Michigan Genomics Initiative Data Freeze 4 Human Leukocyte Antigen Inferences

Brett Vanderwerff[1,2,*], Lars Fritsche[1,2], Anita Pandit[1,2], Snehal Patil[1,2,3], Matthew Zawistowski[1,2], Michael Boehnke[1,2], Xiang Zhou[1,2], and Sebastian Zöllner[1,2,4]

[1] *Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA*
[2] *Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA*
[3] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[4] *Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*
[*] *To whom correspondence regarding data preparation should be addressed:* brettva@umich.edu

August 3, 2021

1 Changes From Data Freeze 3

- Human leukocyte antigen (HLA) inferences for MGI participants in Data Freeze 4 are imputed from a multi-ethnic HLA reference panel of ≈ 20,000 whole genome sequencing samples. HLA inferences in Data Freeze 3 were imputed from only a 2,504 sample subset of this multi-ethnic panel.

2 Data Description

HLA genes are located in the major histocompatibility complex (MHC) region of the human genome and contribute to the regulation of immune function [1].

HLA gene allele and amino acid inferences for three HLA class I genes (HLA-A, -B and -C) and five class II genes (HLA-DQA1 ,-DQB1 , -DRB1 , -DPA1 , -DPB1) are available for 60,215 genotyped MGI participants included in Data Freeze 4. After filtering to exclude poorly imputed variants with estimated imputation quality (Rsq) < 0.3 or very rare variants with a minor allele frequency (MAF) < 0.01% , Data Freeze 4 contains inferences for 742 HLA gene alleles and 3,367 amino acids. Table 1 provides the numbers of inferred HLA gene alleles and amino acids that are available from each gene class with the release of Data Freeze 4.

These data are available in VCF format where the absence or presence of HLA gene allele and amino acid variants are represented by binary markers coded by A and T alleles to designate absence or presence of a given variant, respectively. HLA amino acids in Data Freeze 4 may describe variation at single amino acid residues or at composite sets of amino acid residues. All variant ID nomenclature for HLA gene alleles, amino acids, and intragenic single nucleotide variations (SNVs) in these data follow the conventions outlined in the SNP2HLA v1.0 software manual [2].

To access these data, please apply through our ticketing system (submit a "Custom Data Request" in JIRA): https://doctrjira.med.umich.edu/. You will need to submit an IRB application through IRBMED to access these data, which you can apply for in eResearch Regulatory Management: https://its.umich.edu/academics-research/research/eresearch. For further assistance, please contact the Research Scientific Facilitators at phdatahelp@umich.edu, who can guide you through the data request process.

HLA Gene	Gene Class	# HLA Gene Alleles	# Amino Acids
A	I	121	914
B	I	225	707
C	I	79	442
DPA1	II	25	28
DPB1	II	96	51
DQA1	II	24	113
DQB1	II	47	177
DRB1	II	125	935
Total # HLA Variants:		742	3,367

Table 1: Numbers of HLA gene alleles and amino acids in Data Freeze 4. Counts are specific to the number of well-imputed HLA gene alleles and amino acids remaining after filtering to exclude sites with Rsq < 0.3 or a MAF < 0.01%.

3 Data Production

Production and quality control of genotype data for participants of the MGI was described previously [3]. Following quality control, the genotype data in Freeze 4 contained 4,397 MHC SNVs assayed across 60,215 MGI participants.

HLA imputation estimates unknown HLA gene alleles and amino acids in target samples by comparing MHC region SNVs with a reference panel of samples characterized for HLA. We inferred HLA gene alleles and amino acids in MGI participants from the 4-digit multi-ethnic HLA panel v1.0.0 (build 37) available from the Michigan Imputation Server (MIS, https://imputationserver.sph.umich.edu). This panel is comprised of ≈ 20,000 whole genomes from 5 global populations and contains inferences for HLA gene alleles and amino acids at HLA-A, -B, -C, -DQA1, -DQB1, -DRB1, -DPA1, and -DPB1. [4]. 2,458 MHC region SNVs overlapped between samples of the multi-ethnic HLA reference panel and the target MGI data.

As a post-imputation quality control measure, HLA gene allele and amino acid variants imputed in Data Freeze 4 with a MAF < 0.01% or an Rsq value < 0.3 were excluded.

4 Data Quality Evaluation

We evaluated imputation quality by "Rsq", an imputation quality metric produced by Minimac4, the imputation software used by the MIS. The Rsq metric estimates imputation quality at imputed sites by the formula:

$$Rsq = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1-\hat{p})}$$

where $\hat{p}$ is the frequency of the alternate allele, D_i is the allele dosage for the i^{th} haplotype and n is the number of samples that are evaluated [5]. Rsq for two-field HLA gene alleles and single-position amino acids are summarized in Figure 1 and Table 2. Here we summarize Rsq only at two-field and single-position amino acids in an attempt to limit bias that might result from including hierarchically related and composite alleles.

We compared frequencies of inferred two-field HLA gene alleles and single position amino acids from 9,996 randomly selected European unrelated Freeze 4 MGI participants to expected frequencies reported by the Allele Frequency Net Database (AFND, https://allelefrequencies.net, Figure 2). We determined

MAF Bin [%]	Gene Class	Two-Field HLA Gene Alleles		Single-Position Amino Acids	
		Mean Rsq	# Variants	Mean Rsq	# Variants
[0.01,0.1]	I	0.76	91	0.84	49
(0.1,1]	I	0.93	52	0.68	28
(1,5]	I	0.97	37	0.94	75
(5,50]	I	0.98	18	0.98	328
[0.01,0.1]	II	0.65	51	0.45	60
(0.1,1]	II	0.82	37	0.72	8
(1,5]	II	0.90	30	0.77	62
(5,50]	II	0.94	27	0.92	276

Table 2: Estimated imputation quality by gene class, variant type, and frequency. The mean estimated imputation quality (Rsq) for two-field HLA gene alleles and single-position amino acids. The total number of variants that fall into each frequency bin is also given. Only variants with Rsq ≥ 0.3 and MAF ≥ 0.01% were considered in these calculations.

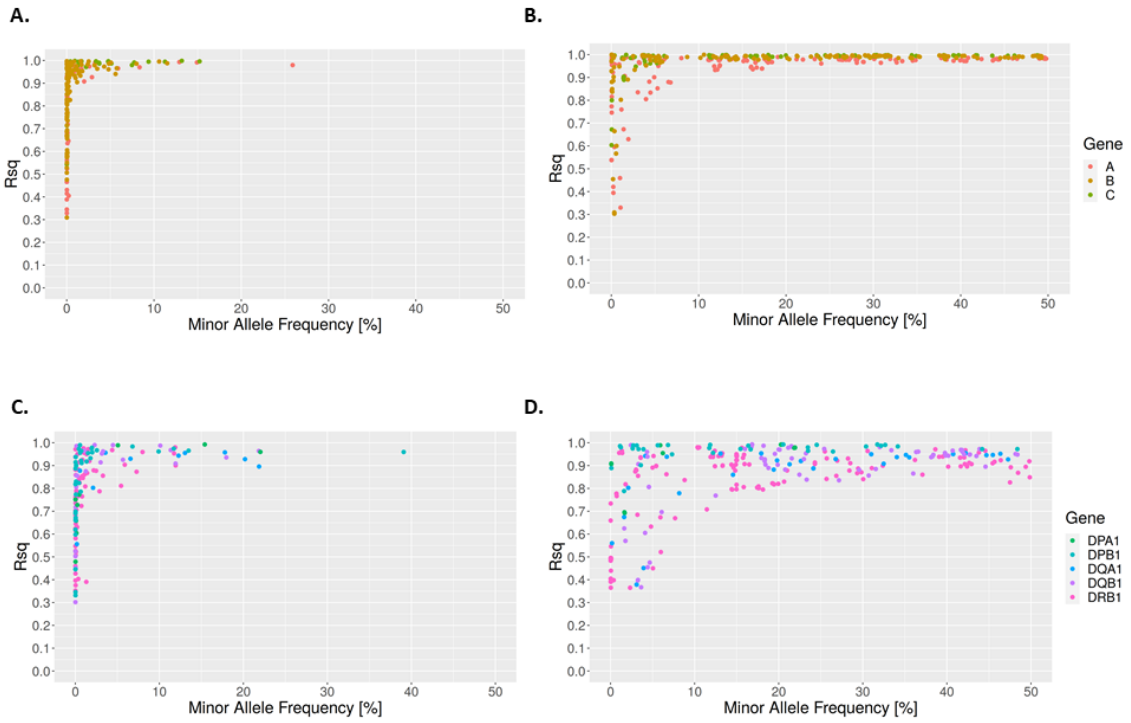

Figure 1: Estimated imputation quality by frequency. Imputation quality as estimated by Rsq for **(A.)** class I two-field HLA gene alleles, **(B.)** class I single-position amino acids, **(C.)** class II two-field HLA gene alleles, and **(D.)** class II single-position amino acids. Only variants with Rsq ≥ 0.3 and MAF ≥ 0.01% are plotted.

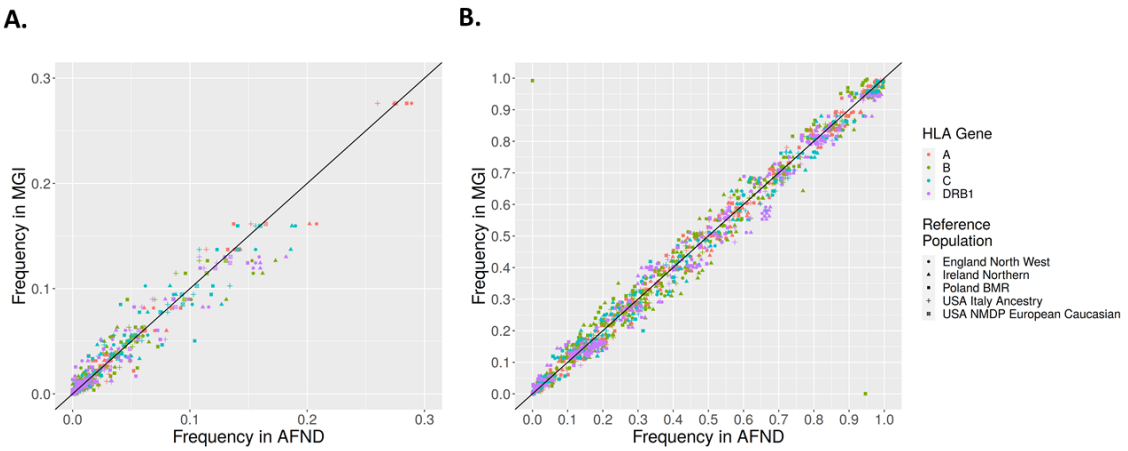

Figure 2: Comparison of MGI HLA variant frequency to AFND. HLA variant frequencies in European unrelated individuals (n=9,996) of the MGI are compared to those reported in several European and European-American populations in the Allele Frequency Net Database (AFND) for **(A.)** 250 two-field HLA gene alleles and **(B.)** 560 HLA gene amino acids. Reference population names and identification numbers (ID) are given as they appear in the AFND. England North West, ID=2837, n=298; Ireland Northern, ID=1243, n=1,000; Poland BMR, ID=3670, n=23,595; USA Italy Ancestry, ID=3714, n=273; USA NMDP European Caucasian, ID=3210, n= 1,242,890. Only variants with Rsq ≥ 0.3 and MAF ≥ 0.01% in MGI are plotted.

the square of the Pearson correlation coefficient (R2) between the frequencies observed in MGI and reported by 5 cohorts of similar ancestry from the AFND. For two-field HLA gene alleles the mean R2 was 0.948 and for single-position amino acids the mean R2 was 0.981.

We tested associations between inferred HLA gene alleles and amino acids in the MGI cohort and autoimmune disease phenotypes to replicate known associations. Cohorts of cases and controls based on European unrelated individuals for type 1 diabetes (T1D), psoriasis, and multiple sclerosis (MS) phenotypes were constructed by parsing patient International Classification of Diseases diagnosis codes for MGI participants. Association tests were then performed by SAIGE [6]. Age, recruiting study, genotype batch, genotype-inferred sex, genotype array, and the first 4 principal components were used as covariates (Figure 3).

The most significant HLA gene allele or amino acid signal for T1D, psoriasis, and MS were the amino acid position 57 of DQB1, HLA-C*06:02:01:01, and HLA-DQB1*06:02, respectively. Each of these associations are well known, indicating that known associations between autoimmune disease phenotypes and HLA gene alleles and amino acids can be recapitulated in the MGI [7, 8, 9, 10].

5 Limitations of These Data

These data do not report HLA gene alleles that were measured by gold standard HLA gene allele typing methods, but rather are inferences based on available MHC region genotypes for participants of the MGI. The uncertainty of these inferences is communicated through the allele "dosages" that are reported in the VCF file.

HLA gene alleles in Freeze 4 may be reported at up to four field resolution based on standard nomenclature, but all HLA inferences in MGI are ultimately based off G-group alleles that were inferred for the multi-ethnic HLA reference panel [11, 4].

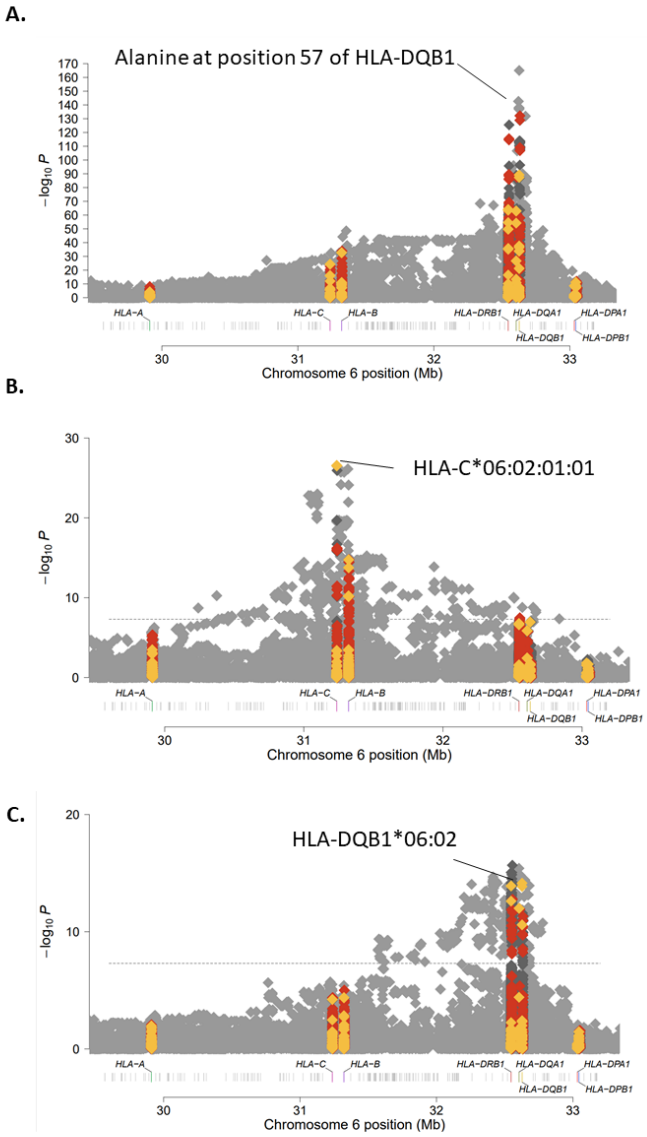

Figure 3: Association of inferred HLA gene alleles and amino acids in MGI with autoimmune disease phenotypes. Manhattan plot for **(A.)** type 1 diabetes (2,250 cases, 35,969 controls), **(B.)** psoriasis (1,330 cases, 41,320 controls), and **(C.)** multiple sclerosis (378 cases, 42,201 controls). For each phenotype, the most significant HLA gene allele or amino acid signal is labeled. Red, yellow, and grey points on the plots represent HLA gene amino acids, HLA gene alleles, and MHC region SNVs, respectively. The dashed line represents a 5 x 10^{-8} significance threshold.

References

[1] Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: Expression, interaction, diversity and disease. *Journal of Human Genetics* **54**, 15–39 (2009).

[2] http://software.broadinstitute.org/mpg/snp2hla/snp2hla_manual.html.

[3] Michigan Genomics Initiative Data Freeze 3 Technical Notes v1.2. https://drive.google.com/drive/folders/11p7d9ahHVYyY5SALthO4L-rH6YM74Vwd.

[4] Luo *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. *medRxiv* (2020). https://www.medrxiv.org/content/early/2020/07/18/2020.07.16.20155606.full.pdf.

[5] Minimac3 Info File - Genome Analysis Wiki. https://genome.sph.umich.edu/wiki/Minimac3_Info_File.

[6] Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *bioRxiv* 583278 (2020).

[7] Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PloS One* **8**, e64683 (2013).

[8] Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature Genetics* **47**, 1107–1113 (2015).

[9] Stuart, P. E. *et al.* A Single SNP Surrogate for Genotyping HLA-C*06:02 in Diverse Populations. *The Journal of investigative dermatology* **135**, 1177–1180 (2015).

[10] Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genetics* **44**, 1341–1348 (2012).

[11] HLA Nomenclature @ hla.alleles.org. http://hla.alleles.org/nomenclature/naming.html.