Supplementary Materials

Contents

Supplementary materials 1

I. **Ethnicity classification by genetic data:** 1

II. **HLA sequencing for DILI cases** 8

III. **Exclusion of potential questionable imputed HLA alleles:** 8

IV. **Information of patients carrying *HLA-B*14:01, HLA-B*14:02*, or *HLA-C*08:02*** 10

V. ***HLA-B*14:01* haplotype association results** 11

VI. ***HLA-B*14:01* distribution in different drug subsets in DILIN** 12

VII. **Patient characteristic comparison between patients with and without *HLA-B*14:01* in European Americans** 13

I. Ethnicity classification by genetic data:

It is known that HLA allele distributions differ by ethnic groups or even different regions within the same ethnicity [1]. Considering that self-reported race/ethnicity is not completely accurate [2], we utilized genome-wide SNP data to infer ethnicity for each person. In the DILIN cohort, 1,886 patients had SNP genotyping performed over the years by different Illumina SNP array platforms, including Human 1MDuo (N=542), Human Core Exome (HCE, N=179)), Multi-Ethnic Global Array (MEGA, N=847), and Expanded Multi-Ethnic Genotyping Array (MEGA^{EX}, N=318). The genotyping protocol was similar among all platforms, as previously described [3-5]. In order to determine ethnicity by genetic variants, we used the 1000 genomes (1KG) data from 26 populations as the reference. We grouped 26 population to five ethnicity categories (Caucasian with European ancestry, African ancestry, Hispanic, East and South Asian) (**Table S1**). Since SNP panel is different among different SNP arrays, the analysis described below was performed for each SNP array used in DILIN cohort separately. First, we combined DILIN SNP data from the same SNP array with 1KG data to compute principal components (PCs) using the Eigensoft program [6]. We then used PCs of 1KG data to generate multiple PC clustering plots by comparing PC1 to other PCs. As the ethnicity is known in the 1KG dataset, we were able to identify key PCs that led to clear ethnicity-specific cluster separation from other clusters. These PCs were selected

as the criteria to determine genetic-inferred ethnicity. For example, South Asian cluster was separated from other clusters in PC1 vs. PC3 (**Figure S1**). Therefore, PC1 and PC3 will be able to infer South Asians in DILIN. In general, Caucasian, African ancestry, South Asian, and East Asian clusters of 1KG data showed more clear clustering and separation with other clusters in certain PC comparisons, but Hispanics group did not have clear pattern of separation. Based on the clustering patterns, we considered the following key PCs for determining each ethnicity: PC1 and PC2 for Caucasians with European ancestry, PC1-PC4 for African ancestry, PC1 and PC3 for South Asians, PC1, PC2, and PC4 for East Asians, and PC1 and PC3 for Hispanics. Next, we set the boundary of each PC as mean±3SD of the PC from the same ethnicity in the 1KG data. These boundaries were used to determine the ethnicity of DILIN subjects using the following criteria. All PCs indicated below refer to the PCs generated from the combined DILIN and 1KG data.

(1) European American: Individuals with their $PC1_{DILIN}$ and $PC2_{DILIN}$ located within the boundaries of $PC1_{1KG}$ and PC_{1kG} where PC boundaries were computed from PCs of Caucasians in 1KG dataset.

(2) African American: Individuals with their $PC1_{DILIN}$ and $PC2_{DILIN}$ located within the boundaries of $PC1_{1KG}$ and $PC2_{1KG}$, $PC1_{DILIN}\ and\ PC3_{DILIN}$ within the boundaries of $PC1_{1KG}\ and\ PC3_{1KG}$, or $PC1_{DILIN}\ and\ PC4_{DILIN}$ within the boundaries of $PC1_{1KG}$ and $PC4_{1KG}$, where the PC boundaries were computed from PCs of Africans in 1KG dataset.

(3) South Asian Individuals with $PC1_{DILIN}$ and $PC3_{DILIN}$ located within the boundaries of $PC1_{1kG}\ and\ PC3_{1KG}$, where PC boundaries were computed from PCs of South Asian in 1KG dataset.

(4) East Asian: Individuals with $PC1_{DILIN}\ and\ PC2_{DILIN}$ located within the boundaries of $PC_{1kG}and\ PC2_{1KG}$, or $PC1_{DILIN}\ and\ PC4_{DILIN}$ within the boundaries of $PC1_{1KG}and\ PC4_{1KG}$, where PC boundaries were computed based on PCs of East Asians in 1KG dataset.

(5) Hispanic: For individuals who are not belong to any of four ethnicity groups above, if their $PC1_{DILIN}and\ PC3_{DILIN}$ are within the boundaries of $PC1_{1KG}\ and\ PC3_{1KG}$, they were designated to Hispanics group. The PC boundaries were computed based on the PCs of Hispanics in 1KG data.

However, 3SD range from the mean of each PC may miss some subjects in the borderline or resulted in some overlapping ethnicity assignment. For these subjects, we compared their locations in the PC clustering plot of DILIN data to 1KG data and make a subjective decision to refer their ethnicity. Finally, for those subjects without PC data due to the lack of SNP array data, self-reported ethnicity was used. **Table S2** summarized the sample count comparison between self-reported and genetic inferred ethnicity, and also list the count of

subjects (3 European American, 2 Asian, and 1 Hispanic) with race assigned by self-reported ethnicity for high confident TMP-SMZ cases. It should be noted that self-reported ethnicity did not distinguish Asian to East and South Asian. The numbers of patients with matched self-reported and genetic inferred ethnicity are 47 European American, 10 African American, 2 East Asian, and 4 Hispanics. One self-reported Hispanics was designated as European American, while one self-reported European American was designed as Hispanic. There are three European Americans, two Asians, and one Hispanics were based on self-reported data. Finally, **Figure S2** contrasts the PC1 vs. PC2 of the inferred ethnicity of DILIN patients to 1KG data by SNP array platforms, which demonstrates the consistent pattern of ethnic clustering indicating the robustness of our approach. In conclusion, 72 TMP-SMX DILI patients were grouped to 51 European American, 10 African American, 4 Asian, and 6 Hispanic groups.

Table S1: ethnicity grouping for 26 populations in the 1000 genomes project

Population Name	Population Abbreviation	1000 Genomes Super population	Ethnicity used in this study
Han Chinese in Beijing, China	CHB	East Asian	East Asian
Japanese in Tokyo, Japan	JPT	East Asian	East Asian
Southern Han Chinese	CHS	East Asian	East Asian
Chinese Dai in Xishuangbanna, China	CDX	East Asian	East Asian
Kinh in Ho Chi Minh City, Vietnam	KHV	East Asian	East Asian
Utah Residents (CEPH) with Northern and Western European Ancestry	CEU	European	European American
Toscani in Italia	TSI	European	European American
Finnish in Finland	FIN	European	European American
British in England and Scotland	GBR	European	European American
Iberian Population in Spain	IBS	European	European American
Yoruba in Ibadan, Nigeria	YRI	African	African American
Luhya in Webuye, Kenya	LWK	African	African American
Gambian in Western Divisions in the Gambia	GWD	African	African American
Mende in Sierra Leone	MSL	African	African American
Esan in Nigeria	ESN	African	African American
Americans of African Ancestry in SW USA	ASW	African	African American
African Caribbeans in Barbados	ACB	African	African American

Mexican Ancestry from Los Angeles USA	MXL	Admixed American	Hispanic
Puerto Ricans from Puerto Rico	PUR	Admixed American	Hispanic
Colombians from Medellin, Colombia	CLM	Admixed American	Hispanic
Peruvians from Lima, Peru	PEL	Admixed American	Hispanic
Gujarati Indian from Houston, Texas	GIH	South Asian	South Asian
Punjabi from Lahore, Pakistan	PJL	South Asian	South Asian
Bengali from Bangladesh	BEB	South Asian	South Asian
Sri Lankan Tamil from the UK	STU	South Asian	South Asian
Indian Telugu from the UK	ITU	South Asian	South Asian

Figure S1: PC clustering by PC1 vs. PC2, PC3, and PC4 of 1000 genomes data, where PCs were computed based on the merged DILIN and 1000 genomes data by each SNP array used in DILIN

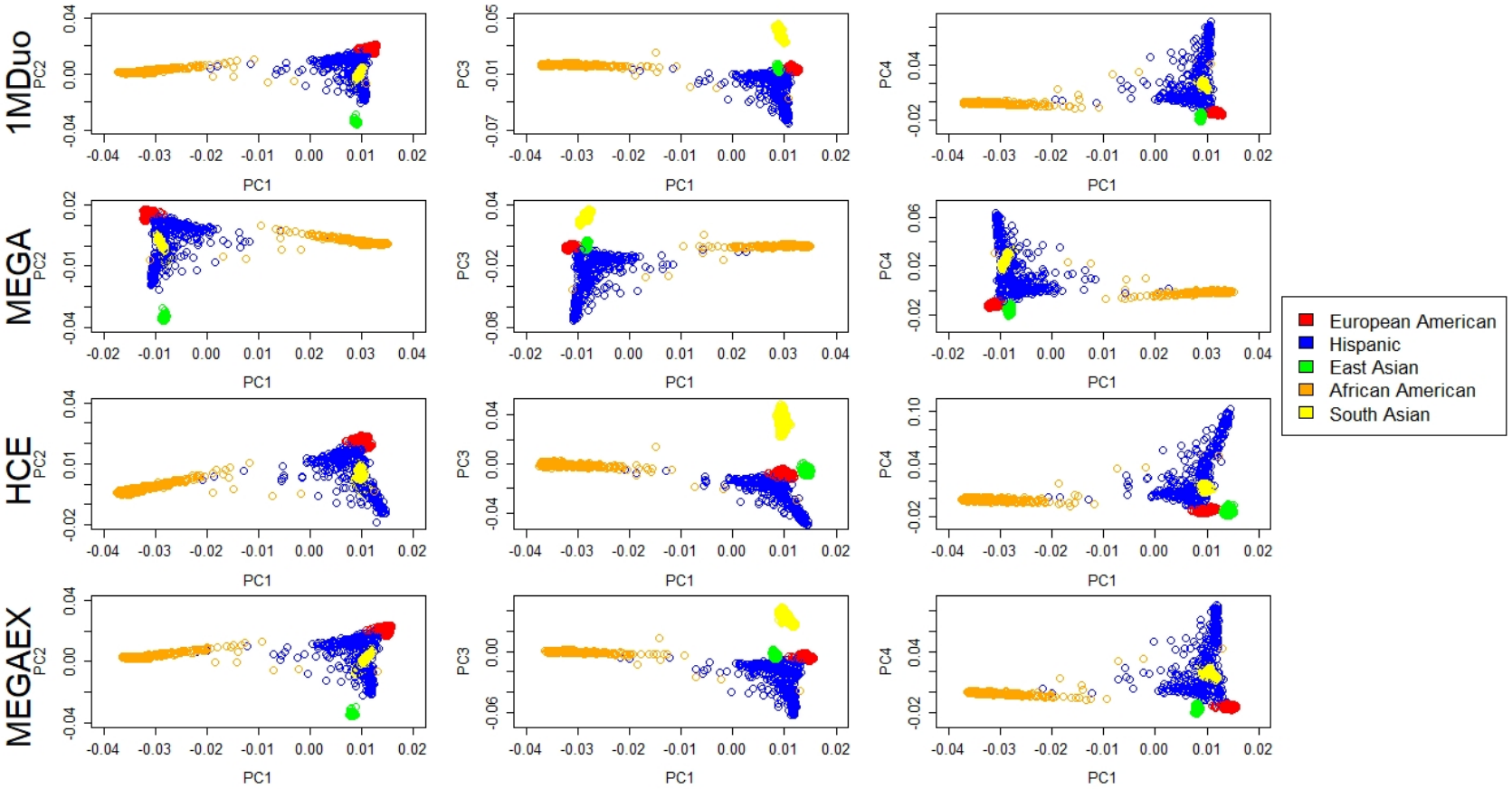
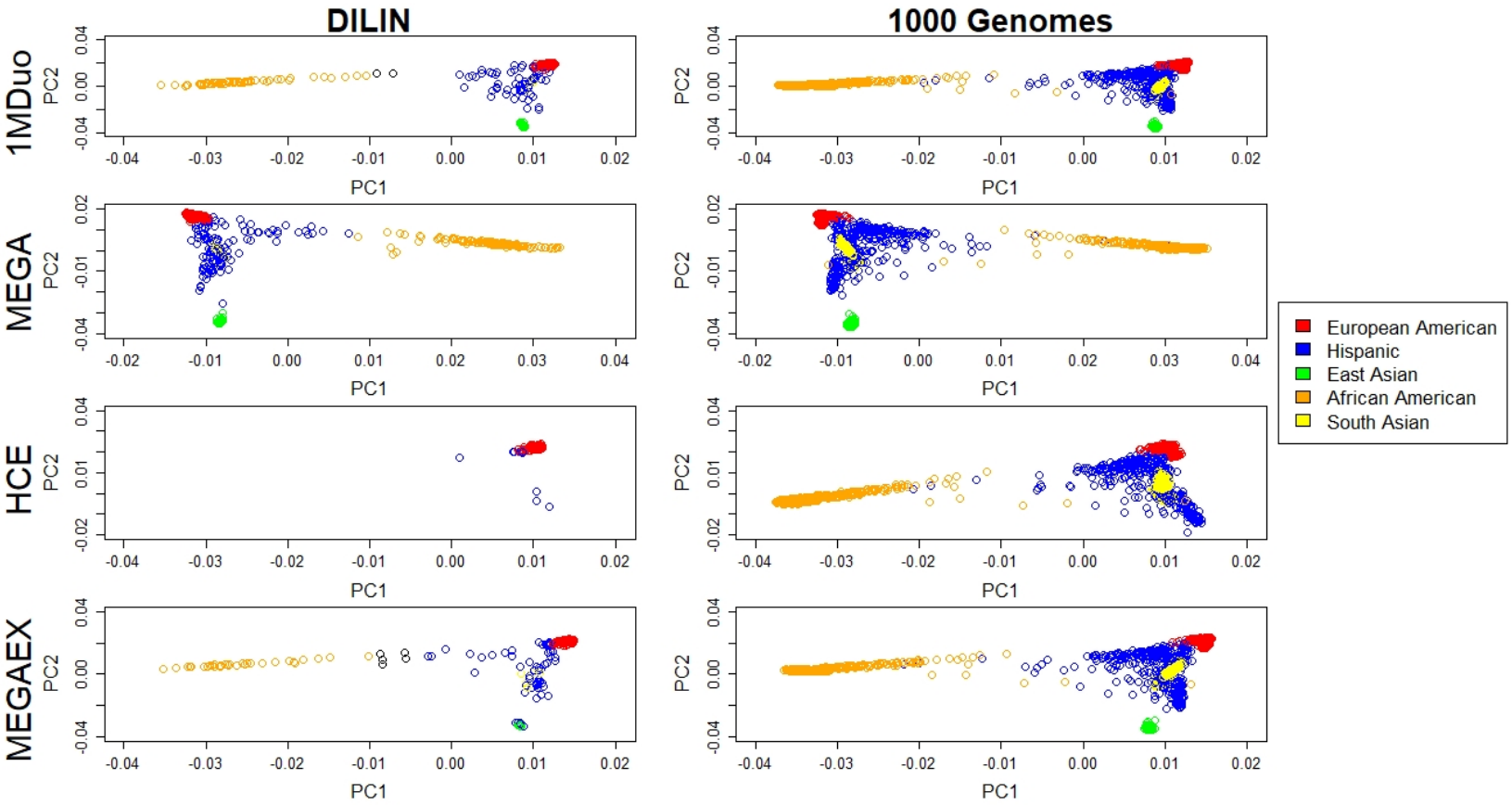


Table S2: Number of patients comparing self-reported vs. genetically-inferred ethnicity in High Confidence Bactrim Patients.

	Genetically-inferred ethnicity				
Self-reported ethnicity	**European American**	**African American**	**East Asian**	**South Asian**	**Hispanics**
European ancestry	47	0	0	0	1
African ancestry	0	10	0	0	0
Asian	0	0	2	0	0
Hispanics	1	0	0	0	4
Assigned by self-reported ethnicity	3	0	2		1
Total	51	10	4		6

One patient not in the table was assigned by self-reported ethnicity as other ethnicity/multiracial.

Figure S2: Comparison of genetically-inferred ethnicity of DILIN data to 1000 genomes data by PC1 vs. PC2, where PCs were computed based on the merged DILIN and 1000 genomes data by each SNP array used in DILIN



II. HLA sequencing for DILI cases

DNA samples extracted from 1,916 DILIN participants were used for HLA deep sequencing. HLA Class I and II gene sequencing was performed in Vanderbilt Immunogenomics, Microbial Genetics and Single Cell Technologies core (IMGSCT), a laboratory accredited by the American Society for Histocompatibility and Immunogenetics (ASHI). Specific HLA loci were PCR amplified using sample specific MID-tagged primers that amplify polymorphic exons from Class I (A, B, C Exons 2 and 3) and Class II (DQ, Exons 2 and 3; DRB and DPB1, Exon 1) MHC genes. MID tagged primers were optimized to minimize allele dropouts and primer bias. Amplified DNA products from unique MID tagged products (up to 48 MIDs) were pooled in equimolar ratios for library preparation. Libraries were quantified using the KAPA library quantitation kit (Kapa Biosystems) and High sensitivity D1000 screentape on an Agilent 2200 Tapestation (Agilent) for concentration and size distribution. Normalized libraries were sequenced on the Illumina MiSeq platform using the MiSeq V3 600-cycle kit (2X300bp reads). Sequences were separated by MID tags and alleles called using an in house accredited HLA allele caller software pipeline that minimizes the influence of sequencing errors. Alleles were called using the IMGT HLA allele sequence database release v3270 (March 24, 2017), IIID HLA analysis suite release v3.11, and IIID allele caller release v2.7 HLA analyse reporting software that performs comprehensive allele balance and contamination checks on the final dataset.

III. Exclusion of potential questionable imputed HLA alleles:

As HLA alleles from eMERGE-I and PAGE were based on imputation, we compared the AF of imputed HLA alleles to population AF. In our analysis we excluded those alleles with absolute AF difference greater than 0.05. **Table S3** listed the HLA alleles excluded in European American and African American subsets, respectively.

Table S3: List of HLA alleles to be excluded due to allele frequency (AF) of imputed HLA allele deviated from population AF more than 0.05

HLA gene		European American			African American		
	HLA Allele	Population	eMERGE1	Absolute difference	Population	PAGE	Absolute difference
DQB1	DQB1*06:03	0	0.066	0.066			

	DQB1*02:01	0.230	0.126	0.104	0.223	0.105	0.118
DQA1	DQA1*05:05	0.003	0.114	0.111	0	0.098	0.098
	DQA1*05:01	0.245	0.126	0.119	0.188	0.109	0.079
DPB1	DPB1*06:02				0	0.132	0.132
	DPB1*04:02				0.111	0	0.111
	DPB1*02:01				0.130	0.183	0.053
	DPB1*01:01				0.272	0.371	0.099

IV. Information of patients carrying *HLA-B*14:01, HLA-B*14:02*, or *HLA-C*08:02*

In Table S4, we provide the basic clinical and Class I gene information for eight patients who carry *HLA-B*14:01, HLA-B*14:02*, or *HLA-C*08:02* in European Americans. A clear strong linkage disequilibrium is observed among *HLA-B*14:01, HLA-B*14:02*, and *HLA-C*08:02*. All five patients with *HLA-B*14:01* and three patients with *HLA*B-14:02* carry *HLA-C*08:02*.

Table S4: List of patients who carry *HLA-B*14:01, HLA-B*14:02*, or *HLA-C*08:02* in European Americans

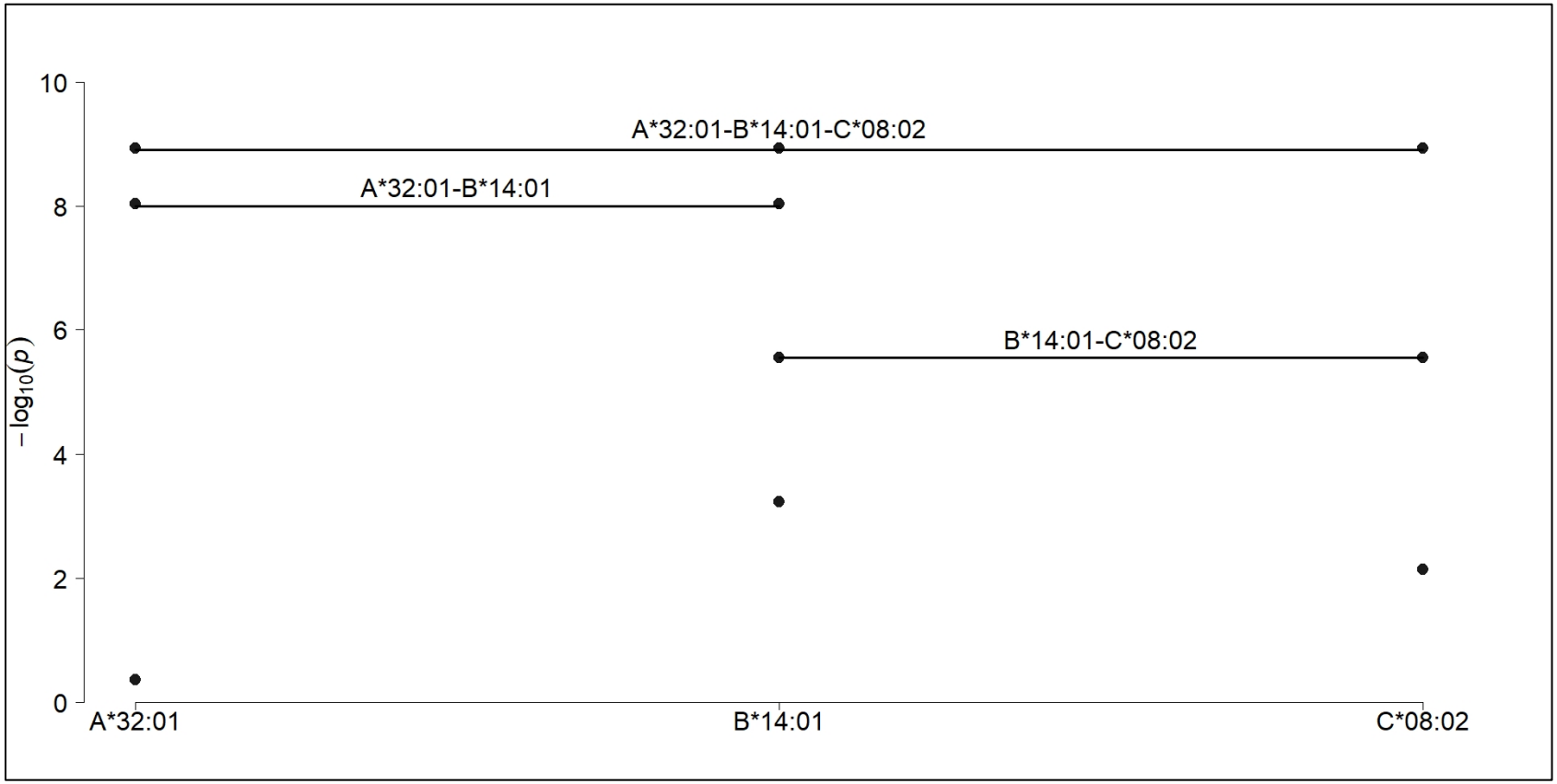
					HLA-A		HLA-B		HLA-C	
ID	**AGE**	**GENDER**	**Injury types**	**Assessment rating**	**ALLELE1**	**ALLELE2**	**ALLELE1**	**ALLELE2**	**ALLELE1**	**ALLELE2**
1	23.3	Female	Cholestatic	Probable 50-75%	A*01:01	A*32:01	B*14:01	B*35:01	C*04:01	C*08:02
2	30.1	Male	Hepatocellular	Highly likely 75-95%	A*01:01	A*26:01	B*08:01	B*14:01	C*07:01	C*08:02
3	67.2	Male	Cholestatic	Probable 50-75%	A*02:01	A*11:01	B*14:02	B*55:01	C*03:03	C*08:02
4	40.9	Male	Mixed	Highly likely 75-95%	A*01:01	A*33:01	B*14:02	B*37:01	C*06:02	C*08:02
5	33.3	Female	Mixed	Highly likely 75-95%	A*02:01	A*11:01	B*14:01	B*35:01	C*04:01	C*08:02
6	51.9	Female	Mixed	Probable 50-75%	A*03:01	A*32:01	B*14:01	B*49:01	C*07:01	C*08:02
7	72.7	Male	Mixed	Highly likely 75-95%	A*01:01	A*68:01	B*08:01	B*14:01	C*07:01	C*08:02
8	62.0	Female	Cholestatic	Probable 50-75%	A*03:01	A*32:01	B*14:02	B*44:02	C*05:01	C*08:02

Note: five patients carrying *HLA-B*14:01*, two patients carrying *HLA-B*14:02*, and eight patients carrying *HLA-C*08:02*

V. *HLA-B*14:01* haplotype association results

Haplotype association tests were performed for HLA-B*14:01 for the combination with HLA Class I A and B genes as the following: A-B, B-C, and A-B-C using haplo.stat program. The most significant haplotypes for HLA-B*14:01 was with *HLA-A*32:01* and *HLA-C*08:02*. In Figure S3, we depicted the haplotype results using $-\log_{10} p$ and compare them to the allelic association results.

Figure S3: Comparison of haplotype and allelic association results for *HLA-A*32:01*, *HLA-B*14:01*, and *HLA-C*08:02.* P-values are presented as $-\log_{10} p$. Haplotypes are connected by horizontal lines while the individual allele association was presented in dots.



VI. *HLA-B*14:01* distribution in different drug subsets in DILIN

Table S5: *HLA-B*14:01* distribution in different drug subsets of the DILIN cohort

Drug	Total Alleles	B*14:01 allele count	AF	Total patients	B14:01 carrier count	CF
Sulfamethoxazole W/Trimethoprim	102	5	0.049	51	5	0.098
Amoxicillin W/Clavulanic Acid	374	3	0.008	187	3	0.016
Nitrofurantoin	120	3	0.025	60	3	0.050
Isoniazid	64	2	0.031	32	2	0.063
Minocycline	64	2	0.031	32	2	0.063
Cefazolin	44	1	0.023	22	1	0.046

AF: allele frequency; CF: carriage frequency

Table S6: Patient counts for non-antibiotic sulfonamides drugs in the DILIN Cohort

Drug	European Americans*	African Americans
Celecoxib	3	0
Darunavir	0	2 (one patient with HLA-B*35:01)
Sulfasalazine	3	0
Zonisamide	1	0

*None of seven European Americans carry *HLA-B*14:01*.

VII. Patient characteristic comparison between patients with and without *HLA-B*14:01* in European Americans

Table S7: Selected clinical and laboratory characteristics of European Americans with and without HLA B*14:01

	without HLA*B 14:01 (N=46)	with HLA*B14:01 (N=5)	P-value
Age (years, mean [SD])	49.3 (20.21)	42.3 (20.04)	0.7
Females	54.3%	60.0%	1.0
BMI (kg/m^2, mean [SD])	25.8 (6.89)	25.8 (2.72)	1.0
Diabetes mellitus	15.2%	0.0%	1.0
Latency (days in median, IQR)	21.0 (10.0, 33.0)	29.0 (22.0, 33.0)	0.56
Jaundice	60.9%	60.0%	1.0
Peripheral eosinophilia	16%	40%	0.4
Liver Biochemistries – Peak values			
ALT (U/L, mean [SD])	1049.9 (1800.87)	541.4 (219.89)	1.0
Alk P (U/L, mean [SD])	524.7 (372.80)	375.2 (59.46)	0.9
Total bilirubin (mg/dl, mean [SD])	10.6 (10.72)	10.7 (3.70)	0.5
INR	1.3 (0.50)	2.2 (2.17)	0.9
Severity of Liver Injury			0.9
Mild	17.4%	0.0%	
Moderate	28.3%	60.0%	
Moderate-hospitalized	30.4%	20.0%	
Severe	19.6%	20.0%	
Fatal	4.3%	0.0%	
Death	4.9%	0.0%	0.9
Liver Transplantation	0.0%	0.0%	-
Chronic DILI	20.5%	0.0%	0.7

Abbreviations: ALT, serum alanine aminotransferase; AP, serum alkaline phosphatase; BMI, body mass index; DILI, drug-induced liver injury; INR, international normalized ratio; IQR, interquartile range (25-75%); ULN, upper limit of normal

Reference

1. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J 2007;48:11-23.
2. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum Genomics 2015;9:1.
3. Urban TJ, Shen Y, Stolz A, et al. Limited contribution of common genetic variants to risk for liver injury due to a variety of drugs. Pharmacogenet Genomics 2012;22:784-95.
4. Nicoletti P, Aithal GP, Bjornsson ES, et al. Association of Liver Injury From Specific Drugs, or Groups of Drugs, With Polymorphisms in HLA and Other Genes in a Genome-Wide Association Study. Gastroenterology 2017;152:1078-1089.
5. Cirulli ET, Nicoletti P, Abramson K, et al. A Missense Variant in PTPN22 is a Risk Factor for Drug-induced Liver Injury. Gastroenterology 2019;156:1707-1716 e2.
6. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-9.