The American Journal of Human Genetics, Volume *106*

Supplemental Data

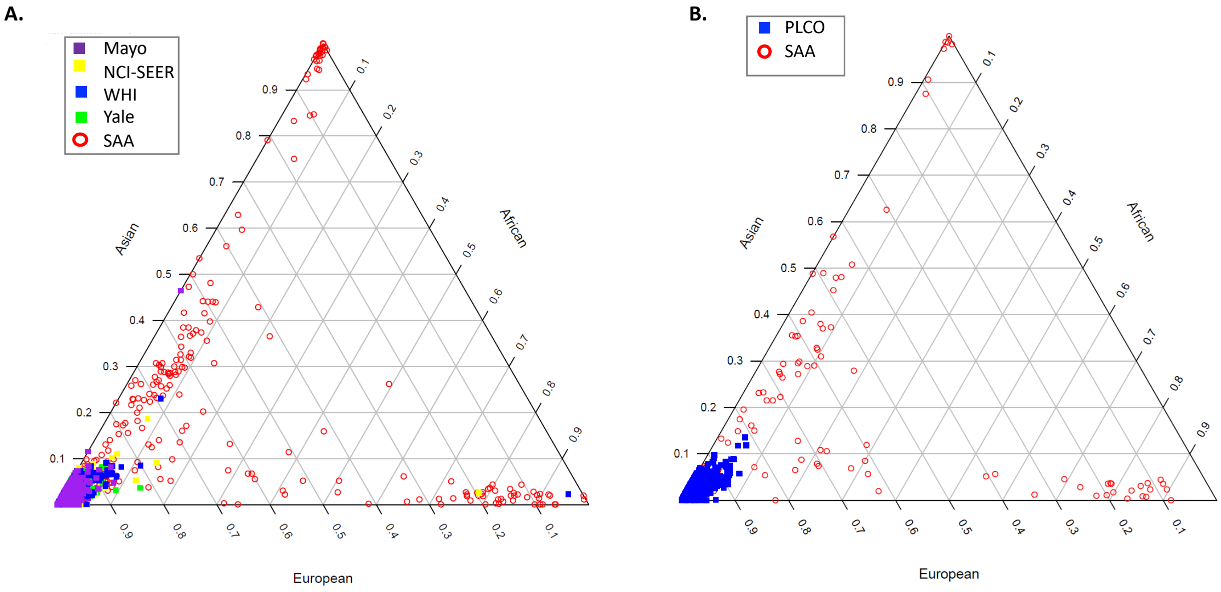
Genome-wide Association Study Identifies *HLA-DPB1* as a Significant Risk Factor for Severe Aplastic Anemia

Sharon A. Savage, Mathias Viard, Colm O'hUigin, Weiyin Zhou, Meredith Yeager, Shengchao Alfred Li, Tao Wang, Veron Ramsuran, Nicolas Vince, Aurelie Vogt, Belynda Hicks, Laurie Burdett, Charles Chung, Michael Dean, Kelvin C. de Andrade, Neal D. Freedman, Sonja I. Berndt, Nathaniel Rothman, Qing Lan, James R. Cerhan, Susan L. Slager, Yawei Zhang, Lauren R. Teras, Michael Haagenson, Stephen J. Chanock, Stephen R. Spellman, Youjin Wang, Amanda Willis, Medhat Askar, Stephanie J. Lee, Mary Carrington, and Shahinaz M. Gadalla

SUPPLEMENTARY FIGURES

Figure S1. Assessment of population admixture in severe aplastic anemia (SAA) cases and controls. Only cases and controls of >80% European ancestry were included in the association analyses **A.** Assessment of admixture in discovery set. **B.** Assessment of admixture in the validation set. **C.** Principal components analysis (PCA) was conducted using GLU struct.pca module based on the same set of population informative SNPs. The PCA was performed only on those subjects determined by GLU struct.admix to have greater than 80% of European ancestry. Abbreviations: Mayo, Mayo Clinic Case-Control Study of Non-Hodgkin Lymphoma and Chronic Lymphocytic Leukemia[1]; NCI-SEER, National Cancer Institute, Surveillance, Epidemiology, and End Results Non-Hodgkin Lymphoma Case-Control Study[2,3]; WHI, Women's Health Initiative[4]; Yale, Population-based Case-Control Study in Connecticut Women[5]; PLCO, The Prostate and Ovarian Cancer Prevention Trial

Supplementary Figure S1.



C.

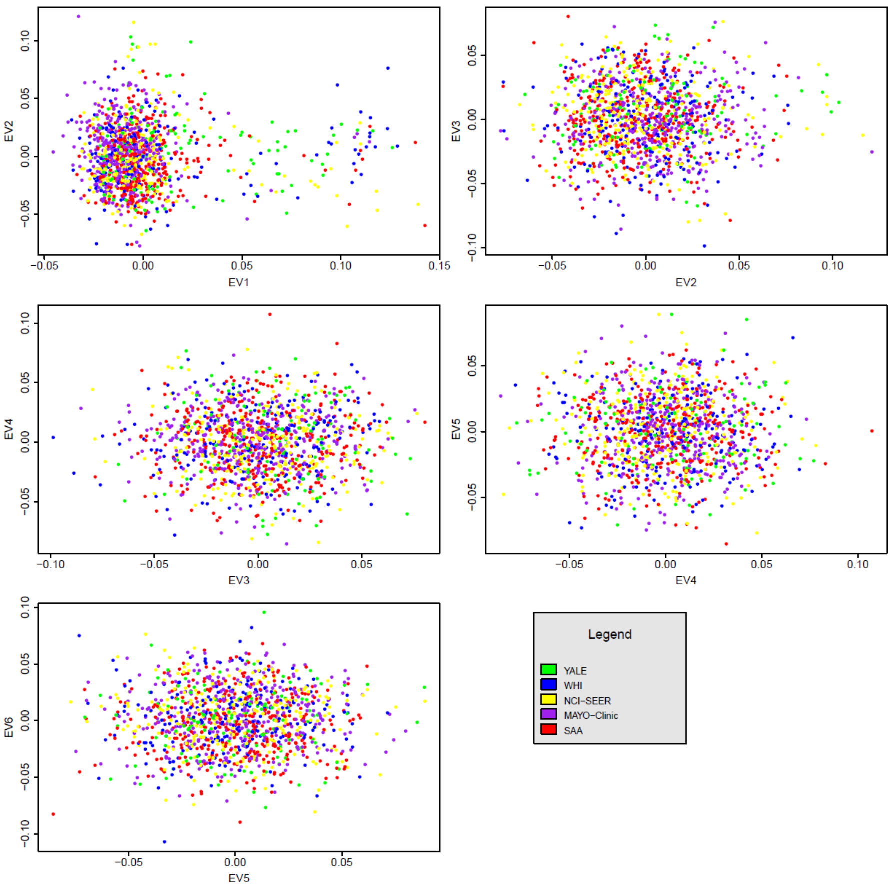

EV1
EV2
EV3
EV4
EV5
EV6
−0.05
0.00
0.05
0.10
0.15
−0.10
Legend
YALE
WHI
NCI−SEER
MAYO−Clinic
SAA

Figure S2. Chromosome 6 copy neutral loss of heterozygosity. **A.** location on chromosome 6p (dark blue bar indicates a second clone)**; B.** Two adjacent mosaic copy neutral loss of heterozygosity events (CNLOH) with different proportion of mosaicism for q arm of chromosome 6 covering HLA region. Each dot in the figure represents one SNP. Red dots represent B allele frequency (BAF, scale on the right side), while black dots show Log R ratio values (LRR, scale on the left side). Chr6-CNLOH event characterized by unchanged Log R ratio (mean of LRR within segment (blue line) = 0) and abnormal heterozygous BAF. The vertical gray lines indicate the breakpoint(s) of the event segment.

Figure S2.

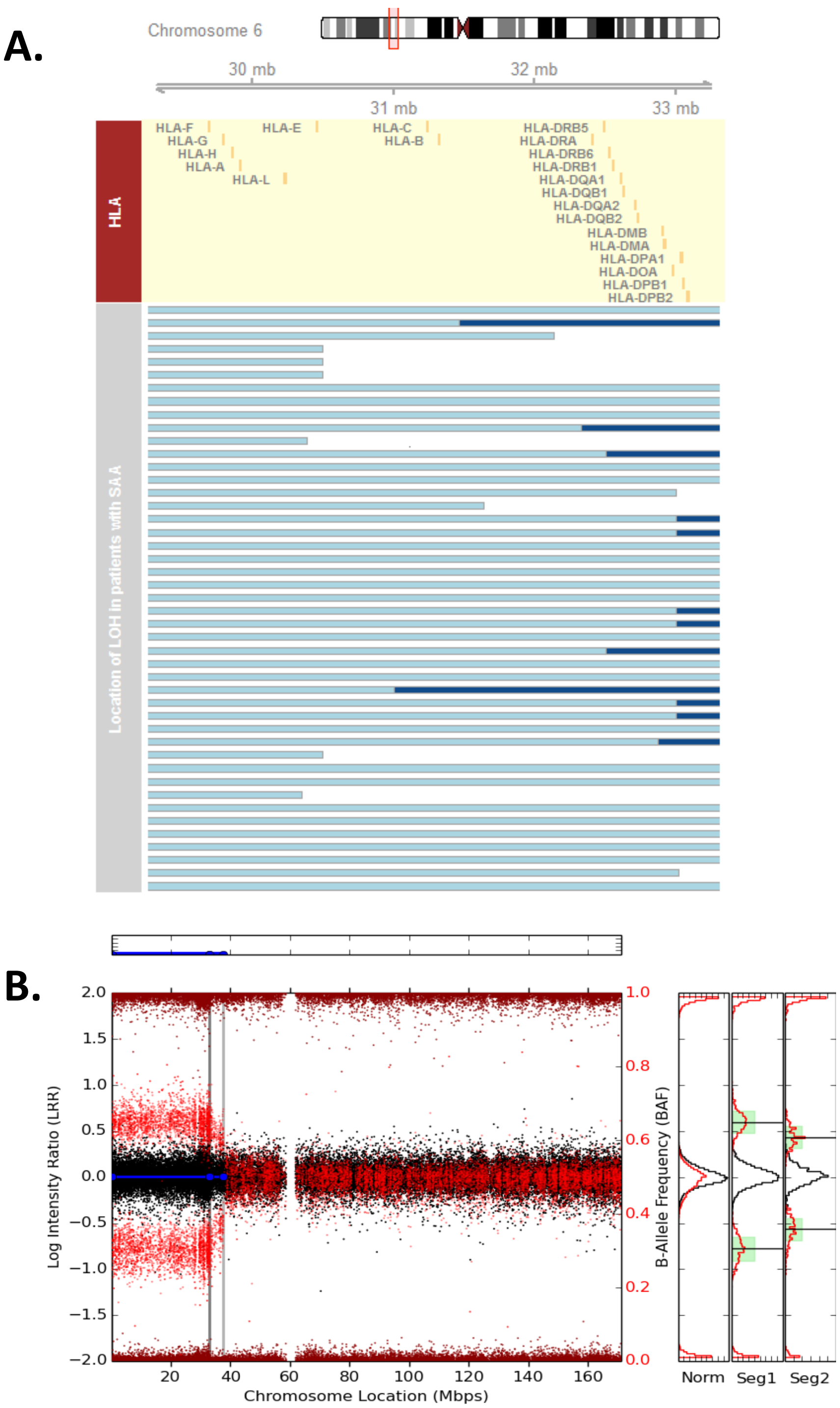
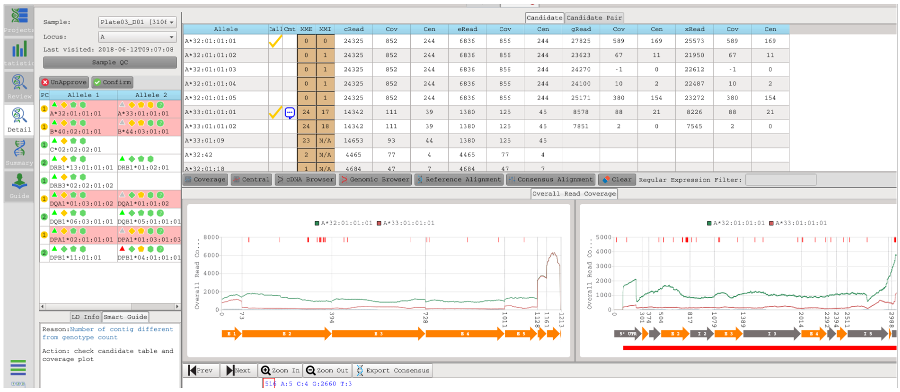

A.
Chromosome 6
30 mb
32 mb
31 mb
33 mb
HLA
HLA-F
HLA-G
HLA-H
HLA-A
HLA-L
HLA-E
HLA-C
HLA-B
HLA-DRB5
HLA-DRA
HLA-DRB6
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DQA2
HLA-DQB2
HLA-DMB
HLA-DMA
HLA-DPA1
HLA-DOA
HLA-DPB1
HLA-DPB2
Location of LOH in patients with SAA
B.
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Log Intensity Ratio (LRR)
20
40
60
80
100
120
140
160
Chromosome Location (Mbps)
1.0
0.8
0.6
0.4
0.2
0.0
B-Allele Frequency (BAF)
Norm
Seg1
Seg2

Figure S3. Example of next generation sequencing results of HLA sequencing showing somatic loss of heterozygosity across multiple loci in the same subject. All loci show more than 5-fold difference in the number of reads among the 2 alleles at a locus that is consistent with mosaicism with a population of microdeletion in the MHC region.

Figure S3.

A. Loss of heterozygosity



Sample: Plate03_D01 [3108
Locus: A
Last visited: 2018-06-12T09:07:08

PC	Allele 1	Allele 2
1	A*32:01:01:01	A*33:01:01:01
1	B*40:02:01:01	B*44:03:01:01
1	C*02:02:02:01	
2	DRB1*13:01:01:01	DRB1*01:02:01
1	DRB3*02:02:01:02	
1	DQA1*01:03:01:02	DQA1*01:01:02
2	DQB1*06:03:01:01	DQB1*05:01:01:01
1	DPA1*02:01:01:01	DPA1*01:03:01:03
2	DPB1*11:01:01	DPB1*04:01:01:01

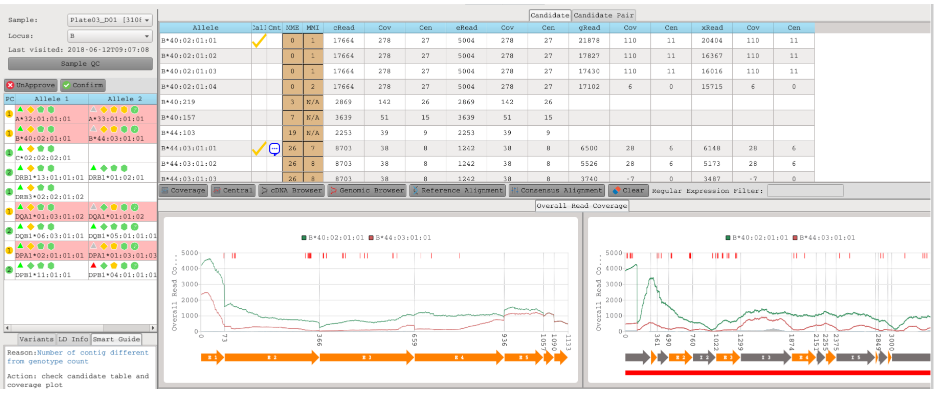
Reason:Number of contig different from genotype count

Action: check candidate table and coverage plot

Allele	Call	Cmt	MME	MMI	cRead	Cov	Cen	eRead	Cov	Cen	gRead	Cov	Cen	xRead	Cov	Cen
A*32:01:01:01	✓		0	0	24325	852	244	6836	856	244	27825	589	169	25573	589	169
A*32:01:01:02			0	1	24325	852	244	6836	856	244	23623	67	11	21950	67	11
A*32:01:01:03			0	1	24325	852	244	6836	856	244	24270	-1	0	22612	-1	0
A*32:01:01:04			0	1	24325	852	244	6836	856	244	24100	10	2	22487	10	2
A*32:01:01:05			0	1	24325	852	244	6836	856	244	25171	380	154	23272	380	154
A*33:01:01:01	✓	💬	24	17	14342	111	39	1380	125	45	8578	88	21	8226	88	21
A*33:01:01:02			24	18	14342	111	39	1380	125	45	7851	2	0	7545	2	0
A*33:01:09			23	N/A	14653	93	44	1380	125	45						
A*32:42			2	N/A	4465	77	4	4465	77	4						
A*32:01:18			1	N/A	4684	47	7	4684	47	7						

516 A:5 C:4 G:2660 T:3

B. Loss of heterozygosity



Sample: Plate03_D01 [3108
Locus: B
Last visited: 2018-06-12T09:07:08

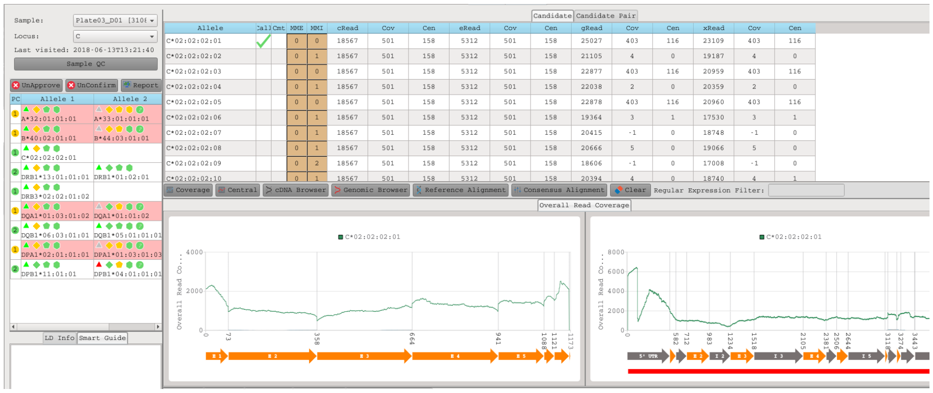
PC	Allele 1	Allele 2
1	A*32:01:01:01	A*33:01:01:01
1	B*40:02:01:01	B*44:03:01:01
1	C*02:02:02:01	
2	DRB1*13:01:01:01	DRB1*01:02:01
1	DRB3*02:02:01:02	
1	DQA1*01:03:01:02	DQA1*01:01:02
2	DQB1*06:03:01:01	DQB1*05:01:01:01
1	DPA1*02:01:01:01	DPA1*01:03:01:03
2	DPB1*11:01:01	DPB1*04:01:01:01

Reason:Number of contig different from genotype count

Action: check candidate table and coverage plot

Allele	Call	Cmt	MME	MMI	cRead	Cov	Cen	eRead	Cov	Cen	gRead	Cov	Cen	xRead	Cov	Cen
B*40:02:01:01	✓		0	1	17664	278	27	5004	278	27	21878	110	11	20404	110	11
B*40:02:01:02			0	1	17664	278	27	5004	278	27	17827	110	11	16367	110	11
B*40:02:01:03			0	1	17664	278	27	5004	278	27	17430	110	11	16016	110	11
B*40:02:01:04			0	2	17664	278	27	5004	278	27	17102	6	0	15715	6	0
B*40:219			3	N/A	2869	142	26	2869	142	26						
B*40:157			7	N/A	3639	51	15	3639	51	15						
B*44:103			19	N/A	2253	39	9	2253	39	9						
B*44:03:01:01	✓	💬	26	7	8703	38	8	1242	38	8	6500	28	6	6148	28	6
B*44:03:01:02			26	8	8703	38	8	1242	38	8	5526	28	6	5173	28	6
B*44:03:01:03			26	8	8703	38	8	1242	38	8	3740	-7	0	3487	-7	0

C. Homozygous

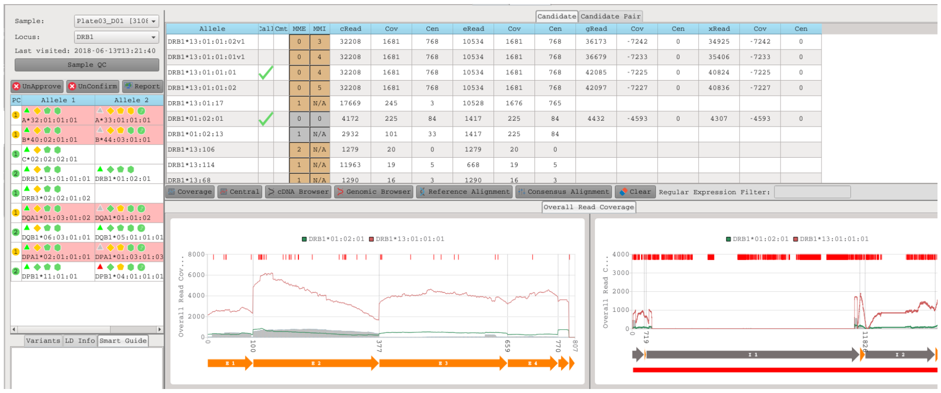


Sample: Plate03_D01 [3108
Locus: C
Last visited: 2018-06-13T13:21:40

PC	Allele 1	Allele 2
1	A*32:01:01:01	A*33:01:01:01
1	B*40:02:01:01	B*44:03:01:01
1	C*02:02:02:01	
2	DRB1*13:01:01:01	DRB1*01:02:01
1	DRB3*02:02:01:02	
1	DQA1*01:03:01:02	DQA1*01:01:02
2	DQB1*06:03:01:01	DQB1*05:01:01:01
1	DPA1*02:01:01:01	DPA1*01:03:01:03
2	DPB1*11:01:01	DPB1*04:01:01:01

Allele	Call	Cmt	MME	MMI	cRead	Cov	Cen	eRead	Cov	Cen	gRead	Cov	Cen	xRead	Cov	Cen
C*02:02:02:01	✓		0	0	18567	501	158	5312	501	158	25027	403	116	23109	403	116
C*02:02:02:02			0	1	18567	501	158	5312	501	158	21105	4	0	19187	4	0
C*02:02:02:03			0	0	18567	501	158	5312	501	158	22877	403	116	20959	403	116
C*02:02:02:04			0	1	18567	501	158	5312	501	158	22038	2	0	20359	2	0
C*02:02:02:05			0	0	18567	501	158	5312	501	158	22878	403	116	20960	403	116
C*02:02:02:06			0	1	18567	501	158	5312	501	158	19364	3	1	17530	3	1
C*02:02:02:07			0	1	18567	501	158	5312	501	158	20415	-1	0	18748	-1	0
C*02:02:02:08			0	1	18567	501	158	5312	501	158	20666	5	0	19066	5	0
C*02:02:02:09			0	2	18567	501	158	5312	501	158	18606	-1	0	17008	-1	0
C*02:02:02:10			0	1	18567	501	158	5312	501	158	20394	4	0	18740	4	1

D. Loss of heterozygosity



Sample: Plate03_D01 [3108
Locus: DRB1
Last visited: 2018-06-13T13:21:40

PC	Allele 1	Allele 2
1	A*32:01:01:01	A*33:01:01:01
1	B*40:02:01:01	B*44:03:01:01
1	C*02:02:02:01	
2	DRB1*13:01:01:01	DRB1*01:02:01
1	DRB3*02:02:01:02	
1	DQA1*01:03:01:02	DQA1*01:01:02
2	DQB1*06:03:01:01	DQB1*05:01:01:01
1	DPA1*02:01:01:01	DPA1*01:03:01:03
2	DPB1*11:01:01	DPB1*04:01:01:01

Allele	Call	Cmt	MME	MMI	cRead	Cov	Cen	eRead	Cov	Cen	gRead	Cov	Cen	xRead	Cov	Cen
DRB1*13:01:01:02v1			0	3	32208	1681	768	10534	1681	768	36173	-7242	0	34925	-7242	0
DRB1*13:01:01:01v1			0	4	32208	1681	768	10534	1681	768	36679	-7233	0	35406	-7233	0
DRB1*13:01:01:01	✓		0	4	32208	1681	768	10534	1681	768	42085	-7225	0	40824	-7225	0
DRB1*13:01:01:02			0	5	32208	1681	768	10534	1681	768	42097	-7227	0	40836	-7227	0
DRB1*13:01:17			1	N/A	17669	245	3	10528	1676	765						
DRB1*01:02:01	✓		0	0	4172	225	84	1417	225	84	4432	-4593	0	4307	-4593	0
DRB1*01:02:13			1	N/A	2932	101	33	1417	225	84						
DRB1*13:106			2	N/A	1279	20	0	1279	20	0						
DRB1*13:114			1	N/A	11963	19	5	668	19	5						
DRB1*13:68			1	N/A	1290	16	3	1290	16	3						

SUPPLEMENTARY METHODS

Genome-wide SNP Genotyping

Genotyping of SAA cases was conducted on the Illumina Infinium OmniExpress BeadChip array at the Cancer Genomics Research Laboratory (CGR) in the Division of Cancer Epidemiology and Genetics (DCEG) at the National Cancer Institute (NCI). Genotyping was done in two stages due to the timing of sample receipt. The first set, called the discovery set, consisted of 640 cases scanned on the Illumina Human OmniExpress-12v1-1_B and Illumina Human OmniExpress -24v1-0_a chip types. The second set, called the validation set, consisted of 255 cases scanned on the Illumina Infinium OmniExpress-24v1-2_A1 chip.

The controls were derived from previously scanned subjects drawn from two large cohort studies (The Prostate, Lung, Colon and Ovarian Cancer Prevention Trial (PLCO)[1] and the American Cancer Society Cancer Prevention Study II (CPSII)[2] scanned on the Illumina Omni 2.5M SNP microarray) as well as 4 other U.S.-based studies [Mayo Clinic Case-Control Study of Non-Hodgkin Lymphoma and Chronic Lymphocytic Leukemia (MAYO)[3], National Cancer Institute, Surveillance, Epidemiology, and End Results Non-Hodgkin Lymphoma (NHL) Case-Control Study (NCI-SEER)[4,5], Women's Health Initiative (WHI)[6], and the Population-based Case-Control Study in Connecticut Women (YALE)[7]] scanned on the Infinium OmniExpress chip. We selected 2,453 controls of European ancestry. The controls for the discovery set consisted of 1,396 subjects drawn from all 6 studies; the controls for the validation set consisted of 1059 subjects drawn from PLCO study.

Quality Control Assessment

We examined the distribution of the sample missing rate and sample mean heterozygosity. In the quality control analysis of discovery set, SNPs with less than a 90% completion rate were excluded from further analysis. Samples were excluded on the basis of (1) completion rates lower than 95% (*n* = 5 samples); (2) abnormal heterozygosity values of less than 25% or greater than 35% (*n* = 3) or; (3) abnormal X-chromosome heterozygosity (*n* = 1). After removing the low performing samples and loci described in above, the data were of high-quality at the sample level. Based on these data sets, we performed the assay concordance analysis and identified all 26 expected duplicates with the average SNP concordance rate at 99.998%. We also detected 15 inter-chip duplicates with concordance rate greater than 99.99%. There were no unexpected duplicate pairs detected. Genotypes for all subject pairs were computed for close relationships (first- and second-degree relatives) using GLU qc.ibds module (http://code.google.com/p/glu-genetics/) with an IBD0 threshold of 0.70. One pair of first degree relative was detected in the cases.

Using a set of 12,000 population informative SNPs that common to both the Illumina and Affymetrix commercial platforms and with low linkage disequilibrium (pair-wise $r^2 < 0.004$)[8] and data from HapMap build 27, we excluded 184 cases with less than 80% European ancestry[9], as determined using GLU strct.admix module. The HapMap I+II CEU, YRI, ASA (JPT+CHB) samples were used as the fixed reference populations[10]; a majority of these subjects represent 73 cases of mixed East Asian and European ancestry, 31 cases of Asian ancestries, 28 cases of African ancestry, and 26 cases of mixed European and African ancestries (Supplementary Figure S1). We also excluded 93 SAA cases reported by CIBMTR with a known inherited bone marrow failure syndrome from association analysis. The final association analysis for the discovery set included 359 cases and 1,396 controls of European ancestry. After quality control filtering, data from 703,857 SNPs were available for the case subjects. The numbers of SNPs overlapping those of pooled controls were 688,067, respectively, and these SNPs were used in the downstream association analyses.

Similar quality control metrics were applied in the validation set; SNPs with less than a 90% completion rate were excluded from further analysis. Samples were excluded on the basis of (1) completion rates lower than 95% (*n* = 2 samples); (2) abnormal heterozygosity values of less than 25% or greater than 38% (*n* = 0); (3) abnormal X-chromosome heterozygosity (*n* = 0). There were nine expected duplicated cases in with the concordance rate of 99.76%. No unexpected duplicate pair detected. No first-degree relative pairs were detected. We excluded 78 case subjects with less than 80% European ancestry.

The association analysis for validation set included 175 cases and 1,059 controls of European ancestry. After quality control filtering, data from 702,117 SNPs were available for the case subjects. The numbers of SNPs overlapping controls were 663,976, respectively, and these SNPs were used in the downstream association analyses. The final association analysis for combined discovery and validation sets included 534 cases and 2,453 controls of European ancestry. After quality control filtering, data from 693,802 SNPs were available for the cases. The numbers of SNPs overlapping controls were 656,416, respectively, and these SNPs were used in the downstream association analyses. dbSNP build GRCh37/hg19 was used for annotations.

All of the controls were drawn from previous scans and passed similar quality control filtering. We excluded two control subjects that were identified as unexpected duplicates between the discovery and validation sets. We also excluded 90 control subjects with less than 80% European ancestry.

TaqMan Genotyping

Fourteen risk SNPs were validated by allele-specific TaqMan® genotyping (ThermoFisher) of a subset of 340 samples (Supplementary Table S3). TaqMan® assay validation and SNP genotyping was performed at the Cancer Genomics Research Facility (http://cgf.nci.nih.gov/). 5 ng of sample DNA, according to Quant-iT PicoGreen dsDNA quantitation (ThermoFisher), was transferred into 384-well plates (ThermoFisher) and dried down. Additionally, 5 ng of assay-specific controls, based on the validation results, were applied to pre-determined wells of the assay plates to guide analysis and overall quality and 5 ng of universal internal controls (NA07057 and NTC) were added to random locations of 384-well plates to provide a unique fingerprint for each plate for overall quality assurance.

Genotyping was performed using 5 uL reaction volumes consisting of: 2.5 uL of 2X KAPA Probe Fast MasterMix (Kapa Biosystems, Woburn, MA), 0.25 uL of 20X TaqMan® assay-specific mix of primers and probes, and 2.25 uL of MBG Water. Plates were heat-sealed using diamond optical seals (ThermoFisher) and PCR was performed using 9700 Thermal Cycler (ThermoFisher) with the following conditions: 95˚C hold for 3 min, 40 cycles of 95˚C for 3 sec and X˚C for 30 sec (where X is the optimized annealing temperature determined in validation for each assay), and 10˚C hold.

Endpoint reads were evaluated using the 7900HT Sequence Detection System (ThermoFisher) and cluster analysis was performed using SDS v2.2.2 software (ThermoFisher). Cluster analysis was performed using the Allelic Discrimination Plot, which is an X-Y scatter plot of FAM and VIC dyes, containing four distinct clusters which represent three possible genotypes: Allele 1

Homozygous (Y-Axis), Allele 2 Homozygous (X-Axis), and Allele 1/Allele 2 Heterozygous (Diagonal Axis) while the fourth cluster was at the origin and contains no amplification (NTCs). Analyzed data was imported into a LIMS where concordance of assay-specific genotyping controls and internal controls are confirmed.

The concordance rates of the 14 genotyped SNPs ranged from 99.7 to 100% (Supplementary Table S3).

GWAS Statistical Analyses

Associations between SNPs and risk of SAA were calculated using the multivariable logistic regression model from GLU assoc.logit1 module (http://code.google.com/p/glu-genetics/) assuming an additive genetic effect on the number of rare alleles presented in each genotype. For the discovery set, when included in the null model (baseline model), principal component analysis (PCA) identified three significant ($P < 0.05$) eigenvectors associated with the case/control status. The main effect model was adjusted for sex and these three eigenvectors to account for the imbalance between cases and controls. For the validation set, none of the 10 eigenvectors was significant ($P<0.05$); the main effect model was not adjusted for any covariates. For the combined discovery and validation sets, the main effect model was adjusted by sex and two eigenvectors, identified on the basis of significance ($P < 0.05$) observed in the null model of the combined sets. The estimated inflation factor λ for the test statistic from discovery, validation, and combined sets were 1.034, 1.032, and 1.047 respectively using estlambda() function with method=median from GenABEL package in R.

Identification of chromosome 6 copy neutral loss of heterozygosity (ch6CN-LOH) using SNP array data

Log R ratio (LRR) and B allele frequency (BAF) were used to assess copy number alteration derived from SNP array intensity data as previous described.[11] The Log R ratio (LRR) value is the normalized measure of total signal intensity and provides data on relative copy number and the B allele frequency derived from the ratio of allelic probe intensity is the proportion of hybridized sample that carries the B allele as designated by the Infinium Assay. The LRR and BAF values from qualified assays were re-normalized[12,13] and data analyzed using custom software pipelines that involved BAF Segmentation package (http://baseplugins.thep.lu.se/wiki/se.lu.onk.BAFsegmentation) to detect copy number change and copy neutral loss of heterozygosity. A deviation from heterozygotes band in BAF with LRR value of zero indicated a copy-neutral loss of heterozygosity (CN-LOH). To minimize false positives, the analysis was restricted to chromosomal abnormalities larger than 2 Mb. All potential events were plotted and visualized, and false positive calls were excluded from the analysis based on manual review of each plot.

Identification of chr6-LOH using HLA next generation sequencing

HLA typing was performed using MIA FORA NGS kit (Immucor, Inc., Norcross, GA). All samples were genotyped for 11 HLA loci, namely HLA-A, -B, -C, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, -DPB1. The coverage for HLA-A, -B, and -C included all exons and introns, at least 200 base pairs of the 5′ UTR and 100–1100 base pairs of the 3′ UTR; coverage for -DPA1, and -DQA1, included all exons and introns, at least 45 base pairs of the 3′ UTR and 25-190 base pairs of the 3′ UTR. Coverage for -DRB1 included all exons, introns 2–6, at least 440 base pairs of the 5′

UTR, 12 base pairs of the 3′ UTR, 275 base pairs of intron 1 adjacent to exon 1, and 210 base pairs of intron 1 adjacent to exon 2; coverage for -DRB3/4/5 included exons 2–6, introns 2–5, and 260 base pairs of intron 1 adjacent to exon 2; coverage for -DQB1 included exons 1–5 and introns 1–4; coverage for -DPB1 included exons 2–4 and introns 2–3. NGS sequencing was performed according to the manufacturer's instructions and described elsewhere [14]. Analysis was performed using MIA FORA software (Immucor, Inc.). Specimens with more than 5-fold difference in the number of reads corresponding to the 2 alleles at a locus that were consistent across all 11 tested loci and determined to be mosaic with a population of microdeletions in the MHC region. Example results from one sample are shown in Supplementary Figure S3A-D.

Imputation of HLA alleles

We performed imputation of HLA class II loci -DPB1* using the published imputation tool HIBAG.[15]

The provided model trained for European ancestry was used for the imputation. The entire set of European ancestry cases and controls were imputed together. High resolution clinical typing available for 401 patients showed a concordance with the imputation of 92.7%.

Cell Surface Expression Analyses of HLA-DP

Cell surface expression levels of HLA-DP were measured using flow cytometry. A total of 175 healthy European American donors were recruited from the Frederick National Laboratory for Cancer Research Blood Donor Program. Samples were stained as follows; fresh blood was processed to generate peripheral blood mononuclear cells (PBMCs). Immediately after processing, one million PBMCs were stained with the following cocktail of antibodies, anti-Human CD19 PE-Cyanine5 (ThermoFisher, Inc, Waltham, MA), anti-human CD3 Brilliant Violet 605™ (BioLegend, Inc., San Diego, CA) and anti-Human HLA-DP Monomorphic R-phycoerythrin (R-PE) clone B7/21 (Leinco Technologies, Inc., Fenton, MO). Cells were stained in the dark at 4 degrees °C for 20 minutes and washed with FACS wash followed by fixing with 300ul of BD Cytofix™ Fixation Buffer (BD Biosciences, Inc., San Jose, CA). Samples were run on the BD LSRFortessa™ and analyzed using FlowJo software (FlowJo, LLC, Ashland, OR).

Figure S3.

SUPPLEMENTARY REFERENCES

1. Thomas G, Jacobs KB, Yeager M, et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet.* 2008;40(3):310-315.
2. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet.* 2009;41(9):986-990.
3. Cerhan JR, Fredericksen ZS, Wang AH, et al. Design and validity of a clinic-based case-control study on the molecular epidemiology of lymphoma. *Int J Mol Epidemiol Genet.* 2011;2(2):95-113.
4. Chatterjee N, Hartge P, Cerhan JR, et al. Risk of non-Hodgkin's lymphoma and family history of lymphatic, hematologic, and other cancers. *Cancer Epidemiol Biomarkers Prev.* 2004;13(9):1415-1421.
5. Wang SS, Cerhan JR, Hartge P, et al. Common genetic variants in proinflammatory and other immunoregulatory genes and risk for non-Hodgkin lymphoma. *Cancer Res.* 2006;66(19):9771-9780.
6. Anderson GL, Manson J, Wallace R, et al. Implementation of the Women's Health Initiative study design. *Ann Epidemiol.* 2003;13(9 Suppl):S5-17.
7. Zhang Y, Hughes KJ, Zahm SH, et al. Genetic variations in xenobiotic metabolic pathway genes, personal hair dye use, and risk of non-Hodgkin lymphoma. *Am J Epidemiol.* 2009;170(10):1222-1230.
8. Yu K, Wang Z, Li Q, et al. Population substructure and control selection in genome-wide association studies. *PLoS One.* 2008;3(7):e2551.

9. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet.* 2015;96(1):37-53.

10. International HapMap C, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007;449(7164):851-861.

11. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* 2006;16(9):1136-1148.

12. Staaf J, Vallon-Christersson J, Lindgren D, et al. Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC bioinformatics.* 2008;9:409.

13. Diskin SJ, Li M, Hou C, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* 2008;36(19):e126.

14. Ehrenberg PK, Geretz A, Sindhu RK, et al. High-throughput next-generation sequencing to genotype six classical HLA loci from 96 donors in a single MiSeq run. *Hla.* 2017;90(5):284-291.

15. Zheng X, Shen J, Cox C, et al. HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* 2014;14(2):192-200.