OXFORD

BIOINFORMATICS ARTICLE

Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles

Frauke Degenhardt[1,†], Mareike Wendorff[1,†], Michael Wittig[1], Eva Ellinghaus[2], Lisa W. Datta[3], John Schembri[4], Siew C. Ng[5], Elisa Rosati[1], Matthias Hübenthal[1], David Ellinghaus[1], Eun Suk Jung[1,6], Wolfgang Lieb[7], Shifteh Abedian[8,9], Reza Malekzadeh[9], Jae Hee Cheon[6], Pierre Ellul[4], Ajit Sood[10], Vandana Midha[10,11], B.K. Thelma[12], Sunny H. Wong[5], Stefan Schreiber[1,13], Keiko Yamazaki[14,15], Michiaki Kubo[16], Gabrielle Boucher[17], John D. Rioux[17,18], Tobias L. Lenz[19], Steven R. Brant[3,20,21] and Andre Franke[1,*]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany, [2]K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo University Hospital, Rikshospitalet, 0424 Oslo, Norway, [3]Department of Medicine, Meyerhoff Inflammatory Bowel Disease Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, [4]Division of Gastroenterology, Mater Dei Hospital, Msida MSD 2090, Malta, [5]Department of Medicine and Therapeutics, Institute of Digestive Disease, LKS Institute of Health Science, State Key Laboratory of Digestive Disease, The Chinese University of Hong Kong, Hong Kong, China, [6]Department of Internal Medicine and Institute of Gastroenterology, Yonsei University College of Medicine, Seoul, 03722, Republic of Korea, [7]Biobank PopGen and Institute of Epidemiology, University Hospital Schleswig-Holstein, Campus Kiel, 24105 Kiel, Germany, [8]Department of Epidemiology, University Medical Center Groningen, 9700 RB Groningen, The Netherlands, [9]Digestive Disease Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, 14117-13135, Tehran, Iran, [10]Department of Gastroenterology, Dayanand Medical College and Hospital, 141001 Ludhiana, Punjab, India, [11]Department of Medicine, Dayanand Medical College and Hospital, 141001 Ludhiana, Punjab, India, [12]Department of Genetics, University of Delhi South Campus, 110021 New Delhi, India, [13]Department of Medicine, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany, [14]Laboratory for Genotyping Development, Center for Integrative Medical Sciences, RIKEN Yokohama Institute, Yokohama, 230-0045, Japan,

[15]Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, 173-8610, Japan, [16]RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan, [17]Montreal Heart Institute, Research Center, Montréal, Québec H1T 1C8, Canada, [18]Université de Montréal Department of Medicine, Montréal, Québec H3C 3J7, Canada, [19]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany, [20]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA, [21]Department of Medicine, Rutgers Robert Wood Johnson Medical School and Department of Genetics, Rutgers University, New Brunswick and Piscataway, NJ 08901, USA

*To whom correspondence should be addressed at: Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Rosalind-Franklin-Street 12, D-24105 Kiel, Germany. Tel: +49 (0) 431/500-15109; Fax: +49 (0) 431/500-15168; E-mail: a.franke@mucosa.de

Abstract

Genotype imputation of the human leukocyte antigen (HLA) region is a cost-effective means to infer classical HLA alleles from inexpensive and dense SNP array data. In the research setting, imputation helps avoid costs for wet lab-based HLA typing and thus renders association analyses of the HLA in large cohorts feasible. Yet, most HLA imputation reference panels target Caucasian ethnicities and multi-ethnic panels are scarce. We compiled a high-quality multi-ethnic reference panel based on genotypes measured with Illumina's Immunochip genotyping array and HLA types established using a high-resolution next generation sequencing approach. Our reference panel includes more than 1,300 samples from Germany, Malta, China, India, Iran, Japan and Korea and samples of African American ancestry for all classical HLA class I and II alleles including *HLA-DRB3/4/5*. Applying extensive cross-validation, we benchmarked the imputation using the HLA imputation tool HIBAG, our multi-ethnic reference and an independent, previously published data set compiled of subpopulations of the 1000 Genomes project. We achieved average imputation accuracies higher than 0.924 for the commonly studied *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* genes across all ethnicities. We investigated allele-specific imputation challenges in regard to geographic origin of the samples using sensitivity and specificity measurements as well as allele frequencies and identified HLA alleles that are challenging to impute for each of the populations separately. In conclusion, our new multi-ethnic reference data set allows for high resolution HLA imputation of genotypes at all classical HLA class I and II genes including the *HLA-DRB3/4/5* loci based on diverse ancestry populations.

Introduction

The major histocompatibility complex, in humans also named human leukocyte antigen (HLA) complex, is a highly variable gene cassette with major functions in the immune system. The HLA region spans ~5 Mb on chromosome 6p21 with genomic positions ranging from 29 Mb to 34 Mb. Genes in this region code for proteins that are involved in many complex functions of the adaptive and innate immune system like the presentation of peptides to the host immune system and also code for proteins that aid peptide presentation or antigen recognition. Results from over 10 years of genome-wide association studies (GWAS) support the HLA as one of the most important disease susceptibility loci for almost every immune-mediated and autoimmune disease. In many cases, the strongest association signals are found within the highly polymorphic classical HLA genes in the class I and II regions, a finding made long before the GWAS era for many of these diseases (1). Therefore, pinpointing the exact genetic variants in the HLA region, which are associated with these diseases, is of utmost importance to disentangle the underlying genetic pathophysiology (2). This is complicated by the highly polymorphic nature of the region, resulting in the need for large disease cohorts to increase statistical power in the detection of genetic association. The costs per sample for Sanger- and next generation sequencing (NGS)-based HLA typing is still at least double that of a genome-wide single nucleotide polymorphism (SNP) array analysis with the new chip platforms. Therefore, imputation methods and reference panels have been developed to provide geneticists with a tool to infer HLA alleles at the classical loci *in silico* using inexpensive and dense SNP array data. These have led to significant advances in fine-mapping of disease relevant genetic variants for many inflammatory and autoimmune diseases (3–5). Published and established HLA imputation tools are amongst others SNP2HLA, HLA Imputation using attribute BAGging (HIBAG) and HLA*IMP (6–8). Imputation of the HLA requires reference panels with high coverage of alleles and genotypes in the region of interest as well as a broad spectrum of samples in order to capture as many different alleles as possible. Additionally, the ancestral background of the reference panel used to impute a data set of interest must be as close as possible to the study population as shown for instance by Jia *et al.* (7). Most HLA imputation reference panels target Caucasian ethnicities and although there has been progress in the development of ancestrally diverse HLA reference panels, studies in which multi-ethnic analyses are performed are still scarce and limited in size (e.g. for chronic inflammatory diseases, (9)). Several imputation references have been published in the past using various genotyping chips and at different resolutions. All reference panels have significantly advanced HLA imputation and analysis conducted with the produced data. However, to date, no full context four-digit multi-ethnic HLA imputation reference panel exists for fine mapping of the HLA region across the totality of the mentioned loci.
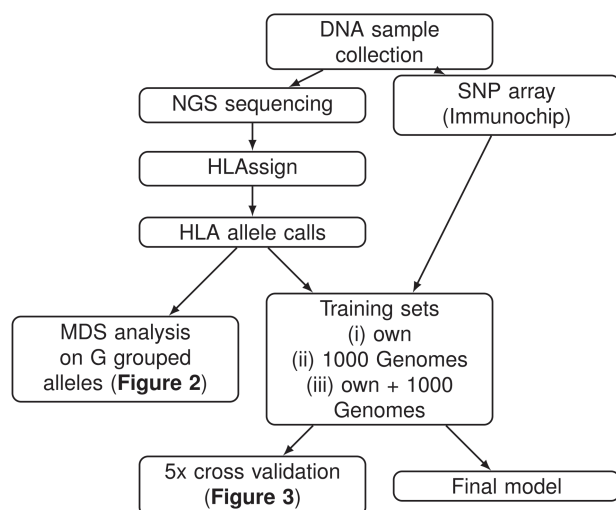
Figure 1. Flowchart of steps taken in preparation and benchmarking of our multi-ethnic reference panel. HLA allele calls were made based on NGS reads. Genotype information was measured using the Illumina Immunochip. These data were combined to train a HIBAG imputation model. Benchmarking was performed using a 5× cross-validation and the independent, previously published, 1000 Genomes data set (24).

With this study, we aimed to create a comprehensive high-quality multi-ethnic HLA reference data set, including *HLA-DPA1*, *-DPB1* and *-DRB3/4/5*, using populations of African American, East Asian (Japan, South Korea and China), European (Germany, Malta) and Middle Eastern (India and Iran) descent.

We generated HLA allele calls from next generation sequencing (NGS) reads for ulcerative colitis (UC) and control individuals of each population, using HLAssign (10) and genotype information using the Illumina Immunochip SNP array [Illumnina, San Diego, CA, USA] (Fig. 1). Using multidimensional scaling (MDS) analysis, we analyzed population structure based on HLA allele frequencies. The combination of called HLA alleles and SNP array genotypes served as training data sets for our new multi-ethnic reference using the HLA imputation tool HIBAG (6). We benchmarked the imputation, applying extensive cross-validation on our multi-ethnic reference panel (Supplementary Material, Fig. S1). The performance of our final model was additionally assessed using the previously published HLA calls of the 1000 Genomes project (11). We also conducted a literature search into the genetic architecture of *HLA-DRB3/4/5* in relation to *HLA-DRB1*, as the presence of the *HLA-DRB3/4/5* are highly dependent on which *HLA-DRB1* allele is carried by an individual. These loci are of particular interest, since they represent a functional variation that has not been considered in many of the previously published reference data sets and hence have been largely excluded in association studies.

Results

### MDS-based clustering of reference samples on HLA allele frequencies

Using MDS analysis on relative frequencies of single HLA G grouped alleles across each cohort, we observed distinct clusters for individuals with East Asian, African and European backgrounds (Fig. 2), except for *HLA-DRB3/4/5* and *HLA-DQB1*. The different subpopulations of our multi-ethnic study population cluster well with respective ethnicities of the 1000 Genomes population. For the 1000 Genomes population, exons 2 and 3

(class I) or exon 2 (class II) were typed only for loci *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* but not for *HLA-DPA1*, *-DPB1* and *-DRB3/4/5*. However, to the best of our knowledge no custom G groups were defined (11). Samples did not show population-specific clustering for *HLA-DQB1*, because frequencies of the HLA alleles in European individuals were similar to those in the Yoruban, African American and European individuals of the 1000 Genomes population. We did not detect consistent clusters for the *HLA-DRB3/4/5* genes, possibly because there was not enough variability to allow good clustering results. In our multi-ethnic data set we only observe four, three and six different four-digit alleles for the *HLA-DRB3/4/5* genes, respectively. In addition, these genes also included a high percentage of null alleles (*HLA-DRB3*, 48.45–81.28%; *HLA-DRB4*, 65.78–84.52%; *HLA-DRB5*, 71.28–85.66%; Table 1) that dominate the frequency spectrum and thus the MDS analysis. With 'null allele' we here refer to the absence of a locus in a given individual. These null alleles are named DRB3*00:00, DRB4*00:00 and DRB5*00:00 throughout this paper. In summary, the MDS analysis reveals significant population heterogeneity for the classical HLA genes and thus, imputation tools should be able to account for this heterogeneity by using population-matched and diverse reference panels.

Imputation benchmark

We performed HLA imputation of the HLA class I loci *HLA-A*, *-B*, *-C* and class II loci *HLA-DQA1*, *-DQB1*, *-DPA1*, *-DPB1*, *-DRB1* and *-DRB3/4/5* using HIBAG and three different constellations: (i) our multi-ethnic reference panel in full four-digit context (Fig. 3 and next paragraph), (ii) our multi-ethnic reference panel combined with the 1000 Genomes data set on G group level (Supplementary Material, Fig. S2 and Supplementary Material, Table S1) and (iii) our multi-ethnic reference panel on G group level as a comparison (Supplementary Material, Fig. S3 and Supplementary Material, Table S2). We also used the 1000 Genomes panel to test the performance of our data (Table 2) with special focus on the imputation for the non-European population panels, as one of the main innovations of this work.

Using a cross-validation approach (Supplementary Material, Fig. S1), we divided the data of each specific population into five random subsamples irrespective of case–control status. For each of the subsets, using the remaining 80% of the population, as well as the HLA allele and genotype information of all other populations, we trained a HIBAG model. The HLA alleles were predicted for the 20% of data from the analyzed population that were not used for training. We calculated accuracies for each of the five subsamples of our population of interest and imputation accuracies for unrelated individuals of the 1000 Genomes population. The results of the cross-validation are depicted in Figure 3 and Table 3. Overall accuracies were high with average accuracies ranging from 0.924 in the Chinese to 0.967 in the Maltese populations (Table 3; Supplementary Material, Table S3). More specifically, high overall accuracies were achieved for the *HLA-C*, *HLA-DP* and *HLA-DQ* loci whereas the *HLA-A*, *-B* and *-DRB1* loci were more challenging to impute across all ethnicities with accuracies as low as 0.862 for *HLA-DRB1* in the Iranian panel. This is also reflected in the posterior probability curves depicted in Figure 3b. Posterior probabilities in HIBAG are used as an additional measure to control prediction accuracies and are generated as an average over all classifiers. Low overall posterior probabilities for a locus indicate that the majority of the alleles were challenging to impute. Note, that correct calls, e.g. for rare alleles, also tend to have smaller posterior probabilities,
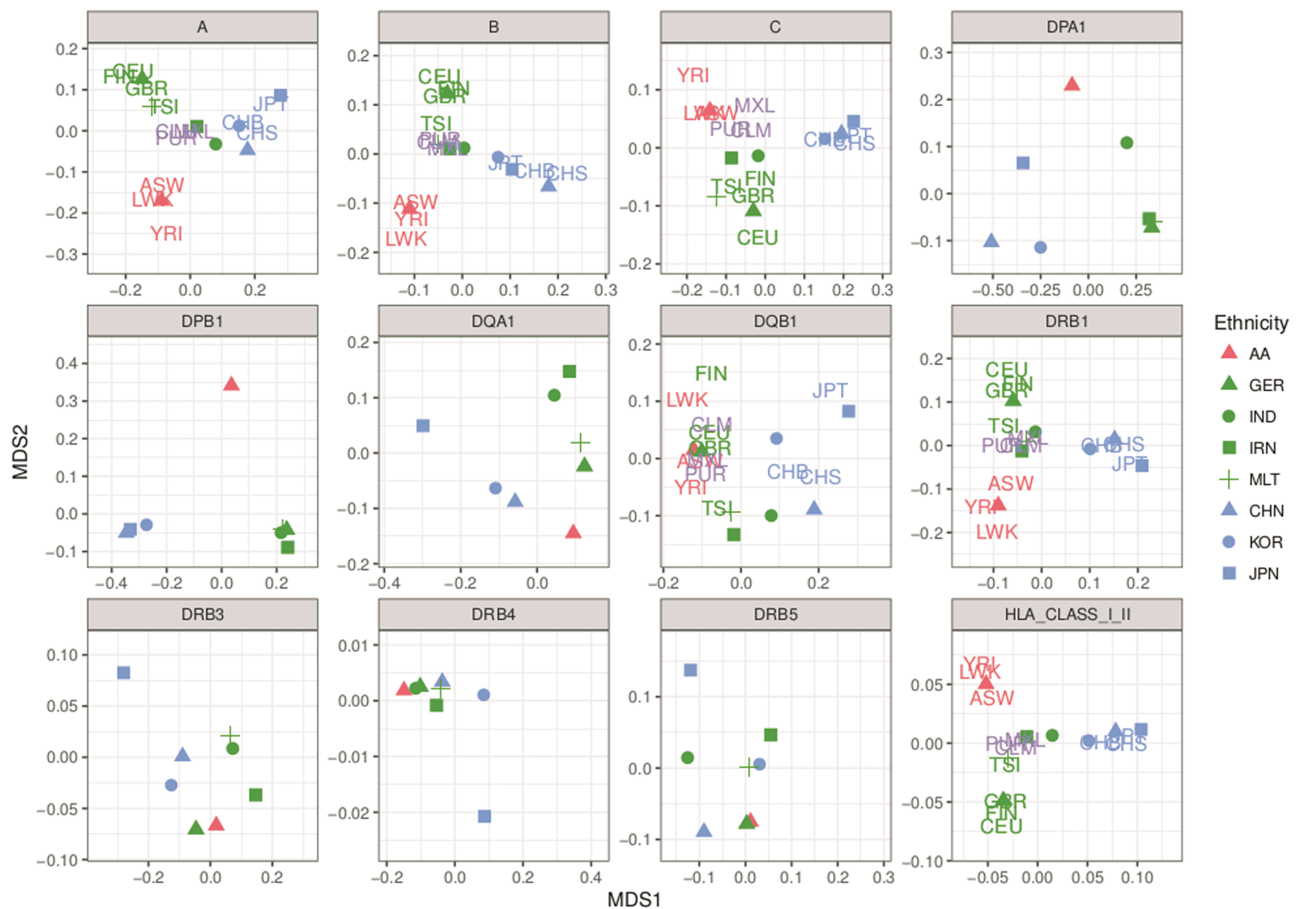
Figure 2. MDS analysis of HLA typed allele data: the MDS analysis was performed using a Euclidean distance measure. Alleles with a frequency <1% were excluded to produce a clustering that is not biased by similarity in low frequency variants. Colors show the origin of the cohort. Red: African American (AA) and African background; Green: European and Middle Eastern background: German (GER), Indian (IND), Iranian (IRN), Maltese (MLT); Blue: Asian background: Hong-Kong Chinese (CHN), South Korean (KOR) and Japanese (JPN); Purple: Non-reference admixed American individuals. Capital acronyms in the panels depict the 1000 Genomes populations as described in Auton *et al.*, (24). The 1000 Genomes populations include Americans of African Ancestry in the Southwest USA (ASW), Africans from Kenya (LWK), Nigeria (YRI), Columbian (CLM), Mexican (MXL) and Puerto Rican (PUR), Han Chinese in Beijing (CHB), Southern Han Chinese (CHS), Japanese in Tokyo (JPT), Finnish (FIN), British (GBR), Tuscan (TSI) and samples with Western European Ancestry collected in the CEPH diversity panel (CEU). For *HLA-DPA1, -DPB1, -DQA1* and the *-DRB3/4/5* loci no data was available in those panels. For the MDS analysis across all loci (HLA CLASS I_II) we included *HLA-A, -B, -C, -DQB1* and *-DRB1*. Samples of our own cohorts cluster well with the corresponding 1000 Genomes population.

while incorrect calls can have a high posterior probability when haplotypes of two alleles are similar across many classifiers. Therefore, we decided to additionally use other measures such as sensitivity and specificity, and allele specific accuracy to evaluate allele specific results in the following analyses. With 29–55 alleles per population, and 75% (Malta) to 82% (Japan) of the alleles having frequencies of <1% (Supplementary Material, Tables S4 and S5), *HLA-B* presented a particular challenge for imputation. Similarly challenging were *HLA-A* and *-DRB1*, which are discussed further below. The remaining loci were not as variable or had a smaller and more even frequency spectrum (Supplementary Material, Table S5), such that posterior probabilities were higher. *HLA-DPA1* and *-DPB1* had the most "on target" SNPs (30 and 51 SNPs, respectively) (Supplementary Material, Table S6), reflecting the fact, that these loci are least variable and therefore better suited to be captured on a SNP genotyping array. Overall, between 682 (*HLA-DPB1*) and 1,794 (*HLA-A*) SNPs were located within the different gene loci including flanking regions of 500 kb upstream and downstream of each gene. A median of 41.5 (*HLA-DRB5*) to 81 (*HLA-A*) SNPs were used by the single classifiers of HIBAG.

In the following, we show the results of the imputation with our own reference data set divided by ethnic background and also compare our data to previously reported HLA imputation accuracies on published data sets from Dilthey *et al.* (8), Jia *et al.* (7), Okada *et al.* (12), Kim *et al.* (13) and Zheng *et al.* (6) (Table 4). It is of importance to note, that high accuracies for a reference panel using a specific benchmarking panel are best achieved when the benchmarking panel follows the same allele nomenclature and grouping as the panel used for imputation. We could not determine to which extent this was considered in each of the above studies, but we estimate that the effect should not be detrimental if differences only occur between slightly different custom allele groupings (i.e. we assume that the allele that a grouping is based on is also the most frequent allele) and not between different levels of grouping (i.e. full context versus G groups). A summary of these data sets is described in Table 4. The following results are specific to the imputation of HLA alleles into the respective populations using our multi-ethnic four-digit full context reference panel. If not stated otherwise, mean accuracies were compared for four-digit allele imputations of *HLA-A, -B, -C, -DQB1* and *-DRB1*. These are the loci that are
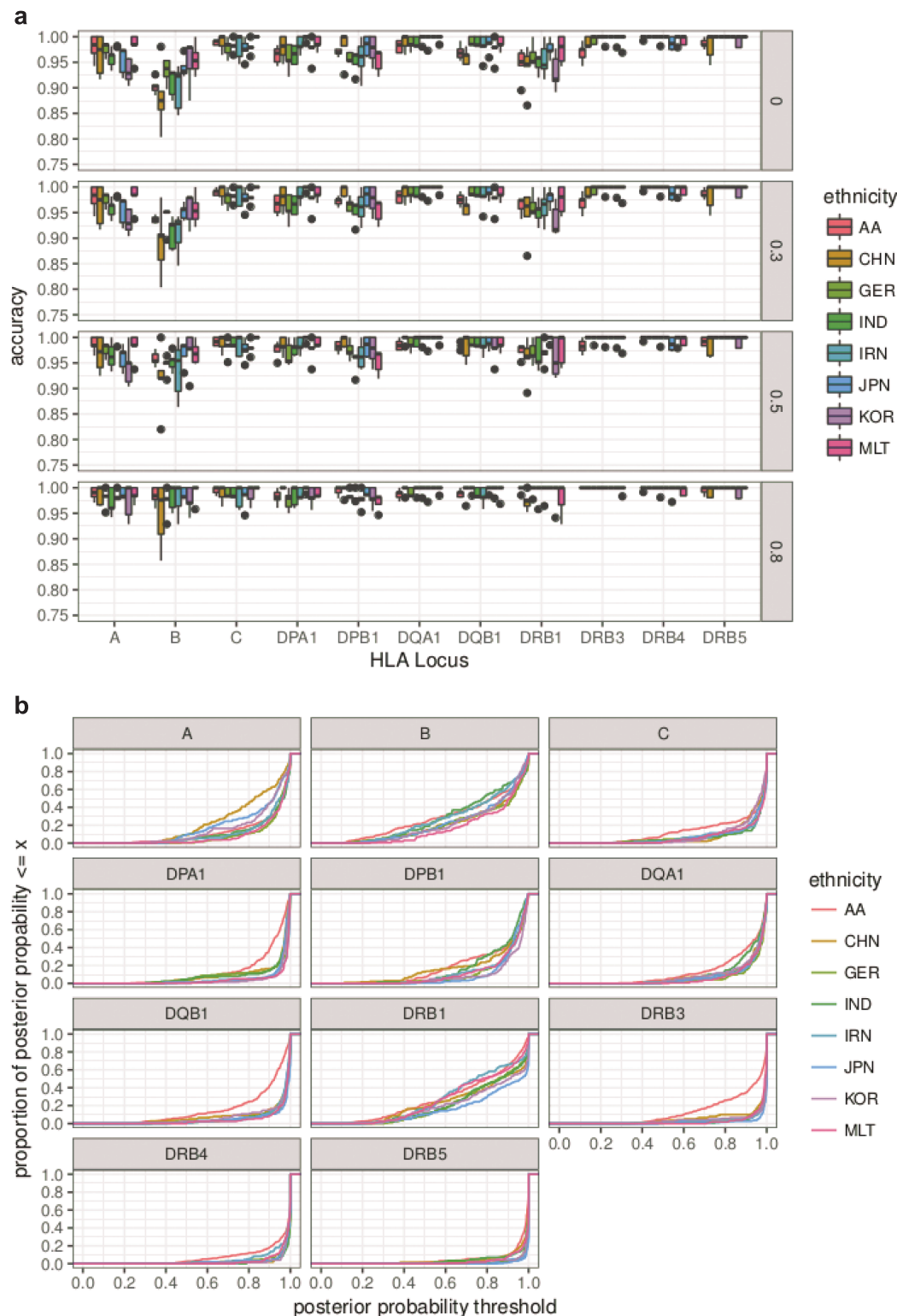
Figure 3. Imputation accuracies employing the multi-ethnic reference panel: accuracies and post-imputation probabilities of HLA imputation with HIBAG using a 5-fold cross-validation scheme and the multi-ethnic data set with full four-digit allele information. 20% of the data with a specific ethnic background were used as the validation set after training a model that used 80% of the remaining data and all data from other ethnic backgrounds. We included 1,360 African American (AA), Hong-Kong Chinese (CHN), German (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) samples in total. **(a)** Accuracies are depicted according to post-imputation probabilities with cut-off thresholds at 0 (no confidence filtering), 0.3, 0.5, 0.8 (only high confidence genotypes). Loci are shown according to alphabetical order. Imputation accuracies are especially high for *HLA-C, -DPA1, -DPB1, -DQB1* and the *-DRB3/4/5. HLA-DRB1* accuracies are especially lowered by misclassifications of DRB1*04:03, DRB1*04:04 and DRB1*11:04. **(b)** Posterior probabilities are depicted as proportion of the number of samples with a posterior probability smaller than a threshold (*x*-axis).

Table 1. Frequencies of HLA-DRB3/4/5 in our multi-ethnic reference panel: frequencies of *HLA-DRB3/4/5* in the typed HLA data for African American (AA), Hong-Kong Chinese (CHN), German (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) populations at full four-digit context. Null alleles have the highest frequencies. For *HLA-DRB4* mainly one other allele, DRB4*01:03, exists. DRB5*01:01 is the second most abundant of the *HLA-DRB5* alleles in all but the Japanese and Iranian panels, where DRB5*01:02 is seen more often.

	AA	CHN	GER	IND	IRN	JPN	KOR	MLT
DRB3*00:00	51.61	64.60	59.88	56.74	48.45	81.28	64.34	55.00
DRB3*01:01	11.13	2.55	14.51	5.32	8.53	4.55	11.07	4.69
DRB3*02:02	27.74	19.34	22.53	32.98	37.98	8.82	16.39	33.75
DRB3*02:24	0.00	0.00	0.62	0.00	0.39	0.00	0.00	0.31
DRB3*03:01	9.52	13.50	2.47	4.96	4.65	5.35	8.20	6.25
DRB4*00:00	84.52	75.91	80.25	80.85	75.97	65.78	68.44	75.63
DRB4*01:01	6.77	0.00	2.47	0.35	1.55	0.00	0.00	3.75
DRB4*01:02	0.00	0.00	0.00	0.00	0.39	2.14	0.41	0.00
DRB4*01:03	8.71	24.09	17.28	18.79	22.09	32.09	31.15	20.31
DRB4*03:01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31
DRB5*00:00	81.94	72.63	80.56	71.28	85.66	71.66	82.38	81.56
DRB5*01:01	15.97	21.53	16.67	15.96	5.43	6.42	11.07	10.00
DRB5*01:02	0.32	1.82	0.62	12.77	6.98	20.59	4.51	3.75
DRB5*01:03	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00
DRB5*01:08	0.32	2.19	0.00	0.00	0.00	0.27	0.41	0.00
DRB5*02:02	0.97	0.36	2.16	0.00	1.94	1.07	1.64	4.69
DRB5*02:03	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00
DRB5*02:13	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. Imputation accuracies for 1000 Genomes populations: population groups are depicted in **bold** and the subpopulations in *italic* type. African (AFR) samples are divided into Americans of African Ancestry in the Southwest USA (ASW), Africans from Kenya (LWK) and Nigeria (YRI). Admixed American (AMR) samples are split into samples with Columbian (CLM), Mexican (MXL) and Puerto Rican (PUR) ancestry. East Asians (EAS) were collected as Han Chinese in Beijing (CHB), Southern Han Chinese (CHS) and Japanese in Tokyo (JPT). Samples with European Ancestry (EUR) are Finnish (FIN), British (GBR), Tuscan (TSI) and samples with Western European Ancestry collected in the CEPH diversity panel (CEU). Accuracies of *HLA-DRB1** are *HLA-DRB1* measured without DRB1*04:03, DRB1*04:04 and DRB1*11:04, which improved accuracies for all ethnicities. HLA-A* are accuracies measured without A*02:03, which improved accuracies for the Chinese samples. Overall accuracies were highest for EUR samples and lowest for the non-AMR, for which no samples with similar backgrounds are included in our novel imputation reference.

	#samples	A	B	C	DQB1	DRB1	mean	A*	DRB1*
AFR	**162**	0.920	0.833	0.932	0.951	0.886	0.904	0.920	**0.906**
ASW	41	0.939	0.805	0.915	0.939	0.902	0.900	0.939	**0.923**
LWK	75	0.880	0.853	0.960	0.980	0.893	0.913	0.880	**0.899**
YRI	46	0.967	0.826	0.902	0.913	0.859	0.893	0.967	**0.902**
AMR	**193**	0.909	0.756	0.972	0.984	0.710	0.866	0.909	**0.766**
CLM	67	0.925	0.709	0.970	0.985	0.687	0.855	0.925	**0.711**
MXL	56	0.857	0.688	0.973	0.991	0.598	0.821	0.857	**0.674**
PUR	70	0.936	0.857	0.971	0.979	0.821	0.913	0.936	**0.888**
EAS	**260**	0.929	0.931	0.975	0.992	0.940	0.953	**0.941**	**0.951**
CHB	82	0.939	0.921	0.988	0.994	0.939	0.956	**0.948**	**0.967**
CHS	92	0.935	0.924	0.967	0.995	0.935	0.951	**0.963**	**0.944**
JPT	86	0.913	0.948	0.971	0.988	0.948	0.953	0.913	0.943
EUR	**322**	0.983	0.944	0.994	0.989	0.890	0.960	0.983	**0.968**
CEU	52	0.981	0.922	0.971	1.000	0.865	0.948	0.981	**0.987**
FIN	95	0.984	0.974	1.000	0.989	0.926	0.975	0.984	**0.959**
GBR	86	0.977	0.959	1.000	0.983	0.884	0.960	0.977	**0.993**
TSI	89	0.989	0.910	0.994	0.989	0.871	0.951	0.989	**0.944**

present for all imputation references (Table 4). Within the cross-validation framework, accuracies for a gene were calculated as an average across the different cross-validation runs as it has been done previously (12,13) and enables better comparison of these values between studies. We also report median, minimum and maximum values in Supplementary Material, Table S3. We report accuracies across all imputed alleles in Table 3, Supplementary Material, Tables S1 and S2. A few alleles were especially challenging to impute, both within our as well as in

previously published reference panels. These alleles usually have comparably lower sensitivity or specificity scores and similar haplotype structures within the same 2-digit allele groups (Supplementary Material, Tables S7 and S8, Supplementary Material, Tables S5–S8 of Zheng *et al.*, (6)). This is especially important in the context of association analyses where the greatest impact from these issues is seen with higher frequency variants (AF >1%) and thus needs to be considered carefully. Note that this also depends on the ethnicity of the samples

Table 3. Imputation accuracies of the imputation with the multi-ethnic reference panel: 20% of the data with a specific ethnic background were used as validation set after training a model with 80% of the remaining data and all data from other ethnic backgrounds. We included 1,360 African American (AA), Hong-Kong Chinese (CHN), German (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) samples in total in the imputation reference. Shown are mean accuracies of the HLA imputation with HIBAG using a 5-fold cross-validation scheme and the multi-ethnic data set with full four-digit allele information. The given mean considers only the loci highlighted in **bold**, as these are loci also analyzed in all previous publications. Accuracies of *HLA-DRB1** are *HLA-DRB1* measured without DRB1*04:03, DRB1*04:04 and DRB1*11:04, which improves accuracies for all ethnicities. *HLA-A** are accuracies measured without A*02:03, which improves accuracies for the Chinese samples. Overall, *HLA-B* is the most challenging to impute. Mean accuracies are higher than 0.925 across all cross-validation runs. Best results are achieved for the GER, JPN and MLT populations.

	AA	CHN	GER	IND	IRN	JPN	KOR	MLT
#samples	312	140	162	143	132	189	122	160
A	0.969	0.900	0.976	0.955	0.973	0.936	0.939	0.984
B	0.877	0.868	0.917	0.875	0.885	0.938	0.934	0.947
C	0.953	0.986	0.975	0.979	0.974	0.973	0.968	0.988
DPA1	0.969	0.979	0.960	0.968	0.985	0.995	0.975	0.988
DPB1	0.925	0.949	0.960	0.944	0.954	0.979	0.963	0.956
DQA1	0.942	0.975	0.975	0.965	0.962	0.968	0.959	0.978
DQB1	0.962	0.964	0.988	0.990	0.981	0.984	0.975	0.984
DRB1	0.925	0.903	0.948	0.924	0.862	0.960	0.918	0.931
DRB3	0.971	1.000	1.000	1.000	1.000	1.000	0.996	0.994
DRB4	0.977	1.000	0.991	0.996	0.996	0.990	1.000	0.988
DRB5	0.987	0.982	1.000	1.000	1.000	1.000	0.992	1.000
mean	0.937	0.924	0.961	0.944	0.935	0.958	0.947	0.967
A*	0.969	0.954	0.976	0.954	0.973	0.935	0.937	0.984
DRB1*	0.930	0.904	0.954	0.952	0.956	0.968	0.926	0.971

evaluated. We describe A*02:01/A*02:03, DRB1*11:01/DRB1*11:04 and DRB1*04:03/DRB1*04:04 below for illustration purposes.

African American panel

The imputation of HLA alleles into our own African American data set achieved an average imputation accuracy on full context four-digit level of 0.951 across all analyzed loci and of 0.937 on average for loci *HLA-A, -B, -C, -DQB1* and *-DRB1* only (Table 3). Employing our multi-ethnic reference data set on G group level (ii), we were able to impute alleles of the genes *HLA-A, -B, -C, -DQB1* and *-DRB1* of the 1000 Genomes African ancestry data with a mean accuracy of 0.904 and highest accuracies for the Luhya Kenyan samples alone (0.880–0.980; mean of 0.913; Table 2). In comparison, Zheng *et al.* (6) imputed HLA alleles of random subsets of their African American HLARES data combined with the Yoruba Nigerians (YRI) HapMap samples with a reported mean accuracy of 0.818 using their tool HIBAG (Table 4b). Jia *et al.* (7) imputed the HLA alleles of YRI HapMap samples using their Caucasian Type 1 Diabetes Genome Consortium (T1DGC) reference panel with accuracies between 0.203 (*HLA-DRB1*) and 0.984 (*HLA-C*) across all loci and an overall mean accuracy of 0.750 (Table 4a).

East Asian panel

Employing our multi-ethnic reference data set (i) to impute HLA alleles into our Chinese samples, we achieved accuracies of 0.868 (*HLA-B*) to 1.000 (*HLA-DRB3/4*) and of 0.924 on average for *HLA-A, -B, -C, -DQB1* and *-DRB1*. We imputed HLA alleles into our Japanese samples with accuracies of 0.936 (*HLA-A*) to 1.000 (*HLA-DRB3/5*) and 0.958 on average for *HLA-A, -B, -C, -DQB1* and *-DRB1*. For our Korean samples imputation accuracies of 0.918 (*HLA-DRB1*) to 1.000 (*HLA-DRB4*) were reached,

with an average accuracy of 0.947 (Table 3). Additionally, we imputed the HLA alleles of the East Asian 1000 Genomes data on G group level (ii) with mean accuracies higher than 0.953 (Table 2).

In comparison, Okada *et al.* (12), Jia *et al.* (7), Kim *et al.* (13) and Zheng *et al.* (6) reported mean accuracies between 0.77 to 0.922 for *HLA-A, -B, -C, -DQB1* and *-DRB1* (Table 4) for East Asian populations using their respective HLA imputation panels. *HLA-DPA1* or *HLA-DRB3/4/5* is not considered in any of the publications for East Asian ethnicities. For single loci the reported imputation accuracies vary between 0.656 (*HLA-B* with T1DGC reference for Han Chinese in Beijing (CHB) and Japanese samples (JPT); (7)) and 0.984 (HLA-C with a Korean reference panel and the same test population; (13)).

In the cross-validation benchmark the accuracy of locus HLA-A in the Chinese population (Fig. 3a) was decreased due to a misclassification of A*02:03 to A*02:01 in 32% of 37 samples in which this allele occurred. This misclassification is due to the high similarity between these alleles (Supplementary Material, Supplementary Text). When excluding A*02:03 from accuracy calculations for *HLA-A*, accuracies improved for the Chinese subpopulation from 0.900 to 0.954 (Table 3).

Iranian and Indian panels

Overall imputation accuracies for our Indian and Iranian panels over all loci were 0.944 and 0.935, respectively. The accuracies were high for all loci except *HLA-B* (0.875 and 0.885, respectively) and *-DRB1* (0.924 and 0.862, respectively) (Table 3).

The accuracy of the Iranian samples in the cross-validation benchmark (Fig. 3a) at *HLA-DRB1* was low due to a misclassification of DRB1*11:04 to DRB1*11:01 in 39% of the 36 Iranian samples in which this allele occurs (Supplementary Material, Supplementary Text). When excluding the DRB1*11:04 as well as the DRB1*04:04 and DRB1*04:03 alleles (see below) from accuracy calculations for *HLA-DRB1*, the accuracies improved from 0.862

Table 4. Previously reported imputation accuracies: accuracies measured for HLA reference panels, which are mainly based on Caucasian and Asian data, with origin of the publications and cohorts used for training and validation as well as a comparison to accuracies achieved with our own multi-ethnic reference panel (i) in the cross-validation experiment on our own data (see also Table 3) and on the 1000 Genomes cohorts (see also Table 2). Accuracies of the cross-validation (own) framework and of the imputation into the 1000 Genomes population are shown. Mean accuracies are calculated across *HLA-A, -B, -C, -DPB1* and *-DRB1* (loci highlighted in **bold**). Mean accuracies of the listed reference panels are lower compared to our own reference panel in the majority of the cases, especially in the non-European population. **(a)** Accuracies published with SNP2HLA. The international T1DGC reference panel (7) published along with SNP2HLA was used to gain the accuracies on the 1948 British Birth Cohort and the HapMap-CEPH Cohort, two European ancestry panels. The T1DGC panel was further used for imputing the Yoruban Nigerian (YRI), the East Asian Han Chinese from Beijing (CHB) and the Japanese from Tokyo (JPT) samples of the 1000 Genomes data sets. For the East Asian 1000 Genomes panels accuracies reached by later-published ethnic-specific references (12,13) are also listed. **(b)** Accuracies published with HIBAG using the HLARES data from GlaxoSmithKline (GSK) clinical trials of specific ethnic background combined with 1000 Genomes data sets (6). **(c)** Accuracies published with HLA*IMP:02 using different combinations of the Golden Set (GS = 1948 Birth Cohort/ HapMap CEU and CEPH CEU+) and the HLARES data as references (8).

(a) SNP2HLA

Source	Jia *et al.* (7)				Okada *et al.* (12)	Kim *et al.* (13)	
imputation reference	T1DGC				Japanese	Korean	Korean
# training samples	5,225				918	330	413
test population	1948 British Birth Cohort	CEPH	YRI	CHB & JPT	JPT	random subset	CHB & JPT
# test samples	918	90	not specified	not specified	44	83	61
A	0.981	0.991	0.699	0.981	0.908	0.908	0.91
B	0.968	0.968	0.905	0.656	0.943	0.859	0.893
C	0.969	0.991	0.984	0.688	0.989	0.928	0.984
DPA1	/	/	/	/	/	/	/
DPB1	/	/	/	/	/	0.95	/
DQA1	/	0.985	0.649	0.963	/	/	/
DQB1	0.983	0.991	0.961	0.964	0.894	0.937	0.893
DRB1	0.933	0.969	0.203	0.923	0.843	0.868	0.893
DRB3	/	/	/	/	/	/	/
DRB4	/	/	/	/	/	/	/
DRB5	/	/	/	/	/	/	/
mean	0.967	0.983	0.729	0.864	0.915	0.908	0.915
mean A-C, DQB1, DRB1	0.967	0.982	0.75	0.842	0.915	0.9	0.915

mean A-C, DQB1, DRB1	own						
	GER 0.961	GER 0.961	AA 0.937	CHN 0.924	CHN 0.924	CHN 0.924	CHN 0.924
	MLT 0.967	MLT 0.967		JPN 0.958	JPN 0.958	JPN 0.958	JPN 0.958
				KOR 0.947	KOR 0.947	KOR 0.947	KOR 0.947
	1000 Genomes						
	EUR 0.96	EUR 0.96	ASW 0.9	CHB 0.956	CHB 0.956	CHB 0.956	CHB 0.956
			LWK 0.913	CHS 0.951	CHS 0.951	CHS 0.951	CHS 0.951
			YRI 0.893	JPT 0.953	JPT 0.953	JPT 0.953	JPT 0.953

(b) HIBAG

Source	Zheng *et al.* (6)			
imputation reference	HLARES data of Asian ancestry & CHB & JPT	HLARES data of Hispanic ancestry	African American HLARES data & 60 African YRI	HLARES data of European ancestry
# training samples	720 + 90 (minus test)	439 (minus test)	173 + 60 (minus test)	2668 (minus test)
test population	random subset	random subset	random subset	random subset
# test samples	subset	subset	subset	subset
A	0.921	0.934	0.924	0.982
B	0.875	0.75	0.768	0.966
C	0.966	0.962	0.885	0.988
DPA1	/	/	/	/

(Continued).

Table 4. *Continued*

(b) HIBAG

DPB1	0.898	0.931	0.8	0.947
DQA1	0.868	0.938	0.794	0.964
DQB1	0.96	0.957	0.742	0.992
DRB1	0.887	0.82	0.771	0.921
DRB3	/	/	/	/
DRB4	/	/	/	/
DRB5	/	/	/	/
mean	0.911	0.899	0.812	0.966
mean A-C, DQB1, DRB1	0.922	0.885	0.818	0.97

mean A-C, DQB1, DRB1

<div align="center">own</div>

	CHN	0.924			AA	0.937	GER	0.961
	JPN	0.958					MLT	0.967
	KOR	0.947						

<div align="center">1000 Genomes</div>

	CHB	0.956	PUR	0.913	ASW	0.9	EUR	0.96
	CHS	0.951			LWK	0.913		
	JPT	0.953			YRI	0.893		

(c) HLA*IMP:02

Source			Dilthey *et al.* (8)			
imputation reference	GS	HLARES _EU	GS & HLARES_ALL			
# training samples	1,585	1,758	2,055			
test population	HLARES_EU	random subset	African Americans of random subset	Asians of random subset	Europeans of random subset	Hispanic of random subset
# test samples	1,060	872	1,008 (all populations)			
A	0.96	0.97	0.73	0.79	0.96	0.82
B	0.9	0.95	0.73	0.68	0.95	0.63
C	0.96	0.96	0.97	0.82	0.97	0.92
DPA1	/	/	/	/	/	/
DPB1	/	0.90 (2-digit)	/	/	/	/
DQA1	0.87	0.97	1	0.73	0.96	0.93
DQB1	0.98	0.98	0.87	0.83	0.97	0.97
DRB1	0.88	0.91	0.71	0.72	0.9	0.8
DRB3	/	0.94 (2 digit)	/	/	/	/
DRB4	/	0.98 (2 digit)	/	/	/	/
DRB5	/	0.99 (2 digit)	/	/	/	/
mean	0.93	0.95	0.84	0.76	0.95	0.85
mean A-C, DQB1, DRB1	0.94	0.95	0.8	0.77	0.95	0.83

mean A-C, DQB1, DRB1

<div align="center">own</div>

GER	0.961	GER	0.961		AA	0.937	CHN	0.924	GER	0.961	
MLT	0.967	MLT	0.967				JPN	0.958	MLT	0.967	
							KOR	0.947			

<div align="center">1000 Genomes</div>

EUR	0.96	EUR	0.96	ASW	0.9	CHB	0.956	EUR	0.96	PUR	0.913
				LWK	0.913	CHS	0.951				
				YRI	0.893	JPT	0.953				

to 0.956 (Table 3). Mean sensitivity values for DRB1*11:04 for the cross-validation runs were 0.307 for the Iranian population and 0.208 for the Indian population (Supplementary Material, Table S8). The frequency of this allele was 2.82% and 13.85%, respectively (Supplementary Material, Table S5).

The improvement of the overall accuracy by excluding these alleles in the Indian samples (0.924 to 0.952) was not as big as in the Iranian samples because of the lower allele frequency (AF). Previously reported sensitivity values for the DRB1*11 alleles (Supplementary Material, Tables S5–S8 of Zheng *et al.* (6)) range

from 0.627 (DRB1*11:04) to 0.993 (DRB1*11:01) in the European population. In this previous study, misclassifications occurred for DRB1*11:04, too, which was called as DRB1*11:01 in 93% of cases when a misclassification occurred in European samples (6). This is in line with our own results.

Imputation for non-reference populations

The Latin American admixed populations of the 1000 Genomes data set (containing Amerindian and European, for Puerto Rico also West African ancestral admixture, here grouped into Mexican, Columbian and Puerto Rican populations) were imputed with mean accuracies ranging from 0.821 for the Mexican, 0.855 for the Columbian to 0.913 for the Puerto Rican population (Table 2). In particular, *HLA-B* and *-DRB1* showed low imputation accuracies (0.688 to 0.857 and 0.598 to 0.821, respectively) while all remaining loci had accuracies higher than 0.857 (Table 2). Overall, the Puerto Rican data set showed highest accuracies and only 40 out of 134 total measured alleles had sensitivity values of lower than 1.000 (Supplementary Material, Table S9). Out of these 40 alleles, 22 have an AF <0.1% in the Puerto Rican panel. Accuracies for loci imputed within the Puerto Rican data set ranged from 0.821 (*HLA-DRB1*) to 0.979 (*HLA-DQB1*) (Table 2).

HLA-DRB3/4/5 haplotypes

Many imputation tools allow the imputation of *HLA-A, -B, -C, -DQB1* and *-DRB1* but only a few studies have reported on the imputation of the *HLA-DRB3, -DRB4 and -DRB5* (*HLA-DRB3/4/5*) loci, such as Dilthey *et al.* (8), who analyzed *HLA-DRB3/4/5* imputation in Caucasian data sets (Table 4c). These genes can be present or absent in an individual depending on the *HLA-DRB1* genotype. For the evaluation of the imputation of these genes and to elucidate which *HLA-DRB3/4/5* loci are known to be located on the same haplotype as a specific *HLA-DRB1*, we conducted an extensive literature review and present the results below. We mainly focus on the information reported by Holdsworth *et al.* (14), Robbins *et al.* (15) and Bontrop *et al.* (16). According to literature, alleles of the *HLA-DRB3/4/5* loci occur within a specific *HLA-DRB1* context, being present in some haplotypes and absent in others. The results of this review are summarized in Figure 4. Haplotypes with *HLA-DRB1* always carry the pseudogene *HLA-DRB9*, which is located downstream of *HLA-DRB1* and that consists of two exons (17). DRB1*01, DRB1*08 and DRB1*10 are not found with any *HLA-DRB3/4/5* allele. Haplotypes with DRB1*03, *11, *12, *13 and *14 are found with *HLA-DRB2* and *-DRB3*. DRB1*04, *07, *09 are found with *HLA-DRB4* as well as *-DRB7* and *-DRB8*. Finally, DRB1*15 and *16 are reported to be located on the same haplotype as *HLA-DRB5*. Exceptions to his rule have been described for DRB1*15 and *16, where especially in African Americans *HLA-DRB5/6* can be missing. DRB1*07 has been reported to occur with a non-expressed form of DRB4*04:01 (15) and DRB1*08 has also been previously identified together with DRB3*03:01 (15).

We investigated our herein-described multi-ethnic data on *HLA-DRB1* and *-DRB3/4/5* for congruence with these previous findings. In short, we determined the *HLA-DRB1* alleles for every sample and checked whether we could also find the expected *HLA-DRB3/4/5* alleles or the absence of these in the same sample. All but four samples followed the haplotype structures depicted in Figure 4. After re-analysis of the remaining four samples we concluded that these samples must have been contaminated, since three or more alleles could plausibly be called for all ana-lyzed loci, with one allele having a smaller number of reads that aligned to it. In further six samples we found one of the exceptions described in the literature. One Maltese sample did not have *HLA-DRB4* while DRB1*07:01 was present and five African American samples did not have *HLA-DRB5* while DRB1*15:03 or DRB1*16:02 was present.

Frequencies of *HLA-DRB3/4/5* are shown in Table 1. Overall, *HLA-DRB3* is the most variable of those genes according to its frequency spectrum, with DRB3*02:02 being the most common non-null allele with an AF ranging from 8.82% in our Japanese panel to 37.98% in our Iranian panel. For *HLA-DRB4*, DRB4*01:03 is the most common non-null allele with frequencies ranging from 8.71% in the African American to 32.09% in the Japanese panel. DRB5*01:01 is the most common non-null allele in all but the Iranian and Japanese panels with frequencies of 5.43% in the Iranian to 21.53% in the Chinese panel, while DRB5*01:02 has a frequency of 20.59% in the Japanese panel and a frequency of 6.98% in the Iranian panel. Our data suggest that DRB1*15:01 is located on the same haplotype as DRB5*01:01, while DRB1*15:02 (which is very common in Japanese samples) is located on the same haplotype as DRB5*01:02 (Supplementary Material, Table S10). Accuracies of the *HLA-DRB3/4/5* imputations are high (>0.971; Table 3 and Fig. 3a). Sensitivity measures for the *HLA-DRB3/4/5* are generally high; however, for low frequency variants (e.g. DRB3*02:24 in the Iranian, Maltese and German panels at frequencies of <0.62%) values as low as 0 were measured. DRB4*01:02 in the Japanese panel, DRB3*01:01 and DRB4*01:01 in the African American panel are common alleles (AF > 1%) classified with mean sensitivity values of lower than 0.800 (0.375, 0.739, 0.690, respectively). We also observed, using the tool Disentangler (18), that the phasing of *HLA-DRB3/4/5* alleles might present a challenge, with many of the null alleles occurring on haplotypes with *HLA-DRB1*, when the respective *HLA-DRB3/4/5* allele is present (Supplementary Material, Fig. S4; *HLA-DRB3/4/5* are excluded here). The analysis of this particular topic, however, is beyond the scope of this paper.

Discussion

We compiled three different imputation panels as pre-trained HIBAG models that can be used for HLA imputation in different ethnicities: (i) a multi-ethnic reference with four-digit full context HLA alleles and (ii) a multi-ethnic reference with four-digit HLA alleles as G groups. Both panels include *HLA-A, -B, -C, -DQA1, -DQB1, -DPA1, -DPB1, -DRB1* and *-DRB3/4/5* and (iii) a multi-ethnic reference panel combined with the 1000 Genomes data (including data from *HLA-A, -B, -C, -DQA1, -DRB1, -DPA1, -DPB1* at a four-digit G group resolution). Our reference panels have high accuracy values across different ethnicities and subsets of the data and also achieve high accuracies in non-reference ethnicities (Tables 2 and 3). The accuracies in non-reference ethnicities are high, but lower than for our reference data sets, as even though our reference is highly diverse the worldwide diversity of the HLA is still not sufficiently captured. Average accuracies of our multi-ethnic reference are larger than 0.924. Tabulated results describing the accuracy measures of panels (ii) and (iii) are presented in Supplementary Material, Tables S1 and S2. Using our reference data, few alleles remain challenging to impute. This affects alleles of the *HLA-DRB1* locus, like the DRB1*11 and DRB1*04 group, which has already been described as problematic in previous benchmarks of other imputation reference panels (6–8) as well as alleles of the highly diverse *HLA-A* and *-C* genes. We therefore recommend using a two-
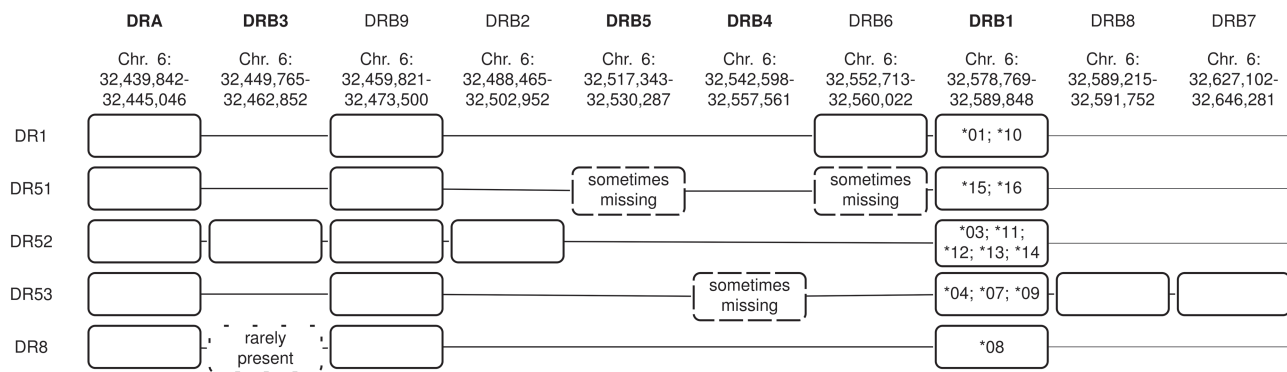
Figure 4. Known architecture of *HLA-DRB3/4/5*: HLA haplotypes that usually contain a specific *HLA-DRB1* allele (*HLA-DRB1* column) are shown. Two-digit alleles are denoted. All loci are depicted in order of their genomic location. *HLA-DRA*, *HLA-DRB1* and *HLA-DRB9* coincide with all haplotypes. The remaining loci are present or absent depending on the haplotype. The most prevalent haplotypes with the known exceptions are shown in the rows below. Exceptions are sometimes seen for *DRB1*08*, *DRB1*07*, *DRB1*15* and *DRB1*16*. *DRB1*08* can occur with *HLA-DRB3*, *DRB1*07* can occur without an expressed form of *HLA-DRB4* and *DRB1*15* and *DRB1*16* can occur without *HLA-DRB5/6*. Loci that usually occur together are joined by a line. The name of the corresponding serotype is shown on the left and haplotypes are ordered by serotype name. Information for this figure was retrieved from Bontrop *et al.*, Holdsworth *et al.* and Robbins *et al.* (14–16).

digit resolution for these alleles and to consider the imputation difficulties in the interpretation of association results for these alleles. We further suggest that the interpretation of specificity and sensitivity measures should be done separately by ethnic background, since measures can vary between ancestries, i.e. haplotypes for an allele that are highly predictive in one ethnicity may not be highly predictive in another ethnicity. We also verified that SNPs missing in the data set for which HLA alleles are imputed—and that exist in the reference— can negatively affect the imputation accuracy. This was the case for DRB1*04:03 and DRB1*04:04, where exclusion of 4.4% of the SNPs used by the HIBAG had a major impact on the imputation accuracy for these alleles (Supplementary Material, Supplementary Text). We therefore suggest, as a general rule, to cautiously investigate the coverage of SNPs used by any imputation reference panel prior to imputation with the respective panel into a data set. Posterior probabilities are often used to improve the quality of the data set. Indeed, we also observe that the accuracies improve when using a posterior probability threshold. However, for some alleles similar haplotype structures can cause incorrect calls despite high posterior probabilities. Especially for rare alleles, correct calls are possible at a very low posterior probability. We therefore suggest using the sensitivity and specificity tables we provide in Supplementary Material, Table S8 to perform data filtering as well as checking the posterior probability.

In summary, imputing HLA alleles into multi-ethnic genome-wide association data sets with our reference panels provides accurate results and can aid HLA fine mapping studies especially in non-Caucasian populations in the future. It allows for HLA imputation using the most recent HLA allele nomenclature at a full context four-digit resolution and a high diversity of different populations.

Nevertheless, larger sample sizes and even more diverse reference panels are needed to adequately cover the existing global HLA polymorphism and frequency spectrum particularly for the ethnicities not included in our panel and also to impute especially rare HLA alleles with high accuracy. DRB1*01:03, for instance, is an allele that has a higher frequency in North American Caucasians (0.9–1.9%) than European Caucasians (~0.6%) (19). As over a million of samples will have been genotyped and whole-genome sequenced in the near future, it is just a matter of warranting global coverage, thus to include representatives from every ethnicity for these efforts. Still, most genetic research focuses on Caucasian ancestry cohorts and neglects large segments of human populations. Decreasing costs of high-resolution NGS-based HLA typing approaches— including phased data sets from long-read technologies—will further fuel the development of more comprehensive and even more accurate imputation reference panels.

Materials and Methods

Resolution of imputation reference panels

Several imputation references have been published in the past using various genotyping chips, allowing for the imputation of different HLA genes at different resolutions, i.e. full context four-digit (two-field), G group and P group resolution (as defined by the IMGT/HLA database) or custom groups (mostly before 2010). Full context four-digit levels provide information on the gene name, their allele group and the protein sequence of the HLA molecule (i.e. A*01:02—Gene: A; allele group: 01; protein: 02). Alleles that are within the same G group have identical nucleotide sequences for exons 2 and 3 (HLA class I) or exon 2 only (HLA class II) and may differ in sequence in the other exons. Alleles that are within the same P group encode for identical amino acid sequences in exons 2 and 3 or exon 2 only. P and G group annotations were introduced in 2010 and a major update in allele naming was conducted (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/Nomenclature_2009.txt), amongst others the separator ':' was introduced and alleles were renamed especially alleles of the *HLA-A*, *-B*, *-C* and *-DPB1* genes. Notably, HLA allele calling conducted before this time, with alleles typed only at exons 2 and 3 or exon 2, may not follow the known G group and P group conventions published by the IMGT/HLA, i.e. HLA alleles might be grouped in custom groups and some of the alleles will carry outdated allele names. This issue should be considered when merging reference panels, such that all included alleles should map to the same allele groups and also in benchmarking studies using external data. G grouping published by the IMGT/HLA database is based on the highest resolution that is recorded for an allele (i.e. eight digits or lower). Note that the post-calling G grouping based on four-digit alleles is problematic for some alleles listed in Supplementary Material, Table S11.

Cohorts & data preparation

Multi-ethnic data set. DNA of 96 healthy individuals and 96 UC patients were collected from different studies of Chinese, German, Indian, Iranian, Japanese, Korean and Maltese populations that have been published and described elsewhere (20,21). In short, Chinese samples were collected in and around Hong Kong (Chinese University of Hong Kong), Korean samples in South Korea (Yonsei University College of Medicine and Asan Medical Centre, Seoul), Japanese samples in Tokyo (Institute of Medical Science, University of Tokyo, RIKEN Yokohama Institute and Japan Biobank), Iranian samples were collected in Tehran (Tehran University of Medical Science), Indian samples in North India (Dayanand Medical College and Hospital, Ludhiana), all self-reported North Indian which was consistent with their genetically determined background, German samples in North Germany and Maltese samples in Malta (Department of Gastroenterology, Mater Dei Hospital, Msida, Malta). In addition to the data from the published UC studies, DNA samples were obtained from 192 healthy controls and 192 UC patients, all self-reported as African American, which was consistent with their genetically determined background as each had an admixture of West African and European ancestry (22). These subjects were recruited in the United States of America and Canada by the Johns Hopkins Multicenter African American IBD Study as well as other Genetics Research Centers of the NIDDK IBD Genetics Consortium. We also received 192 (96 healthy, 96 UC) pre-analyzed Japanese samples directly from RIKEN Yokohama Institute.

High density SNP-array data interrogating a wide proportion of the extended HLA region were produced for these samples using the Illumina, Immunochip (all but Malta) with 196,524 markers addressing immune relevant genes or the Illumina Infimum ImmunoArray 24 (Malta only) with 253,702 markers and subjected to strict quality control criteria as described in the Supplementary Material, Supplementary Methods. DNA was isolated and processed as described previously (10) in preparation for sequencing. Sequencing was performed on an Illumina HiSeq2500 (http://systems.illumina.com) with 100 bp or 125 bp paired-end runs on a panel of both case and control data in a pool of 96 libraries per lane. A total of 192 Japanese samples were provided by the RIKEN Yokohama Institute and sequenced using 125 bp paired-end runs on the HiSeq2500 with pools of 94 libraries per lane. Four-digit HLA alleles for all classical HLA I and HLA II genes *HLA-A, -B, -C, -DQA1, -DQB1, -DPA1, -DPB1, -DRB1* as well as *-DRB3/4/5* were manually curated and called using HLAssign (10). In short, only reads mapping exactly to a reference based on HLA sequences published with the IMGT/HLA database version 3.27.0 (23) were used for calling, taking into consideration evenness of read mapping, read equality and specific read mapping as described by Wittig *et al.* (10). We also cautiously looked at cross-mapping events (reads mapping to multiple HLA loci) and SNP patterns to identify e.g. alleles originating from concatenation of true alleles. In total 1,360 samples were used in this study, having been sequenced and called successfully based on their DNA quality and internal HLAssign measures, i.e. sufficiently large read coverage and also having passed our stringent criteria for the quality control of the Illumina Immunochip array data (Supplementary Material, Supplementary Methods). The *HLA-DRB3/4/5* calls were additionally evaluated for plausibility with respect to the called *HLA-DRB1* genotype. *HLA-DRB3/4/5* alleles, according to reported studies (14–16), occur on certain haplotypes in tight linkage with specific *HLA-DRB1* variants and can either be present or not present at all (i.e. null allele, described

as DRB3*00:00, DRB4*00:00 and DRB5*00:00 in the following) or as one functional *HLA-DRB3/4/5* allele in combination with two of the *HLA-DRB3/4/5* null alleles. For a detailed overview we compiled Figure 4. A total of 312 African American (158 Controls, 154 UC cases), 162 German (78 Controls, 84 Cases), 140 Chinese (68 Controls, 72 Cases), 143 Indian (78 Controls, 65 Cases), 132 Iranian (63 Controls, 69 Cases), 189 Japanese (96 Controls and 93 Cases), 122 South Korean (81 Controls and 41 Cases) and 160 Maltese (75 Controls and 85 Cases) samples were available for construction of HLA imputation models with HIBAG.

1000 Genomes data set. Using the Phase 3 [version from 20130502] 1000 Genomes reference data set (24) and Vcftools (version 0.1.12b), we extracted 174,538 phased SNPs that are present in both the Phase 3 data set and on the Illumina Immunochip used for the main part of our trans-ethnic data. We then performed quality control as described in the Supplementary Material, Supplementary Methods leaving out batch and population stratification analyses. HLA data were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_geotypes/. Publicly available data from the 1000 Genomes data set do not include *HLA-DPA1, -DPB1, -DQA1* and *DRB3/4/5* allele calls. In total 162 samples of African Ancestry, 193 samples of South American Ancestry, 260 samples of East Asian ancestry and 322 samples of European ancestry were available for construction of HLA imputation models with HIBAG. The HapMap data used in other studies (Table 4) are a part of the 1000 Genomes data set.

Calling of HLA-DRB3/4/5 alleles. Data were analyzed visually using HLAssign (10). HLAssign does not calculate phases of the HLA alleles and thus does not make hemizygous calls (i.e. recognize null alleles) such that *HLA-DRB3/4/5* genotypes were edited with respect to the *HLA-DRB1* allele post calling. For consistency with the *HLA-DRB3/4/5* with the literature (Fig. 3), we introduced null alleles DRB3*00:00, DRB4*00:00 or DRB5*00:00 when the *HLA-DRB1* locus was called as DRB1*01, DRB1*08 or DRB1*10, respectively. DRB3*00:00 was assigned if no *HLA-DRB3* was present in the corresponding *HLA-DRB1* haplotype. Equally, DRB4*00:00 and DRB5*00:00 were assigned if haplotypes corresponding to the absence of *HLA-DRB4* or *-DRB5* were called. Samples with inconclusive *HLA-DRB3/4/5* detected during HLAssign analysis were re-analyzed using HLAReporter (25). HLAReporter performs *de novo* assembly on the NGS reads within the investigated HLA locus using the alignment tool TASR (26) and compares these to either G groups or full context alleles known in the IMGT/HLA database with the parameters (-m 50, -o 5, -r 0.7, -u 0, -i 1, -t 0, -e 33, -c 0) for on target reads. Contigs for samples with equal G group predictions were aligned against each other to generate longer overlapping regions using contigs with a coverage higher than 15 and then realigned to the known IMGT/HLA reference alleles.

MDS analysis. Relative allele frequencies were calculated for each allele across the entire multi-ethnic and 1000 Genomes HLA data within the *HLA-A, -B, -C, -DQ* and *-DR* loci. For the MDS analysis alleles with an allele frequency of less than 1% in any subpopulation are excluded to avoid a clustering biased by similarity in low frequency variants. The MDS analysis was performed using R and the stats-Package (cmdscale) with a Euclidean distance measure. For the MDS analysis across all loci we used HLA loci *HLA-A, -B, -C, -DQB1* and *-DRB1*.

HLA imputation benchmark

Training of the reference panel. We performed HLA imputation using the published imputation tool HIBAG (6). This is a machine learning tool implemented in R that employs ensemble classifiers built on bootstrap samples that has been shown to perform with high accuracy in HLA imputation across multi-ethnic data sets (6). In short, a training set with both HLA alleles and SNPs typed in the HLA region on chromosome 6, between 29 and 34 Mb, is used to build several classifiers based on bootstrap samples and a subset of SNPs, similarly to random forest as proposed by Breiman *et al.* (27) that minimize the out-of-bag errors. Once a model is trained, it can be used as reference to predict HLA alleles from unknown samples using their respective SNP genotype information, utilizing the posterior probability as measure of confidence. For the benchmark, we performed a 5× cross-validation using HIBAG (6) and HLA and SNP genotype data from the following two sources: our multi-ethnic cohort described above and the publicly available 1000 Genomes data set (24). The 1000 Genomes data set was typed for *HLA-A, -B, -C, -DPB1* and *-DRB1*, while the multi-ethnic data set contained all classical HLA class I and class II loci and additionally *HLA-DRB3/4/5*. For the 1000 Genomes data set, typed HLA data were available for samples of the following ethnicities: African, South American Ancestry, East Asian and European. We grouped our data into three different data sets: (i) our multi-ethnic reference containing eight different cohorts described above, (ii) the same reference as in (i) with HLA alleles transformed into their respective G groups (G groups combine alleles with identical exon 2 and 3 (HLA Class I) or exon 2 (HLA Class II) nucleotide sequence) using hla_nom_g.txt downloaded from hlaalleles.org date: 2017-07-10, IPD-IMGT/HLA version 3.29.0) and (iii) our multi-ethnic panel and the 1000 Genomes data set combined. In total we used 1,360 samples and 7,428 SNPs within the HLA region for the multi-ethnic reference, as well as 937 samples from the 1000 Genomes data and 7,551 SNPs within the HLA region from the 1000 Genomes data set, with 2,297 samples and 7,126 SNPs for the combined data set as well as their respective HLA calls. For the 1000 Genomes panel, we checked for nomenclature issues, making sure that all of the HLA alleles used in the 1000 Genomes panel mapped to the nomenclature for HLA alleles used since April 2010 (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/Nomenclature_2009.txt). For alleles with unambiguous G groups (Supplementary Material, Table S11), we assigned the lower number allele for reference panels (ii) and (iii). Genotype data were prepared as described in Supplementary Material, Supplementary Methods. Samples with typed HLA information were extracted from each quality-controlled, genotyped data set. The different cohorts were merged and those SNPs with a consistent minor allele frequency (MAF) of <1% (across all cohorts typed for the particular SNP) were excluded. The data were randomly split into five equal parts per cohort with respect to case–control status, thus ensuring that a training set would include both case and control data. Using HIBAG (version.1.8.3), we trained our models using the reference containing the merged subpopulations, excluding 20% of the population of interest and 100 classifiers, as suggested by the authors of the tool (Supplementary Material, Fig. S1).

Validation of the reference panel. The quality-controlled genotype data for each cohort were imputed using Beagle version 4.1 (28) with the cohort itself serving as an internal reference to fill in any remaining missing data. Pretrained HIBAG HLA models (see above) were provided with the respective 20% of the remaining data of each analyzed population (Supplementary Material, Fig. S1), using the genomic position as the identifier. HLA calls were calculated and stored with their respective posterior probabilities. Accuracies and the number of samples to be excluded were calculated for different posterior probability thresholds and compared between the different populations.

Calculation of accuracies. Imputation accuracies were calculated on best-guess alleles compared with the known alleles of the typed data. Accuracies for best-guess alleles were calculated by counting the number of alleles imputed correctly per locus and dividing by the number of samples multiplied by two. Per locus and per allele accuracies were evaluated. We also calculated single allele specificity and sensitivity values if possible. For this we evaluated each allele separately, counting the number of times an allele was predicted correctly as present (True Positive; TP) or absent (True Negative; TN) and the number of times an allele was incorrectly predicted as present (False Positive; FP) or absent (False Negative; FN). We then used the standard definitions to calculate sensitivity and specificity from these values.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

For the calculation of the accuracy, specificity and sensitivity values within the cross-validation, the mean values across the different runs were calculated for each locus or allele, as well as median, minimum and maximum values for comparison. To establish which alleles might have low sensitivity and specificity values in a general setting for (i), we calculated these measures using a model based on the entire population (i).

Imputation reference panels for comparison

A Caucasian reference panel based on genotypes retrieved from the T1DGC (29), as well as a Pan Asian data set (30) using three different Asian populations, were published along with SNP2HLA (7) and are available on request from the SNP2HLA authors. Here, loci *HLA-A, -B, -C, -DQA1, -DQB1, -DPB1* and *-DRB1* were typed (Table 4a). Two additional Asian reference panels based on SNP2HLA were published at a four-digit resolution. First, a Korean reference panel was published in 2014 (13) for the imputation of amino acids and HLA alleles into East Asian populations for *HLA-A, -B, -C, -DQB1, -DPB1* and *-DRB1* and second, a Japanese reference data set was published in 2015 by Okada *et al.* (12) with an evaluation of loci *HLA-A, -B, -C, -DQB1* and *-DRB1*. For these two last reference panels, we assume that they were typed at a full context four-digit resolution. This has not been explicitly mentioned in the respective publications (12,13), but we find that the typed alleles best fit to the full four-digit context based on which alleles are present. Pre-trained multi-ethnic HLA models with European, Asian, Hispanic and African ancestry (based on a total of 3,738 samples) are provided with the HLA imputation tool HIBAG (6). The samples used for these models were obtained from HLARES (samples GlaxoSmithKline clinical trials) (6) and the HapMap project. Loci *HLA-A, -B, -C, -DQA1, -DQB1, -DPB1* and *-DRB1* were evaluated at four-digit resolution (Table 4b). The remaining considered reference panels based on HLA*IMP:02 (8) are based on HLARES data and a study specific "Golden Set" (GS) (Table 4c).

Availability of resources

The herein-described reference data sets are available on request from the authors (email contact: f.degenhardt@ikmb.uni-kiel.de) as pretrained HIBAG models and are mapped to IMGT/HLA database version 3.27.0 with G group definitions derived from IMGT/HLA database version 3.29.0. Note that allele names at four-digit levels did not change between these two releases. The training of these models was performed as described above without exclusion of any samples. A script that will estimate the haplotype similarity between alleles based on the genotype positions available in a data set is also available upon request.

Supplementary Material

Supplementary Material is available at *HMG* online.

References

1. Rose, N.R. (1978) HLA and disease. *Arch. Intern. Med.*, **138**, 527–528.
2. Dendrou, C.A., Petersen, J., Rossjohn, J. and Fugger, L. (2018) HLA variation and disease. *Nat. Rev. Immunol.*, **18**, 325–339.
3. Okada, Y., Yamazaki, K., Umeno, J., Takahashi, A., Kumasaka, N., Ashikawa, K., Aoi, T., Takazoe, M., Matsui, T., Hirano, A. *et al.* (2011) HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology*, **141**(864–871), e861–865.
4. Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E.S., Annese, V., Hauser, S.L. *et al.* (2015) High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.*, **47**, 172–179.
5. Patsopoulos, N.A., Barcellos, L.F., Hintzen, R.Q., Schaefer, C., van Duijn, C.M., Noble, J.A., Raj, T., IMSGC, ANZgene, Gourraud, P.A. *et al.* (2013) Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLoS Genet.*, **9**, e1003926.
6. Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R. and Weir, B.S. (2014) HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.*, **14**, 192–200.
7. Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.M., Concannon, P.J., Rich, S.S., Raychaudhuri, S. and de Bakker, P.I. (2013) Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*, **8**, e64683.
8. 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.
9. Franke, A. (2017) Inflammatory bowel disease: a global disease that needs a broader ensemble of populations. *Gastroenterology*, **152**, 14–16.
10. Wittig, M., Anmarkrud, J.A., Kassens, J.C., Koch, S., Forster, M., Ellinghaus, E., Hov, J.R., Sauer, S., Schimmler, M., Ziemann, M. *et al.* (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.*, **43**, e70.
11. Gourraud, P.A., Khankhanian, P., Cereb, N., Yang, S.Y., Feolo, M., Maiers, M., Rioux, J.D., Hauser, S. and Oksenberg, J. (2014) HLA diversity in the 1000 genomes dataset. *PLoS One*, **9**, e97282.

12. Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., Takahashi, A. and Kubo, M. (2015) Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nat. Genet.*, **47**, 798–802.

13. Kim, K., Bang, S.Y., Lee, H.S. and Bae, S.C. (2014) Construction and application of a Korean reference panel for imputing classical alleles and amino acids of human leukocyte antigen genes. *PLoS One*, **9**, e112546.

14. Holdsworth, R., Hurley, C.K., Marsh, S.G., Lau, M., Noreen, H.J., Kempenich, J.H., Setterholm, M. and Maiers, M. (2009) The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens*, **73**, 95–170.

15. Robbins, F., Hurley, C.K., Tang, T., Yao, H., Lin, Y.S., Wade, J., Goeken, N. and Hartzman, R.J. (1997) Diversity associated with the second expressed HLA-DRB locus in the human population. *Immunogenetics*, **46**, 104–110.

16. Bontrop, R.E., Otting, N., de Groot, N.G. and Doxiadis, G.G. (1999) Major histocompatibility complex class II polymorphisms in primates. *Immunol. Rev.*, **167**, 339–350.

17. Gongora, R., Figueroa, F. and Klein, J. (1996) The HLA-DRB9 gene and the origin of HLA-DR haplotypes. *Hum. Immunol.*, **51**, 23–31.

18. Kumasaka, N., Okada, Y., Takahashi, A., Kubo, M., Nakamura, Y. and Kamatani, N. (2011), In 12th International Congress of Human Genetics/61st Annual Meeting of The American Society of Human Genetics, Montreal, Canada, Abstract/Program #708F. http://kumasakanatsuhiko.jp/projects/disentangler/.

19. Gonzalez-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H., da Silva, A.L., Teles e Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R. *et al.* (2015) Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.*, **43**, D784–788.

20. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T. *et al.* (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, **47**, 979–986.

21. Brant, S.R., Okou, D.T., Simpson, C.L., Cutler, D.J., Haritunians, T., Bradfield, J.P., Chopra, P., Prince, J., Begum, F., Kumar, A. *et al.* (2017) Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease. *Gastroenterology*, **152**(206–217), e202.

22. Huang, C., Haritunians, T., Okou, D.T., Cutler, D.J., Zwick, M.E., Taylor, K.D., Datta, L.W., Maranville, J.C., Liu, Z., Ellis, S. *et al.* (2015) Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in African Americans. *Gastroenterology*, **149**, 1575–1586.

23. Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P. and Marsh, S.G. (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.*, **43**, D423–D431.

24. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

25. Huang, Y., Yang, J., Ying, D., Zhang, Y., Shotelersuk, V., Hirankarn, N., Sham, P.C., Lau, Y.L. and Yang, W. (2015) HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med.*, **7**, 25.

26. Warren, R.L. and Holt, R.A. (2011) Targeted assembly of short sequence reads. *PLoS One*, **6**, e19816.

27. Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

28. Browning, B.L. and Browning, S.R. (2016) Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.*, **98**, 116–126.

29. Mychaleckyj, J.C., Noble, J.A., Moonsamy, P.V., Carlson, J.A., Varney, M.D., Post, J., Helmberg, W., Pierce, J.J., Bonella, P., Fear, A.L. *et al.* (2010) HLA genotyping in the international Type 1 Diabetes Genetics Consortium. *Clin. Trials*, **7**, 75–87.

30. Pillai, N.E., Okada, Y., Saw, W.Y., Ong, R.T., Wang, X., Tantoso, E., Xu, W., Peterson, T.A., Bielawny, T., Ali, M. *et al.* (2014) Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Hum. Mol. Genet.*, **23**, 4443–4451.