PRECISION HEALTH
UNIVERSITY OF MICHIGAN

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN
DATA OFFICE FOR CLINICAL &
TRANSLATIONAL RESEARCH

Human Leukocyte Antigen Gene Allele and Amino Acid Inferences for the Michigan Genomics Initiative

About this guide

This guide provides information about the production and evaluation of human leukocyte antigen (HLA) gene allele and amino acid inferences in the Michigan Genomics Initiative (MGI) cohort.

HLA Overview

HLA gene allele and amino acid inferences for genes HLA-A, -B, -C, -DQA1, -DQB1, -DRB1, -DPA1, and -DPB1 are available for 56,984 genotyped MGI participants. These inferences are based on major histocompatibility complex (MHC) region single nucleotide variations (SNVs) that were assayed for each MGI participant by genotyping array. The inferences were made using a HLA reference panel of 2,504 1000 Genomes Project (1KGP) samples.

Example Use Case

An investigator could use these data from the MGI to test associations between HLA alleles/amino acids and autoimmune disease phenotypes.

FAQ

What are inferences of HLA gene alleles and amino acids?	Human leukocyte antigen (HLA) genes are located in the MHC region of the human genome and contribute to the regulation of immune function[1]. HLA gene alleles are partially defined by the nucleotide and amino acid sequences that are encoded by HLA genes[2]. Some HLA gene allele definitions describe specific amino acid sequences of an HLA gene product[2]. MGI offers data that describe the inferred presence or absence of different HLA gene alleles and amino acids for genotyped MGI participants.
Which HLA genes are represented by these inferences?	These data contain inferred HLA gene allele and amino acid sequences for the following genes: HLA-A, -B, -C, -DQA1, -DQB1, -DRB1, -DPA1, and -DPB1 (Table 1).
What methods were used to produce these data?	Production and quality control of genotype data for participants of the MGI was described previously[3]. Following quality control, the genotype data contained 4,409 MHC region SNVs assayed across 56,984 MGI participants. We inferred HLA gene alleles and amino acids in MGI using the HLA imputation software SNP2HLA and a reference panel of 2,504 1KGP samples with whole genome sequence-based calls for HLA gene alleles and amino acids at HLA-A, -B, -C, -DQA1, -DQB1, -DRB1, -DPA1, and -DPB1 in addition to MHC region SNVs[4,5]. To parallelize imputation, samples from the full MGI cohort were randomly divided into 9 evenly sized sub-cohorts and imputed separately. On average, 2,568 MHC region SNVs overlapped between the reference panel and each MGI sub-cohort. Following imputation, we merged the separately imputed MGI sub-cohorts to form a single data set. A table describing which MGI samples were imputed together is distributed with the merged data set. As a post-imputation quality control measure, variants with a minor allele frequency (MAF) < .0001

	or a Dosage R-Squared (DR2) value < 0.3 were excluded (*see section: What measures were used to evaluate data quality?*).
What measures were used to evaluate data quality?	The estimated imputation quality of HLA allele and amino acid inferences of the MGI cohort were evaluated by DR2, by the imputation engine of SNP2HLA, Beagle (v4.1) [4,6,7]. DR2 represents an estimation of the squared correlation between the estimated allele dose and the true, unobserved allele dose. For HLA gene alleles, the mean DR2 values were .79 (0.1% ≤ MAF ≤ 0.5%), .90 (0.5% < MAF ≤ 5%), and .94 (MAF > 5%). For HLA gene amino acids, the mean DR2 values were .69 (0.1% ≤ MAF ≤ 0.5%), .80 (0.5% < MAF ≤ 5%), and .92 (MAF > 5%) (Fig. 1). We note a trend of slightly reduced DR2 values for the inferred alleles and amino acids of the DRB1 gene compared to other HLA genes. We compared frequencies of inferred HLA gene alleles and amino acids in MGI to expected frequencies reported by the Allele Frequency Net Database (AFND, allelefrequencies.net)[8]. We determined the square of the Pearson correlation coefficient (R^2) between the frequencies observed in MGI and reported by 6 cohorts of similar ancestry from the AFND (Fig. 2). For alleles of HLA-A, -B, -C, -DQB1, and -DRB1 the mean R^2 was .99, .94, .92, .89, and .91, respectively. For amino acids of HLA-A, -B, -C, -DQB1, and -DRB1 the mean R^2 was .99, .94, .99, .97, and .99 respectively. 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 logistic regression (PLINK v1.09)[9]. Age, recruiting study, genotype batch, genotype-inferred sex, and the first 4 principal components were used as covariates (Fig. 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-DRB1*15:01:01:01, 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[6,10–12]. Further, we compared odds ratios of HLA-DRB1-DQA1-DQB1 inferred haplotypes derived from cases and controls of the T1D association study to effect sizes reported in Cucca *et al.* (Table 2 & Fig. 4)[13]. Effect size estimates for haplotypes inferred in MGI carry risk comparable to the literature-derived values.
What format are the data available in?	The data are available in Variant Calling Format (VCF). Variant ID nomenclature for HLA gene alleles, amino acids, and intragenic SNVs follow the conventions outlined in the SNP2HLA manual, with the exception that binary encodings for the absence or presence of HLA gene alleles and amino acids are T = Present, A = Absent[14].
Are there limitations/considerations when using HLA gene allele and amino acid inferences from the MGI?	These data do not report HLA alleles that were measured by gold standard HLA 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. Due to a technical issue, the nomenclature for biallelic amino acid variants used in the VCF file lacks the detail required to easily determine which amino acids are being referred to. For example, the variant with ID \`AA_A_80_29910771_exon2\` refers to a biallelic amino acid variation at position 80 in exon 2 of the HLA-A protein. Although this record reports on the absence or presence of the amino acid isoleucine or threonine at this site, this cannot be easily determined from the VCF file alone. This currently affects ~10% of all amino acid variants that are reported in the VCF file. A resolution is being pursued.

	HLA gene allele and amino acid Inferences in the MGI cohort were made using an HLA reference panel containing 2,504 samples from the 1KGP (*see section: What methods were used to produce these data?*). While the 1KGP HLA reference panel is currently one of the largest panels available to the public without restriction, larger panels that are under restricted access policies do exist, such as the Type 1 Diabetes Genetics Consortium HLA reference panel[6]. The MGI is currently seeking access to larger HLA reference panels, which should allow for improved HLA gene allele and amino acid inference accuracy in the MGI cohort.
How can these data be accessed?	To access MGI HLA gene allele and amino acid inferences, 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, which you can do in eResearch Regulatory Management: https://its.umich.edu/academics-research/research/eresearch. If you need further assistance, please contact the Research Scientific Facilitators at phdatahelp@umich.edu, who can guide you through the data request process.

References

1. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).

2. HLA Nomenclature @ hla.alleles.org. http://hla.alleles.org/nomenclature/naming.html.

3. Vanderwerff, B. *et al.* Michigan Genomics Initiative Data Freeze 3 Technical Notes v1.2. https://drive.google.com/file/d/1tZKmrKxnHxH0MJbvHvKQb8Cx4HovmyyM/view?usp=sharing (2020).
4. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. *medRxiv* 2020.07.16.20155606 (2020) doi:10.1101/2020.07.16.20155606.
5. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
6. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).
7. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
8. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783–D788 (2020).
9. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
10. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44**, 1341–1348 (2012).
11. Stuart, P. E. *et al.* A Single SNP Surrogate for Genotyping HLA-C*06:02 in Diverse Populations. *J. Invest. Dermatol.* **135**, 1177–1180 (2015).

12. Fogdell, A., Hillert, J., Sachs, C. & Olerup, O. The multiple sclerosis- and narcolepsy-associated HLA class II haplotype includes the DRB5*0101 allele. *Tissue Antigens* **46**, 333–336 (1995).

13. Cucca, F. *et al.* A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins. *Hum. Mol. Genet.* **10**, 2025–2037 (2001).

14. SNP2HLA Manual (v1.0). http://software.broadinstitute.org/mpg/snp2hla/snp2hla_manual.html.

Tables

# Inferred HLA Gene Alleles and Amino Acids in MGI			
HLA Gene	*Gene Class*	*# Alleles*	*# Amino Acids*
A	I	64	283
B	I	124	216
C	I	46	232
DPA1	II	22	28
DPB1	II	63	50
DQA1	II	15	62
DQB1	II	25	165
DRB1	II	59	519

Table 1. Counts of HLA gene alleles and amino acids that were inferred in participants of the MGI. HLA alleles of all resolutions contributed to the total allele count.

Risk Level	Haplotype			Previously Reported Risk (Cucca *et al.*)		MGI Risk	
	DRB1	DQA1	DQB1	Odds Ratio	95% CI	Odds Ratio	95% CI
Very High	04:05	03:01	03:02	10.8	5.6-20.6	4.0	2.9-5.5
	04:01	03:01	03:02	7.2	4.6-11.5	4.3	3.6-5.0
	03:01	05:01	02:01	4.3	2.9-6.3	2.7	2.3-3.0
	04:04	03:01	03:02	4.1	2.3-7.1	1.7	1.4-2.0
	04:02	03:01	03:02	3.1	1.4-6.8	3.2	2.4-4.4
Intermediate	08	04:01	04:02	1.6	0.8-3.1	1.5	1.2-1.8
	13:02	01:02	06:04	1.3	0.6-2.7	1.4	1.2-1.8
	09:01	03:01	03:03	1.2	0.5-2.9	1.9	1.4-2.7
	01	**01:01**	**05:01**	**1.0**	**-**	**1.0**	**-**
	16:01	01:02	05:02	0.8	0.5-1.2	1.1	0.8-1.5
	04:01	03:01	03:01	0.8	0.4-1.5	1.1	0.9-1.4
	04:03	03:01	03:02	0.5	0.1-1.4	1.3	0.7-2.4
	13:01	01:03	06:03	0.4	0.2-1.1	0.8	0.7-1.0
	10:01	01:01	05:01	0.2	0.05-0.9	0.6	0.4-1.2
	13:03-13:05	05:01	03:01	0.3	0.1-1.2	0.4	0.2-0.7
Very Low	11-12	05:01	03:01	0.2	0.1-0.3	0.6	0.5-0.8
	15:01	01:02	06:02	0.04	0.01-0.1	0.5	0.4-0.6
	07:01	02:01	03:03	0.1	0.01-0.6	0.6	0.5-0.8
	14:01	01	05:03	0.1	0.01-0.4	0.4	0.3-0.7

Table 2. Odds ratios for type 1 diabetes risk haplotypes. Odds ratios for HLA-DRB1-DQA1-DQB1 type 1 diabetes risk haplotypes that were inferred in MGI are compared to literature-based values from Cucca *et al.* Haplotypes are divided into very high, intermediate, and very low risk categories based on definitions from Cucca *et al.* CI, confidence interval.

Figures

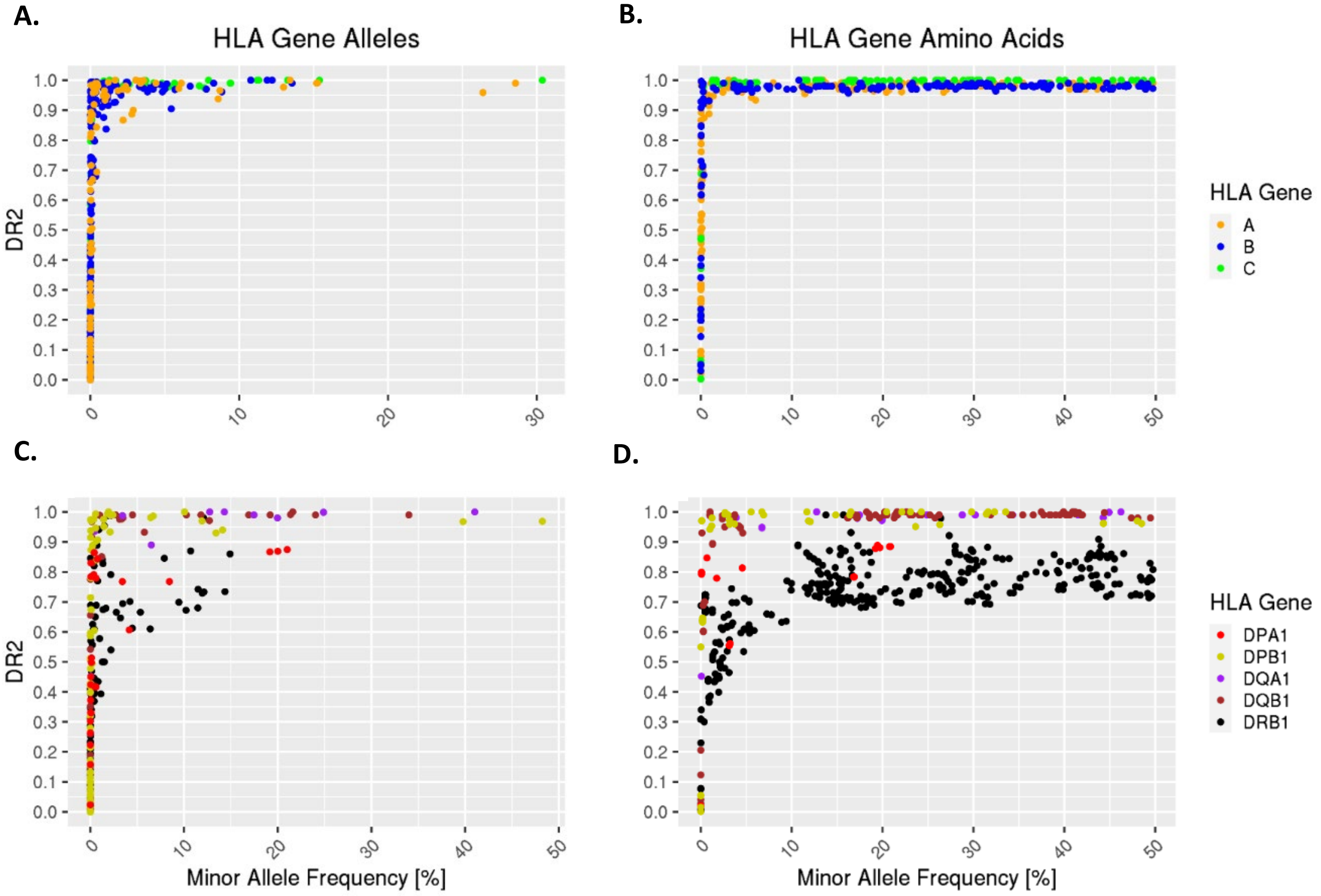

Figure 1. **Estimated HLA gene allele and amino acid imputation quality in MGI.** The estimated squared correlation between the estimated allele dose and the true, unobserved allele dose for **(A-B.)** class I HLA genes and amino acids and **(C-D.)** class II HLA genes and amino acids.

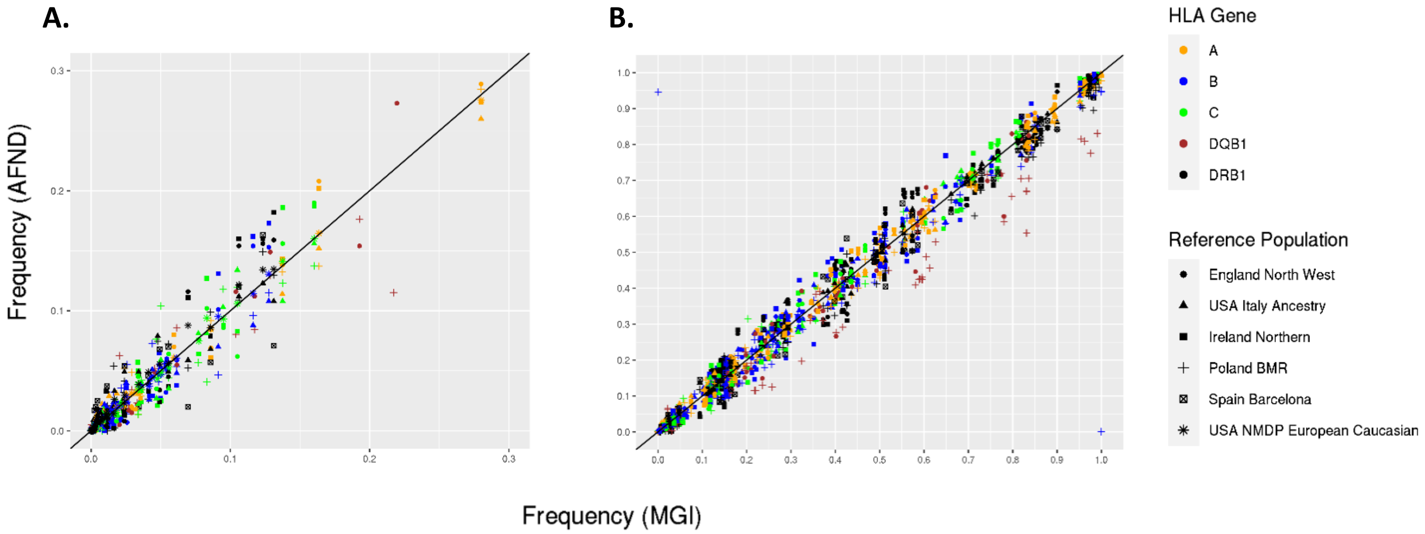

Figure 2. **HLA allele frequencies in MGI compared to the Allele Frequency Net Database.** HLA allele frequencies that were observed in European unrelated individuals (n=45,665) of the MGI are compared to those reported in several European and European-American populations in the Allele Frequency Net Database (AFND) for **(A.)** HLA gene alleles and **(B.)** 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; USA Italy Ancestry, ID=3714, n=273; Ireland Northern, ID=1243, n=1,000; Poland BMR, ID=3670, n=23,595; Spain Barcelona, ID=1242, n=941, USA NMDP European Caucasian, ID=3210, n= 1,242,890.

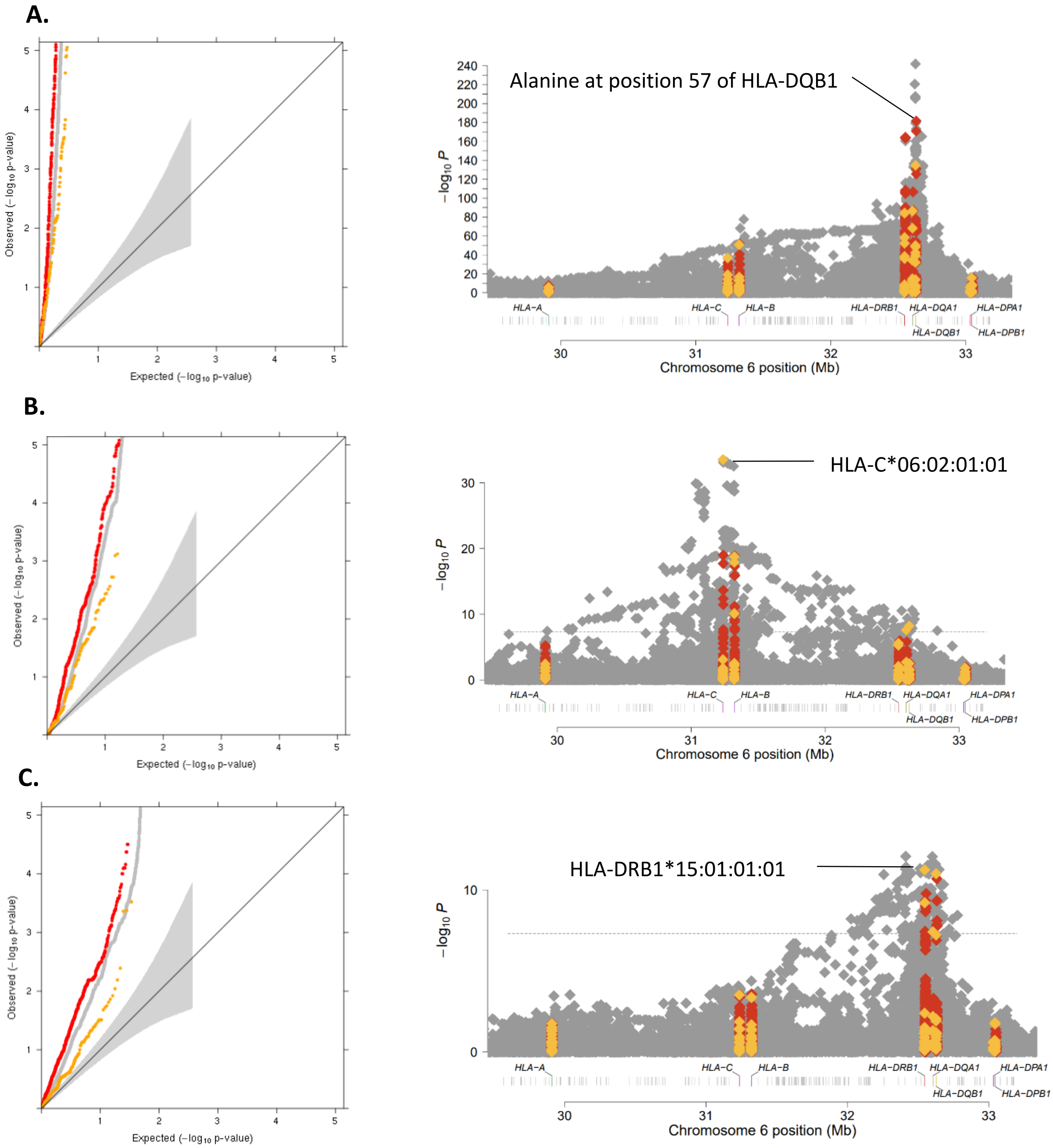

Figure 3. Association of inferred HLA gene alleles and amino acids in MGI with autoimmune disease phenotypes. Quantile-quantile plot (left) and Manhattan plot (right) for **(A.)** type 1 diabetes (1,759 cases, 47,867 controls), **(B.)** psoriasis (1,386 cases, 48,233 controls), and **(C.)** multiple sclerosis (541 cases, 49,085 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.

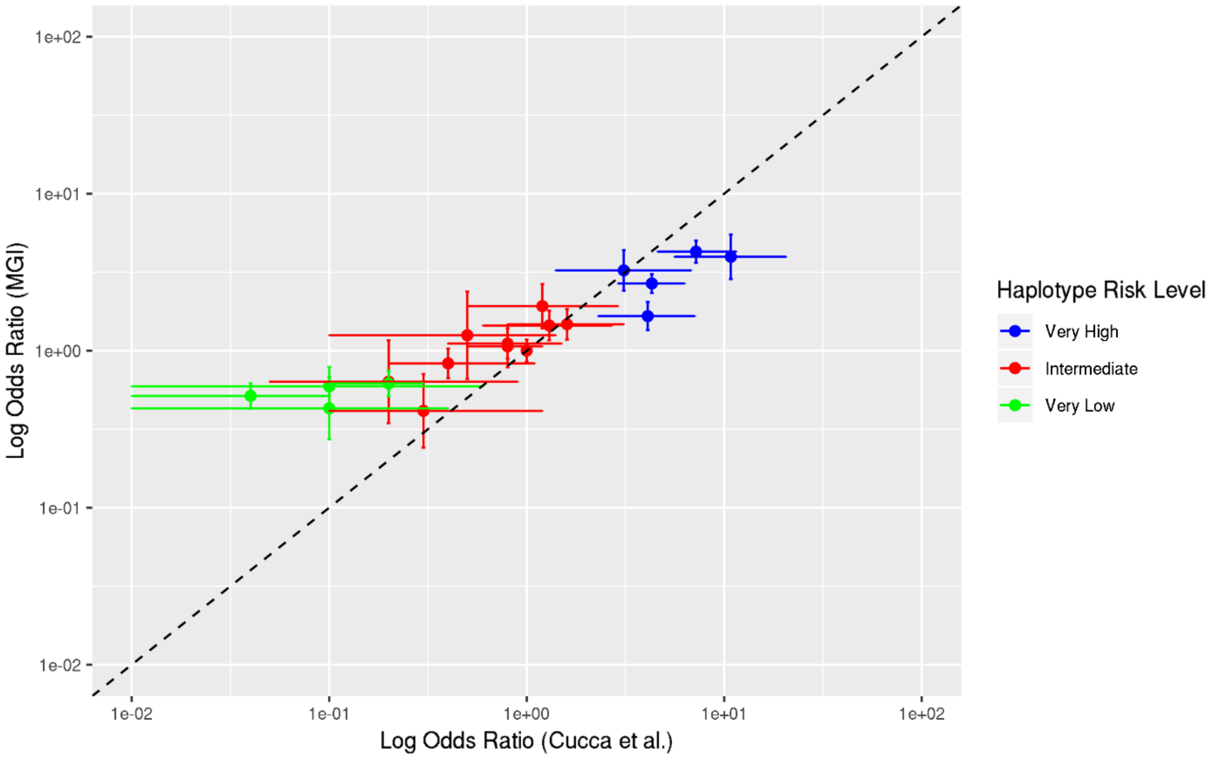

Figure 4. Odds ratios for type 1 diabetes risk haplotypes. Plot depicts the comparison of odds ratios for HLA-DRB1-DQA1-DQB1 type 1 diabetes risk haplotypes that were inferred in MGI to literature-based values from Cucca *et al.* Very high, intermediate, and low haplotype risk level definitions were obtained from Cucca *et.al*. The dashed diagonal line represents the hypothetical perfect match between observed and literature-based odds ratios (y=x). The horizontal and vertical error bars represent the 95% confidence intervals for odds ratios reported in Cucca *et al.* and observed in the MGI cohort, respectively.