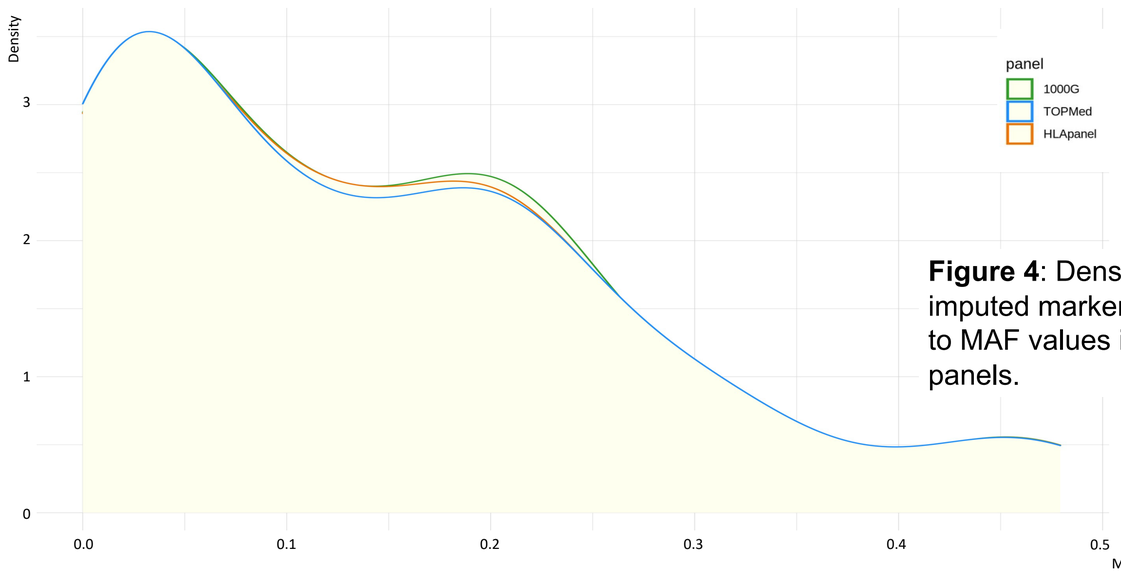


Figure 4: Density of the imputed markers according to MAF values in the three panels.

As shown in table 3, the absolute number of imputed markers was higher when using the HLApanel.
The table 3 shows the distribution of markers according to R2. In the detail, we observed ~70%, ~35%, ~60% of markers with a R2 > 0.9 for the 1000G, TOPMed, and HLApanel respectively.

Table 3: number of markers for each dataset in different R2 ranges.

	0.3 ≤ R2 ≤ 0.6	0.60 ≤ R2 ≤ 0.80	0.80 ≤ R2 ≤ 0.90	R2 > 0.90
1000G				
Dataset 1	17	29	23	380
Dataset 2	10	29	24	372
Dataset 3	19	25	25	362
Dataset 4	9	33	18	380
Dataset 5	16	20	31	359
TOPMed				
Dataset 1	21	11	2	141
Dataset 2	7	12	3	130
Dataset 3	16	0	3	139
Dataset 4	17	2	1	133
Dataset 5	10	3	1	135
HLA panel				
Dataset 1	68	15	8	836
Dataset 2	13	20	18	830
Dataset 3	44	34	36	786
Dataset 4	3	18	30	848
Dataset 5	63	18	17	813

CONCLUSIONS

In general, the R-squared prediction seems to be good in all the three used reference panels.

TOPMed panel resulted not suitable for a detailed HLA region imputation.

The number of markers imputed by the HLA panel results to be extremely high (2.5 folds) than the others, suggesting its use as reference panel for the HLA study.

REFERENCES

1. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze S, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. Nature Genetics 48, 1284–1287 (2016)

2. Taliun, D., Harris, D.N., Kessler, M.D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590,** 290–299 (2021). https://doi.org/10.1038/s41586-021-03205-y

3. Christian Fuchsberger, Gonçalo R. Abecasis, David A. Hinds, minimac2: faster genotype imputation, *Bioinformatics*, Volume 31, Issue 5, 1 March 2015, Pages 782–784, https://doi.org/10.1093/bioinformatics/btu704

4. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet. 2013;14:301-23. doi: 10.1146/annurev-genom-091212-153455. Epub 2013 Jul 15. PMID: 23875801; PMCID: PMC4426292.