AWARD NUMBER: W81XWH-20-1-0686

TITLE: Multiancestral Genomic Approach to SLE Precision Medicine

PRINCIPAL INVESTIGATOR: Carl D. Langefeld

CONTRACTING ORGANIZATION: Wake Forest University Health Sciences

REPORT DATE: Oct 2022

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
Oct 2022	Annual	30SEPT2021 - 29SEPT2022

4. TITLE AND SUBTITLE

Multiancestral Genomic Approach to SLE Precision Medicine

5a. CONTRACT NUMBER
W81XWH-20-1-0686

5b. GRANT NUMBER
W81XWH-20-1-0686

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Carl D. Langefeld, Ph.D.

E-Mail: clangefe@wakehealth.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

WAKE FOREST UNIVERSITY
HEALTH SCIENCES
MEDICAL CENTER BLVD
WINSTON SALEM NC 27157-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We hypothesized that clinical heterogeneity in systemic lupus erythematosus (SLE) is partially due to genetic heterogeneity associated with SLE-risk single nucleotide polymorphisms (SNPs). We have observed considerable differences in SLE-risk allele frequencies across various race/ethnic groups, reflecting their unique microevolutionary histories. As a component of our hypothesis, we posit that unique genetic profiles across ancestral backgrounds will manifest as differences in the relative impact of key biological pathways on SLE-risk; pathways shared by all races/ethnicities but having differing magnitudes of impact at the population level. We have composed a race/ethnicity-specific and trans-race/ethnicity list of SLE-risk polymorphisms from public SLE Immunochip and GWAS studies. We have linked them to genes by function and computed intensive in silico systems biology and bioinformatic analyses across four ancestral groups (European, Amerindian/Hispanic, African, Asian). We have completed similar work using DNA methylation data. Although some pathways are shared (e.g., interferon) between SNP-associated and DNA methylation genes, the nucleic acid sensing pathway was identified in our DNA methylation study. We observed pathway differences by race/ethnicity in the SNP-associated studies; EA-dominant pathways included innate immune, myeloid cell function, and robust interferon response, AA-dominant pathways included aberrant B cell activity accompanied by ER stress and metabolic dysfunction, and Asian-dominant pathways include elevated oxidative stress, altered metabolism and mitochondrial dysfunction. Focused on these hypotheses and corresponding results, we have published two manuscripts, submitted one manuscript, and are preparing two manuscripts.

15. SUBJECT TERMS
Genetic, lupus

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	153	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

1. Introduction

Systemic lupus erythematosus (SLE) is an autoimmune disease in which the immune system mistakenly attacks an individual's own cells, causing inflammation and organ damage. SLE disproportionately affects women, particularly women of child-bearing age and non-European ancestries. There are significant differences in SLE risk and clinical manifestations (heterogeneity) by race/ethnicity. Only a small number of FDA approved drugs are available to treat SLE. In addition to the clinical heterogeneity, there is significant genetic heterogeneity related to the >110 known SLE-risk genetic polymorphisms. Genes form networks which execute specific biological functions. Identifying connections among genes (pathway analysis) may increase the number of relevant drug targets, resulting in novel therapies. The opportunity exists to leverage these genetic associations and genomic differences to identify genetically motivated drug repositioning targets. Our aims are to: 1) Link ancestry-specific and ancestry-shared genomic markers (SNPs, CpG methylation sites, gene transcript expression) associated with SLE risk in women to specific genes (identify genes); 2) complete systems biology and pathway analyses on these genes to identify ancestry-specific and ancestry-shared drug targets (identify/prioritize targets); and 3) identify and prioritize FDA-approved drugs for potential advancement into lupus clinical trials or preclinical models (identify/prioritize drugs).

2. Keywords

Systemic lupus erythematosus, drug repositioning, genetics, transcriptomics, autoimmune disease, gene networks, molecular docking, genetic polymorphisms, methylation, systems biology

3. Accomplishments

What were the major goals of the project?

The first 24 months of the project have shown significant progress in all three specific Aims. We believe we remain on or slightly ahead of schedule relative to the specific aims. There has not been any deviation from the primary objectives in these aims, but we have taken advantage of new protein databases for the *in silico* binding experiments and expanded some analyses using novel methods.

Major accomplishments of the past year built upon the those of the first year of funding. We expanded and updated the identification and compilation of SLE risk SNPs datasets and linked them to their respective plausible gene-targets, across multiple ancestries. A key step forward was the re-analysis of a genome-wide association study for systemic lupus erythematosus in African Americans, complimenting the Immunochip associations in African Americans. The successful implementation of the programming pipelines for high-throughput analysis of *in silico* binding for drug targets with FDA-approved small molecules enabled us to complete these binding experiments in an ongoing manner throughout the year. We have completed these binding experiments for all cGenes (genes where the lupus risk allele results in an amino acid change) with known protein structures. We have identified a priority list of eGenes (genes where the lupus risk allele correlates with gene expression; an eQTL) and completed a set of *in silico* binding experiments for sets of eGenes.

We provide two tables outlining the project's accomplishments and progress. **Table 1** provides a high-level list of progress and major task completions. **Table 2** provides additional detail and relates these accomplishments to the tasks (Specific aims) and subtasks provided in the project's original SOW. It includes estimates of the overall progress for the entire grant period.

Table1: General list of Major Accomplishments (Months 12-24).

Major Accomplishment/Progress Category	Descriptor

Manuscripts and Presentations	• Four accepted/published manuscripts
	• Two additional manuscripts in preparation
	• Two conference posters/presentations
	  ○ Lupus 21st Century 2021 Conference (talk) and
	  ○ American College of Rheumatology 2022 meeting (poster and lightning talk)

Data processing and Analytic Pipelines	• Drug target identification and processing for molecular docking • Integration of SMILES identifiers of FDA approved drugs (to facilitate drug structural similarity analyses) • Expanded summarization of molecular docking results, summaries include generation of $\Delta\Delta G$ for paired docking comparisons (e.g., risk versus non-risk protein isoforms for drug target proteins), cluster analysis of molecular similarity, and graphical illustrations.
Created Datasets	• Expanded and updated SLE risk SNP datasets linked to their most plausible-implicated gene targets (e.g., gene expression, eGenes; and protein coding variants cGenes). • African American systemic lupus erythematosus genome-wide association study results (partially complete) • ZINC 15 (FDA-approved small molecules) IDs linked to public repositories via identifiers (e.g., SMILES, pubchemID, and common drug names). • Database of ΔG and $\Delta\Delta G$ binding energies for ZINC 15 molecules across more than 20 cGene-linked drug targets (2 structures per drug target, corresponding to the risk and non-risk amino acid linked to the SLE risk SNP) and 5 eQTL-based. • Database of Euclidean similarity distance for more than 1,400 FDA approved drugs and linking these distances to the *in silico* binding results for completed targets.
System biology analyses	• Additional pathways and system biology analysis for SLE-associated SNPs in four ancestries/ethnicities and their union and intersections of genes for druggable pathway identification (African ancestry, Asian ancestry, European ancestry, and non-African Hispanic ethnicity). This was the focus of portions of published papers. • Pathway analysis based on results of Mendelian randomization studies with atherosclerosis. Manuscript in press at Cell Reports Medicine

Table 2: Project completion and status (Months 12-24) as related to proposal's original SOW.

Goals and Milestones as listed in the original SOW.		Progress Report 10-2022 Update. Completion Status
Specific Aims (specified in proposal)	**Timeline**	
Specific Aim 1: Identify SLE-risk Genes	**Months**	Reporting Period (Months 13-24) Updates
Subtask 1: Identify SLE-risk single nucleotide polymorphisms (SNPs) in women	1-24	• SNP associations compiled from the African American Immunochip data. • The African American GWAS study is still in progress and the resulting SLE-risk SNPs need to be integrated with Immunochip. **Completion Progress ~85%**
Subtask 2: Link SLE-risk SNPs to genes via eQTL, proximity, transcription factor binding, protein coding, gene-based testing	3-27	• SLE-risk SNPs linked to relevant gene and protein change based on SLE studies across four (European, African, Asian, and Hispanic) ancestries are underway for the cGenes (completed) and eGenes using GTEx and public data (near complete). • SLE-risk SNPs linked via transcription factor binding sites (tGenes) being revised and updated with new information **Completion Progress ~80% complete.**
Subtask 3: Transcriptomic analysis, differential expression of genes identified in subtask 2	3-30	• Comparative analysis of differentially expressed target genes (i.e., cGenes, tGenes, eGenes, pGenes) are summarized and discussed in two manuscripts (first twins paper published in Genes; mendelian randomization in press in Cell Reports Medicine) • Differentially expressed genes identified by annotation from DNA methylation study in MZ female twins discordant for SLE summarized in submitted manuscript. • Identification of genes where SLE-risk allele increases gene expression, that gene's expression is increased in lupus patients within at least one of five relevant tissues or cell types **Completion Progress ~75% complete.**
Subtask 4: DNA methylation analyses, differential methylation of genes identified in subtask 2	1-24	• Developing list of appropriate (e.g., T and B-cell sources) datasets for methylation analysis – completed but will expand if new data is published. • Differential methylation analysis in MZ female twins who are discordant for SLE yielded gene lists which were compared and contrasted with genes in Subtask 2 (published in Genes). **Completion progress ~100%**
Subtask 5: Identify and write potential manuscripts on multi-omic analysis of SLE-risk associated variants and genes.	6-36	• Publication of Mendelian randomization paper identifying shared pathways between systemic lupus erythematosus and coronary artery disease (Cell Reports Medicine, in press – available Nov 4, 2022). • Publication of manuscript exploring the differential analysis of methylation in twins who are discordant for SLE and identifying pathways and listing relevant drug targets from pathway analysis. • Publication, in collaboration with Timothy Niewold MD, on single-cell expression quantitative trait loci (eQTL) analysis of SLE-risk loci in lupus patient monocytes. Focus on select genes of relevance. **Completion Progress - ongoing**
Milestone(s) Achieved: Lists of SLE-risk associated genes informed by ancestry, tissue, and female sex	3-30	• Ongoing expansion of lists of SLE-risk genes informed by ancestry and female-sex. • Monitoring new published results, modification/integration of African American GWAS. **Completion progress ~80%**
Local IRB/IACUC Approval	Completed	**Completed and annually renewed**
Specific Aim 2: For genes and gene lists discovered in Specific Aim 1, complete systems biology, and pathway analysis		Reporting Period (Months 13-24) Updates
Subtask 1: Identify drug targets: Process lists of genes in Aim 1 into one of four Target Groups based on functional criteria, including pathway analyses	3-30	• Drug targets identified via cGenes, eGenes and tGenes (transcription factor binding), with a planned expansion and continuous updating based on African American GWAS and new literature. Largely complete for European, Hispanic, and Asian Ancestries and African American Immunochip-based studies. • Identified targets placed in one of the four Target Groups **Completion progress ~90%**
Subtask 2: Prioritize drug targets (i.e., genes): Prioritize genes first by group assignment and second RILITE's scoring algorithm within each group. Targets with highest prioritization will be assessed for molecular docking (e.g.,	3-30	• Continuous prioritization of targets based on high-quality structures includes the protein databank (PDB) and now, AlphaFold (available as of July 2021). • Current effort aligns SLE-risk allele associated with increased gene

quality protein structures in Protein Data Bank).		expression and that gene's gene expression in one of five tissues/cell types associated with risk of lupus or subtypes **Completion progress ~60%**
Subtask 3: Identify and write potential manuscripts incorporating systems biology and drug target prioritization to evaluate genetic architecture of SLE.	12-36	• See above, multiple publications and annual meeting presentation/posters. • Ongoing task **Completion progress ~60%**
Milestone(s) Achieved: Lists of prioritized drug targets	6-30	• Achieved list of prioritized drug targets for SLE-associations mapping to c-genes (protein-coding changes) and eGenes. • Identified optimal Protein Data Bank structures and AI-based alternative source for missing PDB structures (AlphaFold AI-predicted structures). • Requires manual look up and vetting, time consuming. **Completion progress ~40%**
Specific Aim 3: Identify and prioritize drugs		Reporting Period (Months 13-24) Updates
Subtask 1: Bioinformatic analysis for gene-drug and protein-drug interaction using STITCH, DrugPath, CLUE, etc.	6-36	• Continuous task as list updates • For current list of genes (Aim 1, 2), completed and summarized across multiple manuscripts and presentations at national meetings **Completion progress ~60%**
Subtask 2: Screen libraries of FDA-approved small molecules via molecular docking to identify drugs or small molecules for selected (Aim 2, Subtask 3) SLE drug targets	6-36	• Updated python pipeline to annotate binding data of FDA-approved small molecules with common-names and common-database identifiers (e.g., SMILES, pubChemID) • c-Gene complete with existing PDB structures or AlphaFold structural information (e.g., those without PDB structures). • Exploration of docking sites relative to amino acid change due to SLE-risk allele • Summarized differential binding between risk and non-risk SLE-protein structures using metrics developed in year 1. • Novel approach not in original grant, using SMILES complete various cluster analyses to identify clusters of drugs with similar structures and test for enrichment or "hot spots" in similarity space where there is an enrichment of cGenes. Expanding to eGenes and combining with cGenes. **Completion Status: ~50%**
Subtask 3: Prioritize drugs from Subtasks 1 and 2 using CoLTS scoring algorithm	6-36	• Identified limitations of CoLTS scoring algorithm for off target applications. • Modifying and focus on toxicity reports from CoLTS **Completion Status: ~25%**
Milestone(s) Achieved: Lists of genetically-informed FDA-approved drugs and small molecules, novel to treatment of SLE.	12-36	• List of molecular docking results for over 45,000 analyses. • Current focus on eGenes, tGenes. **Completion status: ~50%**

What was accomplished under these goals?

Significant Results:

Here we summarize key areas of progress. We are at a phase in the project where the synergy across aims is important, leveraging results from Aim 1 and 2 to inform Aim 3 and conversely learning from that experience to update (e.g., integrating new literature, modified pipelines) Aims 1 and 2. We start by briefly summarizing manuscripts submitted, one published (Genes), one in press (Cell Reports Medicine, Nov 4th embargo date), one under review (Nature Genetics), one published in Arthritis Research and Therapy. These manuscripts are listed in the Appendix. We continue by summarizing further progress roughly as proposed in the grant. As noted last year, the full list of cGenes, tGenes, eGenes and pGenes is too long to include in this report. Below as we summarize new work, we will highlight specific sets of genes (drug targets). The published manuscripts provide extensive lists of these genes. We do not repeat descriptions of the ongoing work from the first 12 months (e.g., binding experiments) but focus on additional, new results from the reporting period (months 13-24). Such descriptions are in the previous year's report but are also available upon request.

Published work this year.

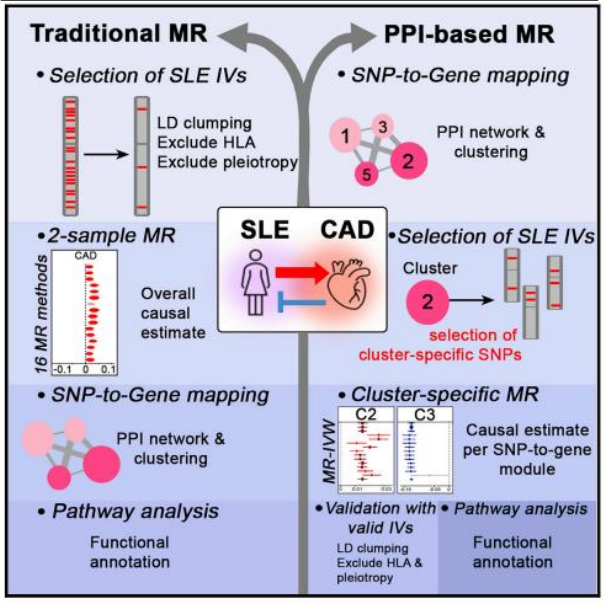
Epigenetics of discordant twins. Last year we summarized the results of the *Nucleic Acid-Sensing and Interferon-Inducible Pathways Show Differential Methylation in MZ Twins Discordant for Lupus and Overexpression in Independent Lupus Samples: Implications for Pathogenic Mechanism and Drug Targeting*. This manuscript was revised, resubmitted, and is now published in Gene (see Appendix)

Pan-autoimmune risk loci. Autoimmune and inflammatory diseases are polygenic disorders of the immune system. Many regions of the genome harbor risk alleles for several diseases, but the limited resolution of genetic mapping prevents determining if the same allele drives risk for multiple diseases or multiple variants within the same region generate distinct risk. If risk alleles are shared across multiple diseases, it suggests there may be a shared underlying mechanism. Using a collection of 129,058 cases and controls across six diseases, including systemic lupus erythematosus, and a novel methods called Joint Likelihood Mapping, we estimate that ~40% of overlapping associations are due to the same allele. We improve fine-mapping resolution for shared alleles by nearly two-fold by combining cases and controls across diseases, allowing us to identify more eQTLs driven by the shared alleles, hence same gene. The patterns of sharing indicate widespread shared mechanisms, but not a single global autoimmune mechanism. The results from this research provide an exciting opportunity for our current grant to include a specific focus on these shared loci in our binding experiments and the downstream drug repositioning pipeline. If so, lupus can be a leader that informs other autoimmune disease research and drug repositioning opportunities. Our paper is again under review at Nature Genetics after a positive initial review but with extensive suggestions/requests. (Please see Appendix: *Genetic mapping across autoimmune diseases reveals shared associations and mechanisms.*)

Singe-cell targeted transcriptomics. In collaboration Timothy Niewold, MD (Hospital for Special Surgery), we completed an expression quantitative trait locus (eQTL) analysis in single classical (CL) and non-classical (NCL) monocytes from patients with systemic lupus erythematosus (SLE) to quantify the impact of well-established genetic risk alleles on transcription at single-cell resolution. Single-cell gene expression was quantified using qPCR in purified monocyte subpopulations ($CD14^{++}CD16^{-}$ CL and $CD14^{dim}CD16^{+}$ NCL) from SLE patients. A novel analysis method, two-part hurdle mixed model, was used to control for the within-person correlations observed while testing for eQTLs between cell types and risk alleles. We observed that the SLE-risk alleles demonstrated significantly more eQTLs in NCLs as compared to CLs ($p = 0.0004$). There were 18 eQTLs exclusive to NCL cells, 5 eQTLs exclusive to CL cells, and only one shared eQTL, supporting large differences in the impact of the risk alleles between these monocyte subsets. The *SPP1* and *TNFAIP3* loci were associated with the greatest number of transcripts. Patterns of shared influence in which different SNPs impacted the same transcript also differed between monocyte subsets, with greater evidence for synergy in NCL cells. *IRF1* expression demonstrated an on/off pattern, in which expression was zero for all monocytes studied from some individuals, and this pattern was associated with a number of SLE risk alleles. We observed corroborating evidence of this IRF1 expression pattern in public data sets. Thus, we observed that multiple SLE-risk allele eQTLs in single monocytes differ greatly between CL and NCL subsets. These data support the importance of the *SPP1* and *TNFAIP3* risk variants and the IRF1 transcript in SLE patient monocyte function. Please see Appendix for details, *Single-cell expression quantitative trait loci (eQTL) analysis of SLE-risk loci in lupus patient monocytes*.)

Systemic lupus erythematosus and coronary artery disease shared loci. We completed and had a new manuscript accepted which explored the shared genetic associations between systemic lupus erythematosus and coronary artery disease (CAD), title: *Mendelian randomization and pathway analysis demonstrate shared genetic associations between systemic lupus erythematosus and coronary artery disease* (please see Appendix for manuscript). In brief, CAD is a leading cause of death in patients with systemic lupus erythematosus (SLE). Despite clinical evidence supporting an association between SLE and CAD, pleiotropy-adjusted genetic association studies are limited and focus on only a few common risk loci. Here, we identify a net positive causal estimate of SLE-associated non-HLA SNPs on CAD by traditional Mendelian randomization (MR) approaches. Pathway analysis using SNP-to-gene mapping followed by unsupervised clustering based on protein-protein interactions (PPIs) identifies biological networks composed of positive and negative causal sets of genes. In addition,

Figure 1. Graphical Abstract for *Mendelian randomization and pathway analysis demonstrate shared genetic associations between systemic lupus erythematosus and coronary artery disease*



we confirm the casual effects of specific SNP-to-gene modules on CAD using only SNP mapping to each PPI-defined functional gene set as instrumental variables. This PPI-based MR approach elucidates various molecular pathways with causal implications between SLE and CAD and identifies biological pathways likely causative of both pathologies, revealing known and novel therapeutic interventions for Q2 managing CAD in SLE. For detailed results please see manuscript in Appendix.

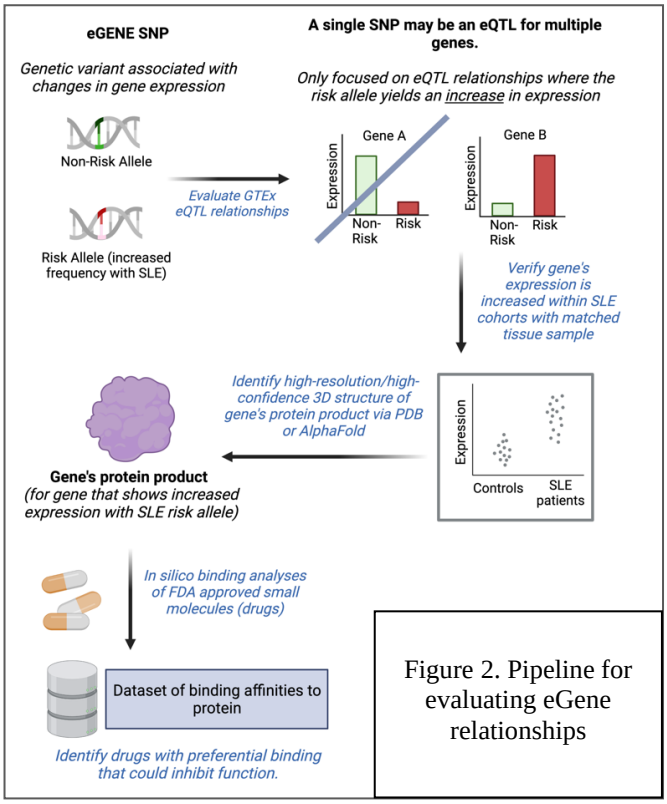


Figure 2. Pipeline for evaluating eGene relationships

Ongoing work this year.

Development and Implementation of eGENES drug-repositioning pipeline. Year 2 of the project continued development and implementation of the eGENEs analysis for drug repositioning (**Figure 2**). While much of this pipeline built upon the cGENEs pipeline (e.g., protein model assessment, *in silico* binding), there are some notable differences that required *de novo* programming implementation. For instance, identification of cGENEs was restricted to SLE-associated variants located within coding regions of the genome. Thus, only a limited number of SNPs met these criteria, mapping to 33 cGENEs. Contrarily, eGENEs are unrestricted by physical location, and thus, *any* SLE-associated SNP has the potential to map to an eGENE. Considering FDR-associated SLE SNPs from the Immunochip (three ancestries: EA, AA, and HA), this provided 1,545 SNPs for assessment, not including expansion due to linkage disequilibrium (identification of highly correlated SNPs). Secondly, while most coding SNPs will map to only a single cGENE; it is common for a single eQTL (SNP associated with gene expression) to map to multiple eGENEs. For example, in the GTEx (V8) dataset, there are 4,632,457 unique eQTLs (SNPs) mapping to 13,791,909 million unique eQTL-eGENE combinations (each eQTL mapping to 1-9 eGENEs). To handle the greater quantity of SLE-SNPs and eGENE mapping, we developed and implemented a pipeline to map SLE associations to the GTEx database and to assess direction of expression relative to SLE-risk allele. Importantly, we filtered our eGENE list to those where the SLE risk allele correlated with *increased* expression of the eGENE under the hypothesis that it is more biologically plausible to inhibit function (by binding), compared to up-regulating function (expression) of a target. From our primary list of 1,545 SLE-associated SNPs (ancestry specific SNPs), we first filtered to non-ambiguous SNPs and then identified 2,275 unique SNP-EQTLs meeting the aforementioned criteria, mapping to a list of 746 unique eGENEs within one or more of the relevant tissues available by the GTEx V8 database (Whole Blood, fibroblasts, leukocytes, kidney cortex, and sun-exposed skin). For 252 of these eGENEs, we were able to assess the relevance of these eGENEs in external, expression datasets of lupus patients. That is, while GTEx identified genes with increased expression with SLE risk alleles, we also verified that we observed increased expression of these eGENES within lupus patients. We leveraged five GEO datasets (**Table 3**) which offered comparable tissue sources for comparison with the GTEx data. We filtered our eGENE list to those genes that had corroborating expression in the lupus patients (increased expression). While this is a stringent filtering criterion, we believe this enables the best initial prioritization of potential drug targets. From this comparison between GTEx and lupus expression datasets, we were able to filter our list of 252 eGENEs (within comparable tissues) to 81 highly prioritized drug targets (**Table 4**). Notably, seven of these genes exhibited consistent evidence and direction in one or more of the GEO-compared tissues (**Table 5**). From these lists (**Table 4**), we are actively identifying high-resolution (PDB) or high-confidence (AlphaFold) three-dimensional structures for *in silico* binding. We are also expanding our eGENEs search to include SNPs from the trans-ancestral meta-analysis and

Table 3: GEO Datasets for comparison to GTEx eGENE results

GEO accession	GEO Tissue	SLE cases	Controls	GTEx Tissue Comparison
GSE39088	Whole blood	78	64	Whole blood
GSE45291	Whole blood	292	20	Whole blood
GSE50772	PBMC	61	22	EBV transformed lymphocytes
GSE81622	PBMC	30	25	EBV transformed lymphocytes
GSE109248	Skin	25	14	Sun exposed lower leg

SNPs in linkage disequilibrium from the Immunochip study.

Table 4. GTEx-identified genes that yielded consistent pattern of expression in SLE-patient datasets (increased expression), in one of three relevant tissues.

SLE Tissue (GTEx Tissue)	eGENES exhibiting increased expression with SLE risk allele in GTEx and increased expression in SLE cases in expression studies (GEO)							
Whole Blood (Whole Blood)	*CCDC136*	*DDX42*	*FDFT1*	*GLS*	*IL12RB2*	*LINC01270*	*MAP3K11*	*MED28**
	MFN2	*MRPS7**	*OASL*	*PPIL3**	*PRMT7**	*PTPRJ*	*RPS6KB1*	*SIPA1**
	*SSBP4**	*VRK2*						
PBMC (EBV-Lymphocytes)	*ARRB2**	*CTTNBP2NL*	*FAM167A*	*IDUA*	*MED28**	*MRPS7**	*PRMT7**	*RMI2*
	SPATS2L	*TIMM10*						
Skin (Sun-Exposed Skin)	*ALDH2*	*ARRB2**	*ATAD3C*	*B3GALT6*	*C17orf107*	*C1QTNF4*	*CAMTA2*	*CAPG*
	CASP10	*CBFA2T2*	*CCDC88B*	*CD38*	*CD79B*	*CNOT3*	*COQ9*	*CTSB*
	DCPS	*EIF6*	*ELOVL7*	*ESRP2*	*EVI5*	*FAM86B3P*	*FBF1*	*FUT11*
	GRB2	*HDLBP*	*ICAM5*	*ILF3*	*IRF5*	*KRI1*	*LCAT*	*LCE1D*
	LCE1E	*LCE3C*	*LRRFIP2*	*MAP1LC3A*	*MED28**	*MRPL45*	*PDIK1L*	*PPIL3**
	PPP1R14B	*PRMT7**	*RAPSN*	*RAVER1*	*RPP25*	*RPTOR*	*SIPA1*	*SLC39A13*
	SLC44A2	*SMARCA4*	*SPRR1B*	*SRP68*	*SSBP4**	*SYN2*	*SYNJ2*	*TIMM29*
	TMEM94	*TRIM65*	*TTC21B*	*UVSSA*	*YIPF2*	*ZC3H3*		

*Gene present in more than one tissue

Table 5. Subset of Genes from Table 4. that were identified in more than one tissue.

	GTEx tissue	Whole Blood	Lymphocytes	Sun-Exposed Skin
	SLE tissue	Whole Blood	PBMC	Skin
Gene	*MED28*	■	■	■
	PRMT7	■	■	■
	MRPS7	■	■	
	PPIL3	■		■
	SIPA1	■		■
	SSBP4	■		■
	ARRB2		■	■

Expansion of protein model search to include high quality AI-predicted structures.

We observed that of the initial 33 cGENEs, nine had high resolution protein structures that covered the amino acid region of interest. Thus, we have integrated Alphafold's AI-based predicted structures into our pipeline (**Figure 3**). Alphafold is a new (released July 2021) database comprised of over 300,000 protein structures generated using a three-track neural network algorithm. Alphafold has released protein models for proteins spanning the UNIPROT database (Universal Protein Resource; Corsortia of EMBL-EBI, SIB, and PIR host institutions). Alphafold represents a major innovation in proteomics and thus, drug-binding studies. This year, we began incorporating AlphaFold into our protein model search pipeline with additional quality control checks on structures (e.g., quality assessment of tertiary structure). So, while experimentally derived structures remain as the preferred source, Alphafold provides a unique opportunity to extend the scope of drug repositioning, beyond current limitations by the PDB. Furthermore, as prediction algorithms continue refinement of tertiary structures (e.g., beyond alpha helices and beta sheets), our drug-repurposing pipelines are readied for their inclusion. For the cGENEs, we identified seven additional protein structures (UHRF, ZACN, WDFY4, LRRC34, CCL22, ATG16L2, and AGBL2) with suitable tertiary structures. However, given that our cGENEs analysis requires the analysis of two isoforms (that is, computationally altering a single amino acid in the structure), we proceeded with a conservative approach and opted to *reserve amino-acid alterations for experimentally derived structures*.

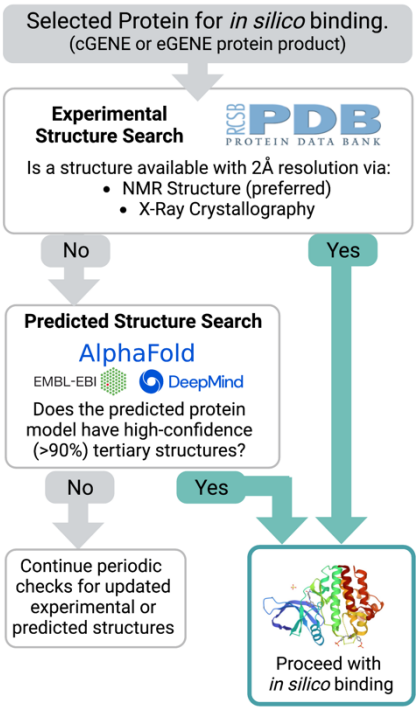


Figure 3. Assessment of protein structures from experimentally and predicted sources.

However, given that several of the aforementioned cGENES are also eGENES (e.g., *WDFY4, CCL22*), we proceeded with in silico binding for the single isoform available from Alphafold. As we incorporate Alphafold into our eGENEs analyses, we anticipate it could help yield as many as 50% more structures than the PDB, alone.

Development of drug-binding prioritization algorithm for cGENES

Figure 4. cGENE drug binding prioritization

Each *in silico drug* binding experiment yields a dataset of continuous binding affinities for 1431 unique FDA-approved small molecules (note: full dataset includes binding affinities for multiple conformers per drug). Work over the past year has focused on refining the produced drug-binding affinities for each gene target. This refinement enables drug prioritization for downstream assessments of biological (e.g., IPA; pathways analyses) and clinical (e.g., electronic health records) relevance. Within the scope of cGENEs, where we compare two isoform binding affinities per cGENE (defined by risk allele versus non-risk allele), we derived an algorithm that incorporates generally accepted thresholds of binding specificities as well as an assessment of the change in binding specificity between isoforms (**Figure 4**). Within cGENEs, this successfully focused our downstream assessments on a refined list of drugs, ranging from 0 to 70 drugs per target (**Table 6**). While our current analyses utilized a set of binary thresholds (**Figure 5**), we are actively exploring effects of continuous measures (e.g., isoclines)

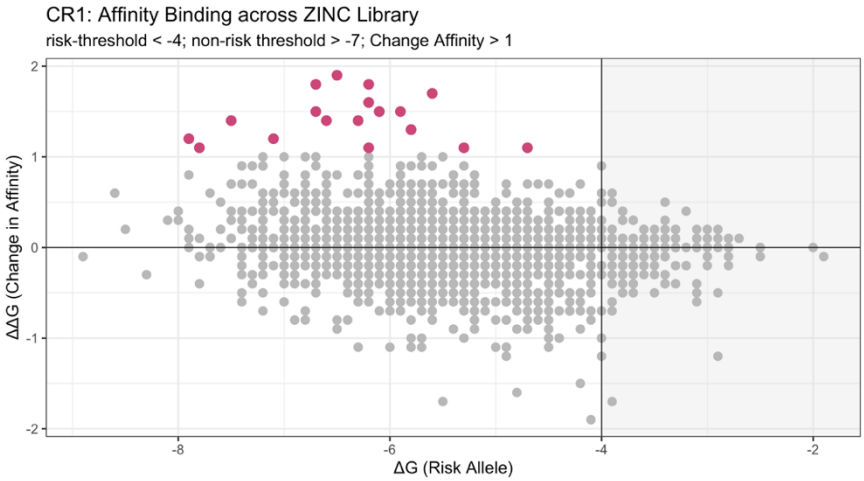


Figure 5. cGENE drug binding prioritization for CR1. X-axis depicts binding affinity for risk allele (risk isoform) while Y-axis depicts the change in binding affinity between the risk and non-risk isoforms. Application of prioritization algorithm limits the 1,431 analyzed drugs to just 20 (shown in red) for CR1. While these criteria represent binary thresholds, we are actively exploring continuous thresholds of prioritization.

Table 6: Prioritized Drugs per cGENE Target

cGENE Target	Prioritized Drugs (from 1,431)
CR1	20
FCGR2A	8
IFIH1	6
IRAK1	2
NT5E	70
PLAT	10
QARS	33
TNFAIP3	24
TYK2	0

Identifying Relationships among Prioritized Drugs.

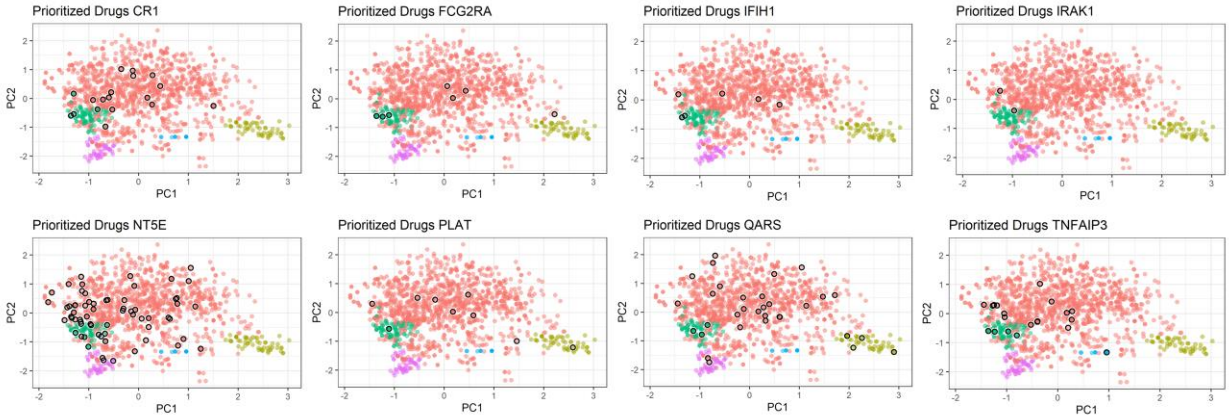
Post prioritization of drugs, a key question is whether any of the drugs relate to one another (within or across targets. For example, for the 20 prioritized drugs for CR1, do these represent 20 different mechanisms of binding and function, or does any overlap exist among these drugs? While future studies could investigate this from a clinical perspective (e.g., investigation of electronic health records for similarity/dissimilarity of outcomes associated with drug lists), within the scope of this project, we assessed the drugs from a structural approach. Given the inherent link between function and structure, we leveraged the drugs' molecular structures to identify potential relationships (similarities and dissimilarities). We utilized GlobalChem, a python-based software suite which uses a natural language algorithm to convert drug's SMILES (simplified molecular-input line entry

system) ID into a binary bit representation of the molecule. These numeric encodings can then be analyzed by dimensionality reduction methods such as principal component analyses (PCA) to describe the relative (structural) relationship among drugs. We applied this method to the 1,431 unique FDA-approved small molecules used within this project for *in silico* binding. GlobalChem then implements k-means clustering on the resulting principal components to identify groups of similarity. Optimal separation via k-means was observed for 5 clusters (in three-dimensional space).

The preliminary results from the k-means cluster analysis on the principal components capturing variation due to similarity were encouraging. However, different clustering algorithms have different strengths and weaknesses, so we compared four hierarchical clustering algorithms (agglomerative and divisive). Specifically, in addition to the k-means algorithm, we completed hierarchical clustering on principal components (HCPC), density-based spatial clustering of applications with noise (DBSCAN), and random forest clustering. Among these four algorithms, we summarize some of the patterns and interpretation of the random forest cluster analysis; we are still working through results of the DBSCAN analysis as that too has interesting results.

We computed an intra-feature random forest cluster (IRFC) analysis based on the first three principal components that capture the dominant 3D structural similarity among the 1,431 unique FDA-approved small molecules (drugs) (**Figure 6**). As with any random forest application, this unsupervised ensemble machine learning method repeatedly randomly samples from the 1,431 molecules many times to obtain an aggregate assignment for each molecule. We selected the IRFC as it can be more robust for complex data structures. Using elbow plots and gap statistics, the IRFC analysis identified five dominant clusters, subclustering is underway; k-means and HCPC also suggested five clusters, albeit with some variation in membership by individual drugs. In **Figure 6**, the salmon colored cluster is the central bulk of the molecular similarity (currently undergoing subcluster analysis), and the remaining four clusters are at the periphery of the PC space.

Figure 6.



We overlaid the drugs that met the above binding parameters (i.e., affinity binding across ZINC library) and computed an enrichment analysis (randomization test) overall and relative to each gene (**Table 7**). For example, in *CR1*, there are 20 drugs meeting the binding experiment parameters and these are highlighted as open black circles in **Figure 6**). We observed a global enrichment of Cluster 3 (green cluster in **Figure 6**); no other clusters exhibited a statistically significant enrichment (P>0.05). **Table 7** provides the count and p-value of the enrichment analysis for each gene. For example, 6 of 20 drugs that bound to the protein from *CR1* were in Cluster 3 (P-value=0.0044).

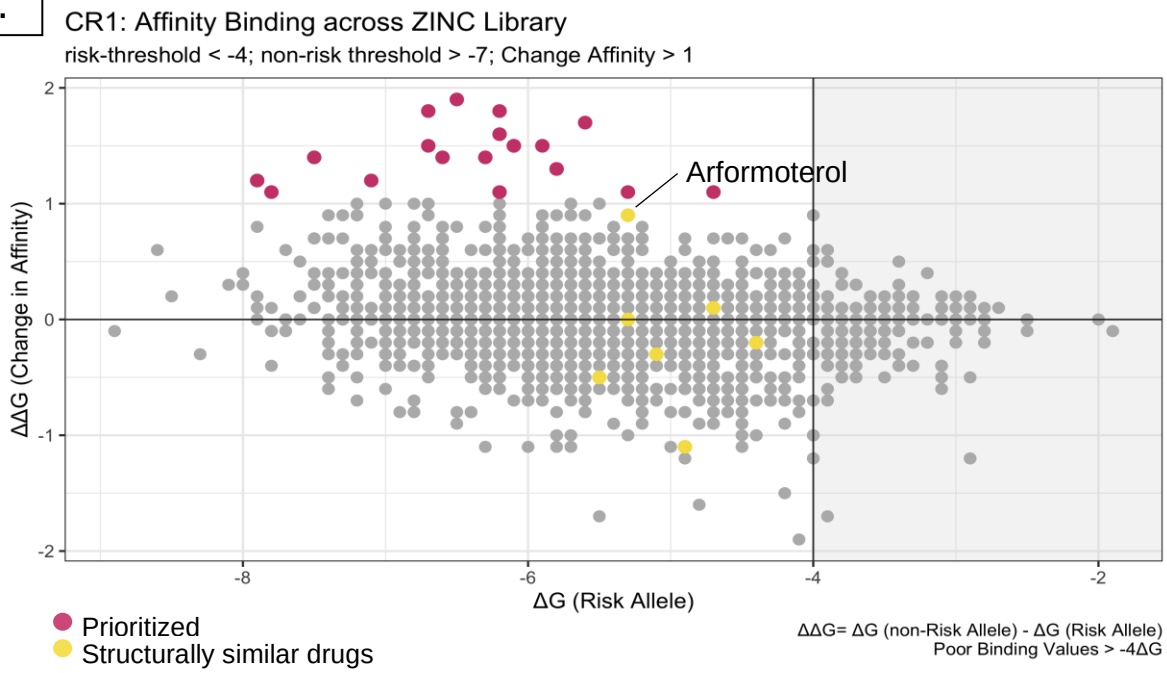
Table 7. Cluster 3 Enrichment of Drug Binding

	Number Drugs in Cluster 3			
Gene	No	Yes	Expected Number of Drugs by chance in Cluster 3	Enrichment P-value
CR1	14	6	1.67	**0.0044**
FCGR2A	4	3	0.58	**0.0156**
IFIH1	4	2	0.50	**0.0854**
IRAK1	2	0	0.17	0.8532
NT5E	52	14	5.66	**0.0014**

PLAT	9	1	0.83	0.5866
QARS	30	3	2.75	0.5344
TNFAIP3	16	6	1.83	**0.0104**

We observed an enrichment (hot spot) on the left side of Cluster 3 (green cluster) that was driving the global enrichment (**Figure 6**). We then identified those drugs in proximity (L_2-norm distance <0.25 from center of that hot spot): Fesoterodine, UNII-ZP145530CI, Arformoterol, (S)-Metoprolol, and (R)-Metoprolol. We explored their relative position in the binding experiment parameter space to determine if altering the parameters might provide more drugs of interest. Our interpretation of these results is that the analysis did not suggest alternative thresholds for these parameters. We illustrate with CR1 binding plots (**Figure 7**).

Figure 7.



Importantly, we emphasize that structurally comparable drugs can have widely varying functions, as evidenced by enantiomers, where often, one of the optical isomers is inert. Thus, innovatively, we are leveraging the structural similarity assessment ***alongside*** an *in silico* functional assessment (as captured by protein binding). For instance, we would not necessarily expect that all drugs tightly clustered in the structural assessment will be clustered in functional graph (**Figure 7**); but by comparing the two, we can identify instances where this does occur. This could indicate drugs that might show functional similarity, driven by some common structural feature. This can be leveraged in future assessments of EHR data (e.g., grouping outcomes across multiple drugs) as well as in future drug discovery, where these complementary assessments can help identify specific structural features for optimal binding to a given target.

In summary, significant progress has been made in the past year in identifying genes and linking those to specific drugs via the binding experiments. Progress has been made in each aim, and we are on or ahead of projected progress due to the successful pipeline development. In addition, we have integrated two significant additional components: 1) AlphaFold, the newly released AI-predicted protein structure that will significantly increase the number of targets explored, and 2) clustering to identify hot spots of binding drugs that allows potential expansion of the list of drugs for evaluation.

What opportunities for training and professional development has the project provided?

Mr. Nolan H. Hamilton completed his one-year training with us and started his graduate training at the University of North Carolina Chapel Hill Bioinformatics graduate program. For this grant, Mr. Nolan has assisted in the preparation of drug derivatives from the ZINC 15 database as SMILES strings for all of the cGenes. He also searched the drug list we identified for links to other autoimmune diseases. This has allowed

Drs. Ainsworth and Langefeld to mentor him in learning basic concepts of biostatistics, statistical genetics, and genetics.

We have hired a new staff biostatistician that has replaced Mr. Hamilton, Ekaterina S. Khvatkova (July 1, 2022, start date). Katya has recently graduated from Wake Forest University, Department of Mathematics and Statistics, with a MS in Mathematical Statistics. As the work outlined here is not standard basic biostatistics, she has been learning about genetics, proteins, while using her generalized linear mixed models and programming skills. She has been leading the programming effort to map eQTLs to genes and check to see that the risk allele is associated with increase gene expression and elevated gene expression is associated with lupus risk (see above).

In the past year, Dr. Langefeld took on a new PhD student, Olamide Arege, an individual with virology background from Nigeria who is seeking to do a PhD in bioinformatics. At present Olamide is still predominately taking the necessary mathematics and statistical theory classes, but he has participated in meetings and is soon to participate in the bioinformatic analyses. Funding for Olamide is from independent sources, including Dr. Langefeld's R&D account.

Dr. Langefeld is also mentoring an undergraduate student (Junior) in the new Department of Statistics at Wake Forest University, Hanna Vaidya. Hanna has worked with Dr. Langefeld and Ainsworth on the cluster analysis of the molecular similarity of drug based on SMILES identifiers. She learned about various hierarchical clustering methods and applied them. She programmed the random forest clustering analysis and the randomization test for cGene enrichment in the individual clusters. Current plans include teaching her the molecular docking analyses and allow her to use parts of the project as her honors thesis.

How were the results disseminated to communities of interest?

We have published multiple papers and presentations. Please see publication list and Appendix.

What do you plan to do during the next reporting period to accomplish the goals?

During the next reporting period (months 25-36), we will build upon the first two years of success by continuing progress towards milestones in each Aim. We will continue to use the pipelines developed, updated, and optimized in the first 24 months (e.g., summarization of ΔG and $\Delta\Delta G$ from molecular docking); leverage drug structural similarity and structural hot spot analysis (novel expansion of Aims); expand the high-throughput method for FDR-approved drug prioritization (Aim 3). **Table 8** outlines planned steps and work for the next reporting period. Planned activities are shown by each specific task and subtask as provided in the Project's original SOW.

Table 8. Goals and milestones

Goals and Milestones as listed in the original SOW.		**Progress Report 10-2022 Update.**
Specific Aims (specified in proposal)	**Timeline**	**Planned Activities for next Period (Months 25-36)**
Specific Aim 1: Identify SLE-risk Genes	**Months**	
Subtask 1: Identify SLE-risk single nucleotide polymorphisms (SNPs) in women	1-24	• Effectively complete except African American GWAS study • Continue curation of SLE risk SNPs across relevant sources of data as they become available.
Subtask 2: Link SLE-risk SNPs to genes via eQTL, proximity, transcription factor binding, protein coding, gene-based testing	3-27	• Evaluate gene-linking for SNPs associations uniquely within respective races/ethnicities. • Evaluate additional links to genes with updated ENCODE regulatory information (e.g., ensuring most plausible gene link for TFBS SNPs). • Gene-based testing, complete, no further work required.
Subtask 3: Transcriptomic analysis, differential expression of genes identified in subtask 2	3-30	• Differential expression analysis in five cell types for eGenes. Focus in next months on tGenes (transcription factor binding site linked genes) for genes not yet analyzed from Subtask 2.
Subtask 4: DNA methylation analyses, differential methylation of genes identified in subtask 2	1-24	• Analysis of differential methylation (in publicly available datasets) for genes identified in Subtask 2 complete.
Subtask 5: Identify and write potential manuscripts on multi-omic analysis of SLE-risk associated variants and genes.	6-36	• Construct additional publications, including mendelian randomization and our analysis pipeline and results.

Milestone(s) Achieved: Lists of SLE-risk associated genes informed by ancestry, tissue, and female sex	3-30	• Largely completed but will continue generation of SLE-risk genes by female sex and tissue-specificity; African American GWAS a task for this cycle.
Local IRB/IACUC Approval	Completed	
Specific Aim 2: For genes and gene lists discovered in Specific Aim 1, complete systems biology, and pathway analysis		
Subtask 1: Identify drug targets: Process lists of genes in Aim 1 into one of four Target Groups based on functional criteria, including pathway analyses	3-30	• Continue processing lists from Aim 1 into relevant Target groups (e.g., SLE risk SNPs from Immunochip study and African American GWAS). • Process to be ongoing as literature emerges.
Subtask 2: Prioritize drug targets (i.e., genes): Prioritize genes first by group assignment and second RILITE's scoring algorithm within each group. Targets with highest prioritization will be assessed for molecular docking (e.g., quality protein structures in Protein Data Bank).	3-30	• Continue prioritization and structure identification for targets from remaining categories (e.g., Groups 1 and 2). • Complete binding experiments for eGenes and tGenes, random forest cluster analysis of relative to eGenes and tGenes. • Prioritize and identify structures for drug targets based on transcription factor binding (Group 2).
Subtask 3: Identify and write potential manuscripts incorporating systems biology and drug target prioritization to evaluate genetic architecture of SLE.	12-36	• As additional groups of targets are identified and prioritized, prepare manuscript related to target identification (Aim 2) and drug identification/prioritization (Aim 3).
Milestone(s) Achieved: Lists of prioritized drug targets	6-30	• Partial list complete. • Generate additional lists of prioritized drug targets based on the four target groups (for analysis in Aim 3).
Specific Aim 3: Identify and prioritize drugs		
Subtask 1: Bioinformatic analysis for gene-drug and protein-drug interaction using STITCH, DrugPath, CLUE, etc.	6-36	• Continue pathways analyses of drug targets to better delineate biological system implicated in SLE and SLE subtypes. • Incorporate existing and continuing analyses into papers.
Subtask 2: Screen libraries of FDA-approved small molecules via molecular docking to identify drugs or small molecules for selected (Aim 2, Subtask 3) SLE drug targets	6-36	• Continue *in silico* binding of FDA-approved small molecules to identified drug targets. This includes both new targets (e.g., as identified by tGenes) and newly available high-quality protein structures (e.g., via PDB or AlphaFold). • Explore random forest clustering of SMILES data for hot spots.
Subtask 3: Prioritize drugs from Subtasks 1 and 2 using CoLTS scoring algorithm	6-36	• Continue prioritization and high-throughput prioritization of drugs identified in Subtask 2.
Milestone(s) Achieved: Lists of genetically informed FDA-approved drugs and small molecules, novel to treatment of SLE.	12-36	• For completed *in silico* molecular docking experiments, prioritize drugs based on CoLTS scoring and other relevant criteria. • Generate additional lists of prioritized drugs based on newly analyzed *in silico* molecular docking and near neighbor / random forest clustering. • Lists of prioritized drugs that are prioritized *across* drug targets (e.g., multiple targets per drug).

4. Impact

What was the impact on the development of the principal discipline(s) of the project?

As noted last year, the analytic and programming pipelines we have established that can be applied to other studies which are exploring precision medicine therapeutics based on disease-associated risk loci. For example, the ZINC 15 database is an established resource of FDA-approved small molecules. The development of efficient annotation of ZINC 15 compounds with publicly available datasets (e.g., SMILES identifiers, PubChem) enables efficient application of these methods, representing a reduction of preparatory

data steps needed. In addition to continuously updating the above, we have generated molecular similarity clusters of FDA approved drugs that can be tested for enrichment. Such molecular similarities might inform novel molecule/drug development. Importantly, many of the identified drug targets originate from genes that are implicated in other diseases (e.g., autoimmune diseases), as we show in our pan-autoimmune manuscript. Thus, molecular docking of FDA-approved compounds to these targets will be valuable datasets for similar studies in other diseases. Further, these pipelines will enable exploration of precision medicine for other lupus phenotypes and other ancestral groups.

What was the impact on other disciplines?

As reported last year, while developing this grant and completing the gene expression studies, we identified two major issues related to single-cell data analyses. We observed that, as a whole, the field of single-cell transcriptomics (and other single-cell omics) were not properly accounting for the correlation that exists within an individual/animal/organism. Thus, tests for differential expression, for example, were reporting too many false associations. Further, investigators were unintentionally grossly overestimating the statistical power of their studies in grants and completing significantly underpowered studies. As described in the listed publications below, of the 30 publications in rheumatological journal employing single cell gene expression, none of them were explicitly and properly accounting for the within person/animal correlation. There are multiple implications related to robustness, reproducibility, and cost-effective science: 1) false associations and inferences using single-cell methods will mislead our scientific efforts and fail to replicate, 2) entrenched ideas driven by enriched false associations from single-cell data will require significantly more resources to correct, 3) perceptions on the robustness of single-cell technology will be significantly damaged reducing acceptance of valid and robust studies. Motivated by these observations during the preparation of this grant and after award, we have published two manuscripts (listed below), are currently writing an invited review paper for the journal Rheumatology, and a response for Nature Communications. Dr. Langefeld's effort on the publications and publication costs for the Hierarchicell and review papers were partially supported by this grant (W81XWH-20-1-0686). This grant is gratefully acknowledged.

1. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun. 2021;12(1):738
2. Zimmerman KD, Langefeld CD. Hierarchicell: An R-package for estimating power for tests of differential expression with single-cell data. BMC Genomics. 2021 May 1;22(1):319. PMCID: PMC8088563.

During the past year, we have expanded upon this work. As two examples, from several, we collaborated with Dr. Timothy Niewold (see publications) to complete analyses that properly account for these correlations. Also, we wrote a rebuttal (tentatively accepted, see draft in Appendix) to a submission to "Rising Issues" that criticized the mixed model approach – the authors are not statisticians and do not understand the underlying mathematics of mixed models (i.e., they are mathematically provably wrong).

Dr. Langefeld also collaborated with other autoimmune genetics researchers to document the overlap in disease loci across several autoimmune diseases (see Appendix). As such, the work we are doing in this grant will potentially affect a broader set to autoimmune diseases.

3. Genetic mapping across autoimmune diseases reveals shared associations and mechanisms. Matthew R Lincoln, Noah Connally, Pierre-Paul Axisa, Christiane Gasperi, Mitja Mitrovic, David van Heel, Cisca Wijmenga, Sebo Withoff, Iris H Jonkers, Leonid Padyukov, International Multiple Sclerosis Genetics Consortium, Stephen S Rich, Robert R Graham, Patrick M Gaffney, Carl D Langefeld, Timothy J Vyse, David A Hafler, Sung Chun, Shamil R Sunyaev, Chris Cotsapas. (Under second review, Nature Genetics)

What was the impact on technology transfer?

Nothing to Report.

What was the impact on society beyond science and technology?

Nothing to Report.

5. Changes/Problems

Changes in approach and reasons for change

The general approach has not changed, but we have added additional, new resources and tools. Specifically, as noted last year, we encountered a limitation in the number of high-quality experimentally derived protein structures in the Protein Databank. As of July 2021, AlphaFold was officially released which contains neural network derived (predicted) structures. We have included this database in our pipeline, which markedly increases our coverage of drug targets to 100%; albeit we are discovering that some of the predicted structures need additional refinement (e.g., for tertiary structural features), and we are reaching out to the AlphaFold team, using their feedback system to prioritize molecules. We will continue to use experimentally derived structures when available; but the inclusion of AlphaFold presents a new opportunity to explore more of our identified targets.

We have also expanded our approach to leverage the structural similarity of the FDA-approved drugs by applying a natural-language learning algorithm to the SMILES (simplified molecular-input line-entry system) identifiers for the FDA-approved drug. From these data, we can generate structurally informed clusters of FDA-approved drug. Using several hierarchical clustering techniques, we found that the results from the Random Forest clustering algorithm enabled us to identify a potential hot spot within Cluster 3 of drugs that bind to the protein product of several cGenes. This is a novel idea that expands upon the immediate binding results and provides additional robustness to the binding results and identifies specific scenarios where structural similarity might be important to consider for drug development for certain targets.

Actual or anticipated problems or delays and actions or plans to resolve them

The Covid-19 pandemic has caused multiple problems, including the restriction from meeting in person and the inability to recruit potential interns or PhD students to the lab. For family reasons, Drs. Langefeld and Ainsworth are not able to travel during the pandemic; both serve as primary caregivers for elderly relatives. As noted last year, the investigators have effectively used Zoom, Microsoft Teams, and WebEx and manuscript productivity is consistent and in good quality journals. Assuming the pandemic eases, we anticipate attending meetings next year.

Changes that had a significant impact on expenditures

Nothing to Report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to Report.

Significant changes in use or care of human subjects

Nothing to Report.

Significant changes in use or care of vertebrate animals.

Nothing to Report.

Significant changes in use of biohazards and/or select agents

Nothing to Report.

6. Products

Publications, conference papers, and presentations

Journal publications.

- Katherine A. Owen[1#], Kristy A. Bell[1], Andrew Price[1], Prathyusha Bachali[1], Hannah Ainsworth[2], Miranda C. Marion[2], Timothy D. Howard[3], Carl D. Langefeld[2], Nan Shen[4], Jinoos Yazdany[5], Maria Dall'era[5], Amrie C. Grammer[1] and Peter E. Lipsky[1] Mendelian randomization id pathway analysis demonstrate shared genetic associations between systemic lupus erythematosus and coronary artery disease. Cell Reports Medicine, In press (online available November 4th).

- Marion MC, Ramos PS, Bachali P, Labonte AC, Zimmerman KD, Ainsworth HC, Heuer SE, Robl RD, Catalina MD, Kelly JA, Howard TD, Lipsky PE, Grammer AC, Langefeld CD. Nucleic Acid-Sensing and Interferon-Inducible Pathways Show Differential Methylation in MZ Twins Discordant for Lupus and Overexpression in Independent Lupus Samples: Implications for Pathogenic Mechanism and Drug Targeting. Genes (Basel). 2021 Nov 26;12(12):1898. PMCID: PMC8701117.

- Ghodke-Puranik Y, Jin Z, Zimmerman KD, Ainsworth HC, Fan W, Jensen MA, Dorschner JM, Vsetecka DM, Amin S, Makol A, Ernste F, Osborn T, Moder K, Chowdhary V, Langefeld CD, Niewold TB. Single-cell expression quantitative trait loci (eQTL) analysis of SLE-risk loci in lupus patient monocytes. Arthritis Res Ther. 2021 Nov 30;23(1):290. PMCID: PMC8630910.

- Genetic mapping across autoimmune diseases reveals shared associations and mechanisms. Matthew R Lincoln, Noah Connally, Pierre-Paul Axisa, Christiane Gasperi, Mitja Mitrovic, David van Heel, Cisca Wijmenga, Sebo Withoff, Iris H Jonkers, Leonid Padyukov, International Multiple Sclerosis Genetics Consortium, Stephen S Rich, Robert R Graham, Patrick M Gaffney, Carl D Langefeld, Timothy J Vyse, David A Hafler, Sung Chun, Shamil R Sunyaev, Chris Cotsapas. (Under second review, Nature Genetics)

Books or other non-periodical, one-time publications.

Nothing to Report.

Other publications, conference papers, and presentations.

Katherine Owen (presenter), Jessica Kain, Miranda Marion, Carl D. Langefeld, Amrie C. Grammer, and Peter Lipsky. "Assessing the genetic risk for atherosclerosis in SLE" Lupus 21st Century 2022, Tuscon, AZ, September 20-23, 2022. Invited talk.

Katherine A. Owen, Kristy A. Bell, Andrew Price, Prathyusha Bachali, Hannah Ainsworth, Miranda C. Marion, Timothy D. Howard, Carl D. Langefeld, Nan Shen, Jinoos Yazdany, Maria Dall'era, Amrie C. Grammer and Peter E. Lipsky. Molecular pathways identified from risk alleles demonstrate mechanistic differences in systemic lupus erythematosus patients of East Asian and European ancestry. Poster and Lightning talk. American College of Rheumatology (ACR) Convergence 2022, Philadelphia, PA, November 10-14, 2022.

Website(s) or other Internet site(s)

Nothing to Report.

Technologies or techniques

Nothing to report

Inventions, patent applications, and/or licenses

Nothing to Report.

Other Products

Other products produced by this work update and expand upon those in the first year and as well as new elements. Collectively this includes:

1. Database of ZINC 15 IDs matched to common database (e.g., pubCHEM) identifiers and common drug names.
2. Continued refinement of growing database of gene linked SLE risk SNPs by ancestry (e.g., linked to most plausible functional gene via gene expression or nonsynonymous variants).
3. Growing database of target-specific molecular docking for FDA-approved small molecules. Currently this dataset will feed into drug prioritization for SLE therapeutics, but since many of the identified drug targets are relevant in other autoimmune diseases, this could eventually be evaluated in the context of other diseases. For example, see submitted paper and pan-autoimmune disease genetics (Appendix).
4. Similarity measures using SMILES of FDA-approved drugs and list of drug in proximity to “hot spot” of enrichment.

7. Participants & Other Collaborating Organizations

What individuals have worked on the project?

Wake Forest investigative team

Name:	Carl D. Langefeld, Ph.D.
Project Role:	Primary Investigator of project
Researcher Identifier (e.g., ORCID ID):	0000-0002-4266-6949
Nearest person month worked:	1.2 months (10% of effort)
Contribution to Project:	Provided overall supervision of administrative and scientific components of the grant. Major contributor of manuscript preparation and presentations.
Funding Support:	Dr. Langefeld volunteered time from his institutional protected time to enable progress on the grant.

Name:	Timothy D. Howard, Ph.D.
Project Role:	Professor
Researcher Identifier (e.g., ORCID ID):	0000-0003-2518-4902
Nearest person month worked:	0.6 calendar months (5% effort)
Contribution to Project:	Dr. Howard has contributed to the characterization of function of the associated SNPs and has contributed to writing and editing of manuscripts.
Funding Support:	

Name:	Hannah C. Ainsworth, Ph.D.
Project Role:	Staff Bioinformatician, promoted to Assistant Professor (tenure track)

Researcher Identifier (e.g., ORCID ID):	0000-0003-1185-0695
Nearest person month worked:	0.6 calendar months (5% effort)
Contribution to Project:	Dr. Ainsworth developed the pipeline and completed the mapping of the cGenes to proteins. She has assisted in writing and editing manuscripts. She has contributed to the characterization of function of the associated SNPs. She has applied her dissertation work on DNA topology as a weighting mechanism to develop credible sets of high priority associated SNP.
Funding Support:	Partial support provided by Dr. Langefeld's personal research and development fund and institutional protected time for junior faculty to pursue the role of DNA topology on identifying plausibly functional variants that might impact gene function, hence targets for protein binding experiments.

Name:	Tony Reeves, Ph.D.
Project Role:	Associate Professor
Researcher Identifier (e.g., ORCID ID):	0000-0002-8209-6020
Nearest person month worked:	0.6 calendar months (5% effort)
Contribution to Project:	Dr. Reeves role on this grant is to complete the molecular docking of the FDA approved drugs to genes for consideration of drug repositioning scoring. He has completed nearly 50% of the cGenes. He also served as a mentor for undergraduate research intern who contributed to the molecular docking analyses.
Funding Support:	

Name:	Miranda C. Marion, MA
Project Role:	Senior Biostatistician
Researcher Identifier (e.g., ORCID ID):	0000-0003-4487-8010
Nearest person month worked:	2.4 calendar months (20% effort)
Contribution to Project:	Ms. Marion completed statistical association analyses to help identify the cGene set of SNPs. She is first author and major contributor to the manuscript reporting DNA methylation patterns in MZ twins discordant for SLE. She has assisted in writing and editing manuscripts. She has been a primary analyst for identifying SNPs associated with SLE across ancestries on the various published arrays.
Funding Support:	

Name:	Nolan Hamilton, BS

Project Role:	Bioinformatician
Researcher Identifier (e.g., ORCID ID):	NA
Nearest person month worked:	1.2 calendar months (10% effort) before leaving (4/1/22).
Contribution to Project:	Mr. Hamilton assisted with the preparation of drug derivatives from the ZINC 15 database as SMILES strings for all of the cGenes. He also searched the drug list we identified for evidence in the literature for links to other autoimmune diseases. Mr. Hamilton is now seeking his PhD in Bioinformatics at the University of North Carolina.
Funding Support:	

Name:	Ekaterina S. Khvatkova, MS
Project Role:	Biostatistician II
Researcher Identifier (e.g., ORCID ID):	NA
Nearest person month worked:	1.2 calendar months (10% effort) since joining the department 7/1/22.
Contribution to Project:	Ms. Khvatkova recently joined our team, July 2022. In addition to training, she has been leading the programming effort to map eQTLs to genes and check to see that the risk allele is associated with increase gene expression and elevated gene expression is associated with lupus risk.
Funding Support:	

Name:	Peter E. Lipsky MD
Project Role:	PI of RILITE subcontract
Researcher Identifier (e.g., ORCID ID):	0000-0002-9287-1676
Nearest person month worked:	4 calendar months (33% effort)
Contribution to Project:	Project oversight
Funding Support:	

Name:	Kate A. Owen, PhD
Project Role:	Data analyst
Researcher Identifier (e.g., ORCID ID):	0000-0002-7530-6130
Nearest person month worked:	4 calendar months (% effort)
Contribution to Project:	Data analyst working on the unique genetic basis of lupus on east Asians and detection of specific drug targets
Funding Support:	

Name:	Jessica Kane
Project Role:	Data analyst
Researcher Identifier (e.g., ORCID ID):	NA
Nearest person month worked:	4 calendar months (33% effort), departed 8/1/22.
Contribution to Project:	Data analyst working primarily on genetic basis of the diversity of lupus as manifested in specific ancestries and individual patient phenotypes. Played central role in the recent Mendelian Randomization paper. Ms. Kane is now attending Stanford University to earn her PhD. RILITE is now recruiting her replacement.
Funding Support:	

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Changes in effort from October 1, 2021 through September 30, 2022

Langefeld, CD

ACTIVE
Grant #: P30 CA012197
Grant Title: Comprehensive Cancer Center of Wake Forest University-Cancer Center Support Grant

Effort: 7.5%

Grant #: 1 R01 NS128425-01
Grant Title: Central And Peripheral STrOke inflammatioN with Exosomes (CAPSTONE)
Effort: 7.5%

Grant #: R01 NS093870
Grant Title: Race/Ethnicity, Hypertension and Prevention of VCID and Stroke after ICH (REACH ICH)
Effort: 5%

COMPLETED
Grant #: 5 P30 CA124723-02
Grant Title: North Carolina Diabetes Research Center
Effort: 9%

Grant #: 5 R21 DK122297-02
Grant Title: Molecular and Epigenetic Predictors of Treatment Response to Topical Steroids in Eosinophilic Esophagitis
Effort: 3%

Grant #: R01 NS093870
Grant Title: Recurrent Hemorrhagic Stroke in Minority Populations
Effort: 5%

Grant #: W81XWH-18-1-0050
Grant Title: Modulation of Epithelial Biomarkers of Breast Cancer Risk by a Community-based intervention for Diabetes Prevention
Effort: 2%

Howard, TD

ACTIVE
Grant #: R01 MD015395
Grant Title: Social Factors, Epigenomics, and Lupus in African American Women (SELA)
Effort: 6%

Grant #: R01 NS120557-01A1
Grant Title: Repeated Assessment of Survivors in ICH (ReASSES)
Effort: 5%

Grant #: 1R01 MD017006-01A1
Grant Title: Migrant and Multi-generational Immigrant Experiences: The effects of Stressors on Epigenetic, Behavioral, and Health-related Outcomes
Effort: 25%

Grant #: 1R01 NS128425-01
Grant Title: Central and Peripheral STrOke inflammation with Exosomes (CAPSTONE)
Effort: 5%

COMPLETED
Grant #: SISL 1612616
Grant Title: An Informal Learning Model of Genetic and Genomic Education for Adult, Bilingual Learners
Effort: 15%

Grant #: P50 AR070591-04
Grant Title: Translational Center of Molecular Profiling in Preclinical and Established Lupus (COMPEL)

Effort: 4%

Reeves, T

ACTIVE
Grant #: Internal Pilot
Grant Title: In Utero Imaging in NHPs
Effort: No salary support

COMPLETED
Grant #: National Science Foundation EIR CHE-2101221
Grant Title: Development of Novel PET Tracers
Effort: 5%

RILITE investigators Drs. Peter Lipsky and Kate Owens report no change in efforts.

What other organizations were involved as partners?

Nothing to Report.

8. Special Reporting Requirements

Collaborative Awards

Not applicable/Nothing to Report

Quad Charts

Not applicable/Nothing to Report

9. Appendices

- Katherine A. Owen[1#], Kristy A. Bell[1], Andrew Price[1], Prathyusha Bachali[1], Hannah Ainsworth[2], Miranda C. Marion[2], Timothy D. Howard[3], Carl D. Langefeld[2], Nan Shen[4], Jinoos Yazdany[5], Maria Dall'era[5], Amrie C. Grammer[1] and Peter E. Lipsky[1] Mendelian randomization id pathway analysis demonstrate shared genetic associations between systemic lupus erythematosus and coronary artery disease. Cell Reports Medicine, In press (online available November 4th).

- Marion MC, Ramos PS, Bachali P, Labonte AC, Zimmerman KD, Ainsworth HC, Heuer SE, Robl RD, Catalina MD, Kelly JA, Howard TD, Lipsky PE, Grammer AC, Langefeld CD. Nucleic Acid-Sensing and Interferon-Inducible Pathways Show Differential Methylation in MZ Twins Discordant for Lupus and Overexpression in Independent Lupus Samples: Implications for Pathogenic Mechanism and Drug Targeting. Genes (Basel). 2021 Nov 26;12(12):1898. PMCID: PMC8701117.

- Ghodke-Puranik Y, Jin Z, Zimmerman KD, Ainsworth HC, Fan W, Jensen MA, Dorschner JM, Vsetecka DM, Amin S, Makol A, Ernste F, Osborn T, Moder K, Chowdhary V, Langefeld CD, Niewold TB. Single-cell expression quantitative trait loci (eQTL) analysis of SLE-risk loci in lupus patient monocytes. Arthritis Res Ther. 2021 Nov 30;23(1):290. PMCID: PMC8630910.

- Genetic mapping across autoimmune diseases reveals shared associations and mechanisms. Matthew R Lincoln, Noah Connally, Pierre-Paul Axisa, Christiane Gasperi, Mitja Mitrovic, David van Heel, Cisca Wijmenga, Sebo Withoff, Iris H Jonkers, Leonid Padyukov, International Multiple Sclerosis Genetics Consortium, Stephen S Rich, Robert R Graham, Patrick M Gaffney, Carl D Langefeld, Timothy J Vyse, David A Hafler, Sung Chun, Shamil R Sunyaev, Chris Cotsapas. (Under second review, Nature Genetics)

Cell Reports Medicine

Mendelian Randomization and Pathway Analysis Demonstrate Shared Genetic Associations Between Systemic Lupus Erythematosus and Coronary Artery Disease

--Manuscript Draft--

Manuscript Number:	CR-MEDICINE-D-22-00062R3
Full Title:	Mendelian Randomization and Pathway Analysis Demonstrate Shared Genetic Associations Between Systemic Lupus Erythematosus and Coronary Artery Disease
Article Type:	Research Article
Keywords:	systemic lupus erythematosus; coronary artery disease; Mendelian Randomization; lupus; genetics; targeted therapeutics
Corresponding Author:	Katherine Anne Owen, Ph.D. AMPEL BioSolutions LLC Charlottesville, VA UNITED STATES
First Author:	Jessica Kain
Order of Authors:	Jessica Kain
	Katherine Anne Owen, Ph.D.
	Miranda Marion
	Carl Langefeld
	Amrie Grammer
	Peter Lipsky
Abstract:	Coronary artery disease (CAD) is a leading cause of death in patients with systemic lupus erythematosus (SLE). Despite clinical evidence supporting an association between SLE and CAD, pleiotropy-adjusted genetic association studies are limited and focus on only a few common risk loci. Here, we identify a net positive causal estimate of SLE-associated non-HLA SNPs on CAD by traditional Mendelian Randomization (MR) approaches. Pathway analysis using SNP-to-gene mapping followed by unsupervised clustering based on protein-protein interactions (PPI) identifies biological networks composed of positive and negative causal sets of genes. In addition, we confirm the casual effects of specific SNP-to-gene modules on CAD using only SNPs mapping to each PPI-defined functional gene set as instrumental variables. This PPI-based MR approach elucidates various molecular pathways with causal implications between SLE and CAD and identifies biologic pathways likely causative of both pathologies revealing known and novel therapeutic interventions for managing CAD in SLE.
Suggested Reviewers:	Laurence Morel, PhD Professor of Pathology, University of Florida morel@ufl.edu Expert in lupus pathogenesis and immunogenetics.
	Edward Wakeland, PhD Professor of Immunology, The University of Texas Southwestern Medical Center edward.wakeland@utsouthwestern.edu Expert in lupus genetics and autoimmune disease.
	Paula Ramos, PhD Professor of Medicine, Medical University of South Carolina ramosp@musc.edu Expert in the genetic etiology of autoimmune diseases and their ethnic disparities.
	Amr Sawalha, MD Professor and Chief, Division of Pediatric Rheumatology, University of Pittsburgh asawalha@pitt.edu

	Expert in the genetic and epigenetic contributions to the pathogenesis of systemic autoimmune and inflammatory diseases.
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Original Code<p>Does this manuscript report original code?	Yes
Standardized datasets<p>A list of datatypes considered standardized under Cell Press policy is available here. Does this manuscript report new standardized datasets?	No
Reviewers must have anonymous access to these original code that is free-of-cost. Please provide code location and instructions for access here.<p>Please consult this Author's guide for more information: "How standardized datasets and original code accompany Cell Press manuscripts from submission through publication" or email us at JOURNAL@cell.com. as follow-up to "Original Code<p>Does this manuscript report original code?"	Original code is now available on figshare (www.figshare.com) http://doi.org/10.6084/m9.figshare.21225251

Sha Yu, PhD
Scientific Editor
Cell Reports Medicine

September 29, 2022

Dear Dr. Yu,

We were delighted to hear that our paper originally entitled “Mendelian Randomization and Molecular Network Analysis Demonstrate Positive and Negative Genetic Causal Associations Between Systemic Lupus Erythematosus and Coronary Artery Disease (CR-MEDICINE-D-22-00062R2) was conditionally accepted for publication. We have now addressed all final reviewer comments as well as those related to the journal guidelines as detailed in our point-by-point response.

Sincerely,

Katherine Owen

AMPEL Biosolutions
250 West Main Street, STE 300
Charlottesville, VA 22902

Reviewer #3:

I would only recommend removing the word "causal" from the title and changing to "… Positive and Negative Shared Genetic Associations Between …"

- At the request of Reviewer 3, we have altered the title of the manuscript to “Mendelian Randomization and Pathway Analysis Demonstrate Shared Genetic Associations Between Systemic Lupus Erythematosus and Coronary Artery Disease.” The changes are located on the title page and are identified by red bold face.

Journal Guidelines:

The title should contain no more than 150 characters, including spaces. Please contact the handling editor if you want to discuss how to revise the title.

- The title has been shortened in accordance with journal guidelines.

Please carefully check STAR Methods as this part of the manuscript will not be copyedited.

- The STAR Methods have been double-checked for completeness and edited for content.

Cell Reports Medicine requires the inclusion of a labeled subsection at the end of Discussion called 'Limitations of the Study'. The inclusion of this of this subsection is not meant to replace discussion of limitations at the natural points in the paper. Rather, the goal of this section is to promote clarity and transparency by highlighting any limitations in the methodology and/or interpretation of the study. This could include limitations of the techniques or models used, assumptions made or limitations around the conclusions and future impact of the research. It can include additional experiments or analysis that would be necessary to definitively prove some conclusions but should be specific to your paper.

- This subsection has now been added to page 20.

Please ensure that there are no priority claims in the paper (including new, novel, unique, ‘the first’ and other such phrases) or in the title, abstract and highlights of the paper.

- All priority claims referring to the analyses described in the manuscript have been removed.

Please make sure that the summary is written in the present tense (currently, your findings are described in past tense). The abstract should not exceed 150 words.

- The summary has been edited to reflect the present tense and does not exceed 150 words.

Figure legends should include information on biological and technical replicates if applicable.

- The analyses performed in this manuscript did not require biological or technical replicates.

Please submit the Supplemental information as a single PDF file. The supplemental section should be organized so that the relevant figure legend is placed beneath each Supplementary figure. The text should be Times New Roman (10 point) with single (left) justification of all text in the Supplemental document.

- A PDF containing all supplemental figures and their legends are included.

Each Supplemental figure or table should be linked to at least one main-text figure or table, and/or to STAR Methods. This is indicated in the legend for the Supplemental figure; for example, "Figure S1. Flies lacking YFG do not exhibit changes in grooming behavior. Related to Figures 1 and 3." Every Supplemental figure must be cited at least once in the main text.

- Supplementary figures and tables are linked to a main-text figure and cited in the main text.

Thank you for including a graphical abstract. There is currently no highlights or eTOC summary. Please refer to our author guidelines for information on these items. Please also ensure that the highlights and eTOC summary should be written in present tense and submitted as a separate Word document.

- A selection of bulleted highlights and an eTOC blub are now included.

You indicated that "Original software code and documentation has been deposited on figshare (www.figshare.com) and is publicly available as of the date of publication". Please make sure the DOIs are listed in the Key Resources Table/Software and Algorithms section.

- We have now included the figshare DOI in the Data Code and Availability section and to the Key Resources Table.

We require that the final bullet point of the Data and Code Availability statement states: “Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.” Please revise the current statement.

- This statement has now been corrected.

Please include this subheading after "Data and Code Availability" section. The followings are the guideline for "EXPERIMENTAL MODEL AND SUBJECT DETAILS". If it is not applicable to this study, please make a brief explanation.

- This section has now been included and states that we did not use any experimental models or subjects.

Please include this subheading after "METHOD DETAILS" section (see some instructions below).

- A subheading for Quantification and Statistical Analysis has now been added.

Please describe all of the statistical analyses and software used in this section and indicate where all of the statistical details of experiments can be found (e.g., in the figure legends,

figures, Results, etc.), including the statistical tests used, exact value of n, what n represents (e.g., number of animals, number of cells, etc.), definition of center, and dispersion and precision measures (e.g., mean, median, SD, SEM, confidence intervals).

- The details of all computational and statistical analyses, as well as software used in the manuscript are included in this section.

In Quantification and Statistical Analysis, please state whether or not any methods were used to determine whether the data met assumptions of the statistical approach. If relevant, please detail the method used.

- Monte Carlo simulations were run to estimate FDRs. The details of these analyses are included.

On behalf of myself and all of the authors, we thank-you for your assistance throughout the review process. We hope the changes made will allow for full acceptance of the manuscript into *Cell Reports Medicine*. Please do not hesitate to contact me for further information.

Sincerely,

Katherine Owen

AMPEL Biosolutions
250 West Main Street, STE 300
Charlottesville, VA 22902

Mendelian Randomization and Pathway Analysis Demonstrate Shared Genetic Associations Between Systemic Lupus Erythematosus and Coronary Artery Disease

Jessica Kain[a], Katherine A. Owen[a1], Miranda C. Marion[b], Carl D. Langefeld[b], Amrie C. Grammer[a], Peter E. Lipsky[a,]

[a]AMPEL BioSolutions LLC, Charlottesville, VA and the RILITE Research Institute, Charlottesville, VA

[b] Department of Biostatistics and Data Science, and Center for Precision Medicine, Wake Forest University School of Medicine, NC

[1] Correspondence: kate.owen@ampelbiosolutions.com

SUMMARY

Coronary artery disease (CAD) is a leading cause of death in patients with systemic lupus erythematosus (SLE). Despite clinical evidence supporting an association between SLE and CAD, pleiotropy-adjusted genetic association studies are limited and focus on only a few common risk loci. Here, we identify a net positive causal estimate of SLE-associated non-HLA SNPs on CAD by traditional Mendelian Randomization (MR) approaches. Pathway analysis using SNP-to-gene mapping followed by unsupervised clustering based on protein-protein interactions (PPI) identifies biological networks composed of positive and negative causal sets of genes. In addition, we confirm the casual effects of specific SNP-to-gene modules on CAD using only SNPs mapping to each PPI-defined functional gene set as instrumental variables. This PPI-based MR approach elucidates various molecular pathways with causal implications between SLE and CAD and identifies biologic pathways likely causative of both pathologies revealing known and novel therapeutic interventions for managing CAD in SLE.

KEY WORDS

Systemic lupus erythematosus, coronary artery disease, mendelian randomization, lupus, genetics, SNPs, targeted therapeutics

INTRODUCTION

Systemic lupus erythematosus (SLE) is a female predominant, autoimmune disease characterized by immune dysregulation and multi-organ inflammation that is frequently associated with the development of cardiovascular disease (CVD)[1,2]. SLE exhibits hyperactivity of the innate and adaptive immune systems, increased production of numerous autoantibodies, and disturbed cytokine balance[3]. Although CVD is not a diagnostic criterion of SLE and was not included in the original descriptions of the disease, it is currently the main cause of death in SLE[4–6] with coronary artery disease (CAD) directly responsible for one-third to one-half of all CVD cases and 30% of deaths[7-9]. Notably, whereas mortality from infections and active disease have decreased in SLE patients, CVD-related death rates have not improved[10] and the standardized mortality ratio related to CVD has actually increased[11]. Women with SLE have a significantly increased risk of stroke and myocardial infarction along with elevated incidence of asymptomatic atherosclerosis compared to the general population[12,13]. Furthermore, traditional CVD risk factors, such as cholesterol, blood pressure, and smoking status fail to fully account for the overall higher risk of acute CVD events in SLE, although the underlying mechanisms remain unknown[14–17]. This lack of an understanding for the increased risk of CVD in SLE has resulted in limited treatment options and the puzzling juxtaposition that despite the efficacy of statins and ACEIs/ARBs in treating the general population, they appear to have little effect on CVD outcomes in SLE patients[5,18]. As a result, even though SLE has a prevalence of only about 70 per 100,000, it ranks among the leading causes of death in young women[1], despite the omission of lupus diagnoses in almost half of SLE patients' death certificates[19,20].

Genetic predisposition imposes important risk factors for both SLE and CVD[21–23]. To date, genetic association studies of SLE patients with and without CVD have been limited in size and have detected only a few common genetic risk loci, including *IRF8*, *STAT4*, *IL19*, and *SRP54-AS1*[22,24,25]. Mendelian Randomization (MR) is a causal inference method using genotypes as "treatments" when randomized controlled trials are not feasible. By measuring and correlating the effect sizes of exposure-associated genetic variants in large-scale genetic association studies on traits of interest, a causal effect of the exposure on the outcome can be estimated. Here, we report the application of multiple, complementary MR methods to identify causal paths from SLE-associated variants to CAD using summary statistics from genetic association studies. Using multiple MR algorithms, we have identified large sets of SLE causal variants that also impart genetic risk for CAD, as well as those that appear to diminish the risk of CAD. Using innovative approaches to build molecular pathways from genetic risk factors[26], we have developed a map of SLE-derived biologic processes with causal implications on CAD that may account for the genetic basis of the association between these two apparently dissimilar clinical entities and may also provide insights into the shared mechanisms underlying each. Understanding the pathogenesis of genetic variants underlying the increased CAD risk in SLE can ultimately provide insight into the immune and inflammatory components of atherosclerosis, as well as reveal opportunities for targeted therapeutics.

RESULTS

Pathway analysis reveals gene networks implicated by genetic variants associated with both SLE & CAD

To explore the shared genetic predispositions for SLE and CAD, we first identified single nucleotide polymorphisms (SNPs) associated with each trait in 5 SLE and 1 CAD multi-ancestral genetic association study[27–32]. In total, 96 SNPs were associated with both conditions (Figure S1A). Notably, the majority of the overlapping SNPs mapped to the HLA region of chromosome 6. To identify putative gene(s) influenced by each of the 96 SNPs associated with both SLE and CAD, we mapped causal SNPs to genes[26], identifying 189 unique genes encoding 135 proteins in STRINGdb (Figure S1B). Stratified linkage disequilibrium score regression (S-LDSC) was then used to validate the biological relevance of SNP-predicted gene and protein sets by assessing whether they captured more disease heritability than expected by chance with respect to all genes and STRINGdb proteins, respectively[33]. Application of S-LDSC using GWAS summary statistics for SLE (GCST003155[27]), CVD (GCST004280[34]) and two CAD datasets (CAD-I, GCST000998[35] and CAD-II, GCST001479[36]) determined that the 189 genes predicted by the 96 overlapping SNPs were significantly ($p<0.05$) enriched for genomic regions capturing the genetic heritability of CAD and CVD (Figure S1C). Nearly identical results were also obtained using the smaller subset of 135 protein-coding genes (Figure S1C). However, application of standard LDSC which was restricted to the use of the relatively small, European-only SLE GWAS[27], did not reveal a significant level of genetic correlation between the diseases (Figure S1D).

To assess molecular networks encoded by the set of 135 protein encoding genes predicted from the overlapping SLE/CAD SNPs, a protein-protein interaction (PPI)

network was generated and unsupervised clustering revealed 12 distinct gene clusters that were functionally enriched in a diverse range of immunological and cellular categories (Figure S1E) many with relevance to SLE and CAD/CVD, including cluster 1 characterized by canonical pathways for *Antigen presentation pathway* and *B cell development*, along with *Sudden cardiac death and* cluster 3 enriched in *Atherosclerosis signaling* and *Lupus erythematosus, systemic*, amongst others (Table S1). Although the molecular pathways associated with SNP-predicted genes suggested a convergence of biological processes underlying SLE and CAD, it remained uncertain whether the finding of overlapping SNPs implied shared genetic causation. The subsequent studies presented here explore this in detail.

MR estimates a positive correlation between effects of SLE-associated non-HLA variants on SLE and CAD

MR methods were employed to estimate the association between effect sizes of relevant variants on SLE and CAD. We first applied six MR methods using various sets of SLE-associated instrumental variables (IVs) to determine whether they tend to confer shared (positive) effects on SLE and CAD, noting that this initial approach did not satisfy all assumptions for IV-validity or IV-independence and therefore could only provide an estimated association (Figure 1A). Initial exploratory analyses employing IVs derived from the Immunochip and GWAS studies suggested a net-positive association for non-HLA SLE-associated SNPs on CAD (Figure 1B-C). Even when using SLE IVs determined with the more stringent significance level and removal of known pleiotropic associations to CVD or confounders such as cholesterol, obesity, blood pressure, insulin resistance and

smoking, the indication of a positive causal relationship between SLE and CAD remained (Figure 1C, bottom row).

To validate the robustness of our estimated associations by satisfying stringent requirement for IV selection, we carried out two-sample MR analyses using multi-ancestral, non-HLA SNPs strongly associated ($p<5x10^{-8}$) with SLE, excluding SNPs weakly associated ($p<10^{-5}$) with CVD or confounders (Table S2), followed by stringent LD-clumping to ensure IV-independence[37] (R^2=0.001, 100kb window, 1000G EA reference population) (Figure 2A). SLE GWAS summary statistics were used for exposure[27], and multiple CAD GWAS were used for outcome (GCST005194[32], CAD-a and GCST005195[32], CAD-b). Since CAD is causative of myocardial infarction (MI) and atherosclerosis is common to CAD and ischemic stroke (IS), MR was carried out using summary statistics for 2 additional MI GWAS (GCST003117[38], MI-a and GCST011365[39], MI-b) and IS (GCST006906[40]). Summary statistics for cardiomyopathy from the FinnGen biobank analysis (finn-b-I9_CARDMYO, CM) and atrial fibrillation (GCST006414[41], AFib), which are not associated with atherosclerosis or CAD, were also included for comparison. After LD-clumping, 60 independent SNPs were included in the SLE GWAS and then harmonized with each outcome-GWAS pair before use as IVs for SLE exposure (Table S9). Between 43-56 harmonized IVs for SLE-exposure were then tested using 16 MR methods, some of which account for additional IV-invalidity, pleiotropy, or heterogeneity, to estimate causal relationships with the various atherosclerotic and cardiac conditions. The majority of MR methods resulted in significant ($p<0.05$) positive causal estimates of SLE-associated variants on CAD-a, and both MI GWAS, but not for cardiomyopathy or Afib (Figure 2B, Figure S2A, Table S3). Directional pleiotropy was only detected between

the SLE and CAD-b GWAS by the MR-Egger intercept test (Figure S2A), indicating potential bias in the causal estimates based upon effect sizes using these summary statistics. Overall, IVW, weighted median, penalized weighted median, maximum likelihood, RAPS and PRESSO were significant in 4 out of 5 outcome GWAS (Figure 2B). These results establish a positive causal effect of SLE on CAD and suggest that the increased CVD-risk associated with SLE is likely to involve atherosclerosis rather than other aspects of cardiac pathology.

To eliminate the possibility that the positive causal estimate of SLE on CAD is bidirectional and therefore unlikely to represent a true causal relationship, MR was also carried out in the reverse direction, with CAD or MI as exposure and SLE as the outcome. Importantly, none of the 14 methods yielded a significant positive causal estimate of CAD or MI on SLE. Of interest, however, significant ($p<0.05$) negative causal estimates of CAD and MI on SLE were observed for approximately half of the 14 MR methods tested (Figure S2B, Table S3).

To understand the pathways underlying the positive causal estimates of SLE on CAD in greater detail, all SLE-associated SNPs included as putative IVs before harmonization with each GWAS were mapped to genes. Consistent with satisfying the exclusion restriction criteria and independence assumption with respect to traits imposing significant CVD-risk, S-LDSC results demonstrated that the 284 genes and 160 predicted proteins captured a significant portion of SLE heritability (p-values = 3.46×10^{-5} and 3×10^{-5}, respectively), but not that of CVD or CAD (Figure 2C).

Proteins predicted from the SLE IVs were then integrated into connectivity networks in STRINGdb (Figure 2D). Cluster annotations were dominated by processes

commonly dysregulated in SLE as expected, including canonical pathways for *Systemic lupus erythematosus in B cell signaling*, *Th1 pathway* and *Th2 pathway*, as well as GO terms for *Regulation of immune response* (GO:0050776) and *Negative regulation of B cell activation* (GO:0050869) (Table S4). Interestingly, disease associations were enriched in various autoimmune diseases (*Lupus erythematosus, systemic*, *Aicardi goutiere's syndrome* and *Hashimoto's disease)*, along with cardiovascular dysfunction, such as *Arterial embolism and thrombosis*, *Hypertension*, *Plaque, atherosclerotic* and *Ischemic heart disease* (Table S4).

Single-SNP MR identifies gene networks implicated by SLE-associated variants with positive and negative causal estimates on CAD

We next employed single-SNP MR (SSMR) to identify specific SLE-associated variants with positive or negative estimates on CAD. SSMR applied to SLE-associated SNPs, including those in the HLA region, reveal that the majority of negative causal SNPs are located on the short arm of chromosome 6; all but one were tightly packed around the HLA region, spanning chr6:28014374-33683352 (Figure S3). When excluding the short-arm of chromosome 6 and SNPs associated with CVD or confounders, SSMR identified 80 and 96 SLE-associated variants with significant ($p<0.05$) positive (Figure 3A, top 25) and negative (Figure 3B, top 25) causal estimates on CAD, respectively (Figure S4, Table S3). The majority of positive-causal SNPs were distributed on chromosomes 1, 2 and 4, whereas over 50% of negative causal SNPs were on chromosomes 7, 11 and 17 (Figure 3C, E).

Non-HLA SLE variants with either significant positive or negative causal estimates on CAD were separately mapped to 236 (Figure 3D) and 244 predicted genes (Figure 3F), respectively, for clustering and pathway analysis. Positive SNP-predicted gene clusters were enriched in the canonical pathway for *Antigen presentation* and functional categories for MHC class I, as well as epigenetic processes, transcription, and endocytosis (Figure 3E,Table S5), whereas negative SNP-predicted gene clusters were dominated by processes related to cell death (*Pyroptosis* (GO:0070269), cluster 2; *Regulation of oxidative stress-induced cell death* (GO:1903201), cluster 6) and protein degradation (Proteasome, cluster 6 and Autophagy, cluster 8; Figure 3F, Table S5). Finally, gene sets predicted by both positive and negative causal SNPs captured a significant portion of SLE heritability, but not that of CAD or CVD, consistent with their selection as IVs for SLE (Figure 3G-H).

Pathway analysis of HLA region variants associated with SLE-risk and protective of CAD

Risk haplotypes in the HLA region heavily contribute to susceptibility for SLE[42] and CAD[43]. However, accurate genotyping of HLA alleles and corresponding GWAS effect size estimates are notoriously unreliable[44]. Additionally, the complex genetic architecture of this region makes mapping HLA variants to genes especially challenging given the extensive LD and high density of genes in this region. Nonetheless, an examination of the HLA area (chr6:28.5–33.5 Mb) revealed 30 SNPs significantly ($p<10^{-6}$) associated with both SLE and CAD in their respective GWAS. While these SNPs are not independently associated variants, all 30 SNPs had positive effect sizes for SLE but were

negative for CAD (Figure S5A), possibly reflecting the extensive LD in this region. Connectivity mapping and clustering of the 69 protein-encoding genes predicted from these 30 SNPs revealed 6 distinct clusters dominated by processes dysregulated in SLE, including the functional categories for MHC class I and MHC class II in clusters 1 and 6, along with canonical pathways for *TH1 and TH2 activation*, *B cell development, Notch signaling* as well as gene ontogeny (GO) terms for *Interferon-gamma mediated signaling pathway* (GO:0060334) (Figure S5B-C). Other pathways of interest involving *Complement system abnormalities, LXR/RXR activation, and 21-hydroxylase deficiency* were predicted by cluster 3 (Figure S5C).

PPI-based MR predicts specific sets of SLE-associated variants and gene pathways causal of CAD

To obtain a more comprehensive view of the possible impact of SLE-derived molecular pathways on atherosclerosis, we mapped SLE-associated, non-HLA Immunochip SNPs with net positive causal estimates on CAD by MR to genes and pathways regardless of their associations with CVD-related traits. In total, 838 SNPs predicted 2,336 putative genes and 1,501 proteins that collectively captured a significant amount of SLE, but not CAD or CVD, heritability (Figure 4A); these 1,501 proteins clustered into 46 distinct clusters based on PPI connectivity (Figure 4B). We then grouped SLE-associated SNPs mapping to genes in each of the 46 PPI-based clusters for use as SLE-IVs to estimate cluster-specific associations with atherosclerotic traits. Initial application of MR-IVW to these 46 subsets of SLE SNP-derived IVs yielded 16 and 9 significant ($p<0.05$) positive and negative causal estimates, respectively (Figure 4B-C, Figure S6A) for CAD.

Additional MR methods, including mode and median based methods, MR-Egger, MR-RAPS, MR-PRESSO, and maximum likelihood, were carried out for further validation of the PPI-based MR-IVW causal estimates (Table S3). Clusters were grouped into tiers with respect to consistency across the various MR methods, with tier 1 clusters yielding significant positive or negative causal estimates by almost all (at least 14/16) MR-methods and tier 2 clusters yielding significant positive or negative causal estimates by MR-IVW or at least 7 MR-methods (Figure 4B-C). Finally, when examined individually, 20 of the 46 clusters specifically captured SLE heritability by PPI-based S-LDSC, many with significant causal estimates on CAD by PPI-based MR (Figure 4D and Table S6).

In an effort to support these results by expanding the size of the network, we added 914 multi-ancestral, non-HLA SNPs associated with SLE on the Phenoscanner database to the analysis. Overall, 1,708 unique SNPs predicted 3,272 putative genes and 1,972 proteins that collectively captured a significant amount of SLE heritability, but not that of CAD or CVD (Figure 5A) and clustered into 67 distinct sets of protein-coding genes (Figure 5B). PPI-based MR-IVW using these 67 clusters of SLE SNPs as IVs yielded 24 and 11 significant ($p<0.05$) positive and negative causal estimates on CAD, respectively (Figure 5C-D, Figure S6B), many of which captured SLE heritability, but not that of CAD or CVD by PPI-based S-LDSC (Figure 5E and Table S6).

To ensure that the majority of predicted causal clusters are not a result of random chance or multiple-hypothesis testing, simulations were carried out to estimate the false discovery rate. Results from these simulations, which account for both LD and pleiotropy, indicate that SLE-derived and PPI-clustered modules, as opposed to randomly generated SNP-to-gene modules, demonstrate a higher rate of significant causal estimates on CAD

(Figure S7). Furthermore, to assess the reproducibility of the cluster-specific causal estimates, PPI-based MR was repeated using CVD-related GWAS datasets on the MR-base platform[45]. The PPI-based MR-IVW causal estimates were highly consistent using summary statistics from 2 CAD and 2 MI GWAS on MR-base, but not cardiomyopathy or AFib (Table S7), suggesting that the stratified causal estimates on CAD are associated with the atherosclerotic component of CVD. Together, these results support the conclusion that the PPI-based MR results are atherosclerosis-specific and unlikely trivial results of random chance or multiple hypothesis testing.

SLE-derived clusters in all positive and negative causal tiers were annotated using multiple functional and cellular composition tools (Figure 5B, Table S8). These results show that a wide range of SNP-predicted biological functions known to be involved in SLE pathogenesis have causal implications on CAD by MR, such as *Neutrophil degranulation* (clusters 2 and 43), *Th1 and Th2 activation*, and/or *Th17 activation* (clusters 3, 5, 8, and 9), *Interferon signaling* (cluster 8), *Leukocyte extravasation signaling* (cluster 12), *Leukocyte trans-endothelial migration* (cluster 28) and *Leukocyte adhesion to endothelial cells* (cluster 2). In addition to immune-related pathways, many of these positive causal clusters were enriched in disease phenotypes associated with cardiovascular disease, including *Th1 cell activation and proliferation in atherosclerosis* (cluster 9) and *Lipid and atherosclerosis* (cluster 12). Interestingly, several clusters were enriched in autonomic nervous system control related to cardiac function (*Cardiac muscle contraction*, clusters 13 and 33) or *Neuroinflammation* (cluster 60) (Figure 5B, Table S8).

In contrast, SLE-derived clusters with negative causal estimates on CAD were enriched for oxidative stress (cluster 10), nitric oxide (clusters 24, 40, and 64), and HDL

cholesterol (clusters 24 and 50) (Figure 5B, Table S8). Pathway enrichment was further reflected in assigned functional categories, with ROS protection (clusters 24 and 45), NR transcription (cluster 40) and ubiquitylation and SUMOylation (cluster 64) dominating clusters with protective estimates on CAD (Figure 5B). These results are highly consistent with the enrichments associated with negative causal variants in our single-SNP MR and HLA-specific pathway analyses.

PPI-based MR stratifies SNPs, genes, and networks underlying the positive and negative causal effects of SLE on CAD

To further validate the causal effects of the 67 SNP-to-gene modules identified by PPI-based MR (Figure 6A), we carried out additional MR analyses with respect to PPI-based MR cluster-groupings after accounting for pleiotropy and LD. Causal estimates of SLE on CAD with IVs derived from clusters meeting the tier 1 or the tier 1 and 2 criteria, as well as those that surpassed the MR-IVW p-value < 0.00075 threshold were universally more positive, significant, and consistent than those based upon all SNPs (Figure 6B-C, Table S3). Similarly, negative causal estimates for SLE on CAD were obtained using IVs meeting the negative tier 1 and MR-IVW p-value < 0.00075 thresholds from the 67-cluster network. In contrast, IVs derived from clusters with insignificant causal estimates generally failed to reach significance in either direction. While these trends were observed using summary statistics from both the SLE GWAS and SLE Immunochip, causal estimates were more significant using the SLE Immunochip, consistent with its larger sample size. Importantly, these results demonstrate that PPI-based MR can be used to

identify independent IVs satisfying MR assumptions that underly both positive and negative causal effects of SLE on CAD.

Pathway analysis facilitates drug prediction

Pathways associated with positive causal clusters were used to facilitate identification of new therapeutic interventions for managing the unique inflammatory environment contributing to CAD in SLE (Figure 7A). Canonical pathways related to immune function in clusters 2, 3, 5 and 8 predicted drugs targeting T and B cells and inflammatory cytokines, including daratumumab (CD38), belimumab (TNFSF13), elotuzumab (SLAMF7), abatacept (CD80/86), iberdimide (IKZF1/IKZF3) and sarliumab (IL6R). Broader analysis of pathway categories also suggested the utility of targeting interferon signaling with anifrolumab (cluster 8), as well as anti-platelet/coagulant therapy to combat dyslipidemia (cluster 5)[46]. Additional noteworthy targets include PCSK9 (cluster 5), a protease involved in the degradation and recycling of the LDL receptor targeted by alirocumab and evolocumab, and oxidated LDL molecules (cluster 5) targeted by orticumab (Figure 7B).

DISCUSSION

Although genetic association studies have been successful in mapping disease loci in both immune and cardiovascular diseases, the genetic and molecular basis for the increased CAD predisposition in SLE patients has remained largely unexplained. Considering the limited data on CAD in SLE, we developed an approach that utilized GWAS summary statistics for both diseases to identify and interpret various sets of SLE-associated variants with causal implications on CAD. New findings suggest the causal

relationship with SLE appears to be focused on the atherosclerotic process, evidenced by positive estimates with CAD, MI and ischemic stroke, but not other cardiac conditions, such as cardiomyopathy or AFib. Furthermore, we developed and carried out PPI-based MR approach to identify specific sets of SLE variants mapping to biologically relevant gene sets with causal implications on CAD. By coupling various MR methods with network modeling and variant interpretation, we not only provided substantial evidence of shared genetic risk but also identified the putative molecular pathways involved in the development of CAD in SLE. Moreover, a number of the immune and inflammatory pathways identified in these analyses could well contribute to the pathogenesis of CAD even in the absence of SLE or other recognized autoimmune conditions. This points to the larger implication that CAD itself is a heterogeneous condition and subpopulations, such as those driven by SLE-associated processes, might require potentially distinct treatment strategies, at least partially motivated by unique genetic predispositions.

Causal inference using traditional MR methods rely on strict assumptions for independent IVs, however given the extensive pleiotropy underlying complex traits such as SLE and CVD, efforts to satisfy these assumptions can result in biasing the analyses by excluding previously established associations. Furthermore, the exclusion of SNPs associated with CVD-related traits results in the loss of relevant molecular information. While the use of SLE IVs that are also associated with CVD or confounders in traditional MR disqualifies the causal estimates from representing an effect on CAD directly through SLE, these SNPs can be just as important with respect to understanding the relevant biological pathways underlying CAD in SLE. Similarly, stringent LD-clumping to obtain an independent set of IVs not only reduces the statistical power of MR[47], but also can omit

additional SNPs, genes, and pathways underlying CAD in SLE. Due to our rigorous efforts to satisfy the assumptions and account for LD in the traditional MR analyses, while also employing numerous MR methods that account for IV-invalidity, pleiotropy, or heterogeneity, these results may give overly-conservative estimates of the causal effects and underlying mechanisms as a result of over-pruning.

To overcome these limitations of traditional MR, we developed and employed a PPI-based MR approach using networks comprehensively derived from large sets of SLE-associated SNPs, regardless of their associations with CVD-related traits. By generating cluster-specific associations between effect sizes on SLE and CAD, biologically relevant SNP-to-gene modules can be categorized as having shared (positive estimates) or opposing (negative estimates) effects on SLE and CAD. Traditional MR using independent, SLE-specific IVs mapping to positive and negative clusters, separately, confirmed that the groups of causal clusters are representative of positive and negative causal effects on CAD through SLE, respectively. We believe that our PPI-based MR approach is particularly beneficial in cases when the exposure is complex and heterogeneous, such as SLE which embodies a diverse range of molecular and pathophysiological mechanisms that we expect to impose unique casual effects on CAD.

Genetic variants are typically mapped to genes with respect to genomic location, identifying genes containing and/or nearby the SNPs of interest. Additionally, more recent advances have given rise to identification of trans-acting genomic regions that can epigenetically and/or transcriptionally influence genes at distant locations. This is especially important for complex, polygenic traits, such as SLE and CAD, of which most associated variants are non-coding. Here, we link SNPs to genes via amino acid changes

in encoded proteins, proximity, expression quantitative trait loci (eQTL) predictions, and regulatory elements in an effort to be as comprehensive as possible. Our subsequent PPI-based clustering elucidated a broad range of biologically relevant molecular networks within the diverse set of implicated genes and importantly served to filter out noise. Furthermore, our PPI-based MR approach served to highlight SNP-to-gene modules contributing most to the causal effects of SLE on CAD. Together, these results demonstrate how SLE genetics can be used to identify both known and novel loci and pathways with causal implications on CAD.

Numerous biologically relevant SNP-to-gene modules were determined to have positive causal effects on CAD through SLE by MR, spanning inflammatory factors, adaptive and innate immunity, intracellular signaling, cell differentiation, microRNA and mRNA processing, mitochondrial function, and more. A wide range of enrichments amongst positive causal clusters have been hypothesized and/or demonstrated to contribute to CVD in SLE patients, including glucocorticoids, neutrophil cell death (NETosis) and degranulation, TNF-like weak inducer of apoptosis (TWEAK) signaling, canonical and alternative complement pathways, Th1 differentiation, lipid and lipoprotein metabolism among others.

Considering the drastically increased prevalence and mortality of CAD in SLE, the considerable portion of SLE-associated risk variants with negative causal effects on CAD was unexpected and suggested that numerous variants contributing to SLE have atheroprotective effects. Further SNP-to-gene mapping and detailed pathway analyses revealed that these variants are involved in various processes, predominantly related to oxidative stress and cholesterol homeostasis, whose atheroprotective effects have been

found to be impaired in certain disease-related contexts, such as SLE. For example, the enzyme responsible for maintaining cholesterol homeostasis though lipoprotein lipase synthesis, cholesterol 27-hydroxylase, has been shown to be decreased in human monocytes and aortic endothelial cells of SLE patients, and is thought to impair the protective mechanism of efflux of cellular cholesterol[48]. Cyp27a1 is the gene that encodes the cholesterol 27-hydroxylase and is an LXR target activated by oxysterols as well as a target of RXR and PPAR in human macrophages[49]. LXR activation has additional proatherogenic and atheroprotective effects, as LXR activation in the liver promotes atherosclerosis via excess lipogenesis, whereas LXR activation in macrophages and dendritic cells has anti-inflammatory effects, linking lipid metabolism, immune cell function, and inflammation[50].

Our approach also has the advantage of identifying “actionable” points of therapeutic intervention with the potential to impact the inflammatory environment associated with CAD in SLE. This is especially important given that CAD risk in SLE cannot be fully accounted for by the increased prevalence of traditional atherosclerotic risk factors. SLE subjects therefore may derive particular benefit from treatments that mitigate inflammatory intermediates such as type I interferons with anifrolumab. Our findings also highlight additional putative targets, including PCSK9 involved in LDL receptor recycling. Inhibitors of PCSK9 activity, such as alirocumab and evolocumab are FDA approved to treat hyperlipidemia and may prove to be effective in controlling atherosclerosis in chronic inflammatory conditions[51]. Finally, recent reports also support targeting oxidized LDL molecules (anti-oxLDL, orticumab) for the prevention of cardiovascular events in SLE[52].

Limitations of the Study

Limitations of this study include those related to the data integrated in our pipeline. First, SLE genetic association studies have been restricted in size and scope, yielding limited power and genomic coverage, especially considering the extensive heterogeneity and polygenicity of lupus. To maximize both power and scope, we used the largest genetic association study for SLE, which is limited to Immunochip SNPs, the largest SLE GWAS, as well as SLE-associated SNPs pooled from the Phenoscanner platform. However, most genetic association studies, including the multi-ancestral data used in this study, are heavily biased towards European ancestries. This is especially problematic given the increased CVD morbidity and mortality in SLE patients of African-ancestry[53] in addition to the ancestry-dependent disparities observed in both SLE and CAD. It is also of note, that certain risk factors leading to distinct phenotypic outcomes such as CAD are likely to be impacted by environmental factors that cannot be accounted for by genetics alone. This is important with respect to the higher disease burden observed in African ancestry patients, where barriers to treatment (such as delayed diagnosis and/or limited access to a specialist) may contribute to elevated mortality in this population and further underscores the importance of generating large datasets with diverse patient populations. In addition, the ability to map genetic variants to implicated genes is limited to known SNP-to-gene relationships included in Ensembl's variant effect predictor (VEP), Genotype-Tissue Expression (GTEx), and Human ACtive Enhancer to interpret Regulatory variants (HACER) databases. Although putative causal pathways associated with the HLA region are intriguing, mapping of the SNPs within the HLA region to genes

is challenging because of the extensive LD across the region. Additionally, genes included in our PPI networks and clusters are limited to protein-coding genes and interactions included in STRINGdb. This is a potential shortcoming of our pipeline especially considering the large number of non-coding genes implicated in our SNP-to-gene predictions in addition to the growing evidence highlighting the contributions of non-coding long RNAs and microRNAs in both SLE and CAD[54,55]. Similarly, the ability to annotate gene clusters functionally is limited and potentially biased by the data underlying the numerous enrichment platforms used in our pathway analyses. IPA, EnrichR, which pools a myriad of public databases, and cell and functional analytic tools were all utilized to obtain orthogonal and reproducible annotations. Ultimately, however, our robust SNP-to-gene mapping approach, which included multiple sources of information in combination with biologically informed clustering employing numerous sources of annotation, enabled comprehensive analysis of both small and large sets of genetic variants to specific pathways with excellent reproducibility.

In summary, we have employed various approaches to clearly identify shared genetic risk factors for SLE and CAD. These results have provided new information about common molecular pathways in SLE and CAD, as well as the genetic and molecular information to consider novel therapeutic interventions in these conditions.

ACKNOWLEDGMENTS

The work presented in this manuscript was funded by a grant awarded to P.E.L. and A.C.G. of the RILITE Research Institute by the John and Marcia Goldman Foundation (jmgoldmanfoundation.org). The funder provided support in the form of salaries for

authors [J.K., K.A.O]. Additional support to all authors was provided by the Department of the Army (W81XWH-20-1-0686) to C.D.L. The authors would like to thank members of the AMPEL/RILITE scientific team for their critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, P.E.L., A.C.G and K.A.O.; Methodology, P.E.L. and J.K.; Software, Formal Analysis and Investigation, J.K.; Validation, C.D.L. and M.M.; Writing – Original Draft, J.K., P.E.L, and K.A.O.; Writing – Review & Editing, C.D.L., M.M., P.E.L. and K.A.O.; Visualization, J.K. and K.A.O.; Funding Acquisition, P.E.L., A.C.G. and C.D.L.; Supervision, P.E.L. and A.C.G.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

MAIN FIGURE LEGENDS

Figure 1. MR demonstrates a positive association of effect sizes of SLE-associated non-HLA SNPs on SLE and CAD. A) Graphical depiction of the 2-stage approach for an initial exploratory analysis using expanded groups of SNPs as IVs followed by a confirmatory analysis using highly curated IVs. **B-C**) Forest plots of 6 MR causal estimates (beta ± standard error). For results, grey indicates insignificant ($p > 0.05$), red, positive causal estimates determined by each MR method. **B**) Immunochip-derived SLE-associated non-HLA SNPs were used as IVs for SLE; summary statistics from both the SLE Immunochip study (left panel) and SLE GWAS (right panel) were used for the

exposure; summary statistics from the CAD GWAS were used for the outcome. **C**) Additional MR analyses using SNPs associated ($p<10^{-6}$) with SLE in the Immunochip and GWAS study (row 1 and 2) or Phenoscanner reaching genome-wide significance ($p <5x10^{-8}$, row 3) were used as IVs; summary statistics for the exposure and outcome are indicated. MR analyses excluding the entire short-arm of chromosome 6 and excluding only the extended HLA region (chr6:27-34Mb, left columns). Right columns show MR analyses using the same sets of SNPs excluding pleiotropic SNPs associated ($p<10^{-5}$) with either CAD- directly or CVD-related confounders, included on the Phenoscanner platform.

Figure 2. MR demonstrates a net positive-causal effect of SLE-associated non-HLA SNPs on CAD. A) MR diagram for testing the causal effects of SLE on CAD with respect to instrument relevance to the exposure, exclusion from the outcomes (i.e. CAD, MI, IS) and independence from confounding factors. LD-clumping ($R^2<0.001$) was used to obtain independent IVs. **B**) Forest plots of MR causal estimates (beta ± standard error) for SLE on CAD (CAD-a, CAD-b), MI (MI-a, MI-b), IS, cardiomyopathy (CM) and atrial fibrillation (AFib) GWAS using 16 MR methods. Missing PRESSO-OC estimates indicate insignificant global tests for horizontal pleiotropy. For results, grey indicates insignificant ($p> 0.05$), red, positive causal ($p<0.05$), and blue, negative causal ($p<0.05$) estimates determined by each MR method. Numbers within forest plots indicate the SNPs used as IVs after harmonization. **C**) Application of S-LDSC using summary statistics for SLE, CVD and CAD GWAS to estimate the heritability (coefficient ± standard error) of the 284 SNP-predicted genes (top panel) and 160 SNP-predicted proteins from STRINGdb (lower panel). Bar color indicates coefficient significance. **D**) Cluster metastructures for the 160

putative protein-coding genes are based on PPI networks. Functional and cell-type enrichments for each cluster were determined using BIG-C (black labels) and I-scope (red labels), respectively. Black labels over colored shadings represent shared functional annotations for the clusters they surround.

Figure 3. Analysis of SLE-associated SNP-predicted genes with causal effects on CAD by single-SNP MR. A-B) Forest plots (beta ± standard error) of the top 25 (by absolute value of causal estimates) positive (**A**) and negative (**B**) causal non-HLA SNPs identified by single-SNP MR (SSMR) using the Wald ratio method. **C** and **E**) Pie charts illustrating the chromosomal distribution of 80 positive (**C**) and 96 negative (**E**) causal SLE SNPs on CAD. **D** and **F**) Cluster metastructures for the 200 (**D**) 184 (**F**) predicted genes from positive and negative causal SNPs identified by single-SNP MR. Functional and cell-type enrichments for each cluster were determined using BIG-C (black labels) and I-scope (red labels), respectively. Bold black labels over colored shadings represent shared functional annotations for the clusters they surround. (**G-H**) S-LDSC using summary statistics for SLE, CVD and CAD GWAS to estimate the heritability (coefficient ± standard error) of genes (open bars) and SNP-predicted proteins (hashed bars) predicted by positive (**G**) and negative (**H**) causal SNPs determined by SSMR. Bar color indicates coefficient significance.

Figure 4. SLE-derived gene network with causal implications on CAD and PPI-based MR. A) S-LDSC using summary statistics for SLE, CVD and CAD GWAS to estimate the heritability (coefficient ± standard error) of the 2,336 genes (open bars) and 1,501 proteins (hashed bars) predicted by 838 Immunochip SNPs associated with SLE.

Bar color indicates coefficient significance. (**B**) Functional and cell-type enrichments for cluster metastructures were determined using BIG-C (black labels) and I-scope (red labels), respectively. Bold black labels over colored shadings represent shared BIG-C functional annotations for the clusters they surround. Node size is proportional to the number of SNPs (height) mapping to the genes in each cluster (width). For node color, red and blue indicate significant positive or negative estimates, respectively for 14/16 MR methods used (tier 1); light red and light blue, significant positive or negative estimates by MR-IVW or at least 7/16 MR-methods (tier 2); grey, insignificant. Thickness of the yellow border is roughly proportional to the negative log of the MR-IVW p-value. Green border indicates clusters with -log(MR-IVW p-value) > 3. **C**) Forest plots from PPI-based MR showing estimates (beta ± standard error) calculated by MR-IVW for select positive and negative clusters. **D**) PPI-based S-LDSC (coefficient ± standard error) using GWAS summary statistics for SLE, CVD and CAD. Bar color indicates coefficient significance.

Figure 5. Comprehensive PPI-based MR predicts sets of SLE associated variants and pathways causal of CAD. A) S-LDSC using GWAS summary statistics for SLE, CVD and CAD to estimate the heritability (coefficient ± standard error) of the 3,272 genes (open bars) and 1,972 protein-coding genes (hashed bars) predicted by 1,708 combined Immunochip and Phenoscanner-derived SNPs. Bar color indicates coefficient significance. (**B**) Functional and cell-type enrichments for cluster metastructures were determined using BIG-C (black labels) and I-scope (red labels), respectively. Bold black labels over colored shadings represent shared BIG-C functional annotations for the clusters they surround. Node size is proportional to the number of SNPs (height) mapping to the genes in each cluster (width). For node color, red and blue indicate significant

positive or negative estimates, respectively for 14/16 MR methods used (tier 1); light red and light blue, significant positive or negative estimates by MR-IVW or at least 7/16 MR-methods (tier 2); purple, mixed estimates; grey, insignificant. Thickness of the yellow border is roughly proportional to the negative log of the MR-IVW p-value. Green border indicates clusters with a negative log of the MR-IVW p-value > 3. **C-D**) Forest plots from PPI-based MR showing estimates (beta ± standard error) calculated by 16 MR methods for select (**C**) positive and (**D**) negative clusters. The number of SNPs used as IVs for each cluster are indicated in the plots. **E**) PPI-based S-LDSC (coefficient ± standard error) using GWAS summary statistics for SLE, CVD and CAD. Bar color indicates coefficient significance.

Figure 6. PPI-based MR identifies SLE SNPs with positive and negative causal effects on CAD. A) Workflow depicting PPI-based MR. **B-C**) PPI-based MR validation. Forest plots (beta ± standard error) from 16 MR methods using summary statistics from the SLE GWAS (**B**) or SLE Immunochip (**C**) as the exposure and CAD GWAS as the outcome. SLE-associated non-HLA SNPs mapping to positive and negative clusters, separately (by tier) and together ("All SNPs") were used as IVs after excluding CVD and confounder-associated SNPs followed by stringent LD-clumping (R^2=0.001) and harmonization. The number of SNPs used as IVs for each SNP set are indicated in the plots. For results, grey indicates insignificant ($p > 0.05$); dark red, positive ($p < 0.00075$); red, positive ($p < 0.05$); dark blue, negative ($p < 0.00075$); blue, negative ($p < 0.05$) by each MR method.

Figure 7. Genes and molecular pathways associated with positive causal clusters identify therapeutic interventions for managing CAD in SLE. A) All tier 1 and a selection of tier 2 clusters were functionally annotated using BIG-C, IPA and the EnrichR database. Select drugs acting on direct gene targets or on any of the associated pathways (italics) are listed. **B**) Venn diagram summarizing therapies that might uniquely impact SLE or CAD and those that may target pathways common to both diseases.

STAR METHODS

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the Lead Contact, Katherine A. Owen (kate.owen@ampelbiosolutions.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. All GWAS and Immunochip studies are referenced.
- Original software code and documentation have been deposited on figshare (www.figshare.com; doi.org/10.6084/m9.figshare.21225251) and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this report is available from the Lead Contact upon request.

Experimental models and subject details

This study did not use any experimental models or enroll human subjects.

METHOD DETAILS

Identification of SLE- and CAD-associated SNPs and overlap

SNPs associated with each disease were obtained from previous GWAS and Immunochip studies. For CAD, we used a comprehensive multi-ancestral meta-analysis of GWAS[32]. For SLE, we included results of multiple GWAS and Immunochip studies to account for as many ancestries as possible[27–31]. In total, 7,222 and 16,163 unique SNPs were significantly ($p<10^{-6}$) associated with SLE and CAD, respectively, and were employed in these studies. A full list of the SNPs, chromosome locations, positions and sources used are detailed in Table S9.

Identification of SNP-predicted genes

Expression quantitative trait loci (eQTLs) were identified using GTEx[56] version 6.8 (GTEXportal.org) and mapped to their associated eQTL expression genes (E-Genes). To find SNPs in enhancers and promoters, and their associated transcription factors and downstream target genes (T- Genes), we queried the atlas of Human Active Enhancers to interpret Regulatory variants[57] (HACER, http://bioinfo.vanderbilt.edu/AE/HACER). To find SNPs in exons of protein-coding genes (C-Genes) and include proximal genes (P-Genes, within 5kb), we queried the human Ensembl genome browser's variant effect predictor[58] (VEP, ensembl.org/info/docs/tools/vep, GRCh38.p12).

Network analysis and visualization

Protein-protein interaction (PPI) networks of SNP-predicted protein-coding genes were generated by STRING[59] (https://string-db.org, version 11.0b), and resulting networks were imported into Cytoscape[60] (version 3.6.1) for visualization and partitioned with MCODE via the clusterMaker2[61] (version 1.2.1) plugin. Metastructures are based on PPI networks. For all metastructures, node gradient shading is proportional to intra-cluster connectivity, cluster size indicates number of genes per cluster and edge weight indicates inter-cluster connections.

Functional gene set analysis

Predicted genes were examined using Biologically Informed Gene Clustering (BIG-C; version 4.4.). BIG-C is a custom functional clustering tool developed to annotate the biological meaning of large lists of genes and has been previously described[62–64]. I-Scope is a custom clustering tool used to identify immune cell types in large gene datasets[65]. The Ingenuity Pathway Analysis (IPA; https://www.qiagenbioinformatics.com) platform and EnrichR[66] (https://maayanlab.cloud/Enrichr/) web server provided additional molecular pathway enrichment analysis.

Drug candidate identification

Drug candidates were identified using LINCS[84], STITCH[85] (v5.0), IPA and literature mining. Each of the database tools includes either a programmatic method of matching existing therapeutics to their targets or else is a list of drugs and targets for achieving the same end.

QUANTIFICATION AND STATISTICAL ANAYSIS

Linkage Disequilibrium Score Regression (LDSC) Genetic Correlations

LDSC[87] was used to estimate genome-wide genetic correlations between traits using GWAS summary statistics. Pre-processed summary statistics from SLE, CAD and CVD GWAS were obtained from the Broad webpage (https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/). Using the LDSC software provided on github (https:// github.com/bulik/ldsc) and reference data on the Broad webpage (https://alkesgroup.broadinstitute.org/LDSCORE/), including European LD scores 'eur_w_ld_chr' or 'weights_hm3_no_hla' as weights for analyses excluding the HLA region. Using standard parameters, the "ldsc.py" (with the "--rg" flag) script was used to generate genome-wide genetic correlation estimates between SLE and CVD or CAD.

Stratified Linkage Disequilibrium Score Regression (S-LDSC)

S-LDSC[33] was used to obtain gene-set specific disease-heritability estimates using GWAS summary statistics. Pre-processed summary statistics from SLE, CAD and CVD GWAS were obtained from Broad webpage (https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/). Using the S-LDSC software provided on github (https:// github.com/bulik/ldsc) and reference data on the Broad webpage (https://alkesgroup.broadinstitute.org/LDSCORE/), annotation and LD score files were generated for each SNP-predicted gene- and protein- set, separately. Using standard parameters, the “make_annot.py” and “ldsc.py” (with the “--l2” flag) scripts were first used to generate the gene-set-specific annotation and LD files, then

the “ldsc.py” (with the “- - h2-cts” flag) script was used to generate stratified heritability scores for each GWAS.

Selection of valid, independent instrumental variables for traditional MR analysis

Traditional MR methods, such as MR-IVW, operate under three strict assumptions for instrumental variable (IV) validity: 1) the relevance assumption, 2) the exclusion restriction criteria assumption, and 3) the independence assumption. To satisfy the relevance assumption, SNPs significantly (genome-wide significance p-value < $5x10^{-8}$) associated with SLE[27,28,74–81,29,67–73] were obtained from the Phenoscanner database (www.phenoscanner.medschl.cam.ac.uk)(82,83) (Table S9). To satisfy the exclusion restriction criteria and independence assumptions, 89,336 SNPs weakly associated (p-value < $1x10^{-5}$) with CVD and confounders including cholesterol, obesity, blood pressure, insulin resistance, smoking, age-related diseases, and many more, were excluded from being IVs for SLE-exposure (see Table S2 for the full list of excluded traits). HLA-region SNPs were conservatively removed from MR analyses by excluding the short-arm of chromosome 6. Stringent LD clumping[37] was employed using the clump data (R^2=0.001, 100kb window, 1000G EA reference population) function to generate an independent set of 60 SLE-IVs harmonized for each GWAS.

Mendelian Randomization (MR)

MR was used to test for causal relationships between SLE and CAD using the MR-Base[45] (https://www.mrbase.org) TwoSampleMR[45] package in R (https://github.com/MRCIEU/TwoSampleMR). Various sets of SLE-associated genetic

variants used as instrumental variables (IVs) and summary statistics for SLE-exposure were manually imported into R and summary statistics were carried out for MR-base compatibility using the ‘format data’ command. All effect sizes and standard errors were obtained from the exposure summary statistics used in each analysis, regardless of the study in which each IV was associated with the exposure. Given the availability of well-powered CAD/MI GWAS on MR-Base, IVs for CAD and MI were directly obtained from each exposure GWAS using the ‘extract instruments’ command for the bidirectional analyses. Data from the SLE and all CVD-related GWAS studies used in our MR analyses are publicly available and also accessible through the MR-Base software, which was used to obtain the outcome summary statistics via the ‘extract outcome data’ command. The ‘allele harmonization’ command was used to ensure the effect estimates of the exposure and outcome are based on matching alleles, excluding SNPs with completely mismatching alleles from the MR analysis or reversing the effect and non-effect alleles along with the effect estimates when applicable. Because of the allele harmonization step and because some SNPs are absent from the available summary statistics, a small proportion of SNPs used as IVs are absent from the final MR calculations. Up to sixteen individual MR methods were carried out through the TwoSampleMR package, including inverse variance weighted (IVW), simple mode, weighted mode, simple median, weighted median (WMedian), MR-Egger, MR-PRESSO (raw and outlier-corrected), MR-RAPS, and two sample maximum likelihood (ML). The ‘MR report’ function was used to generate a summary containing heterogeneity and directional pleiotropy tests and scatterplots (Figure S2). MR-IVW and MR-Egger heterogeneity test results (Q-value) indicate whether significant heterogeneity was detected, which does not necessarily indicate biased causal

estimates. MR-Egger intercept indicate whether significant *directional* horizontal pleiotropy was detected, which usually indicates biased causal estimates. For single-SNP MR, the 'MR single-SNP' function was also carried out using the Wald Ratio method. Full details of all MR results are included in Table S2 and a summary of all the main findings are included in Table S10.

PPI-based MR

SLE-associated variants from the Immunochip[31] and Phenoscanner database[82,83] were linked to their most likely genes, and the genes used to generate PPI-informed gene clusters. The SLE-associated SNPs mapping to genes in each of PPI-based clusters were then extracted to "reverse engineer" subsets of SNPs that could be used separately as SLE-IVs for MR to independently estimate the causal effects of each PPI-informed SNP-to-Gene module on CAD. Up to sixteen MR methods were carried out for each SNP-to-gene module through the TwoSampleMR package.

In additional analyses (related to Figure 6) using the combined Immunochip and Phenoscanner SNP dataset, filtering eliminated SNPs weakly associated (p-value < $1x10^{-5}$) with CVD and confounders including cholesterol, obesity, blood pressure, insulin resistance, smoking, age-related diseases, and many more (Table S2). HLA-region SNPs were conservatively removed from MR analyses by excluding the short-arm of chromosome 6. Stringent LD clumping[37] was employed using the clump_data (R^2=0.001, 100kb window, 1000G EA reference population) function to generate an independent set of SLE-IVs. Various analyses were performed using independent, valid IVs derived from 'All SNPs', SNPs mapping to 'Insignificant' clusters, 'Positive Tier 1' clusters, 'Positive

Tier 1 and Tier 2' clusters, 'Positive MR-IVW p-value<0.00075' clusters, 'Negative MR-IVW p-value<0.00075' clusters, 'Negative Tier 1' clusters, and 'Negative Tier 1 and Tier 2' clusters.

Monte Carlo Simulations for expected MR results using random sets of Immunochip-derived SNP-to-Gene modules

Monte Carlo Simulations were implemented and performed to estimate the false discovery rate with respect to significant PPI-based MR causal estimates. 120,026 Immunochip SNPs included in the SLE summary statistics were mapped to putative genes using the VEP, including regulatory effects, to generate an Immunochip SNP-to-Gene library with 67,211 unique SNPs mapping to 7,602 STRINGdb proteins. In each simulation, a random set of 3 to 152 SNP-predicted proteins were selected from the 7,602 proteins and used to extract up to 400 Immunochip SNPs. MR-IVW was then performed for SLE on CAD using harmonized, non-HLA SNPs (via removal of the entire short-arm of chromosome 6) from the simulated set of Immunochip SNPs as IVs. By using our Immunochip derived SNP-to-Gene dictionary for random selection of protein clusters and associated SNPs to generate random sets of IVs, our simulations account for both a high degree of LD and pleiotropy, especially considering the major influence of loci associated with diabetes in development of the Immunochip.

Supplemental Table Legends

Table S1. Canonical pathway and disease phenotype enrichments for network analysis of SNP-predicted genes overlapping SLE and CAD. Related to Figure 1 and

S1. Gene set enrichments for each cluster were determined using IPA and the EnrichR library database. P-values from Fisher's exact test measures the significance of overlap between genes in each cluster and genes within an annotation.

Table S2. List of all included SLE and excluded CVD/confounder-associated traits from the Phenoscanner database for use as SLE-IVs in MR analyses. Related to Figure 2.

Table S3. Full MR results. Related to Figures 2, 3, 4, 5, and 6.

Table S4. Pathway analysis of SLE IVs determined to be causal of CAD by traditional MR. Related to Figure 2. Gene set enrichments for each cluster were determined using IPA and the EnrichR library database. P-values from Fisher's exact test measures the significance of overlap between genes in each cluster and genes within an annotation.

Table S5. Canonical pathway and disease phenotype enrichments for network analysis of positive and negative causal SNP-predicted genes determined by single-SNP MR. Related to Figure 3. P-values from Fisher's exact test that measures the significance of overlap between genes in each cluster and genes within an annotation

Table S6. PPI-based S-LDSC results for the 46 and 67 PPI-based of genes clusters derived from SLE-associated SNPs. Related to Figure 4.

Table S7. Validation of PPI-based MR-IVW results for 46 and 67 SLE-derived clusters on CAD. Related to Figures 4, 5, 6 and S6. MR-IVW results for SLE on CVD-related summary statistics available on the MR-base platform, including two CAD GWAS, two MI GWAS, Ischemic Stroke, Cardiomyopathy, and Atrial Fibrillation.

Table S8. Pathway analysis of positive and negative-casual Tier 1 and Tier 2 SLE-derived SNP-predicted protein clusters with significant (p-value<0.05) causal estimates on CAD by PPI-based MR for the comprehensive 67-cluster network. Related to Figure 5.

Table S9. Lists of SNPs, chromosome locations, p-values and sources (where available). Related to Figures 1, S1, 2, 3, 4 and 5.

Table S10. Summary of major findings. Related to STAR Methods.

REFERNCES

1. Yen EY, Singh RR. (2018). Brief Report: Lupus—An Unrecognized Leading Cause of Death in Young Females: A Population-Based Study Using Nationwide Death Certificates, 2000–2015. Arthritis Rheumatol. 70(8):1251–5.
2. Liu Y, Kaplan MJ. (2018). Cardiovascular disease in systemic lupus erythematosus: An update. Curr Opin Rheumatol. 30(5):441–8.
3. Gupta S, Kaplan MJ. (2021). Bite of the wolf: innate immune responses propagate autoimmunity in lupus. J. Clin. Inves. 131(3): e144918.
4. Thomas G, Mancini J, Jourde-Chiche N, Sarlon G, Amoura Z, Harlé JR, Jougla E, Chiche, L. (2014). Mortality associated with systemic lupus erythematosus in France assessed by multiple-cause-of-death analysis. Arthritis Rheumatol. 66(9):2503–11.
5. Teixeira V, Tam LS. (2018). Novel Insights in Systemic Lupus Erythematosus and Atherosclerosis. Front. Med. 4:262.
6. Moe SR, Haukeland H, Molberg Ø, Lerang K. (2021). Long-term outcome in systemic lupus erythematosus; knowledge from population-based cohorts. J. Clin. Med. 10(19):4306. doi: 10.3390/jcm10194306.
7. Katz G, Smilowitz NR, Blazer A, Clancy R, Buyon JP, Berger JS. (2019). Systemic Lupus Erythematosus Is Associated With Increased Prevalence of Atherosclerotic Cardiovascular Disease in Hospitalized Patients. Mayo. Clin. Proc. 94(8):1436.
8. Lopez EO, Ballard BD, Jan A. (2021). Cardiovascular Disease. (2021). In StatPearls [Internet], editorial board. (StatPearls Publishing). PMID: 30571040;

Bookshelf ID: NBK535419.

9. Petri M, Perez-Gutthann S, Spence D, Hochberg MC. (1992). Risk factors for coronary artery disease in patients with systemic lupus erythematosus. Am. J. Med. 93(5):513–9.

10. Moghaddam B, Marozoff S, Li L, Sayre EC, Zubieta JAA. (2022). All-cause and cause-specific mortality in systemic lupus erythematosus: a population-based study. Rheumatology. 61(1):367.

11. Bernatsky S, Boivin JF, Joseph L, Manzi S, Ginzler E, Gladman DD, Urowitz M, Fortin PR, Petri M, Barr S, et al. (2006). Mortality in systemic lupus erythematosus. Arthritis. Rheum. 54(8):2550–7.

12. Wu GC, Liu HR, Leng RX, Li XP, Li XM, Pan HF, Ye D-Q. (2016). Subclinical atherosclerosis in patients with systemic lupus erythematosus: A systemic review and meta-analysis. Autoimmun. Rev. 15(1):22–37.

13. Yazdany J, Pooley N, Langham J, Nicholson L, Langham S, Embleton N, Wang X, Desta B, Barut V, Hammond E. (2020). Original research: Systemic lupus erythematosus; stroke and myocardial infarction risk: a systematic review and meta-analysis. RMD Open. 6(2): e001247.

14. Magder LS, Petri M. (2012). Incidence of and Risk Factors for Adverse Cardiovascular Events Among Patients With Systemic Lupus Erythematosus. Am. J. Epidemiol. 176(8):708.

15. Petri MA, Barr E, Magder LS. (2019). Development of a systemic lupus erythematosus cardiovascular risk equation. Lupus Sci. Med. 6(1): e000346.

16. Levinson DJ, Abugroun A, Daoud H, Abdel-Rahman M. (2020). Coronary artery

disease (CAD) risk factor analysis in an age-stratified hospital population with systemic lupus erythematosus (SLE). Int. J. Cardiol. Hypertens. 7: 100056.

17. Ajeganova S, Hafström I, Frostegård J. (2021). Patients with SLE have higher risk of cardiovascular events and mortality in comparison with controls with the same levels of traditional risk factors and intima-media measures, which is related to accumulated disease damage and antiphospholipid syndrome: a case–control study over 10 years. Lupus Sci. Med. [Internet]. 8(1): e000454.
18. Bakshi J, Segura BT, Wincup C, Rahman A. Unmet Needs in the Pathogenesis and Treatment of Systemic Lupus Erythematosus. Clin. Rev. Allergy Immunol. 55(3):352–67.
19. Calvo-Alén J, Alarcón GS, Campbell R, Fernández M, Reveille JD, Cooper GS. (2005). Lack of recording of systemic lupus erythematosus in the death certificates of lupus patients. Rheumatology. 44(9):1186–9.
20. Falasinnu T, Rossides M, Chaichian Y, Simard JF. (2018). Do Death Certificates Underestimate the Burden of Rare Diseases? The Example of Systemic Lupus Erythematosus Mortality, Sweden, 2001-2013. Public Health Rep. 133(4):481–8.
21. Ashley EA, Hershberger RE, Caleshu C, Ellinor PT, Garcia JGN, Herrington DM, Ho CY, Jhonson JA, Kittner SJ, Macrae CA, et al. (2012). Genetics and cardiovascular disease: a policy statement from the American Heart Association. Circulation. 126(1):142–57.
22. Leonard D, Svenungsson E, Dahlqvist J, Alexsson A, Ärlestig L, Taylor KE, Sandlin JK, Bengtsson C, Frolund M, Jonsen A, et al. Novel gene variants associated with cardiovascular disease in systemic lupus erythematosus and

rheumatoid arthritis. Ann. Rheum. Dis. 77(7):1063–9.

23. Ramirez GA. (2018). Genetics in systemic lupus erythematosus: entering the borough of cardiovascular risk. Ann. Transl. Med. 6(Suppl 1):S14.

24. Svenungsson E, Gustafsson J, Leonard D, Sandling J, Gunnarsson I, Nordmark G, Jonsen A, Bengtsson AA, Sturfelt G, Rantapaa-Dahlqvist A, et al. (2010). A STAT4 risk allele is associated with ischaemic cerebrovascular events and anti-phospholipid antibodies in systemic lupus erythematosus. Ann. Rheum. Dis. 69(5):834–40.

25. Leonard D, Svenungsson E, Sandling JK, Berggren O, Jönsen A, Bengtsson C, Wang C, Jensen-Urstad K, Granstam S-O, Bengtsston AA, et al. (2013). Coronary heart disease in systemic lupus erythematosus is associated with interferon regulatory factor-8 gene variants. Circ. Cardiovasc. Genet. 6(3):255–63.

26. Owen KA, Price A, Ainsworth H, Aidukaitis BN, Bachali P, Catalina MD, Dittman JM, Howard TD, Kingsmore KM, Labonte AC, et al. (2021). Analysis of Trans-Ancestral SLE Risk Loci Identifies Unique Biologic Networks and Drug Targets in African and European Ancestries. Am J Hum Genet. 107(5):864–81.

27. Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tombleson P, Behrens TW, Martin J, Fairfax BP, Knight JC, Chen L, Replogle J, et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet. 47(12):1457-64.

28. Lessard CJ, Sajuthi S, Zhao J, Kim K, Ice JA, Li H, Ainsworth H, Rasmussen A, Kelly JA, Marion M, et al. (2016). Identification of a Systemic Lupus

Erythematosus Risk Locus Spanning ATG16L2, FCHSD2, and P2RY2 in Koreans. Arthritis Rheumatol. 68(5):1197–209.

29. Morris DL, Sheng Y, Zhang Y, Wang Y-F, Zhu Z, Tombleson P, Chen L, Cunninghame Graham DS, Bentham J, Roberts AL, et al. (2016). Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nat. Genet . 48(8):940–6.

30. Sun C, Molineros JE, Looger LL, Zhou X-J, Kim K, Okada Y, Ma J, Qi Y-Y, Howard XK, Motghare P, et al. (2016). High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. Nat. Genet. 48(3):323–30.

31. Langefeld CD, Ainsworth HC, Graham DSC, Kelly JA, Comeau ME, Marion MC, Howard TD, Ramos PS, Croker JA, Morris DL, et al. (2017). Transancestral mapping and genetic load in systemic lupus erythematosus. Nat. Commun. 8.:16021.

32. Van Der Harst P, Verweij N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. Circ. Res. 122(3):433–43.

33. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zanh C, Farh K, et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47(11):1228–35. .

34. Warren HR, Evangelou E, Cabrera CP, Gao H, Ren M, Mifsud B, Ntalla I, Surendran P, Liu C, Cook JP, et al. (2017). Genome-wide association analysis

identifies novel blood pressure loci and offers biological insights into cardiovascular risk. Nat. Genet. 49(3):403–15.

35. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AFR, Barbalic M, Gieger C, et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. 43(4):333–40.

36. Howson JMM, Zhao W, Barnes DR, Ho WK, Young R, Paul DS, Waite LL, Freitag DF, Fauman EB, Salfati EL, et al. (2017). Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. Nat. Genet. 49(7):1113–9.

37. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int. J. Methods Psychiatr. Res. 27(2): e1608.

38. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, et al. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. 47(10):1121-30.

39. Hartiala JA, Han Y, Jia Q, Hilser JR, Huang P, Gukasyan J, Schwartzman WS, Cai Z, Biswas S, Tregouet D-A, et al. (2021). Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. Eur. Heart J. 42(9):919–33.

40. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese A-K, van der Laan SW, Gretarsdottir S, et al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat. Genet. 50(4):524-537.

41. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G, et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat. Genet. 50(9):1234–9.

42. Raj P, Rai E, Song R, Khan S, Wakeland BE, Viswanathan K, Arana C, Liang C, Zhang B, Dozmorov I, et al. (2016). Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity. Elife. 5: e12089.

43. Davies RW, Wells GA, Stewart AFR, Erdmann J, Shah SH, Ferguson JF, Hall AS, Anand SS, Burnett MS, Epstein SE, et al. (2012). A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. Circ. Cardiovasc. Genet. 5(2):217–25.

44. Trowsdale J, Knight JC. (2013). Major histocompatibility complex genomics and human disease. Annu. Rev. Genomics Hum. Genet. 14:301–23.

45. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. Elife. 7: e34408.

46. Piranavan P, Perl A. (2020). Management of cardiovascular disease in patients with systemic lupus erythematosus. Expert. Opin. Pharmacother. 21(13):1617.

47. Burgess S, Zuber V, Valdes-Marquez E, Sun BB, Hopewell JC. (2017). Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. Genet. Epidemiol. 41(8):714–25.

48. Reiss AB, Anwar K, Merrill JT, Chan ESL, Awadallah NW, Cronstein BN, Belmont HM, Belilos E, Rosenblum G, Belostocki K, et al. (2010). Plasma from systemic

lupus patients compromises cholesterol homeostasis: a potential mechanism linking autoimmunity to atherosclerotic cardiovascular disease. Rheumatol. Int. 30(5):591–8.

49. Quinn CM, Jessup W, Wong J, Kritharides L, Brown AJ. (2005). Expression and regulation of sterol 27-hydroxylase (CYP27A1) in human macrophages: a role for RXR and PPARgamma ligands. Biochem. J. 385(Pt 3):823–30.

50. Bilotta MT, Petillo S, Santoni A, Cippitelli M. (2020). Liver X Receptors: Regulators of Cholesterol Metabolism, Inflammation, Autoimmunity, and Cancer. Front. Immunol. 11:584303.

51. Liu A, Rahman M, Hafström I, Ajeganova S, Frostegård J. (2020). Proprotein convertase subtilisin kexin 9 is associated with disease activity and is implicated in immune activation in systemic lupus erythematosus. Lupus. 29(8):825–35.

52. Yao Mattisson I, Rattik S, Björkbacka H, Ljungcrantz I, Terrinoni M, Lebens M, Holmgren J, Fredrikson GN, Gullstrand B, Bengtsson AA, et al. (2021). Immune responses against oxidized LDL as possible targets for prevention of atherosclerosis in systemic lupus erythematosus. Vascul. Pharmacol. 140:106863.

53. Barnado A, Carroll RJ, Casey C, Wheless L, Denny JC, Crofford LJ. (2018). Phenome-wide association study identifies marked increased in burden of comorbidities in African Americans with systemic lupus erythematosus. Arthritis Res. Ther. 20(1):69..

54. Esteller M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. 12(12):861–74.

55. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. (2014). Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochim Biophys Acta. 1842(10):1910–22.

56. The GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Gen. 45(6):580-585.

57. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. (2019). HACER: An atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res. 47(D1):D106–12.

58. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. 17:122.

59. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 49(D1):D605–12.

60. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498–504.

61. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics. 12:436.

62. Labonte AC, Kegerreis B, Geraci NS, Bachali P, Madamanchi S, Robl R, Catalina

MD, Lipsky PE, Grammer AC. (2018). Identification of alterations in macrophage activation associated with disease activity in systemic lupus erythematosus. PLoS One. 13(12): e0208132.

63. Catalina MD, Bachali P, Geraci NS, Grammer AC, Lipsky PE. (2019). Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. Commun. Biol. 2019 Dec;2(1).

64. Catalina MD, Owen KA, Labonte AC, Grammer AC, Lipsky PE. (2019). The pathogenesis of systemic lupus erythematosus: Harnessing big data to understand the molecular basis of lupus. J. Autoimmun. 110:102359.

65. Ren J, Catalina MD, Eden K, Liao X, Read KA, Luo X, McMillan RP, Hulver MW, Jarpe M, Bachali P, Grammer AC, et al. (2019). Selective histone deacetylase 6 inhibition normalizes b cell activation and germinal center formation in a model of systemic lupus erythematosus. Front. Immunol. 10:2512.

66. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G V, Clark NR, Ma'ayan A. (2013).Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 14:128.

67. Yang W, Shen N, Ye DQ, Liu Q, Zhang Y, Qian XX, Hirankarn N, Ying D, Pan H-F, Mok CC, et al. (2010). Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. PLoS Genet. 6(2): e1000841.

68. Chung SA, Taylor KE, Graham RR, Nititham J, Lee AT, Ortmann WA, Jacob CO, Alarcon-Riquelme ME, Tsao BP, Harley JB, et al. (2011). Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA

autoantibody production. PLoS Genet. 7(3): e1001323.

69. Yang J, Yang W, Hirankarn N, Ye DQ, Zhang Y, Pan HF, Mok CC, Chan TM, Wong RWS, Mok MY, et al. (2011). ELF1 is associated with systemic lupus erythematosus in Asian populations. Hum. Mol. Genet. 20(3):601–7.

70. Lee YH, Bae SC, Choi SJ, Ji JD, Song GG. (2012). Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. Mol. Biol. Rep. 39(12):10627–35.

71. Okada Y, Shimane K, Kochi Y, Tahira T, Suzuki A, Higasa K, Takahashi A, Horita T, Atsumi T, Ishii T, et al. (2012). A genome-wide association study identified AFF1 as a susceptibility locus for systemic lupus eyrthematosus in Japanese. PLoS Genet. 8(1): e1002455.

72. Yang W, Tang H, Zhang Y, Tang X, Zhang J, Sun L, Yang J, Cui Y, Zhang L, Hirankarn N, et al. (2013). Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. Am. J. Hum. Genet. 92(1):41–51.

73. Armstrong DL, Zidovetzki R, Alarcón-Riquelme ME, Tsao BP, Criswell LA, Kimberly RP, Harley JB, Sivils KL, Vyse TJ, Gaffnet PM, et al. (2014). GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. Genes Immun. 15(6):347–54.

74. Alarcón-Riquelme ME, Ziegler JT, Molineros J, Howard TD, Moreno-Estrada A, Sánchez-Rodríguez E, Ainsworth HC, Ortiz-Tello P, Comeau ME, Rasmussen A, et al. (2016). Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role

of European Admixture. Arthritis Rheumatol. 68(4):932–43.

75. Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S, Lee AT, Chung SA, Ferreira RC, Pant PVK, et al. (2008). Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N. Engl. J. Med. 358(9):900–9.

76. Graham RR, Cotsapas C, Davies L, Hackett R, Lessard CJ, Leon JM, Burtt NP, Guiducci C, Parkin M, Gates C, Plenge RM, et al. (2008). Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat. Genet. 40(9):1059–61.

77. Harley JB, Alarcón-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, Tsao BP, Vyse TJ, Langefeld CD, Nath S, et al. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nat. Genet. 40(2):204–10.

78. Kozyrev S V., Abelson AK, Wojcik J, Zaghlool A, Linga Reddy MVP, Sanchez E, Gunnarsson I, Svenungsson E, Sturfelt G, Jonsen A, et al. (2008). Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. Nat. Genet. 40(2):211–6.

79. Oishi T, Iida A, Otsubo S, Kamatani Y, Usami M, Takei T, Uchida K, Tsuchiya K, Saito S, Ohnisi Y, et al. (2008). A functional SNP in the NKX2.5-binding site of ITPR3 promoter is associated with susceptibility to systemic lupus erythematosus in Japanese population. J. Hum. Genet. 53(2):151–62.

80. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, Sun X, Ortmann W, Kosoy

R, Ferreira RC, Nordmark G, et al. (2009). A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat. Genet. 41(11):1228–33.

81. Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, Xu J-H, Cai Z-M, Huang W, Zhao G-P, et al. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. 41(11):1234–7.

82. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics. 32(20):3207–9.

83. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, Butterworth AS, Staley JR. (2019). PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 35(22):4851–3.

84. Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, Wang Z, Dohlman, Silverstein MC, Lachmann A, et al. (2018). The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst. 6(1):13–24.

85. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, Von Mering C, Jensen LJ, Beyer A, Bork P. (2010). STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res. 38(Database issue):D552-6.

86. Demirci FY, Wang X, Kelly JA, Morris DL, Barmada MM, Feingold E. Kao AH, Sivils KL, Bernatsky S, Pineau C, et al. (2016). Identification of a new

susceptibility locus for systemic lupus erythematosus on chromosome 12 in individuals of European ancestry. Arthritis Rheumatol. 68(1)174-83.

87. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, ReproGen Consortium, Psychiatric Genomic Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. (47) 1236-41.

CellPress

TABLE FOR AUTHOR TO COMPLETE

Please upload the completed table as a separate document. ***Please do not add subheadings to the key resources table.*** *If you wish to make an entry that does not fall into one of the subheadings below, please contact your handling editor.* ***Any subheadings not relevant to your study can be skipped.*** *(**NOTE:** For authors publishing in Cell Genomics, Cell Reports Medicine, Current Biology, and Med, please note that references within the KRT should be in numbered style rather than Harvard.)*

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GTEX v.6.8	The Genotype-Tissue Expression (GTEx) Project	GTEXportal.org
Pre-processed summary statistics	The Broad Institute	https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/
UK Biobank	Biobank UK	https://www.ukbiobank.ac.uk/
Cardiomyopathy GWAS	FinnGen Biobank	Finn-a-I9_CARDMYO; https://www.finngen.fi/fi
MR-Base	Hemani et al., 2018	https://www.mrbase.org
Phenoscanner	Staley et al., 2016; Kamat et al., 2019	www.phenoscanner.medschl.cam.ac.uk
Software and algorithms		
Bioconductor (R)	Open source	https://www.bioconductor.org
Ingenuity pathway analysis (IPA)	Qiagen	https://www.qiagenbioinformatics.com
S-LDSC	Finucane et al., 2015	https:// github.com/bulik/ldsc
LDSC	Bulik-Sullivan et al., 2015	https:// github.com/bulik/ldsc
Cytoscape v.3.9.1	Shannon et al., 2003	https://cytoscape.org
Clustermaker2 v.1.2.1	Morris et al., 2011	https://apps.cytoscape.org
TwoSample MR	Hemani et al., 2018	https://github.com/MRCIEU/TwoSampleMR
PPI-based MR	This manuscript	http://doi.org/10.6084/m9.figshare.21225251
Other		
Human Active Enhancers to interpret Regulatory variants (HACER)	Wang et al., 2019	http://bioinfo.vanderbilt.edu/AE/HACER
Variant Effect Predictor (VEP)	McLaren et al., 2016	ensembl.org/info/docs/tools/vep
Search Tool for the Retrieval of Interacting Genes/proteins (STRING) v. 11.0b	Szklarczyk et al., 2021	https://string-db.org
EnrichR	Chen et al., 2013	https://maayanlab.cloud/Enrichr/
Search Tool for Interacting Chemicals (STITCH)	Kuhn et al., 2009	http://stitch.embl.de

Library of Integrated Network-based Cellular Signatures (LINCS)	Keenan et al., 2018	http://www.lincs.hms.harvard.edu/db/

Figure 1

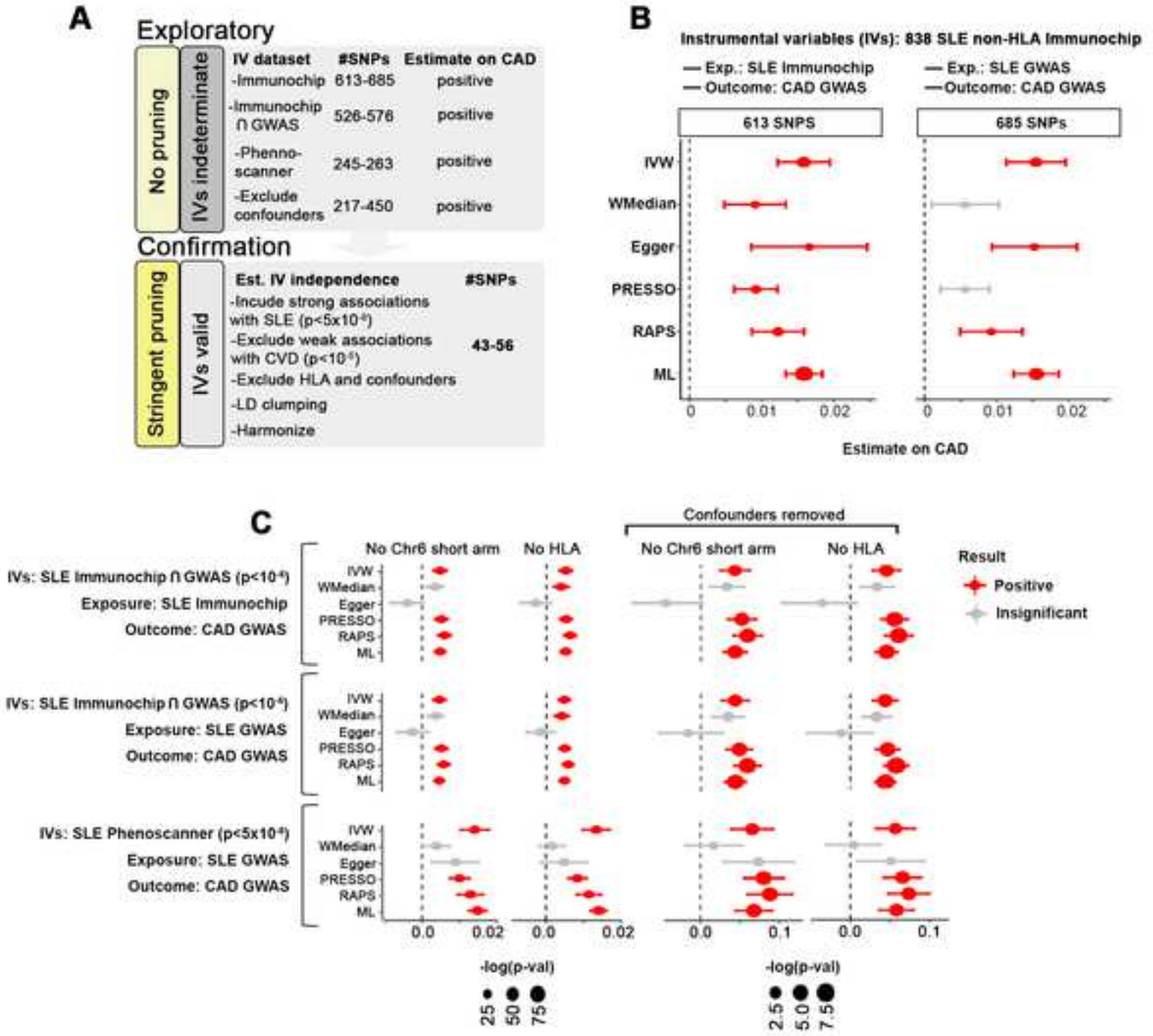


Figure 2

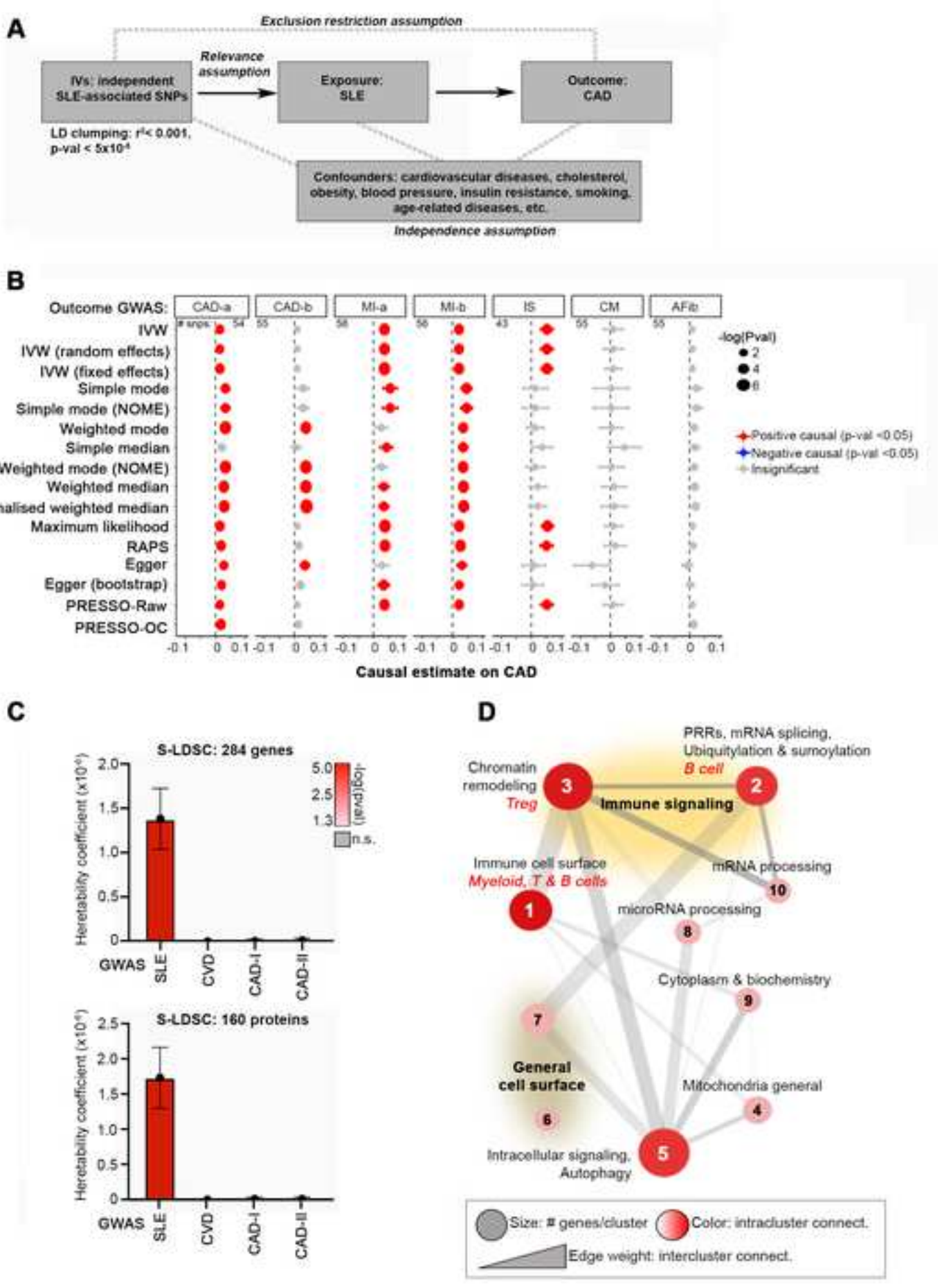


Figure 3

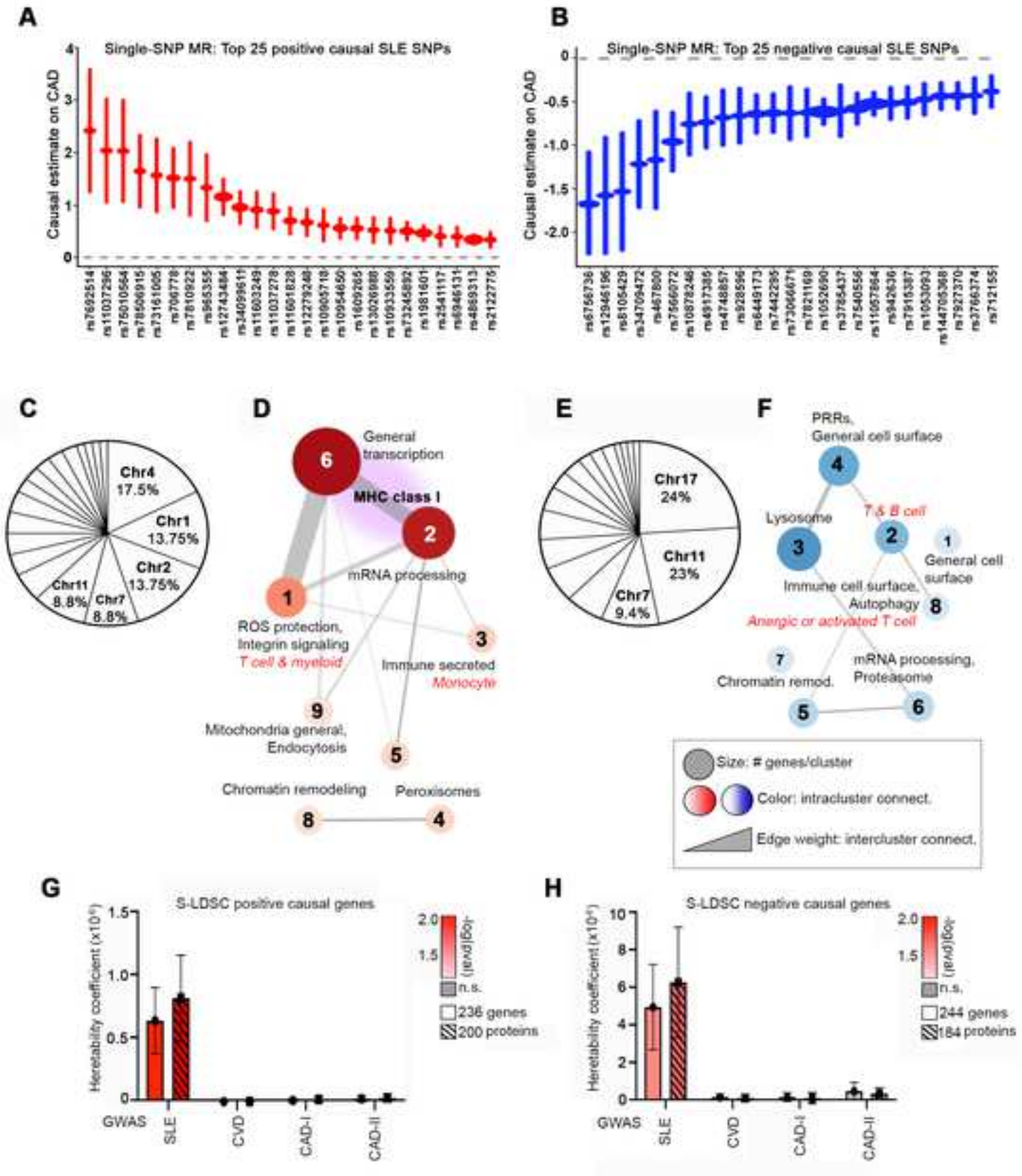


Figure 4

Click here to access/download;Figure;JulyCAD Figure 4.tif

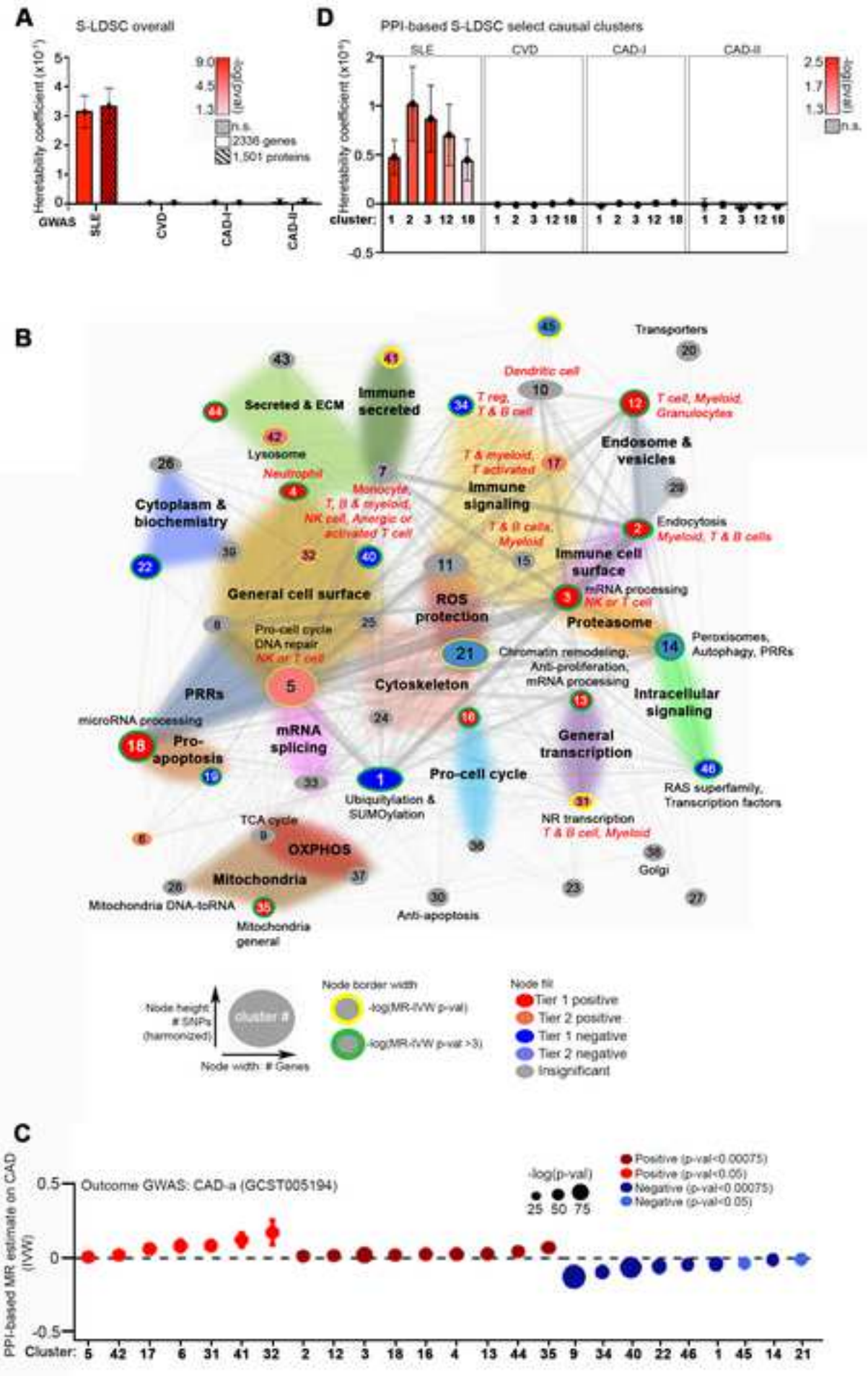


Figure 5

Click here to access/download;Figure;JulyCAD Figure 5.tif

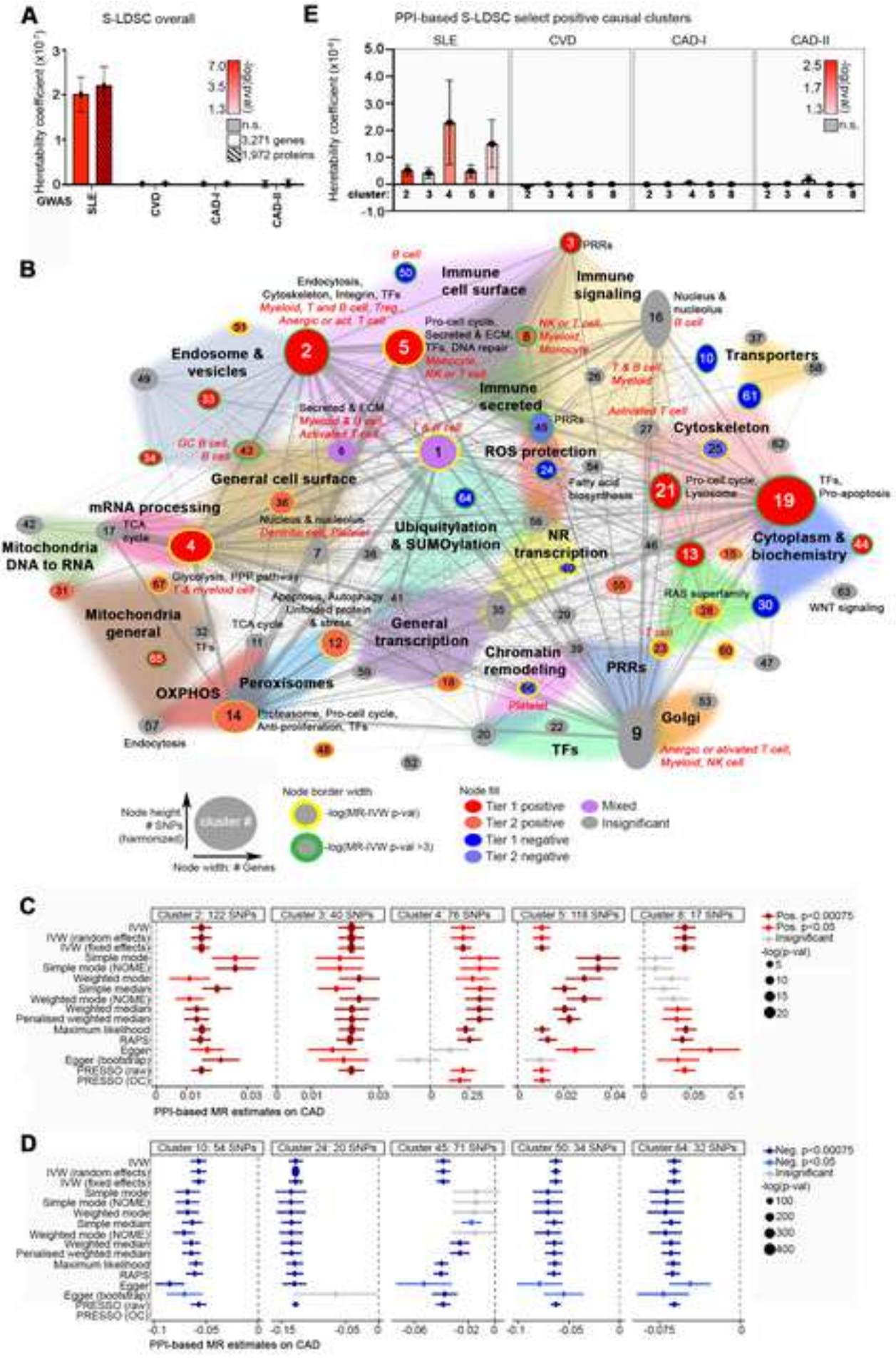


Figure 6

Click here to access/download;Figure;JulyCAD Figure 6.tif

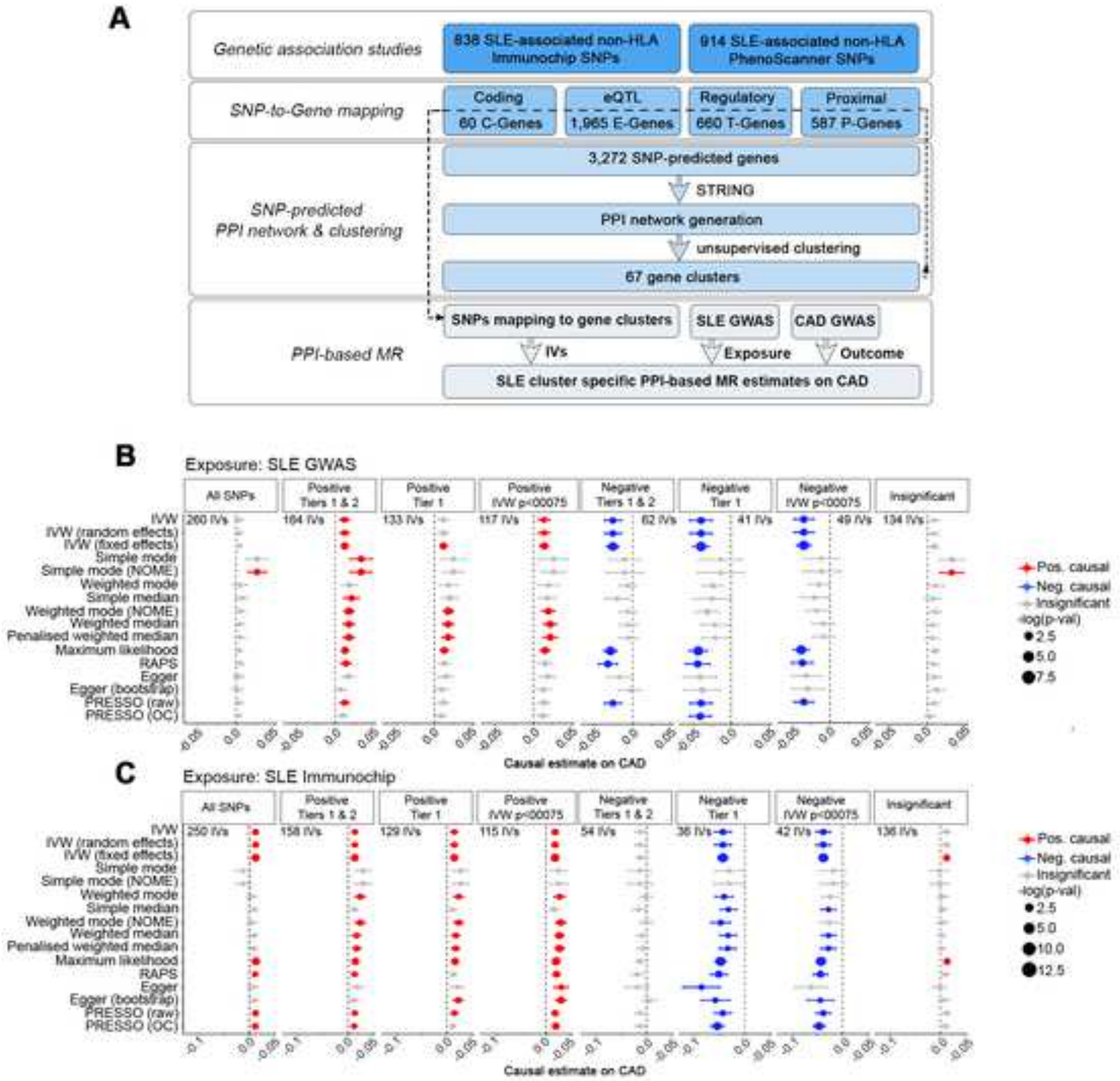
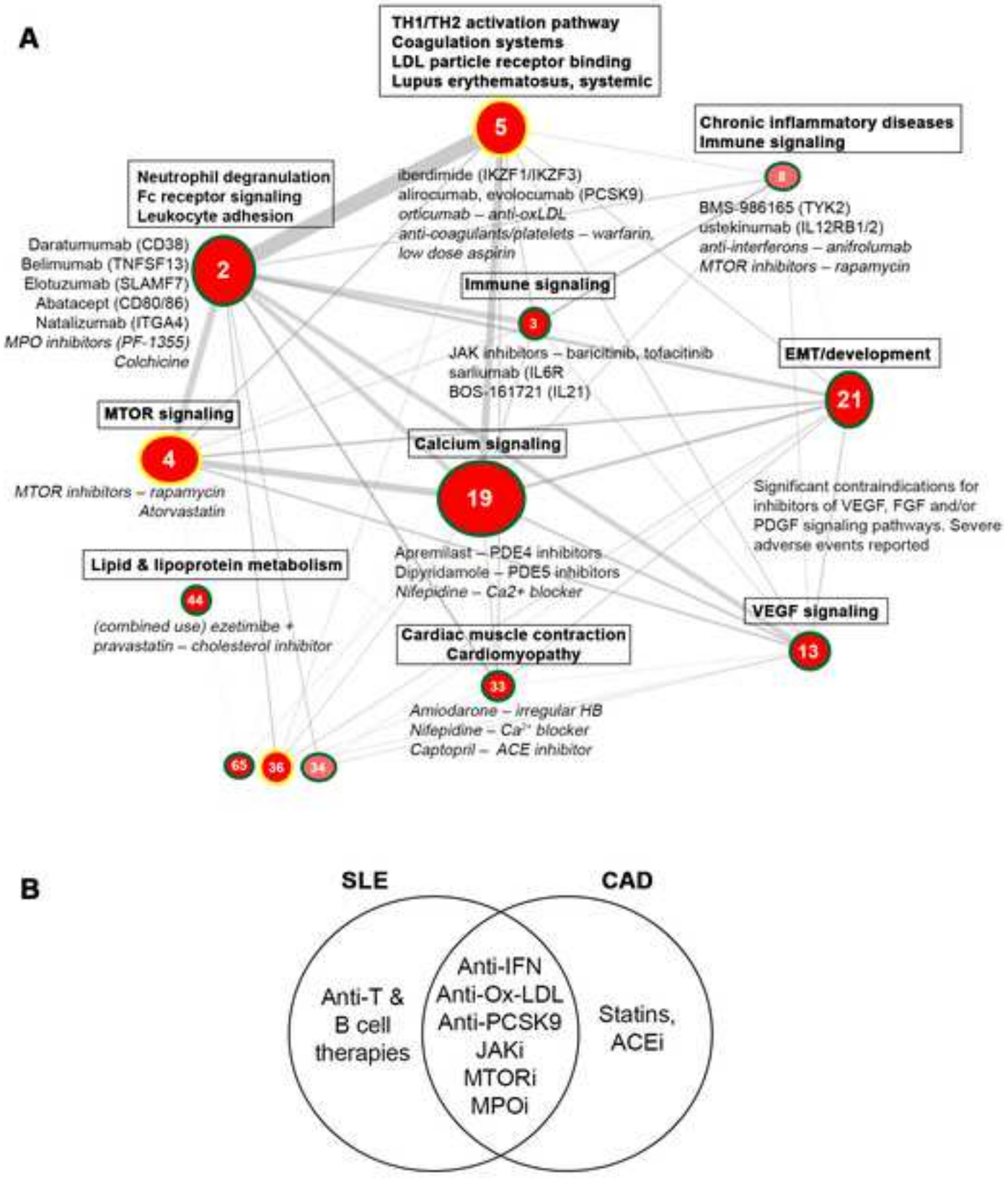


Figure 7



Supplemental Figures and Legends

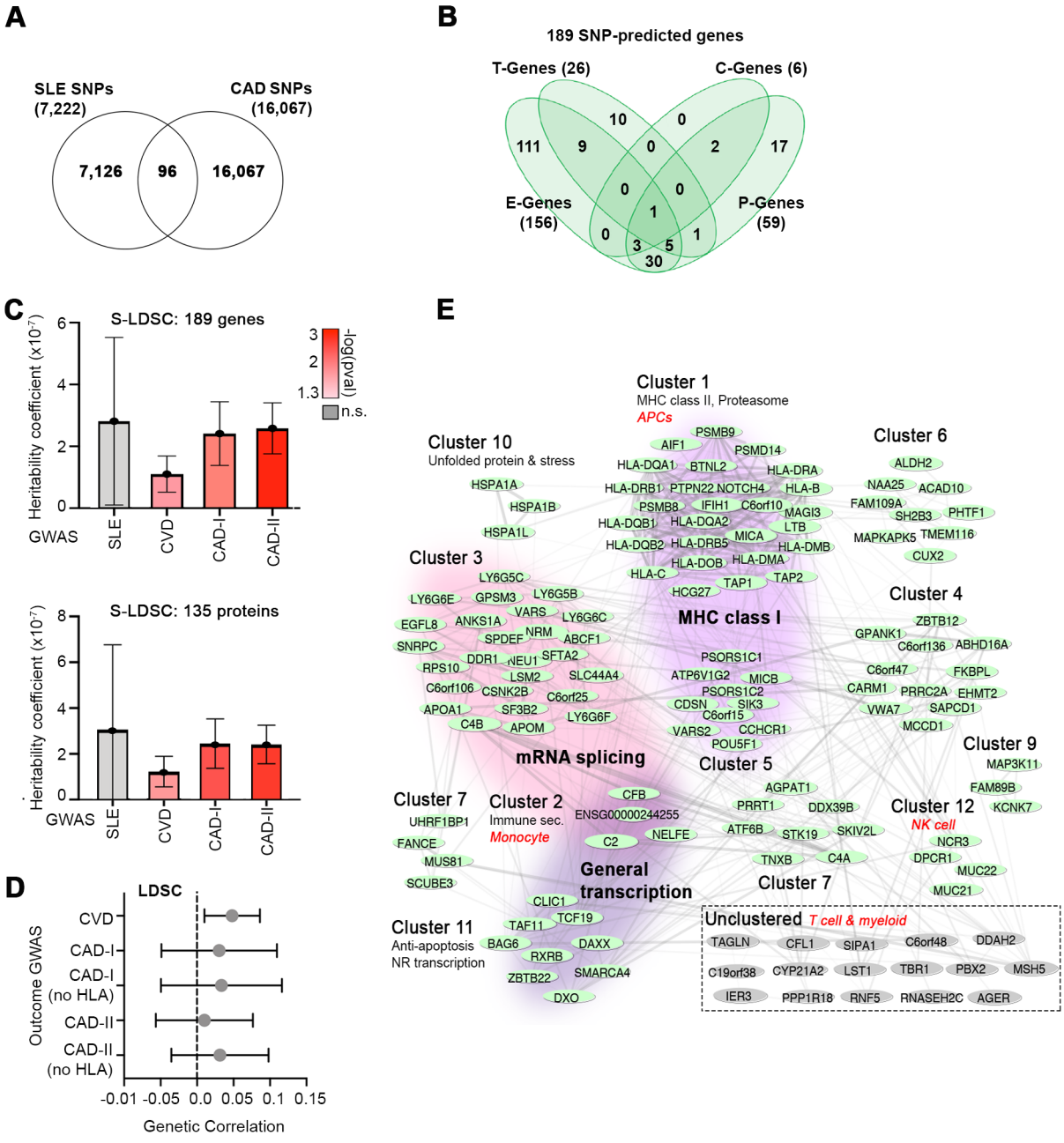


Figure S1. Analysis of SNP-predicted genes associated with both SLE and CAD. Related to Figures 1 and 2. **A**) Venn diagram of overlap between multi-ancestral SLE- and CAD-associated ($p<10^{-6}$) SNPs. **B**) Venn diagram of overlap between SNP-predicted genes derived from regulatory elements (T-Genes), eQTL analysis (E-Genes), coding regions (C-Genes), and proximity within 5kb (P-Genes). **C**) Application of S-LDSC using summary statistics for SLE, CVD, and CAD GWAS to estimate the heritability of the 189 SNP-predicted genes (top panel) and 135 SNP-predicted proteins (lower panel) from STRINGdb. Bar color indicates coefficient significance. **D**) Application of LDSC to estimate the genetic correlation between SLE and CAD or CVD. **E**) PPI network consisting of 135 putative protein-coding genes. Functional and cell-type enrichments for each cluster were determined using BIG-C (black labels) and I-scope (red labels), respectively. Black labels over colored shadings represent shared BIG-C functional annotations for the clusters they surround.

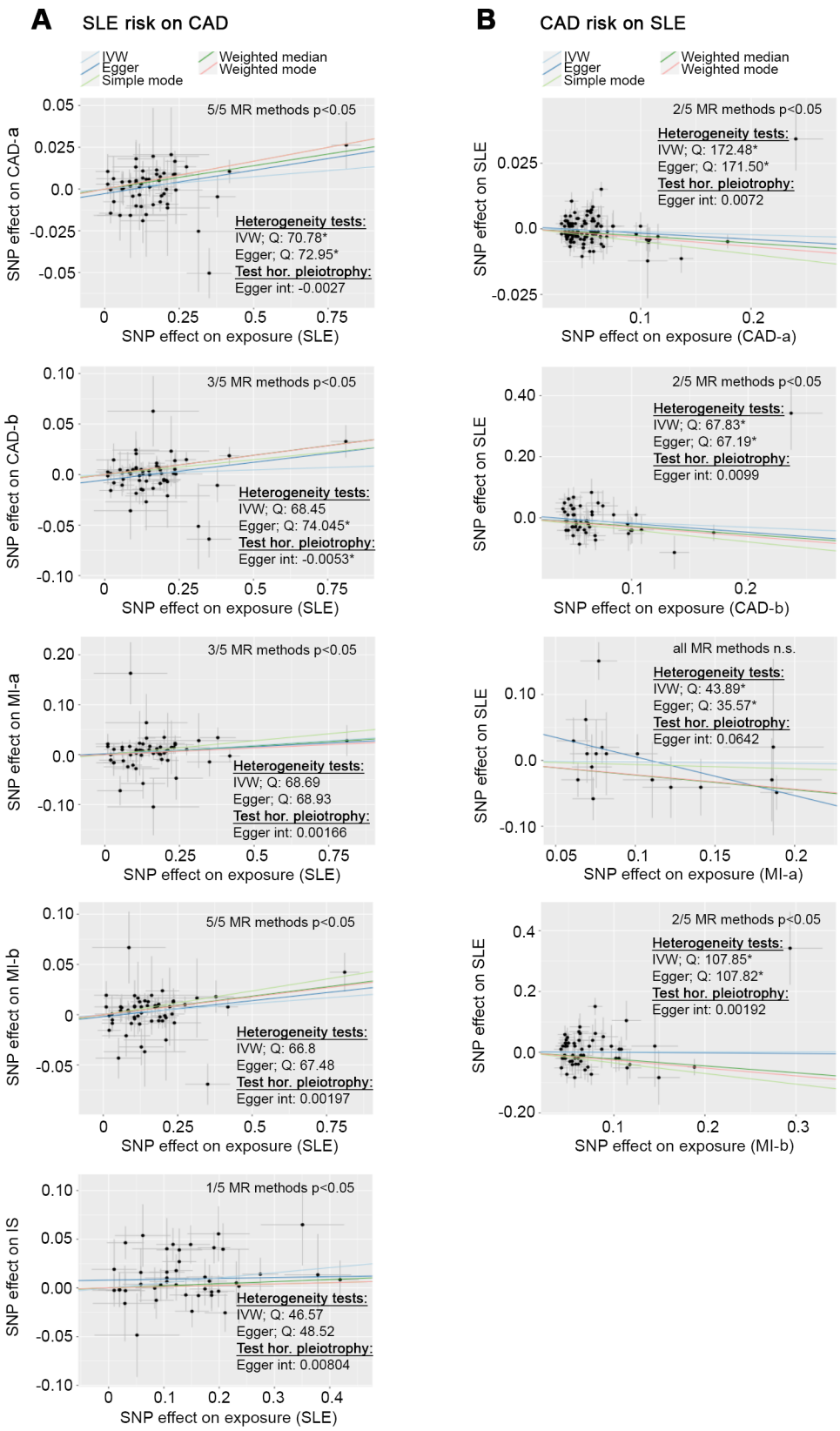

A
SLE risk on CAD
IVW
Egger
Simple mode
Weighted median
Weighted mode
5/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 70.78*
Egger; Q: 72.95*
Test hor. pleiotrophy:
Egger int: -0.0027
SNP effect on CAD-a
SNP effect on exposure (SLE)
3/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 68.45
Egger; Q: 74.045*
Test hor. pleiotrophy:
Egger int: -0.0053*
SNP effect on CAD-b
SNP effect on exposure (SLE)
3/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 68.69
Egger; Q: 68.93
Test hor. pleiotrophy:
Egger int: 0.00166
SNP effect on MI-a
SNP effect on exposure (SLE)
5/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 66.8
Egger; Q: 67.48
Test hor. pleiotrophy:
Egger int: 0.00197
SNP effect on MI-b
SNP effect on exposure (SLE)
1/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 46.57
Egger; Q: 48.52
Test hor. pleiotrophy:
Egger int: 0.00804
SNP effect on IS
SNP effect on exposure (SLE)
B
CAD risk on SLE
IVW
Egger
Simple mode
Weighted median
Weighted mode
2/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 172.48*
Egger; Q: 171.50*
Test hor. pleiotrophy:
Egger int: 0.0072
SNP effect on SLE
SNP effect on exposure (CAD-a)
2/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 67.83*
Egger; Q: 67.19*
Test hor. pleiotrophy:
Egger int: 0.0099
SNP effect on SLE
SNP effect on exposure (CAD-b)
all MR methods n.s.
Heterogeneity tests:
IVW; Q: 43.89*
Egger; Q: 35.57*
Test hor. pleiotrophy:
Egger int: 0.0642
SNP effect on SLE
SNP effect on exposure (MI-a)
2/5 MR methods p<0.05
Heterogeneity tests:
IVW; Q: 107.85*
Egger; Q: 107.82*
Test hor. pleiotrophy:
Egger int: 0.00192
SNP effect on SLE
SNP effect on exposure (MI-b)

Figure S2. Bidirectional MR summaries between SLE and CAD. Related to Figure 2 and Table S3. Scatter plots showing GWAS effect size estimates on the exposure (x-axis) and outcome (y-axis) with each dot representing a SNP and lines representing MR-estimates of SLE on CAD, MI and IS (**A**) and in the reverse direction, with CAD or MI as exposure and SLE as the outcome (**B**). MR-IVW and MR-Egger heterogeneity test results (Q-value) indicate whether significant heterogeneity was detected (asterisks, $p<0.05$), which does not necessarily indicate biased causal estimates. MR-Egger intercept indicates whether significant (asterisks, $p<0.05$) *directional* horizontal pleiotropy was detected, which usually indicates biased causal estimates. N.s., not significant.

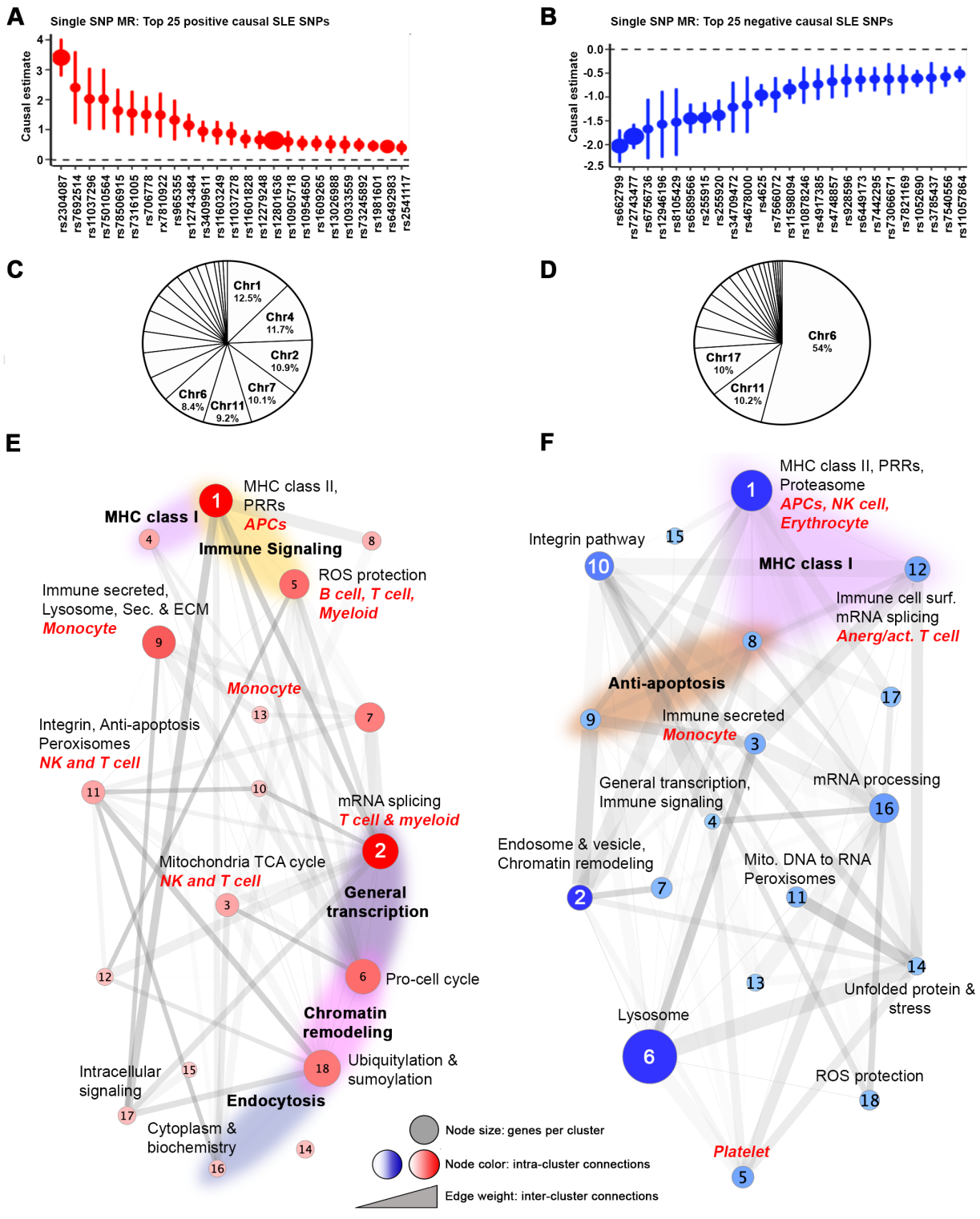

A
Single SNP MR: Top 25 positive causal SLE SNPs
Causal estimate
rs2304087
rs7692514
rs11037296
rs75010564
rs78506915
rs73161005
rs706778
rx7810922
rs965355
rs12743484
rs34099611
rs11603249
rs11037278
rs11601828
rs12279248
rs12801636
rs10905718
rs10954650
rs1609265
rs13026988
rs10933559
rs73245892
rs1981601
rs6492983
rs2541117
B
Single SNP MR: Top 25 negative causal SLE SNPs
Causal estimate
rs662799
rs72743477
rs6756736
rs12946196
rs8105429
rs6589566
rs255915
rs255920
rs34709472
rs4678000
rs4625
rs7566072
rs11598094
rs10878246
rs4917385
rs4748857
rs928596
rs6449173
rs7442295
rs73066671
rs7821169
rs1052690
rs3785437
rs7540556
rs11057864
C
Chr1 12.5%
Chr4 11.7%
Chr2 10.9%
Chr7 10.1%
Chr11 9.2%
Chr6 8.4%
D
Chr6 54%
Chr11 10.2%
Chr17 10%
E
MHC class I
MHC class II, PRRs
APCs
Immune Signaling
ROS protection
B cell, T cell, Myeloid
Immune secreted, Lysosome, Sec. & ECM
Monocyte
Monocyte
Integrin, Anti-apoptosis Peroxisomes
NK and T cell
mRNA splicing
T cell & myeloid
Mitochondria TCA cycle
NK and T cell
General transcription
Pro-cell cycle
Chromatin remodeling
Ubiquitylation & sumoylation
Intracellular signaling
Endocytosis
Cytoplasm & biochemistry
F
MHC class II, PRRs, Proteasome
APCs, NK cell, Erythrocyte
Integrin pathway
MHC class I
Immune cell surf. mRNA splicing
Anerg/act. T cell
Anti-apoptosis
Immune secreted
Monocyte
General transcription, Immune signaling
mRNA processing
Endosome & vesicle, Chromatin remodeling
Mito. DNA to RNA Peroxisomes
Unfolded protein & stress
Lysosome
ROS protection
Platelet
Node size: genes per cluster
Node color: intra-cluster connections
Edge weight: inter-cluster connections

Figure S3. SLE-associated SNPs on chromosome 6 account for the majority of negative causal effects on CAD by SSMR. Related to Figure 3. **A-B**) Forest plots (beta ± standard error) of the top 25 (by absolute value of causal estimates) positive (**A**) and negative (**B**) causal SNPs identified by SSMR using the Wald-ratio method. **C-D**) Pie charts illustrating the distribution of 119 positive (**C**) and 234 negative (**D**) causal SLE SNPs on CAD. **E-F**) Cluster metastructures for the 498 (**E**) 557 (**F**) predicted genes from positive and negative causal SNPs identified by single-SNP MR. Metastructures are based on PPI networks, clustered using MCODE and visualized in Cytoscape. Node gradient shading is proportional to intra-cluster connectivity, cluster size indicates number of genes per cluster and edge weight indicates inter-cluster connections. Functional and cell-type enrichments for each cluster were determined using BIG-C (black labels) and I-scope (red labels), respectively. Bold black labels over colored shadings represent shared functional annotations for the clusters they surround.

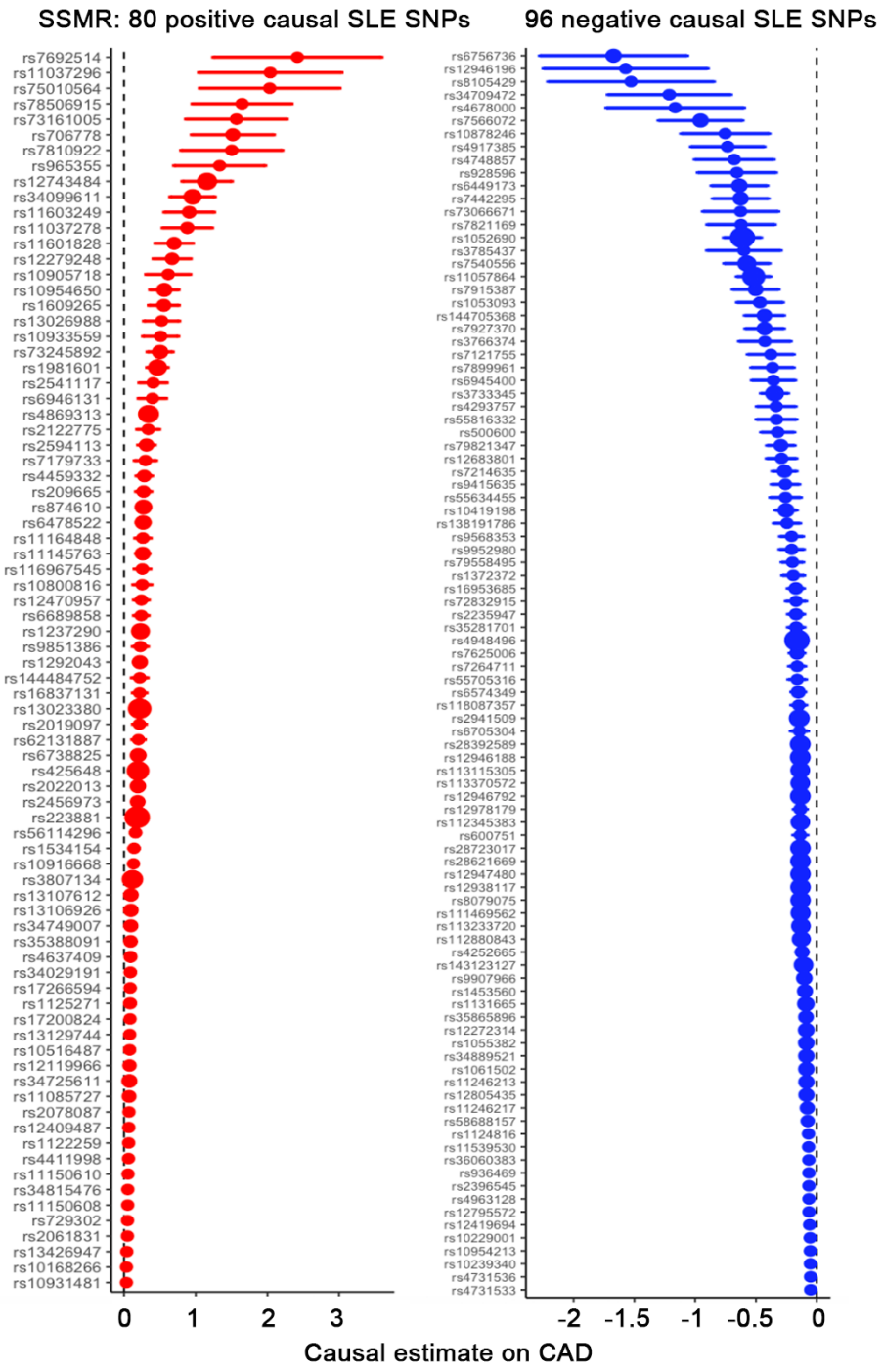
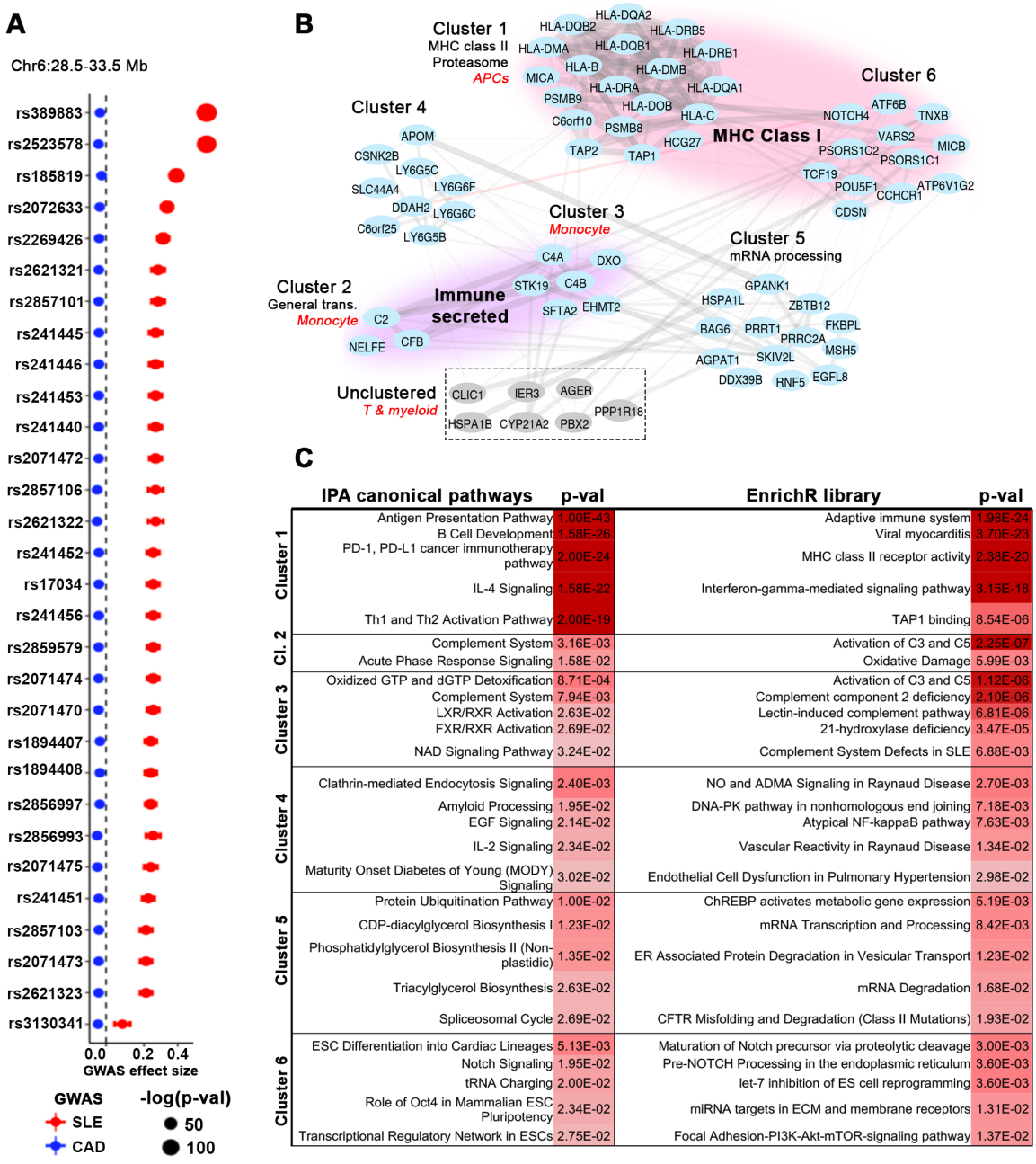
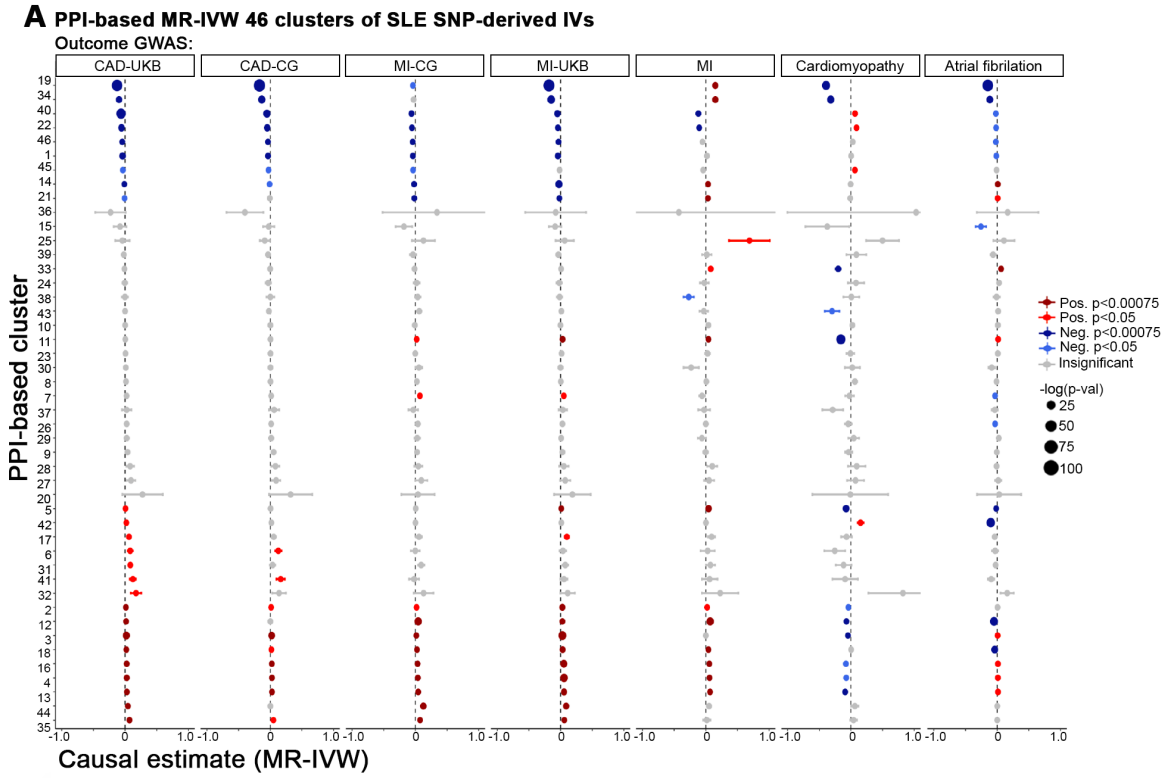


Figure S4. MR analyses for positive and negative causal SNPs determined by SSMR. Related to Figure 3 and Table S3. **A**) Forest plots (beta ± standard error) of the 80 positive (**A**) and 96 negative (**B**) causal non-HLA SNPs identified by SSMR using the Wald ratio method, ordered by absolute value of causal estimates.



	IPA canonical pathways	p-val	EnrichR library	p-val
Cluster 1	Antigen Presentation Pathway	1.00E-43	Adaptive immune system	1.98E-24
	B Cell Development	1.58E-26	Viral myocarditis	3.70E-23
	PD-1, PD-L1 cancer immunotherapy pathway	2.00E-24	MHC class II receptor activity	2.38E-20
	IL-4 Signaling	1.58E-22	Interferon-gamma-mediated signaling pathway	3.15E-18
	Th1 and Th2 Activation Pathway	2.00E-19	TAP1 binding	8.54E-06
Cl. 2	Complement System	3.16E-03	Activation of C3 and C5	2.25E-07
	Acute Phase Response Signaling	1.58E-02	Oxidative Damage	5.99E-03
Cluster 3	Oxidized GTP and dGTP Detoxification	8.71E-04	Activation of C3 and C5	1.12E-06
	Complement System	7.94E-03	Complement component 2 deficiency	2.10E-06
	LXR/RXR Activation	2.63E-02	Lectin-induced complement pathway	6.81E-06
	FXR/RXR Activation	2.69E-02	21-hydroxylase deficiency	3.47E-05
	NAD Signaling Pathway	3.24E-02	Complement System Defects in SLE	6.88E-03
Cluster 4	Clathrin-mediated Endocytosis Signaling	2.40E-03	NO and ADMA Signaling in Raynaud Disease	2.70E-03
	Amyloid Processing	1.95E-02	DNA-PK pathway in nonhomologous end joining	7.18E-03
	EGF Signaling	2.14E-02	Atypical NF-kappaB pathway	7.63E-03
	IL-2 Signaling	2.34E-02	Vascular Reactivity in Raynaud Disease	1.34E-02
	Maturity Onset Diabetes of Young (MODY) Signaling	3.02E-02	Endothelial Cell Dysfunction in Pulmonary Hypertension	2.98E-02
Cluster 5	Protein Ubiquitination Pathway	1.00E-02	ChREBP activates metabolic gene expression	5.19E-03
	CDP-diacylglycerol Biosynthesis I	1.23E-02	mRNA Transcription and Processing	8.42E-03
	Phosphatidylglycerol Biosynthesis II (Non-plastidic)	1.35E-02	ER Associated Protein Degradation in Vesicular Transport	1.23E-02
	Triacylglycerol Biosynthesis	2.63E-02	mRNA Degradation	1.68E-02
	Spliceosomal Cycle	2.69E-02	CFTR Misfolding and Degradation (Class II Mutations)	1.93E-02
Cluster 6	ESC Differentiation into Cardiac Lineages	5.13E-03	Maturation of Notch precursor via proteolytic cleavage	3.00E-03
	Notch Signaling	1.95E-02	Pre-NOTCH Processing in the endoplasmic reticulum	3.60E-03
	tRNA Charging	2.00E-02	let-7 inhibition of ES cell reprogramming	3.60E-03
	Role of Oct4 in Mammalian ESC Pluripotency	2.34E-02	miRNA targets in ECM and membrane receptors	1.31E-02
	Transcriptional Regulatory Network in ESCs	2.75E-02	Focal Adhesion-PI3K-Akt-mTOR-signaling pathway	1.37E-02

Figure S5. Analysis of HLA SNP-predicted genes associated with both SLE and CAD. Related to Figure 3. **A**) Forest plot showing GWAS effect sizes ± standard error for 30 HLA SNPs significantly ($p<10^{-6}$) associated with both SLE (red) and CAD (blue). **B**) PPI network consisting of 69 putative protein-coding genes predicted from the 30 HLA SNPs. Functional and cell-type enrichments for each cluster were determined using BIG-C (black labels) and I-scope (red labels), respectively. Black labels over colored shadings represent shared functional annotations for the clusters they surround. **C**) Gene set enrichments for each cluster were determined using IPA and EnrichR. P-values are from Fisher's exact test that measures the significance of overlap between analysis-ready genes in each cluster and genes within an annotation, with red shading proportional to significance of each enrichment.


A PPI-based MR-IVW 46 clusters of SLE SNP-derived IVs
Outcome GWAS:
CAD-UKB
CAD-CG
MI-CG
MI-UKB
MI
Cardiomyopathy
Atrial fibrilation
PPI-based cluster
Causal estimate (MR-IVW)
Pos. p<0.00075
Pos. p<0.05
Neg. p<0.00075
Neg. p<0.05
Insignificant
-log(p-val)
25
50
75
100

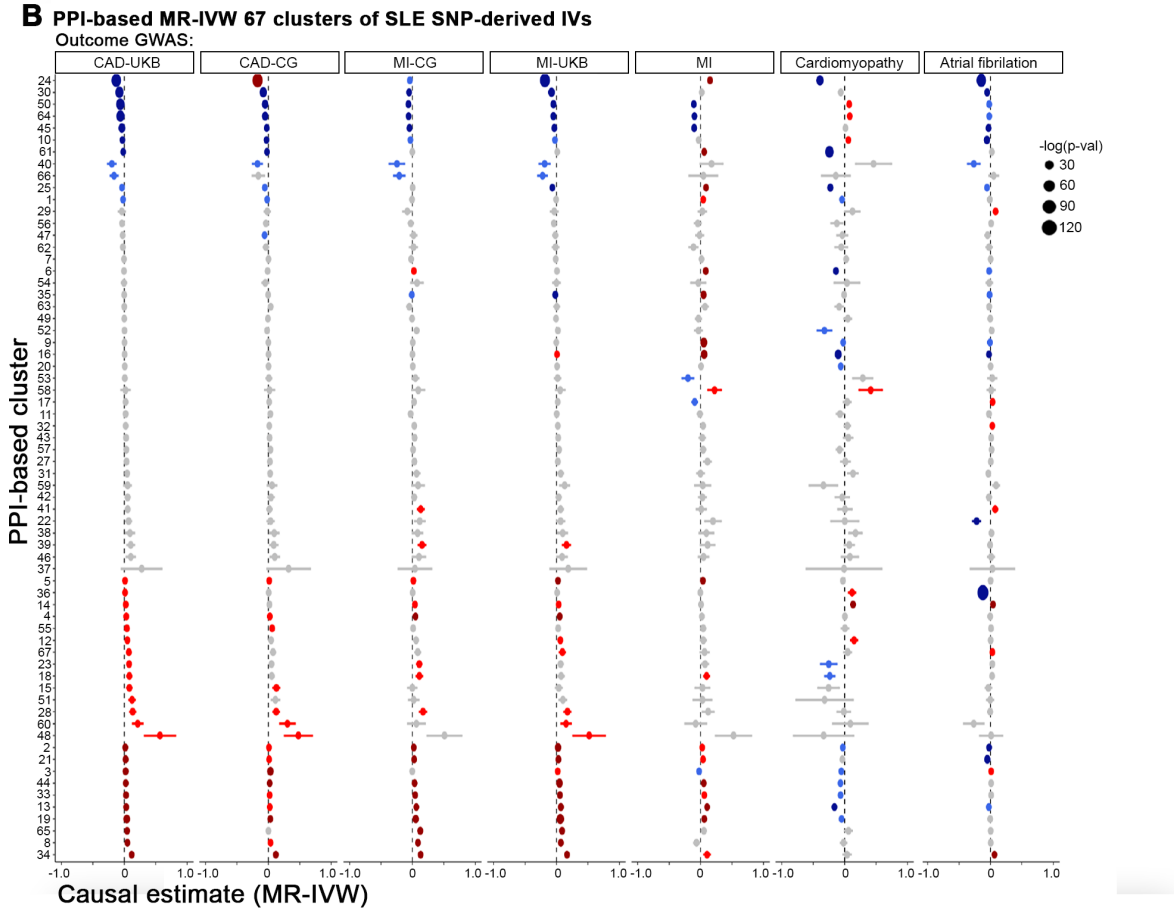

B PPI-based MR-IVW 67 clusters of SLE SNP-derived IVs
Outcome GWAS:
CAD-UKB
CAD-CG
MI-CG
MI-UKB
MI
Cardiomyopathy
Atrial fibrilation
PPI-based cluster
Causal estimate (MR-IVW)
-log(p-val)
30
60
90
120

Figure S6. Positive and negative causal estimates for PPI-based clusters using MR-IVW. Related to Figures 4 and 5. (**A-B**) PPI-based MR-IVW (beta ± standard error) using these the 46 (**A**) and 67 (**B**) clusters of SLE SNP-derived IVs in CAD, MI, IS, cardiomyopathy, and atrial fibrillation GWAS. For results, grey indicates insignificant ($p > 0.05$), dark red and red, positive causal at $p < 0.00075$ and $p < 0.05$, respectively; dark blue and blue, negative causal $p < 0.00075$ and $p < 0.05$, respectively by IVW.

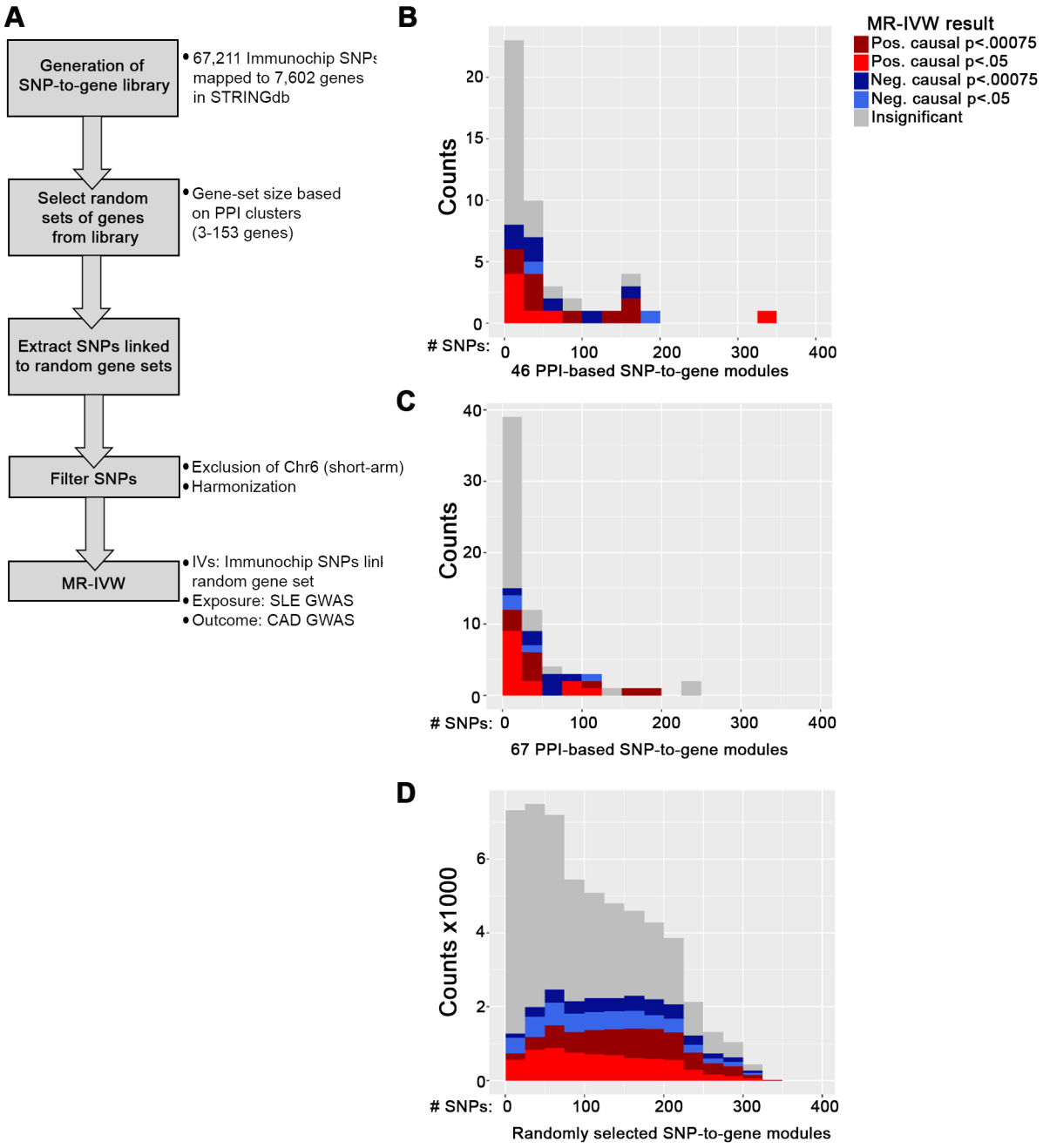


Figure S7. Expected vs. observed MR-IVW casual estimates corresponding to random vs. PPI-based SNP-to-gene modules. Related to Figure 5. **A**) Schematic illustrating the Monte Carlo Simulations for expected MR results using random sets of Immunochip-derived SNP-to-Gene modules. **B-D**) Histograms representing the proportion of insignificant ($p>0.05$, gray), positive causal ($p<0.05$, red), negative causal ($p<0.05$, blue), positive causal ($p<0.00075$, dark red), and negative causal ($p<0.00075$, dark blue) results with respect to number of SNPs used as IVs for SLE-exposure on CAD corresponding to (**B**) the 46 SLE-derived clusters and (**C**) the comprehensive 67 SLE-derived clusters and (**D**) over 50,000 random sets of Immunochip-derived SNP-to-Gene modules.

Figure	Analysis	Purpose	Main findings
1	MR-multiple methods	Exploratory analyses using an expanded SNP repertoire from multiple sources to examine estimated associations between SLE and CAD.	Results suggest a net-positive association for SLE on CAD. Provide justification to confirm estimated association.
2	MR-multiple methods, S-LDSC, pathway analysis	Confirmatory analyses using highly curated IVs, S-LDSC to determine heritability and pathway analysis to examine pathways underlying SLE and causal of CAD.	Majority of MR methods show positive causal estimate on CAD and MI, predicted genes capture SLE heritability and pathways reflect immune and CVD dysfunction.
3	SSMR, S-LDSC, pathway analysis	Orthogonal MR approach to identify single SLE SNPs with positive or negative estimates on CAD.	SSMR identifies positive and negative causal SNPs that capture significant SLE heritability and predict a number of underlying causal and protective pathways.
4	PPI-based MR, S-LDSC, pathway analysis	Development of PPI-based MR (also outlined in the graphical abstract).	PPI-based MR identifies 46 groups of SNPs for use as IVs based on cluster membership. Application of MR-IVW identifies clusters with positive and negative causal estimates.
5	PPI-based MR, S-LDSC, pathways analysis	PPI-based MR using a larger, comprehensive network.	Identification of 67 SNP sets as IVs. Application of MR-IVW identifies clusters with positive and negative causal estimates.
6	MR-IVW	PPI-based MR IV validation after accounting for pleiotropy and LD.	Application of LD clumping to cluster-derived SNPs followed by MR-IVW to confirm clusters with positive and negative causal estimates.
7	Drug matching	Identify new therapeutic interventions.	Pathways linked to positive causal clusters predict novel therapies for managing the inflammatory environment contributing to CAD in SLE.

Table S10. Summary of major findings. Related to STAR Methods.

Click here to access/download

Supplemental Videos and Spreadsheets

CAD Supplementary Data Tables_SEPT26FINAL-2.xlsx

genes

MDPI

Article

Nucleic Acid-Sensing and Interferon-Inducible Pathways Show Differential Methylation in MZ Twins Discordant for Lupus and Overexpression in Independent Lupus Samples: Implications for Pathogenic Mechanism and Drug Targeting

Miranda C. Marion [1,2], Paula S. Ramos [3], Prathyusha Bachali [4], Adam C. Labonte [4], Kip D. Zimmerman [2], Hannah C. Ainsworth [1,2], Sarah E. Heuer [4,5], Robert D. Robl [4], Michelle D. Catalina [4], Jennifer A. Kelly [6], Timothy D. Howard [7], Peter E. Lipsky [4], Amrie C. Grammer [4] and Carl D. Langefeld [1,2,*]

1 Department of Biostatistics and Data Science, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA; mimarion@wakehealth.edu (M.C.M.); hainswor@wakehealth.edu (H.C.A.)
2 Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA; kdzimmer@wakehealth.edu
3 Division of Rheumatology and Immunology, Department of Medicine, Medical University of South Carolina, Charleston, SC 29425, USA; ramosp@musc.edu
4 AMPEL BioSolutions, LLC and RILITE Research Institute, Charlottesville, VA 22902, USA; prathyusha.bachali@ampelbiosolutions.com (P.B.); adam.labonte@ampelbiosolutions.com (A.C.L.); sarah.heuer@tufts.edu (S.E.H.); robert.robl@ampel.org (R.D.R.); michellecatalina@ampel.org (M.D.C.); peterlipsky@ampelbiosolutions.com (P.E.L.); amriegrammer@ampelbiosolutions.com (A.C.G.)
5 The Jackson Laboratory, Tufts Graduate School of Biomedical Sciences, Bar Harbor, ME 04609, USA
6 Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA; Jennifer-Kelly@omrf.org
7 Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA; tdhoward@wakehealth.edu
* Correspondence: clangefe@wakehealth.edu

check for updates

Citation: Marion, M.C.; Ramos, P.S.; Bachali, P.; Labonte, A.C.; Zimmerman, K.D.; Ainsworth, H.C.; Heuer, S.E.; Robl, R.D.; Catalina, M.D.; Kelly, J.A.; et al. Nucleic Acid-Sensing and Interferon-Inducible Pathways Show Differential Methylation in MZ Twins Discordant for Lupus and Overexpression in Independent Lupus Samples: Implications for Pathogenic Mechanism and Drug Targeting. *Genes* **2021**, *12*, 1898. https://doi.org/10.3390/genes12121898

Academic Editors: F. Yesim Demirci and Timothy B. Niewold

Received: 29 October 2021
Accepted: 25 November 2021
Published: 26 November 2021

Abstract: Systemic lupus erythematosus (SLE) is a chronic, multisystem, autoimmune inflammatory disease with genomic and non-genomic contributions to risk. We hypothesize that epigenetic factors are a significant contributor to SLE risk and may be informative for identifying pathogenic mechanisms and therapeutic targets. To test this hypothesis while controlling for genetic background, we performed an epigenome-wide analysis of DNA methylation in genomic DNA from whole blood in three pairs of female monozygotic (MZ) twins of European ancestry, discordant for SLE. Results were replicated on the same array in four cell types from a set of four Danish female MZ twin pairs discordant for SLE. Genes implicated by the epigenetic analyses were then evaluated in 10 independent SLE gene expression datasets from the Gene Expression Omnibus (GEO). There were 59 differentially methylated loci between unaffected and affected MZ twins in whole blood, including 11 novel loci. All but two of these loci were hypomethylated in the SLE twins relative to the unaffected twins. The genes harboring these hypomethylated loci exhibited increased expression in multiple independent datasets of SLE patients. This pattern was largely consistent regardless of disease activity, cell type, or renal tissue type. The genes proximal to CpGs exhibiting differential methylation (DM) in the SLE-discordant MZ twins and exhibiting differential expression (DE) in independent SLE GEO cohorts (DM-DE genes) clustered into two pathways: the nucleic acid-sensing pathway and the type I interferon pathway. The DM-DE genes were also informatically queried for potential gene–drug interactions, yielding a list of 41 drugs including a known SLE therapy. The DM-DE genes delineate two important biologic pathways that are not only reflective of the heterogeneity of SLE but may also correlate with distinct IFN responses that depend on the source, type, and location of nucleic acid molecules and the activated receptors in individual patients. Cell- and tissue-specific analyses will be critical to the understanding of genetic factors dysregulating the nucleic acid-sensing and IFN pathways and whether these factors could be appropriate targets for therapeutic intervention.

Keywords: epigenetics; nucleic acid sensing; methylation; gene expression; SLE; drug repositioning; RIG-I; lupus

1. Introduction

Systemic lupus erythematosus (SLE) is a chronic and severe systemic autoimmune disease characterized by the over-production of autoantibodies and heterogeneous clinical manifestations. With more than 100 risk loci identified, a genetic etiology for SLE is unequivocal [1–8]. In fact, the cumulative effect of these risk loci is substantial; the odds ratio (OR) for SLE in individuals of European ancestry is 30 when comparing individuals with the highest 10% of risk allele genetic load (i.e., polygenetic risk score—the weighted count of the number of risk alleles) to individuals in the lowest 10% of genetic load [6]. Despite the strong genetic contribution to risk, the concordance rate between monozygotic (MZ) twins ranges between 24–35%, suggesting that much of the risk remains unexplained and highlighting the potential importance of epigenetic and environmental factors in SLE susceptibility [9].

There is compelling evidence that epigenetic mechanisms, such as 5′ Cytosine methylation, are involved in the pathogenesis of SLE. For example, promoter demethylation at multiple genes in T cells treated with DNA demethylating agents are sufficient to cause lupus in animal models [10]. In recent years, several studies have investigated DNA methylation in SLE patients on a genome-wide scale. The earliest of these genome-wide studies interrogated 27,578 CpG sites in 12 SLE patients and 12 healthy controls using the Illumina Infinium HumanMethylation27 Beadchip, and identified 336 differentially methylated genes, the majority of which were hypomethylated in the cases relative to the controls [11]. Subsequent studies have examined genome-wide methylation in larger samples of SLE patients using the HumanMethylation450 Beadchip (>485,000 CpG sites) in a number of cell types, including naïve CD4+ T cells [12–16], memory and regulatory T cells [17], CD19+ B cells [17], CD14+ monocytes [14,17], granulocytes [14], neutrophils [18], and whole blood or peripheral blood mononuclear cells (PBMC) [19–25]. Differential methylation has not only been observed when comparing SLE patients to healthy controls, but similar patterns have been identified in SLE patients with nephritis [12,19,22], skin involvement [13], specific antibodies [20], and pediatric SLE [26]. The primary and consistent finding across all these studies has been hypomethylation of interferon-regulated genes across various cell types in cases, regardless of SLE disease activity [27].

The analysis of phenotypically discordant MZ twins represents the ideal design by which to assess the role of epigenetic variation in disease etiology and trait heritability while controlling for genetic background [28] and has revealed the existence of differentially methylated regions associated with several autoimmune diseases, including SLE [29], type 1 diabetes [30], psoriasis [31], and ulcerative colitis [32]. To date, the only previously published twin methylation study in SLE that exclusively used MZ twins quantified DNA methylation in white blood cells from 15 discordant MZ twin pairs at 1505 CpG sites in 807 genes using the Illumina GoldenGate Methylation Cancer Panel I [29]. Here, we performed a genome-wide analysis of DNA methylation in a discovery cohort of MZ twins discordant for SLE. The discovery cohort consisted of three twin pairs of European descent, and methylation was measured in whole blood using Illumina's HumanMethylation450 Beadchip. The two strongest associated signals were validated using pyrosequencing. Findings from the discovery cohort were replicated in an independent set of MZ twins from Denmark. We then evaluated gene expression data from multiple cell types and kidney biopsies from 10 independent SLE cohorts to identify genes proximal to CpGs exhibiting differential methylation (DM) in the SLE-discordant MZ twins and exhibiting differential expression (DE) in independent SLE GEO cohorts (DM-DE genes) for pathway analyses. Together, the methylation, gene expression, and pathway analyses uncovered

two separable yet complimentary molecular pathways of lupus pathogenesis, shedding light on potential drug repositioning opportunities and novel therapeutic targets for SLE.

2. Materials and Methods

2.1. Discovery Cohort

Genomic DNA was extracted from peripheral blood of three female MZ twin pairs of European ancestry discordant for SLE enrolled in the Lupus Family Registry and Repository (LFRR) [33]. All cases met ACR classification criteria for SLE [34].

2.2. Replication Cohort

An SLE study of 15 twin pairs from Denmark, assayed on the HumanMethylation450 Beadchip, in monocytes, CD4+ T cells, CD19+ B cells, and granulocytes, was published in 2018 by Ulff-Moller et al. [14]. These data were downloaded from the Gene Expression Omnibus (GEO, accession no. GSE110607), and all available female MZ twin pairs discordant for SLE were retained for analysis (4 twin pairs). The publication states that of these four female MZ twin pairs discordant for SLE, two of the non-SLE twins had other autoimmune diseases, including Sjogren's syndrome, systemic sclerosis, autoimmune thyroiditis, and primary biliary cirrhosis. However, this clinical information was not available in GEO.

2.3. Genome-Wide DNA Methylation Assay and Array Validation in LFRR Twins

Genomic DNA (1µg) from each individual was treated with sodium bisulfite using the EZ 96-DNA methylation kit (Zymo Research, Irvine, CA, USA), following the manufacturer's standard protocol. Genome-wide DNA methylation was assessed using the Illumina Infinium HumanMethylation450 BeadChip (Illumina, Inc., San Diego, CA, USA), which interrogates over 485,500 CpG sites that cover 99% of RefSeq genes (including the promoter, 5'UTR, first exon, gene body, and 3'UTR), as well as 96% of CpG islands and island shores. Arrays were processed using the manufacturer's standard protocol, with both members of each twin pair being hybridized to the same row on the microarray to minimize batch effects. GenomeStudio software (Illumina, Inc.) was used to perform initial quality control and to calculate the relative methylation level of each interrogated cytosine, which is reported as a β-value given by the ratio of the normalized signal from the methylated probe to the sum of the normalized signals of the methylated and unmethylated probes. This β-value for each CpG site ranges from 0 (unmethylated) to 1 (fully methylated). CpG loci with a stringent detection p-value $> 1.0 \times 10^{-5}$ in any of the samples were excluded (n = 2118 probes) to control for poor-quality assays. Validation of the array data in the LFRR twins was performed by pyrosequencing two of the most significant CpGs probes: cg13304609 (in IFI44L) and cg23570810 (in IFITM1). The correlations between the methylation proportions from the array and pyrosequencing for these two probes were $r^2 = 0.98$ and $r^2 = 0.99$, respectively.

2.4. Collection of Gene Expression Experiments from SLE Patient Datasets

Raw data were downloaded from 10 publicly available gene expression datasets (Supplemental Table S1). Only datasets from female lupus patients were analyzed. Active SLE was defined as a Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) > 6 [35]. This has become the standard threshold for disease activity in recent clinical trials of SLE.

2.5. Data Analysis

To identify differentially methylated genes between unaffected and SLE-affected twins, a paired t-test on the probe-specific β-values was computed separately for the discovery and replication twin datasets. For the discovery set, CpG sites meeting (1) the Benjamini–Hochberg False Discovery Rate (FDR) [36] threshold $P_{FDR} < 0.05$ (equivalent to $p < 1.06 \times 10^{-7}$) and (2) a mean DNA methylation difference of $(\Delta\beta) > |0.085|$ were considered statistically significant; the mean methylation difference threshold was obtained by maximizing the area under the receiver operator characteristic curve (AUC) as a function of

the β-value (described below). The genes related to the differentially methylated CpG sites (as annotated by Illumina for the HumanMethylation450) were queried in the Interferome online database to identify interferon-regulated genes [37]. In addition, significant CpG sites were investigated for evidence of association between DNA methylation pattern and gene expression (mQTL) using the iMETHYL genome browser [38]. These results are based on 100 healthy subjects with RNA-seq data and DNA methylation data in CD4T cells, monocytes, and PBMC.

Statistical analysis of the expression data was completed using the following R packages available from Bioconductor: GEOquery, affy, affycoretools, simpleaffy, gcrma, LIMMA, and GSVA. Non-normalized arrays were first inspected for visual artifacts and poor RNA hybridization using Affymetrix QC plots. Principal component (PC) plots were generated for all cell types in each experiment to identify outliers. After removing outliers, the datasets were normalized using the gcrma package (available in Bioconductor [39], www.bioconductor.org) resulting in log2 intensity values for the R expression set objects (denoted E-sets); an E-set combines several information types in a single structured object: an expression value matrix, phenotypic metadata corresponding to individual samples (phenoData), annotation data describing each feature (probeset) of a microarray platform (featureData), as well as other separate metadata matrices describing the experimental protocol and array platform design. To increase the probability of identifying differentially expressed genes (DE genes), the analyses were completed using normalized datasets prepared using both the native Affymetrix chip definition file (CDF), as well as custom BrainArray Entrez CDFs. Illumina CDFs were used for GSE49454.

The CDF-annotated E-sets were filtered to remove probes with very low intensity values by computing the mean log2 values for each probe across all samples and removing those in the lower half of the range of mean values from the expression set (E-set). Probes missing gene annotation data were also discarded. GCRMA normalized expression values were variance-corrected using local empirical Bayesian shrinkage before calculation of differential expression using the ebayes function in the Bioconductor limma package [40]. The resulting p-values were adjusted for multiple hypothesis testing using Benjamini–Hochberg False Discovery Rate (FDR) [36]. Significant Affymetrix and BrainArray probes within each study were merged and filtered to retain DE probes with a PFDR < 0.2. This list was filtered to retain only the most significant probe per gene.

To identify DM-DE genes, we used a logistic regression model (expression fold change as a binary outcome > 0 versus < 0) to determine cell-type specific thresholds for the difference in the β-value that maximized the area under the ROC curve (AUC) predicting increased differential expression (Figure 1A, Supplemental Figure S1). These thresholds were determined by calculating the area under the receiver operating characteristic curve (AUC) across points at regular intervals between 0 and −0.15 and selecting the values that maximized the AUC. Primary inferences are based on thresholds, which included a logFC in expression > 0 and a mean difference in $\beta < -0.085$, -0.055, -0.08, and -0.055 in whole blood, monocytes, B cells, and T cells, respectively. Figure 1A displays these thresholds as vertical bars. For clarity, genes with differential methylation *p*-values greater than 0.0001 and a mean DNA methylation difference of $(\Delta\beta) > |0.025|$ have been removed from Figure 1A.

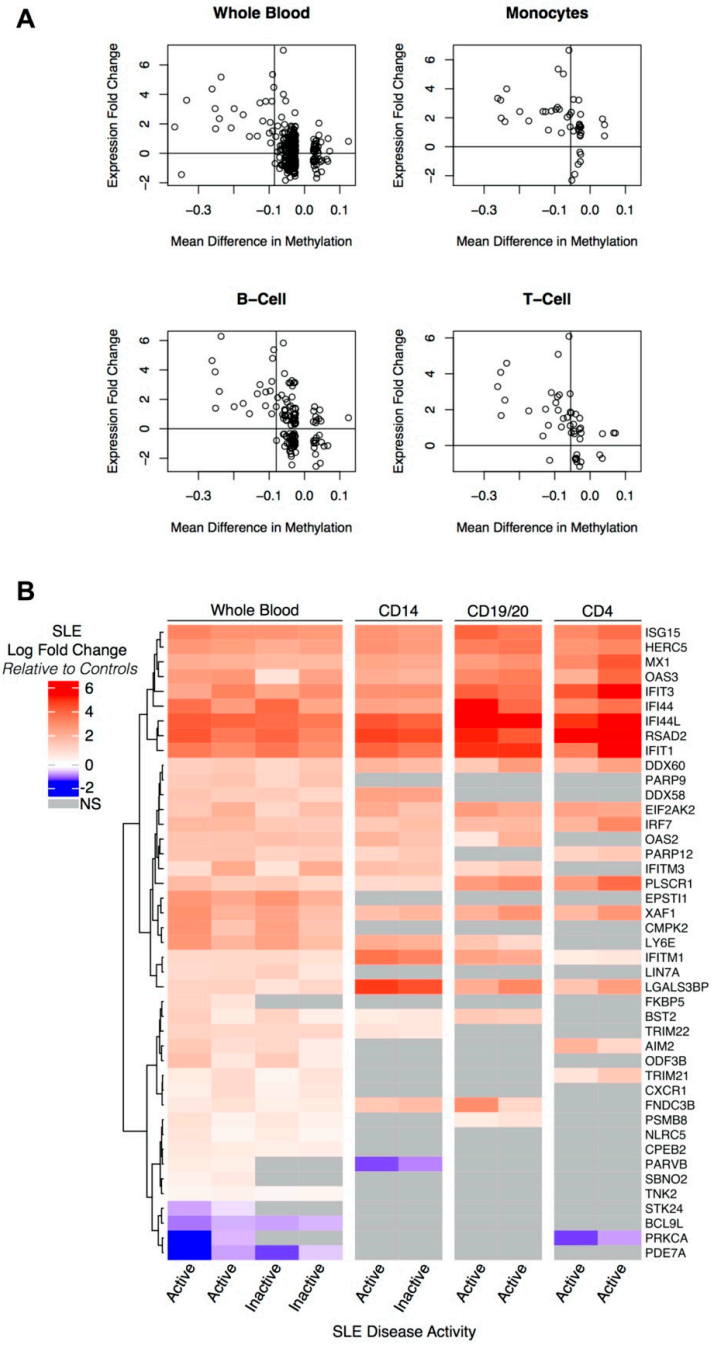


Figure 1. Hypomethylated genes showing differential expression in independent SLE cohorts. (**A**) Specific thresholds for the difference in the β-value (from the discordant twin methylation experiment in whole blood) that maximize the area under the ROC curve predicting increased differential expression in the independent SLE whole blood experiments (GSE39088, GSE49454), monocytes (GSE38351), B-cells (GSE10325, GSE4588), and T cells (GSE10325, GSE51997) are shown as vertical bars. Genes with differential methylation *p*-values greater than 0.0001 and a mean DNA methylation difference of $(\Delta\beta) > |0.025|$ have been removed from the plots. (**B**) Heatmap of 43 genes hypomethylated in the discordant twin data ($\Delta\beta < -0.085$) and differentially expressed between controls and active (SLEDAI $\geq$ 6) or inactive (SLEDAI < 6) lupus patients from two whole blood experiments, monocytes, B cells, and T cells. Hierarchical clustering was performed across rows with Euclidean distance metric and complete linkage. Blue/red gradient represents the log fold change values in lupus patients compared to controls.

The DM-DE genes were analyzed in a pathway analysis using the MCODE [41] clustering algorithm and STRING networking scores [42].

Protein–drug interaction networks were generated for each DM-DE gene individually via STITCH [43], Ingenuity Pathway Analysis (IPA) (Qiagen Bioinformatics: ingenuity.com), and the Drug–Gene Interaction database [44]. Drugs were denoted as (1) known utility in lupus therapy, (2) FDA-approved compound, (3) currently involved in a clinical trial (not necessarily SLE), and (4) generally regarded as safe (GRAS) compounds. Using a hypothesis-driven ranking of the therapeutic potential for SLE applications of specific drugs or compounds, the combined lupus treatment scoring (CoLTS) scores (range −16 to +11) were calculated [45].

3. Results

3.1. Characteristics of the MZ Twins

The LFRR MZ twins were all females of European ancestry, and the SLE-diagnosed twins exhibited a range of SLE clinical conditions (Supplemental Table S2). The Danish MZ twins were also all females of European ancestry. Clinical characteristics such as number of ACR criteria, SLEDAI score, autoantibodies, and medications are described in Ulff-Moller et. al., but were not available in GEO [14].

3.2. Identification of Differentially Methylated Regions in Twins Discordant for SLE

Of the 485,577 CpG sites passing quality control metrics, 59 sites in 33 genes met both a $P_{FDR} < 0.05$ (equivalent to a non-FDR $p < 1.06 \times 10^{-7}$) and a mean DNA methylation difference of $(\Delta\beta) > |0.085|$ (Table 1). Only two of these significant CpG sites showed increased methylation in the affected twins (hypermethylation), while the remaining 57 exhibited lower methylation (hypomethylation). Of the 33 genes represented in Table 1, 22 are regulated at some level by type I interferons (as defined by Interferome [37]). Eleven genes are novel to our study and have not been previously reported as SLE-related in a genome-wide methylation study, five of which are unrelated to the typical interferon signature (*LY6G5C, CXCR1, ATOH8, CACNA1D, MECOM*). Lymphocyte antigen 6 complex, locus G5C (*LY6G5C*), is located within the major histocompatibility complex class III region and codes for a protein associated with the cell membrane by a glycosylphosphatidylinositol linkage and involved in signal transduction [46]. Chemokine (C-X-C motif) receptor 1 (*CXCR1*) encodes for a protein that is a receptor for interleukin 8. Genetic and expression variation in *CXCR1* have been correlated with infections (e.g., active tuberculosis, hepatitis B, *Candida albicans*) and modestly with SLE [6,47–50]. Atonal bHLH transcription factor 8 (*ATOH8*), calcium voltage-gated channel subunit alpha1 D (*CACNA1D*), and MDS1, and EVI1 complex locus (*MECOM*) do not have known links to autoimmune disease or infections. Given the gender bias in SLE, it is interesting to note that none of the differentially methylated probes meeting our significance criteria were located on the X chromosome.

We next examined the 59 differentially methylated CpGs from the discovery cohort (Table 1) in the Danish twin replication cohort. Even with the probable dampening effect generated by two of the Danish non-SLE twins having other autoimmune diseases, we observed very high concordance in the direction of the $\Delta\beta$ values. Specifically, 55 (93%), 54 (92%), 52 (88%), and 54 (92%) of the 59 differentially methylated CpG sites in the LFRR twins were concordant in the Danish twins' monocytes, CD4+ T cells, CD19+ B cells, and granulocytes, respectively. Furthermore, 35, 26, 32, and 33 of the 59 CpG sites were statistically significant (p-value < 0.05) and directionally concordant in the monocyte, CD4+ T cell, CD19+ B cell, and granulocyte expression datasets, respectively; only one of these was statistically significant in the opposite direction (p-value < 0.05; Additional File 1). Thus, the Danish twin data strongly corroborated the global pattern of methylation observed in the LFRR twin data.

Table 1. Differentially methylated probes from three monozygotic twin pairs discordant for SLE.

CpG *	Chr	Pos (bp) †	Gene	Δβ				*p*-Value	Interferon-Regulated ‡	Relation to CpG ††
				Pair1	Pair2	Pair3	Mean			
cg13304609	1	79085162	*IFI44L*	−0.24	−0.27	−0.37	−0.29	1.58×10^{-14}	IRG	
cg06872964	1	79085250	*IFI44L*		−0.26	−0.21	−0.24	1.05×10^{-71}	IRG	
cg03607951	1	79085586	*IFI44L*	−0.27	−0.3	−0.21	−0.26	7.23×10^{-22}	IRG	
cg17515347	1	159047163	*AIM2*	−0.09	−0.11	−0.07	−0.09	3.01×10^{-12}	IRG	
cg08272268	1	200380059	*ZNF281*	−0.08	−0.07	−0.11	−0.09	4.33×10^{-15}		S_Shore
cg01028142	2	7004578	*CMPK2*	−0.22	−0.36	−0.43	−0.33	7.98×10^{-8}	IRG	N_Shore
cg10959651	2	7018020	*RSAD2*	−0.13	−0.1	−0.16	−0.13	3.14×10^{-14}	IRG	
cg10549986	2	7018153	*RSAD2*	−0.08	−0.09	−0.1	−0.09	1.95×10^{-91}	IRG	
cg14126601	2	37384708	*EIF2AK2*	−0.08	−0.1	−0.12	−0.1	5.55×10^{-16}	IRG	S_Shore
cg26337070	2	85999873	*ATOH8*	−0.06	−0.12	−0.11	−0.1	7.55×10^{-9}		
cg04781494	2	202047246	*CASP10*	−0.07	−0.13	−0.08	−0.09	8.39×10^{-8}	IRG	
cg15768138	2	219030752	*CXCR1*	−0.09	−0.12	−0.11	−0.11	7.38×10^{-27}		
cg13411554	3	53700276	*CACNA1D*	−0.06	−0.12	−0.09	−0.09	8.66×10^{-8}		
cg22930808	3	122281881	*PARP9-DTX3L*	−0.36	−0.34	−0.4	−0.37	6.74×10^{-126}	IRG	N_Shore
cg08122652	3	122281939	*PARP9-DTX3L*	−0.34	−0.31	−0.51	−0.38	1.11×10^{-9}	IRG	N_Shore
cg00959259	3	122281975	*PARP9-DTX3L*	−0.37	−0.3	−0.34	−0.34	1.32×10^{-56}	IRG	N_Shore
cg06981309	3	146260954	*PLSCR1*	−0.24	−0.28	−0.21	−0.24	6.41×10^{-31}	IRG	N_Shore
cg02556393	3	168866705	*MECOM*	−0.08	−0.09	−0.1	−0.09	3.14×10^{-95}		N_Shore
cg07809027	4	15007205	*CPEB2*	−0.07	−0.1	−0.12	−0.1	2.08×10^{-14}		S_Shore
cg02215171	4	89379156	*HERC5*	−0.08	−0.09	−0.11	−0.09	4.48×10^{-18}	IRG	S_Shore
cg17786255	4	108814389	*SGMS2*	−0.07	−0.09	−0.11	−0.09	2.01×10^{-16}	IRG	
cg21873524	4	190942744		−0.1	−0.1	−0.12	−0.11	1.03×10^{-55}		Island
cg24740632	5	134486678		−0.11	−0.12	−0.14	−0.12	2.26×10^{-60}		
cg06012695	6	28770593			−0.1	−0.13	−0.11	3.59×10^{-16}		
cg25138053	6	31368016		−0.11	−0.09	−0.07	−0.09	3.67×10^{-15}		S_Shore
cg22708150	6	31649619	*LY6G5C*	−0.12	−0.14	−0.17	−0.14	1.05×10^{-19}		N_Shore
cg07292773	6	156718177		0.07	0.1	0.11	0.1	2.22×10^{-17}		Island
cg12013713	7	139760671	*PARP12*	−0.12	−0.14	−0.09	−0.12	1.44×10^{-16}	IRG	N_Shore
cg20190772	8	48572496	*KIAA0146*	−0.08	−0.07	−0.13	−0.09	1.40×10^{-8}		
cg14864167	8	66751182	*PDE7A*	−0.25	−0.35	−0.45	−0.35	1.21×10^{-9}		N_Shelf
cg06102678	8	81491328		−0.08	−0.12	−0.07	−0.09	1.00×10^{-8}		Island
cg12110437	8	144098888	*LY6E*	−0.16	−0.17	−0.27	−0.2	3.14×10^{-9}	IRG	N_Shore
cg17555806	10	74448117		−0.08	−0.12	−0.07	−0.09	1.51×10^{-8}		N_Shelf
cg02314339	10	91020653		−0.08	−0.14	−0.11	−0.11	1.72×10^{-8}		
cg06188083	10	91093005	*IFIT3*	−0.29	−0.16	−0.31	−0.25	6.18×10^{-8}	IRG	
cg05552874	10	91153143	*IFIT1*	−0.2	−0.28	−0.3	−0.26	6.01×10^{-16}	IRG	
cg14910175	10	131840954		−0.07	−0.11	−0.08	−0.09	1.56×10^{-11}		N_Shelf
cg10552523	11	313478	*IFITM1*	−0.14	−0.12	−0.14	−0.13	5.90×10^{-115}	IRG	N_Shelf
cg20566897	11	313527	*IFITM1*	−0.11	−0.11	−0.09	−0.1	7.00×10^{-62}	IRG	N_Shelf
cg23570810	11	315102	*IFITM1*	−0.24	−0.25	−0.34	−0.27	1.43×10^{-18}	IRG	N_Shore
cg03038262	11	315262	*IFITM1*	−0.24	−0.22	−0.29	−0.25	4.41×10^{-40}	IRG	N_Shore
cg20045320	11	319555		−0.19	−0.13	−0.2	−0.18	4.85×10^{-17}		S_Shore
cg17990365	11	319718	*IFITM3*	−0.16	−0.15	−0.15	−0.16	8.78×10^{-295}	IRG	S_Shore
cg08926253	11	614761	*IRF7*	−0.15	−0.14	−0.23	−0.17	2.01×10^{-9}	IRG	Island
cg12461141	11	5710654	*TRIM22*	−0.1	−0.08	−0.12	−0.1	6.35×10^{-25}	IRG	
cg23571857	17	6658898	*XAF1*	−0.07	−0.13	−0.11	−0.1	1.46×10^{-8}	IRG	
cg04927537	17	76976091	*LGALS3BP*	−0.14	−0.11	−0.2	−0.15	2.77×10^{-10}	IRG	
cg25178683	17	76976267	*LGALS3BP*	−0.15	−0.11	−0.21	−0.16	2.01×10^{-8}	IRG	
cg16503797	18	19476805		−0.08	−0.12	−0.08	−0.09	5.39×10^{-12}		N_Shore
cg15871086	18	56526595		−0.07	−0.11	−0.08	−0.09	2.08×10^{-11}		N_Shelf
cg23352030	20	62198469	*PRIC285*	0.13	0.19	0.11	0.14	2.36×10^{-11}		Island

Table 1. *Cont.*

CpG *	Chr	Pos (bp) †	Gene	Δβ				*p*-Value	Interferon-Regulated ‡	Relation to CpG ††
				Pair1	Pair2	Pair3	Mean			
cg16785077	21	42791867	*MX1*	−0.11	−0.09	−0.12	−0.11	8.45×10^{-27}	IRG	N_Shore
cg22862003	21	42797588	*MX1*	−0.31	−0.25	−0.35	−0.31	1.62×10^{-25}	IRG	N_Shore
cg26312951	21	42797847	*MX1*	−0.26	−0.17	−0.2	−0.21	6.28×10^{-15}	IRG	N_Shore
cg21549285	21	42799141	*MX1*	−0.5	−0.35	−0.57	−0.47	6.59×10^{-13}	IRG	S_Shore
cg05543864	22	24979755	*GGT1*	−0.08	−0.08	−0.1	−0.09	1.44×10^{-45}		
cg20098015	22	50971140	*ODF3B*	−0.19	−0.22	−0.21	−0.21	9.88×10^{-83}	IRG	S_Shore
cg05523603	22	50973101		−0.17	−0.23	−0.27	−0.22	5.51×10^{-14}		S_Shelf
cg02247863	22	50983415		−0.07	−0.1	−0.11	−0.09	2.51×10^{-13}		N_Shore

* CpGs meeting the $P_{FDR} < 0.05$ threshold (equivalent to $p < 1.06 \times 10^{-7}$) and having $|\Delta\beta| > 0.085$. † Positions are from Build 37. ‡ IRG as defined by Interferome [37]. †† Island: CpG sites > 200 bp, with GC content > 55% and observed to expected ratio > 0.6. N_shore: 0–2 kb upstream from island; S-shore 0–2 kb downstream from island; N_shelf 2–4 kb upstream from island; S_shelf 2–4 kb downstream from island.

We also sought to determine if the dominating presence of the interferon signature might have masked more modest signals from other individual (non-IFN) loci. After regressing out the mean β-value (methylation value) for the most significant CpG site in each interferon-regulated gene in Table 1 (as defined by Interferome [37]), no additional CpG sites across the genome met an FDR threshold of significance ($P_{FDR} > 0.05$).

We considered the genomic context of the CpG sites showing aberrant methylation in the LFRR MZ twins. Here, a CpG island was defined as a cluster of CpG sites of greater than 200 bp, with GC content >55%, and the observed-to-expected (under mathematical independence of the Gs and Cs) ratio >0.6 [51]. Interestingly, out of 59 CpG sites differentially methylated, the majority (54%, $n = 32$) were located in a CpG shore (0–2 kb from island) or shelf (2–4 kb from island), whereas only 8% ($n = 5$) were located in a CpG island (Table 1). This is in contrast to the composition of the 450k chip in which about one third of the CpG sites reside in islands (Supplemental Figure S2). Notably, the only two hypermethylated CpG sites (relative to the unaffected twin) meeting our significance thresholds reside in CpG islands.

3.3. Hypomethylated Genes Are Overexpressed in Independent Cohorts

Methylation at CpG sites influences gene expression. Thus, linking differential methylation to changes in gene expression by showing that the same genes were associated with SLE in both types of experiments (even in independent samples) would provide further evidence of the importance of these genes and could identify potential actionable mechanisms.

Genes harboring a CpG site with $\Delta\beta < -0.085$ and $p < 0.01$ (for differential methylation) were tested for differential expression in whole blood from two independent cohorts, each comparing SLE patients to healthy controls (GSE39088 and GSE49454) (Table 2). Relative to controls, overexpression was observed in both active and inactive SLE patients within almost all of these genes, and the level of expression was highly correlated within the gene expression experiments (experiment 1, $r = 0.95$; experiment 2, $r = 0.99$). *IFI44L, RADS2,* and *IFIT1* showed the highest fold changes and comparable increases in expression in active and inactive SLE patients; *IFI44L* is noteworthy as it has been reported to be predictive of SLE status relative to healthy controls and other autoimmune diseases [52]. Cohorts with expression data derived from monocytes (GSE38351), CD19+, and CD20+ B cells (GSE10325, GSE4588), and CD4+ T cells (GSE10325, GSE51997) reflected a consistent pattern of increased expression in genes meeting the mean (methylation) $\Delta\beta$ threshold of −0.085 (Figure 1B). Upon extending $\Delta\beta$ to <−0.055, the statistically appropriate threshold for detecting differential expression in monocytes and T cells in our dataset (see Methods), an additional 54 hypomethylated genes were evaluated in the gene expression datasets (Supplemental Table S3). Overall, the pattern of differential expression of hypomethylated genes was very similar across the cell subtypes examined (Figure 1B, Supplemental Table S3). Thus, the differential expression results in independent cohorts in

multiple cell types provide a multi-omic, independent pseudo-replication, and translational interpretation of the methylation results (Table 2).

Hierarchical clustering (Euclidean distance, complete linkage) of the DM-DE genes using the log fold change (LFC) identified a cluster of nine genes with markedly higher LFC (Figure 1B). This cluster shows a consistent pattern across whole blood, monocytes, B cells, and T cells, as well as in both active and inactive SLE disease. In fact, the LFC remained largely consistent between active and inactive disease across all DM-DE genes. Exceptions to this pattern include FK506 binding protein 5 (*FKBP5*), parvin beta (*PARVB*), and strawberry notch homolog 2 (*SBNO2*) in whole blood, where there is upregulation in active patients and non-significant change in inactive patients. This pattern was not replicated in any of the individual cell types.

Table 2. Differential expression of hypomethylated genes in whole blood from two independent SLE cohorts.

							Active SLE [§]		Inactive SLE [§]	
CpG *	**Chr**	**Pos (bp)** [†]	**Gene**	**Mean** $\Delta\beta$	**Methylation** p**-Value**	**Interferon-Regulated** [‡]	**Log FC Expt 1**	**Log FC Expt 2**	**Log FC Expt 1**	**Log FC Expt 2**
cg16526047	1	949893	*ISG15*	−0.11	1.28×10^{-4}	IRG	3.1	2.77	2.74	2.59
cg05696877	1	79088769	*IFI44L*	−0.3	6.60×10^{-6}	IRG	3.98	3.8	3.64	3.4
cg01079652	1	79118191	*IFI44*	−0.34	5.34×10^{-4}	IRG	3.54	2.53	3.7	2.33
cg17515347	1	159047163	*AIM2*	−0.09	3.01×10^{-12}	IRG	1.39	0.86	1.08	0.49
cg01028142	2	7004578	*CMPK2*	−0.33	7.98×10^{-8}	IRG	2.76	1.5	2.43	1.51
cg10959651	2	7018020	*RSAD2*	−0.13	3.14×10^{-14}	IRG	4.04	3.32	3.76	3.04
cg14126601	2	37384708	*EIF2AK2*	−0.1	5.55×10^{-16}	IRG	1.47	2.02	1.08	1.68
cg15768138	2	219030752	*CXCR1*	−0.11	7.38×10^{-27}		0.43	0.96	0.38	0.66
cg08122652	3	122281939	*PARP9-DTX3L*	−0.38	1.11×10^{-9}	IRG	1.36	1.56	1.07	1.55
cg06981309	3	146260954	*PLSCR1*	−0.24	6.41×10^{-31}	IRG	1.77	1.25	1.38	1.07
cg02694620	3	172109284	*FNDC3B*	−0.11	3.80×10^{-3}		0.57	0.82	0.41	0.52
cg15065340	3	195632915	*TNK2*	−0.16	4.04×10^{-3}		0.22	0.31	0.2	0.25
cg07809027	4	15007205	*CPEB2*	−0.1	2.08×10^{-14}		0.66	0.52	0.42	0.45
cg02215171	4	89379156	*HERC5*	−0.09	4.48×10^{-18}	IRG	2.62	2.48	2.14	2.36
cg05883128	4	169239131	*DDX60*	−0.25	2.13×10^{-5}	IRG	1.24	1.38	1.06	1.46
cg08099136	6	32811251	*PSMB8*	−0.11	1.43×10^{-4}	IRG	−0.39	−0.13	NS	NS
cg00052684	6	35694245	*FKBP5*	−0.16	1.65×10^{-3}		1.11	0.71	NS	NS
cg05994974	7	139761087	*PARP12*	−0.15	6.89×10^{-5}	IRG	1.52	1.57	1.14	1.25
cg14864167	8	66751182	*PDE7A*	−0.35	1.21×10^{-9}		−1.24	−0.41	−0.82	−0.23
cg12110437	8	144098888	*LY6E*	−0.2	3.14×10^{-9}	IRG	2.66	1.92	2.43	1.7
cg03848588	9	32525008	*DDX58*	−0.1	4.34×10^{-4}	IRG	1.48	1.3	1.32	1.07
cg06188083	10	91093005	*IFIT3*	−0.25	6.18×10^{-8}	IRG	2.25	3.15	2.3	2.87
cg05552874	10	91153143	*IFIT1*	−0.26	6.01×10^{-16}	IRG	3.39	2.94	3.42	2.81
cg23570810	11	315102	*IFITM1*	−0.27	1.43×10^{-18}	IRG	1	1.03	1.03	0.81
cg17990365	11	319718	*IFITM3*	−0.16	8.78×10^{-295}	IRG	0.92	2.23	0.71	2.13
cg08926253	11	614761	*IRF7*	−0.17	2.01×10^{-9}	IRG	1.84	1.79	1.4	1.37
cg08577913	11	4415193	*TRIM21*	−0.1	1.74×10^{-3}	IRG	0.56	0.93	0.28	0.75
cg12461141	11	5710654	*TRIM22*	−0.1	6.35×10^{-25}	IRG	1.14	1	0.99	1.05
cg26811705	11	118781408	*BCL9L*	−0.09	1.64×10^{-3}		−0.6	−0.35	−0.41	−0.32
cg19347790	12	81332050	*LIN7A*	−0.09	1.87×10^{-4}		0.93	0.99	1.24	0.61
cg25800166	12	113375896	*OAS3*	−0.13	5.36×10^{-5}	IRG	2.52	2.69	0.73	2.35
cg19371652	12	113415883	*OAS2*	−0.11	2.24×10^{-5}	IRG	1.48	1.56	1.64	1.53
cg03753191	13	43566902	*EPSTI1*	−0.1	9.23×10^{-5}	IRG	2.65	2.26	2.71	2.02
cg00246969	13	99159656	*STK24*	−0.11	6.26×10^{-6}		0.81	0.32	0.66	0.36
cg07839457	16	57023022	*NLRC5*	−0.23	6.10×10^{-6}	IRG	0.7	0.23	0.53	0.27
cg23571857	17	6658898	*XAF1*	−0.1	1.46×10^{-8}	IRG	2.85	1.96	2.35	1.68
cg23378941	17	64361956	*PRKCA*	−0.11	6.89×10^{-5}	IRG	−1.11	−0.3	NS	NS
cg25178683	17	76976267	*LGALS3BP*	−0.16	2.0×10^{-8}	IRG	1.16	1.21	0.72	1.05
cg07573872	19	1126342	*SBNO2*	−0.15	2.77×10^{-3}	IRG	0.38	0.58	NS	NS
cg07839313	19	17514600	*BST2*	−0.12	3.48×10^{-3}	IRG	1.24	0.49	1.17	0.41
cg21549285	21	42799141	*MX1*	−0.47	6.59×10^{-13}	IRG	2.12	2	1.86	1.79
cg19460508	22	44422195	*PARVB*	−0.1	1.64×10^{-3}		0.54	0.39	NS	NS
cg20098015	22	50971140	*ODF3B*	−0.21	9.88×10^{-83}	IRG	1.61	0.61	1.36	0.47

Differential gene expression values come from GSE39088 (Expt 1) and GSE49454 (Expt 2) in whole blood of lupus patients compared with controls. * CpGs with $p < 0.01$ and $|\Delta\beta| > 0.085$. [†] Positions are from Build 37. [‡] As defined by Interferome [37]. [§] Active disease is defined as ≥6 on the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) [35].

Although only one of the three affected MZ twins in the discovery cohort had renal involvement, almost all of the genes mapping to differentially methylated CpG sites showed overexpression in both the kidney glomerulus and tubulointerstitium from independent lupus nephritis patients (Table 3). In the glomerulus, 28 genes were overexpressed, 2 were under expressed, and 14 were not significantly differentially expressed in lupus nephritis samples compared to healthy controls. In the tubulointerstitium, 27 were overexpressed, 5 under expressed, and 12 not significantly differentially expressed. *IFI44L*, *MX1*, and *IFI44* showed the highest levels of overexpression across the two tissues. The fold change was correlated between the two tissues ($r = 0.66$, $p < 0.0001$).

Significant DNA methylation sites were further investigated for evidence of association between DNA methylation at a specific CpG site and gene expression (eQTM) using the iMETHYL genome browser with data on 100 healthy Japanese subjects with RNA-seq data and DNA methylation data in CD4T cells, monocytes, and PBMC [38] (Supplemental Table S4). Most of the CpGs from Table 1 that are identified in iMETHYL are eQTMs for the gene in which they reside. In contrast, some are eQTMs for additional genes of interest. For example, cg17515347 is in physical proximity to *AIM1*, which has an important role in T cell regulation in autoimmune diseases. However, this CpG site is also an eQTM for five other genes in CD4+ T cells (*TAGLN2, SLAMF8, DUSP23, PHYIN1 FCRL6*), several of which have established autoimmune disease connections. Transgelin-2 may help regulate activation and migration of B cells in lymph node follicles, exhibits increased expression in B cells from lymph nodes in SLE patients, and appears important in host defense [53,54]. SLAM family member 8 (*SLAMF8*) is a member of the SLAM family of genes of which several members have been associated with multiple autoimmune diseases [55]. FcR-like 6 (*FCRL6*), a receptor that binds to major histocompatibility complex (MHC) class II HLA-DR, is expressed in B cells and has a tyrosine-based immunoregulatory function [56,57]. Dual-specificity protein phosphate 23 (*DUSP23*) expression is reportedly higher in CD4+ T cells from SLE patients compared to healthy controls [58]. Thus, DNA methylation in these regions, and potentially others, may have a complex and multifaceted impact on autoimmunity. Annotation of cg20098015 on chromosome 22 is linked to Outer Dense Fiber of Sperm Tails 3 (*ODF3B*). However, this CpG is an eQTM for SCO2 homolog, mitochondrial and SCO cytochrome oxidase deficient homolog 2 (*SCO2*), and thymidine phosphorylase (*TYMP*), both involved in mitochondrial functions.

3.4. Pathway Analysis of DM-DE Genes

Pathway, clustering, and networking analyses were completed to elucidate patterns among the DM-DE genes. Ingenuity Pathway Analysis (IPA) identified two primary canonical pathways: (1) interferon signaling and (2) pattern recognition receptor (PRR) (Figure 2A). The overlap *p*-value, which tests for independence between known targets of each transcription regulator in a pathway and the list of genes provided, shows very strong association for these two pathways. Other significant pathways of note include the activation of interferon regulatory factors (IRFs) by pattern recognition receptors, retinoic acid-inducible gene I protein (RIG-I)-like receptors in innate immunity, and NF-κB activation by viruses. Figure 2B illustrates the IFN signaling pathway determined by IPA. Notably, in this pathway all of the DM-DE genes are downstream, and none were identified as upstream signaling molecules. IPA also identified 39 upstream regulators ($|\text{Z-score}| \geq 2$) of the DM-DE genes that showed differential expression between SLE cases and controls in whole blood (Figure 2C).

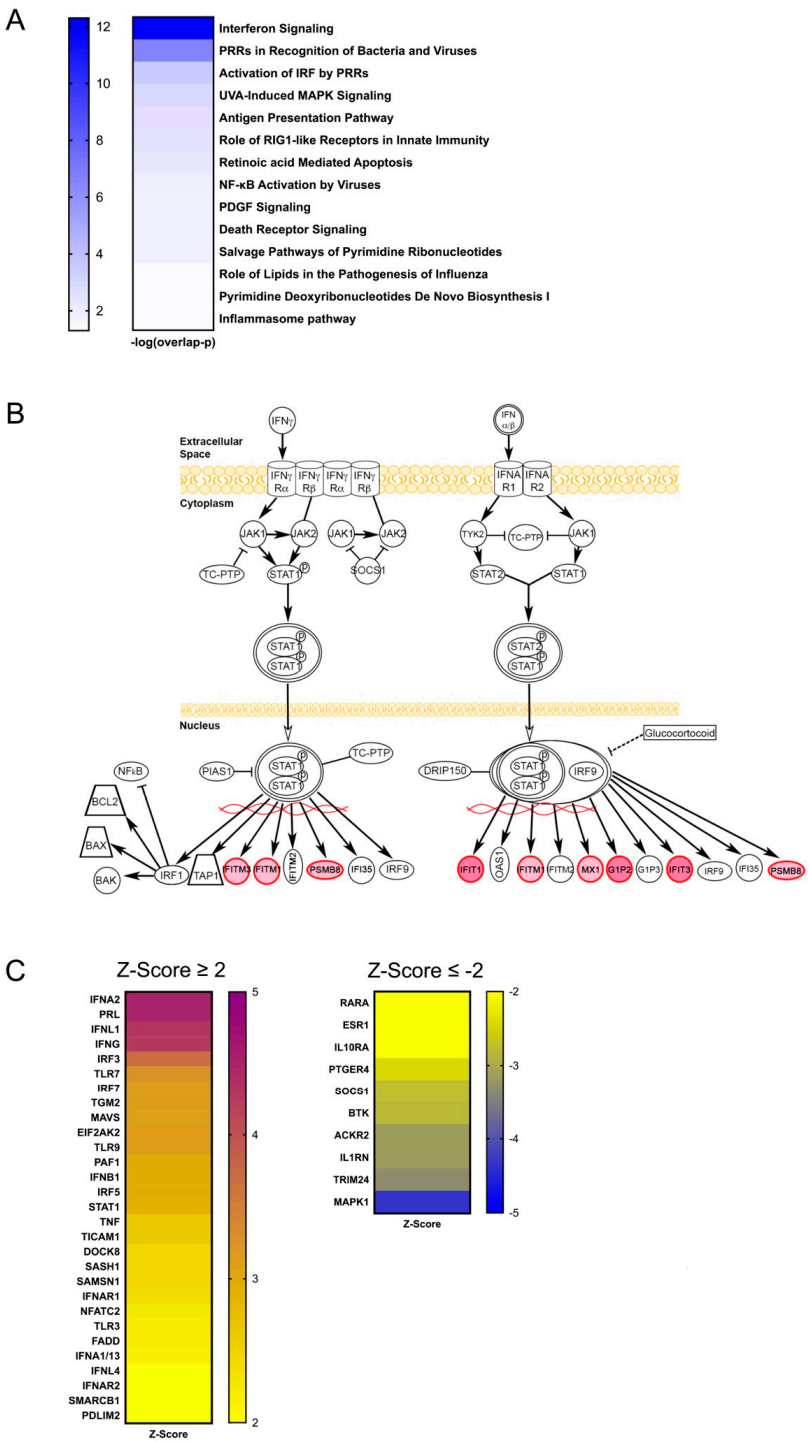


Figure 2. Pathway analyses of hypomethylated genes showing differential expression in independent SLE cohorts. (**A**) List and statistical significance of the overlap of the IPA canonical pathways comprised of hypomethylated genes showing differential expression in whole blood of independent SLE patients. (**B**) IPA canonical IFN signaling of hypomethylated genes showing differential expression (increased expression in SLE cases in red) in whole blood of independent SLE patients. (**C**) Activation Z-scores of genes predicted as upstream regulators of genes hypomethylated in the discordant twin data ($\Delta\beta < -0.085$) and differentially expressed in whole blood between independent SLE cases and controls. A positive (negative) Z-score indicates that a regulator has significantly more (fewer) activated predictions than inhibited predictions.

Table 3. Differential expression of hypomethylated genes in kidney biopsies from independent SLE patients with lupus nephritis.

CpG *	Chr	Pos (bp) †	Gene	Mean $\Delta\beta$	Methylation *p*-Value	Interferon-Regulated ‡	Log FC Glomerulus	Log FC Tubulointerstitium
cg16526047	1	949893	*ISG15* ‖	−0.11	1.28×10^{-4}	IRG	3.32	4.7
cg05696877	1	79088769	*IFI44L* Δ	−0.3	6.60×10^{-6}	IRG	5.14	5.94
cg01079652	1	79118191	*IFI44* ‖	−0.34	5.34×10^{-4}	IRG	3.94	4.76
cg17515347	1	159047163	*AIM2*	−0.09	3.01×10^{-12}	IRG	0.58	NS
cg01028142	2	7004578	*CMPK2* Δ	−0.33	7.98×10^{-8}	IRG	NS	NS
cg10959651	2	7018020	*RSAD2*	−0.13	3.14×10^{-14}	IRG	4.31	3.36
cg14126601	2	37384708	*EIF2AK2* Δ	−0.1	5.55×10^{-16}	IRG	1.54	1.72
cg15768138	2	219030752	*CXCR1*	−0.11	7.38×10^{-27}		0.68	−0.17
cg08122652	3	122281939	*PARP9-DTX3L* Δ	−0.38	1.11×10^{-9}	IRG	NS	NS
cg06981309	3	146260954	*PLSCR1* Δ	−0.24	6.41×10^{-31}	IRG	1.92	2.07
cg02694620	3	172109284	*FNDC3B*	−0.11	3.80×10^{-3}		NS	0.47
cg15065340	3	195632915	*TNK2* ‖	−0.16	4.04×10^{-3}		0.38	−0.4
cg07809027	4	15007205	*CPEB2*	−0.1	2.08×10^{-14}		NS	NS
cg02215171	4	89379156	*HERC5* ¶	−0.09	4.48×10^{-18}	IRG	3.16	1.96
cg05883128	4	169239131	*DDX60*	−0.25	2.13×10^{-5}	IRG	1.11	2.31
cg08099136	6	32811251	*PSMB8*	−0.11	1.43×10^{-4}	IRG	0.76	2.51
cg00052684	6	35694245	*FKBP5* §	−0.16	1.65×10^{-3}		−1.27	−2.77
cg05994974	7	139761087	*PARP12* ‖	−0.15	6.89×10^{-5}	IRG	2.26	1.86
cg14864167	8	66751182	*PDE7A*	−0.35	1.21×10^{-9}		NS	NS
cg12110437	8	144098888	*LY6E* ◊	−0.2	3.14×10^{-9}	IRG	1.28	1.23
cg03848588	9	32525008	*DDX58* ◊	−0.1	4.34×10^{-4}	IRG	2.89	2.59
cg06188083	10	91093005	*IFIT3* ◊	−0.25	6.18×10^{-8}	IRG	2.59	3.14
cg05552874	10	91153143	*IFIT1* ◊	−0.26	6.01×10^{-16}	IRG	2.24	2.77
cg23570810	11	315102	*IFITM1*	−0.27	1.43×10^{-18}	IRG	2.24	3.29
cg17990365	11	319718	*IFITM3*	−0.16	8.78×10^{-295}	IRG	2.24	2
cg08926253	11	614761	*IRF7* ‖	−0.17	2.01×10^{-9}	IRG	2.8	1
cg08577913	11	4415193	*TRIM21*	−0.1	1.74×10^{-3}	IRG	1.35	0.77
cg12461141	11	5710654	*TRIM22*	−0.1	6.35×10^{-25}	IRG	1.73	2.86
cg26811705	11	118781408	*BCL9L*	−0.09	1.64×10^{-3}		NS	NS
cg19347790	12	81332050	*LIN7A*	−0.09	1.87×10^{-4}		NS	−0.57
cg25800166	12	113375896	*OAS3*	−0.13	5.36×10^{-5}	IRG	3.77	1.1
cg19371652	12	113415883	*OAS2*	−0.11	2.24×10^{-5}	IRG	4.86	1.74
cg03753191	13	43566902	*EPSTI1* ¶	−0.1	9.23×10^{-5}	IRG	NS	NS
cg00246969	13	99159656	*STK24*	−0.11	6.26×10^{-6}		NS	0.28
cg07839457	16	57023022	*NLRC5*	−0.23	6.10×10^{-6}	IRG	NS	NS
cg23571857	17	6658898	*XAF1*	−0.1	1.46×10^{-8}	IRG	3.14	3.05
cg23378941	17	64361956	*PRKCA*	−0.11	6.89×10^{-5}	IRG	−0.48	−0.08
cg25178683	17	76976267	*LGALS3BP*	−0.16	2.0×10^{-8}	IRG	0.57	1.49
cg07573872	19	1126342	*SBNO2*	−0.15	2.77×10^{-3}	IRG	NS	NS
cg07839313	19	17514600	*BST2* ‖	−0.12	3.48×10^{-3}	IRG	NS	2.91
cg21549285	21	42799141	*MX1* Δ	−0.47	6.59×10^{-13}	IRG	4.05	4.64
cg19460508	22	44422195	*PARVB*	−0.1	1.64×10^{-3}		0.28	NS
cg20098015	22	50971140	*ODF3B*	−0.21	9.88×10^{-83}	IRG	NS	NS

Differential gene expression values come from GSE32591: kidney glomerulus and tubulointerstitium WHO class 3/4 lupus nephritis versus control samples. NS indicates not significant FDR *p*-value > 0.2). * CpGs with $p < 0.01$ and $\Delta\beta < -0.085$. † Positions are from Build 37. ‡ As defined by Interferome [37]. § SLE patients show decreased expression in both kidney tissues. ‖ Hypomethylation of this gene at the same CpG site has been reported in SLE patients with renal involvement [12]. ¶ Hypomethylation of this gene at a different CpG site has been reported in SLE patients with renal involvement [12]. Δ Hypomethylation of this gene at the same CpG site has been reported in SLE patients with and without renal involvement [12]. ◊ Hypomethylation of this gene at a different CpG site has been reported in SLE patients with renal involvement [12].

The DM-DE genes were further analyzed in an additional pathway analysis using the MCODE clustering algorithm and STRING networking scores. Two distinct yet related clusters emerged (Figure 3). As expected, there was an enrichment of genes in the IFN-inducible/pattern recognition receptor pathway. As visually represented by the colors of the nodes and node outlines in Figure 3, all genes in this cluster were upregulated in both active and inactive SLE patients; all of these except *PARP9* were overexpressed in both kidney tissues.

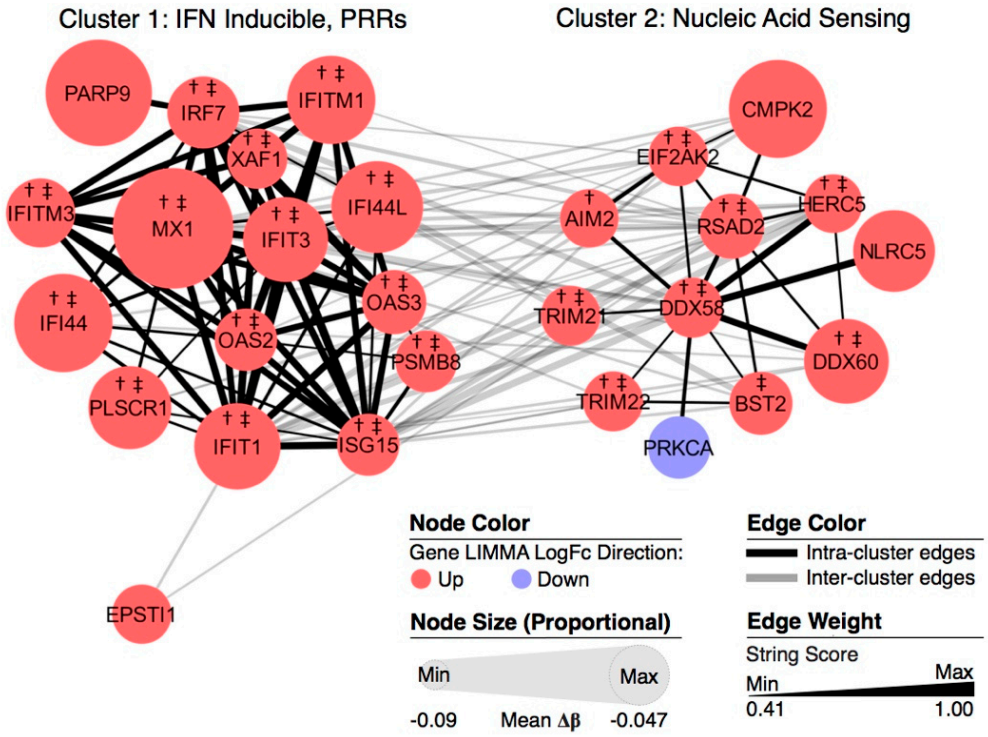


Figure 3. MCODE clustering of hypomethylated genes showing differential expression in independent SLE cohorts. A network scoring degree cutoff of 2, node score cutoff of 0.2, k-Core of 2, and a max depth of 100 were applied. Node color indicates log2(FC) direction and node size is inversely scaled with Δβ (larger nodes are more strongly hypomethylated). Edge weight is scaled by STRING protein–protein connectivity score. All upregulated genes present in clusters were also upregulated in inactive SLE WB samples. †, upregulated in kidney glomerulus, WHO class 3/4. ‡, upregulated in kidney tubulointerstitium, WHO class 3/4.

The second cluster was comprised of genes involved in the nucleic acid-sensing pathway, a primary antiviral defense in vertebrates as well as a mechanism to respond to intracellular nucleic acids of cellular origin. There were strong links among the genes in these two clusters as this nucleic acid response of the innate immune system results in the production of type 1 interferon (i.e., INF-α and INF-β) and expression of interferon stimulated genes [59]. These hypomethylated genes showed increased expression in both active and inactive SLE patients; the lone exception observed was the reduced expression of *PRKCA* in active SLE patients. As in the IFN-inducible/pattern recognition receptor pathway, the majority of these nucleic acid-sensing pathway genes were expressed in both kidney tissues. The gene DEAD H-box helicase 58 (*DDX58*), which encodes for retinoic acid-inducible gene I (*RIG-I*) [60], was the central node and exhibited the strongest and most numerous links to other genes within the cluster.

3.5. Potential Drug Targets

The DM-DE genes were analyzed for potential gene–drug interactions (Table 4). As evidence of its potential utility, this approach identified methotrexate, a lupus therapy, targeting *EPSTI1*. Twelve of the DM-DE genes are linked to drugs that are currently in ongoing clinical trials, primarily trials related to cancer (Table 4). The drug target analysis also identified 24 additional FDA-approved drugs linked to genes associated with the nucleic acid-sensing or the interferon-inducible pathways. These drugs could merit careful consideration for future clinical trials in SLE.

Table 4. Predicted drugs targeting hypomethylated genes and associated pathways with $\Delta\beta < -0.085$.

CpG *	Chr	Pos(bp) †	Gene	Mean Δβ	*p*-Value	STITCH [43]	IPA ‡	DGIdb [44]
cg16526047	1	949893	*ISG15*	−0.11	1.28×10^{-4}			Irinotecan [F]
cg10959651	2	7018020	*RSAD2*	−0.13	3.14×10^{-14}	Fludarabine [F]		
cg14126601	2	37384708	*EIF2AK2*	−0.1	5.55×10^{-16}			Indirubin derivative E804
cg15768138	2	219030752	*CXCR1*	−0.11	7.38×10^{-27}	Reparixin [D]	Reparixin [D]	SCH-527123, Ketoprofen [F]
cg06981309	3	146260954	*PLSCR1*	−0.24	6.41×10^{-31}	Wogonin [G]		
cg15065340	3	195632915	*TNK2*	−0.16	4.04×10^{-3}	Dasatinib^{-1} [F]	Osimertinib [F], Vemurafenib[F]	Debromohymenialdisine
cg08099136	6	32811251	*PSMB8*	−0.11	1.43×10^{-4}	Carfilzomib^{4} [F], Oprozomib [D], Bortezomib^{6} [F]	Carfilzomib^{4} [F]	Carfilzomib^{4} [F],
cg00052684	6	35694245	*FKBP5*	−0.16	1.65×10^{-3}	Rapamycin/ Sirolimus^{2} [F], Tacrolimus^{5} [F]		Venlafaxine [F], Clomipramine [F]
cg14864167	8	66751182	*PDE7A*	−0.35	1.21×10^{-9}			Ketotifen [F], Dyphylline [F]
cg12110437	8	144098888	*LY6E*	−0.2	3.14×10^{-9}		DLYE5953A[D]	
cg06188083	10	91093005	*IFIT3*	−0.25	6.18×10^{-8}	Imidazoles [D]		
cg08926253	11	614761	*IRF7*	−0.17	2.01×10^{-9}	Hesperidin [D]		
cg03753191	13	43566902	*EPSTI1*	−0.1	9.23×10^{-5}	Methotrexate [F T], Vinblastine [F], Doxorubicin [F], Cisplatin [F]		
cg00246969	13	99159656	*STK24*	−0.11	6.26×10^{-6}	Staurosporine [D]		
cg23378941	17	64361956	*PRKCA*	−0.11	6.89×10^{-5}	Staurosporine [D]	Aprinocarsen	Midostaurin [F], Enzastaurin [D], Quercetin [D G], Aprinocarsen, Ruboxistaurin [D], Ingenol Mebutate [FW], Bryostatin [D], Sotrastaurin Acetate [D], Tamoxifen^{2} [F]
cg07839313	19	17514600	*BST2*	−0.12	3.48×10^{-3}	Resveratrol^{6} [D G]		
cg21549285	21	42799141	*MX1*	−0.47	6.59×10^{-13}	Mitomycin C [F], Colchicine [F]		
cg19460508	22	44422195	*PARVB*	−0.1	1.64×10^{-3}	Lovastatin^{3} [F]		Bortezomib^{6} [F]

* CpGs with $p < 1 \times 10^{-3}$ and $\Delta\beta < -0.085$. † Positions are from Build 37. ‡ Qiagen Bioinformatics: ingenuity.com [F] FDA approved. [D] Ongoing clinical trial or DiD [G] GRAS. [T] Known utility in lupus therapy. [FW] Ingenol mebutate is FDA-approved in the US but withdrawn in the EU. Numbers in superscript are CoLTS scores and range from −16 to +11.

4. Discussion

Environmental challenges coupled with genetic susceptibility are often hypothesized to cause the innate and adaptive immune system to become chronically active, causing failure to recognize subsequent autoimmune disease [61]. Aging and environmental exposures such as smoking, chemicals, diet, and viral pathogens predictably trigger methylation or demethylation at CpG sites. Altered methylation of a CpG site changes the accessibility of transcriptional elements to specific regions, which leads to regulation of gene expression. The relationship between DNA methylation and gene expression is complex, including being influenced by specific tissues/cells [62–64]. However, in general, DNA methylation in promoter regions is often inversely correlated with gene expression. The above paradigm is consistent with the results of this multi-omic study, which has demonstrated that genes involved in the nucleic acid-sensing and interferon-inducible pathways were observed to be hypomethylated in SLE-affected MZ twins and upregulated in independent SLE cohorts. Despite the clear biological importance of tissue-specific methylation and gene expression, here, the high concordance of hypomethylated genes in whole blood with increased gene expression across a variety of tissues from multiple independent cohorts suggests a high fidelity of the DNA methylation-gene expression relationship at these loci.

Every epigenome-wide study of SLE to date, including this one, has identified hypomethylation of multiple type I IFN-related genes. While there is no doubt that stimulation of the type I IFN pathway is important in SLE, the mechanism by which this stimulation occurs will be unique for each SLE patient. Interferon induction occurs due to activation of one of several types of pattern recognition receptors, which are programmed to respond to double-stranded DNA (dsDNA), double-stranded RNA (dsRNA), or single-stranded RNA (ssRNA). The type of nucleic acid (NA) present will depend on the species and cell type producing the NA. Furthermore, the NA may leak into the cytosome where its recognition is again specific to the receptor activated. In our study, bioinformatic analysis identified the NA-sensing pathway, with DEAD/H-Box helicase 58 (*DDX58*) as the central node (Figure 3). *DDX58* encodes for retinoic acid-inducible gene I (*RIG-I*), which recognizes ssRNA. In contrast to Toll-like receptors (TLRs), which recognize NAs in the endosome, RIG-I-like receptors (RLRs) interact with mitochondrial antiviral signaling protein (*MAVS*) in the cytosol [65]. *MAVS* subsequently phosphorylates interferon regulatory factors 3 (*IRF3*) to stimulate type 1 IFN expression. The NA-sensing pathway generated by our analysis also included absent in melanoma 2 (*AIM2*), a cytosolic dsDNA-sensing protein that activates the inflammasome, further emphasizing the plausible role of this pathway in initiating lupus inflammation [66,67].

The cascade of functional consequences resulting from genetic variation and unique environmental exposures will differ for each individual SLE patient. While some SLE patients (10–30%) will present no IFN signature [68], others will overexpress IFN through one of the several mechanisms described above. The DM-SE gene list we prioritized may be a useful tool in grouping SLE patients into DA receptor groups, or "endotypes" as they have been termed by Mustelin et al. [68] Therapies targeting helicases such as *RIG-I*, *MAV*S, or *AIM2* could prove useful for SLE. One such inhibitor of *RIG-I*, enhancer of zeste homolog 2 (*EZH2*), has been shown to play an epigenetic role in SLE and was proposed as a therapeutic target by Tsou et al. [60]. Network analyses and public database queries of our DM-DE genes yielded a list of genes whose products predict gene–drug interactions. The resulting list includes methotrexate, a drug used for the treatment of lupus. The remaining gene–drug interactions we identified merit thorough scrutiny as they could be candidates for future trials.

Three recent studies have observed aberrant methylation of IFN genes in SLE patients with renal involvement [12,19,22]. A summary of the literature (Additional File 2) shows our study's consistencies with these published findings. While hypomethylation of these genes has been confirmed in CD4+ T cells and peripheral blood, no SLE study to date has examined genome-wide DNA methylation in kidney biopsies. By considering differential gene expression derived from the micro-dissected glomerulus and tubulointerstitium kidneys in an independent cohort of SLE patients, in conjunction with the significance of aberrant methylation in the MZ twin data, this study corroborates many of the loci previously published as being hypomethylated in lupus nephritis patients.

The lack of any differentially methylated genes on the X chromosome is noteworthy given the 9:1 female to male gender bias in SLE. This result is not fully explained by the fact that older female MZ twins show a strong tendency for the same X chromosome to be inactivated [69,70] as the lack of differentially methylated sites on the X chromosome in this study is consistent with previous studies of unrelated individuals [11,15,17–21,23,52]. Jeffries et al., using the Illumina Infinium Human Methylation27 array, did observe differential methylation of CpGs in *PCTK1, ARAF, RRAGB*, and *SNX12* on the X chromosome [11], but no studies utilizing the more recent arrays replicate these findings. In our MZ twin study, CpG sites associated with *SNX12* had a minimum p-value = 0.02 (change in $\beta = -0.04$), but none of the other three genes had p-values < 0.05. Thus, to date, methylation patterns among genes located on the X chromosome do not appear to explain a substantial portion of the risk of SLE.

Within this study, the genomic locations of hypomethylated CpG sites were highly skewed toward CpG shores (0–2 kb from island) and shelves (2–4 kb from island) instead

of islands. Here, only 5 of 59 CpG sites were in a CpG island, despite nearly one third of the CpG sites on the Illumina HumanMethylation450 BeadChip being in a CpG island (Supplemental Figure S2). Our findings are consistent with those of Yeung et al., who demonstrated that most CpG sites hypomethylated in their lupus patients, when compared to controls, were located in CpG shores [21]. These data corroborate the hypothesis that CpG islands tend to have lower methylation rates than less dense CpG regions (e.g., shores and shelves) and that lower density allows for greater methylation autonomy in response to the environment, leading to increases in potential functional significance of the shores and shelves.

There are several limitations of this multi-omics study. One limitation was the modest sample size, as a larger sample would provide the potential to identify additional differentially methylated regions and pathways. However, it is important to recognize the power and value of a discordant MZ twin study design to reduce confounding based on genetic and environmental background. Further, the modest sample size does not negate the positive findings. There were only three discordant MZ twin pairs in the discovery cohort, but we replicated these results in an independent cohort of four MZ twin pairs. Given the number of samples, we were unable to construct and adjust for the full cell composition of the peripheral blood samples as the limited degrees of freedom precluded the robust use of deconvolution methods. Adjusting for latent methylation components in our analysis, while dampening the associations slightly, still identified the same IFN signature. Further, the collective results are supported by larger, independent case–control studies (described in Additional File 2), and we have shown that our methylation results correlate with gene expression in multiple cell types and tissues in independent SLE case–control studies; many were also identified as eQTMs in a Japanese cohort of 100 healthy individuals. We recognize that our cross-sectional study design (i.e., discovery, replication) cannot separate causality from response to disease, but the consistency of differentially methylated regions with the differentially expressed genes from independent gene expression studies is informative and helps identify epigenetically modified genes and pathways that are important in SLE.

5. Conclusions

The intersection of hypomethylated genes from MZ twins and upregulated genes from multiple independent cohorts and cell types were attributed to two distinct but integrated biologic pathways: the nucleic acid-sensing pathway and the IFN-inducing pathway. The source, type, and location of nucleic acids found in an SLE patient determine how and by which receptor the NA is recognized, and ultimately which IRF is stimulated. A multi-omics approach could allow classification of patients into different endotypes and possible treatment groups. Informatically linking the DM-DE genes to drug therapies identified a list of compounds that could be critically evaluated as potential candidates for future trials, either broadly for SLE or for individuals with specific hypomethylation signatures.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12121898/s1: Table S1: GEO repository datasets interrogated for differential expression between SLE cases and controls; Table S2: Clinical characteristics of monozygotic twins discordant for SLE; Table S3: Differential expression of hypomethylated genes in peripheral cell types from independent SLE cohorts; Table S4: Differentially methylated probes from three monozygotic twin pairs discordant for SLE; Figure S1: Area under the receiver operating characteristic curve (AUC) calculated at regular intervals between 0 and −0.15 in four cell types; Figure S2: Proportions of significantly associated CpGs (as defined in Table 1) located in islands, shores, shelves, and other content categories; Additional File 1: Differential methylation in MZ twins discordant for SLE; Additional File 2: Replication of CpG sites in Table 1 across published epigenome-wide SLE studies.

Author Contributions: Conceptualization, A.C.G., T.D.H., C.D.L. and P.E.L.; Acquisition of data, A.C.G., T.D.H., J.A.K., C.D.L. and P.E.L.; Analysis and interpretation, R.D.R., H.C.A., P.B., M.D.C., A.C.G., S.E.H., T.D.H., A.C.L., C.D.L., P.E.L., M.C.M., P.S.R. and K.D.Z. Drafting of the manuscript, H.C.A., A.C.G., T.D.H., C.D.L., M.C.M. and P.E.L. All authors have read and agreed to the published version of the manuscript.

Funding: The research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases under award numbers P30AR073750 and N01AR62277, the US Department of Defense under award number W81XWH-20-1-0686, the Wake Forest School of Medicine Center for Public Health Genomics, and research funds from CDL. AC Labonte, SE Heuer, RD Robl, MD Catalina, PE Lipsky, and AC Grammer received financial support from the ALR/LRI (now the Lupus Research Alliance) for drug repositioning work and from the John and Marcia Goldman Foundation for BigData analysis of gene expression from lupus patients compared to normal individuals.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Wake Forest School of Medicine Institutional Review Board (IRB).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All gene expression datasets are publicly available from the Gene Expression Omnibus (GEO) (Supplemental Table S1). The methylation data from the three monozygotic twins are available upon request from the authors. Samples, including DNA, are available from the Lupus Family Registry and Repository.

Acknowledgments: We would like to thank the Lupus Family Registry and Repository for providing the LFRR twin samples. We also thank the Wake Forest School of Medicine Center for Public Health Genomics for providing computational resources.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN); Harley, J.B.; Alarcón-Riquelme, M.E.; Criswell, L.A.; Jacob, C.O.; Kimberly, R.P.; Moser, K.L.; Tsao, B.P.; Vyse, T.J.; Langefeld, C.D.; et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.* **2008**, *40*, 204–210. [CrossRef] [PubMed]
2. Morris, D.L.; Sheng, Y.; Zhang, Y.; Wang, Y.-F.; Zhu, Z.; Tombleson, P.; Chen, L.; Cunninghame Graham, D.S.; Bentham, J.; Roberts, A.L.; et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat. Genet.* **2016**, *48*, 940–946. [CrossRef]
3. Lessard, C.J.; Sajuthi, S.; Zhao, J.; Kim, K.; Ice, J.A.; Li, H.; Ainsworth, H.; Rasmussen, A.; Kelly, J.A.; Marion, M.; et al. Identification of a Systemic Lupus Erythematosus Risk Locus Spanning ATG16L2, FCHSD2, and P2RY2 in Koreans. *Arthritis Rheumatol.* **2016**, *68*, 1197–1209. [CrossRef]
4. Sun, C.; Molineros, J.E.; Looger, L.L.; Zhou, X.-J.; Kim, K.; Okada, Y.; Ma, J.; Qi, Y.-Y.; Kim-Howard, X.; Motghare, P.; et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat. Genet.* **2016**, *48*, 323–330. [CrossRef]
5. Alarcón-Riquelme, M.E.; Ziegler, J.T.; Molineros, J.; Howard, T.D.; Moreno-Estrada, A.; Sánchez-Rodríguez, E.; Ainsworth, H.C.; Ortiz-Tello, P.; Comeau, M.E.; Rasmussen, A.; et al. Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. *Arthritis Rheumatol.* **2016**, *68*, 932–943. [CrossRef] [PubMed]
6. Langefeld, C.D.; Ainsworth, H.C.; Cunninghame Graham, D.S.; Kelly, J.A.; Comeau, M.E.; Marion, M.C.; Howard, T.D.; Ramos, P.S.; Croker, J.A.; Morris, D.L.; et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **2017**, *8*, 16021. [CrossRef]
7. Deng, Y.; Tsao, B.P. Updates in Lupus Genetics. *Curr. Rheumatol. Rep.* **2017**, *19*, 68. [CrossRef] [PubMed]
8. Chen, L.; Morris, D.L.; Vyse, T.J. Genetic advances in systemic lupus erythematosus: An update. *Curr. Opin. Rheumatol.* **2017**, *29*, 423–433. [CrossRef]
9. Deapen, D.; Escalante, A.; Weinrib, L.; Horwitz, D.; Bachman, B.; Roy-Burman, P.; Walker, A.; Mack, T.M. A revised estimate of twin concordance in systemic lupus erythematosus. *Arthritis Rheum.* **1992**, *35*, 311–318.
10. Hughes, T.; Sawalha, A.H. The role of epigenetic variation in the pathogenesis of systemic lupus erythematosus. *Arthritis Res. Ther.* **2011**, *13*, 245. [CrossRef]
11. Jeffries, M.A.; Dozmorov, M.; Tang, Y.; Merrill, J.T.; Wren, J.D.; Sawalha, A.H. Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. *Epigenetics* **2011**, *6*, 593–601. [CrossRef]
12. Coit, P.; Renauer, P.; Jeffries, M.A.; Merrill, J.T.; McCune, W.J.; Maksimowicz-McKinnon, K.; Sawalha, A.H. Renal involvement in lupus is characterized by unique DNA methylation changes in naïve CD4+ T cells. *J. Autoimmun.* **2015**, *61*, 29–35. [CrossRef]

13. Renauer, P.; Coit, P.; Jeffries, M.A.; Merrill, J.T.; McCune, W.J.; Maksimowicz-McKinnon, K.; Sawalha, A.H. DNA methylation patterns in naïve CD4+ T cells identify epigenetic susceptibility loci for malar rash and discoid rash in systemic lupus erythematosus. *Lupus Sci. Med.* **2015**, *2*, e000101. [CrossRef]
14. Ulff-Møller, C.J.; Asmar, F.; Liu, Y.; Svendsen, A.J.; Busato, F.; Grønbaek, K.; Tost, J.; Jacobsen, S. Twin DNA Methylation Profiling Reveals Flare-Dependent Interferon Signature and B Cell Promoter Hypermethylation in Systemic Lupus Erythematosus. *Arthritis Rheumatol.* **2018**, *70*, 878–890. [CrossRef]
15. Coit, P.; Jeffries, M.; Altorok, N.; Dozmorov, M.G.; Koelsch, K.A.; Wren, J.D.; Merrill, J.T.; McCune, W.J.; Sawalha, A.H. Genome-wide DNA methylation study suggests epigenetic accessibility and transcriptional poising of interferon-regulated genes in naïve CD4+ T cells from lupus patients. *J. Autoimmun.* **2013**, *43*, 78–84. [CrossRef]
16. Coit, P.; Ognenovski, M.; Gensterblum, E.; Maksimowicz-McKinnon, K.; Wren, J.D.; Sawalha, A.H. Ethnicity-specific epigenetic variation in naïve CD4+ T cells and the susceptibility to autoimmunity. *Epigenetics Chromatin* **2015**, *8*, 49. [CrossRef]
17. Absher, D.M.; Li, X.; Waite, L.L.; Gibson, A.; Roberts, K.; Edberg, J.; Chatham, W.W.; Kimberly, R.P. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS Genet.* **2013**, *9*, e1003678. [CrossRef] [PubMed]
18. Coit, P.; Yalavarthi, S.; Ognenovski, M.; Zhao, W.; Hasni, S.; Wren, J.D.; Kaplan, M.J.; Sawalha, A.H. Epigenome profiling reveals significant DNA demethylation of interferon signature genes in lupus neutrophils. *J. Autoimmun.* **2015**, *58*, 59–66. [CrossRef] [PubMed]
19. Mok, A.; Solomon, O.; Nayak, R.R.; Coit, P.; Quach, H.L.; Nititham, J.; Sawalha, A.H.; Barcellos, L.F.; Criswell, L.A.; Chung, S.A. Genome-wide profiling identifies associations between lupus nephritis and differential methylation of genes regulating tissue hypoxia and type 1 interferon responses. *Lupus Sci. Med.* **2016**, *3*, e000183. [CrossRef]
20. Chung, S.A.; Nititham, J.; Elboudwarej, E.; Quach, H.L.; Taylor, K.E.; Barcellos, L.F.; Criswell, L.A. Genome-Wide Assessment of Differential DNA Methylation Associated with Autoantibody Production in Systemic Lupus Erythematosus. *PLoS ONE* **2015**, *10*, e0129813. [CrossRef] [PubMed]
21. Yeung, K.S.; Chung, B.H.-Y.; Choufani, S.; Mok, M.Y.; Wong, W.L.; Mak, C.C.Y.; Yang, W.; Lee, P.P.W.; Wong, W.H.S.; Chen, Y.-A.; et al. Genome-Wide DNA Methylation Analysis of Chinese Patients with Systemic Lupus Erythematosus Identified Hypomethylation in Genes Related to the Type I Interferon Pathway. *PLoS ONE* **2017**, *12*, e0169553. [CrossRef]
22. Zhu, H.; Mi, W.; Luo, H.; Chen, T.; Liu, S.; Raman, I.; Zuo, X.; Li, Q.-Z. Whole-genome transcription and DNA methylation analysis of peripheral blood mononuclear cells identified aberrant gene regulation pathways in systemic lupus erythematosus. *Arthritis Res. Ther.* **2016**, *18*, 162. [CrossRef]
23. Imgenberg-Kreuz, J.; Carlsson Almlöf, J.; Leonard, D.; Alexsson, A.; Nordmark, G.; Eloranta, M.-L.; Rantapää-Dahlqvist, S.; Bengtsson, A.A.; Jönsen, A.; Padyukov, L.; et al. DNA methylation mapping identifies gene regulatory effects in patients with systemic lupus erythematosus. *Ann. Rheum. Dis.* **2018**, *77*, 736–743. [CrossRef] [PubMed]
24. Joseph, S.; George, N.I.; Green-Knox, B.; Treadwell, E.L.; Word, B.; Yim, S.; Lyn-Cook, B. Epigenome-wide association study of peripheral blood mononuclear cells in systemic lupus erythematosus: Identifying DNA methylation signatures associated with interferon-related genes based on ethnicity and SLEDAI. *J. Autoimmun.* **2019**, *96*, 147–157. [CrossRef]
25. Breitbach, M.E.; Ramaker, R.C.; Roberts, K.; Kimberly, R.P.; Absher, D. Population-Specific Patterns of Epigenetic Defects in the B Cell Lineage in Patients with Systemic Lupus Erythematosus. *Arthritis Rheumatol.* **2020**, *72*, 282–291. [CrossRef]
26. Yeung, K.S.; Lee, T.L.; Mok, M.Y.; Mak, C.C.Y.; Yang, W.; Chong, P.C.Y.; Lee, P.P.W.; Ho, M.H.K.; Choufani, S.; Lau, C.S.; et al. Cell Lineage-specific Genome-wide DNA Methylation Analysis of Patients with Paediatric-onset Systemic Lupus Erythematosus. *Epigenetics* **2019**, *14*, 314–351. [CrossRef] [PubMed]
27. Weeding, E.; Sawalha, A.H. Deoxyribonucleic Acid Methylation in Systemic Lupus Erythematosus: Implications for Future Clinical Practice. *Front. Immunol.* **2018**, *9*, 875. [CrossRef] [PubMed]
28. Castillo-Fernandez, J.E.; Spector, T.D.; Bell, J.T. Epigenetics of discordant monozygotic twins: Implications for disease. *Genome Med.* **2014**, *6*, 60. [CrossRef]
29. Javierre, B.M.; Fernandez, A.F.; Richter, J.; Al-Shahrour, F.; Martin-Subero, J.I.; Rodriguez-Ubreva, J.; Berdasco, M.; Fraga, M.F.; O'Hanlon, T.P.; Rider, L.G.; et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* **2010**, *20*, 170–179. [CrossRef] [PubMed]
30. Rakyan, V.K.; Beyan, H.; Down, T.A.; Hawa, M.I.; Maslau, S.; Aden, D.; Daunay, A.; Busato, F.; Mein, C.A.; Manfras, B.; et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet.* **2011**, *7*, e1002300. [CrossRef]
31. Gervin, K.; Vigeland, M.D.; Mattingsdal, M.; Hammerø, M.; Nygård, H.; Olsen, A.O.; Brandt, I.; Harris, J.R.; Undlien, D.E.; Lyle, R. DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: Identification of epigenetically dysregulated genes. *PLoS Genet.* **2012**, *8*, e1002454. [CrossRef]
32. Häsler, R.; Feng, Z.; Bäckdahl, L.; Spehlmann, M.E.; Franke, A.; Teschendorff, A.; Rakyan, V.K.; Down, T.A.; Wilson, G.A.; Feber, A.; et al. A functional methylome map of ulcerative colitis. *Genome Res.* **2012**, *22*, 2130–2137. [CrossRef] [PubMed]
33. Rasmussen, A.; Sevier, S.; Kelly, J.A.; Glenn, S.B.; Aberle, T.; Cooney, C.M.; Grether, A.; James, E.; Ning, J.; Tesiram, J.; et al. The lupus family registry and repository. *Rheumatology* **2011**, *50*, 47–59. [CrossRef] [PubMed]
34. Hochberg, M.C. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* **1997**, *40*, 1725. [CrossRef] [PubMed]

35. Bombardier, C.; Gladman, D.D.; Urowitz, M.B.; Caron, D.; Chang, C.H. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. *Arthritis Rheum.* **1992**, *35*, 630–640. [CrossRef] [PubMed]
36. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802. [CrossRef]
37. Rusinova, I.; Forster, S.; Yu, S.; Kannan, A.; Masse, M.; Cumming, H.; Chapman, R.; Hertzog, P.J. Interferome v2.0: An updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **2013**, *41*, D1040–D1046. [CrossRef] [PubMed]
38. Hachiya, T.; Furukawa, R.; Shiwa, Y.; Ohmomo, H.; Ono, K.; Katsuoka, F.; Nagasaki, M.; Yasuda, J.; Fuse, N.; Kinoshita, K.; et al. Genome-wide identification of inter-individually variable DNA methylation sites improves the efficacy of epigenetic association studies. *NPJ Genom Med.* **2017**, *2*, 11. [CrossRef] [PubMed]
39. Wu, J.; Irizarry, R.; MacDonald, J.; Gentry, J. Gcrma: Background adjustment using sequence information. *R Package Version* **2012**, *2200*, 3–10. [CrossRef]
40. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
41. Bader, G.D.; Hogue, C.W.V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2. [CrossRef] [PubMed]
42. Snel, B.; Lehmann, G.; Bork, P.; Huynen, M.A. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **2000**, *28*, 3442–3444. [CrossRef] [PubMed]
43. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452. [CrossRef] [PubMed]
44. Griffith, M.; Griffith, O.L.; Coffman, A.C.; Weible, J.V.; McMichael, J.F.; Spies, N.C.; Koval, J.; Das, I.; Callaway, M.B.; Eldred, J.M.; et al. DGIdb: Mining the druggable genome. *Nat. Methods* **2013**, *10*, 1209–1210. [CrossRef] [PubMed]
45. Grammer, A.C.; Ryals, M.M.; Heuer, S.E.; Robl, R.D.; Madamanchi, S.; Davis, L.S.; Lauwerys, B.; Catalina, M.D.; Lipsky, P.E. Drug repositioning in SLE: Crowd-sourcing, literature-mining and Big Data analysis. *Lupus* **2016**, *25*, 1150–1170. [CrossRef]
46. Mallya, M.; Campbell, R.D.; Aguado, B. Characterization of the five novel Ly-6 superfamily members encoded in the MHC, and detection of cells expressing their potential ligands. *Protein Sci.* **2006**, *15*, 2244–2256. [CrossRef]
47. Alaridah, N.; Winqvist, N.; Håkansson, G.; Tenland, E.; Rönnholm, A.; Sturegård, E.; Björkman, P.; Godaly, G. Impaired CXCR1-dependent oxidative defence in active tuberculosis patients. *Tuberculosis* **2015**, *95*, 744–750. [CrossRef]
48. Xu, R.; Bao, C.; Huang, H.; Lin, F.; Yuan, Y.; Wang, S.; Jin, L.; Yang, T.; Shi, M.; Zhang, Z.; et al. Low expression of CXCR1/2 on neutrophils predicts poor survival in patients with hepatitis B virus-related acute-on-chronic liver failure. *Sci. Rep.* **2016**, *6*, 38714. [CrossRef]
49. Almajhdi, F.N.; Al-Ahdal, M.; Abdo, A.A.; Sanai, F.M.; Al-Anazi, M.; Khalaf, N.; Viswan, N.A.; Al-Ashgar, H.; Al-Kahtani, K.; Al-Humaidan, H.; et al. Single nucleotide polymorphisms in CXCR1 gene and its association with hepatitis B infected patients in Saudi Arabia. *Ann. Hepatol.* **2013**, *12*, 220–227. [CrossRef]
50. Swamydas, M.; Gao, J.-L.; Break, T.J.; Johnson, M.D.; Jaeger, M.; Rodriguez, C.A.; Lim, J.K.; Green, N.M.; Collar, A.L.; Fischer, B.G.; et al. CXCR1-mediated neutrophil degranulation and fungal killing promote Candida clearance and host survival. *Sci. Transl. Med.* **2016**, *8*, 322ra10. [CrossRef]
51. Gardiner-Garden, M.; Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **1987**, *196*, 261–282. [CrossRef]
52. Zhao, M.; Zhou, Y.; Zhu, B.; Wan, M.; Jiang, T.; Tan, Q.; Liu, Y.; Jiang, J.; Luo, S.; Tan, Y.; et al. IFI44L promoter methylation as a blood biomarker for systemic lupus erythematosus. *Ann. Rheum. Dis.* **2016**, *75*, 1998–2006. [CrossRef]
53. Kiso, K.; Yoshifuji, H.; Oku, T.; Hikida, M.; Kitagori, K.; Hirayama, Y.; Nakajima, T.; Haga, H.; Tsuruyama, T.; Miyagawa-Hayashino, A. Transgelin-2 is upregulated on activated B-cells and expressed in hyperplastic follicles in lupus erythematosus patients. *PLoS ONE* **2017**, *12*, e0184738. [CrossRef]
54. Kim, H.-R.; Lee, H.-S.; Lee, K.-S.; Jung, I.D.; Kwon, M.-S.; Kim, C.-H.; Kim, S.-M.; Yoon, M.-H.; Park, Y.-M.; Lee, S.-M.; et al. An Essential Role for TAGLN2 in Phagocytosis of Lipopolysaccharide-activated Macrophages. *Sci. Rep.* **2017**, *7*, 8731. [CrossRef]
55. Dragovich, M.A.; Mor, A. The SLAM family receptors: Potential therapeutic targets for inflammatory and autoimmune diseases. *Autoimmun. Rev.* **2018**, *17*, 674–682. [CrossRef]
56. Li, F.J.; Won, W.J.; Becker, E.J.; Easlick, J.L.; Tabengwa, E.M.; Li, R.; Shakhmatov, M.; Honjo, K.; Burrows, P.D.; Davis, R.S. Emerging roles for the FCRL family members in lymphocyte biology and disease. *Curr. Top. Microbiol. Immunol.* **2014**, *382*, 29–50. [CrossRef]
57. Schreeder, D.M.; Cannon, J.P.; Wu, J.; Li, R.; Shakhmatov, M.A.; Davis, R.S. Cutting edge: FcR-like 6 is an MHC class II receptor. *J. Immunol.* **2010**, *185*, 23–27. [CrossRef] [PubMed]
58. Balada, E.; Felip, L.; Ordi-Ros, J.; Vilardell-Tarrés, M. DUSP23 is over-expressed and linked to the expression of DNMTs in CD4+ T cells from systemic lupus erythematosus patients. *Clin. Exp. Immunol.* **2017**, *187*, 242–250. [CrossRef] [PubMed]
59. Cavlar, T.; Ablasser, A.; Hornung, V. Induction of type I IFNs by intracellular DNA-sensing pathways. *Immunol. Cell Biol.* **2012**, *90*, 474–482. [CrossRef]
60. Suárez-Calvet, X.; Gallardo, E.; Nogales-Gadea, G.; Querol, L.; Navas, M.; Díaz-Manera, J.; Rojas-Garcia, R.; Illa, I. Altered RIG-I/DDX58-mediated innate immunity in dermatomyositis. *J. Pathol.* **2014**, *233*, 258–268. [CrossRef] [PubMed]
61. Crow, M.K. Collaboration, genetic associations, and lupus erythematosus. *N. Engl. J. Med.* **2008**, *358*, 956–961. [CrossRef]
62. Ambrosi, C.; Manzo, M.; Baubec, T. Dynamics and Context-Dependent Roles of DNA Methylation. *J. Mol. Biol.* **2017**, *429*, 1459–1475. [CrossRef]

63. Crowl, J.T.; Gray, E.E.; Pestal, K.; Volkman, H.E.; Stetson, D.B. Intracellular Nucleic Acid Detection in Autoimmunity. *Annu. Rev. Immunol.* **2017**, *35*, 313–336. [CrossRef]
64. Schübeler, D. Function and information content of DNA methylation. *Nature* **2015**, *517*, 321–326. [CrossRef] [PubMed]
65. Shrivastav, M.; Niewold, T.B. Nucleic Acid sensors and type I interferon production in systemic lupus erythematosus. *Front. Immunol.* **2013**, *4*, 319. [CrossRef]
66. Rathinam, V.A.K.; Jiang, Z.; Waggoner, S.N.; Sharma, S.; Cole, L.E.; Waggoner, L.; Vanaja, S.K.; Monks, B.G.; Ganesan, S.; Latz, E.; et al. The AIM2 inflammasome is essential for host defense against cytosolic bacteria and DNA viruses. *Nat. Immunol.* **2010**, *11*, 395–402. [CrossRef] [PubMed]
67. Man, S.M.; Karki, R.; Kanneganti, T.-D. AIM2 inflammasome in infection, cancer, and autoimmunity: Role in DNA sensing, inflammation, and innate immunity. *Eur. J. Immunol.* **2016**, *46*, 269–280. [CrossRef]
68. Mustelin, T.; Lood, C.; Giltiay, N.V. Sources of Pathogenic Nucleic Acids in Systemic Lupus Erythematosus. *Front. Immunol.* **2019**, *10*, 1028. [CrossRef] [PubMed]
69. Christensen, K.; Kristiansen, M.; Hagen-Larsen, H.; Skytthe, A.; Bathum, L.; Jeune, B.; Andersen-Ranberg, K.; Vaupel, J.W.; Orstavik, K.H. X-linked genetic factors regulate hematopoietic stem-cell kinetics in females. *Blood* **2000**, *95*, 2449–2451. [CrossRef] [PubMed]
70. Huang, Q.; Parfitt, A.; Grennan, D.M.; Manolios, N. X-chromosome inactivation in monozygotic twins with systemic lupus erythematosus. *Autoimmunity* **1997**, *26*, 85–93. [CrossRef] [PubMed]

Ghodke-Puranik *et al.*
Arthritis Research & Therapy (2021) 23:290
https://doi.org/10.1186/s13075-021-02660-2

Arthritis Research & Therapy

RESEARCH ARTICLE Open Access

Single-cell expression quantitative trait loci (eQTL) analysis of SLE-risk loci in lupus patient monocytes

Yogita Ghodke-Puranik[1], Zhongbo Jin[2], Kip D. Zimmerman[3], Hannah C. Ainsworth[3], Wei Fan[4], Mark A. Jensen[1], Jessica M. Dorschner[5], Danielle M. Vsetecka[5], Shreyasee Amin[6], Ashima Makol[6], Floranne Ernste[6], Thomas Osborn[6], Kevin Moder[6], Vaidehi Chowdhary[7], Carl D. Langefeld[3] and Timothy B. Niewold[1*]

Abstract

Background: We performed expression quantitative trait locus (eQTL) analysis in single classical (CL) and non-classical (NCL) monocytes from patients with systemic lupus erythematosus (SLE) to quantify the impact of well-established genetic risk alleles on transcription at single-cell resolution.

Methods: Single-cell gene expression was quantified using qPCR in purified monocyte subpopulations ($CD14^{++}CD16^{-}$ CL and $CD14^{dim}CD16^{+}$ NCL) from SLE patients. Novel analysis methods were used to control for the within-person correlations observed, and eQTLs were compared between cell types and risk alleles.

Results: The SLE-risk alleles demonstrated significantly more eQTLs in NCLs as compared to CLs ($p = 0.0004$). There were 18 eQTLs exclusive to NCL cells, 5 eQTLs exclusive to CL cells, and only one shared eQTL, supporting large differences in the impact of the risk alleles between these monocyte subsets. The *SPP1* and *TNFAIP3* loci were associated with the greatest number of transcripts. Patterns of shared influence in which different SNPs impacted the same transcript also differed between monocyte subsets, with greater evidence for synergy in NCL cells. *IRF1* expression demonstrated an on/off pattern, in which expression was zero in all of the monocytes studied from some individuals, and this pattern was associated with a number of SLE risk alleles. We observed corroborating evidence of this *IRF1* expression pattern in public data sets.

Conclusions: We document multiple SLE-risk allele eQTLs in single monocytes which differ greatly between CL and NCL subsets. These data support the importance of the *SPP1* and *TNFAIP3* risk variants and the IRF1 transcript in SLE patient monocyte function.

Introduction

Systemic lupus erythematosus (SLE) is a poorly understood autoimmune syndrome driven by the interplay of genetic and environmental influences, which lead to a break in immunologic self-tolerance. Genetic studies in SLE have been successful in identifying more than 100 SLE susceptibility loci [1, 2]. Most of the genetic polymorphisms associated with SLE are not coding-change variants [3, 4]. They are either located in non-coding regulatory regions near the 5′ and 3′ regions of genes, in DNAse hyper-sensitivity sites, or are in perfect LD with DNAse hypersensitivity sites. This suggests modulation of transcription as a likely mechanism by which many SLE-risk loci impact immune system biology [2], and data from many complex diseases support this idea [5]. Importantly, there is substantial variation in the pattern of DNAse hyper-sensitivity among different human cell

*Correspondence: Timothy.Niewold@nyumc.org
[1] Colton Center for Autoimmunity, NYU Grossman School of Medicine, 550 1st Ave, New York, NY 10016, USA
Full list of author information is available at the end of the article

types, supporting the idea that polymorphisms can tune gene expression in a highly cell-specific manner [5]. Thus, examining multiple cell types will be critical in determining the function of SLE-risk loci, as it is likely that the regulatory influence of these polymorphisms vary across cell types.

Transcriptomic studies in SLE using whole blood, peripheral mononuclear cells, or whole tissue are confounded by variations in the numbers and types of cells found within different samples and between individuals. In such studies, the relative proportion of contributing cell subsets can influence gene expression profile based on the unique gene signature related to their functions [6, 7], making it more difficult to interpret the biological significance of the observed differential gene expression. For example, it is impossible to determine if the difference in gene expression is shared homogeneously in all cells, or if the observed difference in gene expression is primarily driven by divergent gene expression in one particular cell subset, or if the difference arises solely due to a difference in proportions of specific cell types [8]. Similarly, an impact of the risk locus on gene expression in a minor cell subset may not be observed within a bulk cell data set. The situation could be even more complex, as each of these possibilities could be present in varying proportional degrees across samples within a given study. While de-convolution methods can be used, it is easy to envision scenarios in which de-convolution would be of limited use (e.g., the same transcript is simultaneously up- and down-regulated in different cell types, to varying degrees) [9, 10]. An additional strength of single-cell gene expression studies is that correlations between transcripts represent within-cell correlations, while co-expression in mixed cell bulk samples could represent some within cell correlations, but also could be the result of complex relationships between cells of different types.

While most of the confirmed SLE-risk loci are located in or near genes with immune system function, for the vast majority, we do not understand their impact on cell biology and immune responses nor their influence on various immune cell subsets. For risk loci near genes of unknown molecular function, it is difficult to identify the relevant biological pathway and cell type(s) when considering functional follow-up experiments. This is a major challenge in SLE genetics as many risk loci have been definitively implicated in SLE pathogenesis, but their molecular function is poorly understood [2]. When considering using gene expression data in eQTL studies, the above advantages of single-cell gene expression data from purified cell populations are intriguing and would suggest that single-cell expression studies would more accurately indicate the biological impact of risk loci. Our group and others have previously studied gene expression in sorted immune cell populations as well as at single-cell level in SLE patients and found striking between-individual differences in gene expression between immune cell subsets and within the same immune cell types [1, 7, 9]. In this study, we use single-cell gene expression data from two important SLE monocyte subsets and perform a single-cell eQTL analysis. We selected seven SNPs from six established SLE risk loci and 90 target genes for this analysis. We observed many eQTLs that met statistical significance after adjusting for the within-individual correlation by modeling the individual as a random effect in a linear model and applying multiple testing correction. These results demonstrate the efficiency of single-cell eQTL approach to effectively detect the biological impact of risk loci. The associated eQTL transcripts largely differed between the two closely related monocyte subsets, making the case that risk locus function differently depending upon cell type. We also observed a great deal of diversity in the transcript lists associated with each risk SNP.

Methods

Patients and samples

Whole blood samples from 15 Female SLE patients fulfilling the American College of Rheumatology criteria for the diagnosis of SLE [11, 12] and five age-sex matched healthy controls were procured from the Mayo Clinic, Rochester, MN. Exclusion criteria included pregnancy, active acute infection, chronic infection (e.g., hepatitis C, HIV, etc.), and current intravenous therapy (e.g., methylprednisolone or cyclophosphamide). The institutional review board approved the study and all patients provided informed consent. The patient data were used for all eQTL analyses, and the control data were only used in the comparison of IRF1 expression. The control set was too small to analyze separately for eQTLs, and combining patient and control cells together for eQTL analysis could result in confounding due to the expected differences in gene expression between patients and controls. Control data were only used in the IRF1 expression analysis.

Purification of classical ($CD14^{++}CD16^{-}$) and non-classical ($CD14^{dim}CD16^{+}$) monocytes

As previously described [9], $CD14^{++}CD16^{-}$classical (CL) monocytes and $CD14^{dim}CD16^{+}$ non-classical (NCL) monocytes were isolated from peripheral blood and purified using magnetic separation. Briefly, CL monocytes were first purified by negative selection using a modified Human Pan-Monocyte Isolation protocol (Miltenyi) with addition of anti-CD16-biotin (Miltenyi) into the biotin-antibody cocktail. The purity was further increased using subsequent CD14 positive selection (Miltenyi). NCL monocytes were purified

similarly with addition of anti-CD14-biotin (Miltenyi) to the antibody cocktail for negative selection followed by CD16 microbeads (Miltenyi) for positive selection. Flow cytometry analysis showed that of the CL and NCL populations obtained, each contained >95% of each desired cell type (Supplemental Fig. 1).

C1 single-cell capture

Single-cells from each bulk monocyte subset were isolated using Fluidigm C1 Single-Cell Auto Prep System. Purified CL monocytes were stained with Molecular Probes™ CellTracker™ Green CMFDA Dye (Life Technologies), while NCL monocytes were unstained before loading to C1 Single-Cell Auto Prep Array Integrated Fluidic Circuits (IFCs). CL and NCL monocytes were then sequentially loaded onto the C1 Integrated Fluidic Circuit (IFC). CL vs. NCL monocyte lineage of individual cells was determined by direct visualization using fluorescent microscopy, and at the same time, empty wells and wells that contained more than one cell were marked to exclude from later analysis. The IFCs were then examined using fluorescent microscopy, and the captured cells were identified as CL (stained) or NCL (not stained). Wells that contained more than one cell were also noted to exclude from later analysis. We captured 470 CL and 394 NCL cells from the SLE patients in total, averaging between 50 to 60 single cells per patient across both monocyte subsets, after excluding doublets and fragments. These results represent a 60% capture site efficiency.

Single cell PCR gene expression

A total of 90-target genes, relevant to monocyte function, that included major cytokines and pathway proteins involved in inflammation were selected for pre-amplification in the IFCs using the Fluidigm C1 Single-Cell Auto Prep System according to the manufacturer's protocol. qPCR-based gene expression assay of the target gene pre-amplified cDNAs were carried out using 96.96 IFCs on the BioMark HD System (Fluidigm) as described in the protocol. Raw data was analyzed using the Fluidigm Real-Time PCR Analysis software (v. 4.1.2) and quality check was performed by inspecting melt curves, amplification curves. A failure score was calculated for each cell as described previously [9, 13]. Cells with failure score (total CT value) greater than two standard deviations above the mean were excluded from downstream analysis. The limit of detection CT values was set at 28 [10]; CT values greater than or equal to 28 were considered non-detected and were assigned a value of zero for analysis. Gene expression values were calculated by subtracting the threshold cycle value for each gene for each cell from the number of cycles in the PCR reaction. In this way, higher numbers represent greater gene expression, and lower numbers indicate less expression.

Genotyping

Seven lupus risk single nucleotide polymorphisms (SNPs) in six gene loci, IRF5, IRF7, ITGAM, PTPN22, SPP1, and TNFAIP3 were genotyped for eQTL analysis. We selected well-established lupus risk polymorphisms from the literature which we thought may have function in monocytes [2]. The polymorphisms studied were as follows: IRF5 (rs10488631), IRF7 (rs1061502), ITGAM (rs1143679, rs1143689), PTPN22 (rs2476601), SPP1 (rs9138), and TNFAIP3 (rs2230926). Genotyping was performed using PCR allelic discrimination assays on a BioMark HD System (Fluidigm). The observed genotype frequencies of the studied SNPs did not deviate significantly from Hardy–Weinberg equilibrium.

Statistical analysis

For the initial univariate analysis, gene expression data was separated in to three genotype categories for each bi-allelic SNP for each patient (homozygous minor allele, heterozygous, and homozygous major allele). Data in CL and NCL populations were separately analyzed, using non-parametric analyses (Mann-Whitney *U*). Even when considering eQTL associations that surpassed a Bonferroni correction for the number of comparisons ($P = 8 \times 10^{-5}$), this was found to be too permissive with respect to type I error (Supplemental Fig. 2) [14]. This was due to distributional properties of the data that demonstrated patterns of normal expression mixed with varying degrees of dropout data and significant within-person correlation in transcript values. To deal with these properties, data was reanalyzed for eQTL associations utilizing four separate approaches [15]. The first approach used a tweedie mixed-effects model [16] to simultaneously account for the dropout and the person-specific heterogeneity. Gene expression was modeled as the outcome and genotypes were modeled as predictors along with a random effect for individual. The second approach used a logistic mixed-effects model [17], where all non-zero gene expression values were assigned as ones and modeled as a binary outcome to compare the proportion of genes turned "on" or "off" for each SNP. Genes where the average proportion turned "on" exceeded 98% were dropped. The third approach also computed a mixed-effects model with just the non-zero gene expression values, assuming an underlying Gaussian distribution. Lastly, the proportion of genes turned "on" or "off" was computed within each individual and a simple analysis of variance was computed where the proportion was modeled as the outcome and the genotype as the predictor.

A Benjamini-Hochberg false discovery rate was used to control for multiple comparisons and results meeting an FDR < 0.1 were retained [18]. As shown in Table 1, the logistic and proportional models provided the strongest ability to detect eQTLs (12 and 9 eQTLs respectively), followed by Gaussian (3 eQTLs), and tweedie (1 eQTL) models. eQTL lists were compared among risk alleles and between cell types to understand the degree to which effects were shared between cells types and the degree to which SLE-risk loci coordinately regulated the same transcripts. eQTLs were considered shared if they met the significance cutoff in both monocyte subsets and were in the same direction of association. These patterns of sharing are represented using Venn diagrams.

Analyses to detect modules of gene co-expression in the single-cell data were completed in each cell type separately (CL and NCL). Using the intersection of genes (common across all individuals), we built a pairwise gene-by-gene correlation matrix for each individual and each cell type. Each correlation matrix was averaged into a single correlation matrix to find a mean correlation across all individuals while removing the inter-individual differences. The mean correlation was then used to compute eigenvectors and eigenvalues and build a principal component analysis. From there, each individual cell was projected on to that principal component space and observed for differences by individual. Gene sets were retained if the absolute value of the individual loadings associated with highly explanatory principal components were greater than 0.7.

Table 1 List of significant eQTL associations detected by the various statistical methods in classical and non-classical monocytes at < 0.1 FDR

Gene (SNP rsID)	Associated transcript	Method	Monocyte subset
ITGAM (rs1143679)	*TLR7*	Logistic	Classical
ITGAM (rs1143683)	*JAK1*	Logistic	Classical
TNFAIP3 (rs2230926)	*IRF8*	Proportion	Classical
SPP1 (rs9138)	*ARG1*	Logistic	Classical
SPP1 (rs9138)	*IRF1*	Logistic	Classical
SPP1 (rs9138)	*IRF4*	Logistic	Classical
IRF5 (rs10488631)	*IRF1*	Logistic	Non-classical
IRF7 (rs1061502)	*IRF1*	Logistic	Non-classical
ITGAM (rs1143679)	*ARG1*	Gaussian	Non-classical
ITGAM (rs1143679)	*TCF4*	Logistic	Non-classical
ITGAM (rs1143683)	*IL1B*	Gaussian	Non-classical
ITGAM (rs1143683)	*TNFA*	Gaussian	Non-classical
TNFAIP3 (rs2230926)	*CD274*	Logistic	Non-classical
TNFAIP3 (rs2230926)	*FCER1G*	Proportion	Non-classical
TNFAIP3 (rs2230926)	*IL7R*	Proportion	Non-classical
TNFAIP3 (rs2230926)	*STAT1*	Proportion	Non-classical
TNFAIP3 (rs2230926)	*STAT2*	Tweedie	Non-classical
TNFAIP3 (rs2230926)	*TNFA*	Logistic	Non-classical
TNFAIP3 (rs2230926)	*TYK2*	Proportion	Non-classical
PTPN22 (rs2476601)	*IL5*	Logistic	Non-classical
SPP1 (rs9138)	*IFIT5*	Proportion	Non-classical
SPP1 (rs9138)	*IL1A*	Proportion	Non-classical
SPP1 (rs9138)	*IRF1*	Logistic	Non-classical
SPP1 (rs9138)	*TLR3*	Proportion	Non-classical
SPP1 (rs9138)	*TYK2*	Proportion	Non-classical

Results

Unique eQTL associations between CL and NCL monocytes

Using the four different analysis methods to query the data resulted in a total of 25 eQTL associations meeting a FDR < 0.1 (Table 1, Fig. 1). Interestingly, these largely differed between the two related monocyte subsets. There were 18 eQTLs exclusive to NCL cells, 5 eQTLs exclusive to CL cells, and one shared eQTL (Fig. 1, $p = 0.0007$ for a difference between the observed degree of sharing and a model in which 50% of eQTLs are shared between cell types). The SLE-associated SNPs demonstrated more eQTLs in NCLs compared to CLs ($p = 0.0004$). For a given SNP, the eQTL associated transcripts largely differed between cell types, with only one transcript-eQTL shared between CL and NCL cells (*SPP1* rs9138 with the IRF1 transcript). The greatest number of eQTLs was observed with the *SPP1* and *TNFAIP3* loci (7 and 8 eQTLs respectively). We included two missense SNPs in the *ITGAM* locus that have been shown evidence for independent biological function [19], and these two SNPs in the same locus were associated with different transcripts. These data indicate that the same risk allele had a different biological impact between the two monocyte subsets. This is striking given that the two monocyte subsets would largely be more closely related, than to B cells or T cells. These data suggest the importance of studying risk alleles within very specific cellular subsets to understand their biological roles. The different analysis methods used to detect eQTLs performed differently in the single-cell data, with logistic and proportional models detecting the greatest number of eQTLs (Table 1).

Degree of eQTL transcript sharing between SLE-risk alleles

Next, we assessed whether different SNPs modulated the same transcripts (transcript sharing), as this could indicate different risk alleles converging on similar biological pathways. There were no transcripts shared among SNPs in CL cells (Fig. 2). In NCLs, two transcripts were common between two SNPs (*TNFA*, *TYK2*), and one transcript was common to three SNPs

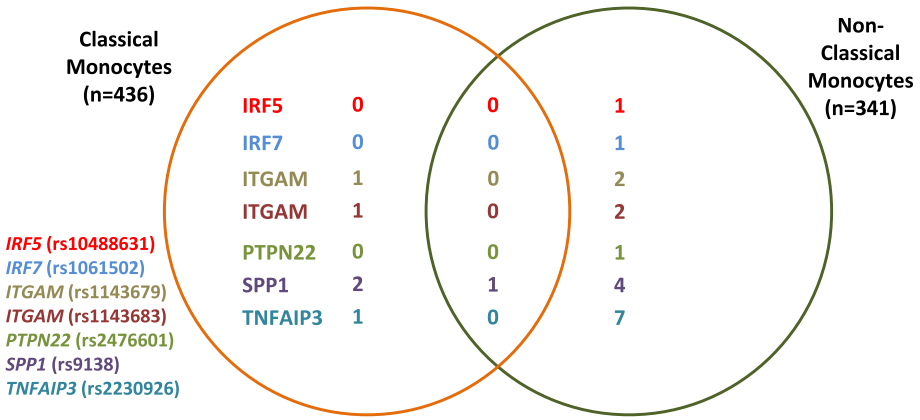


Fig. 1 Venn diagram showing unique and shared eQTL associated transcripts between CL and NCL for each lupus risk SNP. Numbers indicate the number of transcripts associated with each SNP, with the numbers inside the overlap indicating transcript associations which are shared across the two monocyte subsets and those outside the overlap indicating unique SNP-transcript associations for each monocyte subset. The orange circle represents CL monocytes and the green circle represents NCL monocytes. Each lupus risk SNP is represented with different color

(*IRF1*). Interestingly, in the NCL cells SNPs in IRFs (*IRF5* and *IRF7*) are associated with *IRF1* expression. It is also notable that *IRF1* was the one eQTL that was shared between CL and NCL cells in the analyses above. Thus, while genetic variation in *IRF1* has not been associated with SLE, these analyses support the idea that *IRF1* expression is modulated by SLE genetic risk factors in monocyte lineage cells.

On/off pattern of gene expression

Interestingly, the *IRF1* transcript demonstrated a highly binary expressed/not expressed pattern for all cells

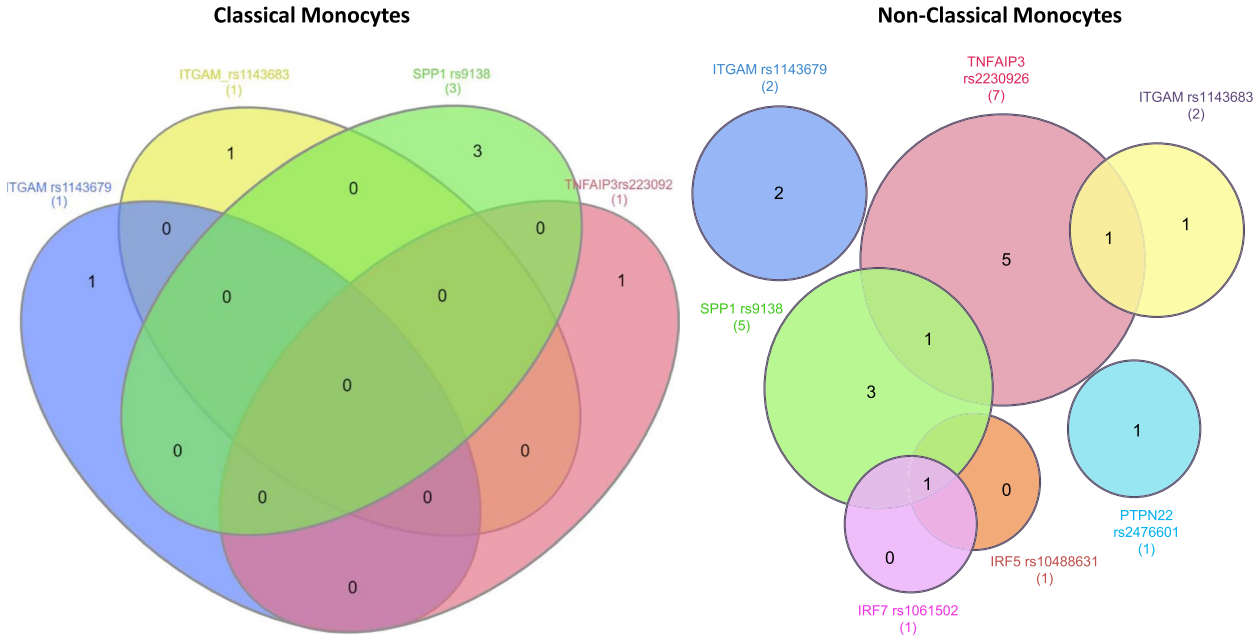


Fig. 2 Comparison of eQTL lists for the different SLE-risk SNPs in two monocyte subsets. Venn diagram showing unique and shared eQTL transcripts associated with each risk allele for **A** CL and **B** NCL monocytes. The circles indicated by each color to represent one lupus risk SNP. Numbers in each area of the diagram represent the number of transcripts significantly associated with that risk allele, either separately or overlapping between risk alleles.

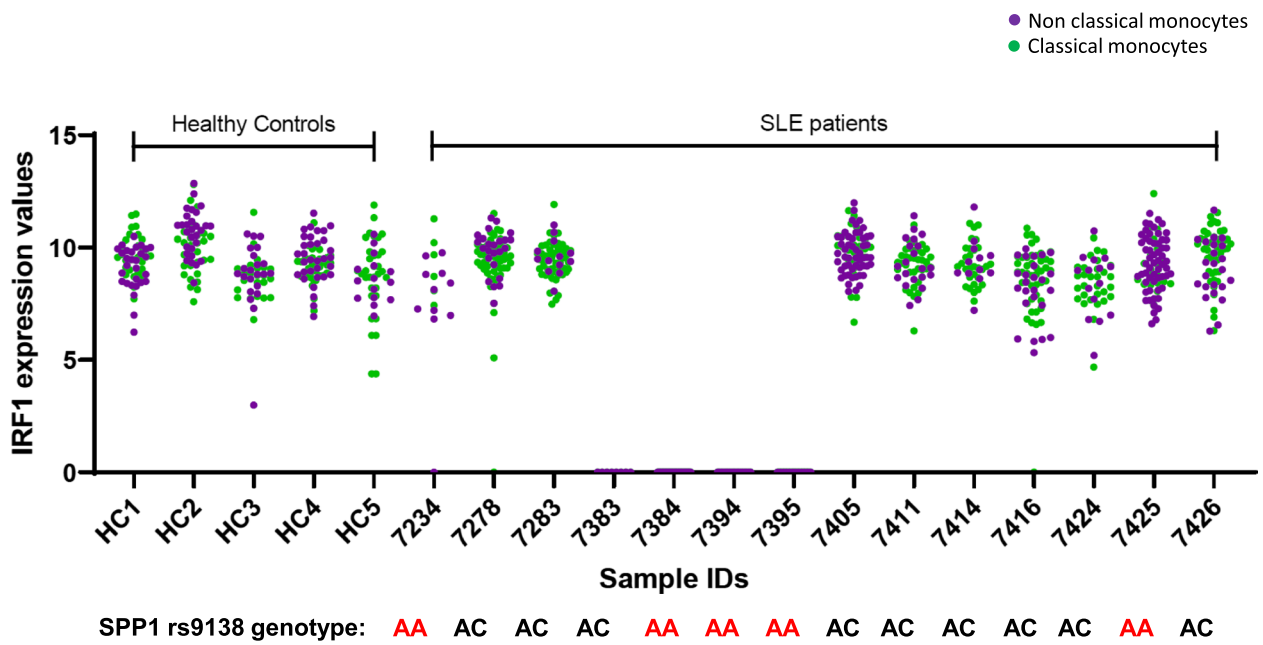


Fig. 3 *IRF1* expression in CL and NCL monocytes in each individual separately. Gene expression values for *IRF1* are shown, with the cells from each individual in the study in a separate column. CL monocytes are shown in blue and NCLs in green, with each dot representing one cell. The genotypes under each column represent the *SPP1* rs9138 genotype in each person

from a given individual, such that either all of the individual's cells did not express the gene or the majority of cells showed *IRF1* expression (Fig. 3). This was true of both the CL and NCL monocytes from the same person. This pattern was restricted to patients and not observed in the controls in our study. We have seen this pattern in other single-cell qPCR studies of other diseases [20]. For example in a study of rheumatoid arthritis monocytes, we observed that *JAK1* expression fit this pattern, in patients only and not in controls [20]. *JAK1* did not fit this pattern in the present study of lupus patients, suggesting that this pattern of gene expression may be specific to the disease state. We searched public databases for other precedents of this on/off pattern of gene expression using Bio Turing browser version 2.5.3 [21]. We found a similar pattern for *IRF1* in monocytes from a single-cell RNA sequencing study examining patients with myeloma [22] (Supplemental Fig. 3). While our PCR data have a wider dynamic range of values than the public RNA-seq data, the on/off pattern appears similar between these two studies. This suggests that examining gene expression patterns in an individual is important, as this type of pattern is likely to be lost when individuals are pooled for analysis. The strength of the pattern in our data compared to RNA-seq data sets may indicate that these patterns are more efficiently detected in single cell qPCR data than in single cell RNA-seq data.

Modular co-expression analysis of the single-cell data

The principal component analyses revealed much higher overlap of cells when correcting for inter-individual differences than not (Fig. 4). For classical cells, the first principal component explained 39.5% of the variance and the second principal component explained only 3.05% of the variance. Similarly, in non-classical cells, the first principal component explained 35.6% of the variance and the second principal component explained only 3.49% of the variance (Fig. 4). Thirty-two genes were associated with lower principal component 1 scores across both of the cell types (|loadings|>0.7) (Table 2). Sixteen genes were associated with lower principal component 1 scores in non-classical cells (Table 2). Of those, 15 were shared in both cell types and only one (IFNG) was unique to NCLs, demonstrating a core set of co-expressed genes that are in common across both cell types (Table 2). In the CL cells, there were 16 additional genes that were co-expressed, supporting a larger co-expression network in this cell type.

Discussion

In this study, we document a number of eQTLs associated with common autoimmune risk alleles for SLE in human monocytes, at a single-cell resolution. We studied patients, which may have increased our ability to detect eQTLs associated with these alleles, as the other requisite genetic background for SLE is also present in

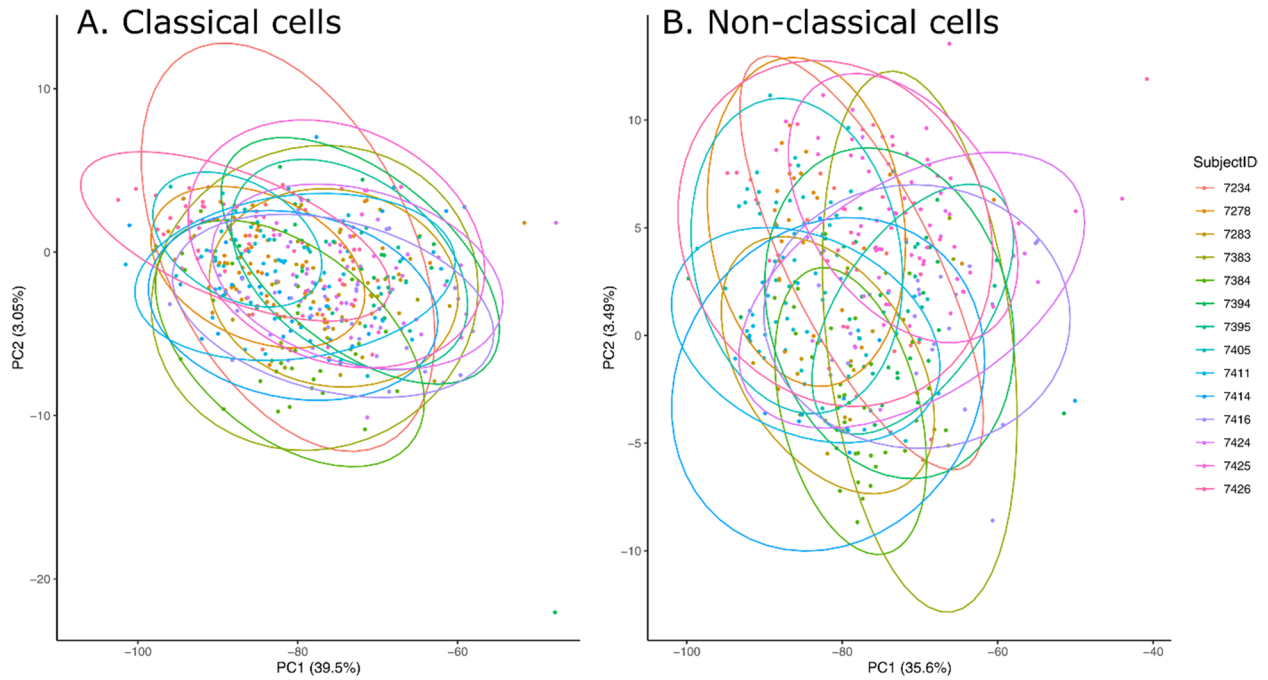


Fig. 4 Principal component analyses of classical and non-classical cells. Each cell is a dot, and data are shown after adjusting for the inter-individual differences by averaging gene-gene correlation matrices across each individual and subsequently projecting cells onto to the principal component space. Cells are color-coded and circled by 95% confidence ellipses by subject identifiers. Large overlap demonstrates the removal of the individual-specific heterogeneity

Table 2 Co-expression networks, genes associated with lower principal component 1 scores (|loadings|> 0.7). These gene sets represent a set of co-expressed genes that explain the most variance in each dataset. A large portion of the genes are shared; however, classical cells demonstrate a much larger co-expression network

Classical	Shared	Non-classical
CCR6	*CCR2*	*IFNG*
ITGAE	*CCR5*	
CD36	*IDO1*	
CD86	*IFIH1*	
FCER1G	*IFIT3*	
GMCSF	*IL23A*	
IFIT2	*STAT3*	
IFNB1	*STAT5*	
IL15	*TLR3*	
IL2	*IL12B*	
LILRA4	*TRAF6*	
PRDM1	*FLT3*	
STAT6	*CTLA4*	
TLR8	*CXCR7*	
TICAM1	*CD80*	
TYK2		
VCAN		

these individuals. The degree of difference in eQTL lists between monocyte subsets was striking, as these two cell types are more closely related to each other than other common immune cell types such as T cells and B cells. These data suggest that highly cell-type specific patterns of eQTLs are present in immune cells. Therefore, choosing the right cell types and including multiple cell types will be critical when studying risk alleles in immune mediated diseases. Screening of single-cell eQTL data [23] across multiple cell types would be an important strategy to decide upon which cell type to study in functional experiments, and our data support the limitations of gene annotation and presumed functions when considering the biological impact of the risk allele. One example of this would be the large number of trans associations we observe, which could not be predicted based upon the sequence location of the risk variant (e.g. SPP1(rs9138) associated with *IRF1* and *TYK2* transcripts).

It is interesting that we observed more eQTLs in the NCLs as compared to the CLs, as the cell numbers were similar between the two cell subsets and this is not related to statistical power. It could suggest that these risk alleles mediate their risk of disease to a greater degree via the NCL lineage as compared to the CL lineage. The structure of shared transcript modulation

shown in Fig. 2 provides a map of the interactions between risk alleles at the biological level, and these data suggest greater coordination between risk alleles in NCL monocytes at least with respect to the variants and transcripts that we studied. Interestingly, while the number of eQTLs observed in CL cells was fewer, the co-expression network observed in this cell type contained a larger number of transcripts. This taken together with the analysis above would suggest that a fewer number of risk alleles are operative in CL cells but that these alleles result in a larger number of co-expressed transcripts. This finding should be tested in an RNA-seq experiment, as this conclusion is limited by the fact that we tested a prescribed set of transcripts in this study. Our data also support the overall importance of the *SPP1* and *TNFAIP3* risk alleles with respect to gene transcription in both CL and NCL monocytes. These data support the idea that different risk alleles will have their greatest effects in specific cell types, which will not be predictable from the magnitude of the effect size in case-control genetic association. The *SPP1* risk variant has been linked to innate immune system cytokine production in SLE previously [24], while *TNFAIP3* variants have been associated with differential TNFAIP3 function in monocyte lineage cells [25].

The on-off pattern of gene expression observed with *IRF1* is striking, and in comparison with public RNA-seq data sets it seems that the qPCR approach we have used illustrates this pattern more dramatically. This could be due to the more quantitative nature of PCR vs. shotgun sequencing. Biologically, this could relate to a strong transcriptional repressor, and it is interesting that we have observed this phenomenon in disease but not in controls, and in multiple disease states and with different transcripts [20]. This could indicate that the on/off gene expression pattern is related to either medication or to the underlying disease process. In our study, the *IRF1* transcript which was expressed in an on/off pattern was an eQTL. This could suggest genetic variation as a cause of the on/off pattern, although it is a trans-eQTL and thus would not represent a simple impact upon a *cis*-regulatory element. We have observed trans-eQTLs in this study despite measuring some of the transcripts for the annotated *cis*-gene variants being studied. We did not include each transcript in the region of the SNPs studied, and thus, we did not emphasize *cis*-eQTLs, but instead focused on monocyte-relevant transcripts that result from pathway activation events in the cell.

There are some limitations of this study. We have studied limited number of target genes and well-established SLE risk alleles; however, future studies are needed to include additional risk alleles and more diverse transcripts related to SLE pathogenesis. This will help in identifying additional eQTLs and in delineating the effect of risk variants in different cell types through *cis* or *trans* transcript regulation. Second, it is will be interesting to follow up the surprising on/off gene expression pattern in other disease states, larger control samples, and across different cell types. We expect that this should be done using single cell qPCR along with single cell RNA-seq, and the qPCR method may be more sensitive to detect this pattern.

Conclusions

Studying single-cell eQTLs in SLE patient immune cells has allowed for novel insights which could not be achieved using previous mixed immune cell gene expression methods. These data support the importance of the *SPP1* and *TNFAIP3* risk variants and the IRF1 transcript in SLE patient monocyte function. This approach would be of great utility to detect differential transcription related to SLE-risk loci across multiple primary human cell types. This approach addresses a major frontier in complex autoimmune disease genetics, allowing us to understand how the function of a given risk allele varies by cell type in humans.

Abbreviations

eQTL: Expression quantitative trait locus; CL: Classical; NCL: Non-classical; SLE: Systemic lupus erythematosus; IFN: Interferon; qPCR: Quantitative polymerase chain reaction; LD: Linkage disequilibrium; IFC: Integrated fluidic circuit; IRF1: Interferon regulatory factor 1; IRF5: Interferon regulatory factor 5; IRF7: Interferon regulatory factor 7; ITGAM: Integrin alpha M; PTPN22: Protein tyrosine phosphatase, non-receptor type 22; SPP1: Secreted phosphoprotein 1; TNFAIP3: Tumor necrosis factor alpha induced protein 3.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13075-021-02660-2.

Additional file 1: Supplemental Figure 1. Scatter plots of purified classical and non-classical monocytes. PBMCs = peripheral blood mononuclear cells, percentages in the gates for classical and non-classical monocytes indicate the purity of populations purified. **Supplemental Figure 2.** Quantile − quantile plot showing eQTL associations with SLE risk loci. A) Classical monocytes; B) Non-classical monocytes. A very early deviation of the observed from the expected P value for eQTL associations suggested high type I error, which we found to relate to within-person correlations and distributional properties of the single cell data as described in the paper. eQTL associations that remained significant after using the approaches described in the Methods to correct for the distributional properties of the data are highlighted in orange for classical and magenta for non-classical monocytes. **Supplemental Figure 3.** On/off pattern of IRF1 gene expression in monocytes from a single cell RNA sequencing study examining patients with multiple myeloma. X-axis shows individual patients and the Y-axis shows gene expression values for the IRF1 transcript. Each plotted dot represents the expression level for IRF1 in one cell.

Bars show the median, error bars show the interquartile range. Data from public database as reported in Haradhvala, N.J., et al., Cancer Research, 2019.

Acknowledgements
The authors thank the funding sources noted above and the patients for their participation in this study.

Authors' contributions
YGP, ZJ, WF, MAJ, and JMD generated experimental data, analyzed data, and participated in drafting the manuscript. KDZ, HCA, and CDL analyzed data and participated in drafting the manuscript. DMV, SA, AM, FE, TO, KM, and VC assisted in recruiting the patients and providing clinical data and participated in drafting the manuscript. TBN designed the study, analyzed the data, and participated in drafting the manuscript. The author(s) read and approved the final manuscript.

Funding
Niewold: Colton Center for Autoimmunity, NIH (AR060861, AR057781, AR065964, AI071651), the Lupus Research Foundation, and the Lupus Research Alliance. Ghodke-Puranik: Colton Center for Autoimmunity. Langefeld: Department of the Army (W81XWH-20-1-0686).

Availability of data and materials
Primary data are available upon request to qualified investigators.

Declarations

Ethics approval and consent to participate
The institutional review board at the Mayo Clinic approved the study, and all patients provided informed consent.

Consent for publication
All authors have reviewed the final manuscript and approve of the submission for publication.

Competing interests
TBN has received research grants from EMD Serono, Inc., and has consulted for Thermo Fisher, Progenetec, Roivant Sciences, Ventus, Toran, and Inova, all unrelated to the current manuscript. Other authors have no conflict of interest to declare.

Author details
[1]Colton Center for Autoimmunity, NYU Grossman School of Medicine, 550 1st Ave, New York, NY 10016, USA. [2]Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA. [3]Department of Biostatistics and Data Science and Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA. [4]Department of Rheumatology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. [5]Department of Immunology and Division of Rheumatology, Mayo Clinic, Rochester, MN, USA. [6]Division of Rheumatology, Mayo Clinic, Rochester, MN, USA. [7]Division of Rheumatology, Allergy and Immunology, Yale University School of Medicine, New Haven, USA.

Received: 5 April 2021 Accepted: 17 October 2021
Published online: 30 November 2021

References

1. Catalina MD, Owen KA, Labonte AC, Grammer AC, Lipsky PE. The pathogenesis of systemic lupus erythematosus: Harnessing big data to understand the molecular basis of lupus. J Autoimmun. 2020;110:102359.
2. Ghodke-Puranik Y, Niewold TB. Immunogenetics of systemic lupus erythematosus: a comprehensive review. J Autoimmun. 2015;64:125–36.
3. Langefeld CD, Ainsworth HC, Cunninghame Graham DS, Kelly JA, Comeau ME, Marion MC, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. Nat Commun. 2017;8:16021.
4. Owen KA, Price A, Ainsworth H, Aidukaitis BN, Bachali P, Catalina MD, et al. Analysis of trans-ancestral SLE risk loci identifies unique biologic networks and drug targets in African and European ancestries. Am J Hum Genet. 2020;107(5):864–81.
5. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5.
6. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A. 2003;100(4):1896–901.
7. Sharma S, Jin Z, Rosenzweig E, Rao S, Ko K, Niewold TB. Widely divergent transcriptional patterns between SLE patients of different ancestral backgrounds in sorted immune cell populations. J Autoimmun. 2015;60:51–8.
8. Flatz L, Roychoudhuri R, Honda M, Filali-Mouhim A, Goulet JP, Kettaf N, et al. Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. Proc Natl Acad Sci U S A. 2011;108(14):5724–9.
9. Jin Z, Fan W, Jensen MA, Dorschner JM, Bonadurer GF 3rd, Vsetecka DM, et al. Single-cell gene expression patterns in lupus monocytes independently indicate disease activity, interferon and therapy. Lupus Sci Med. 2017;4(1):e000202.
10. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat Biotechnol. 2013;31(8):748–52.
11. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum. 1997;40(9):1725.
12. Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, et al. 2019 European league against rheumatism/American college of rheumatology classification criteria for systemic lupus erythematosus. Arthritis Rheumatol. 2019;71(9):1400–12.
13. Livak KJ, Wills QF, Tipping AJ, Datta K, Mittal R, Goldson AJ, et al. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. Methods. 2013;59(1):71–9.
14. Zimmerman KD, Langefeld CD. Hierarchicell: an R-package for estimating power for tests of differential expression with single-cell data. BMC Genomics. 2021;22(1):319.
15. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun. 2021;12(1):738.
16. Mollie E, Brooks KK, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. Bolker: glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R Journal. 2017;9(2):378–400.
17. Bates DMM, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Software. 2015;67(1):1–48.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Series B (Methodological). 1995;57(1):289–300.
19. Faridi MH, Khan SQ, Zhao W, Lee HW, Altintas MM, Zhang K, et al. CD11b activation suppresses TLR-dependent inflammation and autoimmunity in systemic lupus erythematosus. J Clin Invest. 2017;127(4):1271–83.
20. Wampler Muskardin TL, Fan W, Jin Z, Jensen MA, Dorschner JM, Ghodke-Puranik Y, et al. Distinct single cell gene expression in peripheral blood monocytes correlates with tumor necrosis factor inhibitor treatment response groups defined by type i interferon in rheumatoid arthritis. Front Immunol. 2020;11:1384.
21. BioTuring Inc. BioTuring Browser: a platform for single-cell data analysis. V. 2.6.0. 2020. https://bioturing.com/bbrowser
22. Zavidij O, Haradhvala NJ, Mouhieddine TH, Sklavenitis-Pistofidis R, Cai S, Reidy M, et al. Singlecell RNA sequencing reveals compromised immune microenvironment in precursor stages of multiple myeloma. Nat Cancer. 2020;1(5):493–506.
23. Monique GPvdW, Dylan H. de Vries, Hilde E. Groot, Gosia Trynka, Chung-Chau Hon, Martijn C. Nawijn, Youssef Idaghdour, Pim van der Harst, Chun J. Ye, Joseph Powell, Fabian J. Theis, Ahmed Mahfouz, Matthias Heinig, Lude Franke: Single-cell eQTLGen Consortium: a personalized understanding of disease. https://arxiv.org/abs/1909.12550.
24. Kariuki SN, Moore JG, Kirou KA, Crow MK, Utset TO, Niewold TB. Age- and gender-specific modulation of serum osteopontin and interferon-alpha by osteopontin genotype in systemic lupus erythematosus. Genes Immun. 2009;10(5):487–94.

25. Adrianto I, Wen F, Templeton A, Wiley G, King JB, Lessard CJ, et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nature genetics. 2011;43(3):253–8.

Publisher's Note

Genetic mapping across autoimmune diseases reveals shared associations and mechanisms

Matthew R Lincoln[1,2,3], Noah Connally[1], Pierre-Paul Axisa[1], Christiane Gasperi[1], Mitja Mitrovic[1], David van Heel[4], Cisca Wijmenga[5], Sebo Withoff[5], Iris H Jonkers[5], Leonid Padyukov[6], International Multiple Sclerosis Genetics Consortium, Stephen S Rich[7,8], Robert R Graham[9,10], Patrick M Gaffney[11], Carl D Langefeld[12,13], Timothy J Vyse[14], David A Hafler[1], Sung Chun[15,16], Shamil R Sunyaev[15,16], Chris Cotsapas[1,17,*]

[1]Department of Neurology, Yale School of Medicine, New Haven, CT
[2]Division of Neurology, Department of Medicine, University of Toronto, Toronto ON
[3]Keenan Research Centre for Biomedical Science, St. Michael's Hospital, Toronto ON
[4]Blizard Institute, Queen Mary University of London, London, UK
[5]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[6]Division of Rheumatology, Department of Medicine, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden
[7]Center for Public Health Genomics, University of Virginia, Charlottesville, VA
[8]Department of Public Health Sciences, University of Virginia, Charlottesville, VA
[9]Maze Therapeutics, South San Francisco, CA
[10]Genentech, South San Francisco, CA
[11]Genes and Human Disease Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK
[12]Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, NC
[13]Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC
[14]Department of Medical and Molecular Genetics, Kings College London, London, UK
[15]Division of Genetics, Brigham and Women's Hospital, Boston, MA
[16]Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[17]Department of Genetics, Yale School of Medicine, New Haven, CT
[*]Correspondence: cotsapas@broadinstitute.org

Abstract

Autoimmune and inflammatory diseases are polygenic disorders of the immune system. Many genomic loci harbor risk alleles for several diseases, but the limited resolution of genetic mapping prevents determining if the same allele is responsible, indicating a shared underlying mechanism. Using a collection of 129,058 cases and controls across six diseases, we show that ~40% of overlapping associations are due to the same allele. We improve fine-mapping resolution for shared alleles two-fold by combining cases and controls across diseases, allowing us to identify more eQTLs driven by the shared alleles. The patterns of sharing indicate widespread shared mechanisms, but not a single global autoimmune mechanism. Our approach can be applied to any set of traits, and is particularly valuable as sample collections become depleted.

Keywords

Autoimmune disease, Genetics, Molecular mechanism, Gene expression

Autoimmune and inflammatory diseases are a heterogeneous group of disorders, where activation of both the adaptive and innate immune system coupled with loss of self-tolerance leads to target tissue destruction[1]. These diseases are heritable, and genome-wide association studies (GWAS) have identified hundreds of susceptibility loci, confirming their polygenic nature[2,3]. Like other complex disease risk traits, heritability is strongly enriched in gene regulatory regions active in specific cell populations[4–6], suggesting risk is mediated to a large extent by altering gene expression in specific cell types under specific conditions. These diseases are also comorbid[7,8], with dual diagnoses being more frequent in individuals

than expected by chance, and multiple diseases aggregating in families[9]. We and others have shown that many genetic loci harbor risk variants for multiple autoimmune diseases[10–12], suggesting that comorbidity may be due to shared genetic liability and, hence, shared mechanisms of disease. In particular, a previous survey of five chronic inflammatory conditions found widespread sharing of risk effects, indicating the presence of such shared pathways[11].

Instances of pleiotropy, where the same variant influences risk to more than one disease, would by definition point to a shared molecular effect, and thus a shared mechanism. The limited resolution of genetic mapping has made it difficult to distinguish such cases from situations where distinct genetic variants in the same locus mediate risk to different diseases. This limited resolution restricts our ability to uncover shared pathogenic mechanisms, understand why modulating some immune functions can increase risk to one disease whilst decreasing risk to others, or make inferences about the origins of these diseases and their different prevalence rates around the world. Several methods have been developed to leverage pleiotropy to look for shared associations; when applied to autoimmune and inflammatory diseases, these methods have shown substantial sharing, which mirrors the overall shared heritability of these diseases[13–16].

An important driver of the limited resolution of genetic mapping is disease cohort sample size[17]. Currently available disease cohorts, most of which have been extensively studied already, are the result of decades-long international recruitment efforts. Meaningful increases in sample size are thus difficult to envision in the immediate future. An alternative way to increase sample size, and thus genetic mapping resolution, would be to jointly analyze cohorts across diseases. Previous studies have not taken this opportunity to explicitly increase fine-mapping resolution through pleiotropy, which we pursue here. In conventional meta-analyses of cohorts with the same disease, we assume that any associations are shared across strata. Most pleiotropy mapping methods also assume that effects are shared across diseases[18]; however, it is unclear whether this approach will work at loci that contain multiple associations that vary across diseases (e.g. *IL2RA*). We cannot make this assumption across diseases. It is thus crucial to ensure that the same allele drives risk to two or more diseases, rather than separate alleles in the same genomic locus.

Here, we first show substantial genome-wide shared heritability between autoimmune and inflammatory diseases. We then look at 224 instances where genetic associations to multiple diseases occur in the same genomic region, and show that 41.5% of these observed associations are due to pleiotropic variants, with the remainder being due to different alleles in the region. When we combine cases and controls across diseases to map each shared association, we increase fine-mapping resolution two-fold on average. This increase in resolution is meaningful, as it reveals new instances where a shared disease risk effect is pleiotropic with an immune cell subtype eQTL. Comorbidity is widespread between diseases of all organ systems, and sample sizes are limited, so this strategy is widely applicable beyond the immune-mediated diseases. Thus, this approach to careful dissection of shared effects can reveal mechanisms that are common across diseases and pinpoint key genes driving shared biology.

Results

Autoimmune and inflammatory diseases share heritability

We first assessed the evidence for genome-wide shared heritability between 17 autoimmune and inflammatory diseases from GWAS summary data. After quality control, we used LD score regression[19] to estimate heritability (h_g^2) for each trait (Supplementary Fig. 2a). We found that 11/17 diseases had sufficient heritability captured by common variants to make these comparisons meaningful (Z-score > 4)[20], so we restricted our analysis to this subset. We then calculated the proportion of shared heritability between each pair of diseases, again using LD score regression, which is robust to sample overlaps between cohorts[20]. We found a broad pattern of shared heritability (Fig. 1), with the strongest overlaps between atopic dermatitis, asthma and allergic traits ($0.51 \leq r_g \leq 0.91$), which may represent a shared

basis for atopic inflammatory disease. We saw a strong correlation between systemic sclerosis and systemic lupus erythematosus, which were also correlated with primary biliary cirrhosis ($0.42 \leq r_g \leq 0.86$). In line with our previous findings[10,21], these results indicate that autoimmune and inflammatory diseases share a substantial portion of genetic risk factors, even when accounting for the major histocompatibility locus (MHC), where overlapping haplotypes confer risk to different autoimmune and inflammatory diseases[22]. Overall, this suggests that some mechanisms are common between sets of diseases, but we find no evidence of universal sharing indicative of a large core autoimmune susceptibility component[10,21].

[Fig. 1: Joint analysis of shared autoimmune disease risk alleles improves fine-mapping two-fold.]

Autoimmune diseases share genetic associations

While this shared heritability gives an overall impression of the relationship between diseases, it cannot identify specific genetic risk factors—and thus, genes and pathways—shared between diseases. To compare samples from different collections genotyped at different centers, it was important to minimize batch effects by ensuring all samples were profiled on the same platform. We therefore chose six autoimmune and inflammatory diseases with large numbers of samples genotyped on the ImmunoChip[23] (celiac disease, inflammatory bowel disease, multiple sclerosis, rheumatoid arthritis, systemic lupus erythematosus and type 1 diabetes; Supplementary Fig.1). This targeted array interrogates variants in 188 known risk loci to saturation, representing only 1.9% of the genome but capturing 38-86% of risk loci that have been identified in the six diseases (Supplementary Fig. 2a). Using partitioned LD score regression, we confirmed that ImmunoChip regions account for 27.6% (MS) to 46.3% (CeD) of the estimated heritability for five of the six diseases for which GWAS data was available (Supplementary Fig. 3b). After quality control, removal of population outliers, resolution of duplicate and related samples, and imputation to the 1,000 Genomes reference haplotypes, we analyzed a total of 104,302 SNPs in 188 non-MHC genomic regions for association with disease in 82,630 cases and 104,573 controls (Supplementary Fig. 1).

We first identified associations across the 188 loci in each disease independently by assembling cases and controls into homogeneous population strata and meta-analyzing across these groups. As multiple independent associations at a locus have been described in all diseases, we used stepwise logistic regression followed by fixed-effects meta-analysis to allow for such effects. Results from fixed-effects and random-effects models were not meaningfully different (Supplementary Fig. 4). An orthogonal backward selection analysis with GCTA[24] recapitulated at least 90% of our findings (Supplementary Fig. 5). We found 197 independent associations in 123 different ImmunoChip loci at genome-wide significance ($P < 5 \times 10^{-8}$), and 361 associations at 166 loci with suggestive association evidence ($P < 10^{-5}$, Supplementary Fig. 2b). Overall, we find some level of support for essentially all known genome-wide significant effects in the ImmunoChip regions.

We found substantial evidence for multiple independent associations within loci, with 7% (RA) to 30% (IBD) of loci exhibiting more than one independent effect (Supplementary Fig. 2c). This included three instances of associations that have not been reported before (Supplementary Fig. 6). In celiac disease, we found suggestive unconditioned associations at two loci: a variant intronic to *ANKS1A* on chromosome 6 (rs12206298; $P = 4.1 \times 10^{-7}$), and a variant intronic to *CTSH* on chromosome 15 (rs3784539; $P = 1.3 \times 10^{-5}$). After conditional association, both these associations passed the genome-wide significance threshold ($P = 4.9 \times 10^{-8}$ and 1.1×10^{-8} respectively). We found evidence of a second, independent effect in each locus (rs4713844, $P = 9.9 \times 10^{-8}$; and rs7181033, $P = 8.7 \times 10^{-5}$). Similarly, in IBD, we found that a suggestive association in the *CLEC16A* locus on chromosome 16 (rs7201325, $P = 1.4 \times 10^{-7}$) reached genome-wide significance after conditioning ($P = 1.1 \times 10^{-10}$), with evidence of a secondary, independent effect (rs55773334, $P = 7.6 \times 10^{-5}$). The presence of multiple masked independent effects highlights the need to look carefully at suggestive associations.

Having ensured we were capturing most of the known associations in ImmunoChip loci for each of the six diseases, we looked for shared effects across diseases, i.e. whether the same variant mediates risk to more than one disease. We found 218 overlapping conditionally independent associations at 98 loci (a lead variant associated to one disease at $P < 10^{-5}$, and a lead variant for another disease $P < 10^{-4}$; both lead variants being in LD $r^2 > 0.5$ with at least one common SNP). Using joint likelihood mapping (JLIM), we found evidence of a shared effect in 90/218 (41.3%) such overlaps, involving a total of 56 conditionally independent shared effects spanning 52 unique loci (Fig. 1b). Of these, 42 effects were shared between two diseases, nine between three, and five between four diseases (Fig. 1c). Unlike previous reports, which could not distinguish between shared and distinct associations with multiple diseases in a locus, these observations indicate that many mechanisms are shared between autoimmune and inflammatory diseases.

We found three loci where multiple conditionally independent associations for one disease were shared. In the *STAT4* locus, we found two independent effects each for RA and SLE were shared (Supplementary Fig. 7a). In the *CD28-CTLA4* locus, one T1D risk association near *CD28* is shared with CeD, whereas another, an intronic variant in *CTLA4* is shared with RA (Supplementary Fig. 7b). In the *TYK2* locus, one RA risk association is shared with SLE, IBD, and T1D; a second association, localizing to *ICAM3*, is shared with SLE alone (Supplementary Fig. 7c). Cumulatively, these examples demonstrate that disease-associated alleles in the same locus can have different consequences, and that careful comparisons across diseases can distinguish each effect.

Shared associations improve fine-mapping resolution

We next assessed if joint analysis across diseases could improve fine-mapping resolution. For each of the 56 shared associations, we assembled conditionally independent association data across all disease cohorts sharing that association, and combined them with fixed-effects, inverse variance-weighted meta-analysis. In a subset of loci, we saw an unexpected decrease in significance and increase in heterogeneity in the meta-analysis; we found these to be shared associations with opposite effects, where an allele increases risk for one disease and decreases it for the other (Supplementary Figs. 8-16). In five of these nine cases, variants with opposing effects were shared between MS and IBD. After inverting the association statistics to account for these effects, our meta-analysis resulted in higher significance for 122/131 (93.1%) associations across all 52 loci harboring a shared effect, demonstrating the potential to bolster association findings with our approach.

To establish if this increase in sample size provides a meaningful increase in fine-mapping resolution, we used FINEMAP[25] to calculate posterior inclusion probabilities for SNPs at each of the 56 shared effects, both in individual diseases and in the cross-disease meta-analysis. We then calculated 95% credible sets for each disease, both before and after cross-disease meta-analysis. We found a substantial decrease in the mean credible interval size, from 36.6 (s.d. 46.8) to 16.5 (s.d. 20.0), representing an improvement of 55% (Fig. 1e). We saw resolution improvement across the spectrum of initial association evidence, with the largest gains where an effect had relatively weak evidence of association in a disease: for associations below genome-wide significance in a single disease, our resolution increased from a mean of 50.8 SNPs to 18.0 SNPs after cross-disease meta-analysis; for associations already above genome-wide significance in a single disease, we saw improvement from a mean of 21.8 SNPs to 14.9 SNPs. This is exemplified by a shared association in the *C1orf106* locus on chromosome 1, where credible intervals of 28, 8, and 11 SNPs for CeD, IBD and MS respectively are reduced to eight variants in very tight linkage disequilibrium (minimum $r^2 = 0.976$) on cross-disease meta-analysis (Fig. 2). In this case, there are genome-wide significant associations in each disease independently, but increasing sample size from symmetric equivalent 19,026 (CeD), 53,312 (IBD), 35,618 (MS) to a cross-disease meta-analysis 93,001 (symmetric equivalent) increases the resolution for both CeD and MS, identifying a core risk haplotype within *C1orf106*. As the eight variants in this haplotype are in near-perfect LD, we may have reached the limit of fine-mapping resolution at this locus using samples of a single ancestry.

[Fig. 2: A shared effect on chromosome 1 can be fine-mapped to eight variants across celiac disease, inflammatory bowel disease, and multiple sclerosis.]

Shared associations indicate common mechanisms

The ultimate promise of increasing fine-mapping resolution is to increase the interpretability of association signals. We and others have shown that disease risk associations are enriched in non-coding regions with gene regulatory potential[4,5,26,27]. We have used the JLIM approach to show that autoimmune disease associations are sometimes shared with expression quantitative trait locus (eQTL) signals[28], indicating the risk allele also influences gene expression. However, most associations are not shared with an eQTL, nor are they attributable to coding variants. To assess if this is due to limitations in fine-mapping resolution, we looked for shared associations between the 56 shared effects we discovered and *cis*-eQTLs for nearby genes in naïve T cells, monocytes and neutrophils in the BLUEPRINT[29] dataset. We found 137 shared effects between each of 131 shared conditionally independent association signals in a single disease and eQTLs for nearby genes. We then looked for shared effects between the better-powered cross-disease meta-analysis data in each of the 52 loci, and can attribute 19 new disease/eQTL effects to the underlying diseases (Fig. 3; Supplementary Table 6). Most of the implicated eQTLs are present in only one of the three cell types we interrogated, with T cells providing the largest number. We also exclude 13/137 disease/eQTL shared effects as no longer relevant because we do not find evidence of shared association between the cross-disease meta-analysis and eQTL data. Our gains primarily occur in cases where the cross-disease meta-analysis reduces the credible interval size (Fig. 3c), indicating that this gain of resolution drives these new observations.

The direction of shared eQTL effects indicate whether we should expect increases or decreases in expression for those genes to increase disease risk. We reasoned that we might also see the same direction of effect between cases and controls, where the risk state is magnified. We therefore looked at single cell RNA-seq data derived from T cells collected from a cohort of MS patients and healthy controls[30]. After quality control, we were able to detect twelve genes that were targets of eQTLs shared with MS risk signals in our analysis. We found a significant pattern of correlation (P = 0.018): when a disease risk allele increased expression of a target gene, we saw higher expression in cases than in controls, and when it decreased expression we saw lower levels in cases than in controls (Fig. 3d). This suggests that shared associations that drive risk-altering changes to gene regulation do in fact alter disease risk, and our results are uncovering pathogenic mechanisms.

[Fig. 3: The increased resolution of fine-mapping shared associations across diseases allows identification of more disease-eQTL overlaps.]

The relative direction of the disease and eQTL associations can also suggest specific mechanistic hypotheses. This is exemplified by an association in the *RGS1* locus, shared between celiac disease and MS (Fig. 4). *RGS1* encodes a regulator of G-protein mediated signaling active in immune cell populations. We had previously reported a shared association between an *RGS1* eQTL in macrophages and both MS and CeD risk. We now show that the effect is shared between the two diseases; this better fine-mapped shared effect also overlaps with *RGS1* eQTLs in multiple cell populations, not just macrophages. The cross-disease meta-analysis reduces the credible interval to 10 variants overlapping the promoter region of *RGS1*. The lead credible interval variant overlaps a region of accessible chromatin within an active enhancer immediately upstream of the *RGS1* promoter. Further, this variant lies in a predicted binding site for *ZNF263*, and position-weight matrix analysis suggests the minor allele abrogates binding[31].

[Fig. 4: Jointly analyzing an association shared between multiple sclerosis and celiac disease improves fine-mapping resolution and identifies a shared eQTL for *RGS1*.]

Discussion

We have quantified the shared heritability between autoimmune and inflammatory diseases, and demonstrated that we can leverage this to identify genetic variants that alter risk to multiple diseases. In previous work, Ellinghaus *et al.* looked for shared effects between five chronic inflammatory diseases, of which only the subsets of IBD are in common with our work[11]. We observe, as they did, widespread sharing between diseases. Uniquely, we then meta-analyze across diseases and show this significantly increases fine-mapping resolution, compared to considering each disease in isolation: the number of effects where a single variant explains 95% of the posterior probability of association increases from 13 to 20 (a 54% increase); for 50% of the posterior probability, we see an increase from 31 to 54 (74%). Furthermore, we see an increase in the number of eQTLs, with evidence of sharing an effect with disease risk, from 137 to 143 (4.4%). Thus, in terms of identifying causal variants and functional interpretation, meta-analyzing across diseases meaningfully increases our ability to interpret genetic associations. This sets the stage for variant-to-function efforts to uncover key pathogenic mechanisms, as we provide high-value targets relevant to multiple diseases.

This approach can be applied to any set of traits sharing associations; we therefore suggest this is a fruitful avenue to maximize the interpretability of existing genetic studies of human complex traits, especially as shared mechanisms are applicable to multiple conditions. It is particularly valuable as sample collections, particularly of diseases that are difficult to diagnose or not especially common in the population, become depleted. Disease cohorts are often genotyped on different platforms, and the majority of common variants imputed. This can introduce a substantial bias, if cohorts of samples with different diseases have differential genome coverage. We have avoided this in our study by using a common platform, at the expense of not covering the entire genome. These technical hurdles will diminish as genotyping platforms coalesce around a standard set of variants, and as the community shifts to whole-genome sequencing rather than genotyping. We note that biological interpretation of genetic associations, shared or otherwise, is dependent on access to molecular and cellular phenotype studies such as eQTLs, which require profiling a wide array of tissues or cell types under diverse stimuli in order to identify the consequences of disease-associated variants. The BLUEPRINT dataset, which we used here, covers three very different blood cell types, but dozens more exist, in which the variants we have identified could act. This context specificity may be one reason we cannot always assign a cognate eQTL to each well-resolved association[32].

In terms of understanding the common mechanisms in autoimmunity, we and others have reported that many loci harbor associations to multiple autoimmune diseases. However, these approaches have relied on simple proximity of variants to infer that the underlying mechanisms must be shared. We have quantified the shared heritability between autoimmune and inflammatory diseases, and shown that a substantial proportion of shared loci harbor pleiotropic effects influencing risk to multiple diseases, which represent shared mechanisms. Many loci, however, harbor multiple independent effects, indicative of distinct mechanisms driving risk to different diseases; this is consistent either with the same underlying genes being influenced in different contexts to induce risk for different diseases, or with different genes which happen to be encoded near each other. Previous studies by us and others were not designed with this resolution[10], and could only identify loci harboring potentially different effects to multiple diseases.

Our results reveal complex patterns of shared heritability between autoimmune diseases. In particular, we find many opposite effects shared between IBD and MS, where the same allele increases risk for one disease but decreases risk for the other. This is reminiscent of the differential outcomes of anti-TNFα therapies, which are beneficial in IBD but exacerbate MS symptoms[33], as initially suggested by dissection of risk effects impacting the TNF receptor 1[34]. Further, it suggests that some disease mechanisms may have an optimum state, and either hypermorphism or hypomorphism are deleterious, as previously suggested[15]. However, these two diseases also have the largest number of cases in our analysis. We therefore cannot completely exclude the possibility that opposing effects are widespread but we lack

power to detect them. Overall, we see no evidence for a substantial component of risk shared across all six diseases, which would be indicative of a pan-autoimmunity mechanism. Our benchmarking suggests this is not due to a lack of power to detect shared effects[28], and our results strongly support independent effects in most loci. As our results argue against a single, shared autoimmune mechanism, they also dispute a single evolutionary origin for autoimmune and inflammatory diseases, which would have resulted in a set of risk alleles driving broad autoimmunity[16].

Methods

Shared heritability analysis

We downloaded complete summary statistics for all autoimmune and inflammatory disease GWAS available in the NHGRI-EBI GWAS catalog[35]. We focused on European ancestry studies with at least 2,000 subjects for which signed summary statistics were available. Where multiple studies were available for a given trait, we chose the study with the largest cohort size. By applying these filters, we obtained GWAS statistics for atopic dermatitis (AtD)[36], allergic traits (All)[37], asthma (Ast)[38], celiac disease (CeD)[39], eosinophilic granulomatosis with polyangiitis (EGPA)[40], selective IgA deficiency (sIgAD)[41], inflammatory bowel disease (IBD)[3], latent autoimmune diabetes in adults (LADA)[42], primary biliary cirrhosis (PBC)[43], primary sclerosing cholangitis (PSC)[44], psoriatic arthritis (PsA)[45], systemic lupus erythematosus (SLE)[46], systemic sclerosis (SSc)[47], and vitiligo (Vit)[48]. IBD summary statistics also included results for Crohn's disease (CD) and ulcerative colitis (UC); as expected, these exhibited high correlation as they share some, but not all of their genetic architecture (data not shown). We downloaded summary statistics for psoriasis (Ps)[49] from dbGaP and summary statistics for rheumatoid arthritis (RA)[50] from GRASP. We obtained multiple sclerosis (MS) summary statistics from the International MS Genetics Consortium[2]. Sources and accession numbers for included studies are documented in Supplementary Table 2.

We first removed indels and single nucleotide polymorphisms (SNPs) inconsistent with the 1,000 Genomes Project (Phase 3) reference panel[51]. We next filtered for strand-unambiguous biallelic SNPs with minor allele frequency (MAF) > 0.01 in the 1,000 Genomes European (EUR) reference subjects. Following Bulik-Sullivan et al.[20], we removed variants with INFO < 0.9 where this information was available. As INFO scores were not available for most datasets, we uniformly filtered on SNPs present in the HapMap 3[52] reference panel. Where differing effective sample sizes were provided for each variant, we removed SNPs genotyped in fewer than two-thirds of the 90th percentile population size.

After quality control, we used linkage disequilibrium (LD) score regression[19] to estimate heritability (h_g^2) for each trait from summary statistics, using the 1,000 Genomes EUR individuals as reference. As recommended by the developers[20], we excluded traits with heritability Z-scores < 4 from further analysis. We next used LD score regression to calculate correlations (r_g) among all pairs of the remaining traits.

ImmunoChip datasets

We obtained raw ImmunoChip genotypes for six autoimmune and inflammatory diseases (Supplementary Fig. 1; Supplementary Table 2), including CeD[53], IBD[54], MS[55], RA[56], SLE[57,58] and type 1 diabetes (T1D)[59]. Each of the participating disease consortia provided data, including two separate SLE consortia (OMRF and Genentech). For the CeD and SLE datasets, we used GenomeStudio to call genotypes from intensity files.

After resolving conflicting SNP nomenclature and allelic encoding across datasets, we lifted these over from GRCh36 (hg18) to GRCh37 (hg19). We excluded SNPs that could not be mapped unambiguously to the newer assembly.

All datasets consisted of multiple strata, typically divided by country of origin (Supplementary Table 3). We therefore divided datasets into country-level strata and performed quality control independently within

each stratum. As one of the T1D datasets consisted of affected sibling pairs, we processed this separately.

Genotype quality control

Quality control and association analysis are discussed in detail in the Supplementary Materials and here in brief. We used PLINK[60] to perform initial quality control. Within each stratum, we first removed individuals missing >10% of genotypes, and SNPs that were missing in >5% of individuals. We then assessed the remaining samples for sex inconsistencies. Where X chromosome genotypes were available, we calculated X chromosome homozygosity for each individual. We then used Mclust[61] to divide samples into male and female clusters assuming a Gaussian mixture model with two components. Inferred sex was used where these were not specified in the original datasets. Individuals were removed if their recorded sex differed from the model-inferred sex.

We focused our analysis on individuals of European descent. To identify population outliers, we merged our genotype data with reference data from the 1,000 Genomes Project[51] and performed principal component analysis[62]. We removed samples with a smaller Euclidean distance to the EAS or AFR centroids than to the EUR centroid. We then repeated principal component analysis and removed samples with a smaller Euclidean distance to the SAS centroid than to the EUR centroid. At each step, samples that did not correspond to an identifiable reference population were removed empirically.

After removing population outliers, we removed SNPs exhibiting deviation from Hardy-Weinberg equilibrium expectation ($P < 10^{-8}$). We next identified and removed subjects with extreme homozygosity. We used PLINK to calculate inbreeding coefficients (F) for each individual. Within each stratum, we removed individuals with F > 2.5 s.d. from the stratum mean. We next applied a second, more stringent filter for missing values, removing individuals with >1% missing data and SNPs missing in >1% of individuals.

We then identified close relatives ($\hat{\pi} > 0.185$) and duplicates ($\hat{\pi} > 0.90$) within each disease dataset. Duplicates were removed from further analysis. Relatives were excluded from a second Hardy-Weinberg equilibrium assessment, where we filtered SNPs that violated Hardy-Weinberg equilibrium at $P < 10^{-5}$. We included relatives to provide additional chromosomes for phasing and imputation; we removed these before association testing. After quality control, we excluded strata with fewer than 150 remaining cases or controls.

A total of 168,928 subjects were available for analysis after quality control. After identifying and removing control samples that were shared among studies, we were left with 129,058 unique individuals. The numbers of subjects supplied and analyzed are indicated in Supplementary Fig. 1. More detail on the numbers of subjects per stratum are given in Supplementary Table 3. A detailed accounting of all subjects (Supplementary Table 4) and SNPs (Supplementary Table 5) can also be found in the Supplementary Data.

Imputation and association analysis

Before imputation, we removed indels, rare SNPs (MAF < 0.05) and SNPs that were missing differentially between cases and controls ($P < 10^{-5}$). We then used SHAPEIT2[63] to remove SNPs that were inconsistent with the 1,000 Genomes (Phase 3) reference haplotypes and to phase remaining SNPs. We used IMPUTE2[64] to impute reference SNPs that were not genotyped, and to fill in sporadically missing genotypes. A subset of genotyped SNPs could not be reliably imputed from the reference haplotypes (concord_type0 < 0.75 despite info_type0 > 0.8). As these were either mis-mapped or unreliably genotyped, we excluded these SNPs and performed a second round of imputation as above.

We removed imputed SNPs if they did not have two alleles present, if they were imputed with INFO < 0.75, or if they had MAF < 0.05. We also removed variants that violated Hardy-Weinberg equilibrium at $P < 0.001$, and those that exhibited differential missingness between cases and controls at $P < 0.01$.

We used SNPTEST[65] to perform logistic regression of imputed genotype dosages against phenotype in each stratum, incorporating the first two principal components as covariates into an additive model. We then combined association statistics into a fixed-effects, inverse variance-weighted meta-analysis[66] for each disease. The extended MHC (6:28-34 Mb, GRCh37 coordinates) was excluded from analysis.

To allow for multiple independent effects at a given locus, we used iterative stepwise conditional logistic regression. For each iteration after the first, we repeated logistic regression in each stratum, this time conditioning on all previously identified meta-analysis lead SNPs with $P < 0.0001$. Results were again combined in a fixed-effects meta-analysis. We restricted our search for lead variants to SNPs present in all strata, with $I^2 < 50$ and $r^2 < 0.9$ to all previous lead SNPs. Where such a lead SNP could be identified, we added this to the list of conditioning variants and proceeded with another round of association testing. We continued conditioning until we detected three independent signals or no variants with $P < 0.0001$ remained.

Our iterative conditioning approach produced a set of independent associations for each trait at each ImmunoChip locus. To identify conditionally independent association signals at each locus, we iterated over the set of lead variants, this time conditioning on the all-but-one variant. For this analysis, we again required SNPs to be present in all strata, with $I^2 < 50$. We validated our forward selection results using a reverse selection method implemented in GCTA. As this analysis produced results nearly identical to our reverse selection model (Supplementary Fig. 5), we used the reverse stepwise conditioning results in subsequent analyses.

Identification of shared genetic effects

We used Joint Likelihood Mapping (JLIM)[28] to identify genetic effects that are shared across multiple diseases at each ImmunoChip locus. The method relies on permutation of genotype-level data, so we restricted our analysis to diseases with data available for large numbers of samples. We wished to analyze trait pairs exhibiting at least moderate strength of association; we therefore identified pairs of traits at each locus with lead variants significant at $P_1 < 0.00001$ and $P_2 < 0.0001$. To ensure a moderate degree of linkage disequilibrium between assessed traits, we identified the set of variants with $r^2 \geq 0.5$ to each lead variant; trait pairs were assessed for a common underlying genetic effect where these sets shared at least one variant. To this initial set of candidates, we added additional trait pairs that appeared similar based on their Manhattan plots.

For each analyzed trait pair, we identified and removed shared controls before JLIM analysis. We then applied JLIM in both directions, using each trait alternately as the primary and secondary trait. We set the analysis window to be the maximal coordinates of the union of the $r^2 \geq 0.5$ windows; resolution was set to the default $r^2 = 0.8$. Linkage disequilibrium in the primary trait was estimated from 1,000 Genomes reference data; for the secondary trait, we estimated this directly from best-guess genotypes.

We estimated JLIM significance by permutation. For each permutation, we shuffled phenotype labels independently in each disease stratum and repeated logistic regression and meta-analysis as above. A minimum of 10,000 permutations were performed for each trait pair.

To identify clusters of traits sharing a common genetic effect, we analyzed pairwise JLIM results as graphs. Edges were defined between traits where JLIM was significant at $P < 0.05$. Maximal connected undirected subgraphs were then identified at each locus. For subgraphs of size greater than two, we

repeated JLIM for each ordered pair of traits in the subgraph, this time using a common analysis window defined by the union of the $r^2 \geq 0.5$ windows for all traits in the subgraph.

We validated our results by comparing with coloc2 (Supplementary Table 1). We identified a single false positive (the *SOCS1* locus) and removed this from our analysis.

Fine-mapping of susceptibility loci

For each cluster of traits sharing a common genetic effect, we combined data by meta-analysis. Duplicate samples were identified and removed from each cluster. We then repeated logistic regression and meta-analysis as described above. We used the I^2 statistic to assess variants for heterogeneity. Within each disease, variants were excluded if $I^2 \geq 50$, or if they were present in fewer than half of the constituent strata. To be included in fine-mapping analysis, variants were required to have survived filtering in all diseases of a given cluster.

Trait pairs with opposing effect directions were identified by linear regression. For each pair, we regressed SNP Z scores for the first trait against corresponding Z scores for the second trait. We considered trait pairs to have opposing effect directions when their slope term was negative and statistically significant. For such trait pairs, we reversed the direction of the effect for one trait and repeated meta-analysis. Opposing trait pairs were confirmed by comparing heterogeneity statistics before and after reversal.

For each shared effect cluster, we used FINEMAP[25] to estimate posterior inclusion probabilities for each variant within our shared effect clusters. SNP correlation matrices were calculated from genotypes for each trait. We assumed a single causal variant at each locus and performed an exhaustive search. We quantified fine-mapping improvement by comparing the number of SNPs in 95% credible intervals for individual disease traits, and for meta-analyzed clusters.

Expression quantitative trait locus (eQTL) data

We obtained and quantitated raw RNA-sequencing reads from three human immune cell types from the BLUEPRINT consortium[29]: neutrophils (CD66b+ CD16+; 196 individuals), monocytes (CD14+ CD16-; 193 individuals), and naïve CD4 T cells (CD4+ CD45RA+; 169 individuals). Subjects in this study were ascertained to be free of disease and were representative of the United Kingdom population. We used the GTEx Analysis V8 pipeline (https://gtexportal.org)[67] to align FASTQ files, filter for quality control and quantitate gene expression. Briefly, we used STAR v2.5.3a to align reads to GRCh38. We quantitated expression to the gene level with RNA-SeQC v1.1.9, using the GENCODE 26 gene model. We included genes with expression values >0.1 TPM and ≥6 reads in at least 20% of samples; we then normalized counts using the trimmed mean of M-values (TMM) method implemented in edgeR[68]. We then normalized expression across samples using an inverse normal transformation. We retained all samples for analysis as none had fewer than the minimum 10 million reads.

We obtained genotype data for all individuals with available gene expression data; a total of 7,008,524 variants were available. Whole-genome sequencing, alignment, variant calling and quality control were performed previously[29]. All SNPs were biallelic. We removed indels and SNPs with MAF < 0.05 or that violated Hardy-Weinberg equilibrium at $P < 0.00001$. There were no heterozygosity outliers, defined as samples with heterozygosity > 5 standard deviations from the sample mean. Similarly, there were no cryptic relatives ($\hat{\pi} > 0.1875$) or population outliers (>4 standard deviations in the first four PCs). A total of 4,853,096 variants were available for analysis in 197 subjects.

Identification of shared susceptibility-eQTL loci

We used JLIM to identify shared eQTL–disease susceptibility loci. Disease susceptibility summary statistics were lifted over to GRCh38 coordinates. For each shared effect, we assessed all genes with transcription start site within 1 Mb of any susceptibility lead SNP. We regressed normalized expression

values for these genes against genotype in a linear model, assuming an additive model of inheritance. We used covariates to adjust for age, sex the first 5 principal components and 30 PEER factors[69]. The same covariates were used to generate permutation data for JLIM.

To allow for multiple independent eQTLs within a given locus, we performed conditional cis-eQTL analyses. For eQTL with $P < 0.001$, we repeated linear regression modelling, this time conditioning on the lead SNP from the first model. We continued adding lead variants to our model until either (a) the lead variant $P \geq 0.001$ or (b) three conditioning SNPs had been included. To identify conditionally independent eQTL signals, we again iterated on the set of lead variants, this time conditioning on the all-but-one variant.

After identifying conditionally independent eQTL signals for each gene, we used JLIM to assess for a common underlying genetic effect between disease susceptibility loci and eQTLs. We lifted summary statistics for susceptibility loci over to GRCh38 and used these as primary traits. Expression QTLs were used as secondary traits. For each primary trait, the JLIM analysis window was chosen to be the union of all SNPs ±100 kb from the lead SNP. We estimated significance by permuting eQTL expression values 100,000 times for each trait. Within a given cluster of disease associations, we used the Benjamini-Hochberg procedure to correct *P*-values for the number of genes and cell types assessed.

Single sell RNA-seq quality control

We obtained raw sequencing data from a previously-published single cell RNA-seq (scRNA-seq) study of multiple sclerosis[30]. Sample collection and data preparation are described in detail in the original publication. Briefly, peripheral blood mononuclear cells (PBMCs) and cerebrospinal fluid (CSF) cells were obtained from 6 healthy donors and 5 new-onset multiple sclerosis patients. For each donor, single cell suspensions were prepared for analysis using the 10x Genomics platform. We used unique molecular identifier (UMI) count matrices as described[30]. We filtered extreme outliers by excluding droplets with (a) <1000 UMI counts or <500 unique genes detected, or (b) >15,000 UMI counts or >5,000 genes detected. To exclude low-quality cells and potential doublets from our analysis, we examined the distributions of UMI counts and number of detected genes per cell. As distributions of these parameters varied across emulsions, we quantile-normalized $\log_{10}$-transformed UMI counts and $\log_{10}$-transformed number of detected genes per cell. Using quantile-normalized values and the percentage of counts mapping to mitochondrial genes, we excluded low-quality cells with <2,000 UMI counts, <900 genes detected, or >12.5% counts mapping to mitochondrial genes. We also excluded doublets with >8,000 UMI counts or <2000 genes detected.

Dimensionality reduction and clustering. For cells passing quality control, we normalized UMI counts using a count per million approach, dividing each count by the total number of counts per cell. We then multiplied normalized counts by 10,000 and added a pseudo count of 1 before log-transformation. We then applied a variance-stabilizing transformation (VST) to account for variation in gene expression levels across the dataset, and used genes with stabilized variance >1 and stabilized mean expression >10^-3^ as input for principal component analysis (PCA). Genes mapping to the T cell receptor (TCR), the B cell receptor (BCR) and the Y chromosome were excluded from PCA. We computed the first 50 principal components (PCs) using a partial singular value decomposition method, based on the implicitly restarted Lanczos bidiagonalization algorithm (IRLBA), as implemented in the Seurat package[70].

To correct for systematic differences across samples we applied Harmony integration[71] to the first 50 PC loadings. We then retained the first 30 harmony-corrected PCs, and used PC loadings as input for visualization using UMAP (minimum distance = 0.5, spread = 10), and clustering by applying the Louvain algorithm to a shared nearest neighbors (SNN) graph (resolution = 0.01), as implemented in Seurat. This low-resolution clustering separated T and NK cells from B cells and monocytes. We then selected T and NK cells and re-applied the same pipeline to the raw UMI counts to obtain a dedicated UMAP visualization

and clusters of T cells (SNN k = 20 and Louvain resolution = 0.5), enabling us to distinguish between different T cell sub-populations. Using normalized log-transformed UMI counts, we computed the area under a receiver operating curve (auROC) to define diferentially expressed genes between each cluster pairs. Manual inspection of gene markers enabled us to define several sub-populations (by order of abundance): central memory CD4 T cells (cluster 0: *CD4*, *CXCR5*, *LTB*, *KLRB1*), naïve CD4 T cells (cluster 1: *TCF7*, *CCR7*, *LEF1*, *CD4*, *STAB1*, *TSHZ2*, *NPM1*, *SELL*), central memory CD8 T cells 1 (cluster 2: *CCL5*, *CD8A*, *GZMA*, *NKG7*, *GLNY*, *CD8A*), naïve CD8 T cells (cluster 3: *TCF7*, *NELL2*, *SELL*, *CCR7*, *LEF1*, *CD8A*, *CD8B*), effector CD8 T cells 1 (cluster 4: *CCL4*, *CCL5*, *GZMA*, *CST7*, *PRF1*, *CD8A*, *CD8B*), natural killer cells (cluster 5: *KLRG1*, *NKG7*, *PRF1*, *KLRB1*, *GZMK*), regulatory T cells (cluster 6: *FOXP3*, *IL10RA*, *TIGIT*, *CD4*), gamma-delta T cells (cluster 7: *TRDC*, *KLRB1*, *GLNY*, *KLRC1*, *CCL5*, *GZMA*, *PRF1*), T follicular helper CD4 T cells (cluster 8: *FAU*, *FTH1*, *VIM*, *CD4*), megakaryocytes (cluster 9: *NRGN*, *PPBP*, *TUBB1*, *SPARC*), type I interferon activated CD4 T cells (cluster 10: *MX1*, *ISG15*, *IRF7*, *XAF1*, *IFI6*), central memory CD8 T cells 2 (cluster 11: *ZNF683*, *CD7*, *KLRC3*, *LEF1*, *CD8A*, *CD8B*), central memory CD8 T cells 3 (cluster 12: *TCF7*, *CD27*, *GZMK*, *CD8B*).

Pseudo-bulk analysis. We used cluster assignments to sum UMI counts across cell types, disease status, tissue and donor. For each cell type in the blood compartment (PBMCs), we used a negative binomial distribution with a local fit, as implemented in DESeq2[72] to model gene expression and test differences between cases and controls, while controlling for sex as a covariate. We used shrinkage to account for $\log_2$-fold change inflation on genes with low counts and used shrunken $\log_2$-fold change for subsequent analyses. We focused on cluster 0 for validation of T cell eQTL predictions as this cluster was most abundant.

Resource availability

Data availability. This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in Supplementary Table 2.

Code availability. Code used in this analysis is available on GitHub (https://github.com/cotsapaslab/CrossDiseaseImmunochip).

Acknowledgements

We thank the EAGLE eczema consortium for providing GWAS summary statistics. This research utilizes resources provided by the T1DGC, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development, and JDRF and supported by U01 DK062418. We thank the RACI consortium for access to RA data and the International IBD Genetics Consortium for access to IBD data. De-identified data were provided from a total of 4617 samples (2563 SLE cases and 2054 population controls) in the Lupus Family Registry and Repository collection[73] at the Oklahoma Medical Research Foundation. The SLE Genentech samples were originally genotyped and analyzed as part of a large SLEGEN Consortium ImmunoChip study[57]. The Alliance for Lupus Research (now Lupus Research Alliance) provided funds for the SLE ImmunoChip study.

MRL is supported by a Career Transition Fellowship from the Consortium of MS Centers and the National MS Society. CG received a research fellowship from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for this project. She further received funding from the Hans und Klementia Langmatz-Stifung and the Hertie Network of Excellence in Clinical Neuroscience, not related to this study. CW was supported by an ERC Advanced grant (FP/2007-2013/ERC grant 2012-322698), the NWO Spinoza prize grant (NWO SPI 92-266), a grant from Stiftelsen K. G. Jebsen, and The Netherlands

Organ-on-Chip Initiative - an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of The Netherlands. SW was supported by The Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of The Netherlands. I.H.J. is supported by a Rosalind Franklin Fellowship from the University of Groningen and an NWO VIDI grant (No. 016.171.047).

Author contributions

MRL, NC, CG, and MM curated and analyzed data;
DvH, CW, SW, IHJ, LP, IMSGC, SSR, RRG, PMG, CDL, TJV and DAH provided data;
MRL and CC wrote and edited the manuscript, with input from all co-authors;
SC and SRS designed and implemented analytical methods;
CC conceptualized and oversaw the project.

Competing interests

The authors declare no competing interests.

Figure Legends

Fig. 1: Joint analysis of shared autoimmune disease risk alleles improves fine-mapping two-fold. (a) We find broad genome-wide correlation between association statistics for susceptibility to thirteen autoimmune and inflammatory diseases for which genome-wide association data were available (allergy, All; asthma, Ast; atopic dermatitis, AtD; celiac disease, CeD; inflammatory bowel disease, IBD; multiple sclerosis, MS; primary biliary cirrhosis, PBC; rheumatoid arthritis, RA; systemic lupus erythematosus, SLE; systemic sclerosis, SSc; and vitiligo, Vit). Alleles were not available for type 1 diabetes (T1D). (b) This correlation is reflected in many loci harboring risk alleles to more than one of six diseases with available ImmunoChip data (lower triangle). Of these 218 pairs of associations, 90 are driven by the same underlying allele (upper triangle). (c) Risk alleles are mostly shared between two diseases (42 cases), with nine shared between three, and five between four diseases. (d) Nine shared alleles have opposite effect directions, increasing risk of one disease and decreasing risk of another. This is most frequent between MS and IBD. (e) Combining cases and controls across diseases increases fine-mapping resolution for these shared associations. We assess resolution as the number of variants required to explain 95% of the posterior probability of association. This credible interval decreases by 55% when combining samples across diseases (pink) compared to using only samples for one disease (orange). Associations that are not shared across diseases have similar credible interval distributions in individual diseases (teal). Outlying values (triangles) are winsorized; sizes of these credible intervals are indicated.

Fig. 2: A shared effect on chromosome 1 can be fine-mapped to eight variants across celiac disease, inflammatory bowel disease, and multiple sclerosis. (a) Overlapping associations in celiac disease, inflammatory bowel disease and multiple sclerosis on chromosome 1, with 95% credible intervals varying both in number of variants and physical span. (b) For each pair of diseases, the strength of association (vertical axis) for the first trait decays in a linear fashion as a function of r^2 to the lead SNP in the second trait, consistent with a shared causal variant. (c) Meta-analyzing across the three diseases gives a stronger association signal, which can be fine-mapped to a narrow interval within *C1orf106*. (d) We find strong pairwise evidence that the association is shared between all three diseases; JLIM is asymmetric, so we run comparisons in both directions.

Fig.3: The increased resolution of fine-mapping shared associations across diseases allows identification of more disease-eQTL overlaps. (a) We looked for shared effects between disease associations and expression QTLs in loci harboring shared disease effects. When considering each disease separately, we find 137 significant disease—eQTL overlaps across monocytes, neutrophils and T cells from the BLUEPRINT consortium (left panel). When comparing eQTLs to cross-disease meta-analyses, we find new overlaps (blue, middle panel) and no longer find evidence for some eQTLs (red, middle panel), for a grand total of 143 disease-eQTL overlaps (4.4% net discovery increase, right panel). (b) Some of the shared eQTL effects can be detected in multiple tissues, but most are restricted to a single cell type, indicating substantial effect specificity. (c) We find new eQTL shared effects in loci where the cross-disease meta-analysis decreases the credible interval substantially, suggesting this resolution drives new discoveries. Disease associations where an eQTL is lost after meta-analysis also have smaller credible intervals, suggesting these may have been false positive findings due to lack of resolution in individual disease datasets. (d) The effects of risk-increasing shared alleles on gene expression is mirrored in expression differences between multiple sclerosis cases and controls. This suggests that risk states imparted due to small changes in gene expression persist during active disease, and provide validation that our eQTL discoveries are relevant to pathogenesis.

Fig. 4: Jointly analyzing an association shared between multiple sclerosis and celiac disease improves fine-mapping resolution and identifies a shared eQTL for *RGS1*. (a) Overlapping associations for the two diseases are due to a shared effect (JLIM $P = 5 \times 10^{-5}$ for CeD as primary trait; $P < 5 \times 10^{-5}$ for MS as primary trait). Meta-analyzing across the two diseases increases the overall significance and produces a narrower credible interval (credible interval variants for each panel are in dark grey; the physical span of the credible interval is shaded grey). The credible interval focuses on the intergenic region proximal to *RGS1*. (b) This shared association is also shared with an eQTL for *RGS1* in naïve CD4 T cells (JLIM $P = 0.015$). (c) The lead disease-associated variant lies in a region of accessible chromatin in naïve CD4 T cells and total T cells. This is marked with H3K27ac in total T cells and with H3K4me1 in naïve CD4 T cells, suggesting this is an active, primed enhancer element. (d) The *RGS1* eQTL lead variant predicts the disease association *P* value, further indicating this is a shared effect. (e) Disease and eQTL association effects are negatively correlated, indicating that disease risk is associated with lower *RGS1* expression. (f) *RGS1* is expressed at lower levels in T cells obtained from MS patients compared to healthy controls, confirming this risk effect direction.

Supplementary Figure Legends

Supplementary Fig. 1: Methods overview. (a) Schematic overview of cross-disease fine-mapping. We used joint likelihood mapping (JLIM) to identify susceptibility loci that are likely to share a common underlying causal variant across multiple autoimmune diseases (left panels). At such loci, we performed cross-disease meta-analysis, combining data for co-localized diseases into a fixed-effects model (middle panel, top). We then performed statistical fine-mapping; SNPs contained within the 95% credible interval are shown as filled black circles in this schematic. In general, cross-disease fine-mapping produced smaller credible intervals, in this schematic represented as a single causal variant. We next used JLIM to assess each susceptibility locus for overlap with cis-eQTLs in the BLUEPRINT dataset of naïve CD4 T cells, monocytes and neutrophils. In this schematic, the meta-analysis signal overlaps with an eQTL signal (lower middle panel). At such overlaps, we expect the disease susceptibility signal to decay as a linear function of correlation to the eQTL lead variant. We also expect a linear correlation between corresponding effect sizes for susceptibility and eQTL gene expression (right panels). (b) Overview of the number of subjects assessed and which passed quality control.

Supplementary Fig. 2: ImmunoChip covers a significant fraction of GWAS loci for six autoimmune diseases. (a) Between 38.5% (MS) and 85.7% (CeD) of previously identified, non-MHC GWAS susceptibility loci for CeD, IBD, MS, RA, SLE and T1D fall within high-density ImmunoChip regions. (b) The total number of conditionally independent associations declines as a function of *P*-value threshold, with 495 independent effects identified at $P < 0.0001$ and 202 independent effects identified at $P < 5 \times 10^{-8}$. At all thresholds, the largest number of effects are identified for IBD, the largest dataset. (c) The majority of associated loci exhibit a single genetic effect at all thresholds. At $P < 5 \times 10^{-8}$, up to 13.6% of associated loci in IBD exhibit more than one effect; at the lowest threshold, multiple effects are seen at between 7.3% (RA) and 30.1% (IBD) of associated loci.

Supplementary Fig. 3: Heritability of immune-mediated disorders is enriched in ImmunoChip regions. We used LD score regression to estimate heritabilities (h_g^2) for 17 immune-mediated disorders for which GWAS summary statistics were available. We excluded traits with heritability *Z*-scores < 4 (indicated in pink) from further analysis. (a) Heritability estimates for the remaining 11 traits (observed scale) are highly variable, ranging from 0.068 (All) to 0.47 (SLE). While heritability is sensitive to population and method of estimation, we see that several estimates are smaller than expected, reflecting the influence of genomic control correction used in the original association studies. As this downward bias affects both numerator and denominator equally[20], it does not influence genetic correlation analysis. (b) Using partitioned LD score regression[6] to measure the proportion of heritability that can be attributed to ImmunoChip regions (~2% of the genome), we see broad patterns of enrichment, ranging from 16.8% (Vit) to 46.3% (CeD). AtD, atopic dermatitis; All, allergy; Ast, asthma; CeD, celiac disease; EGPA, eosinophilic granulomatosis with polyangiitis; IBD, inflammatory bowel disease; sIgAD, selective IgA deficiency; LADA, latent autoimmune diabetes in adults; MS, multiple sclerosis; PBC, primary biliary cirrhosis; Ps, psoriasis; PsA, psoriatic arthritis; PSC, primary sclerosing cholangitis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; SSc, systemic sclerosis; Vit, vitiligo.

Supplementary Fig. 4: Fixed-effects and random-effects meta-analysis produced similar results. We conducted both fixed-effects and random-effects meta-analysis. We observe differences in effect size estimates for only a handful of variants, all are from one locus and all for SLE (highlighted in red). The effects are in the same direction, but magnitude is generally smaller in the fixed-effect analysis.

Supplementary Fig. 5: Backward selection with GCTA recapitulates at least 90% of results obtained with JLIM. We reanalyzed all shared traits identified through forward conditional logistic regression with a reverse selection model implemented in GCTA. Most effects in our analysis were in tight linkage disequilibrium ($r^2 \geq 0.8$) with a GCTA lead variant (120/134, 89.6%). Of these, 117 GCTA associations were identical.

Supplementary Fig. 6: Previously unreported ImmunoChip associations. Using conditional logistic regression to allow for multiple associated variants at a single locus, we identified three genome-wide significant associations (two for CeD, one for IBD) that were not reported in the respective publications. (a) Unconditional testing at the *ANKS1A* locus on chromosome 6 produced evidence of association that did not reach genome-wide significance in our analysis. Conditional testing produced genome-wide significance for a single variant within *ANKS1A* (rs12206298; $P = 4.9 \times 10^{-8}$), and suggestive evidence for a second effect (rs4713844; $P = 9.9 \times 10^{-8}$) in the region. Published summary statistics (lines, top panel) were also significant. The region may have been excluded from the initial publication as it is within the extended MHC. (b) Unconditional testing at the *ADAMTS7—MORF4L1—CTSH* locus on chromosome 15 in CeD did not reach genome-wide significance. Conditional testing revealed genome-wide significance for a variant in *CTSH* (rs3784539; $P = 1.1 \times 10^{-8}$) and suggestive evidence for a second effect in *MORF4L1* (rs7181033, $P = 8.7 \times 10^{-5}$). (c) Unconditional testing at the *CLEC16A* locus on chromosome 16 in IBD did not reach genome-wide significance. Conditional testing produced evidence

for a single effect in *CLEC16A* (rs7201325, $P = 1.1 \times 10^{-10}$) and modest evidence for a second effect near *RP11-396B14.2* (rs55773334, $P = 7.6 \times 10^{-5}$).

Supplementary Fig. 7: Distinct conditional associations are shared with distinct sets of diseases. At three loci, multiple conditionally independent associations for a single disease are shared. (a) At the *STAT4* locus, two independent effects are each shared between RA and SLE. (b) At the *CD28-CTLA4* locus, the second independent association in T1D near *CD28* is shared with CeD and the third independent association within *CTLA4* is shared with RA. (c) At the *TYK2* locus, the first independent association of RA is shared with SLE, IBD and T1D, while the second association in RA (near *ICAM3*) is shared with SLE.

Supplementary Fig. 8: An association signal near *TNFRSF9*, *PARK7* and *ERRFI1* exhibits opposing effects in CeD and IBD. (a) Meta-analysis of CeD and IBD association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for CeD against corresponding *Z* scores for IBD reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for CeD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in CeD and IBD are opposed at this locus.

Supplementary Fig. 9: An association signal near *PLEK* exhibits opposing effects in CeD and MS. (a) Meta-analysis of CeD and MS association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for CeD against corresponding *Z* scores for MS reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for CeD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in CeD and MS are opposed at this locus.

Supplementary Fig. 10: An association signal near *IL12A* exhibits opposing effects in MS and SLE. (a) Meta-analysis of MS and SLE association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for MS against corresponding *Z* scores for SLE reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for MS and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in MS and SLE are opposed at this locus.

Supplementary Fig. 11: An association signal near *ERAP2* exhibits opposing effects in IBD and MS. (a) Meta-analysis of IBD and MS association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for IBD against corresponding *Z* scores for MS reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for IBD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in IBD and MS are opposed at this locus.

Supplementary Fig. 12: An association signal near *TNFSF8* exhibits opposing effects in IBD and MS. (a) Meta-analysis of IBD and MS association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for IBD against corresponding *Z* scores for MS reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for IBD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in IBD and MS are opposed at this locus.

Supplementary Fig. 13: An association signal near *ZNF365* exhibits opposing effects in IBD and MS. (a) Meta-analysis of IBD and MS association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for IBD against corresponding *Z* scores for MS reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for IBD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in IBD and MS are opposed at this locus.

Supplementary Fig. 14: An association signal near *LTBR* exhibits opposing effects in MS and T1D. (a) Meta-analysis of MS and T1D association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for MS against corresponding *Z* scores for T1D reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for MS and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in MS and T1D are opposed at this locus.

Supplementary Fig. 15: An association signal near *CLEC2D*, *CLECL1* and *CD69* exhibits opposing effects in IBD and MS. (a) Meta-analysis of IBD and MS association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for IBD against corresponding *Z* scores for MS reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for IBD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in IBD and MS are opposed at this locus.

Supplementary Fig. 16: An association signal near *STAT3* exhibits opposing effects in IBD and MS. (a) Meta-analysis of IBD and MS association data (third panel) reduces significance (points) and increases heterogeneity (lines, Cochran's *Q*) of the association signal. (b) Regression of *Z* scores for IBD against corresponding *Z* scores for MS reveals an inverse linear relationship, suggesting opposing directions of effect in the two diseases. After reversing effects for IBD and repeating meta-analysis (a, fourth panel), significance increases and heterogeneity decreases, confirming that the effects in IBD and MS are opposed at this locus.

Supplementary Table Legends

Supplementary Table 1: Posterior probabilities of colocalization from coloc2.

Supplementary Table 2: GWAS and ImmunoChip studies used in the analysis. Sources for datasets used in this study. Databases and accession numbers are provided for publicly available datasets. For datasets not available in public datasets, contact details for relevant consortia are provided. NHGRI-EBI, NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/); dbGAP, Database of Genotypes and Phenotypes (https://www.ncbi.nlm.nih.gov/gap/); GRASP, Genome-Wide Repository of Associations Between SNPs and Phenotypes (https://grasp.nhlbi.nih.gov/Overview.aspx); IBDGC, NIDDK Inflammatory Bowel Disease Genetics Consortium (https://ibdgc.uchicago.edu/); IMSGC, International MS Genetics Consortium (https://imsgc.net); RACI, Rheumatoid Arthritis Consortium International)

Supplementary Table 3: Subjects by disease, stratum and phenotype class. Case and control counts are provided for each disease and stratum, before and after quality control filtering.

Supplementary Table 4: Subject quality control. Inclusion status for each subject. For each subject failing quality control, the failed step is indicated.

Supplementary Table 5: SNP quality control. Inclusion status for each SNP. For each SNP failing quality control, the failed step is indicated.

Supplementary Table 6: eQTLs identified at disease level and after cross-disease meta-analysis. Shared disease susceptibility—eQTL associations are indicated. Conditioning SNPs for the eQTL data are indicated where appropriate. Where an eQTL is identified at the disease level and also in cross disease meta-analysis, this is indicated as “Stable eQTL.” eQTLs that are newly identified after cross disease meta-analysis are labelled “New eQTL.”

References

1. Rosenblum, M. D., Remedios, K. A. & Abbas, A. K. Mechanisms of human autoimmunity. *J. Clin. Invest.* **125**, 2228–2233 (2015).
2. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).
3. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
4. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
5. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
6. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
7. Hemminki, K., Li, X., Sundquist, K. & Sundquist, J. Shared familial aggregation of susceptibility to autoimmune diseases. *Arthritis Rheum.* **60**, 2845–2847 (2009).
8. Eaton, W. W., Rose, N. R., Kalaydjian, A., Pedersen, M. G. & Mortensen, P. B. Epidemiology of autoimmune diseases in Denmark. *J. Autoimmun.* **29**, 1–9 (2007).
9. Kuo, C.-F. *et al.* Familial Aggregation of Systemic Lupus Erythematosus and Coaggregation of Autoimmune Diseases in Affected Families. *JAMA Intern. Med.* **175**, 1518–1526 (2015).
10. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).
11. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).
12. Sirota, M., Schaub, M. A., Batzoglou, S., Robinson, W. H. & Butte, A. J. Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet.* **5**, e1000792 (2009).
13. Pouget, J. G. *et al.* Cross-disorder analysis of schizophrenia and 19 immune-mediated diseases identifies shared genetic risk. *Hum. Mol. Genet.* **28**, 3498–3513 (2019).
14. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
15. Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* **14**, 661–673 (2013).
16. Burren, O. S. *et al.* Genetic feature engineering enables characterisation of shared risk factors in immune-mediated diseases. *Genome Med.* **12**, 106 (2020).
17. Bunt, M. van de *et al.* Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLOS Genet.* **11**, e1005535 (2015).
18. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
19. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
20. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
21. Zhernakova, A., Withoff, S. & Wijmenga, C. Clinical implications of shared genetics and pathogenesis in autoimmune diseases. *Nat. Rev. Endocrinol.* **9**, 646–659 (2013).
22. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).
23. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
24. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
25. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinforma. Oxf. Engl.* **32**, 1493–1501 (2016).

26. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
27. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
28. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
29. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
30. Pappalardo, J. L. *et al.* Transcriptomic and clonal characterization of T cells in the human central nervous system. *Sci. Immunol.* **5**, (2020).
31. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).
32. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet. TIG* **37**, 109–124 (2021).
33. The Lenercept Multiple Sclerosis Study Group & The University of British Columbia MS/MRI Analysis Group. TNF neutralization in MS: Results of a randomized, placebo-controlled multicenter study. *Neurology* **53**, 457–457 (1999).
34. Gregory, A. P. *et al.* TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature* **488**, 508–511 (2012).
35. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
36. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–1456 (2015).
37. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
38. Han, Y. *et al.* Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat. Commun.* **11**, 1776–13 (2020).
39. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
40. Lyons, P. A. *et al.* Genome-wide association study of eosinophilic granulomatosis with polyangiitis reveals genomic loci stratified by ANCA status. *Nat. Commun.* **10**, 5120 (2019).
41. Bronson, P. G. *et al.* Common variants at PVT1 , ATG13 – AMBRA1 , AHI1 and CLEC16A are associated with selective IgA deficiency. *Nat. Genet.* **48**, 1425–1429 (2016).
42. Cousminer, D. L. *et al.* First Genome-Wide Association Study of Latent Autoimmune Diabetes in Adults Reveals Novel Insights Linking Immune and Metabolic Diabetes. *Diabetes Care* **41**, 2396–2403 (2018).
43. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019–11 (2015).
44. Ji, S.-G. *et al.* Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat. Genet.* **49**, 269–273 (2017).
45. Aterido, A. *et al.* Genetic variation at the glycosaminoglycan metabolism pathway contributes to the risk of psoriatic arthritis but not psoriasis. *Ann. Rheum. Dis.* **78**, 355–364 (2019).
46. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
47. López-Isac, E. *et al.* GWAS for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. *Nat. Commun.* **10**, 4955–14 (2019).
48. Jin, Y. *et al.* Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat. Genet.* **48**, 1418–1424 (2016).

49. Feng, B.-J. *et al.* Multiple Loci within the Major Histocompatibility Complex Confer Risk of Psoriasis. *PLOS Genet.* **5**, e1000606 (2009).
50. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
51. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
52. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
53. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
54. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
55. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
56. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).
57. Langefeld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
58. Zhao, J. *et al.* A missense variant in NCF1 is associated with susceptibility to multiple autoimmune diseases. *Nat. Genet.* **49**, 433–437 (2017).
59. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
60. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
61. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289–317 (2016).
62. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PloS One* **9**, e93766 (2014).
63. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
64. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
65. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
66. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).
67. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
68. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2009).
69. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
70. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
71. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
73. Rasmussen, A. *et al.* The lupus family registry and repository. *Rheumatol. Oxf. Engl.* **50**, 47–59 (2011).