Human Molecular Genetics, 2014, Vol. 23, No. 16 **4443–4451**
doi:10.1093/hmg/ddu149
Advance Access published on April 3, 2014

Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations

Nisha Esakimuthu Pillai[1], Yukinori Okada[3,4,5], Woei-Yuh Saw[1], Rick Twee-Hee Ong[1], Xu Wang[1], Erwin Tantoso[1], Wenting Xu[1], Trevor A. Peterson[6,7], Thomas Bielawny[6,7], Mohammad Ali[8], Koon-Yong Tay[8], Wan-Ting Poh[1], Linda Wei-Lin Tan[1], Seok-Hwee Koo[2], Wei-Yen Lim[1], Richie Soong[10], Markus Wenk[11,12], Soumya Raychaudhuri[3,4,5,13], Peter Little[8], Francis A. Plummer[6,7], Edmund J. D. Lee[2], Kee-Seng Chia[1], Ma Luo[6,7], Paul I. W. De Bakker[5,13] and Yik-Ying Teo[1,8,9,14,15,*]

[1]Saw Swee Hock School of Public Health, [2]Department of Pharmacology, National University of Singapore, Singapore 117597, Singapore, [3]Division of Rheumatology, Immunology, and Allergy, [4]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [5]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA, [6]Department of Medical Microbiology, University of Manitoba, 730 William Avenue, Winnipeg, Manitoba R3E 0Z2, Canada, [7]National Microbiology Laboratory, Winnipeg, Manitoba, Canada, [8]Life Sciences Institute, [9]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore, [10]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore, [11]Department of Biochemistry, National University of Singapore, Singapore 117596, Singapore, [12]Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore, [13]Centre for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114-2790, USA, [14]NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore and [15]Agency for Science, Technology and Research, Genome Institute of Singapore, Singapore 138672, Singapore

Received March 3, 2014; Revised March 3, 2014; Accepted March 31, 2014

The major histocompatibility complex (MHC) containing the classical human leukocyte antigen (HLA) Class I and Class II genes is among the most polymorphic and diverse regions in the human genome. Despite the clinical importance of identifying the HLA types, very few databases jointly characterize densely genotyped single nucleotide polymorphisms (SNPs) and HLA alleles in the same samples. To date, the HapMap presents the only public resource that provides a SNP reference panel for predicting HLA alleles, constructed with four collections of individuals of north-western European, northern Han Chinese, cosmopolitan Japanese and Yoruba Nigerian ancestry. Owing to complex patterns of linkage disequilibrium in this region, it is unclear whether the HapMap reference panels can be appropriately utilized for other populations. Here, we describe a public resource for the Singapore Genome Variation Project with: (i) dense genotyping across ∼9000 SNPs in the MHC; (ii) four-digit HLA typing for eight Class I and Class II loci, in 96 southern Han Chinese, 89 Southeast Asian Malays and 83 Tamil Indians. This resource provides population estimates of the frequencies of HLA alleles at these eight loci in the three population groups, particularly for *HLA-DPA1* and *HLA-DPB1* that were not assayed in HapMap. Comparing between population-specific reference panels and a cosmopolitan panel created from all four HapMap populations, we demonstrate that more accurate imputation is obtained with population-specific panels than with the cosmopolitan panel, especially for the Malays and Indians but even when imputing between northern and southern Han Chinese. As with SNP imputation, common HLA alleles were imputed with greater accuracy than low-frequency variants.

*To whom correspondence should be addressed at: School of Public Health, MD3 16 Medical Drive, National University of Singapore, Singapore 117597. Tel: +65 65162760; Fax: +65 68723919; Email: statyy@nus.edu.sg

INTRODUCTION

The major histocompatibility complex (MHC) is a region of ~4 Mb in size on chromosome 6 of the human genome that spans over 160 genes including the classical human leukocyte antigen (HLA) Class I and Class II genes. With a significantly higher single nucleotide polymorphisms (SNP) density than most regions, the MHC is among the most polymorphic regions in the human genome (1), and exhibits considerable diversity between populations (Supplementary Material, Table S1). With a significant fraction of the HLA genes encoding proteins that are involved in the immune system and self versus non-self-autoimmune responses (2), the degree of HLA matching is an important predictor of transplant rejection and genome-wide association studies have implicated the HLA genes in numerous infectious and autoimmune diseases (3–10). Several HLA alleles have also been identified to be strongly associated with increased susceptibility to adverse reactions to particular drugs (11,12), such as the B*57 : 01 allele for abacavir hypersensitivity (13) and the B*15 : 02 allele for carbamazepine-induced Stevens–Johnson syndrome (14).

Despite the clinical importance of identifying and matching the HLA types, there are very few databases that are dedicated to characterizing the HLA alleles across multiple populations. This is partly due to the high costs of determining the HLA types, especially for high-resolution typing across multiple HLA loci that are necessary for bone marrow or stem cell transplants. The prospect of statistically inferring the HLA alleles from genotype data is promising and cost-effective, and several statistical methods have been developed to leverage on HLA reference panels built by jointly characterizing SNPs and HLA loci for the same set of samples (15–19) such as the International HapMap Project (HapMap) (20) and the MHC Working Group of the Type 1 Diabetes Genetics Consortium (21). In the second phase of the HapMap project, dense genotyping yielded a database of almost 2.4 million SNPs and HLA typing was performed across six HLA loci (*-A, -B, -C, -DQA1, -DQB1, -DRB1*) for 301 samples from four ancestry groups: (i) Utah residents with northern and western European ancestry; (ii) Han Chinese in Beijing, China; (iii) Japanese in Tokyo, Japan and (iv) Yoruba in Ibadan, Nigeria.

Statistical imputation of the HLA types from genetic data fundamentally relies on matching the patterns of linkage disequilibrium (LD) that are found in the target data with those present in the reference panel, and variations in the extent or patterns of such genetic correlation can confound this analysis (22). Several studies have reported complex patterns of LD across the MHC region that differ significantly between populations (23–26), with reports of differential evidence of positive selection at specific HLA loci even between closely related populations such as northern and southern Han Chinese (27). It is thus unlikely that the reference panels built with the four populations in the HapMap resource can be universally representative of other global populations.

Here, we introduce another HLA resource built from the Singapore Genome Variation Project (SGVP) (28) which genotyped 268 subjects from three population groups in Southeast Asia on the Illumina 1 M and Affymetrix 6.0 microarrays, yielding a genotype database of ~9000 SNPs in the 25–35 Mb region in the MHC. High-resolution sequence typing using sequence-based typing and taxonomy-based sequence analysis was performed across three loci in Class I (*-A, -B, -C*) and five loci in Class II (*-DPA1, -DPB1, -DQA1, -DQB1, -DRB1*), respectively, to produce four-digit HLA alleles for each of the 268 subjects, which permitted the construction of three additional HLA reference panels from the 96 southern Han Chinese, 89 Southeast Asian Malays and 83 Tamil Indians from Singapore. Using an independent set of samples from the three population groups, we compared the accuracy of imputing the HLA types using the HapMap and the SGVP samples. In particular, we combined all four HapMap panels into one cosmopolitan reference panel and assessed how well such a cosmopolitan panel will perform in imputing the HLA alleles for populations such as the Malays and the Indians that are considerably different from the HapMap populations. We also evaluated the accuracy of the imputation in the southern Han validation samples when imputed with either the HapMap northern Chinese reference panel or the SGVP southern Chinese panel. We expect our findings will provide a benchmark for deciding on the appropriate HLA reference panels to impute against, especially for Asian populations with ancestry similar to those of southern Chinese, Malays and Tamil Indians. The SGVP HLA resource is made publicly available at http://www.statgen.nus.edu.sg/~SGVP/hla.html.

RESULTS

The SGVP created a SNP resource for 96 southern Han Chinese (CHS), 89 Southeast Asian Malays (MAS) and 83 Tamil Indians (INS) from Singapore by genotyping each sample on the Illumina 1 M and Affymetrix 6.0 microarrays, producing at least 1.5 million SNPs for each population group. This included ~9000 SNPs that were present in all three populations in the MHC region between 25 and 35 Mb on chromosome 6. HLA typing was performed in all 268 subjects to yield at least four-digit resolution HLA alleles for *HLA-A*, *HLA-B* and *HLA-C* in Class I, and *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* in Class II. By phasing the SNP genotype data together with the HLA alleles in each population separately with BEAGLE, three population-specific haplotype reference panels were created for the purpose of imputing HLA alleles at these eight loci. We similarly constructed four reference panels for the populations in Phase 2 of the HapMap by integrating the resource from at least 16 000 SNPs in the MHC with six HLA loci (*-A, -B, -C, -DQA1, -DQB1, -DRB1*) in the Caucasian (CEU), Nigerian African (YRI), Han Chinese (CHB) and Japanese (JPT) samples. A cosmopolitan reference panel was constructed by combining all four population-specific panels for HapMap.

Comparing the HLA alleles between SGVP and HapMap

We observed that *HLA-B* is the most polymorphic locus out of the eight HLA loci, where the number of alleles ranged from 30 in CHS to 40 in MAS. This concurred with the findings from HapMap that *HLA-B* was the most polymorphic loci, with 28 alleles in CHB and 56 alleles in the combined cosmopolitan panel. The number of HLA alleles present across the eight HLA loci was similar between the CHS and MAS samples, while the INS samples exhibited more alleles in all loci except at *DPA1* and *DRB1* (Table 1). However, there were significant

Table 1. Overview of the samples and number of HLA alleles

	SGVP			HapMap		iOmics validation		
	CHS	MAS	INS	CHB	Cosmo[a]	Chinese	Malay	Indian
No. of samples	96	89	83	54	301	20	20	20
No. of SNPs	10 391	10 381	10 869	16 012	18 015	26 376	26 376	26 376
HLA locus	Number of distinct four-digit HLA alleles							
A	18	18	23	14	32	10	10	11
B	30	40	34	28	56	17	18	16
C	18	17	20	13	26	12	11	14
DPA1[b]	5	5	4	—	—	4	4	4
DPB1[b]	18	15	19	—	—	10	10	8
DQA1[c]	3	3	3	8	8	7	7	7
DQB1[d]	12	11	9	13	17	9	10	11
DRB1[e]	7	7	3	23	42	16	12	14

[a]Refers to a cosmopolitan panel obtained by combining the four population-specific HLA reference panels for CEU, CHB, JPT and YRI.
[b]No information is available for the HapMap samples as these two loci were not surveyed.
[c]Only 47 CHS, 66 MAS and 88 INS chromosomes had unambiguous four-digit HLA alleles for this locus.
[d]Only 107 CHS, 91 MAS and 103 INS chromosomes had unambiguous four-digit HLA alleles for this locus.
[e]Only 20 CHS, 27 MAS and 31 INS chromosomes had unambiguous four-digit HLA alleles for this locus.

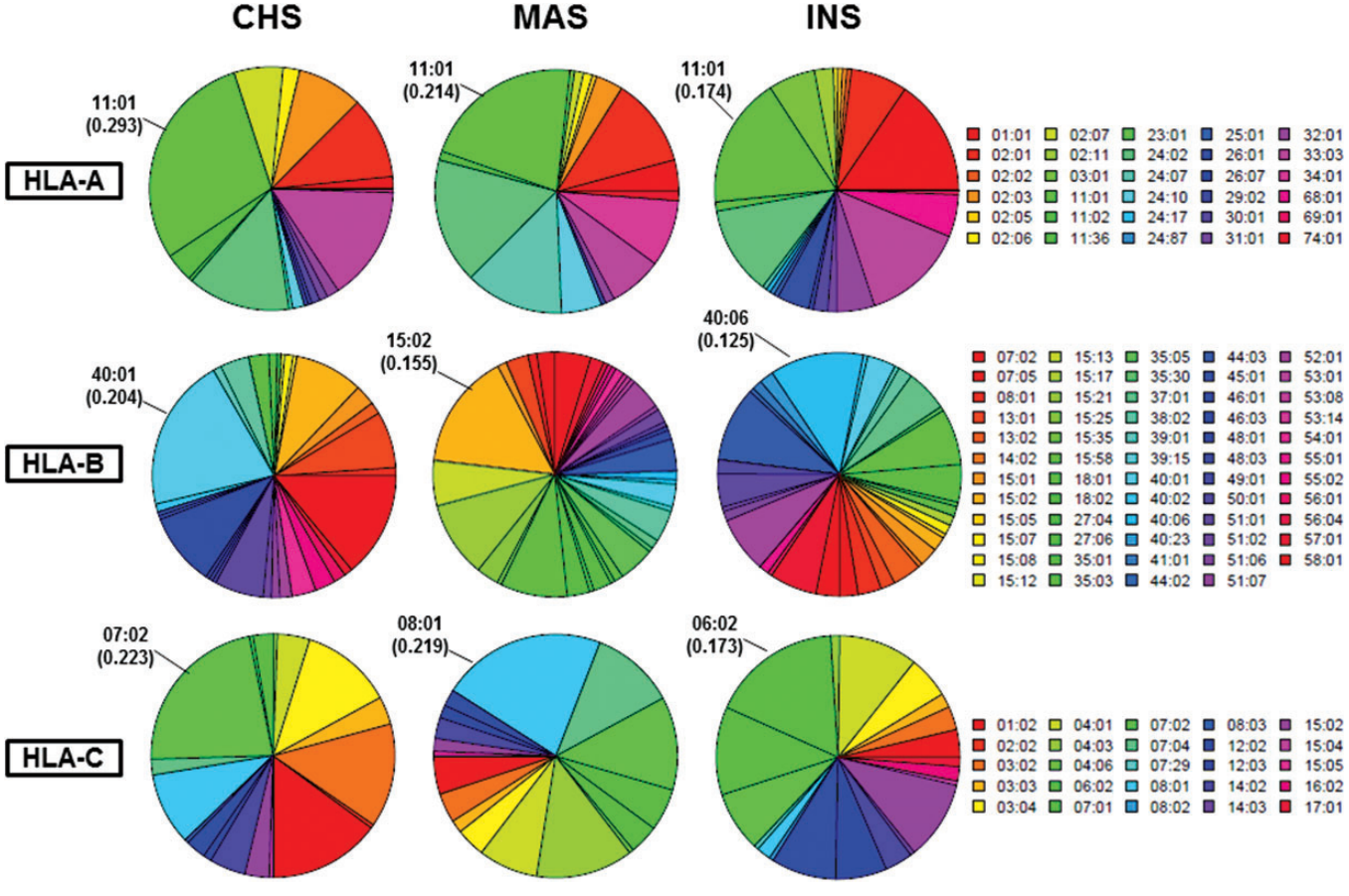

Figure 1. Allelic diversity and distribution in HLA Class I genes. Pie charts illustrating the allelic diversity of the three genes in HLA Class I in the three populations, corresponding to southern Han Chinese (CHS), Southeast Asian Malays (MAS) and Tamil Indians (INS) in Singapore. The frequency distribution of the alleles in each population is represented by the area of the segments in the respective pie chart. Estimates of the population-specific HLA allele frequencies can be found in Supplementary Material, Table S1.

differences in the distributions of the alleles within each locus (Fig. 1 and Supplementary Material Fig. S1 and Table S2). For example, while A*11:01 was the most frequently occurring allele in *HLA-A* in all three SGVP populations, different alleles were found to be most common in *HLA-B* with B*40:01 in CHS at a frequency of 20.4%, B*15:02 in MAS at 15.5% and

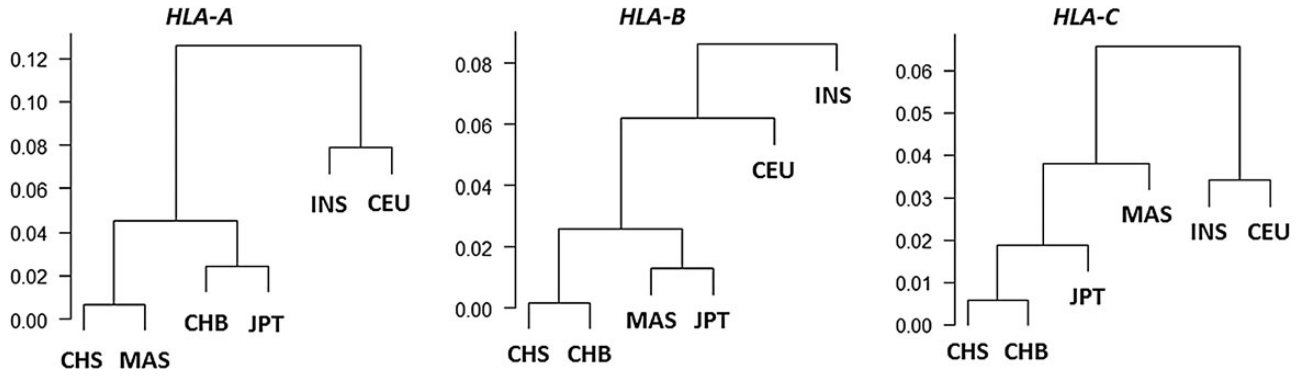

Figure 2. Population dendrograms for HLA Class I genes. Genetic distance between every pair of populations for each of the three HLA Class I loci is calculated using the multi-allelic F_{ST}, and this distance is used to perform an agglomerative hierarchical clustering of the six populations from SGVP (CHS, MAS, INS) and from the HapMap (CEU, CHB, JPT). The dendrogram clustering trees display the results of the agglomerative clustering, where the vertical coordinate where two branches join provides a measure of the dissimilarity between the two clusters.

B*40:06 in INS at 12.5%. In fact, other than *DQB1* where DQB1*06:01 was most common in all three populations, different alleles were observed to be most frequent in the three populations for the remaining five HLA loci. The number of alleles across *HLA-A*, *HLA-B*, *HLA-C* and *HLA-DQB1* was consistent between the SGVP and HapMap populations, although there were less number of alleles reported in the SGVP populations for *HLA-DQA1* and *HLA-DRB1* due to the greater extent of missing or ambiguous HLA data in the SGVP samples (Table 1). No comparison could be made for *DPA1* and *DPB1* as HapMap did not survey these two loci.

We investigated the diversity of the Class I HLA genes across the three SGVP populations (CHS, MAS and INS) and three HapMap populations (CEU, CHB and JPT) by calculating the multi-allelic F_{ST} between every pair of populations. The choice of the populations was motivated by the interest to compare between: (i) the southern Chinese (CHS) with the northern Chinese (CHB) and another East Asian population (JPT); (ii) the Europeans (CEU) with the Tamil Indians (INS) given previous reports that the Indians were genetically closer to the Europeans than to the Chinese and Malays. The results of the F_{ST} estimation at each Class I gene were illustrated with a population dendrogram obtained through agglomerative hierarchical clustering (Fig. 2). While the relationships between the six populations were broadly consistent across all three loci, subtle variations existed that may be insightful to the representativeness, or lack thereof, of the HapMap HLA data for the SGVP populations. For example, while the two Han Chinese cohorts were most homogeneous of the six populations for *HLA-B* and *HLA-C* ($F_{ST} = 0.19$ and 0.59%, respectively, Supplementary Material, Table S3), the southern Chinese (CHB) was found to be closer to the Malays (MAS, $F_{ST} = 0.69\%$) than to the northern Chinese ($F_{ST} = 1.22\%$) at *HLA-A*. The Tamil Indians (INS) were generally found to be closer to the Europeans (CEU) than Chinese and Malays (F_{ST} between CEU and INS for *HLA-A* = 7.35%, *HLA-B* = 5.31%, *HLA-C* = 5.48%, compared with F_{ST} between CEU and CHS for *HLA-A* = 13.03%, *HLA-B* = 7.49%, *HLC-C* = 6.97%) although the magnitude of the F_{ST} estimates suggested that considerable diversity likely existed in the HLA profiles of these Class I genes between Europeans and Indians (Supplementary Material, Table S3).

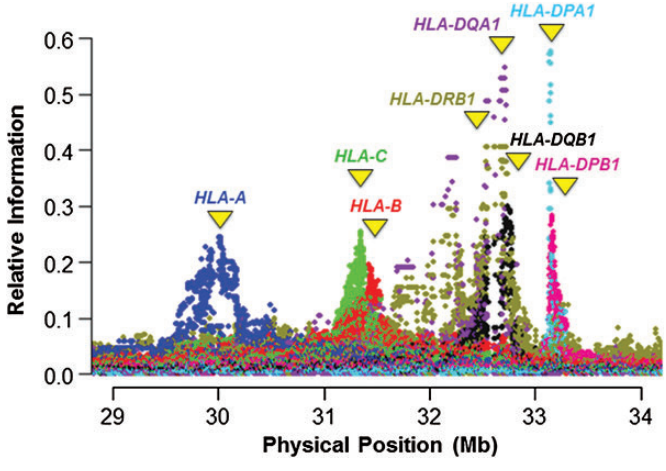

Figure 3. Association between SNPs and HLA alleles. Statistical association between HLA alleles in the eight Class I and Class II loci, and SNPs present in the SGVP between 29.0 and 34.0 Mb on chromosome 6. Every SNP is assessed for the ability to predict the alleles at each HLA locus, which is measured by relative information (see Materials and Methods). The assessment for three loci in Class II (*HLA-DQA1*: purple; *HLA-DQB1*: black; *HLA-DRB1*:olive) is noisier as exhibited by the greater spread due to greater degree of missingness in the HLA typing, compared with the almost complete typing for the Class I loci (*HLA-A*: blue; *HLA-B*: red; *HLA-C*: green) and two Class II loci (*HLA-DPA1*: cyan; *HLA-DPB1*: magenta).

Tagging HLA alleles with SNPs

Joint phasing of the HLA alleles and the SNPs allowed the assessment of the correlation between particular combinations of SNP variants with the HLA alleles, thus quantifying the ability for particular sets of SNPs to tag the HLA alleles. In general, the HLA alleles were tagged by SNPs that were located in the vicinity of the respective genes (Fig. 3), with 46.8% (CHS),

48.3% (MAS) and 56.5% (INS) of the HLA alleles exhibiting a correlation of $r^2 \geq 0.80$ with up to three SNPs (Supplementary Material, Table S4). The SNPs identified exhibited different tagging abilities in terms of: (i) the extent of correlation between the same SNP identified to be tagging a particular HLA allele in all three populations; (ii) the set of SNPs identified to be associated with the HLA allele in each population. An

example of the former is the allele HLA-B*57:01 which was perfectly tagged by the guanine allele of rs2395029 ($r^2 = 1$) in CHS and INS, but the same SNP only had a tagging efficiency of $r^2 = 0.746$ in MAS and required an additional SNP rs4418214 for the guanine–cytosine haplotype of the two SNPs to achieve perfect tagging; an example of the latter is allele HLA-B*44:03 which is tagged (with $r^2 \geq 0.80$) by seven SNPs in CHS but was instead tagged by an entirely discordant set of 27 SNPs in MAS.

Predicting HLA types in independent sample sets

To assess the performance of the SGVP reference panels in imputing the HLA alleles, we performed a validation experiment with 120 samples from the Singapore Integrative Omics Study (iOmics), where 40 individuals from each of the same three population groups in the SGVP have been (i) genotyped on the Illumina HumanOmni2.5 and Illumina Exome microarrays and (ii) assayed for the same eight HLA loci in Class I and Class II. This presented a denser panel of 23 451 SNPs across the 25 and 35 Mb region of chromosome 6, which we used as input to predict the alleles at the eight HLA loci for the 120 samples using the population-specific and cosmopolitan reference panels that we have constructed. The predicted four-digit HLA alleles were then compared against the experimentally determined four-digit HLA types to evaluate the error rates of the imputation. *In silico* HLA alleles were similarly predicted using the HapMap reference panels, where we considered the CHB and cosmopolitan panels for imputing the 40 southern Chinese samples, and only the cosmopolitan panels for imputing the Malays and the Indians samples.

For the southern Chinese samples, the HapMap cosmopolitan panel obtained similarly or more accurate four-digit imputation as the population-specific CHS panel for six loci (*-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*) (Table 2). No comparisons can be made for *DPA* and *DPB* as HapMap did not assay these two Class II loci. Comparing between the SGVP and HapMap cosmopolitan panels, the SGVP panel produced more accurate imputation for all three Class I loci, especially at *HLA-B* where the error rate decreased from 35% with the HapMap panel to 23% with the SGVP panel. As expected, the best imputation performance was observed at the Grand cosmopolitan panel that combined both the SGVP and HapMap cosmopolitan panels. The northern Chinese panel (CHB) was inferior to the southern Chinese panel (CHS) for all three Class I loci but outperformed the CHS panel for the three Class II loci although this was likely an effect of the small number of samples with available data to construct the SGVP panels at *-DQA1*, *-DQB1*, *-DRB1* (see Table 1). For the Malays, the HapMap cosmopolitan panel was clearly inappropriate with error rates of 55, 65 and 14% for *HLA-A*, *HLA-B* and *HLA-C*, respectively, compared with the much lower error rates of 20, 34 and 4% when imputed using the MAS panel (Table 2). The SGVP cosmopolitan panel produced even better performance with error rates of 11, 28 and 4%, respectively, although surprisingly the Grand cosmopolitan panel produced higher error rates than the SGVP panel. The population-specific INS panel yielded better performance than the HapMap cosmopolitan panel for the polymorphic *HLA-B* (25 versus 48%, respectively), although the HapMap cosmopolitan panel had lower error rates for both *HLA-A* and *HLA-C* (11 and 16% for HapMap, 18 and 29% for INS, respectively). The SGVP panel offered the lowest error rates at both *HLA-A* and *HLA-B*, although the Grand cosmopolitan panel performed the best for *HLA-C*. The HapMap or Grand cosmopolitan panels consistently performed better than the population-specific or SGVP panels for the three Class II loci, since the number of alleles in the SGVP populations for these loci was considerably smaller than the number of alleles in the target iOmics data set due to poor assays for most of the samples at these three loci.

Imputation accuracy by HLA allele frequency

In order to understand how imputation accuracy varies with the frequency of the HLA alleles in the populations, we binned all the HLA alleles observed across the eight loci in the iOmics samples into two categories: (i) low-frequency, defined as a HLA allele frequency of between 0 and 10% and (ii) common, defined as a HLA allele frequency of $>10\%$. Regardless of the choice of the reference panel, common HLA alleles were imputed with greater accuracy than low-frequency variants,

Table 2. Four-digit HLA imputation error rates for 120 iOmics samples when imputed using SGVP and HapMap reference panels

HLA locus	iOmics southern Chinese					iOmics Malays				iOmics Indians			
	CHS	CHB	SGVP[a]	Cosmo[b]	Grand[c]	MAS	SGVP[a]	Cosmo[b]	Grand[c]	INS	SGVP[a]	Cosmo[b]	Grand[c]
A	0.26	0.28	0.14	0.23	**0.04**	0.20	**0.11**	0.55	0.26	0.18	**0.09**	0.11	0.13
B	0.35	0.49	0.23	0.35	**0.16**	0.34	**0.28**	0.65	0.29	0.25	**0.11**	0.48	0.13
C	0.13	0.39	0.05	0.10	**0.04**	**0.04**	**0.04**	0.14	**0.04**	0.29	0.19	0.16	**0.11**
DPA1[d]	**0.01**	—	**0.01**	—	**0.01**	0.04	**0.01**	—	**0.01**	0.04	**0.01**	—	**0.01**
DPB1[d]	0.08	—	**0.04**	—	**0.04**	**0.08**	0.11	—	0.11	0.16	**0.11**	—	**0.11**
DQA1	0.75	0.09	0.73	0.09	**0.08**	0.68	0.64	**0.13**	0.40	0.48	0.56	0.33	**0.21**
DQB1	0.55	0.36	0.51	**0.21**	0.24	0.73	0.64	**0.19**	0.43	0.40	0.39	0.54	**0.20**
DRB1	0.90	0.43	0.76	0.29	**0.21**	0.86	0.46	**0.29**	0.18	0.66	0.65	0.41	**0.38**

The error rate corresponding to the reference panel(s) with the most accurate imputation for each HLA locus in each population is highlighted in bold. The error rates were calculated using 40 iOmics samples in each of the three population groups.

[a]SGVP refers to combined panel constructed by merging the three SGVP panels for Singapore Chinese (CHS), Singapore Malay (MAS) and Singapore Indian (INS).

[b]Cosmo refers to the combined panel constructed by merging the four HapMap panels for Europeans (CEU), northern Chinese (CHB), Japanese (JPT) and Nigerian Africans (YRI).

[c]Grand refers to the combined panel constructed by merging **SGVP[1]** and **Cosmo[2]**.

[d]HapMap did not assay this locus.

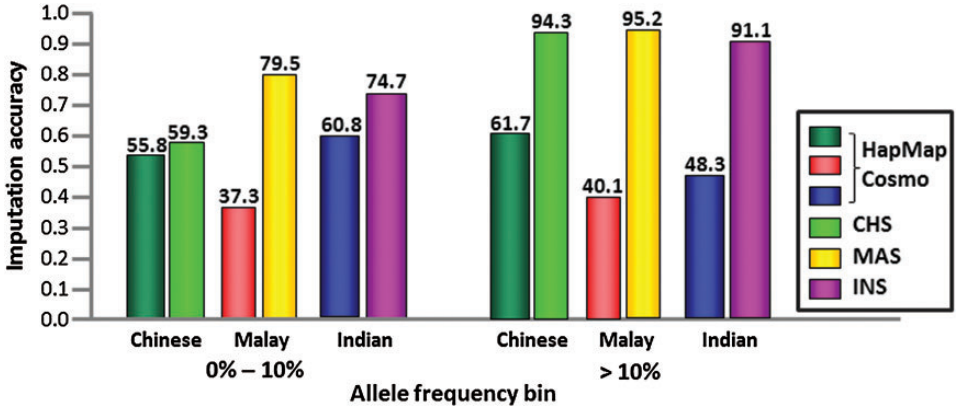

Figure 4. Imputation accuracy by HLA allele frequency bin. Each of the HLA alleles observed across all 8 loci in the 120 validation iOmics samples was categorized according to the allele frequency present in the original SGVP data set as either: (i) low-frequency, defined as an allele frequency between 0 and 10%; or (ii) common, with a frequency greater than 10%. The HLA types for these samples were masked and imputed using SNP data against two reference panels—the HapMap cosmopolitan panel which combines the HLA reference panels from all four populations in Phase 2 of the HapMap; or the population-specific reference panel from SGVP that matches the ethnic group of the samples. Imputation accuracy is measured by the extent of concordance between the imputed HLA alleles and the typed allele.

although the accuracy was significantly higher when the population-specific panels were used instead of the HapMap cosmopolitan panel (Fig. 4). The gain in accuracy was particularly greater for common alleles than for low-frequency alleles, and for Malay and Tamil Indian samples that were considerably distinct from the makeup of the HapMap cosmopolitan panel.

DISCUSSION

We have introduced a genomic resource that integrates data from dense SNP genotyping with high-resolution four-digit HLA allelotyping at three HLA Class I loci and five Class II loci, for three populations of southern Han Chinese, Southeast Asian Malay and Tamil Indian origins. This is developed with SNPs located in the Illumina1M and Affy6.0 microarrays, and has enabled three population-specific reference panels to be constructed for the purpose of imputing HLA types with SNP data, especially for the Malays and Tamil Indians for which there are currently no suitable reference panels. However, only the reference panels for the three Class I loci (*-A*, *-B*, *-C*) and two loci in Class II (*-DPA1*, *-DPB1*) are suitable for use, as poor HLA sequencing results meant that there were significantly higher levels of missing allele calls for *-DQA1*, *-DQB1* and *-DRB1*. When we compared the performance of these population-specific panels against that of a cosmopolitan reference panel built from combining all four HapMap Phase 2 populations, they consistently delivered more accurate imputation in an independent set of samples from the three population groups. However, the best imputation performance was obtained with a grand cosmopolitan panel obtained from using all available samples from the SGVP and HapMap. As with SNP imputation, HLA alleles that are more common in the population are imputed with greater accuracy, whereas low-frequency alleles are more likely to be wrongly imputed regardless of the choice of reference panel.

Numerous studies evaluating the use of SNP imputation in diverse populations have indicated that: (i) increasing the size of the reference panel improved SNP imputation accuracy, especially for non-African target populations (29–33) and (ii) a cosmopolitan reference panel can often be relevant for imputing SNPs in populations not present in the makeup of the cosmopolitan panel. The former is especially true for East Asian populations such as the Han Chinese and Japanese, where across the genome the set of distinct haplotypes is usually smaller due to the relative homogeneity between these populations. Therefore, one naturally expects the CHB samples from HapMap may serve as an appropriate reference panel to impute the HLA types for other Han Chinese samples. However, the southern Han Chinese samples experienced higher rates of errors when imputed against the northern Han Chinese, or even with the cosmopolitan panel in the case of *HLA-B*. Also, the HapMap and the Grand cosmopolitan panels were clearly inappropriate to infer the HLA types for the new populations of the Malays and Tamil Indians, despite the inclusion of the SGVP data in the Grand cosmopolitan panel. In addition, the iOmics samples used to evaluate the reference panels had >23 000 SNPs between 25 and 35 Mb, while the HapMap reference panel was built using ~13 000 SNPs compared with the SGVP panel of ~9000 SNPs. Despite the two disadvantages against the SGVP panel (less SNPs and smaller sample size per population-specific panel), the population-specificity of the SGVP panels still delivered better imputation performance, and this indicates that having the appropriate reference panel is particularly important in HLA imputation.

Several HLA alleles have been demonstrated to be highly predictive of adverse drug reactions (ADRs), particularly those related to immunoallergic reactions. From a public health perspective, understanding the burden of predictable ADRs in the population begins with measuring how common the predicting triggers are in a target group, which for HLA-associated ADRs will be the frequencies of the particular HLA alleles in the population (Table 3). By knowing how common the associated HLA alleles are in the population, as well as the positive and negative predictive values for the presence of the specific HLA alleles (11), health economic modeling can be undertaken to evaluate the cost-effectiveness of prospective HLA screening. For example, information from this resource for the frequencies of the HLA-B*15:02 allele in the three population groups in Singapore was used in a cost-effectiveness analysis of a public health

Table 3. Drug-associated HLA alleles

HLA allele	Drug	Adverse reaction[a]	Allele frequency (%)			Predictive value[b] (%)	
			CHS	MAS	INS	Positive	Negative
A*31:01	Carbamazepine	Rash	1.1	0.0	1.2	60	92
B*15:02	Carbamazepine	SJS	9.1	15.5	1.9	94	100
	Phenytoin					33	100
B*38:02	Sulfomethoxazole	SJS/TEN	3.8	4.2	1.9	50	99
B*57:01	Abacavir	HSS	1.1	6.3	1.8	47	100
	Flucloxacilin	DILI				91	88
B*58:01	Allopurinol	SJS	14.0	3.1	4.8	77	100

[a]DILI, drug-induced liver injury; HSS, hypersensitivity syndrome; SJS, Stevens–Johnson syndrome; TEN, toxic epidermal necrolysis.
[b]Predictive values obtained from Becquemont (11).

policy around genomic screening to prevent Stevens–Johnson syndrome induced by the anti-epileptic drugs carbamazepine and phenytoin. The economic evaluation revealed that screening was cost-effective to the healthcare system for the Chinese and Malays, but not for the Tamil Indians, as the frequency of the HLA-B*15:02 allele was present in only 1.9% of the Indians, when compared with 9.1 and 15.5% in the Chinese and Malays, respectively (34).

Our study has augmented the SNP resource of the SGVP with high-resolution HLA data for three additional populations in Asia. To date, this is the first comprehensive survey of HLA Class I and Class II loci in an Austronesian group (Malay) and a South Asian population. Given the multi-allelic and diverse nature of the MHC, we do not anticipate that the imputation reference panels that we have introduced will be applicable to other Austronesian and South Asian populations that are not Malays or Tamil Indians, respectively. Indeed, we caution against the use of these reference panels for populations other than those defined to be southern Han Chinese, Southeast Asian Malays or Tamil Indians. We envisage our panels will be a timely complement to those from the HapMap, as the genetics community begins to evaluate the biological significance of disease association findings located within the HLA regions, which will require progressing from SNP association to association with classical HLA alleles, and finally to identify the specific amino acid changes involved (35,36).

MATERIALS AND METHODS

Sample collection, genotyping and HLA typing

The SGVP surveyed 100 individuals from each of southern Han Chinese, Southeast Asian Malay and Tamil Indian ancestries. The subjects were randomly and anonymously chosen from a study on inter-population variation to drug response. Self-reported gender and population membership are available for each of the 300 samples, with the latter determined through self-reports that all four grandparents belong to the same population group. Out of the 300 samples identified for SGVP, 99 Chinese, 98 Malays and 95 Indians were genotyped on both the Illumina HumanHap1M and Affymetrix SNP 6.0 microarrays, of which 96 Chinese (CHS), 89 Malays (MAS) and 83 Indians (INS) across, respectively, 1 584 040, 1 580 905 and 1 583 454 autosomal SNPs were retained for further analyses after quality assessments. A set of 9766 SNPs that are present in all three populations in the 25–35 Mb region of the MHC is extracted to develop the HLA reference panels in this study. Details of the SNP-level quality checks are available in the original SGVP publication (28). High-resolution sequence-based HLA typing was performed for the three Class I loci (*-A, -B, -C*) and five Class II loci (*-DPA1, -DPB1, -DQA1, -DQB1, -DRB1*) using a sequence-based typing method with taxonomy-based sequence analysis, with a target resolution of at least four digits (37,38).

Ethics statement

Ethical consents for the SNP and HLA typing were obtained from the Institutional Review Board of the National University of Singapore, and informed consent was obtained from all participants for the inter-population study on genetic variability to drug response.

Data from the International HapMap Project

Phase 2 of the International HapMap Project surveyed ~3.1 million SNPs in: (i) 30 parent-offspring trios of northern and western European ancestry from the Centre d'Etude du Po,ymorphisme Humain collection (CEU); (ii) 30 parent-offspring trios of the Yoruba people from Ibadan, Nigeria (YRI); (iii) 45 unrelated Han Chinese from Beijing, China (CHB) and (iv) 45 unrelated Japanese from Tokyo, Japan (JPT) (20). The MHC data set for six HLA loci (*-A, -B, -C, -DQA1, -DQB1, -DRB1*) in 301 of these samples were obtained from the HapMap resource and 18 015 SNPs were available in the 25–35 Mb region of the MHC for the four population groups. The HapMap HLA resource is available at http://www.inflammgen.org.

Imputation of classical HLA alleles

To build the HLA reference panels for the populations in SGVP and HapMap, SNPs located between 25 and 35 Mb of chromosome 6 were considered although we excluded SNPs with minor allele frequency $<1\%$, missing genotype $>5\%$, or where the Hardy–Weinberg equilibrium P-value $<10^{-6}$. All SNP allele annotations were mapped to the forward strand, and physical coordinates corresponding to hg18 were used in this study. Each of the unique four-digit HLA alleles was encoded as a biallelic marker in each subject with the entry effectively counting the number of alleles present (0, 1 or 2). Any HLA

allele with frequency <0.01 was excluded and such alleles were encoded as missing in the samples. The HLA alleles at each locus were assigned a genetic position that corresponded to the center of the respective gene, except when the position coincided with the location of an actual SNP in which case the genetic position of the HLA alleles is shifted by one base. Phasing was performed with BEAGLE (36,39) on the SNP and HLA data to generate the reference panels for the populations in HapMap and SGVP. In addition, we combined the four population-specific HLA reference panels from HapMap to generate a HapMap cosmopolitan reference panel and combined the three population-specific HLA reference panels from SGVP to generate a SGVP cosmopolitan reference panel, as well as the combined cosmopolitan reference panel of SGVP cosmopolitan panel and HapMap cosmopolitan panel. In imputing the HLA alleles for a target data set possessing a similar set (or a subset) of the SNPs present in the reference panel, the binary markers at the HLA loci were encoded as missing values and BEAGLE was subsequently used to impute the outcome at these markers. This imputation process produces posterior probabilities for each possible HLA allele as well as the best-guess allele for each individual, and the posterior probabilities can be used to calculate the dosages for the respective HLA alleles. Default parameter settings for BEAGLE were used in the phasing and imputation except we assumed a larger maximum window size of 1000 consecutive markers for building the haplotype frequency model. These steps are embedded in the SNP2HLA tool (36), and the details of commands used can be found in the Supplementary Material.

LD calculation and tagSNP identification

The pairwise LD between each of the HLA alleles with surrounding SNPs was calculated using Haploview (40) applied to the phased haplotypes with the binary-encoded HLA alleles, and the extent of LD was quantified using the genetic correlation coefficient r^2. In order to identify the tagging SNPs for each of the HLA alleles, we first identified the set of HLA alleles that were in perfect LD with single SNPs and performed an aggressive search for tagging haplotypes consisting of either two or three SNPs for the remaining HLA alleles with the 'Tagger' algorithm implemented in Haploview. If none of the one to three-marker haplotype combinations provide an $r^2 \geq 0.8$, the specific combination that yields the highest r^2 is reported.

Multi-allelic F_{ST} calculation and population dendrogram construction

We quantified the genetic distance between two populations at each of the three HLA loci in Class I by calculating the multi-allelic F_{ST}, which is a weighted-average implementation of the standard biallelic-SNP F_{ST} metric defined by Weir and Cockerham (41) with weights defined by the pooled frequencies of the alleles across the two populations. The F_{ST} calculations were performed using Arlequin (42) for every pair of the following six populations: CHS, MAS and INS from SGVP, and CEU, CHB and JPT from HapMap. This yielded a 6×6 distance matrix for each of the three HLA loci in Class I (*-A, -B, -C*), which is subsequently used as the input to perform an agglomerative hierarchical clustering of the six populations using the *agnes* command in R. Briefly, each population is assigned to a separate cluster and the algorithm iterates between merging two clusters with the smallest between-cluster dissimilarity and updating the dissimilarity between the newly formed cluster with all remaining clusters. A dendrogram clustering tree is subsequently constructed to display the results of the agglomerative clustering, where the vertical coordinate where two branches join provides a measure of the dissimilarity between the two corresponding clusters.

Validation samples from the Singapore Integrative Omics Study

To assess the transferability of the SGVP and HapMap reference panels for HLA allele prediction, we identified an independent collection of 120 samples from the Singapore Integrative Omics Study (iOmics), consisting of 40 southern Han Chinese, 40 Southeast Asian Malays and 40 Tamil Indians from Singapore. As with the SGVP samples, population membership was determined by ascertaining that all four grandparents self-reported to belong to the same population group. Each of these samples has been genotyped on both the Illumina HumanOmni2.5 and HumanExome microarrays which provided a total of 23 451 SNPs across 25 and 35 Mb of chromosome 6. Each sample has also been typed for the same eight HLA loci as the SGVP samples with the same methodology as described previously.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

This project acknowledges the support of the Saw Swee Hock School of Public Health, the Yong Loo Lin School of Medicine, the National University Health System, the Life Science Institute and the Office of Deputy President (Research and Technology) from the National University of Singapore. N.E.P., R.T.H.O., W.T.P. and Y.Y.T. additionally acknowledge support from the National Research Foundation Singapore (NRF-RF-2010-05).

Conflict of Interest statement. None declared.

WEB RESOURCES

The URLs for data and software presented herein as follows: International HapMap Project: http:/hapmap.ncbi.nlm.nih.gov/; Singapore Genome Variation Project, http:/www.statgen.nus.edu.sg/~SGVP; The HapMap HLA resource, http://www.inflammgen.org.

REFERENCES

1. Wong, L.P., Ong, R.T., Poh, W.T., Liu, X., Chen, P., Li, R., Lam, K.K., Pillai, N.E., Sim, K.S., Xu, H. *et al.* (2013) Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.*, **92**, 52–66.
2. Dupont, B. and Svejgaard, A. (1977) HLA and disease. *Transplant Proc.*, **9**, 1271–1274.
3. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H.,

Samani, N.J. *et al.* (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.*, **39**, 1329–1337.
4. Khor, C.C., Chau, T.N., Pang, J., Davila, S., Long, H.T., Ong, R.T., Dunstan, S.J., Wills, B., Farrar, J., Van Tram, T. *et al.* (2011) Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nat. Genet.*, **43**, 1139–1141.
5. Ferreira, M.A., Mangino, M., Brumme, C.J., Zhao, Z.Z., Medland, S.E., Wright, M.J., Nyholt, D.R., Gordon, S., Campbell, M., McEvoy, B.P. *et al.* (2010) Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. *Am. J. Hum. Genet.*, **86**, 88–92.
6. Ross, I., Boulle, A., Soule, S., Levitt, N., Pirie, F., Karlsson, A., Mienie, J., Yang, P., Wang, H., She, J.X. *et al.* (2010) Autoimmunity predominates in a large South African cohort with Addison's disease of mainly European descent despite long-standing disease and is associated with HLA DQB*0201. *Clin. Endocrinol. (Oxf)*, **73**, 291–298.
7. Rioux, J.D., Goyette, P., Vyse, T.J., Hammarstrom, L., Fernando, M.M., Green, T., De Jager, P.L., Foisy, S., Wang, J., de Bakker, P.I. *et al.* (2009) Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc. Natl Acad. Sci. USA*, **106**, 18680–18685.
8. Monsuur, A.J., de Bakker, P.I., Zhernakova, A., Pinto, D., Verduijn, W., Romanos, J., Auricchio, R., Lopez, A., van Heel, D.A., Crusius, J.B. *et al.* (2008) Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS One*, **3**, e2270.
9. Pappu, B.P., Borodovsky, A., Zheng, T.S., Yang, X., Wu, P., Dong, X., Weng, S., Browning, B., Scott, M.L., Ma, L. *et al.* (2008) TL1A-DR3 interaction regulates Th17 cell function and Th17-mediated autoimmune disease. *J. Exp. Med.*, **205**, 1049–1062.
10. Bishof, N.A., Welch, T.R., Beischel, L.S., Carson, D. and Donnelly, P.A. (1993) DP polymorphism in HLA-A1,-B8,-DR3 extended haplotypes associated with membranoproliferative glomerulonephritis and systemic lupus erythematosus. *Pediatr. Nephrol.*, **7**, 243–246.
11. Becquemont, L. (2010) HLA: a pharmacogenomics success story. *Pharmacogenomics*, **11**, 277–281.
12. Pavlos, R., Mallal, S. and Phillips, E. (2012) HLA and pharmacogenetics of drug hypersensitivity. *Pharmacogenomics*, **13**, 1285–1306.
13. Mallal, S., Phillips, E., Carosi, G., Molina, J.M., Workman, C., Tomazic, J., Jagel-Guedes, E., Rugina, S., Kozyrev, O., Cid, J.F. *et al.* (2008) HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.*, **358**, 568–579.
14. Ferrell, P.B. Jr. and McLeod, H.L. (2008) Carbamazepine, HLA-B*1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. *Pharmacogenomics*, **9**, 1543–1546.
15. de Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M. *et al.* (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.*, **38**, 1166–1172.
16. Leslie, S., Donnelly, P. and McVean, G. (2008) A statistical method for predicting classical HLA alleles from SNP data. *Am. J. Hum. Genet.*, **82**, 48–56.
17. Listgarten, J., Brumme, Z., Kadie, C., Xiaojiang, G., Walker, B., Carrington, M., Goulder, P. and Heckerman, D. (2008) Statistical resolution of ambiguous HLA typing data. *PLoS Comput. Biol.*, **4**, e1000016.
18. 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.
19. Xie, M., Li, J. and Jiang, T. (2010) Accurate HLA type inference using a weighted similarity graph. *BMC Bioinformatics*, **11**(Suppl. 11), S10.
20. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
21. Brown, W.M., Pierce, J., Hilner, J.E., Perdue, L.H., Lohman, K., Li, L., Venkatesh, R.B., Hunt, S., Mychaleckyj, J.C. and Deloukas, P. (2009) Overview of the MHC fine mapping data. *Diabetes Obes. Metab.*, **11**(Suppl. 1), 2–7.
22. Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J., Pinder, M. *et al.* (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet.*, **41**, 657–665.
23. Miretti, M.M., Walsh, E.C., Ke, X., Delgado, M., Griffiths, M., Hunt, S., Morrison, J., Whittaker, P., Lander, E.S., Cardon, L.R. *et al.* (2005) A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **76**, 634–646.
24. Blomhoff, A., Olsson, M., Johansson, S., Akselsen, H.E., Pociot, F., Nerup, J., Kockum, I., Cambon-Thomsen, A., Thorsby, E., Undlien, D.E. *et al.* (2006) Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype specific manner assessed mainly by DRB1*03 and DRB1*04 haplotypes. *Genes Immun.*, **7**, 130–140.
25. Teo, Y.Y., Fry, A.E., Bhattacharya, K., Small, K.S., Kwiatkowski, D.P. and Clark, T.G. (2009) Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.*, **19**, 1849–1860.
26. Alper, C.A., Larsen, C.E., Dubey, D.P., Awdeh, Z.L., Fici, D.A. and Yunis, E.J. (2006) The haplotype structure of the human major histocompatibility complex. *Hum. Immunol.*, **67**, 73–84.
27. Suo, C., Xu, H., Khor, C.C., Ong, R.T., Sim, X., Chen, J., Tay, W.T., Sim, K.S., Zeng, Y.X., Zhang, X. *et al.* (2012) Natural positive selection and north-south genetic diversity in East Asia. *Eur. J. Hum. Genet.*, **20**, 102–110.
28. Teo, Y.Y., Sim, X., Ong, R.T., Tan, A.K., Chen, J., Tantoso, E., Small, K.S., Ku, C.S., Lee, E.J., Seielstad, M. *et al.* (2009) Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.*, **19**, 2154–2162.
29. Huang, L., Jakobsson, M., Pemberton, T.J., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J.K., Tishkoff, S.A. and Rosenberg, N.A. (2011) Haplotype variation and genotype imputation in African populations. *Genet. Epidemiol.*, **35**, 766–780.
30. Huang, L., Wang, C. and Rosenberg, N.A. (2009) The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am. J. Hum. Genet.*, **85**, 692–698.
31. Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A. and Scheet, P. (2009) Genotype-imputation accuracy across worldwide human populations. *Am. J.f Hum. Genet.*, **84**, 235–250.
32. Howie, B., Marchini, J. and Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3 (Bethesda)*, **1**, 457–470.
33. Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet*, **11**, 499–511.
34. Dong, D., Sung, C. and Finkelstein, E.A. (2012) Cost-effectiveness of HLA-B*1502 genotyping in adult patients with newly diagnosed epilepsy in Singapore. *Neurology*, **79**, 1259–1267.
35. Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I., Walker, B.D., Ripke, S., Brumme, C.J., Pulit, S.L., Carrington, M. *et al.* (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, **330**, 1551–1557.
36. 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.
37. Luo, M., Bamforth, J., Gill, K., Cohen, C., Brunham, R.C. and Plummer, F.A. (2005) High-resolution sequence-based DPA1 typing identified two novel DPA1 alleles, DPA1*010303 and DPA1*0303, from a Kenyan population. *Tissue Antigens*, **65**, 120–122.
38. Luo, M., Ramdahin, S., Iqbal, S., Pan, Y., Jacobson, K., Narayansingh, M.J., Schroeder, M., Brunham, R.C., Embree, J. and Plummer, F.A. (2003) High resolution sequence-based DPB1 typing identified two novel DPB1 alleles, DPB1*9401 and DPB1*9501, from a Kenyan population. *Tissue Antigens*, **62**, 182–184.
39. Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.
40. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
41. Weir, B.S. and Cockerham, C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
42. Excoffier, L. and Lischer, H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, **10**, 564–567.