Supplementary Note

A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response

Luo et al.

Contents

1. Deep-coverage whole-genome sequencing cohort descriptions

Study participants for constructing the multi-ethnic MHC reference panel were from the Jackson Heart Study (JHS, $N = 3,024$), the Multi-Ethnic Study of Atherosclerosis cohort (MESA, $N = 4,260$), the Chronic Obstructive Pulmonary Disease Gene study (COPDGene, $N = 10,674$), the Estonian Biobank (EST, $N = 2,244$), the Japan Biological Informatics Consortium (JPN, $N = 295$), the Biobank Japan (JPN, $N = 1,025$) and the 1000 Genomes Project (1KG, $N = 2,504$). Each study was previously approved by respective institutional review boards (IRBs), including for the generation of whole genome sequencing (WGS) data and association with phenotypes. All participants provided written consent. Details of each cohort included in the reference panel construction is described below.

1.1 Jackson Heart Study (JHS)

The Jackson Heart Study (JHS) (`https://www.jacksonheartstudy.org`) is a large, population-based observational study evaluating the etiology of cardiovascular, renal, and respiratory diseases among African Americans residing in the three counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi metropolitan area. Data and biologic materials have been collected from 5,306 participants, including a nested family cohort of 1,498 members of 264 families. The age at enrollment for the unrelated cohort was 35-84 years; the family cohort included related individuals > 21 years old. Participants provided extensive medical and social history, had an array of physical and biochemical measurements and diagnostic procedures, and provided genomic DNA during a baseline examination (2000-2004) and two follow-up examinations (2005-2008 and 2009-2012). The study

population is characterized by a high prevalence of diabetes, hypertension, obesity, and related disorders. Annual follow-up interviews and cohort surveillance are ongoing, and preparation for a fourth examination is in progress.

1.2 Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis (MESA) is a study of the characteristics of subclinical cardiovascular disease and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MESA consisted of a diverse, population-based sample of an initial 6,814 asymptomatic men and women aged 45-84. 38 percent of the recruited participants were white, 28 percent African American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and University of California - Los Angeles. Each participant received an extensive physical exam and determination of coronary calcification, ventricular mass and function, flow-mediated endothelial vasodilation, carotid intimal-medial wall thickness and presence of echogenic lucencies in the carotid artery, lower extremity vascular insufficiency, arterial wave forms, electrocardiographic (ECG) measures, standard coronary risk factors, sociodemographic factors, lifestyle factors, and psychosocial factors. Selected repetition of subclinical disease measures and risk factors at follow-up visits allowed study of the progression of disease. Participants are being followed for identification and characterization of cardiovascular disease events, including acute myocardial infarction and other forms of coronary heart disease (CHD), stroke, and congestive heart failure; for cardiovascular disease interventions; and for mortality. The first examination took place over two years,

from July 2000 - July 2002. It was followed by four examination periods that were 17-20 months in length. Participants have been contacted every 9 to 12 months throughout the study to assess clinical morbidity and mortality.

1.3 The Chronic Obstructive Pulmonary Disease Gene (COPDGene) study

Eligible subjects in The Chronic Obstructive Pulmonary Disease Gene (COPDGene) study Study (NCT00608764, `www.copdgene.org`) were of non-Hispanic white (NHW) or African-American (AA) ancestry, aged 45-80 years old, with at least 10 pack-years of smoking and no diagnosed lung disease other than COPD or asthma [1]. IRB approval was obtained at all study centers, and all study participants provided written informed consent.

M.H.C. was supported by NHLBI grants R01HL113264, R01HL137927, and R01HL135142. The COPDGene project (NCT00608764) was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

COPD Foundation Funding
The COPDGene® project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

COPDGene® Investigators – Core Units

Administrative Center: James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J. Make, MD; Elizabeth A. Regan, MD, PhD

Genetic Analysis Center: Terri Beaty, PhD; Ferdouse Begum, PhD; Peter J. Castaldi, MD, MSc; Michael Cho, MD; Dawn L. DeMeo, MD, MPH; Adel R. Boueiz, MD; Marilyn G. Foreman, MD, MS; Eitan Halper-Stromberg; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS, MPH; Brian D. Hobbs, MD; John E. Hokanson, MPH, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Margaret M. Parker, PhD; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; Sungho Won, Ph.D.; Phuwanat Sakornsakolpat, M.D.; Dmitry Prokopenko, Ph.D.

Imaging Center: Mustafa Al Qaisi, MD; Harvey O. Coxson, PhD; Teresa Gray; MeiLan K. Han, MD, MS; Eric A. Hoffman, PhD; Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber; David A. Lynch, MB; John D. Newell, Jr., MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Douglas Stinson; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; George Washko, MD; Carla G. Wilson, MS;

PFT QA Center, Salt Lake City, UT: Robert Jensen, PhD

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO: Douglas Everett, PhD; Jim Crooks, PhD; Camille Moore, PhD; Matt Strand, PhD; Carla G. Wilson, MS

Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO: John E. Hokanson, MPH, PhD; John Hughes, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Katherine Pratte, MSPH; Kendra A. Young, PhD

Mortality Adjudication Core: Surya Bhatt, MD; Jessica Bon, MD; MeiLan K. Han,

MD, MS; Barry Make, MD; Carlos Martinez, MD, MS; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; Carla G. Wilson, MS

Biomarker Core: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush Banaei-Kashani, Ph.D.

COPDGene® Investigators – Clinical Centers

Ann Arbor VA: Jeffrey L. Curtis, MD; Carlos H. Martinez, MD, MPH; Perry G. Pernicano, MD.

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Philip Alapat, MD; Mustafa Atik, MD; Venkata Bandi, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Arun Nachiappan, MD; Amit Parulekar, MD;

Brigham and Women's Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD

Columbia University, New York, NY: R. Graham Barr, MD, DrPH; John Austin, MD; Belinda D'Souza, MD; Gregory D.N. Pearson, MD; Anna Rozenshtein, MD, MPH, FACR; Byron Thomashow, MD Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD; H. Page McAdams, MD; Lacey Washington, MD

HealthPartners Research Institute, Minneapolis, MN: Charlene McEvoy, MD, MPH; Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD; Robert Brown, MD; Nadia N. Hansel, MD, MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD, MHS

Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer, MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD

Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD; Charlie

Lan, DO Minneapolis VA: Christine Wendt, MD; Brian Bell, MD

Morehouse School of Medicine, Atlanta, GA: Marilyn G. Foreman, MD, MS; Eugene Berkowitz, MD, PhD; Gloria Westney, MD, MS.

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD; David A. Lynch, MB

Reliant Medical Group, Worcester, MA: Richard Rosiello, MD; David Pace, MD

Temple University, Philadelphia, PA: Gerard Criner, MD; David Ciccolella, MD; Francis Cordova, MD; Chandra Dass, MD; Gilbert D'Alonzo, DO; Parag Desai, MD; Michael Jacobs, PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD; Irene Swift, MD; Maria Elena Vega-Sanchez, MD

University of Alabama, Birmingham, AL: Mark Dransfield, MD; William Bailey, MD; Surya Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD

University of California, San Diego, CA: Joe Ramsdell, MD; Paul Friedman, MD; Xavier Soler, MD, PhD; Andrew Yen, MD

University of Iowa, Iowa City, IA: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John Newell, Jr., MD; Brad Thompson, MD

University of Michigan, Ann Arbor, MI: MeiLan K. Han, MD, MS; Ella Kazerooni, MD; Carlos H. Martinez, MD, MPH

University of Minnesota, Minneapolis, MN: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD; Jessica Bon, MD; Divay Chandra, MD, MSc; Carl Fuhrman, MD; Joel Weissfeld, MD, MPH

University of Texas Health Science Center at San Antonio, San Antonio, TX: Antonio Anzueto, MD; Sandra Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD

1.4 Description for the 1,320 Japanese subjects (JPN)

Genomic DNA of 295 Japanese subjects were obtained from Epstein-Barr virus transformed B-lymphoblast cell lines of unrelated Japanese individuals established by the Japan Biological Informatics Consortium (JBIC, `http://www.jbic.or.jp/english`). WGS analysis was conducted as described elsewhere [2]. Briefly, WGS library was constructed using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) according to the manufacturer's protocols. All subjects were sequenced using 2×150-bp paired end reads on a HiSeq X Five (Illumina). The sequence reads were converted to the FASTQ format using `bcl2fastq2` (version 2.17.1.14) and trimmed to clip Illumina adapters using `Trimmomatic` (version 0.36). They were aligned to the reference human genome with the decoy sequence (GRCh37/hg19, human_g1k_v37_decoy) using `BWA-MEM` (version 0.7.5a).

A separate 1,025 Japanese individuals were enrolled from BioBank Japan Project (BBJ), as described elsewhere [3]. They were affected with any of the five diseases (acute myocardial infarction, drug eruption, colorectal cancer, breast cancer, prostate cancer). WGS data of the subjects were obtained from the National Bioscience Database Center (NBDC) Human Database (`https://humandbs.biosciencedbc.jp/en/`, ID: hum0014) and Japanese Genotype-phenotype Archive (JGA; `https://www.ddbj.nig.ac.jp/jga/index.html`, ID: JGAS00000000114). WGS analysis was conducted as described elsewhere [3]. Briefly, WGS library was constructed using the TruSeq Nano DNA Library Preparation Kit (Illumina) according to the manufacturer's protocols. All subjects were sequenced using 2×160-bp paired end reads on a HiSeq 2500 (Illumina). The sequence reads were converted to the FASTQ format using `bcl2fastq2` (version 2.17.1.14) and trimmed to clip Illumina adapters using `Trimmomatic` (version 0.36). They were

aligned to the reference human genome with the decoy sequence (GRCh37/hg19, human_g1k_v37_decoy) using `BWA-MEM` (version 0.7.5a).

1.5 Description of the 2,244 Estonian subjects (EST)

The 2,244 Estonian subjects were enrolled from the Estonian Biobank of the Estonian Genome Center. All samples followed a PCR-free sample preparation. Libraries sequenced on the Illumina HiSeq X Ten at 30x coverage. The details of sample selection and processing were described previously[4].

1.6 1000 Genomes Project (1KG)

All deep-coverage whole genome sequencing data of the 1000 Genomes Project as described in [5] were downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage`. A subset of 1,267 individuals covers 14 populations and four major continental groups underwent Sequence-based typing (SBT) for typing HLA three class I genes (*HLA-A*, *HLA-B* and *HLA-C*) and two class II genes (*HLA-DRB1* and *HLA-DQB1*) at G-group resolution[6]. The SBT HLA genotype was downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_genotypes/20140702_hla_diversity.txt`.

2. Construction of a multi-ethnic HLA reference panel

2.1 Read mapping

Whole genome-sequencing reads alignment was performed by each cohort. Due to changes in the informatics pipeline over the course of different projects, two different versions of human reference were used:

- The primary GRCh37 primary assembly with human herpesvirus and the concatenated decoy sequences (GRCh37/hg19, human_g1k_v37_decoy). Cohorts mapped to this version of reference includes the JHS, EST, and JPN.
- The primary GRCh38 assembly. Cohorts mapped to this version of reference includes the MESA, the COPDGene study and 1GK.

2.2 Inference of HLA classical alleles from whole-genome sequencing

To performed HLA typing using whole-genome sequencing data, we first used `samtools` (version 1.3) to extract the extend MHC region (chr6:25,000,000-35,000,000) and all unmapped reads from aligned reads:

```
samtools view -h $inputfile.bam  6:25000000-35000000 \
-O bam -o $inputfile.MHC.bam
samtools view -f0x4 -h $inputfile.bam -O bam -o $inputfile.MHC_unmapped.bam
```

We then merged the extended MHC region and all unmapped reads with:

```
samtools merge -h $sample.MHC.bam $sample\_merged.bam \
$sample.MHC.bam $sample.unmapped.bam
samtools sort $sample\_merged.bam -O b -o $sample.bam
samtools index $sample.bam
```

We used HLA*PRG [7] to perform HLA typing in six classical HLA genes (HLA-*A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*) in the earlier collection of the two cohorts (JHS and EST) when HLA*LA [8] was not available. We then performed HLA typing in eight classical HLA genes using HLA*LA (https://github.com/DiltheyLab/HLA-LA/blob/master/) for all samples included in the JPN, COPDGene, MESA, 1000G and the 1000 Genomes Project (HLA-*A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*, *-DPA1*, *-DPB1*).

```
  perl5.8.9 HLA-PRG-LA.pl \
--BAM $sample.bam \
--graph PRG_MHC_GRCh38_withIMGT \
--sampleID $sample \
--maxThreads $n
```

Briefly, both HLA*PRG and HLA*LA constructs a directed graph in which alternative alleles, insertions and deletions are represented as alternative paths using whole-genome sequencing data.

2.3 Variant calling

To construct a HLA imputation we used variants (SNPs and Indels) called within the extend MHC region (chr6:25,000,000-35,000,000) either from vcfs provided by each individual cohort (MESA, COPDGene and 1KG) or performed using mapped reads described in Section 2.1 following GATK [9] (version 3.6) best practices (EST, JHS and JPN).

Briefly, we first used HaplotypeCaller to perform per-sample variant calling with the following command:

```
java -jar GenomeAnalysisTK.jar \
  -T HaplotypeCaller -R Homo_sapiens_assembly19.fasta \
  -I $sample.bam \
  --emitRefConfidence GVCF \
  -L 6:25000000-35000000 \
  -o $sample.g.vcf.gz \
  -variant_index_type LINEAR \
  -variant_index_parameter 128000
```

Then we combined multiple sample *gvcfs* using `GenotypeGVCFs`:

```
java -jar GenomeAnalysisTK.jar \
  -T GenotypeGVCFs \
  -R Homo_sapiens_assembly19.fasta \
  --variant sample.list \
  -o all.vcf.gz \
  --max_alternate_alleles 2
```

We next unified all variants called in individual cohort using `CombineVariants`.

2.4 Quality control

2.4.1 Variant QC

Post variant calling, we applied the following quality control (QC) criteria to each variant that had been called:

- variant must overlap with sites discovered in 1000 Genomes Phase III release [5].
- in the extended MHC region (chr6:28-34Mb)

- minor allele frequency $\geq 0.5\%$ in each cohort
- SNPs only
- missingness rate $\leq 5\%$
- variants within the eight classical HLA gene body (**Supplementary Table 17**)

These QC criteria have been applied both on individual cohorts. We did not filter variants that are out of Hardy-Weinberg equilibrium due to extreme high linkage disequilibrium in this region[10]. In total, there are 38,398 SNPs included in the reference panel.

2.4.2 Sample QC

Since individual cohort were sequenced at different depth, to ensure quality of inferred HLA types, we first restrict samples that have more than 20x in each of the eight HLA genes. The average coverage in the MHC region is summarized in **Supplementary Table 1**. Next we excluded samples that failed genome-wide QC. In total, there are 21,546 individuals included in the reference panel.

References

[1] Regan, E. A. *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD* **7**, 32–43 (2010).

[2] Okada, Y. *et al.* Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of japanese. *Nat. Commun.* **9**, 1631 (2018).

[3] Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the japanese population. *Nat. Genet.* (2019).

[4] Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).

[5] 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[6] Gourraud, P.-A. *et al.* HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, e97282 (2014).

[7] Dilthey, A. T. *et al.* High-Accuracy HLA type inference from Whole-Genome sequencing data using population reference graphs. *PLoS Comput. Biol.* **12**, e1005151 (2016).

[8] Dilthey, A. T. *et al.* HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019).

[9] McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

[10] Graffelman, J., Jain, D. & Weir, B. A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. *Hum. Genet.* **136**, 727–741 (2017).