Imputation of classical HLA types from UK Biobank genotype data

Interim Data Release, March 2016

Details of the analysis

Imputation of four-digit HLA alleles from SNPs was carried out using HLA*IMP:02 (Dilthey et al., 2013) with modified settings to accommodate the addition of samples to the reference panel, while achieving feasible computational and memory requirements. The modified settings were: localization feature turned off (previously turned on at all loci except HLA-B and -DRB1); graph sampling error (mS) and graph building error (mB) probabilities both set to 0.001 (previously both were set to 0.002); and the number of sampled haplotype pairs, NS, set to 5 (previously 50).

The reference datasets used (described fully in Motyer et al. (in preparation, 2016)) are shown in Table 1.

Dataset	Description
CEU+58	Dataset composed of the British 1958 Birth Cohort, HapMap CEU individuals, and CEPH CEU+ additional individuals
GSK	Dataset provided by GlaxoSmithKline (also known as 'HLA_RES')
YRI	HapMap YRI individuals
1000G	1000 Genomes Project dataset
T1DGC	Type 1 Diabetes Genetics Consortium dataset
KC	African-American individuals provided by King's College, London, UK (unpublished).
SW	Swedish individuals provided by Karolinska Institutet, Sweden (unpublished).
PA	Pan-Asian dataset made available by Pillai et al (2014).

Table 1 Reference datasets

The reference datasets were typed on various SNP arrays and had QC applied to each individually. QC procedures for 'CEU+58', 'GSK' and 'YRI' are as described in Dilthey et al. (2013). QC for '1000G' was the standard HLA*IMP dataset QC with the HLA*IMP interface, using a 20% missing data threshold on SNPs and individuals. 'T1DGC' had QC performed as per Jia et al. (2013). 'KC' had Sequence Based Typing (SBT) of HLA performed at Oklahoma Medical Research Foundation and University of Alabama and SNP QC as per '1000G' with a missing data threshold of 5%. 'SW' had high-resolution HLA typing (SBT) and QC as for 'KC'. 'PA' had QC applied as per Pillai et al (2014).

In addition to QC applied individually to the data from each source, further QC steps were performed to enable combining these datasets into a single reference panel. This included: converting the SNPs to GRCh37 coordinates using 'liftOver'; converting to a common strand alignment using 'PLINK'; and excluding SNPs that could not be aligned or had very different allele frequencies between the datasets. As part of the merging process we excluded duplicate individuals (there was overlap between some of the datasets, e.g. CEU and 1000G). Because the extent of lab-based typing differed for each HLA locus, we created a separate version of the merged dataset for each locus in order to retain the maximal number of SNPs. Specifically, for each locus

we included only the individuals which had lab-based HLA types for that locus, and only the SNPs that were polymorphic and were typed in at least 98% of that set of individuals. We treated ambiguous HLA types, e.g. G and P coded alleles, as the most frequent allele.

At each HLA locus, the reference datasets that were merged and the number of individuals in the reference panel (total and by self-reported ancestry) are listed in the Table 2.

HLA imputation was performed using only the SNPs in the reference panel that were also typed on the UK Biobank Axiom array and within 1 Mb of the the HLA locus. The number of SNPs used for each HLA locus is listed in Table 2.

HLA locus	Reference datasets merged	Number of SNPs used	Number of reference individuals	Number of reference Europeans	Number of reference Africans/African Americans	Number of reference Asians	Number of reference Latinos
HLA-A	CEU+58,GSK,YRI,1000G,T1DGC	661	8,085	7,347	208	307	223
HLA-B	CEU+58,GSK,YRI,1000G,T1DGC	927	9,120	8,112	236	417	355
HLA-C	CEU+58,GSK,YRI,1000G,T1DGC	908	7,732	6,984	212	313	223
HLA-DRB1	CEU+58,GSK,YRI,1000G,T1DGC	626	8,869	7,896	226	403	344
HLA-DRB3	GSK,KC,SW	849	880	484	345	17	34
HLA-DRB4	GSK,KC,SW	849	865	467	346	20	32
HLA-DRB5	GSK,KC,SW	801	808	408	346	18	36
HLA-DQA1	CEU+58,GSK,YRI,T1DGC,PA	747	6,242	5,640	27	503	72
HLA-DQB1	CEU+58,GSK,YRI,1000G,T1DGC	623	8,491	7,676	217	335	263
HLA-DPA1	T1DGC,PA,SW	794	6,067	5,615	0	452	0
HLA-DPB1	GSK,T1DGC,PA,SW	691	6,176	5,687	0	463	26

Table 2 The number of SNPs used for each HLA locus and the number of individuals in the reference panel

We assessed accuracy of HLA*IMP:02 with these settings and reference panel by performing five-fold cross-validation. The estimate of 4-digit accuracy with HLA imputations called for every individual (i.e. a posterior probability call threshold of 0) are given for each of the major populations in Table 3.

HLA locus	Europeans	Africans/African American	Asians	Latinos
HLA-A	97.2	94.4	89.2	90.5
HLA-B	94	81.7	86.3	74.5
HLA-C	97.8	92.8	94.4	94.1
HLA-DRB1	93.9	87.9	87.6	82.4
HLA-DRB3	97.8	96.5	93.1	94.7
HLA-DRB4	97.7	98.4	100	100
HLA-DRB5	99.2	99.3	100	100
HLA-DQA1	98.4	94	94.8	81.6
HLA-DQB1	97.8	87.6	95.1	92.5
HLA-DPA1	99.5	-	98.8	-
HLA-DPB1	94.5	-	86.2	88.5

Table 3 Estimate of 4-digit accuracy (%) of HLA imputation with a posterior probability call threshold of 0

When interpreting these accuracy estimates it is important to note that 94% of the UK Biobank individuals are self-reported as White. Thus, in assembling the reference panel we have sought to optimise accuracy for Europeans by, at some HLA loci, not making use of all available non-European reference datasets. This was done in order to maximise the number of SNPs available for imputation.

Imputation accuracy is generally higher if calls are restricted to those imputations with a posterior probability greater than a specified threshold. Accuracy and call rate from five-fold cross-validation with a call threshold of 0.7 is given in Table 4.

HLA locus	Europeans	Africans/African American	Asians	Latinos
HLA-A	97.8/98.7	95.8/97.6	91.7/94.3	91.4/97.3
HLA-B	96.5/95.5	89.3/87.5	92.4/88.7	85.8/80.7
HLA-C	98.2/98.9	94.7/95.4	95.0/98.6	95.9/98.0
HLA-DRB1	96.3/95.3	90.8/92.8	92.6/91.0	89.9/85.9
HLA-DRB3	98.0/99.5	97.3/97.7	93.1/100.0	96.4/96.5
HLA-DRB4	98.2/98.6	98.8/98.4	100.0/96.8	100.0/100.0
HLA-DRB5	99.5/99.0	99.6/99.6	100.0/96.9	100.0/100.0
HLA-DQA1	98.8/99.1	97.9/94.0	96.2/98.0	82.9/97.4
HLA-DQB1	98.4/98.6	90.6/94.3	95.9/97.4	93.1/97.0
HLA-DPA1	99.6/99.7	-	99.0/99.6	-
HLA-DPB1	96.1/95.1	-	89.5/92.8	88.2/98.1

Table 4 Estimate of 4-digit accuracy (%) of HLA imputation/Call rate(%) with a posterior probability call threshold of 0.7.

Notes for researchers

The data is available as a table of HLA*IMP:02 imputations. These contain the most likely HLA types for each individual in each of the 11 imputed HLA loci. There is one row for each individual, with each row containing a single imputed HLA allele for each of the two chromosomes and the 11 loci. The columns are labelled with the locus name and chromosome number. For each genotype call one quality metric is reported, Q is the absolute posterior probability of the allele inference (Q2 from the HLA*IMP:02 output). We suggest using a Q threshold of 0.7, that is, setting to missing the imputations with Q below this value.

The imputations will need to be converted by the user to an appropriate file format for association analyses, e.g. PLINK, SNPTEST. The imputations can be converted to the required format with a tool such as R. One possible approach is to read the HLA*IMP:02 imputations into R, impose posterior thresholding at 0.7, then for each HLA locus create a marker representing the presence/absence of each HLA allele at the HLA locus which can be coded like a SNP.

For HLA-DRB3, -DRB4, and -DRB5 the allele '9901' indicates that no allele is present (n.b. these loci can have copy number 0).

Acknowledgements

The following people contributed to the analysis described in this document

Dr Allan Motyer, Dr Damjan Vukcevic, Dr Adrian Cortes, Prof Gil McVean, Dr Stephen Leslie

References

Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M.R., and McVean, G. (2013). Multi-Population Classical HLA Type Imputation. PLoS Comput Biol 9, e1002877.

Jia, X. et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. PLoS One 8, e64683 (2013).

Motyer, A., Vukcevic, D., Dilthey, A., Donnelly, P., McVean, G., and Leslie, S. (2016). An Evaluation of Methods for Imputation of HLA Alleles from SNP Genotype Data. (in preparation).

Pillai, N. E. et al. Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. Hum. Mol. Genet. 23, 4443–51 (2014).