PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Michigan Genomics Initiative Freeze 6 Human Leukocyte Antigen Inferences

Brett Vanderwerff[1*], Matthew Zawistowski[1], Lars G. Fritsche[1], Emily Bertucci-Richter[1], Snehal Patil[1,2], Michael Boehnke[1], Xiang Zhou[1], and Sebastian Zöllner[1,3]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. [3]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

1 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] This data release contains HLA gene allele and amino acid inferences for three HLA class I genes (HLA-A, -B and -C), five class II genes (HLA-DQA1 ,-DQB1 , -DRB1 , -DPA1 , and -DPB1), and MHC region single nucleotide variations (SNVs) for 80,529 MGI participants included in Data Freeze 6. HLA gene alleles in Freeze 6 may be reported at up to two field resolution, which describes a unique HLA protein amino acid sequence. More information on HLA allele nomenclature can be found in resources from the HLA informatics group: https://hla.alleles.org/nomenclature/naming.html.

After filtering to exclude poorly imputed variants with estimated imputation quality (Rsq) < 0.7 or very rare variants with a minor allele frequency (MAF) < 0.01%, Freeze 6 contains inferences for 392 HLA gene alleles, 2,326 amino acids, and 18,535 MHC region SNVs. Figure 1 provides the counts of inferred HLA gene alleles and amino acids that are available from each gene class with the release of Freeze 6.

These HLA 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 the absence or presence of a given variant, respectively. HLA amino acids in Freeze 6 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 SNVs in these data follow the conventions outlined in the SNP2HLA v1.0 software manual.[2]

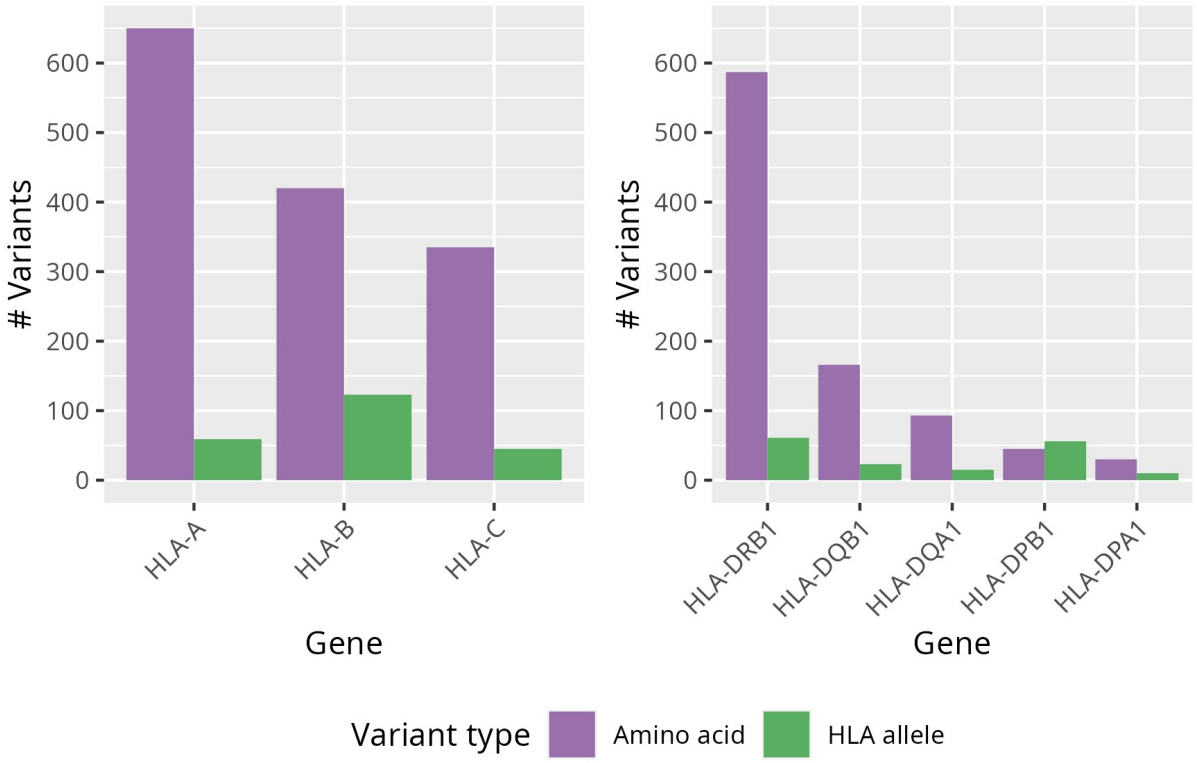

Figure 1: HLA gene allele and amino acid counts. Counts of HLA gene alleles and amino acids in Freeze 6 for (left) class I and (right) class II HLA genes. Counts are specific to the number of well-imputed HLA gene alleles and amino acids remaining after filtering to exclude sites with Rsq < 0.7 or MAF < 0.01%.

2 Data Access

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.

3 Data Production

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 v2@1.0.0 (build 37) available from the Michigan Imputation Server (

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 and MHC region SNVs.[3]

We describe the production and quality control of genotype data for participants of the MGI and our approach for merging separately imputed data elsewhere.[4] Briefly, MGI participants in Freeze 6 are genotyped on one of two different arrays, the CoreExome array (n=60,715) or the Global Screening Array (GSA, n=19,814). We performed HLA imputation separately for participants assayed on the CoreExome or GSA using target haplotypes that were pre-phased using the Trans-Omics for Precision Medicine (TOPMed) panel as reference.[5] Following HLA imputation in MGI, we merged data from all samples using the Michigan Imputation Server post-processing tool "hds-util".[6]

4 Data Quality Control

As a post-imputation quality control measure, we excluded HLA gene allele, amino acids, and SNVs imputed in Freeze 6 with a MAF < 0.01% or an Rsq value < 0.7.

5 Data Quality Evaluation

We used the "Rsq" and "EmpRsq" metrics produced by Minimac4, the genotype imputation software used by the Michigan Imputation Server, to evaluate imputation quality.[7] The Rsq metric estimates imputation accuracy at all imputed sites by the formula:

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

where $\hat{p}$ is the estimated 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.[8] Rsq for two-field HLA gene alleles and single-position amino acids are summarized in Figure 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.

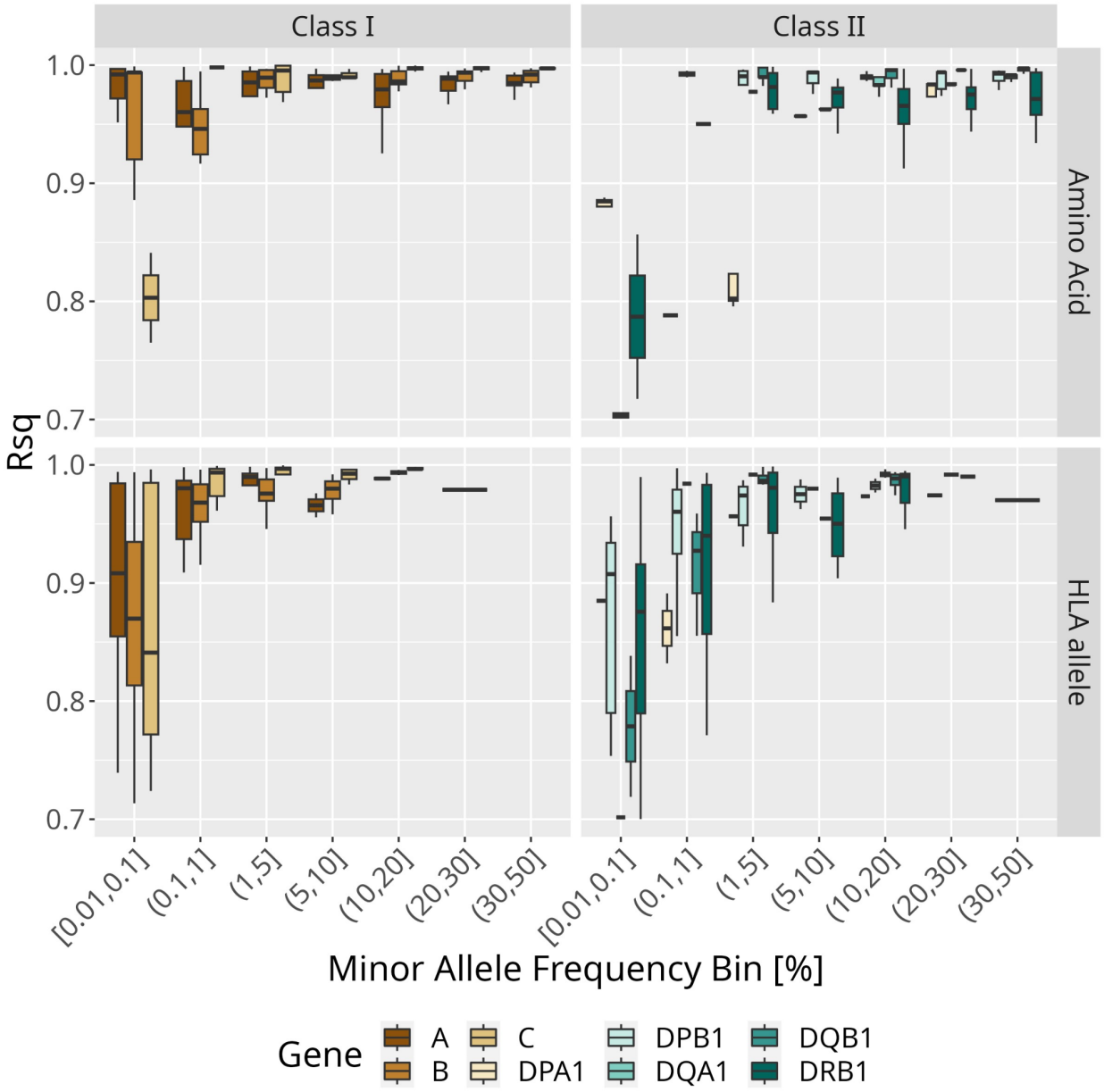

Figure 2: Estimated imputation quality. Estimated imputation quality (Rsq) by minor allele frequency for class I and class II two-field HLA gene alleles and single-position amino acids. Only variants with Rsq ≥ 0.7 and MAF ≥ 0.01% are plotted.

We compared frequencies of inferred two-field HLA gene alleles and single position amino acids from 10,000 randomly selected European ancestry Freeze 6 MGI participants to expected frequencies reported by the Allele Frequency Net Database (AFND, https://allelefrequencies.net) for HLA-A, -B, -C, and -DRB1. We determined the square of the Pearson correlation coefficient (R^2) between the frequencies observed in MGI and reported by five cohorts of similar ancestry from the AFND. For two-field HLA gene alleles the R^2 was 0.9457 and for single-position amino acids the R^2 was 0.9853 (Figure 3).

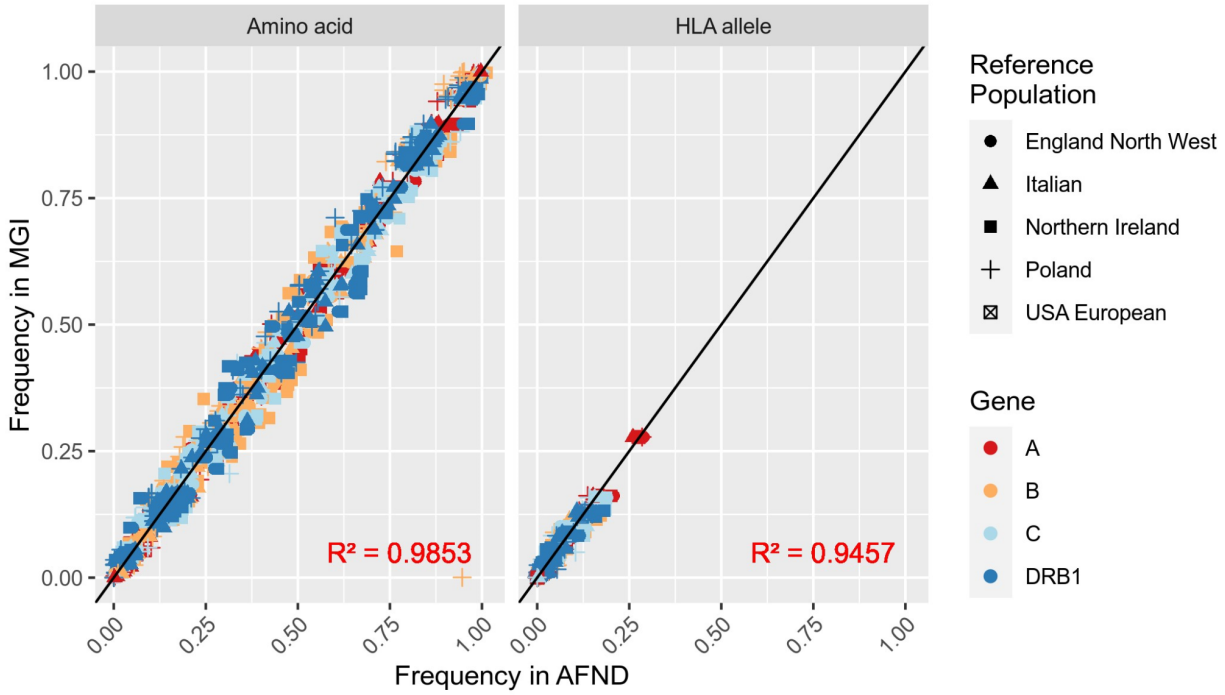

Figure 3: Comparison of MGI HLA variant frequency to AFND. HLA variant frequencies in European individuals (n=10,000) of the MGI are compared to those reported in several European and European-American populations in the Allele Frequency Net Database (AFND) for (left) 541 HLA gene amino acids and (right) 207 two-field HLA gene alleles. Reference population names and identification numbers (ID) are given as they appear in the AFND. England North West, ID=2837, n=298; Italian, ID=3714, n=273; Northern Ireland, ID=1243, n=1,000; Poland, ID=3670, n=23,595; USA European, ID=3210, n= 1,242,890. Only variants with Rsq ≥ 0.7 and MAF ≥ 0.01% in MGI are plotted. R^2 is the square of the Pearson correlation coefficient between MGI and AFND allele frequencies.

We tested associations between inferred HLA gene alleles, HLA amino acids, and MHC region SNVs in the MGI cohort and autoimmune disease phenotypes to replicate known associations. We constructed cohorts of cases and controls based on multi-ancestry MGI participants for type 1 diabetes, psoriasis, and multiple sclerosis phenotypes by converting patient International Classification of Diseases diagnosis codes to phecodes using the R PheWAS package.[9] We then performed association tests between inferred HLA genotypes and the phecode phenotypes using SAIGE v 1.3 and controlling for age, genotype-inferred sex, genotype array, and the first 4 genetic principal components (Figure 4).[10]

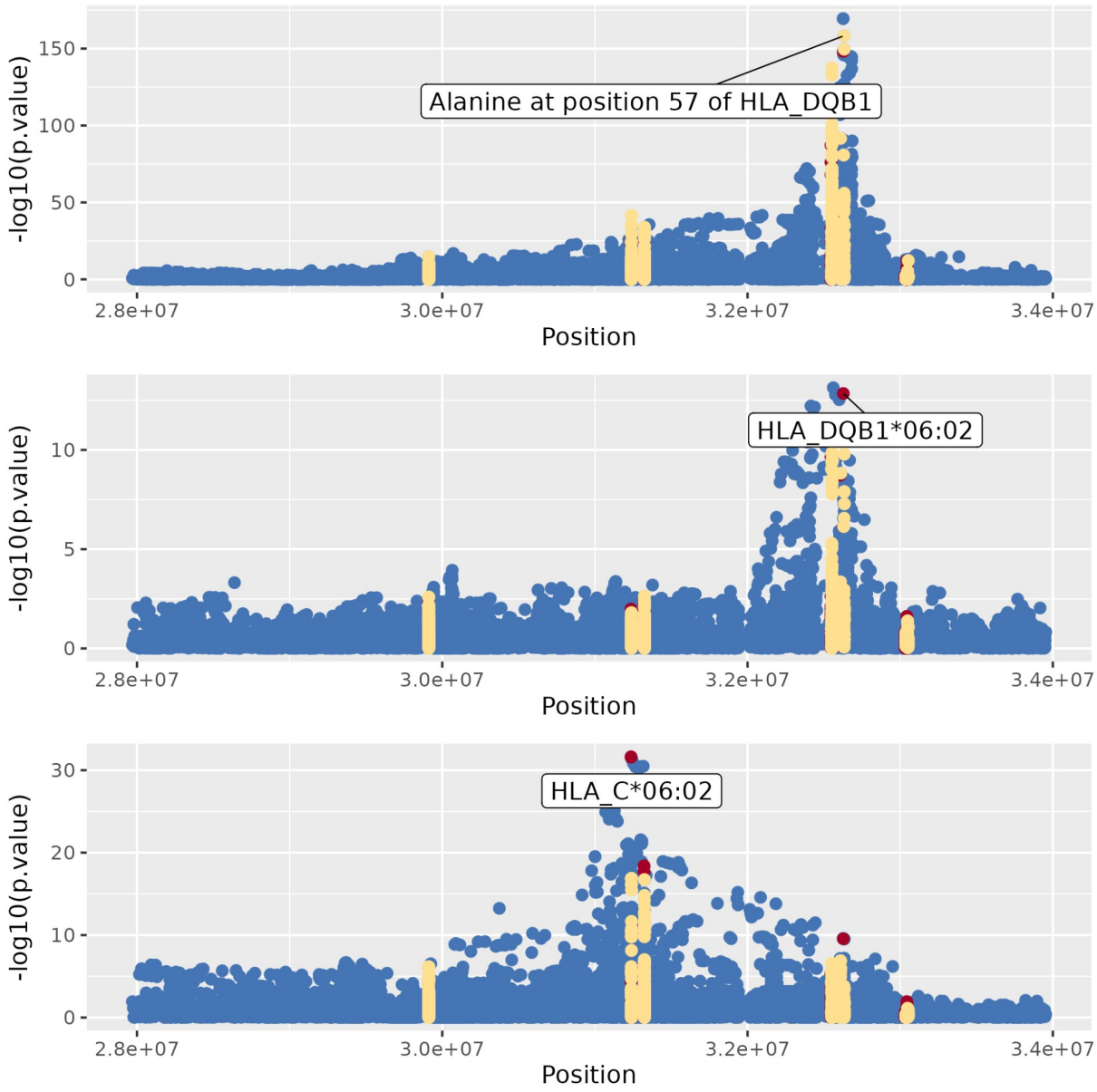

Figure 4: Association of inferred HLA genotypes in MGI with autoimmune disease phenotypes. Manhattan plot for (top panel) type 1 diabetes (3,586 cases & 53,267 controls), (middle panel) multiple sclerosis (615 cases & 63,598 controls), and (bottom panel) psoriasis (2,218 cases & 62,821 controls). For each phenotype, the most significant HLA gene allele or amino acid signal is labeled. Yellow, red, and blue points on the plots represent HLA gene amino acids, HLA gene alleles, and MHC region SNVs, respectively.

The most significant HLA gene allele or amino acid signal for type 1 diabetes, psoriasis, and multiple sclerosis were the amino acid position 57 of DQB1, HLA-C*06:02, and HLA-DQB1*06:02, respectively. Each of these associations are well known, indicating that

known associations between autoimmune disease phenotypes and inferred HLA gene alleles and amino acids can be recapitulated in the MGI.[11–14]

6 Recommendations for Analyzing Across Genotyping Arrays

We generated these HLA inferences based on genotypes that were directly measured on either the GSA or CoreExome arrays, which makes the array a potential confounding factor in analyses. Two options for analyzing these HLA data across genotyping arrays are 1.) meta-analysis of summary statistics from analyses run on imputed HLA data from each array or 2.) joint-analysis of imputed HLA data pooled across all arrays.

We empirically compared these options for analysis of HLA data using an approach we described previously for genome-wide imputed genotypes.[4] Briefly, we assigned case/control status among European ancestry participants for type 1 diabetes (3,524 cases & 46,828 controls), multiple sclerosis (607 cases & 54,714 controls), and psoriasis (2,203 cases & 54,513 controls). To perform meta-analysis, we first ran association tests using SAIGE on imputed HLA data collected from participants assayed on the CoreExome array or GSA separately and controlling for age, genotyping array version, sex, and first 4 genetic principal components.[15] For each phenotype, we then meta-analyzed each pair of summary statistics generated from SAIGE by running METAL in inverse variance weighted mode.[16] To perform joint-analysis, we ran association tests as described above with the exception that we provided imputed HLA data pooled from participants assayed on either the CoreExome array or GSA as input for SAIGE.

We compared p.values and effect sizes (betas) between the meta- and joint-analysis approaches at sites with p-value < .05 in either the meta-analysis or joint-analysis approach. The -log10(p-value) and beta concordance between each approach increased with MAF and had high concordance with R^2 > .999 among sites with MAF > 1% for both p-values and betas (Figure 5). These data suggest that meta-analysis of the HLA imputation performs nearly identical to joint-analysis, particularly for sites with MAF > 1%. Given that joint-analysis requires less computational steps than meta-analysis, we recommend that users testing variants with MAF > 1% use joint-analysis and to consider meta-analysis when evaluating rarer variants.

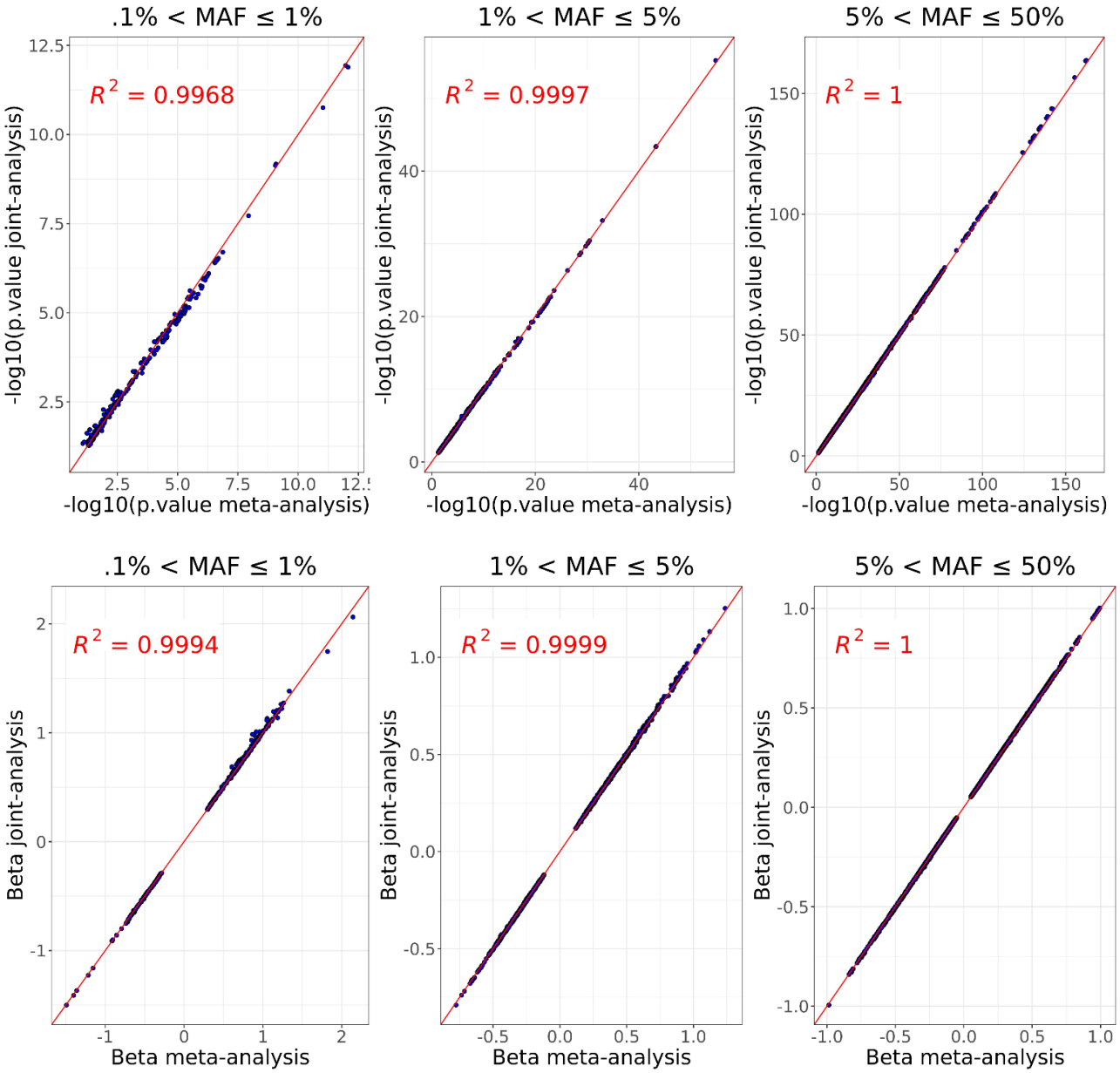

Figure 5. Comparison of p-values and betas from meta- and joint-analysis. Results for -log10(p-values) (top row) and betas (bottom row). Points plotted are any hit with a p-value < .05 in either the joint-analysis or meta-analysis across any of the type 1 diabetes, multiple sclerosis, or psoriasis phenotypes evaluated. R^2 is the square of the Pearson correlation coefficient between meta- and joint-analysis. MAF, minor allele frequency.

7 Limitations of These Data

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

HLA gene alleles in Freeze 6 may be reported at up to two 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.[3]

8 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. SNP2HLA Manual (v1.0). http://software.broadinstitute.org/mpg/snp2hla/snp2hla_manual.html.
3. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet* **53**, 1504–1516 (2021).
4. Vanderwerff, B. *et al.* Michigan Genomics Initiative Freeze 6 Genome-Wide Genotypes.
5. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
6. statgen/hds-util. https://github.com/statgen/hds-util (2023).
7. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
8. Minimac3 Info File - Genome Analysis Wiki. https://genome.sph.umich.edu/wiki/Minimac3_Info_File (2021).
9. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).
10. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat Genet* **52**, 634–639 (2020).

11. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).

12. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat Genet* **47**, 1107–1113 (2015).

13. 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).

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

15. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).

16. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).