Genome-wide association study identifies multiple HLA loci for sarcoidosis susceptibility

Liao SY[1,2,3]; Jacobson S[1]; Hamzeh, NY[4]; Culver, DA[5]; Barkes, BQ[1]; Mroz, P[1]; Macphail K[1], Pacheco, K[1,2,3]; Patel, DC[6]; Wasfi, YS[7]; Koth LL[8]; Langefeld, CD[9,10]; Leach S[1], White E[1], Montgomery, C[11], Maier LA[1,2,3]; Fingerlin TE[1,2,3,12]; and GRADs investigators[13]

[1]National Jewish Health, Department of Medicine, Denver, CO

[2]University of Colorado Anschutz Medical Campus, Department of Medicine, Aurora, CO

[3]Colorado School of Public Health, University of Colorado Denver – Anschutz Medical Campus, Aurora, CO

[4]University of Iowa, Department of Medicine, Iowa City, IA

[5]Cleveland Clinic, Cleveland, OH

[6]University of Florida, Division of Pulmonary, Critical Care and Sleep Medicine, Gainesville, FL

[7]Johnson & Johnson, Spring House, PA

[8]University of California-San Fransisco, Department of Medicine, San Fransisco, CA

[9]Wake Forest University School of Medicine, Department of Biostatistics and Data Science

[10]Wake Forest University School of Medicine, Center for Precision Medicine

[11]Oklahoma Medical Research Foundation, Oklahoma City, OK

[12]National Jewish Health, Department of Immunology and Genomic Medicine, Denver, CO

[13]Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis (GRADS)

Corresponding author information:

Tasha E. Fingerlin, PhD

1400 Jackson Street, Denver, CO 80206. Email: fingerlint@njhealth.org

Authors' contribution:

TEF had full access to all data in the study and took full responsibility for the integrity of the data and the accuracy of the data analysis. SL, TEF, and LAM wrote the manuscript. SL, TEF, and LAM developed the analysis plan, and TEF supervised SL and SJ with the data analysis. TEF, and LAM designed the study. CM provided the data collection and analysis of African American cohort. NYH, DAC, BB, P.M., KM, KP, DP, YSW, LK, CDL, SL (Leach), and EW were involved in the sample enrollment, processing, and data collection. TEF and LAM supervised the research.

Take home message:

Using a genome-wide integrative analysis of genetics and transcriptomics, we found that SNPs located in *HLA-DRA, -DRB9, -DRB5, -DQA1*, and *BRD2* genes, along with classic HLA alleles DRB1*0101, DQA1*0101, and DQB1*0501, are associated with sarcoidosis.

Key Words:

Sarcoidosis; GWAS; HLA; eQTL

Abstract word count: 249

Manuscript word count: 3,000

Abstract

Sarcoidosis is a complex systemic disease. Our study aimed to 1) identify novel alleles associated with sarcoidosis susceptibility; 2) provide an in-depth evaluation of HLA alleles and sarcoidosis susceptibility; 3) integrate genetic and transcription data to identify risk loci that may more directly impact disease pathogenesis.

We report a genome-wide association study of 1,335 sarcoidosis cases and 1,264 controls of European descent (EA) and investigate associated alleles in a study of African Americans (AA: 1,487 cases and 1,504 controls). The EA cohort was recruited from National Jewish Health, Cleveland Clinic, University of California San Francisco, and Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis. The AA cohort was from a previous study with subjects enrolled from multiple United States sites. HLA alleles were imputed and tested for association with sarcoidosis susceptibility. Expression quantitative locus and colocalization analysis were performed using a subset of subjects with transcriptome data.

49 SNPs in *HLA-DRA, -DRB9, -DRB5, -DQA1*, and *BRD2* genes were significantly associated with sarcoidosis susceptibility in EA. Among them, rs3129888 was also a risk variant for sarcoidosis in AA. Classical HLA alleles DRB1*0101, DQA1*0101, and DQB1*0501, which are highly correlated, were also associated with sarcoidosis. rs3135287 near *HLA-DRA* was associated with *HLA-DRA* expression in peripheral blood mononuclear cells and bronchoalveolar lavage.

In summary, we identified several novel SNPs and three HLA alleles associated with sarcoidosis susceptibility in the largest EA population evaluated to date using an integrative analysis of genetics and transcriptomics. We also replicated our findings in an AA population.

Introduction

Sarcoidosis is a complex systemic disease affecting between 45-300/100,000 persons in the United States[1]. While environmental and genetic susceptibility factors have been associated with sarcoidosis[2-4], the driving risk factors for disease are still largely undefined. Previous genome-wide association studies (GWAS) have consistently implicated the HLA region as harboring genetic risk loci based on single nucleotide polymorphisms (SNPs)[5-12], while candidate gene studies have implicated specific HLA alleles such as DRB1*1101, DRB1*1501, and DQB1*0602[13, 14]. The strong role of the HLA region for sarcoidosis disease risk is consistent with exposure and immune-based associations observed in patients, as well as the heterogeneous disease course.

Most sarcoidosis genetic studies have focused on genotype data, limiting understanding of potential functional genetic features associated with disease status. Studying the impact of risk alleles on gene expression can provide the first step in understanding their potential biological impact[15]. This is especially true for complex diseases since the majority of genetic variants robustly associated with these diseases fall in non-coding regions of the genome[16]. While the function of non-coding regions was unknown in the past, there is substantial evidence that many of them influence disease risk through regulatory effects, including those that impact gene expression[15]. Hence, integrating genetic and transcriptomic data may identify loci with a more direct or functional effect on disease pathogenesis.

To identify genetic risk factors for disease and study their effects on gene expression, we conducted a genome-wide association study in European American (EA) population that includes follow-up of risk alleles on gene expression in peripheral blood as well as lung cells. We confirm the strong role of the HLA region in sarcoidosis risk and demonstrate both novel

SNPs and classical HLA alleles associated with disease and influencing HLA expression. We replicated some of these findings in a large African American (AA) population.

Materials and Methods

Study design and population:

This GWAS used a two-phase approach to identify genome-wide significant SNPs associated with sarcoidosis; additional details are present in Supplemental Methods. DNA samples from sarcoidosis cases and controls were obtained from National Jewish Health (NJH), University of California, San Francisco, Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis (GRADS) consortium [17], and Cleveland Clinic Foundation. Sarcoidosis case definition was based on the ATS/ERS/WASOG statement[18]. After quality control (see below), 818 cases and 981 controls (Phase 1) and 517 cases and 283 controls (Phase 2), all self-reported EA ancestry (**Table** 1 and **Figure 1**), reflected the majority of the race/ethnicities seen in our clinics. A subset of study participants had peripheral blood mononuclear cell (PBMC) and/or bronchoalveolar lavage (BAL) cell RNA sequencing data available through GRADS study[17, 19]. We tested significant SNPs identified in our EA two-phase approach in an AA study population. The AA GWAS summary statistics were obtained from a published study[5] with updated imputations since publication of those data.

Genome-wide genotyping:

DNA was extracted from whole blood using the PAXgene Blood DNA kit. Genotyping was performed using the Illumina HumanOmini 2.5 BeadChip to interrogate ~ 2.4 million markers. The markers were derived from the 1000 Genomes Project[21], including all three HapMap phases, 19K SNPs across the MHC, and over 41K non-synonymous SNPs. Genotyping was conducted at Hudson Alpha Biotechnology Institute (Huntsville, AL https://hudsonalpha.org/).

Genotype quality control:

We used 1000 genomes data[21] to infer ancestry-informative principal components (PCs), which were projected onto cases and controls. We prioritized SNPs with minor allele frequency (MAF)>0.03 and Hardy-Weinberg Equilibrium p>0.001 in cases and controls evaluated separately and <10% missing data.

Imputation of additional genotypes and HLA variants for the EA population:

We imputed genotypes using combined case and control discovery samples for all 1000 genomes SNPs. We imputed classical HLA alleles using R package HLA Genotype Imputation with Attribute Bagging (HIBAG)[27].

RNA sequencing and quality control:

Total RNA was extracted, and RNA-sequencing was conducted as outlined[19]. We followed a similar quality control procedure for both PBMC and BAL samples and removed RNA samples with an unmapped read rate >20% and mitochondrial read rate >0%. We removed outlier samples through PC analysis; the EA individuals who also had gene expression data available were included in the expression quantitative trait analyses.

Statistical analysis:

Single-SNP association test and meta-analysis

We tested for association between each SNP and sarcoidosis using Snptest (v2)[28] as described previously[29]. To obtain an overall measure of association with sarcoidosis, we performed a meta-analysis of Phase 1 and Phase 2 using summary statistic data and the weighted inverse normal method[30] as implemented in the software METAL[31]. Genome-wide significance was defined as meta-analysis $p<5\times10^{-8}$. Genome-wide significant SNPs identified in our EA population were tested in the AA population. Statistical significance for these SNPs

was defined as p<0.05/(number of significant SNPs in EA). We compared our GWAS results to previously identified loci in other studies including SNPs in *ANXA11* (rs1049550, rs1953600, rs2573346, rs2784773)[32], *RAB23* (rs1040461)[9], *C100RF67* (rs1398024)[33], *OS9* (rs1050045)[8], *CCDC88B* (rs479777)[7], and *NOTCH4* (rs715299)[5]. Statistical significance for these *a priori* SNPs was defined as p<0.0056(0.05/9 SNPs).

Classic HLA alleles analysis

We used logistic regression models to test for association between dosage of each imputed HLA allele and sarcoidosis. Given strong *a priori* associations with the HLA region, we used p<0.00011(0.05/448 HLA alleles tested) as statistical significance. We compared our HLA results to previously identified HLA alleles in other studies, including DRB1*1101, DRB1*1501, and DQB1*0602[13, 14]. We used a p=0.016(0.05/3 alleles) to determine statistical significance for *a priori* alleles.

Conditional Models

To assess the independence of single-SNP effects from HLA risk alleles, we computed multivariable logistic regression models where HLA risk alleles were included as covariates in the model, and each SNP, one at a time, was tested for association (i.e., association adjusted for HLA risk alleles).

Expression quantitative locus (eQTL) and colocalization analysis

For those sarcoidosis cases with gene expression data from GRADS, we performed colocalization analysis using eCAVIAR[34] to identify variants with evidence for colocalization of disease and *cis* eQTL associations. The algorithm estimates the posterior probability that the same variant is casual in both GWAS and eQTL studies while accounting for linkage

disequilibrium (LD). We included all SNPs within the defined gene boundary of the significant SNP in the analysis (+/-25K base pairs) and tested for association between those SNPs and gene expression. The threshold for significance was set as a colocalization posterior probability (CLPP)>0.001, as suggested by eCAVIAR[34]. The same approach was applied to imputed HLA alleles. We tested for association between HLA alleles and gene expression in two tissues (BAL and PBMC). In addition, using GRADS data, we conducted a comprehensive cis-eQTL search using the publicly available database, Genotype-Tissue Expression (GTEx)[36, 37]. The GTEx project was supported by the Common Fund of the Office of the Director of National Institutes of Health, NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for analyses described in this manuscript were obtained from the GTEx Portal on 05/31/2022.

Results

Descriptive analysis of GWAS in EA:

We enrolled 3,141 subjects genotyped on the Illumina array. After quality control (**Figure 1**), we included n=2,599 subjects (1,335 cases and 1,264 controls) in the analysis (**Table 1**); 818 cases (49% female/51% male) and 981 controls (37% female/63% male) in Phase 1 and 517 cases (53% female/47% male) and 283 controls (71% male/29 % female) in Phase 2. For eQTL analyses, paired genotype-transcription data was available for 136 genotype-BAL transcription and 193 genotype-PBMC transcription samples.

GWAS identifies HLA region associations with sarcoidosis:

The meta-analysis of Phase 1 and 2 data identified 49 SNPs reaching GWAS significance. The r^2 LD plot of those SNPs and the top seven SNPs representing all significant SNPs (r^2>0.70) are shown in **Figure S1**. Those top seven SNPs were rs9269233, rs9271346, rs35656642, rs28589559, rs9276935, rs3129888, and rs71549283 (**Table 2**), located across *HLA-DRA, -DRB9, -DRB5, -DQA1*, and *BRD2 on* chromosome 6. The remaining significant SNPs are shown in **Table S1**. With the much-reduced Phase 2 sample size compared to Phase 1, the Phase 2 p-value is not nominally significant, although effect sizes (odds ratios) were comparable, and both contribute proportionally to meta-analysis p-values. The Manhattan plot for SNP associations **is shown in Figure 2A**, and **Figure 2B** shows the locus-specific plot for all significant SNPs highlighting the seven top SNPs. For the seven top SNPs, we used a stepwise approach to adjust for other SNPs (**Table S2**), and found rs9269233, rs9276935, rs28589559, and rs3129888 still nominally significant after adjustment (all p<0.01). Other top SNPs with $p<5 \times 10^{-5}$ are shown in **Table S3**. In the AA cohort, rs3129888 was also significantly associated with increased risk of sarcoidosis (**Table 2**). When we compared our results to

previously-identified loci from other GWAS studies, we found SNPs in *ANXA11* were nominally significant (**Table S4**) in the meta-analysis (p<0.0056).

Classic HLA alleles are associated with sarcoidosis:

Nine HLA alleles were significantly associated with sarcoidosis susceptibility (p<0.00011, **Table 3 and Table S5**). Three HLA alleles had a p-value<5 x 10^{-8}: DRB1*0101, DQA1*0101, and DQB1*0501. All three HLA alleles were highly correlated and protective against sarcoidosis. We found three candidate HLA alleles, DRB1*1101, DRB1*1501, and DQB1*0602, significantly associated with increased risk of sarcoidosis (**Table 3**).

GWAS SNPs in HLA are independent of HLA allele associations:

Although the p-values were attenuated slightly, each genome-wide significant SNP remained associated with sarcoidosis after adjustment for each of the HLA risk alleles (**Table 4)**. We then adjusted for all three HLA risk alleles, and the effects of each SNP on odds ratios were largely unchanged after adjustment (**Table S6**). Interestingly, there was a significant interaction between rs9271346 and DRB1*0101 (p=0.02, **Table S7**).

eQTL and colocalization analyses demonstrate an association between SNPs and gene expression:

Colocalization analysis was conducted across five genes: *HLA-DRA, -DRB9, -DRB5, -DQA1*, and *BRD2.* We found no significant colocalization when we assumed only one casual SNP in the region. The model assuming two casual SNPs in the region demonstrated that rs3135387 colocalized with both PBMC and BAL cell expression levels of *HLA-DRA* (CLPP=0.003 and 0.002, respectively). Other nearby SNPs, rs3129888, and rs3135390, within the *HLA-DRA* region also showed significant colocalization posterior probability with PBMC gene expression

(CLPP=0.002, **Figure S2).** A comprehensive cis-eQTL search of whole blood and lung from GTEx is shown in **Table 5**. In brief, rs9269233 demonstrated significant eQTL for *HLA-DRB9* expression in whole blood, while rs9271346, rs35656642, and rs28589559 were significant eQTLs for expression of H*LA-DQA1* in whole blood and lung. The colocalization analysis for the three genome-wide significant HLA alleles showed that DRB1*0101 was the most significantly associated with sarcoidosis susceptibility and with expression levels in PBMC and BAL among sarcoidosis patients for *DRB1, DQA1, DQB1*, and *DRB9* (all CLPPs>0.001). The DRB1*0101 risk allele was positively associated with PBMC *DQB1* gene expression (p=0.03, **Figure 3A**) and negatively associated with BAL *DRB9* gene expression (p=0.03, **Figure 3B**).

Discussion

We identified 49 SNPs associated with sarcoidosis in this largest EA GWAS of sarcoidosis; one of these SNPs, rs3129888, was also associated with sarcoidosis in an AA GWAS. All SNPs are on Chromosome 6 in the well-known HLA risk region. While we evaluated previously identified GWAS SNPs *a priori* and found an association with *ANXA11*, we found no other SNPs outside the HLA region associated with sarcoidosis. Using colocalization and eQTL analysis, we found rs3135387 colocalizes with PBMC and BAL gene expression of *HLA-DRA* in our EA population. HLA DRB1*0101, DQA1*0101, and DQB1*05*01 were significantly associated with sarcoidosis, while DRB1*0101 was also associated with expression of PBMC *HLA-DQB1* and BAL *HLA-DRB9*. These results suggest that in sarcoidosis, the HLA region is likely functional, impacting gene expression and disease development.

The most significant SNP associated with sarcoidosis in our study, rs9269233, is between *HLA-DRB9* and *HLA-DRB5*. This association with rs9269233 was only modestly attenuated after adjustment for sarcoidosis-related HLA alleles (adjusted $p=2.15x10^{-7}$), suggesting that rs9269233 is an independent risk allele for sarcoidosis in the region. The A allele of this SNP showed increased risk of sarcoidosis, has not been reported in other studies and was not significant in AA cohort. Potentially this may indicate that sarcoidosis pathogenesis differs between EA and AA, and in fact, other studies have found different risk variants in EA and AA subjects[38]. In a previous GWAS[38], rs1964995, also located between *HLA-DRB9* and *HLA-DRB5* (r^2 with rs9269233=0.53), showed a protective effect in non-Lofgren sarcoidosis vs. controls in a white Swedish cohort. Interestingly, rs1964995 increases the risk of rheumatoid arthritis (RA) in AA, although not in EA[39]. rs9269233 has been found to be an eQTL for *HLA-DRB9* gene expression in whole blood in our analysis using the GTEx database. In addition to rs9269233, the presence of at least one DRB1*0101 allele was significantly associated with *HLA-DRB9* gene expression in BAL cells. *HLA-DRB9* is a pseudogene that is transcribed into

RNA but does not encode proteins. However, pseudogenes can regulate other protein-coding genes[40]; our findings suggest this may be the case with DRB9 in our population. In previous studies, *HLA-DRB* loci were found to group into five major haplogroups (DR1, DR8, DR51, DR52, and DR53), which differ by alleles at functional DRB genes (DRB3, DRB4, or DRB5) and DRB pseudogenes (DRB2, DRB6, DRB7, DRB8, or DRB9)[41]. The HLA-DRs are associated with various diseases, including sarcoidosis[42], which is usually thought to have implications for antigen presentation. Our data suggests that these associations may reflect gene expression regulation that is not currently known to be functional but may impact immune responses in sarcoidosis.

Two other significant SNPs associated with sarcoidosis, rs9271346, and rs35656642, are both near *HLA-DQA1* (<25K base pairs). rs9271346 was nominally significant in AA and EA in a previous study[5], with the same allele associated with risk in each, although it did not reach genome-wide significance (p=0.007 in AA and 8.63×10^{-6} in EA). rs35656642 is a novel risk variant for sarcoidosis that has not been described previously, and we found the same allele associated with sarcoidosis in AA. Both SNPs are also eQTLs in lung tissue and whole blood based on publicly available datasets, although we did not find associations in lung BAL cells in our study. A study found rs2187668 near gene *HLA-DQA1* significantly associated with Lofgren's syndrome[38], while another reported three SNPs (rs28609302, rs9273113, rs9272594) in this region associated with ocular sarcoidosis vs. controls in US AA(2 x 10-7 to 6 x 10^{-6})[43]. Of note, a subpopulation from the AA cohort was used in our study. We found low LD between those previously reported SNPs and the SNPs in our EA population(absolute r^2 0.06-0.58), indicating that the findings in our EA population are unlikely simple replications of previously-identified SNPs.

Except for rs3129888 ($p=1.52 \times 10^{-6}$), significant SNPs in EA were not significantly associated with sarcoidosis in AA. The rs3129888 G allele, located in an *HLA-DRA* intron, increased

sarcoidosis risk. This finding is consistent with a study on Lofgren's syndrome in Sweden, Germany, and a US AA population[38]. Our US EA samples are distinct from the European cohort, but our AA population largely overlaps with the US AA population[38].

We found three HLA alleles significantly associated with sarcoidosis, all of which are highly correlated with each other. DRB1*0101 and DQB1*0501 showed a protective effect on sarcoidosis risk, consistent with a previous study from the UK, Netherlands, and Japan[44]. In this previous study, DRB1*0101 was protective against lung-predominant sarcoidosis, Lofgren's syndrome, and uveitis. The sarcoidosis patients in our study included all subtypes of disease. When we searched the database from another large study using 13,835 EA individuals from five US sites of the Electronic Medical Records and Genomics (eMERGE) network[45] (using the International Classification of Diseases code as the definition for diseases), DRB1*0101, DQB1*0501, and DQA1*0101 were all protective for sarcoidosis but increased risk of RA (**Table S8**). These opposite genetic effects are consistent with a study of pleiotropy between sarcoidosis and RA, which demonstrates higher RA polygenic risk score associated with a protective effect of sarcoidosis[46], and another epidemiologic study demonstrating a lower prevalence of RA in a British sarcoidosis cohort vs. the general population[47]. While RA and sarcoidosis are both inflammatory diseases, this may imply that their disease pathogenesis is distinct and drivers of one disease protect from development of the other. Each of the previously-reported sarcoidosis HLA risk alleles, DRB1*1101, DRB1*1501, and DQB1*0602, showed a nominal increase in the risk of sarcoidosis in our study. DRB1*1101 allele was previously associated with increased risk of sarcoidosis in AA and European descent individuals[14]. DRB1*1501, which is in high LD with DQB1*0602, has been associated with increased risk of severe pulmonary sarcoidosis in individuals of European descent[13]. Of note, these HLA alleles were not the strongest risk alleles in our study, and this might be due to sub-populations (e.g., race/ethnicity) or varied phenotypes in our cohort compared to others. For

example, our cohort is likely a mixture of different sarcoidosis phenotypes (e.g., cardiac, ocular, cutaneous, etc.) as we did not restrict enrollment to specific phenotypes. Future GWAS studies would benefit from focusing on specific sarcoidosis phenotypes or those with specific organ involvement to explore genetic drivers of sarcoidosis manifestations, including neurological, cardiac, or specific pulmonary phenotypes of sarcoidosis.

There are limitations to our study. First, we included a heterogeneous population of sarcoidosis patients we various phenotypes; this could reduce our study's power to identify novel SNPs versus European studies focused on Lofgren's syndrome. Regardless, we found HLA associations linked to other sarcoidosis phenotypes in previous studies, like DRB1*1101 and *1501. Second, we have gene expression data available on only a subset of participants for the eQTL analyses, impacting power to identify other eQTLs in our study. To help mitigate this limitation, we used the GTEx database to enhance our evaluation of associations between SNPs and gene expression.

In summary, our findings provide convincing evidence that HLA alleles are an important contributor to risk of sarcoidosis not only in our EA population but also in an AA population with genotyping already available. In addition to traditional GWAS SNPs and imputed HLA variants, we also explored how these genotypes are associated with gene expression in both PBMC and BAL cells using integrated analysis and demonstrated showed potential gene expressions impacted by these risk variants. Our future studies will explore these potential variants/genes and their mechanistic implications.

Acknowledgments

We would like to thank all the participants of this study, as well as for the administrative support that we received from NJH and research coordination assistance from Christina Riley. We also

appreciate the grant support from National Institutes of Health Grants U01 HL112707, U01 HL112694, U01 HL112695, U01 HL112696, U01 HL112702, U01 HL112708, U01 HL112711, U01 HL112712, UL1 RR029882, UL1 RR025780, R01 HL110883, R01 HL114587, R01HL114587; Clinical and Translational Science Institute Grant U54 9 UL1 TR000005; and Centers for Disease Control and Prevention National Mesothelioma Virtual Bank for Translational Research Grant 5 U24 OH009077; and Foundation for Sarcoidosis Research (FSR).

Declaration of interests: All authors report no conflicts of interest related to this work

References

1. Erdal BS, Clymer BD, Yildiz VO, Julian MW, Crouser ED. Unexpectedly high prevalence of sarcoidosis in a representative U.S. Metropolitan population. *Respir Med* 2012: 106(6): 893-899.
2. Liu H, Patel D, Welch AM, Wilson C, Mroz MM, Li L, Rose CS, Van Dyke M, Swigris JJ, Hamzeh N, Maier LA. Association Between Occupational Exposures and Sarcoidosis: An Analysis From Death Certificates in the United States, 1988-1999. *Chest* 2016: 150(2): 289-298.
3. Fingerlin TE, Hamzeh N, Maier LA. Genetics of Sarcoidosis. *Clin Chest Med* 2015: 36(4): 569-584.
4. Rybicki BA, Iannuzzi MC, Frederick MM, Thompson BW, Rossman MD, Bresnitz EA, Terrin ML, Moller DR, Barnard J, Baughman RP, DePalo L, Hunninghake G, Johns C, Judson MA, Knatterud GL, McLennan G, Newman LS, Rabin DL, Rose C, Teirstein AS, Weinberger SE, Yeager H, Cherniack R, Group AR. Familial aggregation of sarcoidosis. A case-control etiologic study of sarcoidosis (ACCESS). *Am J Respir Crit Care Med* 2001: 164(11): 2085-2091.
5. Adrianto I, Lin CP, Hale JJ, Levin AM, Datta I, Parker R, Adler A, Kelly JA, Kaufman KM, Lessard CJ, Moser KL, Kimberly RP, Harley JB, Iannuzzi MC, Rybicki BA, Montgomery CG. Genome-wide association study of African and European Americans implicates multiple shared and ethnic specific loci in sarcoidosis susceptibility. *PLoS One* 2012: 7(8): e43907.
6. Cozier YC, Ruiz-Narvaez EA, McKinnon CJ, Berman JS, Rosenberg L, Palmer JR. Fine-mapping in African-American women confirms the importance of the 10p12 locus to sarcoidosis. *Genes Immun* 2012: 13(7): 573-578.
7. Fischer A, Schmid B, Ellinghaus D, Nothnagel M, Gaede KI, Schurmann M, Lipinski S, Rosenstiel P, Zissel G, Hohne K, Petrek M, Kolek V, Pabst S, Grohe C, Grunewald J, Ronninger M, Eklund A, Padyukov L, Gieger C, Wichmann HE, Nebel A, Franke A, Muller-Quernheim J, Hofmann S, Schreiber S. A novel sarcoidosis risk locus for Europeans on chromosome 11q13.1. *Am J Respir Crit Care Med* 2012: 186(9): 877-885.
8. Hofmann S, Fischer A, Nothnagel M, Jacobs G, Schmid B, Wittig M, Franke A, Gaede KI, Schurmann M, Petrek M, Mrazek F, Pabst S, Grohe C, Grunewald J, Ronninger M, Eklund A, Rosenstiel P, Hohne K, Zissel G, Muller-Quernheim J, Schreiber S. Genome-wide association analysis reveals 12q13.3-q14.1 as new risk locus for sarcoidosis. *Eur Respir J* 2013: 41(4): 888-900.
9. Hofmann S, Fischer A, Till A, Muller-Quernheim J, Hasler R, Franke A, Gade KI, Schaarschmidt H, Rosenstiel P, Nebel A, Schurmann M, Nothnagel M, Schreiber S, GenPhenReSa C. A genome-wide association study reveals evidence of association with sarcoidosis at 6p12.1. *Eur Respir J* 2011: 38(5): 1127-1135.
10. Hofmann S, Franke A, Fischer A, Jacobs G, Nothnagel M, Gaede KI, Schurmann M, Muller-Quernheim J, Krawczak M, Rosenstiel P, Schreiber S. Genome-wide association study identifies ANXA11 as a new susceptibility locus for sarcoidosis. *Nat Genet* 2008: 40(9): 1103-1106.
11. Levin AM, Iannuzzi MC, Montgomery CG, Trudeau S, Datta I, Adrianto I, Chitale DA, McKeigue P, Rybicki BA. Admixture fine-mapping in African Americans implicates XAF1 as a possible sarcoidosis risk gene. *PLoS One* 2014: 9(3): e92646.
12. Levin AM, Iannuzzi MC, Montgomery CG, Trudeau S, Datta I, McKeigue P, Fischer A, Nebel A, Rybicki BA. Association of ANXA11 genetic variation with sarcoidosis in African Americans and European Americans. *Genes Immun* 2013: 14(1): 13-18.
13. Voorter CE, Drent M, van den Berg-Loonen EM. Severe pulmonary sarcoidosis is strongly associated with the haplotype HLA-DQB1*0602-DRB1*150101. *Hum Immunol* 2005: 66(7): 826-835.
14. Rossman MD, Thompson B, Frederick M, Maliarik M, Iannuzzi MC, Rybicki BA, Pandey JP, Newman LS, Magira E, Beznik-Cizman B, Monos D, Group A. HLA-DRB1*1101: a significant risk factor for sarcoidosis in blacks and whites. *Am J Hum Genet* 2003: 73(4): 720-735.

15. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* 2013: 368(1620): 20120362.
16. Ward LD, Kellis M. Interpreting non-coding genetic variation in complex traits and human disease. *Nat Biotechnol* 2012: 30(11): 1095-1106.
17. Moller DR, Koth LL, Maier LA, Morris A, Drake W, Rossman M, Leader JK, Collman RG, Hamzeh N, Sweiss NJ, Zhang Y, O'Neal S, Senior RM, Becich M, Hochheiser HS, Kaminski N, Wisniewski SR, Gibson KF, Group GSS. Rationale and Design of the Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis (GRADS) Study. Sarcoidosis Protocol. *Ann Am Thorac Soc* 2015: 12(10): 1561-1571.
18. Costabel U, Hunninghake GW. ATS/ERS/WASOG statement on sarcoidosis. Sarcoidosis Statement Committee. American Thoracic Society. European Respiratory Society. World Association for Sarcoidosis and Other Granulomatous Disorders. *Eur Respir J* 1999: 14(4): 735-737.
19. Vukmirovic M, Yan X, Gibson KF, Gulati M, Schupp JC, Deluliis G, Adams TS, Hu B, Mihaljinec A, Woolard TN, Lynn H, Emeagwali N, Herzog EL, Chen ES, Morris A, Leader JK, Zhang Y, Garcia JGN, Maier LA, Collman RG, Drake WP, Becich MJ, Hochheiser H, Wisniewski SR, Benos PV, Moller DR, Prasse A, Koth LL, Kaminski N, Investigators G. Transcriptomics of bronchoalveolar lavage cells identifies new molecular endotypes of sarcoidosis. *Eur Respir J* 2021.
20. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee SB, Tian X, Browning BL, Das S, Emde AK, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen YI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Kottgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin KH, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell BD, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Pleiness J, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Schwartz DA, Seo JS, Seshadri S, Sheehan VA, Sheu WH, Shoemaker MB, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Van Den Berg DJ, Vasan RS, Viaud-Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng LC, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman EK, Qasba P, Gan W, Consortium NT-OfPM, Papanicolaou GJ, Nickerson DA, Browning SR, Zody MC, Zollner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, Abecasis GR. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021: 590(7845): 290-299.
21. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature* 2015: 526(7571): 68-74.
22. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistic Software* 2004.
23. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013: 10(1): 5-6.
24. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2011: 9(2): 179-181.

25. Wendland J, Walther A. Genome evolution in the eremothecium clade of the Saccharomyces complex revealed by comparative genomics. *G3 (Bethesda)* 2011: 1(7): 539-548.
26. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009: 5(6): e1000529.
27. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, Weir BS. HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J* 2014: 14(2): 192-200.
28. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010: 11(7): 499-511.
29. Fingerlin TE, Zhang W, Yang IV, Ainsworth HC, Russell PH, Blumhagen RZ, Schwarz MI, Brown KK, Steele MP, Loyd JE, Cosgrove GP, Lynch DA, Groshong S, Collard HR, Wolters PJ, Bradford WZ, Kossen K, Seiwert SD, du Bois RM, Garcia CK, Devine MS, Gudmundsson G, Isaksson HJ, Kaminski N, Zhang Y, Gibson KF, Lancaster LH, Maher TM, Molyneaux PL, Wells AU, Moffatt MF, Selman M, Pardo A, Kim DS, Crapo JD, Make BJ, Regan EA, Walek DS, Daniel JJ, Kamatani Y, Zelenika D, Murphy E, Smith K, McKean D, Pedersen BS, Talbert J, Powers J, Markin CR, Beckman KB, Lathrop M, Freed B, Langefeld CD, Schwartz DA. Genome-wide imputation study identifies novel HLA locus for pulmonary fibrosis and potential role for auto-immunity in fibrotic idiopathic interstitial pneumonia. *BMC Genet* 2016: 17(1): 74.
30. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, Loyd JE, Cosgrove GP, Lynch D, Groshong S, Collard HR, Wolters PJ, Bradford WZ, Kossen K, Seiwert SD, du Bois RM, Garcia CK, Devine MS, Gudmundsson G, Isaksson HJ, Kaminski N, Zhang Y, Gibson KF, Lancaster LH, Cogan JD, Mason WR, Maher TM, Molyneaux PL, Wells AU, Moffatt MF, Selman M, Pardo A, Kim DS, Crapo JD, Make BJ, Regan EA, Walek DS, Daniel JJ, Kamatani Y, Zelenika D, Smith K, McKean D, Pedersen BS, Talbert J, Kidd RN, Markin CR, Beckman KB, Lathrop M, Schwarz MI, Schwartz DA. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* 2013: 45(6): 613-620.
31. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genome-wide association scans. *Bioinformatics* 2010: 26(17): 2190-2191.
32. Yang SH, Andl T, Grachtchouk V, Wang A, Liu J, Syu LJ, Ferris J, Wang TS, Glick AB, Millar SE, Dlugosz AA. Pathological responses to oncogenic Hedgehog signaling in skin are dependent on canonical Wnt/beta3-catenin signaling. *Nat Genet* 2008: 40(9): 1130-1135.
33. Franke A, Fischer A, Nothnagel M, Becker C, Grabe N, Till A, Lu T, Muller-Quernheim J, Wittig M, Hermann A, Balschun T, Hofmann S, Niemiec R, Schulz S, Hampe J, Nikolaus S, Nurnberg P, Krawczak M, Schurmann M, Rosenstiel P, Nebel A, Schreiber S. Genome-wide association analysis in sarcoidosis and Crohn's disease unravels a common susceptibility locus on 10p12.2. *Gastroenterology* 2008: 135(4): 1207-1215.
34. Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* 2016: 99(6): 1245-1260.
35. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015: 4: 7.
36. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013: 45(6): 580-585.
37. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015: 348(6235): 648-660.
38. Rivera NV, Ronninger M, Shchetynsky K, Franke A, Nothen MM, Muller-Quernheim J, Schreiber S, Adrianto I, Karakaya B, van Moorsel CH, Navratilova Z, Kolek V, Rybicki BA, Iannuzzi MC, Petrek M, Grutters JC, Montgomery C, Fischer A, Eklund A, Padyukov L, Grunewald J. High-Density Genetic Mapping Identifies New Susceptibility Variants in Sarcoidosis Phenotypes and Shows Genomic-driven Phenotypic Differences. *Am J Respir Crit Care Med* 2016: 193(9): 1008-1022.

39. Danila MI, Laufer VA, Reynolds RJ, Yan Q, Liu N, Gregersen PK, Lee A, Kern M, Langefeld CD, Arnett DK, Bridges SL, Jr. Dense Genotyping of Immune-Related Regions Identifies Loci for Rheumatoid Arthritis Risk and Damage in African Americans. *Mol Med* 2017: 23: 177-187.
40. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DR. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* 2011: 17(5): 792-798.
41. Handunnetthi L, Ramagopalan SV, Ebers GC, Knight JC. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes Immun* 2010: 11(2): 99-112.
42. Abe S, Yamaguchi E, Makimura S, Okazaki N, Kunikane H, Kawakami Y. Association of HLA-DR with sarcoidosis. Correlation with clinical course. *Chest* 1987: 92(3): 488-490.
43. Garman L, Pezant N, Pastori A, Savoy KA, Li C, Levin AM, Iannuzzi MC, Rybicki BA, Adrianto I, Montgomery CG. Genome-Wide Association Study of Ocular Sarcoidosis Confirms HLA Associations and Implicates Barrier Function and Autoimmunity in African Americans. *Ocul Immunol Inflamm* 2021: 29(2): 244-249.
44. Sato H, Woodhead FA, Ahmad T, Grutters JC, Spagnolo P, van den Bosch JM, Maier LA, Newman LS, Nagai S, Izumi T, Wells AU, du Bois RM, Welsh KI. Sarcoidosis HLA class II genotyping distinguishes differences of clinical phenotype across ethnic groups. *Hum Mol Genet* 2010: 19(20): 4100-4111.
45. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP, Wolf WA, e MT. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011: 4: 13.
46. Lareau CA, DeWeese CF, Adrianto I, Lessard CJ, Gaffney PM, Iannuzzi MC, Rybicki BA, Levin AM, Montgomery CG. Polygenic risk assessment reveals pleiotropy between sarcoidosis and inflammatory disorders in the context of genetic ancestry. *Genes Immun* 2017: 18(2): 88-94.
47. Rajoriya N, Wotton CJ, Yeates DG, Travis SP, Goldacre MJ. Immune-mediated and chronic inflammatory disease in people with sarcoidosis: disease associations in a large UK database. *Postgrad Med J* 2009: 85(1003): 233-237.
48. Li L, Silveira LJ, Hamzeh N, Gillespie M, Mroz PM, Mayer AS, Fingerlin TE, Maier LA. Beryllium-induced lung disease exhibits expression profiles similar to sarcoidosis. *Eur Respir J* 2016: 47(6): 1797-1808.

Table 1: Sample size and sex for Phase 1 and Phase 2*

	All (N= 2599)		Phase 1 (N=1799)		Phase 2 (N=800)	
Sex	Sarcoidosis (n= 1335)	Control (n=1264)	Sarcoidosis (n=818)	Control (n=981)	Sarcoidosis (n=517)	Control (n=283)
Female, n (%)	671 (50 %)	565 (45%)	399 (49%)	364 (37%)	272 (53%)	201 (71%)
Male, n (%)	664 (50%)	699 (55%)	419 (51%)	617 (63%)	245 (47%)	82 (29%)

*Phase 2 DNA samples became available after the Phase 1 group had been genotyped (see Materials and Methods).

Table 2. Summary of top 7 genome-wide significant SNPs of the genome-wide association study of sarcoidosis (p-value<$5x10^{-8}$)

					European American Cohort									African American Cohort		
						Phase 1			Phase 2			Meta-analysis				
SNP	Chr	position	Minor allele	Nearest gene (+/-25K)	MAF case	OR (95% CI)	p-value	MAF case	OR (95% CI)	p-value	OR (95% CI)	p-value	MAF case	OR (95% CI)	p-value
rs9269233	6	32451762	A	HLA-DRB9	0.38	1.73 (1.48,2.01)	5.67E-13	0.34	1.08 (0.85,1.36)	0.54	1.50 (1.32,1.71)	2.32E-10	0.32	1.09 (0.81,1.47)	0.56*
rs9271346	6	32583468	C	HLA-DQA1	0.27	1.71 (1.45,2.02)	1.53E-10	0.24	1.14 (0.89,1.47)	0.30	1.51 (1.32,1.74)	3.73E-09	0.24	1.19 (1.05,1.34)	0.01
rs35656642	6	32583610	A	HLA-DQA1	0.32	0.66 (0.58,0.76)	5.73E-09	0.34	0.84 (0.67,1.05)	0.12	0.71 (0.63,0.80)	1.19E-08	0.29	0.87 (0.78,0.98)	0.02
rs28589559	6	32587716	T	HLA-DQA1	0.08	0.50 (0.40,0.63)	2.90E-10	0.13	0.91 (0.67,1.23)	0.55	0.62 (0.52,0.74)	2.46E-08	0.13	0.93 (0.80,1.08)	0.33
rs9276935	6	32936441	C	BRD2 (inside gene)	0.05	0.51 (0.39,0.68)	1.26E-06	0.05	0.56 (0.37,0.84)	0.006	0.53 (0.42,0.66)	2.72E-08	0.01	0.87 (0.55,1.40)	0.57
rs3129888	6	32411726	G	HLA-DRA (inside gene)	0.27	1.63 (1.38,1.91)	3.71E-09	0.25	1.15 (0.90,1.48)	0.26	1.47 (1.28,1.68)	3.21E-08	0.27	1.36 (1.20,1.54)	1.52E-06
rs71549283	6	32505038	A	HLA-DRB5	0.33	0.61 (0.51,0.72)	9.70E-09	0.34	0.83 (0.63,1.10)	0.20	0.66 (0.57,0.77)	4.22E-08	0.20	0.86 (0.64,1.17)	0.34*

SNP: single nucleotide polymorphism; Chr: chromosome; MAF: minor allele frequency; OR: odds ratio; 95% CI: 95% confidence interval

*The results of rs9269233 and rs71549283 in the African American cohort did not pass the imputation quality control. The result was obtained from the sequencing data on 932 subjects.

Table 3. Association between HLA alleles and sarcoidosis.

	Phase 1				Phase 2					
	Dosage frequency		Univariate results		Dosage frequency		Univariate results		Meta-analysis	
HLA allele	Cases	Controls	OR (95% CI)	p-value	Cases	Controls	OR (95% CI)	p-value	OR (95% CI)	p-value
Novel alleles*										
DRB1*0101	0.07	0.18	0.35 (0.26, 0.48)	1.46E-11	0.09	0.23	0.31 (0.20, 0.47)	3.51E-08	0.34 (0.26, 0.43)	2.27E-18
DQA1*0101	0.12	0.22	0.53 (0.42, 0.68)	2.61E-07	0.12	0.29	0.34 (0.24, 0.50)	1.09E-08	0.47 (0.38, 0.57)	1.09E-13
DQB1*0501	0.13	0.24	0.61 (0.49, 0.75)	5.45E-06	0.12	0.30	0.39 (0.28, 0.55)	6.03E-08	0.53 (0.45, 0.64)	1.37E-11
Candidate alleles										
DRB1*1101	0.15	0.11	1.57 (1.13, 2.18)	7.05E-03	0.14	0.12	1.58 (0.91, 2.73)	0.10	1.57 (1.19, 2.08)	1.67E-03
DRB1*1501	0.40	0.27	1.56 (1.30, 1.86)	1.03E-06	0.35	0.29	1.20 (0.90, 1.61)	0.21	1.45 (1.25, 1.69)	1.75E-06
DQB1*0602	0.39	0.26	1.57 (1.31, 1.88)	1.11E-06	0.34	0.27	1.23 (0.91, 1.66)	0.19	1.47 (1.26, 1.72)	1.53E-06

OR: odds ratio; 95% CI: 95% confidence interval
*P-value < 5 x 10^{-8}, remaining significant HLA alleles (P<0.00011) are listed in Table S5

Table 4. P-values of the three genome-wide significant SNPs adjusted by HLA alleles (one at a time)

					Without HLA adjustment	DRB1*0101	DQA1*0101	DQB1*0501
SNP	Chr	Position	Minor allele	Nearest gene (+/-25K)	p-value	p-value	p-value	p-value
rs9269233	6	32451762	A	HLA-DRB9	2.32E-10	2.22E-07	2.41E-07	1.35E-07
rs9271346	6	32583468	C	HLA-DQA1	3.73E-09	1.85E-07	5.93E-08	2.17E-08
rs35656642	6	32583610	A	HLA-DQA1	1.19E-08	7.22E-05	4.81E-05	1.87E-05
rs28589559	6	32587716	T	HLA-DQA1	2.46E-08	0.03	0.02	3.88E-03
rs9276935	6	32936441	C	BRD2 (inside gene)	2.72E-08	2.43E-07	1.46E-07	1.49E-07
rs3129888	6	32411726	G	HLA-DRA (inside gene)	3.21E-08	1.38E-06	1.45E-06	1.22E-06
rs71549283	6	32505038	A	HLA-DRB5	4.22E-08	4.26E-05	3.96E-05	4.66E-05

SNP: single nucleotide polymorphism; Chr: chromosome; MAF: minor allele frequency; OR: odds ratio; 95% CI: 95% confidence interval

Table 5. Genotype-Tissue Expression Portal (GTEx) *Cis*-eQTL in lung and whole blood for associated SNPs.

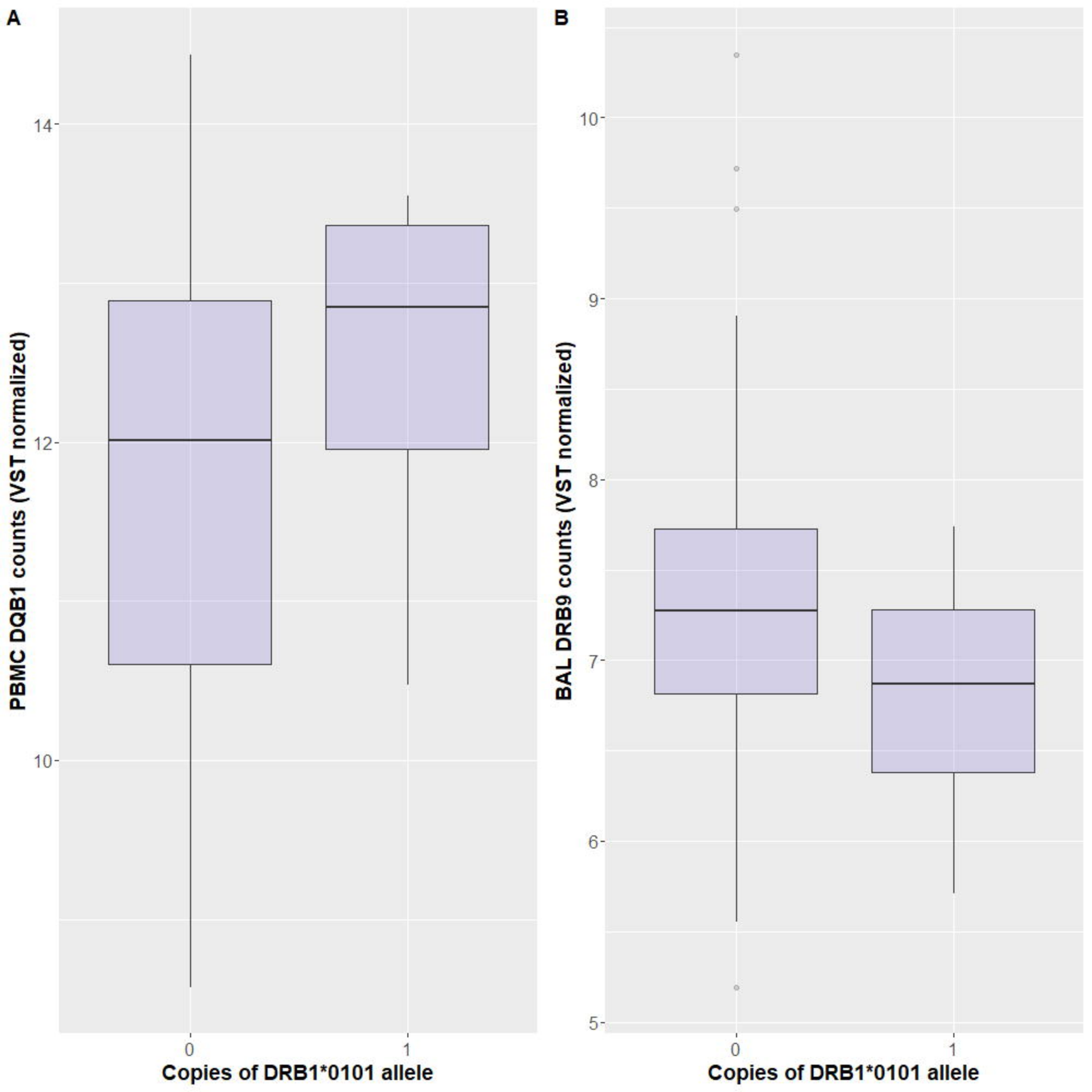
SNP	Nearest gene (+/- 25K)	Common for Lung and Whole blood	*Unique for Lung	*Unique for Whole blood
rs9269233	HLA-DRB9	HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DRB1, HLA-DRB5, HLA-DRB6	LY6G5C, STK19B	C4A, CYP21A2, HLA-DQB2, HLA-DRB9
rs9271346	HLA-DQA1	CYP21A2, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HLA-DRB6, HLA-DRB9	HCG23, XXbac-BPG154L12.4	LY6G5B, TNXB
rs35656642	HLA-DQA1	HLA-DQA1, HLA-DQA2, HLA-DRB5, HLA-DRB6, LY6G5B	HCG23, LY6G5C, XXbac-BPG154L12.4	CYP21A1P, HLA-DRB1
rs28589559	HLA-DQA1	HLA-DQA1, HLA-DQA2, HLA-DQB2, HLA-DRB6, HLA-DRB9	HLA-DOB, NOTCH4, TAP2	HLA-DQB1, PBX2
rs3129888	HLA-DRA (inside gene)	C4A, CYP21A2, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DRB5, HLA-DRB6, HLA-DRB9, STK19B	HCG23, HLA-DQA1, HLA-DRB1, XXbac-BPG154L12.4	HLA-DQB2

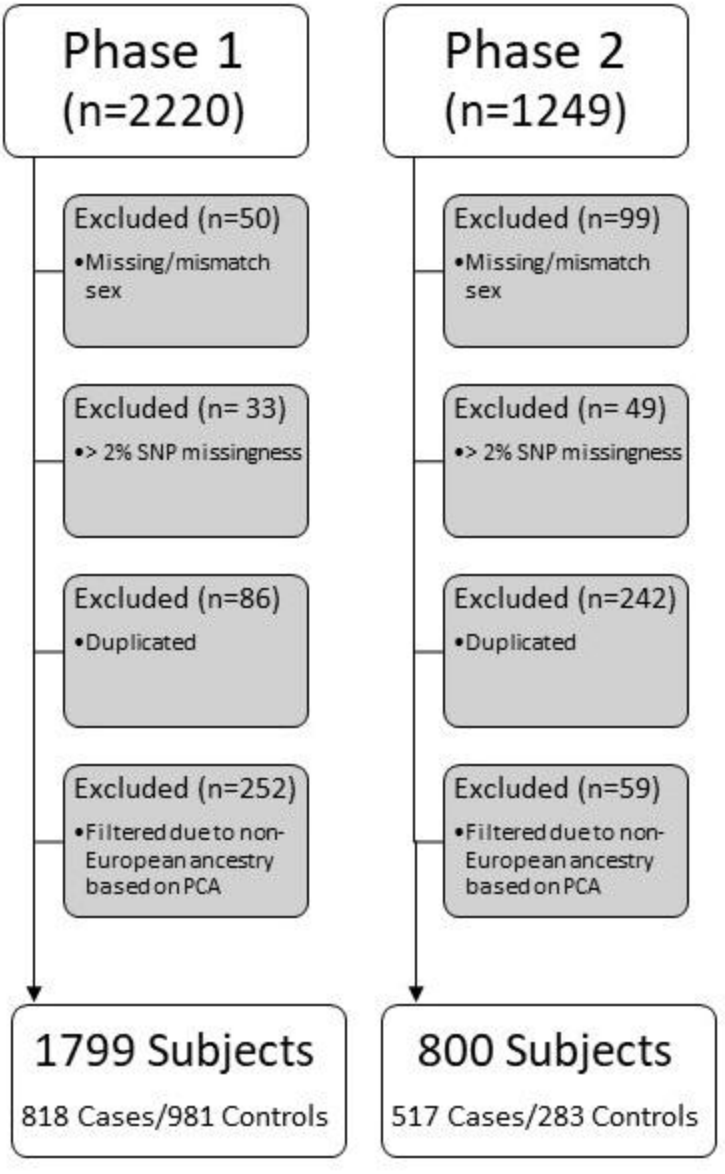
*Unique gene with its gene expression affected by the specific SNPs only in the specific tissue. For example. rs9269233 only affects *STK19B* gene expression in the lung but not in the whole blood.
rs9276935 were not eQTLs in lung or whole blood tissue; rs71549283 was not found in the GTEx database as of 5/31/2022

Figure 1. CONSORT flow chart for the GWAS samples. This figure outlines the exclusion of the subject in Phase 1 and Phase 2 respectively.

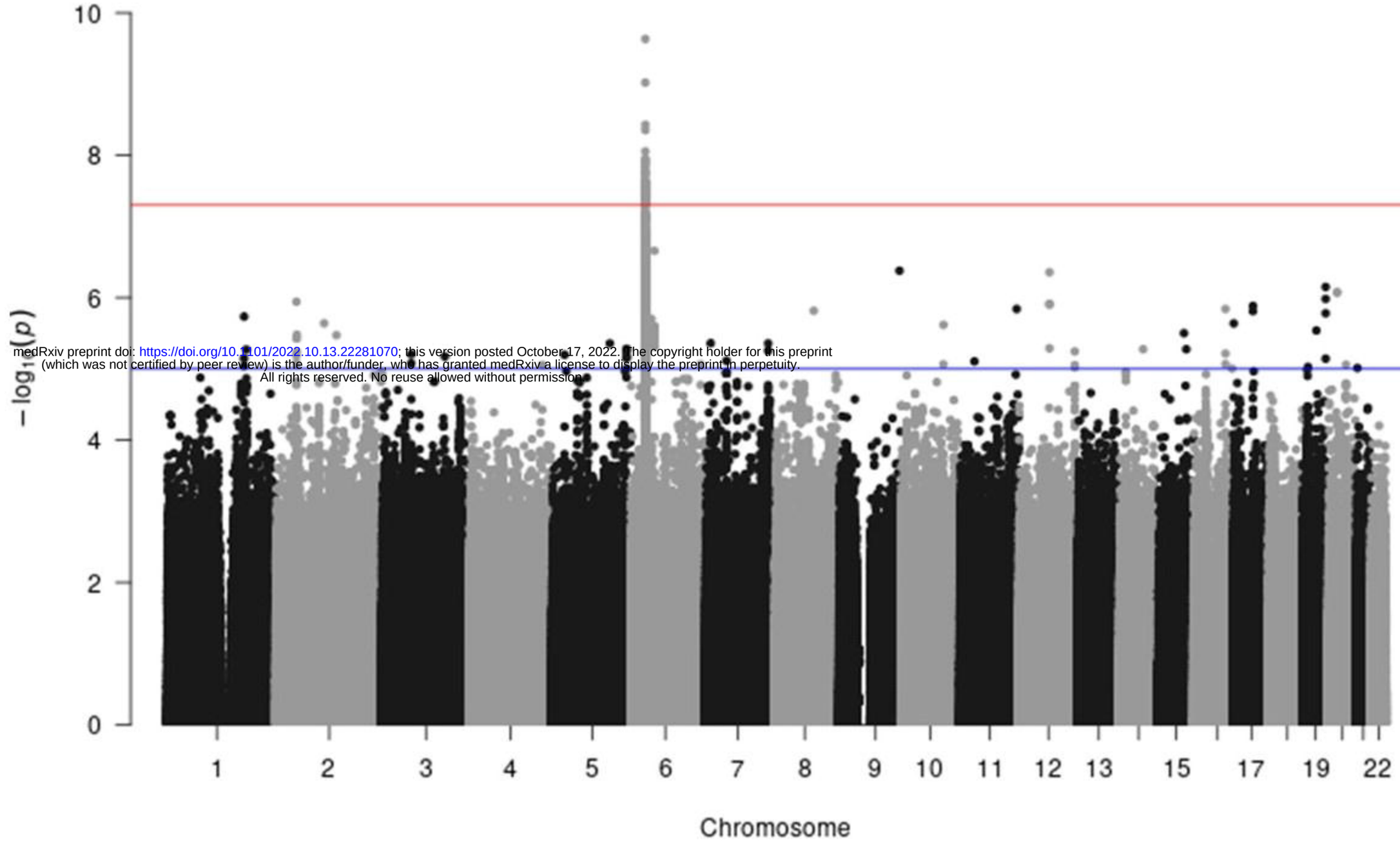
Figure 2. 2A) Manhattan plot for SNP associations, with significant associations in the HLA region of chromosome 6. 2B) Locus-zoom plot for significant SNPs in chromosome 6 (position: 32381726-32984689); rs9269233 is the most significant SNP and r^2 values with other significant SNPs are low.

Figure 3. A) PBMC DQB1 gene expression (variance stabilizing transformation [VST] normalized counts) and presence/absence of DRB1*0101 allele. Individuals with one DRB1*0101 allele have higher PBMC DQB1 gene expression compared to those with none. B) BAL DRB9 gene expression (VST normalized counts) and presence/absence of DRB1*0101 allele. Individuals with one DRB1*0101 allele have higher BAL DRB9 gene expression.


A
PBMC DQB1 counts (VST normalized)
14
12
10
0
1
Copies of DRB1*0101 allele
B
BAL DRB9 counts (VST normalized)
10
9
8
7
6
5
0
1
Copies of DRB1*0101 allele


Phase 1
(n=2220)
Excluded (n=50)
•Missing/mismatch sex
Excluded (n= 33)
•> 2% SNP missingness
Excluded (n=86)
•Duplicated
Excluded (n=252)
•Filtered due to non-European ancestry based on PCA
1799 Subjects
818 Cases/981 Controls
Phase 2
(n=1249)
Excluded (n=99)
•Missing/mismatch sex
Excluded (n= 49)
•> 2% SNP missingness
Excluded (n=242)
•Duplicated
Excluded (n=59)
•Filtered due to non-European ancestry based on PCA
800 Subjects
517 Cases/283 Controls

A



B