

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: OCTOBER 2023

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

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE October 2023		2. REPORT TYPE Annual		3. DATES COVERED 30SEPT2022 - 29SEPT2023	
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 UU	18. NUMBER OF PAGES 60	19a. NAME OF RESPONSIBLE PERSON USAMRDC
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code)

TABLE OF CONTENTS

	<u>Page</u>
1. Introduction	4
2. Keywords	4
3. Accomplishments	4
4. Impact	15
5. Changes/Problems	16
6. Products	17
7. Participants & Other Collaborating Organizations	18
8. Special Reporting Requirements	22
9. Appendices	23

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 self-reported 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 >120 known SLE-risk genetic polymorphisms; this list is anticipated to grow with ongoing work in northern and southern Asian continent populations. 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 36 months of the project have shown significant progress in all three specific Aims. We believe we have remained on schedule relative to almost all 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. Dr. Langefeld and Ainsworth have also been aggressive in sharing results with the ClinGen effort for rheumatic diseases. ClinGen is funded by the NIH to build a central resource describing the clinical significance of genetic polymorphisms for precision medicine and research. We received no compensation.

Major accomplishments of the past year built upon the those of the previous years and leveraged new methods we are developing. We completed the identification and compilation of SLE risk SNPs datasets, expanded beyond original list of associations by incorporating new publications, 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, complementing the ImmunoChip associations in African Americans; manuscript to be submitted over coming months. 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 are converting the pipeline to a Python-based pipeline to use high performance computing. This greatly reduces computing time from days to hours for each set of binding experiments to a given target. 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 prioritized 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 are now intensely working on associations mapped to transcription factor binding sites, leveraging DNA topology, novel methods in development by Drs. Ainsworth and Langefeld.

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 24-36).

Major Accomplishment/Progress Category	Descriptor
Manuscripts and Presentations	• Ongoing. Six accepted/published manuscripts duration of grant. • Two additional manuscripts in preparation • Three conference posters/presentations at the Lupus 21st Century 2023 Conference (2 talks; 2 posters)
Data Processing and Analytic Pipelines	• Drug target identification and processing for molecular docking; conversion to Python-based pipeline for high performance computing cluster for significantly faster docking experiments. • Extracted principal components and variance explained (i.e., eigenvalues of the molecular fingerprints based on the SMILES identifiers of FDA approved drugs (to facilitate drug structural similarity analyses). • Contributed the Python code to enable the principal component and eigenvalue extraction to GitHub for GlobalChem. • Utilized alternative dimensionality reduction methods: uniform manifold approximation and projection (UMAP). • Expanded summarization of molecular docking results, summaries include generation of $\Delta\Delta G$ for paired docking comparisons across cGenes, eGenes, and tGenes, cluster analysis of molecular similarity, and graphical illustrations.
Created Datasets	• Continued the expansion of SLE risk SNP datasets linked to their most plausible-implicated gene targets (e.g., gene expression, eGenes; transcription factor binding site, tGenes; and protein coding variants cGenes). • Finalized African American systemic lupus erythematosus genome-wide association stud, final circulation to co-authors for approval to submit. • Building database of experimental structures and their properties (e.g., resolution) from the Protein Data Bank (PDB). • Expanded database of Euclidean similarity distance to include both UMAP and principal components for the 1,430 FDA approved drugs and linking these distances to the <i>in silico</i> binding results for completed targets.
System biology analyses	• Continued pathways and system biology analysis for SLE-associated SNPs across the four ancestries/ethnicities for druggable pathway identification (African ancestry, Asian ancestry, European ancestry, and non-African Hispanic ethnicity). This was the focus of published papers. • Pathway analysis based on results of Mendelian randomization studies with atherosclerosis; manuscript revised, resubmitted and published.

Table 2: Project completion and status (Months 24-36) 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 24-36) 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 100%
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 95% 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 100% 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 can 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 100%
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 24-36) 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 100%
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).

quality protein structures in Protein Data Bank).		- Current effort aligns SLE-risk allele associated with increased gene expression and that gene's gene expression in one of five tissues/cell types associated with risk of lupus or subtypes. - Database of PDB structures and characteristics for prioritizing binding experiments. Completion progress ~90%
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 ~90%
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-intensive tasks. Completion progress ~95%
Specific Aim 3: Identify and prioritize drugs		Reporting Period (Months 24-36) Updates
Subtask 1: Bioinformatic analysis for gene-drug and protein-drug interaction using STITCH, DrugPath, CLUE, etc.	6-36	For current list of genes (Aim 1, 2), completed and summarized across multiple manuscripts and presentations at national meetings. Completion progress 100%
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 tGenes, eGenes and combining with cGenes. Now including UMAP for drug clustering. Completion Status: ~90%
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 as it was not effective for high-throughput screening. - Modifying and focus on toxicity reports from CoLTS, no longer pursued per above. Completion Status: 100%
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 . - Current focus on eGenes, tGenes. Completion status: ~75%

What was accomplished under these goals?

Significant Results:

Here we summarize key areas of progress. We are at a phase in the project where most of the foundational components of the grant are complete and we are leveraging them to complete Aim 3 tasks. We start by briefly summarizing manuscripts submitted or published. The published manuscripts are listed in the Appendix. We continue by summarizing further progress roughly as proposed in the grant. As noted in previous reports, 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 include highlighting pathways and drug targets motivated by the genetic associations. We do not repeat descriptions of the ongoing work from the first 24 months but focus on additional, new results from the reporting period (months 25-36). Previous years' progress updates are in earlier reports, but details are also available upon request.

Published work this year.

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 method 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 drug repositioning pipeline; this work is underway. If so, lupus can be a leader that informs other autoimmune disease research and drug repositioning opportunities. We had anticipated this would be published in Nature Genetics last year, *Genetic mapping across autoimmune diseases reveals shared associations and mechanisms*. However, additional reviewer comments required more analyses. The manuscript is now tentatively accepted pending editorial recommendations and is in MedRxiv (see appendix).

Molecular pathway analyses suggest mechanistic differences in systemic lupus erythematosus in patients of Asian and European ancestry.

Individuals of Asian ancestry (AsA) tend to experience more severe lupus symptoms, including renal involvement and tissue damage, than individuals of European ancestry (EA). They also exhibit shared and distinct genetic risk loci. We used the public gene expression data at the GTEx version 8 (GTExportal.org) to link risk loci SNPs to genes (eGenes), the Human Active Enhancers to interpret Regulatory variants (HACER) to link SNPs to transcription factor binding sites, and the Ensembl genome browser to protein coding SNPs (cGene). The resulting “tagged” genes were partitioned into EA, AsA, or both, depending on what ancestries exhibited the evidence for association, and analyzed using bioinformatic tools. Genes lists were examined using Big-C, proprietary database at RILITE, for function. Pathway and network analyses was completed using STRINGdb, Cytoscape, MCODE. These tools identified different biological pathways within the two ancestral groups (Figure 1 from Owen et al., 2023; Sci Reports 13:5339). These results motivate novel hypotheses for animal and laboratory work, but also have translational and drug repositioning implications. As examples noted in the publication, mTOR pathway modulators such as N-acetylcysteine and rapamycin might be targets for investigation of whether they can reduce disease activity. In lupus-prone mice, pioglitazone, a peroxisome proliferator-activated receptor (PPARγ) eased nephritis symptoms inhibitor. It is important to understand that individuals from across many ancestries have these SLE-risk alleles, but we were able to detect the association with SLE in the specific ancestry. Thus, depending on other genetic and environmental backgrounds, individuals with the risk allele of any ancestry likely exhibit increased risk to SLE. Thus, these pathways might not be abnormally active in every SLE patient within the identifying ancestry (e.g., AsA) and might be comparably abnormally active in SLE patients of other ancestries. Hence, the discovery in one ancestry has generalizability to other ancestries and offers potential

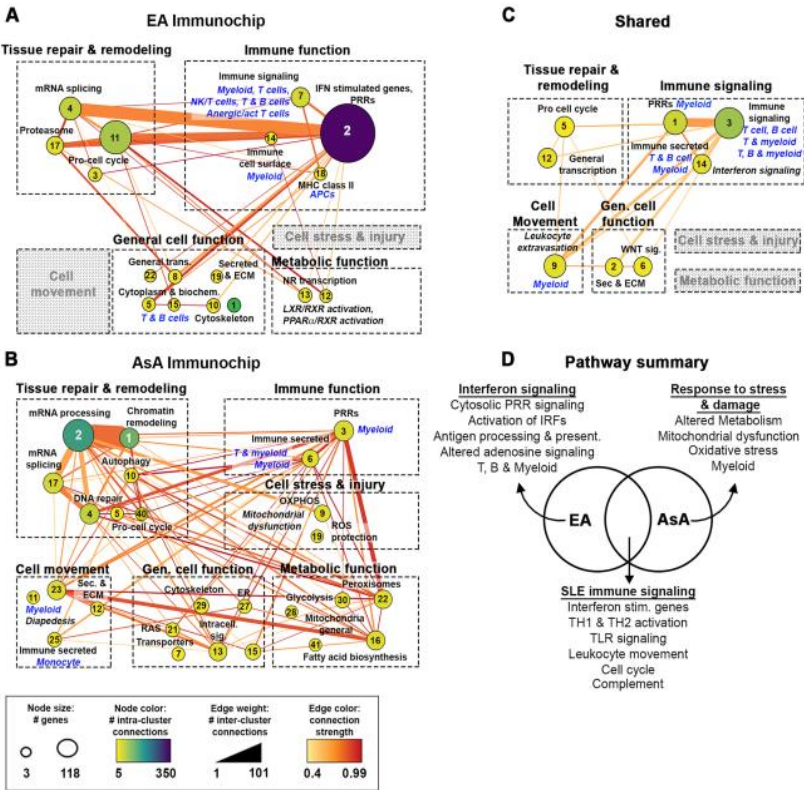


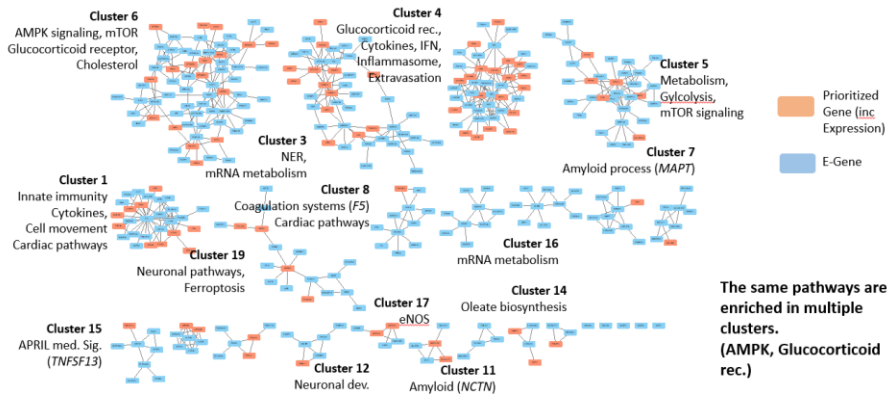
Figure 3. Key pathways determined by EA and AsA-associated genes. Cluster metastructures for EA (A), AsA (B) and the shared gene cohort (C) were generated based on PPI networks, clustered using MCODE and visualized in Cytoscape. Cluster size indicates the number of genes per cluster, edge weight indicates the number of inter-cluster connections and color indicates the number of intra-cluster connections. Enrichment for each cluster was determined by BIG-C and IPA; clusters were then grouped and categorized according to overall function (immune, tissue repair, metabolic, motility or general). Grey boxes indicate categories lacking relevant clusters. (D) Venn diagram showing the number of overlapping pathways motivated by EA or AsA predicted genes. Representative pathways are listed.

Figure 1. Molecular Pathways analysis from published manuscript. Via Owen et al., 2023; Sci Reports 13:5339

precision medicine approaches via biomarkers (e.g., SNP panels) which may be more informative of SLE risk (e.g., compared to self-reported race or ancestry)

Ongoing work this year.

The past year has focused on completing and leveraging the results in Aims 1 and 2 to provide insight into the key biological pathways and prioritizing targets (i.e., genes). We explored 746 eGenes meeting criteria of statistically significant increased expression, as defined by false discovery rate (FDR) thresholds. A key set of pan-ancestry results are displayed in Figure 2, which provides a high-level view of the collection of networks identified by our bioinformatic analyses (e.g., STRINGdb, Cytoscape, MCODE) similar to that described in the above paper.



AMPK signaling (10 genes)

AK5, AK9, BRD7, PPP2CA, RPS6KB1, RPTOR, SIRT1, SMARCA4, STRADA, CHRNE

Glucocorticoid receptor signaling (25 genes)

ICAM1, IL18R1, IL6, PLAUG, GTF2H1, GTF2H5, SUMO1, TAF11, CD247, CXCR1, CXCR2, GRB2, IL12RB1, IL12RB2, IL18RAP, IL6R, RAC1, RAF1, TYK2, BRD7, H3-3A/H3-3B, POLR2C, RARA, SMARCA4

mTOR signaling (6 genes)

RPS11, RPS20, PLD2, PPP2CA, RPS6KB1, RPTOR

Cardiac hypertrophy (19)

ADCY7, ADSS2, RPS6KB1, GRB2, IL6R, RAC1, RAF1, HSPB2, IL18R1, IL37, IL6, TNFSF12, NFSF12, CACNA2D3, CHP1, MAP3K11, MYL9, OPN1SW, PLCL1

Neurological

ADCY, ARPC2, ARRB2, BLK, CACNA2D3, CASP8, CD38, CD40, CFLAR, CHP1, FARP2, GBP, AR1, GRB2, ICAM1, IL37, IL6, IL6R, MAPT, NCF2, NCSTN, NOTCH2, OPN1SW, P2RX4, PLAU, PLCL1, PLD2, PPP2CA, RND2

The respective clusters are driven by SLE risk loci linked to genes and these genes differential expression related to SLE status (i.e., case/control); upregulated in cases. Many of these pathways are either supported by the literature or logical given the clinical manifestations of SLE. The products of these genes, regardless of whether cGenes, eGenes, or tGenes, are now input for the high throughput FDA-approved drug docking pipeline.

Development and implementation of drug-repositioning pipeline.

We have continued the development and implementation of the eGENES analysis for drug repositioning and expanded it to include the tGenes. Recall that we are also focusing on SLE risk alleles 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. While much of this pipeline was built upon the cGENES pipeline (e.g., protein model assessment, *in silico* binding), there are notable differences that additional required *de novo* programming implementation. First, identification of

Figure 2. High-level view of the collection of networks identified by our bioinformatic analyses (e.g., STRINGdb, Cytoscape, MCODE)

cGENES was restricted to SLE-associated variants located *within coding regions* of the genome. Thus, only a limited number of SNPs met these criteria, resulting in 34 cGENES. However, both eGENES and tGenes are unrestricted by physical location. Thus, *any* SLE-associated SNP has the potential to map to an eGENE and/or tGene. Considering FDR-associated SLE-risk 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). Second, generally coding SNPs map to only a single cGENE. In contrast, it is common for a SNP to map to multiple eGenes; a SNP is an eQTL for multiple genes. Recall from last year's report that in the GTEx (V8) dataset, there are 4,632,457 unique eQTLs (SNPs) mapping to 13,791,909 million unique eQTL-eGENE combinations. The eQTLs map to anywhere between one and nine eGenes. Thus, a high-throughput processing pipeline is necessary that can accommodate cGenes, eGenes and tGenes.


```
python3 run_autodock_vina_parallel.py --help
usage: run_autodock_vina_parallel.py [-h] [--ligand_dir LIGAND_DIR]
                                     [--ligand_list LIGAND_LIST]
                                     [--target_pdb_ID TARGET_PDB_ID]
                                     [--output_dir_prefix OUTPUT_DIR_PREFIX]

This python script reads in multiple ligands as listed in your specified
LIGAND_FILE and then completes docking for each of these ligands on the
specified TARGET. It is assumed that your target (.pdbqt) and grid and config
files will be in a subdir with the targets same name. Will output a log file
for each ligand as well as a pdbqt file to specified output dir.

optional arguments:
  -h, --help            show this help message and exit
  --ligand_dir LIGAND_DIR
                        Directory name where the ligand.pdbqt files are
                        stored. Example: LIGAND_FILES (note: Default is
                        current dir.)
  --ligand_list LIGAND_LIST
                        Full path and filename of the ligand list file (one
                        ligand file name per line).
  --target_pdb_ID TARGET_PDB_ID
                        PDB name of of docking target (e.g., 6jp5apo)
  --output_dir_prefix OUTPUT_DIR_PREFIX
                        Directory name for output files. Default is
                        VINA_OUTPUT/ note: will append target PDB ID to end of
                        dir name
  --num_processes NUM_PROCESSES
                        Number of processes for parallelization (e.g., number
                        nodes * number tasks per node * CPUs per task)
```

Figure 5. Drug docking with Autodock Vina is now automated for usage on our HPC system (SLURM job manager) with new command line python wrapper (user help shown).

PDB (<https://www.rcsb.org/>). The cGENEs targets list was of a size suitable for manual curation via the RCSB PDB website; however, as we have expanded to several 100s of potential targets (eGENES), a solely manual curation was not feasible for efficient processing. Thus, a portion of our efforts this past year has focused on transitioning parts of the pdb curation towards automated components, where feasible. Specifically, we have instituted a python script that leverages HTML parsing to identify pdb IDs, requiring only a gene name for the search query. This process enabled efficient identification of whether an experimental structure existed for the target of interest. Of course, not all PDB structures will be usable for docking purposes and the next step required better automated processing of potential structures (e.g., ensuring source data = *homo sapiens*; ensuring structure was not a fusion/artificially created protein; structure resolution was of sufficient quality.) Once we have the gene target and the pdb ID,

we can now automatically download the structure's corresponding 'no-atom XML file' from the RCSB PDB. This provides a file with the majority of the information required for filtering structures. Importantly, these files can be parsed using python, which has greatly streamlined our process (Figure 4).

The second stage in the drug docking pipeline involves file preparation for AutoDock Vina. There are several steps to this stage, which were previously limited to manual execution. One of the first steps is identifying which chain(s) of the PDB structure relate to our specific target (e.g., excluding chains of ligands bound to the target within the PDB entry. While previously, this information was obtained via the website, this information can now also be queried from the aforementioned downloaded XML files. Having the relevant chains identified then proceeds to the next steps of file preparation, including: extracting the relevant chains from the pdb file, removing water molecules, adding polar hydrogens, adding Kollman charges, ultimately creating the final .pdbqt file for AutoDock Vina. These steps were previously completed manually, (e.g., GUI interface via structure prep software). However, we have been testing conversion of these steps towards command-line execution for better efficiency.

Table 1: Runtime Performance of AutoDock Vina with varied HPC resources				
Number of Nodes	Tasks per node	CPUs per task	Total Processes	Time (hours)
1	1	1	1	22.4
1	8	2	16	3.02
1	16	2	32	1.57
2	16	2	64	1.54

The third stage in the drug docking pipeline is the initiation and running of AutoDock Vina. Our initial pipeline utilized Perl scripts that initiated AutoDock Vina on a local machine. One of the biggest advantages to AutoDock Vina is its ability to leverage parallelization. So, although this enabled analyzing up to four targets on a local machine (unfathomable with other docking algorithms); this still required considerable time, per target. For instance, docking 2,114 molecules (representing ~1400 unique drugs; multiple conformers, per drug) to a single target took approximately 48 hours. Running in parallel with up to four targets required a week of run time. Thus, we transitioned these analyses to our High-Performance Computing (HPC) system which offers vastly more modern hardware, memory, and CPUs. Transitioning to our HPC system with no additional optimization reduced the run time by over half (i.e., 48 hours to 22.4 hours) per target. However, to better leverage the multi-threading capacity via the HPC, we recoded the Perl scripts to a Python wrapper that enabled designation of the number of processes to use for each target's analyses (e.g., number of resources to dock 2,114 compounds). An added advantage to the python wrapper, is the streamlined submission process, where a single line can initiate AutoDock Vina with the relevant options and file input/outputs (Figure 5). This wrapper seamlessly integrates with our HPC's SLURM job manager. Runtime performances by

resource designation are shown in Table 1; runtime benefits were capped at 32 processes as we observed minimal improvement with higher resource designation. Overall, this represents a 95% improvement in runtime.

Clustering by Molecular Similarity. This past year, we also worked on further uncovering the structural similarity among the list of FDA-approved small molecules. This work will greatly aid in understanding underlying relationships of drug docking to targets (e.g., protein products of cGenes and eGenes). This work offers the potential to uncover sub-groupings of drugs (defined by structure) that may collectively share similar binding patterns and thus similar patterns of efficacy. This is beneficial towards both current drug-repositioning (e.g., identifying different drugs with similar effects) and future drug-discovery (e.g., optimizing structures for new products) efforts. Previously, we leveraged principal component analyses (PCA) for dimensionality reduction of the molecular structures captured by molecular fingerprinting based on SMILES (simplified molecular input line-entry system) IDs, as this was the method implemented within the GlobalChem-Chemical graph network methods (<https://globalchem.gitbook.io/globalchem-your-chemical-graph-network/>). In the past year, we added custom code to GlobalChem python modules and determined that each principal component (PC) was only explaining a small portion of the overall variance of the FDA-approved small molecules dataset (e.g., first three PCs only explaining 13.23% of the overall variance). Given our focus on the relative relationships among these drugs, we explored an alternative dimensionality reduction method: Uniform Manifold Approximation and Projection (UMAP). UMAP has shown greater utility in capturing and preserving non-linear relationships among high-dimensional datasets [Becht *et al.* Nat. Biotech. 2019]. We explored using UMAP with distance metrics designed for binary, sparse data (e.g., as we have with molecular fingerprinting data), including Jaccard distance (**Figure 6**). We identified clearer patterns of separation among groupings of drugs and are in the process of optimizing and naming clustering assignment (e.g., Figure X shows K-Means clustering, the most commonly used cluster method, but we previously found and reported our success with Random Forests clustering algorithms).

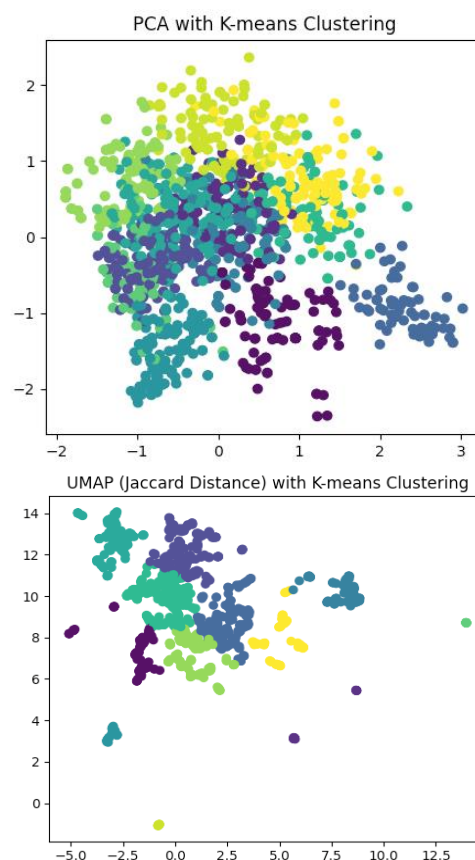


Figure 6. Comparison of Dimensionality reduction of FDA-small molecules data using PCA and UMAP. UMAP offers clearer separation of clusters of molecular similarity, compared to PCA, this will aid in the identification of sub-classes of drug activity (combined with in silico docking).

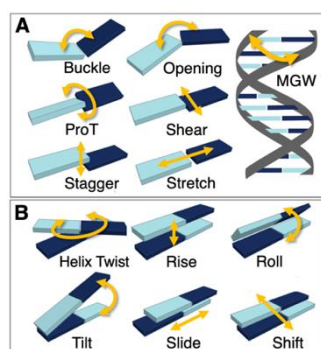


Figure 7. DNA shape features. A) Intra-features: between paired bases. B) Inter-features, between adjacent base pairs.

DNA topology. To prioritize functional SNPs in the association and eQTL analyses, we are incorporating sequence-dependent DNA topology (or DNA shape) into fine-mapping analyses. DNA shape refers to the molecular geometries created by a specific sequence of nucleotides (Olson *et al.*, 2001, *J. Mol. Biol.* **313**, 229–237 (2001)). These topological features (e.g., minor groove width -**Figure 7**) have been well-studied, and their function in specific contexts are well-documented (Lazarovici *et al.*, 2013, *Proc. Natl. Acad. Sci.* **110**, 6376–6381; Morgunova 2015, *at. Commun.* **6**, 10050; Duan *et al.*, 2018, *Genome Biol.* **19**, 1–12). Recent advancements in prediction algorithms now enable high-throughput generation of DNA shape on a genome-wide scale (Li *et al.*, 2017, *Nucleic Acids Res.* **45**, 12877–12887). As an intrinsic property of DNA, DNA shape can **readily annotate any region of the genome**, overcoming database biases (e.g., incomplete data), common among other annotation sources, which have contributed to health disparities in omics research (Marari *et al.*, 2016, *N. Engl. J.*

Med. **375**, 655–665; Long *et al.*, 2022, *Nat. Commun.* **13**, 2907). [ref] We have been investigating the impact of genotyped SNPs on local DNA topology for SLE risk loci, potentially identifying the SNP's potential impact on function. This method is especially beneficial within genomic regions of high LD where SNP associations (e.g.,

phenotypic outcomes, eQTLs) are largely indistinguishable from one another. Previously, we leveraged the magnitude of topological change (imposed by a SNP) to fine-map plausibly causal SLE-associations in select regions (Ainsworth et al., 2020, *Nucleic Acids Res.* **48**, 11304–11321). Here, we will summarize each SNP's impact on DNA shape as a measure of magnitude (i.e., Euclidean Distance -**Figure 8**), hypothesizing, that SNPs which 1) cause large disruptions in local topology and 2) are associated with a SLE (or related phenotype), or functional (e.g., eQTL) outcome are more likely to be causal. This could help in prioritizing genes. We are also linking changes in DNA topology to specific transcription factor binding sites, to inform additional tGenes. A SNP's location within a TFBS does not necessarily disrupt binding, and thus it is important to accurately predict allele-specific effects Kumar et al., 2017, *Nucleic Acids Res.* **45**, D139–D144. TFs are highly sequence-dependent, and binding sites can be characterized as “motifs”, summarized as position weight matrices (PWM). While PWMs can be used to predict SNP effects on TF binding, predictive performance can vary depending upon the transcription factor Yan, et al., 2021, *Nature* **591**, 147–151). There is an established role of intrinsic DNA topology in transcription factor binding. For genotyped SNPs, we first identify whether they fall within a likely TFBS. We completed this using two publicly available datasets: JASPAR2022 predicted TF motifs, a dataset of over 678 million predicted TFBS across the genome (Castro-Mondragon 2021, *Nucleic Acids Res.* 2022 Jan 7;50(D1):D165-D173) and ReMap2022 a curated database of over 10,000 ChIP-seq datasets (Hammal et al., 2022, *Nucleic Acids Res.* 7;50(D1):D316-D325). We are then assessing the differences in shape profiles (e.g., between two alleles at a SNP) to determine potential binding disruption. This novel idea is within the conceptual agenda of the aims of the grant, but will not be fully completable within the grant. In fact, we are very pleased to announce that Dr. Hannah Ainsworth, new assistant professor, leveraged the work we did as part of this Department of Defense grant to successfully apply for a grant, *Intrinsic DNA Topology for Causal Variant Identification*. This is only one of two such grants annually provided by the Daryl and Marguerite Errett Foundation.

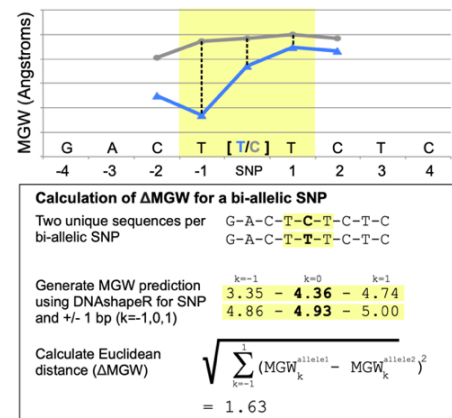


Figure 8. Calculation of MGW disruption. Approach can be used for all 13 features.

ClinGen. Drs. Langefeld and Ainsworth were invited to participate in the ClinGen initiative. The ClinGen initiative is a large, international collaborative effort to curate genetic associations and genes toward informing on their clinical relevance of genes and variants for precision medicine and research (clinicalgenome.org). Our primary focus is to assist with the rheumatic diseases and systemic lupus erythematosus (and other lupus phenotypes). To that end, based on the work funded by this grant, we have presented to the ClinGen investigators our curated list of SLE-risk SNPs and their assignment to genes across five ancestral groups: several hundred associations representing years of our work. Historically, ClinGen was focused more on monogenic disorders, and our work on lupus is requiring significant evolution in thinking to complex, multigenic disorders. We note that we receive no funding from ClinGen and have provided our lists even before publishing them.

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 three significant additional components: 1) AlphaFold, the newly released AI-predicted protein structure that will significantly increase the number of targets explored, 2) clustering to identify hot spots of binding drugs that allows potential expansion of the list of drugs for evaluation, and 3) DNA topology as an additional metric toward function.

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

Over the three years of the grant, several students, staff, and junior faculty have benefited from participating in the science and mentoring.

Mr. Nolan H. Hamilton completed 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 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 Ekaterina S. Khvatkova to replace Mr. Hamilton. 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. The past several months she has been working on the DNA topology component under the guidance and mentoring of Dr. Ainsworth.

We also hired Sawyer Welden, a Python programmer for 20% effort to assist in the automation of the pipelines for the high-performance computing cluster. However, unexpectedly and without forewarning, he was released from employment by our department chair, with the approval of the institution's Human Resources Department. Thus, we are seeking to replace him with another Python programmer who has some chemistry background. Despite the unclear circumstances Mr. Welden received opportunities to expand his programming skills and participate in our project.

Dr. Langefeld accepted a new PhD student, Olamide Arege, an individual with virology background from Nigeria who is seeking to do a PhD in bioinformatics. Mr. Arege participated in planning and implementation of several aspects of the project. Funding for Olamide is from independent sources, including Dr. Langefeld's R&D account. The project, among others, were helpful for Mr. Arege to determine that bioinformatics is not his primary research interest and he has moved to a virology lab and a different advisor.

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. Mr. Haoqing Wu completed a summer internship focused with us focused on the docking experiments. He presented his work at our annual student science poster session. His work was titled *Optimization of FDA-approved Drugs for Lupus Targets*.

As previously noted, we are very pleased that Dr. Hannah Ainsworth, new assistant professor, leveraged the work we did as part of this Department of Defense grant to successfully apply for a grant, *Intrinsic DNA Topology for Causal Variant Identification*. Only two such grants funded by the Daryl and Marguerite Errett Foundation are awarded each year.

How were the results disseminated to communities of interest?

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

As described above, we have presented our curated gene lists to the ClinGen Rheumatic Disease working group.

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

During the next reporting period (months 36-48), we will build upon the previous years and focus on the eGenes and tGenes docking experiments and the completion of a complete version of the high-performance computing pipeline for the docking experiments. We will continue to work toward manuscript and participate in the ClinGen consortium. **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 36-48)
Specific Aim 1: Identify SLE-risk Genes	Months	
Subtask 1: Identify SLE-risk single nucleotide polymorphisms (SNPs) in women	1-24	- Effectively complete except submitting African American GWAS study manuscript. - This is complete as of Jan 2024.
Subtask 2: Link SLE-risk SNPs to genes via eQTL, proximity, transcription factor binding, protein coding, gene-based testing	3-27	Completed
Subtask 3: Transcriptomic analysis, differential expression of genes identified in subtask 2	3-30	Completed
Subtask 4: DNA methylation analyses, differential methylation of genes identified in subtask 2	1-24	Completed
Subtask 5: Identify and write potential manuscripts on multi-omic analysis of SLE-risk associated variants and genes.	6-48	Construct additional publications, including Python-based HPC high throughput screening algorithm for docking experiments of FDA-approved drugs.

Milestone(s) Achieved: Lists of SLE-risk associated genes informed by ancestry, tissue, and female sex	3-30	• Completed
Local IRB/IACUC Approval	Completed	Complete/
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	• Largely Completed • Incorporating DNA topology analysis to finalize identification of tGenes. • Completed as of Jan 2024
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	• Completed
Subtask 3: Identify and write potential manuscripts incorporating systems biology and drug target prioritization to evaluate genetic architecture of SLE.	12-48	• Ongoing for Python-based docking pipeline
Milestone(s) Achieved: Lists of prioritized drug targets	6-30	• Completed
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	• Completed
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	• To be completed Feb 2024
Subtask 3: Prioritize drugs from Subtasks 1 and 2 using CoLTS scoring algorithm	6-36	• Completed. Not pursuing CoLTS
Milestone(s) Achieved: Lists of genetically informed FDA-approved drugs and small molecules, novel to treatment of SLE.	12-46	• Partially completed and published lists • Ongoing docking experiments

4. Impact

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

The continued curation of the SLE-risk genetic polymorphisms, linking them to genes and elevated gene expression in SLE cases, and subsequent network and pathway analyses continues motivate targets for drug repositioning. Collectively, the moves the field toward the identification of genetically identified subpopulations that might benefit by specific drug repurposing. The analytic and programming pipelines we have established that can be applied to other lupus and 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 in other ancestral groups.

We note that we contributed to the completion of the African American genome-wide association study, which is out to co-authors for final approval to submit.

What was the impact on other disciplines?

Last year we reported that, motivated by reviewing single cell transcriptomic data for the work in this grant, 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. 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 three manuscripts (listed below), are currently writing an invited review paper for the journal Rheumatology. Dr. Langefeld's effort on the publications and publication costs for these papers were partially supported by this grant (W81XWH-20-1-0686). This grant is gratefully acknowledged. The latter paper, new this cycle of the grant review, responds to a paper that incorrectly attacks a component of our papers by not understanding the underlying statistical theory of hypothesis testing.

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. PMID: PMC8088563.
3. Zimmerman KD, Evans C, Langefeld CD. Reply to: A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat Commun.* 2022 Dec 22;13(1):7852. doi: 10.1038/s41467-022-35520-x. PMID: 36550147; PMID: PMC9780219.

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.

4. 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*)

The pipelines we are and have developed as part of this grant is also being informative for another Department of Health and Human Services grant, *Lung Organ Tissue Equivalent Platform for Modeling Chlorine Gas Toxicology and Medical Countermeasure Efficacy*. (PI Atala; Project Number: OS245776; Source of support: BARDA). The major goals of the grant are to advance the development of a micro- tissue engineered lung organ tissue equivalent platform as tool for modeling pulmonary chlorine gas toxicity and evaluating medical countermeasure efficacy. As part of our (Drs. Langefeld and Ainsworth) effort on that grant, we are leveraging the pipelines developed here, including clustering algorithm and approaches to identification of plausible drug targets and binding experiments.

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 previously noted, 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; 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 (e.g., for prediction refinement). We will continue to use high quality experimentally derived structures when available; but the inclusion of AlphaFold presents an 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. Previously we levered Principal Component Analyses to generate a set of PCs for clustering. This year, we expanded to also use UMAP which has been shown to better capture both local and global data

structures. 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.

Finally, the volume of results we are examining are requiring the development of a high-performance computing Python-based pipeline for the drug binding. This pipeline is still under development but is resulting in substantial increases volume of docking analyses that can be completed.

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

My grants administrator, Nora Snively, died unexpectedly as we were originally preparing this summary. A wonderfully kind, highly competent person, with a wealth of knowledge. She was central to managing the financial for all of my grants. I am grateful others in the department who stepped in and helped stabilize the situation, despite being down 1.75 positions in our administrative team. We are still in the process of interviewing candidates for the positions.

As noted previously, 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. However, we have effectively used Zoom, Microsoft Teams, and WebEx and manuscript productivity is consistent and in good quality journals. Drs. Langefeld and Owens attended the Lupus21 Century meeting this year and both presented our work.

The unanticipated termination of our Python programmer, Mr. Sawyer, has slowed the Python-based pipeline development and we are seeking a Python programmer who has some knowledge of basic undergraduate biology and chemistry.

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.

- Owen KA, Bell KA, Price A, Bachali P, Ainsworth H, Marion MC, Howard TD, Langefeld CD, Shen N, Yazdany J, Dall'era M, Grammer AC, Lipsky PE. Molecular pathways identified from single nucleotide

polymorphisms demonstrate mechanistic differences in systemic lupus erythematosus patients of Asian and European ancestry. Sci Rep. 2023 Apr 1;13(1):5339. doi: 10.1038/s41598-023-32569-6. PMID: 37005464; PMCID: PMC10067935.

- Kain J, Owen KA, Marion MC, Langefeld CD, Grammer AC, Lipsky PE. Mendelian randomization and pathway analysis demonstrate shared genetic associations between lupus and coronary artery disease. Cell Rep Med. 2022 Nov 15;3(11):100805. doi: 10.1016/j.xcrm.2022.100805. Epub 2022 Nov 4. PMID: 36334592; PMCID: PMC9729823.
- Zimmerman KD, Evans C, Langefeld CD. Reply to: A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun. 2022 Dec 22;13(1):7852. doi: 10.1038/s41467-022-35520-x. PMID: 36550147; PMCID: PMC9780219.
- Joint analysis reveals shared autoimmune disease associations and identifies common 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 third review, Nature Genetics) Available at medRxiv

Invited oral presentations:

Langefeld CD. The effect of race and ethnicity on lupus genetics. (2023, September 27-30) Lupus 21st Century, Naples FL, USA.

Poster:

Owen K, Kerns S, Kain J, Marion MC, Langefeld CD, Grammer AG, Lipsky PL. (2023, September 27-30) Assessing genetic risk for cardiovascular disease in SLE using mendelian randomization approaches. Lupus 21st Century, Naples, FL, USA.

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

Nothing to Report.

Other publications, conference papers, and presentations.

Nothing else to report.

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

Nothing further to report over previous year.

7. Participants & Other Collaborating Organizations

What individuals have worked on the project?

Wake Forest University School of Medicine and RILITE investigative teams.

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:	0.6 months (5% 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:	Tim Howard, Ph.D.
Project Role:	Professor
Researcher Identifier (e.g., ORCID ID):	0000-0003-2518-4902
Nearest person month worked:	0.24 months (4% effort for six months)
Contribution to Project:	
Funding Support:	

Name:	Tony Reeves, Ph.D.
Project Role:	Assistant Professor
Researcher Identifier (e.g., ORCID ID):	0000-0002-8209-6020
Nearest person month worked:	0.6 months (5% of effort)
Contribution to Project:	
Funding Support:	

Name:	Miranda C. Marion, MA
Project Role:	Senior Biostatistician
Researcher Identifier (e.g., ORCID ID):	<u>0000-0003-4487-8010</u>
Nearest person month worked:	0.6 calendar months (20% effort for three months)
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:	

RILITE investigative team

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% of 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 (33% of 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:	

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, 2022 through September 30, 2023

Langefeld, CD

Changes in effort from October 1, 2022 through September 30, 2023

ACTIVE

Grant#: 1 R01 DC020715-01

Grant Title: Computer-assisted diagnosis of ear pathologies by combining digital otoscopy with complementary data using machine learning.

Effort: 3%

Grant#: U54 DA049113

Grant Title: Integrated Omics Analysis of Pain: Omics Data Generation Center

Effort: 2%

COMPLETED

Grant#: 5 R01 HD089928-05

Grant Title: Identifying Pharmacogenomic Predictors of Methotrexate Response and Metabolite Biotransformation in JIA.

Effort: 1%

Grant#: 5 R01 NS100417-05

Grant Title: ICH Recovery Grant

Effort: 1%

Howard, TD

ACTIVE

Grant #: 1R01 MD017006-01A1

Grant Title: Migrant and Multi-generational Immigrant Experiences: The effects of Stressors on Epigenetic, Behavioral, and Health-related Outcomes

Effort: 23%

Grant #: R01 EY030521

Grant Title: Predicting Uveitis Onset in Children with Juvenile Idiopathic Arthritis

Effort: 5%

Grant #: U54 DA049113

Grant Title: Integrated Omics Analysis of Pain; Omics Data Generation Center

Effort: 5%

Grant #: Development of a Wake Forest Multi-Species NHP Biorepository to Support Interdisciplinary Aging Studies

Grant Title: 4R33AG073733

Effort: 15%

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

None to report

COMPLETED

Grant #: Internal Pilot

Grant Title: In Utero Imaging in NHPs

Effort: No salary support.

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

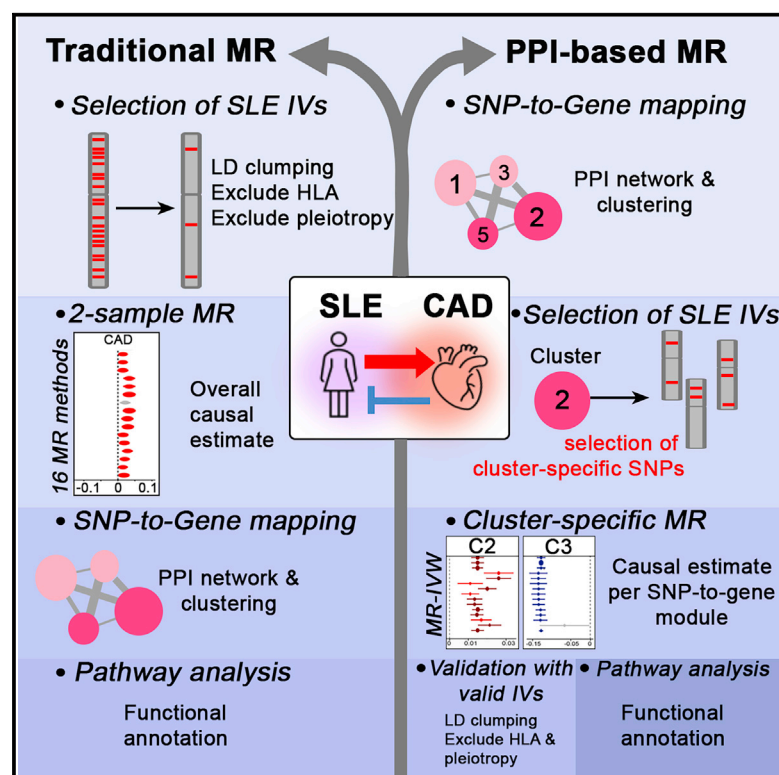
Not applicable/Nothing to report.

9. Appendices

- Kain J, Owen KA, Marion MC, Langefeld CD, Grammer AC, Lipsky PE. Mendelian randomization and pathway analysis demonstrate shared genetic associations between lupus and coronary artery disease. Cell Rep Med. 2022 Nov 15;3(11):100805. doi: 10.1016/j.xcrm.2022.100805. Epub 2022 Nov 4. PMID: 36334592; PMCID: PMC9729823.
- Owen KA, Bell KA, Price A, Bachali P, Ainsworth H, Marion MC, Howard TD, Langefeld CD, Shen N, Yazdany J, Dall'era M, Grammer AC, Lipsky PE. Molecular pathways identified from single nucleotide polymorphisms demonstrate mechanistic differences in systemic lupus erythematosus patients of Asian and European ancestry. Sci Rep. 2023 Apr 1;13(1):5339. doi: 10.1038/s41598-023-32569-6. PMID: 37005464; PMCID: PMC10067935.
- Zimmerman KD, Evans C, Langefeld CD. Reply to: A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun. 2022 Dec 22;13(1):7852. doi: 10.1038/s41467-022-35520-x. PMID: 36550147; PMCID: PMC9780219.

Mendelian randomization and pathway analysis demonstrate shared genetic associations between lupus and coronary artery disease

Graphical abstract



Authors

Jessica Kain, Katherine A. Owen, Miranda C. Marion, Carl D. Langefeld, Amrie C. Grammer, Peter E. Lipsky

Correspondence

kate.owen@ampelbiosolutions.com

In brief

Kain et al. report a positive genetic relationship between SLE and CAD using traditional and comprehensive Mendelian randomization (MR) approaches. MR analysis coupled with network modeling highlights the shared genetic risk underlying the two diseases and identifies putative molecular pathways involved in the development of CAD in SLE.

Highlights

- Genetic predisposition to SLE confers risk of CAD
- Increased CVD risk in SLE involves atherosclerosis rather than cardiac dysfunction
- PPI-based MR identifies pathways with causal implications between SLE and CAD
- Pathway analysis informs therapeutic selection for managing CAD in SLE



Article

Mendelian randomization and pathway analysis demonstrate shared genetic associations between lupus and coronary artery disease

Jessica Kain,^{1,2} Katherine A. Owen,^{1,2,4,*} Miranda C. Marion,³ Carl D. Langefeld,³ Amrie C. Grammer,^{1,2} and Peter E. Lipsky^{1,2}

¹AMPEL BioSolutions, LLC, Charlottesville, VA, USA

²The RILITE Research Institute, Charlottesville, VA, USA

³Department of Biostatistics and Data Science, and Center for Precision Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA

⁴Lead contact

*Correspondence: kate.owen@ampelbiosolutions.com

<https://doi.org/10.1016/j.xcrm.2022.100805>

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 (PPIs) identifies biological networks composed of positive and negative causal sets of genes. In addition, we confirm the causal 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 managing CAD in SLE.

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 has 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 with 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 angiotensin-converting enzyme inhibitors/angiotensin receptor blockers 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 biological processes with causal implications for 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 revealing opportunities for targeted therapeutics.

RESULTS

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

To explore the shared genetic predispositions for SLE and CAD, we first identified single nucleotide polymorphisms (SNPs) associated with each trait in five SLE and one 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 human leukocyte antigen (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 genome-wide association study (GWAS) summary statistics for SLE (GCST003155²⁷), CVD (GCST004280³⁴), and two CAD datasets (CAD-I, GCST000998³⁵ and CAD-II, GCST001479³⁶) 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,” among 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 GWASs suggested a net positive association for non-HLA SLE-associated SNPs on CAD (Figures 1B and 1C). 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 ($p < 5 \times 10^{-8}$) with SLE, excluding SNPs weakly associated ($p < 10^{-5}$) with CVD or confounders (Table S2), followed by stringent linkage disequilibrium (LD) clumping to ensure IV independence³⁷ ($R^2 = 0.001$, 100 kb window, 1000 Genomes EA reference population) (Figure 2A). SLE GWAS summary statistics were used for exposure,²⁷ and multiple CAD GWASs 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 two additional MI GWASs (GCST003117,³⁸ MI-a and GCST011365,³⁹ MI-b) and IS (GCST006906⁴⁰). 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 and 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 GWASs, but not for cardiomyopathy or AFib (Figures 2B and S2A; Table S3). Directional pleiotropy was only detected between the SLE and CAD-b GWASs by the MR-Egger intercept test (Figure S2A), indicating potential bias in the causal estimates based on effect sizes using these summary statistics. Overall, inverse variance weighting (IVW), weighted median, penalized weighted median, maximum likelihood, robust adjusted profile score (RAPS), and pleiotropy residual sum and outlier (PRESSO) were significant in four out of five outcome GWASs (Figure 2B). These results establish a positive causal effect of SLE on CAD and suggest that the increased CVD risk associated

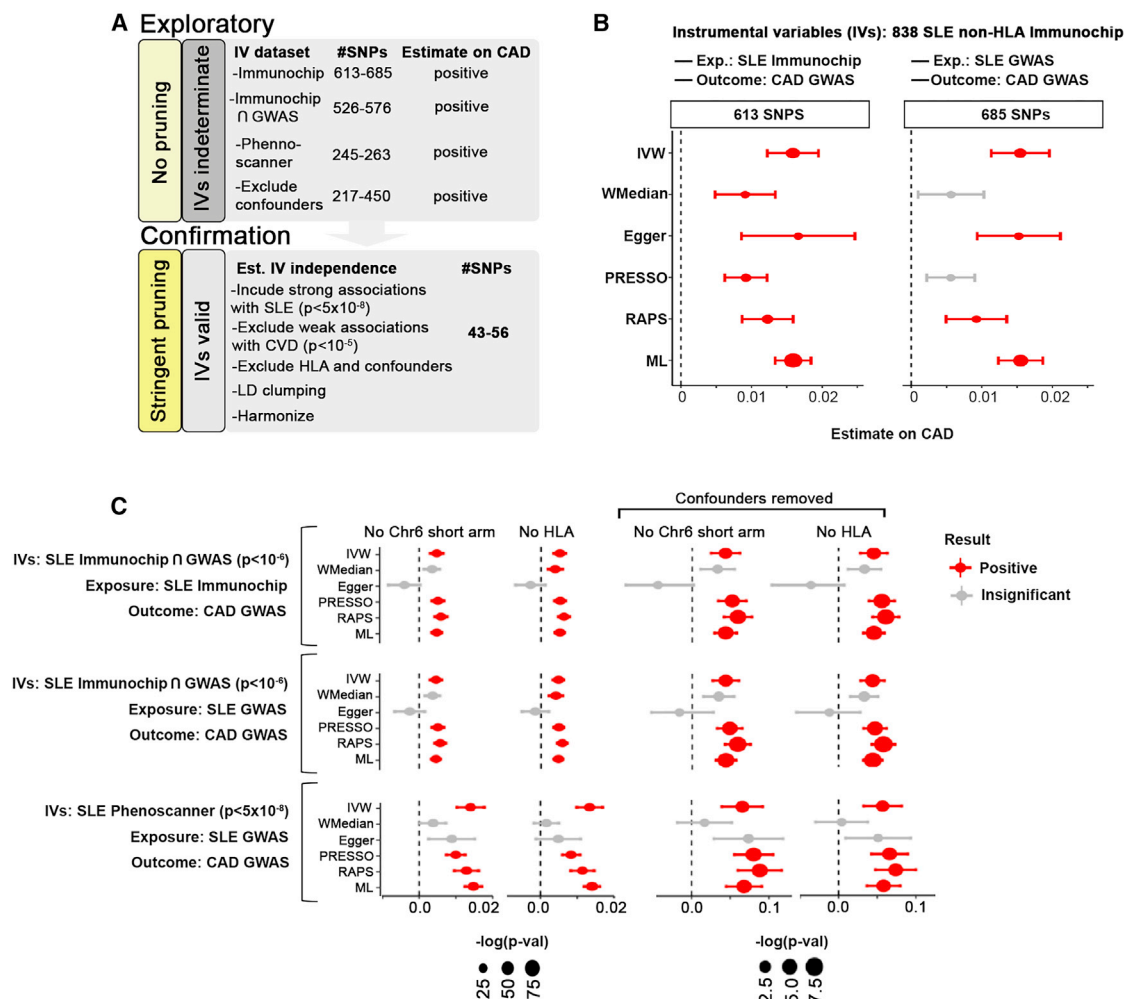


Figure 1. MR demonstrates a positive association of effect sizes of SLE-associated non-HLA SNPs on SLE and CAD

(A) Graphical depiction of the two-stage approach for an initial exploratory analysis using expanded groups of SNPs as IVs followed by a confirmatory analysis using highly curated IVs.

(B and C) Forest plots of six MR causal estimates ($\beta \pm$ standard error). For results, gray indicates insignificant ($p > 0.05$) and red indicates 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) and SLE GWAS (right) 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 (rows 1 and 2) or Phenoscanner reaching genome-wide significance ($p < 5 \times 10^{-8}$, 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: 27–34 Mb, 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.

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 and 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 ($p = 3.46 \times 10^{-5}$ 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

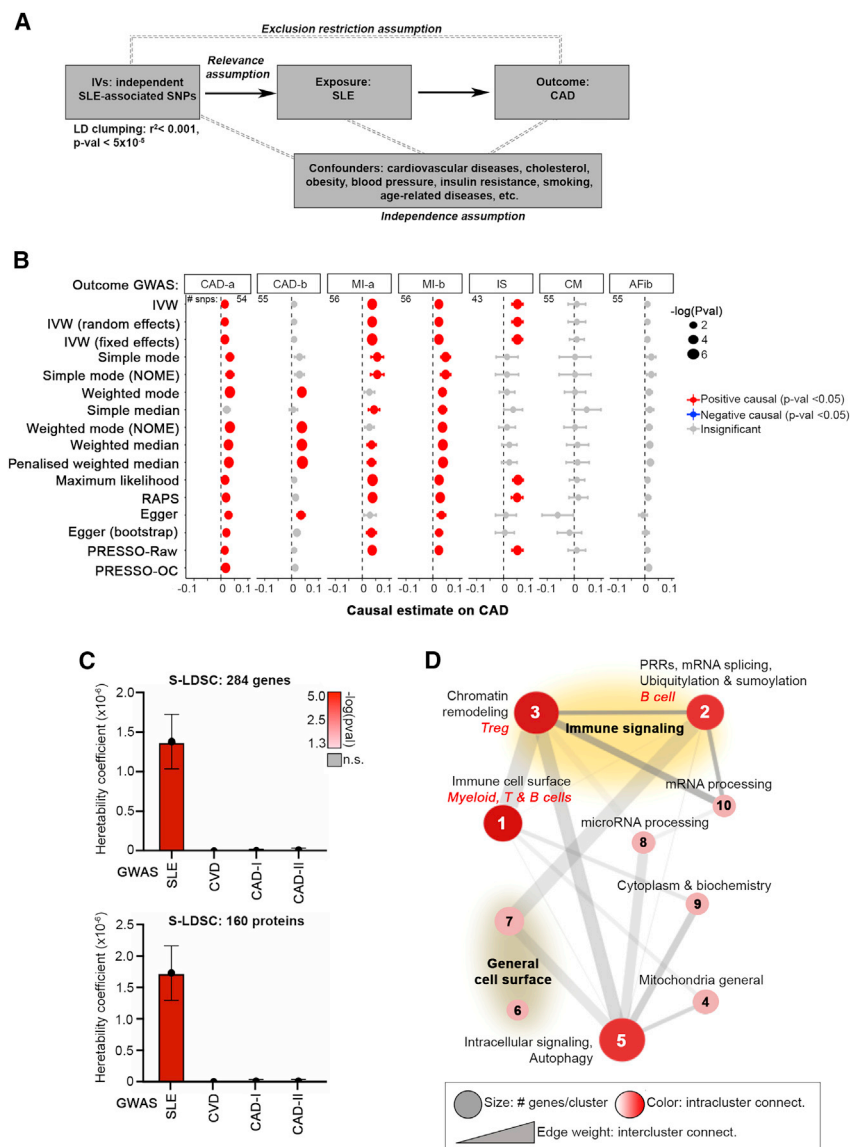


Figure 2. MR demonstrates a net positive causal effect of SLE-associated non-HLA SNPs on CAD

(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 ($\beta \pm$ 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 (outlier correction) estimates indicate insignificant global tests for horizontal pleiotropy. For results, gray indicates insignificant ($p > 0.05$), red indicates positive ($p < 0.05$), and blue indicates negative ($p < 0.05$) causal estimates determined by each MR method. Numbers within forest plots indicate the SNPs used as IVs after harmonization. NOME, no measurement error.

(C) Application of S-LDSC using summary statistics for SLE, CVD, and CAD GWAS to estimate the heritability (coefficient $\pm$ standard error) of the 284 SNP-predicted genes (top) and 160 SNP-predicted proteins from STRINGdb (bottom). 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.

“Th2 pathway,” as well as gene ontology (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-Goutières 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 chr6:28,014,374–33,683,352 (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 and 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 (Figures 3C and 3E).

Non-HLA SLE variants with either significant positive or negative causal estimates on CAD were separately mapped to 236 (Figure 3D) and 244 (Figure 3F) predicted genes, respectively, for clustering and pathway analysis. Positive SNP-predicted gene clusters were enriched in the canonical pathway for “Antigen presentation” and functional categories for major histocompatibility complex (MHC) class I, as well as epigenetic processes, transcription, and endocytosis (Figure 3E and 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

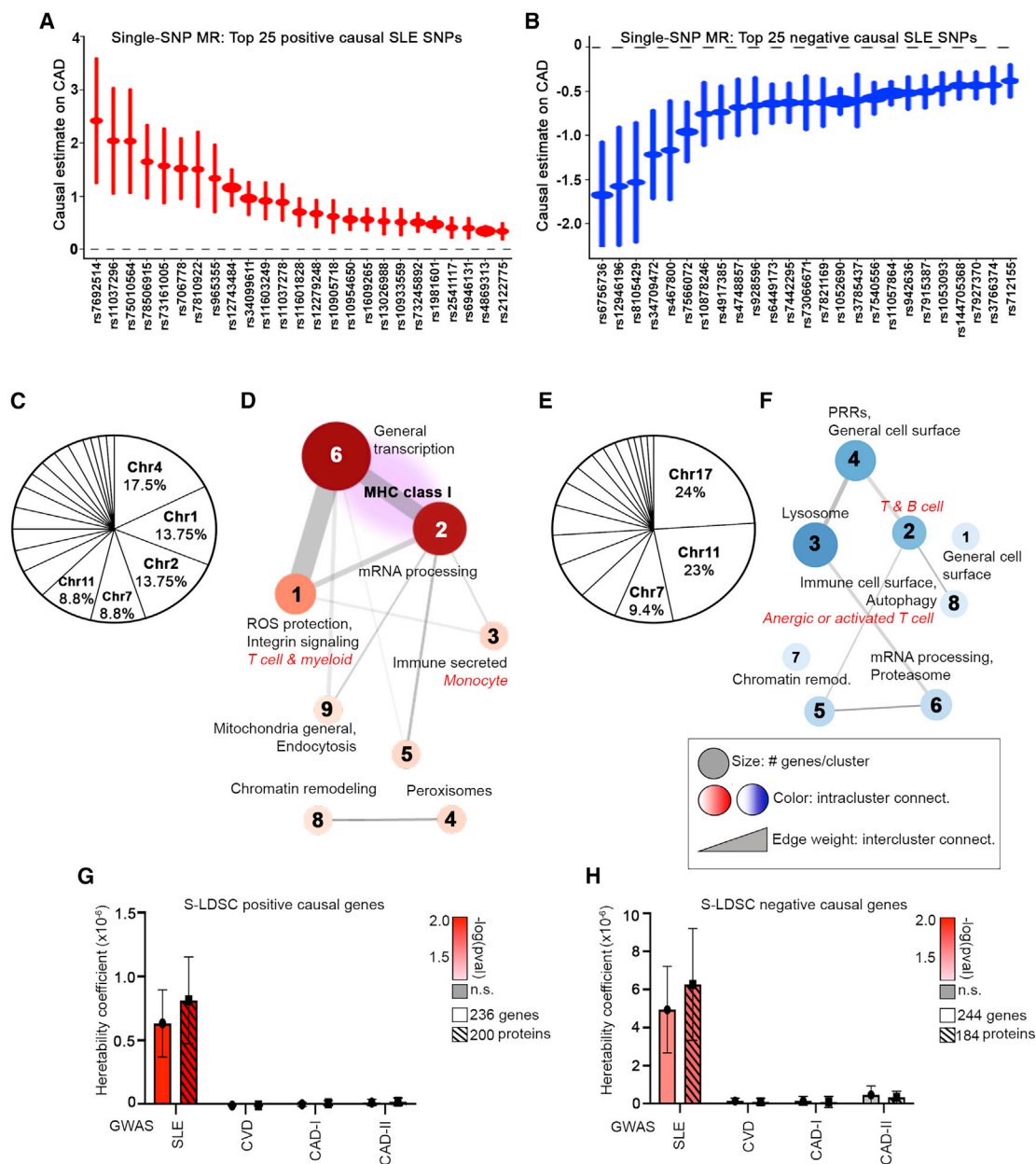


Figure 3. Analysis of SLE-associated SNP-predicted genes with causal effects on CAD by single-SNP MR

(A and B) Forest plots ($\beta \pm$ 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) and 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 and H) S-LDSC using summary statistics for SLE, CVD, and CAD GWAS to estimate the heritability (coefficient $\pm$ 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.

protein degradation (“Proteasome,” cluster 6 and “Autophagy,” cluster 8; Figure 3F and 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 (Figures 3G and 3H).

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 SLE⁴² 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 six 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,” and “Notch signaling,” as well as GO term for “Interferon-gamma mediated signaling pathway” (GO:0060334) (Figures S5B and S5C). 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 SNP 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 (Figures 4B, 4C, and 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 seven MR methods (Figures 4B and 4C). 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 signifi-

cant 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 (Figures 5C, 5D, and 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 two CAD and two 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 to be 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 and Table S8). These results show that a wide range of SNP-predicted biological functions known to be involved in SLE pathogenesis have causal implications for 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 transendothelial 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 CVD, 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 and 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 high-density lipoprotein cholesterol (clusters 24 and 50) (Figure 5B and Table S8). Pathway enrichment was further reflected in assigned functional categories, with reactive oxygen species protection (clusters 24 and 45), nuclear receptor 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.

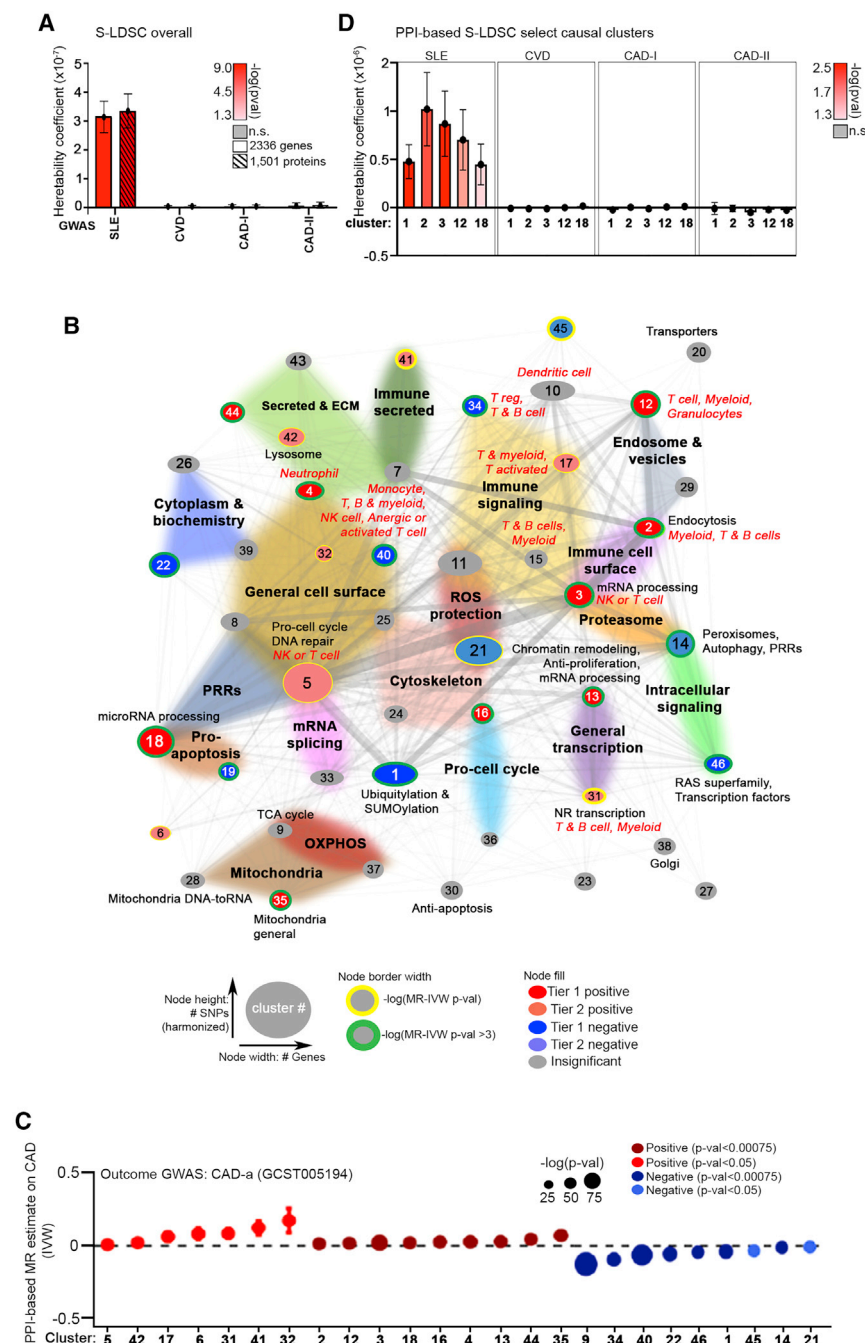


Figure 4. SLE-derived gene network with causal implications for CAD and PPI-based MR

(A) S-LDSC using summary statistics for SLE, CVD, and CAD GWAS to estimate the heritability (coefficient $\pm$ 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 indicate significant positive or negative estimates by MR-IVW or at least 7/16 MR methods (tier 2); gray indicates insignificant. Thickness of the yellow border is roughly proportional to the negative log of the MR-IVW p value. Green border indicates clusters with $-\log(\text{MR-IVW } p\text{ value}) > 3$.

(C) Forest plots from PPI-based MR showing estimates ($\beta \pm$ standard error) calculated by MR-IVW for select positive and negative clusters.

(D) PPI-based S-LDSC (coefficient $\pm$ standard error) using GWAS summary statistics for SLE, CVD, and CAD. Bar color indicates coefficient significance.

positive, significant, and consistent than those based on all SNPs (Figures 6B and 6C; 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

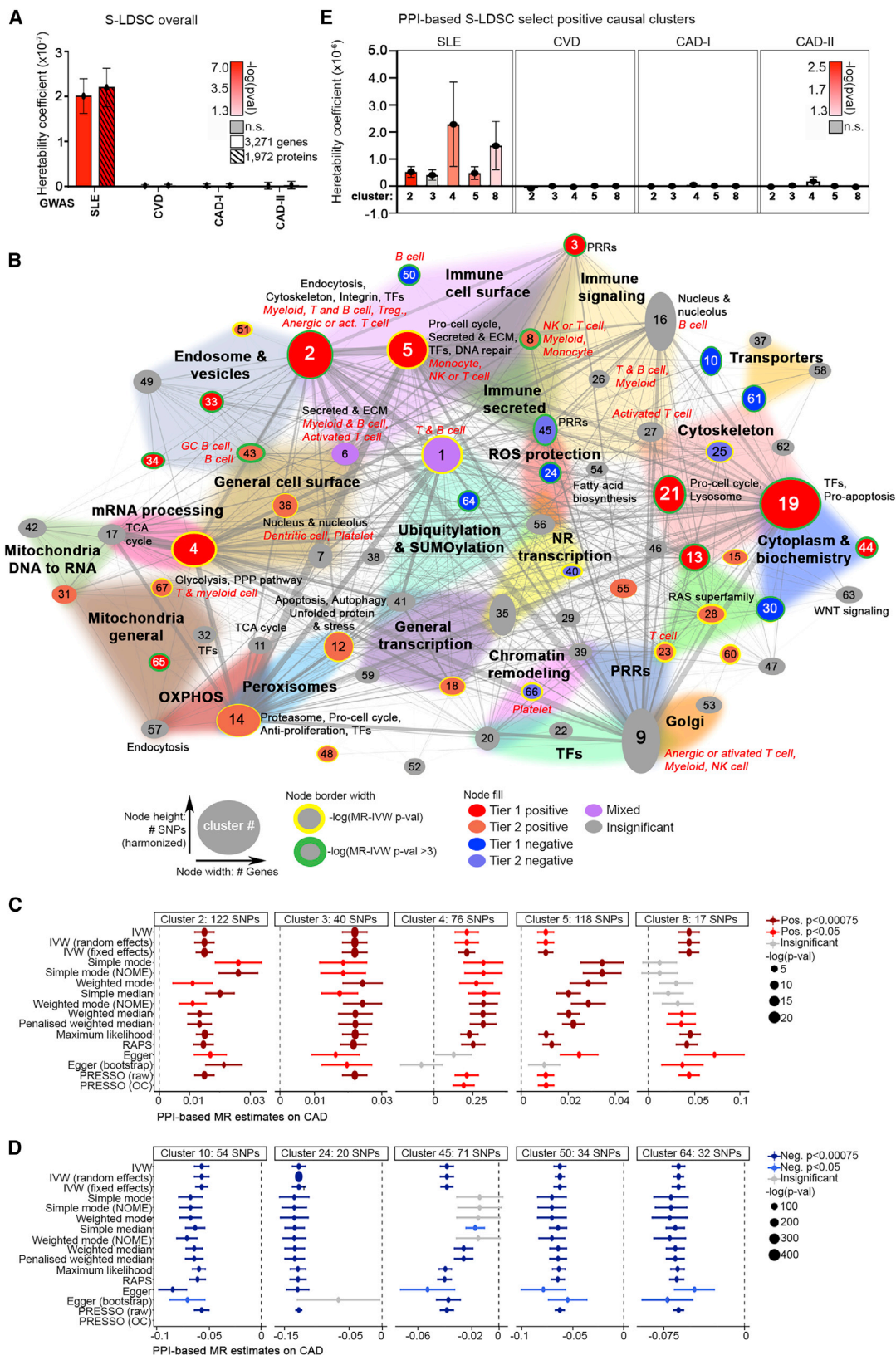
PPI-based MR stratifies SNPs, genes, and networks underlying the positive and negative causal effects of SLE on CAD

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

demonstrate that PPI-based MR can be used to identify independent IVs satisfying MR assumptions that underlie 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



(legend on next page)

targeting T and B cells and inflammatory cytokines, including daratumumab (CD38), belimumab (TNFSF13), elotuzumab (SLAMF7), abatacept (CD80/86), iberdimide (IKZF1/IKZF3), and sarilumab (IL6R). Broader analysis of pathway categories also suggested the utility of targeting interferon signaling with anifrolumab (cluster 8), as well as antiplatelet/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 low-density lipoprotein (LDL) receptor targeted by alirocumab and evolocumab, and oxidized 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 for CAD. New findings suggest that the causal relationship with SLE appears to be focused on the atherosclerotic process, evidenced by positive estimates with CAD, MI, and IS, 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 for 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, ef-

forts 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 MR⁴⁷ 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 overpruning.

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 causal 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 predictions, and regulatory elements in an effort to

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 $\pm$ 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 indicate significant positive or negative estimates by MR-IVW or at least 7/16 MR methods (tier 2); purple indicates mixed estimates; gray indicates 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 and D) Forest plots from PPI-based MR showing estimates ($\beta \pm$ standard error) calculated by 16 MR methods for select positive (C) and negative (D) clusters. The number of SNPs used as IVs for each cluster are indicated in the plots.

(E) PPI-based S-LDSC (coefficient $\pm$ standard error) using GWAS summary statistics for SLE, CVD, and CAD. Bar color indicates coefficient significance.

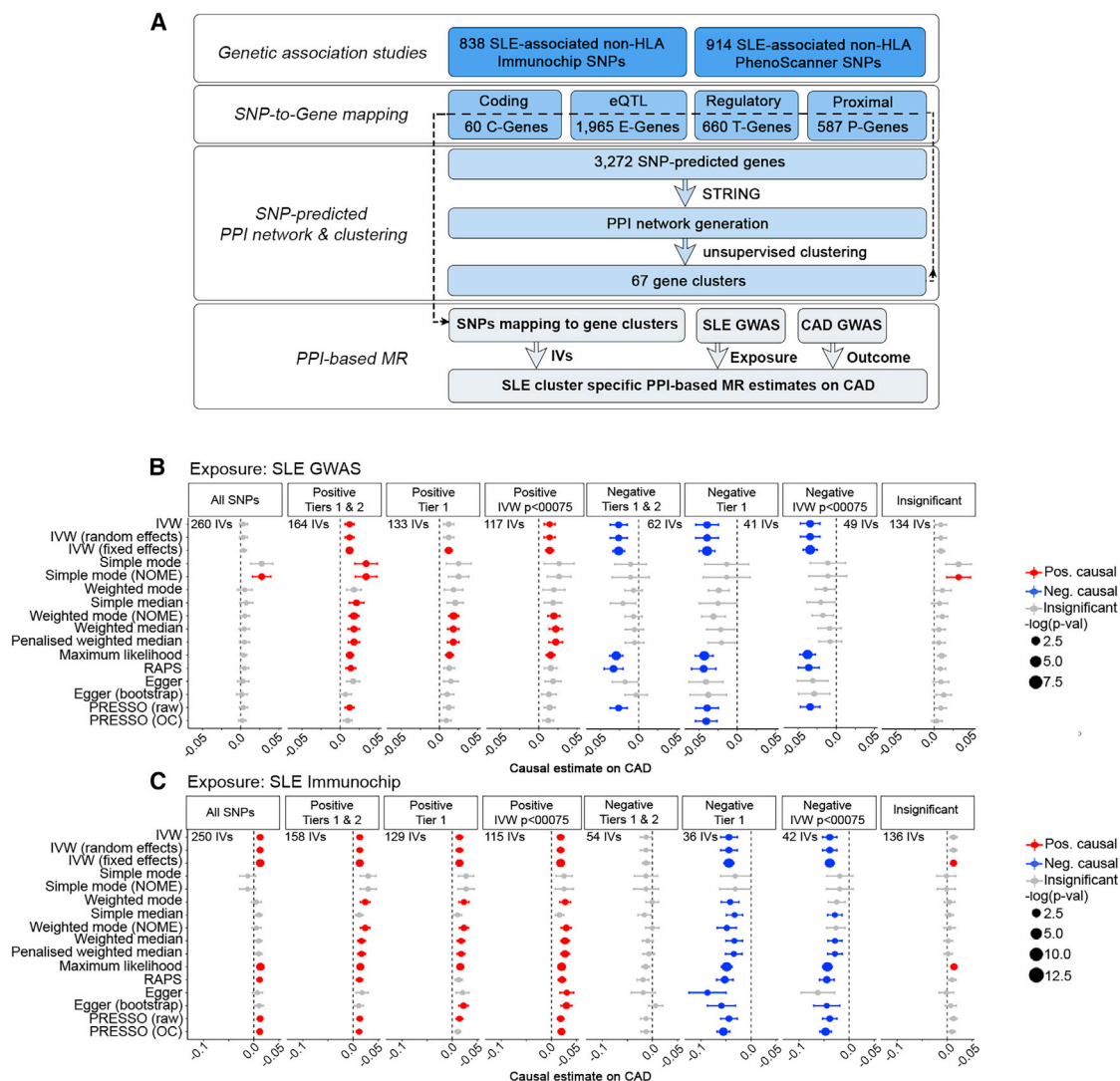


Figure 6. PPI-based MR identifies SLE SNPs with positive and negative causal effects on CAD

(A) Workflow depicting PPI-based MR.

(B and C) PPI-based MR validation. Forest plots ($\beta \pm$ 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, gray indicates insignificant ($p > 0.05$), dark red indicates positive ($p < 0.00075$), red indicates positive ($p < 0.05$), dark blue indicates negative ($p < 0.00075$), and blue indicates negative ($p < 0.05$) by each MR method.

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 for 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 among positive causal clusters have been hypothesized and/or demonstrated to contribute to CVD in SLE patients, including glucocorticoids, neutrophil cell death (NETosis) and degranulation, tumor necrosis factor-like weak inducer of apoptosis (TWEAK) signaling, canonical and alternative complement pathways, Th1 differentiation, and 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

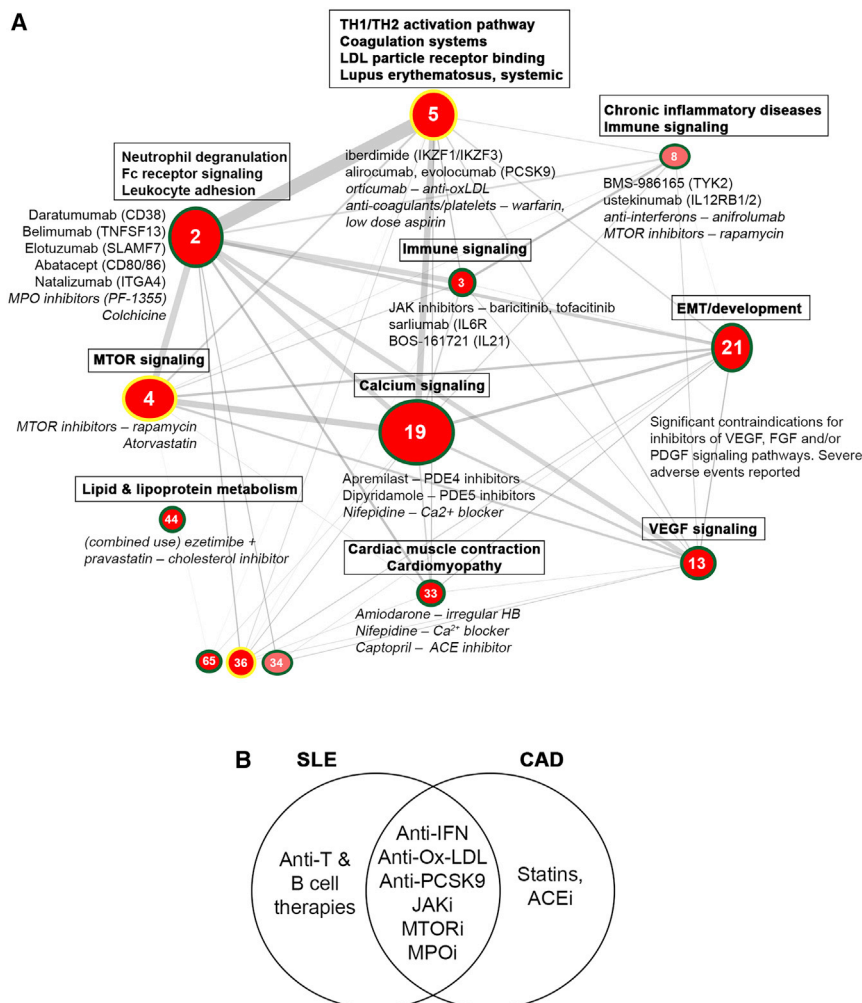


Figure 7. Genes and molecular pathways associated with positive causal clusters identify therapeutic interventions for managing CAD in SLE

(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.

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 approved by the Food and Drug Administration 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 toward European ancestries⁵³. This is especially problematic given the increased CVD morbidity and mortality in SLE patients of African ancestry⁵⁴ 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.

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 through 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 cholesterol 27-hydroxylase and is a liver X receptor (LXR) target activated by oxysterols as well as a target of retinoid X receptor (RXR) and peroxisome proliferator activated receptor (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

This is important with respect to the higher disease burden observed in patients of African ancestry, 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), Geno-type-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.^{55,56} 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. Ingenuity Pathway Analysis (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 needed to consider novel therapeutic interventions in these conditions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **EXPERIMENTAL MODELS AND SUBJECT DETAILS**
- **METHOD DETAILS**
 - Identification of SLE- and CAD-associated SNPs and overlap
 - Identification of SNP-predicted genes
 - Network analysis and visualization
 - Functional gene set analysis
 - Drug candidate identification
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

- Linkage disequilibrium score regression (LDSC) genetic correlations
- Stratified linkage disequilibrium score regression (S-LDSC)
- Selection of valid, independent instrumental variables for traditional MR analysis
- Mendelian randomization (MR)
- PPI-based MR
- Monte Carlo simulations for expected MR results using random sets of Immunochip-derived SNP-to-gene modules

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xcrm.2022.100805>.

ACKNOWLEDGMENTS

The work presented in this paper 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 (imgoldmanfoundation.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.C.M.; writing – original draft, J.K., P.E.L., and K.A.O.; writing – review & editing, C.D.L., M.C.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.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 10, 2022

Revised: July 8, 2022

Accepted: October 12, 2022

Published: November 4, 2022

REFERENCES

1. Yen, E.Y., and Singh, R.R. (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, 1251–1255.
2. Liu, Y., and Kaplan, M.J. (2018). Cardiovascular disease in systemic lupus erythematosus: an update. *Curr. Opin. Rheumatol.* 30, 441–448.
3. Gupta, S., and Kaplan, M.J. (2021). Bite of the wolf: innate immune responses propagate autoimmunity in lupus. *J. Clin. Invest.* 131, e144918.
4. Thomas, G., Mancini, J., Jourde-Chiche, N., Sarlon, G., Amoura, Z., Harlé, J.R., Jougla, E., and Chiche, L. (2014). Mortality associated with systemic lupus erythematosus in France assessed by multiple-cause-of-death analysis. *Arthritis Rheumatol.* 66, 2503–2511.
5. Teixeira, V., and Tam, L.S. (2017). Novel insights in systemic lupus erythematosus and atherosclerosis. *Front. Med.* 4, 262.
6. Reppe Moe, S., Haukeland, H., Molberg, Ø., and Lerang, K. (2021). Long-term outcome in systemic lupus erythematosus; knowledge from population-based cohorts. *J. Clin. Med.* 10, 4306. <https://doi.org/10.3390/jcm10194306>.

7. Katz, G., Smilowitz, N.R., Blazer, A., Clancy, R., Buyon, J.P., and Berger, J.S. (2019). Systemic lupus erythematosus is associated with increased prevalence of atherosclerotic cardiovascular disease in hospitalized patients. *Mayo Clin. Proc.* 94, 1436–1443.
8. Lopez, E.O., Ballard, B.D., and Jan, A. (2021). Cardiovascular disease. In *StatPearls* [Internet], editorial board (StatPearls Publishing). PMID: 30571040; Bookshelf ID: NBK535419.
9. Petri, M., Perez-Gutthann, S., Spence, D., and Hochberg, M.C. (1992). Risk factors for coronary artery disease in patients with systemic lupus erythematosus. *Am. J. Med.* 93, 513–519.
10. Moghaddam, B., Marozoff, S., Li, L., Sayre, E.C., and Zubieta, J.A.A. (2021). All-cause and cause-specific mortality in systemic lupus erythematosus: a population-based study. *Rheumatology* 61, 367–376.
11. Bernatsky, S., Boivin, J.F., Joseph, L., Manzi, S., Ginzler, E., Gladman, D.D., Urowitz, M., Fortin, P.R., Petri, M., Barr, S., et al. (2006). Mortality in systemic lupus erythematosus. *Arthritis Rheum.* 54, 2550–2557.
12. Wu, G.C., Liu, H.R., Leng, R.X., Li, X.P., Li, X.M., Pan, H.F., and Ye, D.-Q. (2016). Subclinical atherosclerosis in patients with systemic lupus erythematosus: a systemic review and meta-analysis. *Autoimmun. Rev.* 15, 22–37.
13. Yazdany, J., Pooley, N., Langham, J., Nicholson, L., Langham, S., Embleton, N., Wang, X., Desta, B., Barut, V., and Hammond, E. (2020). Original research: systemic lupus erythematosus; stroke and myocardial infarction risk: a systematic review and meta-analysis. *RMD Open* 6, e001247.
14. Magder, L.S., and Petri, M. (2012). Incidence of and risk factors for adverse cardiovascular events among patients with systemic lupus erythematosus. *Am. J. Epidemiol.* 176, 708–719.
15. Petri, M.A., Barr, E., and Magder, L.S. (2019). Development of a systemic lupus erythematosus cardiovascular risk equation. *Lupus Sci. Med.* 6, e000346.
16. Levinson, D.J., Abugroun, A., Daoud, H., and 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., and 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 anti-phospholipid syndrome: a case-control study over 10 years. *Lupus Sci. Med.* [Internet] 8, e000454.
18. Bakshi, J., Segura, B.T., Wincup, C., and Rahman, A. (2018). Unmet needs in the pathogenesis and treatment of systemic lupus erythematosus. *Clin. Rev. Allergy Immunol.* 55, 352–367.
19. Calvo-Alén, J., Alarcón, G.S., Campbell, R., Fernández, M., Reveille, J.D., and Cooper, G.S. (2005). Lack of recording of systemic lupus erythematosus in the death certificates of lupus patients. *Rheumatology* 44, 1186–1189.
20. Falasinnu, T., Rossides, M., Chaichian, Y., and Simard, J.F. (2018). Do death certificates underestimate the burden of rare diseases? The example of systemic lupus erythematosus mortality, Sweden, 2001–2013. *Public Health Rep.* 133, 481–488.
21. Ashley, E.A., Hershberger, R.E., Caleshu, C., Ellinor, P.T., Garcia, J.G.N., Herrington, D.M., Ho, C.Y., Johnson, J.A., Kittner, S.J., Macrae, C.A., et al. (2012). Genetics and cardiovascular disease: a policy statement from the American Heart Association. *Circulation* 126, 142–157.
22. Leonard, D., Svenungsson, E., Dahlqvist, J., Alexsson, A., Årlestig, L., Taylor, K.E., Sandling, J.K., Bengtsson, C., Frodlund, M., Jönsen, A., et al. (2018). Novel gene variants associated with cardiovascular disease in systemic lupus erythematosus and rheumatoid arthritis. *Ann. Rheum. Dis.* 77, 1063–1069.
23. Ramirez, G.A. (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., Jönsen, A., Bengtsson, A.A., Sturfelt, G., Rantapää-Dahlqvist, S., 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, 834–840.
25. Leonard, D., Svenungsson, E., Sandling, J.K., Berggren, O., Jönsen, A., Bengtsson, C., Wang, C., Jensen-Ustad, K., Granstam, S.-O., Bengtsson, A.A., et al. (2013). Coronary heart disease in systemic lupus erythematosus is associated with interferon regulatory factor-8 gene variants. *Circ. Cardiovasc. Genet.* 6, 255–263.
26. Owen, K.A., Price, A., Ainsworth, H., Aidukaitis, B.N., Bachali, P., Catalina, M.D., Dittman, J.M., Howard, T.D., Kingsmore, K.M., Labonte, A.C., et al. (2020). Analysis of trans-ancestral SLE risk loci identifies unique biologic networks and drug targets in African and European ancestries. *Am. J. Hum. Genet.* 107, 864–881.
27. Benthams, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, P., Behrens, T.W., Martin, J., Fairfax, B.P., Knight, J.C., Chen, L., 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, 1457–1464.
28. 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. (2016). Identification of a systemic lupus erythematosus risk locus spanning ATG16L2, FCHSD2, and P2RY2 in Koreans. *Arthritis Rheumatol.* 68, 1197–1209.
29. Morris, D.L., Sheng, Y., Zhang, Y., Wang, Y.-F., Zhu, Z., Tombleson, P., Chen, L., Cunningham Graham, D.S., Benthams, J., Roberts, A.L., 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, 940–946.
30. 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. (2016). High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat. Genet.* 48, 323–330.
31. Langefeld, C.D., Ainsworth, H.C., Cunningham Graham, D.S., Kelly, J.A., Comeau, M.E., Marion, M.C., Howard, T.D., Ramos, P.S., Coker, J.A., Morris, D.L., et al. (2017). Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* 8, 16021.
32. Van Der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* 122, 433–443.
33. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235.
34. Warren, H.R., Evangelou, E., Cabrera, C.P., Gao, H., Ren, M., Mifsud, B., Ntalla, I., Surendran, P., Liu, C., Cook, J.P., et al. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.* 49, 403–415.
35. Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F.R., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* 43, 333–338.
36. Howson, J.M.M., Zhao, W., Barnes, D.R., Ho, W.K., Young, R., Paul, D.S., Waite, L.L., Freitag, D.F., Fauman, E.B., Salfati, E.L., et al. (2017). Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* 49, 1113–1119.
37. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E.M. (2018). A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 27, e1608.
38. Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C., et al. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47, 1121–1130.

39. Hartiala, J.A., Han, Y., Jia, Q., Hilser, J.R., Huang, P., Gukasyan, J., Schwartzman, W.S., Cai, Z., Biswas, S., Tréguët, D.A., et al. (2021). Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. *Eur. Heart J.* **42**, 919–933.
40. Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L., Giese, A.-K., van der Laan, S.W., 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**, 524–537.
41. Nielsen, J.B., Thoroldsdottir, R.B., Fritsche, L.G., Zhou, W., Skov, M.W., Graham, S.E., Herron, T.J., McCarthy, S., Schmidt, E.M., Sveinbjornsson, G., et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239.
42. Raj, P., Rai, E., Song, R., Khan, S., Wakeland, B.E., 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, R.W., Wells, G.A., Stewart, A.F.R., Erdmann, J., Shah, S.H., Ferguson, J.F., Hall, A.S., Anand, S.S., Burnett, M.S., Epstein, S.E., 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**, 217–225.
44. Trowsdale, J., and Knight, J.C. (2013). Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323.
45. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., 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., and Perl, A. (2020). Management of cardiovascular disease in patients with systemic lupus erythematosus. *Expert Opin. Pharmacother.* **21**, 1617–1628.
47. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B.B., and Hopewell, J.C. (2017). Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables. *Genet. Epidemiol.* **41**, 714–725.
48. Reiss, A.B., Anwar, K., Merrill, J.T., Chan, E.S.L., Awadallah, N.W., Cronstein, B.N., Michael Belmont, H., 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**, 591–598.
49. Quinn, C.M., Jessup, W., Wong, J., Kritharides, L., and Brown, A.J. (2005). Expression and regulation of sterol 27-hydroxylase (CYP27A1) in human macrophages: a role for RXR and PPARgamma ligands. *Biochem. J.* **385** (Pt 3), 823–830.
50. Bilotta, M.T., Petillo, S., Santoni, A., and Cipitelli, 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., and 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**, 825–835.
52. Yao Mattisson, I., Rattik, S., Björkbacka, H., Ljungcrantz, I., Terrinoni, M., Levens, M., Holmgren, J., Fredrikson, G.N., Gullstrand, B., Bengtsson, A.A., et al. (2021). Immune responses against oxidized LDL as possible targets for prevention of atherosclerosis in systemic lupus erythematosus. *Vascul. Pharmacol.* **140**, 106863.
53. Demirci, F.Y., Wang, X., Kelly, J.A., Morris, D.L., Barmada, M.M., Feingold, E., Kao, A.H., Sivits, K.L., 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**, 174–183.
54. Barnado, A., Carroll, R.J., Casey, C., Wheless, L., Denny, J.C., and Croford, L.J. (2018). Phenome-wide association study identifies marked increased in burden of comorbidities in African Americans with systemic lupus erythematosus. *Arthritis Res. Ther.* **20**, 69.
55. Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12**, 861–874.
56. Hrdlickova, B., de Almeida, R.C., Borek, Z., and 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**, 1910–1922.
57. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J., et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209.
58. Kamat, M.A., Blackshaw, J.A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A.S., and Staley, J.R. (2019). PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853.
59. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3; Duncan, L., Perry, J.R.B., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241.
60. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.
61. Morris, J.H., Apeltsin, L., Newman, A.M., Baumbach, J., Wittkop, T., Su, G., Bader, G.D., and Ferrin, T.E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinf.* **12**, 436.
62. Wang, J., Dai, X., Berry, L.D., Cogan, J.D., Liu, Q., and Shyr, Y. (2019). HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112.
63. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. *Genome Biol.* **17**, 122.
64. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., 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**, D605–D612.
65. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.* **14**, 128.
66. Kuhn, M., Szklarczyk, D., Franceschini, A., Campillos, M., Von Mering, C., Jensen, L.J., Beyer, A., and Bork, P. (2010). Stitch 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.* **38**, D552–D556.
67. Keenan, A.B., Jenkins, S.L., Jagodnik, K.M., Koplev, S., He, E., Torre, D., Wang, Z., Dohman, A.B., Silverstein, M.C., 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**, 13–24.
68. The GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. *Nat. Gen.* **45**, 580–585.
69. Labonte, A.C., Kegerreis, B., Geraci, N.S., Bachali, P., Madamanchi, S., Robl, R., Catalina, M.D., Lipsky, P.E., and Grammer, A.C. (2018). Identification of alterations in macrophage activation associated with disease activity in systemic lupus erythematosus. *PLoS One* **13**, e0208132.
70. Catalina, M.D., Bachali, P., Geraci, N.S., Grammer, A.C., and Lipsky, P.E. (2019). Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. *Commun. Biol.* **2**, 140.

71. Catalina, M.D., Owen, K.A., Labonte, A.C., Grammer, A.C., and Lipsky, P.E. (2020). The pathogenesis of systemic lupus erythematosus: harnessing big data to understand the molecular basis of lupus. *J. Autoimmun.* **110**, 102359.
72. Ren, J., Catalina, M.D., Eden, K., Liao, X., Read, K.A., Luo, X., McMillan, R.P., Hulver, M.W., Jarpe, M., Bachali, P., 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.
73. Yang, W., Shen, N., Ye, D.Q., Liu, Q., Zhang, Y., Qian, X.X., Hirankarn, N., Ying, D., Pan, H.-F., Mok, C.C., et al. (2010). Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet.* **6**, e1000841.
74. Chung, S.A., Taylor, K.E., Graham, R.R., Nititham, J., Lee, A.T., Ortmann, W.A., Jacob, C.O., Alarcón-Riquelme, M.E., Tsao, B.P., Harley, J.B., et al. (2011). Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet.* **7**, e1001323.
75. Yang, J., Yang, W., Hirankarn, N., Ye, D.Q., Zhang, Y., Pan, H.F., Mok, C.C., Chan, T.M., Wong, R.W.S., Mok, M.Y., et al. (2011). ELF1 is associated with systemic lupus erythematosus in Asian populations. *Hum. Mol. Genet.* **20**, 601–607.
76. Lee, Y.H., Bae, S.C., Choi, S.J., Ji, J.D., and Song, G.G. (2012). Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. *Mol. Biol. Rep.* **39**, 10627–10635.
77. 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 erythematosus in Japanese. *PLoS Genet.* **8**, e1002455.
78. 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**, 41–51.
79. Armstrong, D.L., Zidovetzki, R., Alarcón-Riquelme, M.E., Tsao, B.P., Criswell, L.A., Kimberly, R.P., Harley, J.B., Sivits, K.L., Vyse, T.J., Gaffney, P.M., et al. (2014). GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. *Genes Immun.* **15**, 347–354.
80. 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. (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**, 932–943.
81. Hom, G., Graham, R.R., Modrek, B., Taylor, K.E., Ortmann, W., Garnier, S., Lee, A.T., Chung, S.A., Ferreira, R.C., Pant, P.V.K., et al. (2008). Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N. Engl. J. Med.* **358**, 900–909.
82. Graham, R.R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C.J., Leon, J.M., Burt, N.P., Guiducci, C., Parkin, M., Gates, C., et al. (2008). Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 1059–1061.
83. 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. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PTK, KIAA1542 and other loci. *Nat. Genet.* **40**, 204–210.
84. Kozyrev, S.V., Abelson, A.K., Wojcik, J., Zaghlool, A., Linga Reddy, M.V.P., Sanchez, E., Gunnarsson, I., Svenungsson, E., Sturfelt, G., Jönsen, A., et al. (2008). Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 211–216.
85. Oishi, T., Iida, A., Otsubo, S., Kamatani, Y., Usami, M., Takei, T., Uchida, K., Tsuchiya, K., Saito, S., Ohnishi, 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**, 151–162.
86. Gateva, V., Sandling, J.K., Hom, G., Taylor, K.E., Chung, S.A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R.C., 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**, 1228–1233.
87. Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., 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**, 1234–1237.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GTEX v.6.8	The Geno-type-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.finnngen.fi/fi
MR-Base	Hemani et al., 2018 ⁴⁵	https://www.mrbase.org
Phenoscanner	Staley et al., 2016 ⁵⁷ ; Kamat et al., 2019 ⁵⁸	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 ³³	https://github.com/bulik/ldsc
LDSC	Bulik-Sullivan et al., 2015 ⁵⁹	https://github.com/bulik/ldsc
Cytoscape v.3.9.1	Shannon et al., 2003 ⁶⁰	https://cytoscape.org
Clustermaker2 v.1.2.1	Morris et al., 2011 ⁶¹	https://apps.cytoscape.org
TwoSample MR	Hemani et al., 2018 ⁴⁵	https://github.com/MRCIEU/TwoSampleMR
PPI-based MR	This manuscript	https://doi.org/10.6084/Lm9.figshare.21225251
Other		
Human Active Enhancers to interpret Regulatory variants (HACER)	Wang et al., 2019 ⁶²	http://bioinfo.vanderbilt.edu/AE/HACER
Variant Effect Predictor (VEP)	McLaren et al., 2016 ⁶³	ensembl.org/info/docs/tools/vep
Search Tool for the Retrieval of Interacting Genes/proteins (STRING) v. 11.0b	Szklarczyk et al., 2021 ⁶⁴	https://string-db.org
EnrichR	Chen et al., 2013 ⁶⁵	https://maayanlab.cloud/Enrichr/
Search Tool for Interacting Chemicals (STITCH)	Kuhn et al., 2009 ⁶⁶	http://stitch.embl.de
Library of Integrated Network-based Cellular Signatures (LINCS)	Keenan et al., 2018 ⁶⁷	http://www.lincs.hms.harvard.edu/db/

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 to conduct PPI-based MR analyses have been deposited on figshare (www.figshare.com; <https://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 GTEx⁶⁸ 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 variants⁶² (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 predictor⁶³ (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 STRING⁶⁴ (<https://string-db.org>, version 11.0b), and resulting networks were imported into Cytoscape⁶⁰ (version 3.6.1) for visualization and partitioned with MCODE via the clusterMaker2⁶¹ (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.^{69–71} I-Scope is a custom clustering tool used to identify immune cell types in large gene datasets.⁷² The Ingenuity Pathway Analysis (IPA; <https://www.qiagenbioinformatics.com>) platform and EnrichR⁶⁵ (<https://maayanlab.cloud/Enrichr/>) web server provided additional molecular pathway enrichment analysis.

Drug candidate identification

Drug candidates were identified using LINCS,⁶⁷ STITCH⁶⁶ (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 ANALYSIS

Linkage disequilibrium score regression (LDSC) genetic correlations

LDSC⁵⁹ 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-LDSC³³ 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 $< 5 \times 10^{-8}$) associated with SLE^{27–29,73–87} 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 $< 1 \times 10^{-5}$) 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 clumping³⁷ 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-Base⁴⁵ (<https://www.mrbase.org>) TwoSampleMR⁴⁵ 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 Immunochip³¹ and Phenoscanner database^{57,58} 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 $< 1 \times 10^{-5}$) 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 clumping³⁷ 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 3–152 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.



OPEN Molecular pathways identified from single nucleotide polymorphisms demonstrate mechanistic differences in systemic lupus erythematosus patients of Asian and European ancestry

Katherine A. Owen^{1✉}, Kristy A. Bell¹, Andrew Price¹, Prathyusha Bachali¹, Hannah Ainsworth², Miranda C. Marion², Timothy D. Howard³, Carl D. Langefeld², Nan Shen⁴, Jinoos Yazdany⁵, Maria Dall'era⁵, Amrie C. Grammer¹ & Peter E. Lipsky¹

Systemic lupus erythematosus (SLE) is a multi-organ autoimmune disorder with a prominent genetic component. Individuals of Asian-Ancestry (AsA) disproportionately experience more severe SLE compared to individuals of European-Ancestry (EA), including increased renal involvement and tissue damage. However, the mechanisms underlying elevated severity in the AsA population remain unclear. Here, we utilized available gene expression data and genotype data based on all non-HLA SNP associations in EA and AsA SLE patients detected using the Immunochip genotyping array. We identified 2778 ancestry-specific and 327 trans-ancestry SLE-risk polymorphisms. Genetic associations were examined using connectivity mapping and gene signatures based on predicted biological pathways and were used to interrogate gene expression datasets. SLE-associated pathways in AsA patients included elevated oxidative stress, altered metabolism and mitochondrial dysfunction, whereas SLE-associated pathways in EA patients included a robust interferon response (type I and II) related to enhanced cytosolic nucleic acid sensing and signaling. An independent dataset derived from summary genome-wide association data in an AsA cohort was interrogated and identified similar molecular pathways. Finally, gene expression data from AsA SLE patients corroborated the molecular pathways predicted by SNP associations. Identifying ancestry-related molecular pathways predicted by genetic SLE risk may help to disentangle the population differences in clinical severity that impact AsA and EA individuals with SLE.

Systemic lupus erythematosus (SLE) (OMIM:152700) is a complex autoimmune disease characterized by clinical and genetic heterogeneity. Individuals of East Asian ancestry (AsA) have a greater prevalence of renal involvement, infections and cardiovascular complications compared to individuals of European ancestry (EA)¹. In particular, lupus nephritis and end stage renal disease (LN/ESRD) are severe complications of SLE that are more prevalent in patients of AsA ancestry than patients of EA ancestry^{2–4}. Whereas some of this variation may be accounted for by confounding environmental and/or socioeconomic factors⁵, it is unclear why AsA ancestry remains associated with clinical severity and sub-phenotypes in SLE.

Immunochip-based and genome-wide association (GWA) studies have revealed important ancestry-specific and trans-ancestral risk associations predisposing to SLE^{6–10}. Recent meta-analyses of European and Chinese

¹AMPEL BioSolutions LLC and the RILITE Research Institute, Charlottesville, VA 22902, USA. ²Department of Biostatistics and Data Science, Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27109, USA. ³Department of Biochemistry, Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27109, USA. ⁴Shanghai Institute of Rheumatology, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. ⁵University of California San Francisco, San Francisco, CA 94117, USA. ✉email: kate.owen@ampelbiosolutions.com

GWAS data suggest that the greater disease burden evident in East Asian populations is at least partially a consequence of different risk variant frequencies⁹. Although these and other studies enable a better understanding of the genetic architecture of SLE, they tend to focus on only the most significant associations and linked genes, and thereby do not capture the totality of genetic variation. Critically, genetic analyses to date have been unable to provide a clear path toward novel therapeutic development. This shortcoming is of particular concern with respect to AsA patients, where the control of disease activity remains suboptimal⁵. Here, we undertook a bioinformatics-driven approach to identify a comprehensive list of ancestry-specific and shared SLE-associated genes, using eQTL mapping, the identification of functional variants in coding regions and variants impacting transcription factor binding site occupancy, as well as SNP-gene proximity. Together, this approach identified 3105 potential SLE-associated genes in one or more ancestral groups (1349 EA, 1429 AsA and 327 trans-ancestral). Connectivity mapping and network analysis were used to identify ancestry-enriched biological pathways and inform ancestry-specific pharmacological targets. This trans-ancestral analysis strategy not only identified additional SLE-associated molecular pathways but, due to the underlying differences between AsA and EA in risk-allele frequencies, may enable a deeper understanding of the differences in the prevalence of SLE risk, severity, and clinical phenotypes. Such an understanding may motivate population-specific clinical trials and interventions.

Results

Identification of ancestry-dependent and independent non-HLA SLE-associated variants and downstream target genes. Despite the success achieved by GWAS in mapping polygenic disease risk loci in SLE, the biological implications of the majority of identified variants has remained unknown. To gain a broader view of how inherited genetic variation impacts disease risk, we took the global approach of integrating SNPs with a range of association significance to generate a cohort of predicted genes that could ultimately be pruned and mapped to functional pathways for analysis. ImmunoChip-based association analyses have identified 700 single-nucleotide polymorphisms (SNPs) reported as significantly associated with SLE in patients of East Asian (AsA) ancestry⁶ and 757 SNPs associated with disease in European (EA) populations (Fig. 1A)⁸. Twenty SNP associations (<1.5%) were shared between ancestries. In both ancestries, approximately 70% of SNPs were found in non-coding regions (intergenic and intronic), and 8% of SNPs were in coding regions

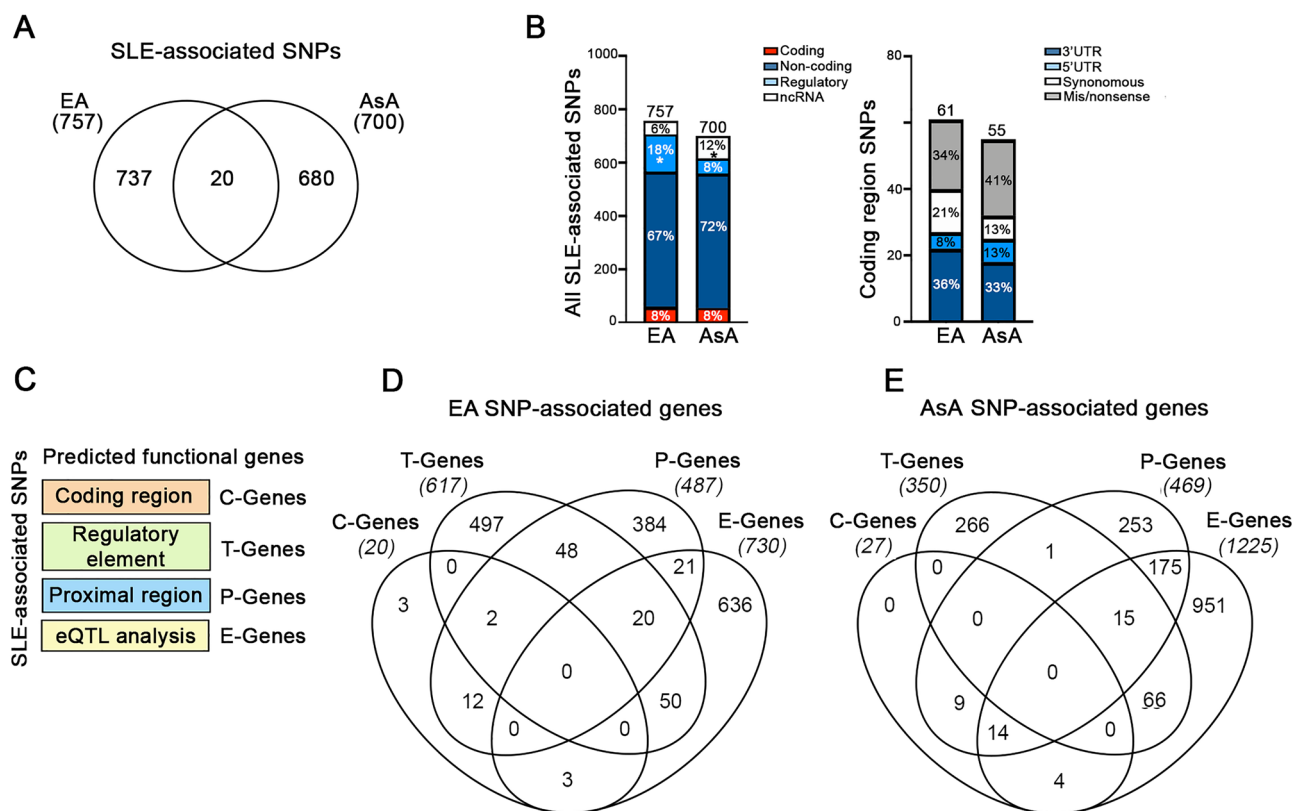


Figure 1. Mapping the functional genes associated with SLE-ImmunoChip SNPs. **(A)** Venn diagram depicting the ancestral overlap of all SLE-associated ImmunoChip SNPs. **(B)** Distribution of genomic functional categories for all EA and AsA non-HLA associated SLE SNPs. Genomic category comparisons between ancestral groups were performed using a 2-proportion z test. P values were 2-tailed, and asterisks indicate a significance threshold of $p < 0.05$. **(C)** Functional SNP-associated genes are derived from 4 sources, including eQTL analysis (E-Genes), regulatory regions (T-Genes), coding regions (C-Genes) and proximal gene-SNP annotation (P-Genes). **(D, E)** Venn diagrams showing the overlap of all EA **(D)** and AsA **(E)** associated E-, T-, C- and P-Genes.

(3'UTRs, 5'UTRs, synonymous and non-synonymous) (Fig. 1B and Supplemental Table S1). AsA populations had a significantly higher percentage of SLE-associated SNPs in non-coding (nc)RNAs (lncRNA and miRNA), whereas EA populations had more SLE-associated SNPs located within regulatory regions, including enhancers, promoters, open chromatin, and transcription factor binding sites (Fig. 1B).

We used a bioinformatics-based approach to identify the most plausible genes affected by the SLE-SNP association. As previously described^{11,12}, we first determined whether there was evidence that the SNP was an expression quantitative trait loci (eQTL) using the GTEx database (version 8) and the Blood eQTL browser¹³. 226 EA- and 405 AsA- SLE-associated eQTLs linked to 730 and 1225 expression genes (E-Genes), respectively (Fig. 1C–E and Supplemental Table S2). eQTLs were identified for nearly 60% of the total number of AsA Immunochip SNPs, compared to 29% of SLE-associated SNPs in the EA data (Supplemental Fig. S1). Whereas the number of AsA eQTLs was higher across all genomic categories (coding, non-coding and regulatory regions as well as ncRNAs), the proportion of AsA SLE-associated SNPs linked to ncRNAs (51%; 42/82) was nearly three times higher than that observed in EA populations (18%; 8/44) (Supplemental Fig. S1 and Supplemental Table S1). Next, we sought to identify SNPs within distal and *cis* regulatory elements (e.g., enhancers and promoters), using the computational tools GeneHancer and HACER (Human Active Enhancers to interpret Regulatory variants), both of which connect regulatory SNPs with downstream target genes (T-Genes)^{14,15}. Together, GeneHancer and HACER identified 105 SLE-associated SNPs (59 EA, 36 AsA) overlapping distal regulatory elements or promoters predicted to impact the expression of 964 T-Genes (617 EA, 350 AsA) (Fig. 1C–E and Supplemental Table S2). For variants located in coding regions, 44 SNPs (21 EA, 23 AsA) were associated with either non-synonymous or nonsense changes in 47 genes (C-Genes; 20 EA, 27 AsA) (Supplemental Table S2). The remaining SNPs that were not linked to E-, T- or C-Genes were assigned to the closest proximal gene (P-Gene), identifying an additional 956 P-Genes (487 EA, 469 AsA) (Supplemental Table S2).

Overlapping EA and AsA SNP-linked E-, T-, C- and P-Genes are depicted in Fig. 1D,E, respectively. No genes were shared within all four groups within either ancestry, and we observed limited commonality between T-, P- and E-Genes, with only 20 genes shared among the three groups in EA and 15 genes shared in AsA. It is notable that of the total of 3,432 SNP-linked genes, < 10% (327) overlapped between AsA and EA lupus cohorts (Fig. 2A).

Characterization of gene signatures. We next completed a series of bioinformatic analyses to determine the overall biological function of the 1349 genes associated with SLE in EA and 1429 genes associated with SLE in AsA, as well as the 327 SLE-associated genes common to both ancestries. Analysis of genes linked through EA revealed enrichment in processes related primarily to adaptive immune function, including the functional category for interferon stimulated genes, canonical pathways for *TH1 and TH2 activation pathway* and *Macrophage classical activation signaling pathway*, and GO terms for the Regulation of B cell proliferation (GO:0030888) and the Regulation of T cell proliferation (GO:0042129). Genes linked to SNPs associated with SLE in the AsA cohort were enriched in categories related to pathogen-influenced signaling, such as *Role of PRRs in the recognition of bacteria and viruses*, and the Positive regulation of lymphocyte differentiation (GO:0045621), as well as those representing more diverse biological functions, such as Regulation of oxidative stress-induced neuron death (GO:1903203) and DNA ligation involved in DNA repair (0051103). Shared genes were distributed in a range of adaptive and innate immune gene categories (Fig. 2B,D,E).

In addition, EA- and AsA-derived gene sets were examined using a clustering program that detects immune and inflammatory cell type signatures within large gene lists to identify dominant immune cell populations driving disease pathology within each ancestry¹⁶. Consistent with our pathway analysis, EA exhibited strong enrichment in cellular categories for myeloid, T, and B cells, whereas SLE-associated genes in AsA were not enriched in any cellular category (Fig. 2C). Independent analysis of shared genes revealed enrichment in the T, B and myeloid, and the NK or T cell categories. Finally, parallel analyses examining P-Genes separately from E-, T-, and C-Genes were conducted to assess the potential overrepresentation of immune-based processes because of the Immunochip design bias¹⁷. As expected, P-Genes (384 EA, 253 AsA) were enriched in immunologically-driven functional categories and pathways; exclusion of P-Genes resulted in only minor alterations to overall categorization in either ancestral background (Supplemental Fig. S2).

Delineation of signaling pathways identified by ancestry-specific SNP-associated genes. To assess ancestry-driven key signaling pathways in greater detail, ancestry-based protein–protein interaction (PPI) networks consisting of EA-associated, AsA-associated, or ancestry-independent genes were constructed using STRING, visualized in Cytoscape and clustered using MCODE (Supplemental Table S3). To provide an additional level of functional annotation, clusters contributing to overall immune function, tissue repair, mechanisms of cellular stress, cell motility, metabolic function or general cell function were grouped together. EA-associated genes were dominated by the functional category for interferon stimulated genes observed in cluster 2 (118 genes) (Fig. 3A), along with multiple canonical pathways related to the activation of pattern recognition receptors and downstream type I interferon signaling (Supplemental Table S4). Cluster 7 revealed additional enrichment in lymphocyte activation and differentiation, such as the *TH1 and TH2 activation pathway* that was also represented in the shared gene network, and cellular enrichment for cells of myeloid and/or lymphoid origin. Notably, the EA-associated network lacked evidence of cell motility and cell stress/injury, whereas metabolic function was represented by clusters 12 and 13 enriched in retinoid X receptor activation (*LXR/RXR activation*, *PPARα/RXRα activation*) involved in the regulation of lipid metabolism, inflammation, and cholesterol bile acid catabolism. Pathways associated with SLE in AsA were indicative of a diverse range of biological processes with protein metabolic functions dominating clusters 2 and 17 (Fig. 3B), whereas clusters 3 and 6 were enriched in multiple canonical pathways related to cytokine production and signaling (Supplemental Table S4). Interestingly, genes linked to SNPs associated in the AsA cohort did not include a unique interferon signature, but instead

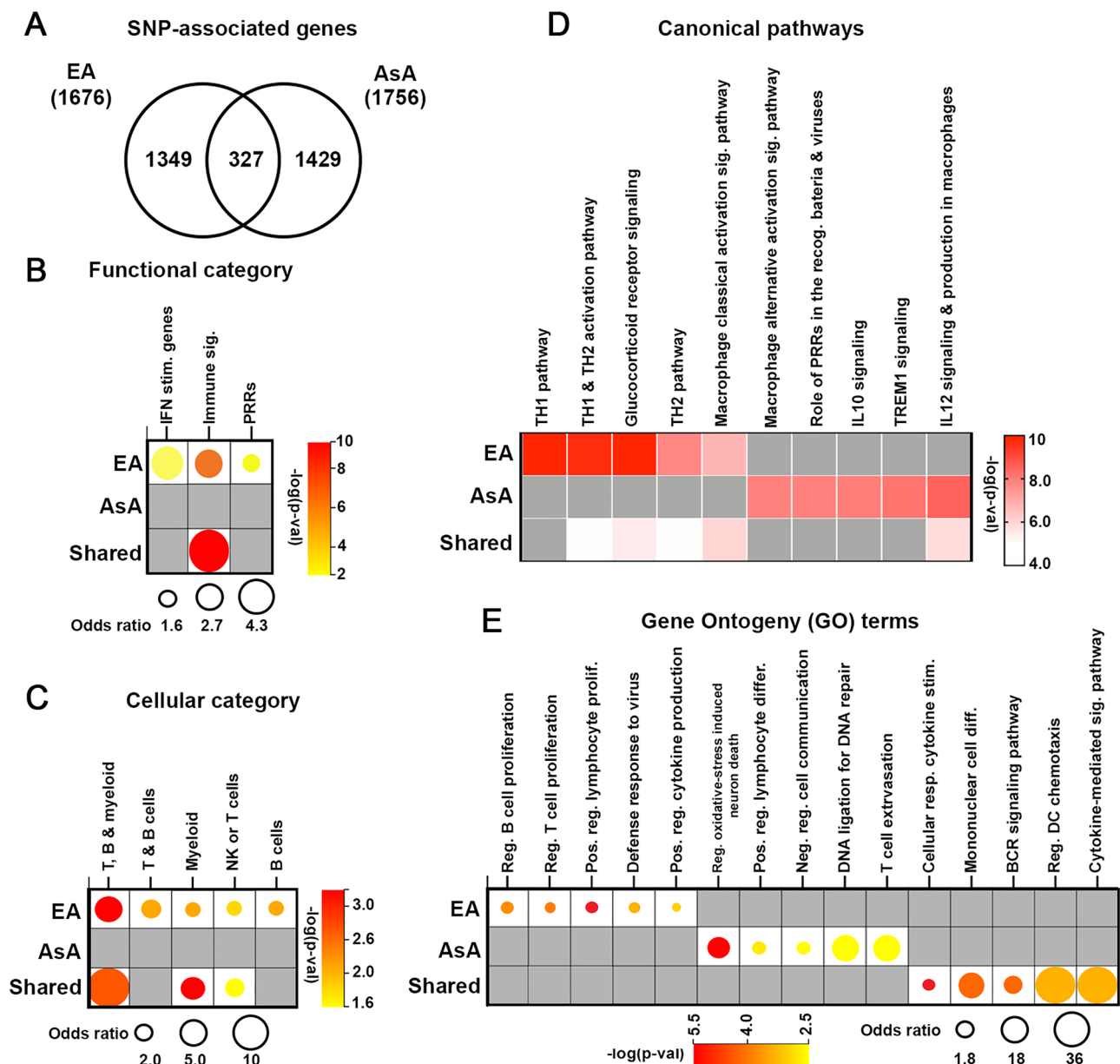


Figure 2. Functional characterization of SNP-associated genes. **(A)** Venn diagram depicting the overlap between all EA- and AsA-SNP associated genes. **(B, C)** Bubble plots depict ancestry-dependent and independent SNP-associated genes analyzed to determine enrichment using functional definitions from the BIG-C (Biologically Informed Gene Clustering) annotation library and I-Scope for hematopoietic cell enrichment. Enrichment was defined as any category with an odds ratio (OR) > 1 and a $-\log(p\text{-value}) > 1.33$. **(D)** Heatmap (generated by GraphPad Prism 8.3; www.graphpad.com) visualization of the top five significant IPA canonical pathways and **(E)** bubble plot showing gene ontology (GO) terms for each gene list organized by ancestry. Top pathways with OR > 1 and $-\log(p\text{-value}) > 1.33$ are listed.

coalesced into multiple small clusters related to mitochondrial dysfunction (clusters 9 and 19) and metabolism, evident in clusters 16, 22 and 30. Additionally, AsA-associated gene clusters were enriched in chromatin remodeling found in cluster 1, along with evidence of cell motility (clusters 11, 12, 23 and 25). AsA cellular enrichment was dominated by monocytes and myeloid lineage cells.

Pathways exemplified by SLE-linked genes in both EA and AsA appear to be a blend of the pathways enriched within each ancestry. Common pathways included *Interferon signaling*, *TH1 and TH2 activation pathway*, *Complement system* and *Leukocyte extravasation* (Fig. 3C and Supplemental Table S5). Figure 3D depicts a selection of both the unique and overlapping canonical pathways motivated by the EA-associated and AsA-associated gene sets. We also carried out a parallel series of bioinformatics analyses to determine the biological function of the full array of EA (1676) and AsA (1756) SNP-predicted genes, including those associated with both ancestries (Supplemental Fig. S3 and Supplemental Table S5). As expected, shared genes were evenly distributed throughout each large network and subsequent connectivity mapping revealed the addition of several new clusters to both

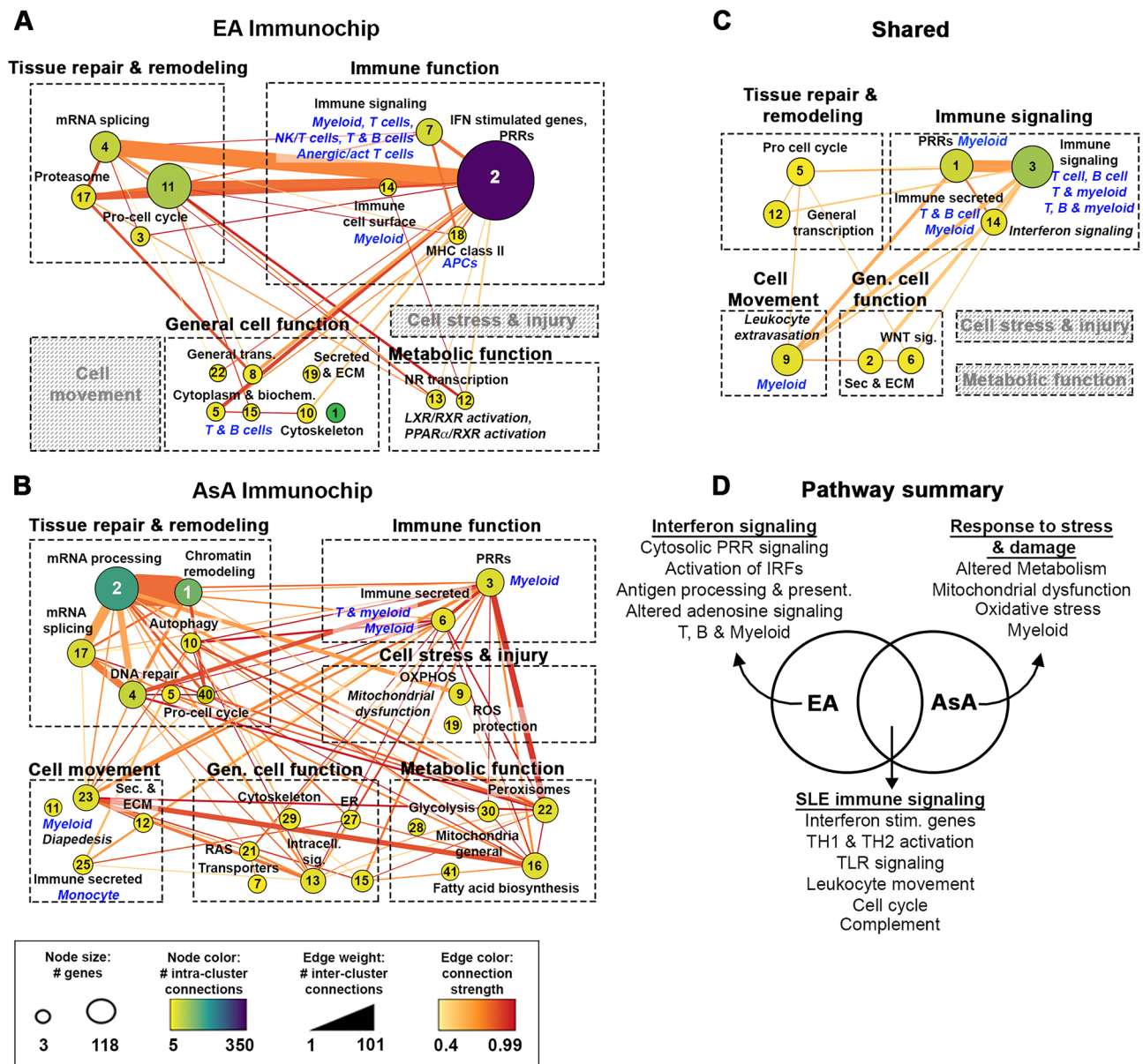


Figure 3. Key pathways determined by EA and AsA-associated genes. Cluster metastructures for EA (A), AsA (B) and the shared gene cohort (C) were generated based on PPI networks, clustered using MCODE and visualized in Cytoscape. Cluster size indicates the number of genes per cluster, edge weight indicates the number of inter-cluster connections and color indicates the number of intra-cluster connections. Enrichment for each cluster was determined by BIG-C and IPA; clusters were then grouped and categorized according to overall function (immune, tissue repair, metabolic, motility or general). Grey boxes indicate categories lacking relevant clusters. (D) Venn diagram showing the number of overlapping pathways motivated by EA or AsA predicted genes. Representative pathways are listed.

the EA and AsA networks. For example, the full EA network gained several clusters contributing to cell motility enriched in integrin signaling and granulocyte diapedesis (clusters 34 and 35), whereas the enlarged AsA network gained multiple clusters enriched in immune function (clusters 9, 12 and 31) and interferon signaling (cluster 3), as well as enrichment in a more diverse array of cell types, including T and B cells, neutrophils and NK/T cells. Both networks acquired a small cluster enriched in the functional category for reactive oxygen species (ROS) protection (EA cluster 22, AsA cluster 24) driven by the glutathione peroxidases *GPX4*, an antioxidant enzyme involved in ferroptosis¹⁸, and *GPX3* an ROS scavenger. Overall, the addition of the shared gene cohort to each network highlights many common pathways and biological functions, but still revealed ancestry-driven differences. Taken together, unique pathways identified via EA analyses appear dominated by immune-related signaling as well as T, B and myeloid cells, whereas those in AsA analyses are dominated by biological processes related to altered metabolism and mitochondrial dysfunction.

Validation of AsA-enriched molecular pathways using summary GWAS data. To test our pathway predictions, we combined summary data from previously reported GWA studies^{7,9} that identified 1350 SNPs associated with SLE in patients of East AsA ancestry (Supplemental Table S6). Of these SNPs, 68% were located in non-coding regions, 6.5% were in coding regions, 2.7% were in regulatory regions and 22% were located within or proximal to non-coding RNAs (Supplemental Fig. S4). Validation AsA-associated GWAS SNPs exhibited limited commonality when compared to Immunochip SNPs, with <1% of either EA- or AsA-associated Immunochip SNPs overlapping GWAS SNPs, and only 3 SNPs common to all 3 datasets (Supplemental Fig. S4). We next applied our same bioinformatics-driven methodology to generate a validation gene cohort composed of 1321 E-Genes, 307 T-Genes, 17 C-Genes and 974 P-Genes (Supplemental Table S7). Connectivity mapping of all validation genes were used to create PPI networks that were clustered as described above (Fig. 4A). Examination of each cluster revealed functional similarity to those derived from AsA Immunochip-associated genes. For example, clusters 1, 3, 4, 5 and 6 share hallmarks of tissue repair and remodeling exemplified by categories for mRNA processing, pro-cell cycle and protein degradation (proteasome, lysosome, endocytosis). Additionally, we observed smaller clusters (21, 27 and 28) representative of processes involved in metabolic function, and clusters (13, 18 and 24) characteristic of cell stress and injury, including the *Inhibition of ARE-mediated degradation pathway* and *Mitochondrial dysfunction* canonical pathways (Supplemental Table S8). Cluster 9 contained a small interferon-stimulated gene signature consisting of *IFI27*, *IFI44* and *RSAD2* (Supplemental Table S3). Cellular categories were again dominated by monocytes, T cells, NK cells, B cells and plasmacytoid (p)DCs and are consistent with findings observed with AsA Immunochip-associated genes. In contrast to AsA-associated genes, where we observed large, highly connected clusters, an equivalent cohort of apparently random genes generally formed smaller clusters, exhibited fewer intra- and inter-cluster connections, and were primarily enriched in functional categories mostly related to basic cell function (general cell surface, secreted and ECM) (Fig. 4B,C). As shown in Fig. 4D, which displays the number of genes (and percentage of total genes) assigned to each functional category, random genes are skewed toward general cell function, whereas AsA-associated genes are more prevalent in the overall immune (15.3% of genes), tissue repair (53.4%) and cell stress (7%) categories. The random gene network also lacked evidence of cell movement and the diversity of cellular enrichment identified from AsA SNP-associated genes (Fig. 4B).

Validation of AsA-enriched molecular pathways using gene expression data. SNP-based pathway predictions were also tested in differential gene expression analyses from whole blood samples collected from AsA SLE patients with active disease and renal involvement compared with healthy controls (E-MTAB-11191, Supplementary Table S11). Overall, differential expression (DE) analysis revealed 5886 DE genes (DEGs) enriched in functional categories for interferon stimulated genes, gene expression, RNA processing and metabolism (Fig. 4 and Supplemental Fig. S5). A total of 685 AsA and 300 EA SNP-predicted genes were shared with AsA SLE DEGs, and 144 genes, representative of type I and type II interferon signaling, were shared among all three groups (Supplemental Fig. S6). Genes common to AsA DEGs and AsA SNP-predicted genes were enriched in RNA processing and translation, whereas DEGs shared with EA SNP-predicted genes were specifically enriched in type I interferon/cytokine signaling. Connectivity mapping and pathway analysis of AsA DEGs again revealed striking commonality with AsA SNP-associated genes from both the Immunochip and summary GWAS, exemplified by cluster enrichment in RNA splicing/processing, ubiquitylation, chromatin remodeling and metabolic function (glycolysis, TCA cycle, fatty acid biosynthesis and peroxisomes), and canonical pathways for *Spliceosome cycle* (cluster 2), *Inhibition of ARE-mediated RNA degradation* (clusters 12, 20 and 25), and *Mitochondrial dysfunction* (cluster 16) (Fig. 5A and Supplemental Table S9). The distribution of DE genes in AsA for each overall category were also similar to that observed with AsA-associated genes (Immunochip and GWAS) but differed from both the EA-Immunochip predicted or random genes (Fig. 5B).

To test the pathway predictions, Gene Set Variation Analysis (GSVA)¹⁹ was applied to determine the relative enrichment of gene signatures identified in peripheral blood mononuclear cell (PBMC) samples from SLE patients (EA and AsA) and controls (Supplemental Table S10). In FDAPBMC1, a dataset composed of EA patients (Supplemental Table S11), all 7 IFN gene signatures (IGS) and signatures for the RIG-I pathway and DNA/RNA sensors were strongly enriched in SLE PBMCs compared to controls (Fig. 6A). In contrast, only the signatures for IFNA2, IFNB1, IFNW1 and the Type I core were enriched in SLE PBMCs from AsA patients in GSE81622 (Fig. 6B). GSVA using a random group of genes did not separate SLE from controls in either dataset.

GSVA enrichment scores using signatures for complement activation and metabolic pathways, including mitochondrial oxidative phosphorylation (OXPHOS), the TCA cycle and glycolysis were able to separate AsA SLE patients, but not EA patients, from healthy controls (Fig. 6C,D). Consistent with our predictions, AsA SLE patients exhibited significant enrichment in signatures for mitochondrial dysfunction and oxidative stress. As shown in Fig. 6E–H, enrichment of a number of pathways and cell types in both EA and AsA SLE cohorts were noted, including TLR signaling, inflammasome, TNF signaling and monocyte/myeloid lineage cells, whereas AsA patients exhibited additional enrichment in pro-cell cycle, low density granulocytes (LDGs) and B cells. Varying degrees of T cell/NK cell lymphopenia were evident in both ancestral populations.

Differential Inflammatory cell metabolism and activation status by ancestry. Because dysfunctional metabolic reprogramming can directly influence and exacerbate defective immune responses, we carried out linear regression analysis between the GSVA scores for individual cell signatures and the glycolysis and oxidative phosphorylation gene signatures to interrogate the metabolic status of immune cells from EA and AsA PBMC datasets. Enrichment scores for the glycolysis gene signature in AsA SLE patients exhibited positive correlation with both the monocyte/myeloid ($R^2 = 0.14$, $p = 0.03$) and B cell signatures ($R^2 = 0.12$, $p = 0.051$) (Fig. 7A). In contrast, glycolysis was not associated with gene signatures representing monocyte/myeloid cells,

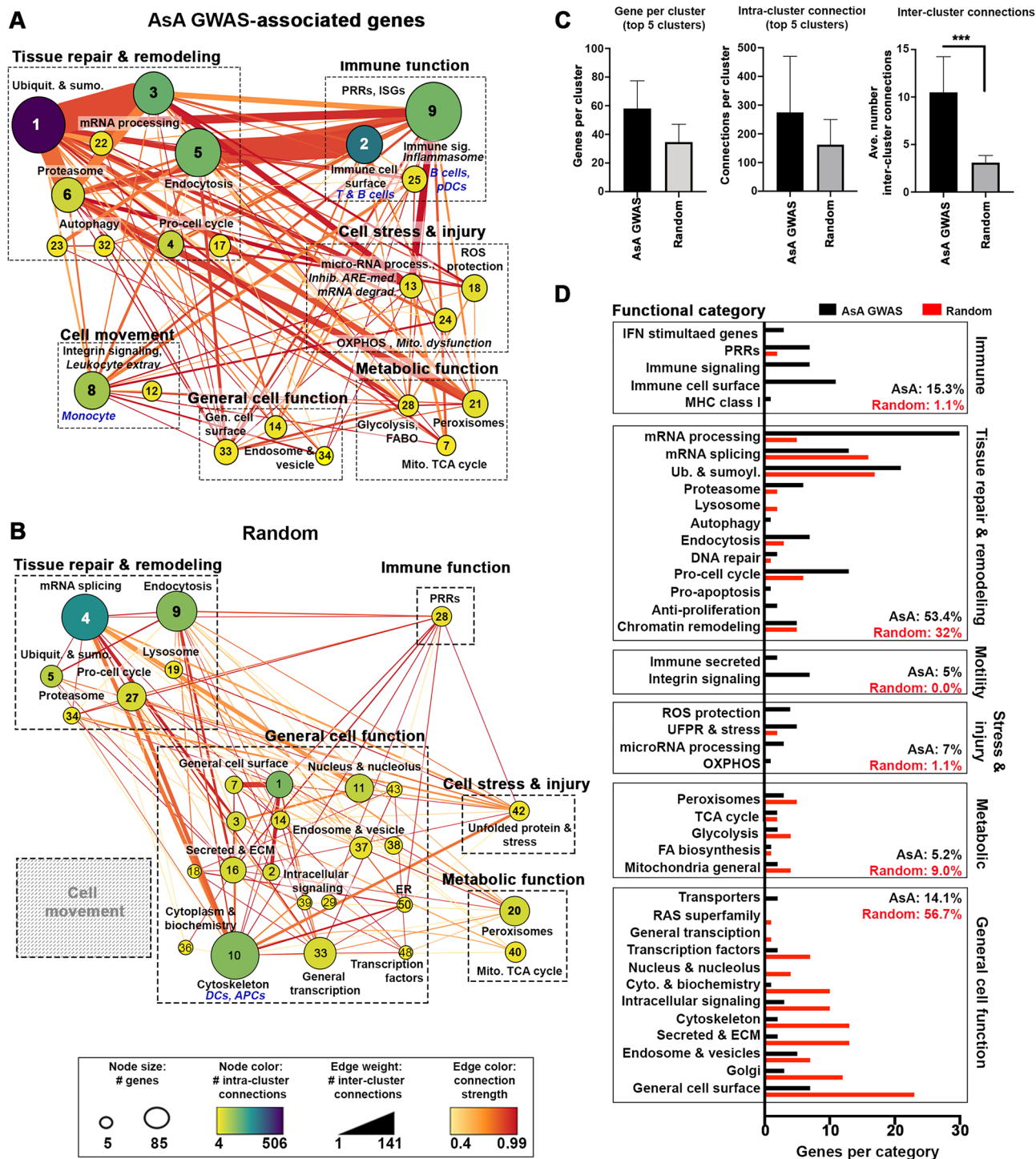


Figure 4. AsA ImmunoChip-based pathways are supported by summary GWAS from AsA SLE patients. Using SNP-predicted genes from the AsA GWAS validation SNP-set (A) or an equivalently sized cohort of random genes (B) metastructures were generated based on PPI networks, clustered using MCODE and visualized in Cytoscape. Cluster size indicates the number of genes per cluster, edge weight indicates the number of inter-cluster connections and color indicates the number of intra-cluster connections. Enrichment for each cluster was determined by BIG-C and IPA; clusters were then grouped and categorized according to overall function (immune, tissue repair, metabolic, motility or general). Grey boxes indicate categories lacking relevant clusters. (C) Quantitation of cluster size, intra-cluster connections and inter-cluster connections network is displayed. Error bars represent the 95% confidence interval; asterisks (***) indicate a p-value < 0.001 using Welch's t-test. (D) Quantitation of AsA GWAS (black bars) and random (red) genes falling into each BIG-C category and grouped by overall functionality.

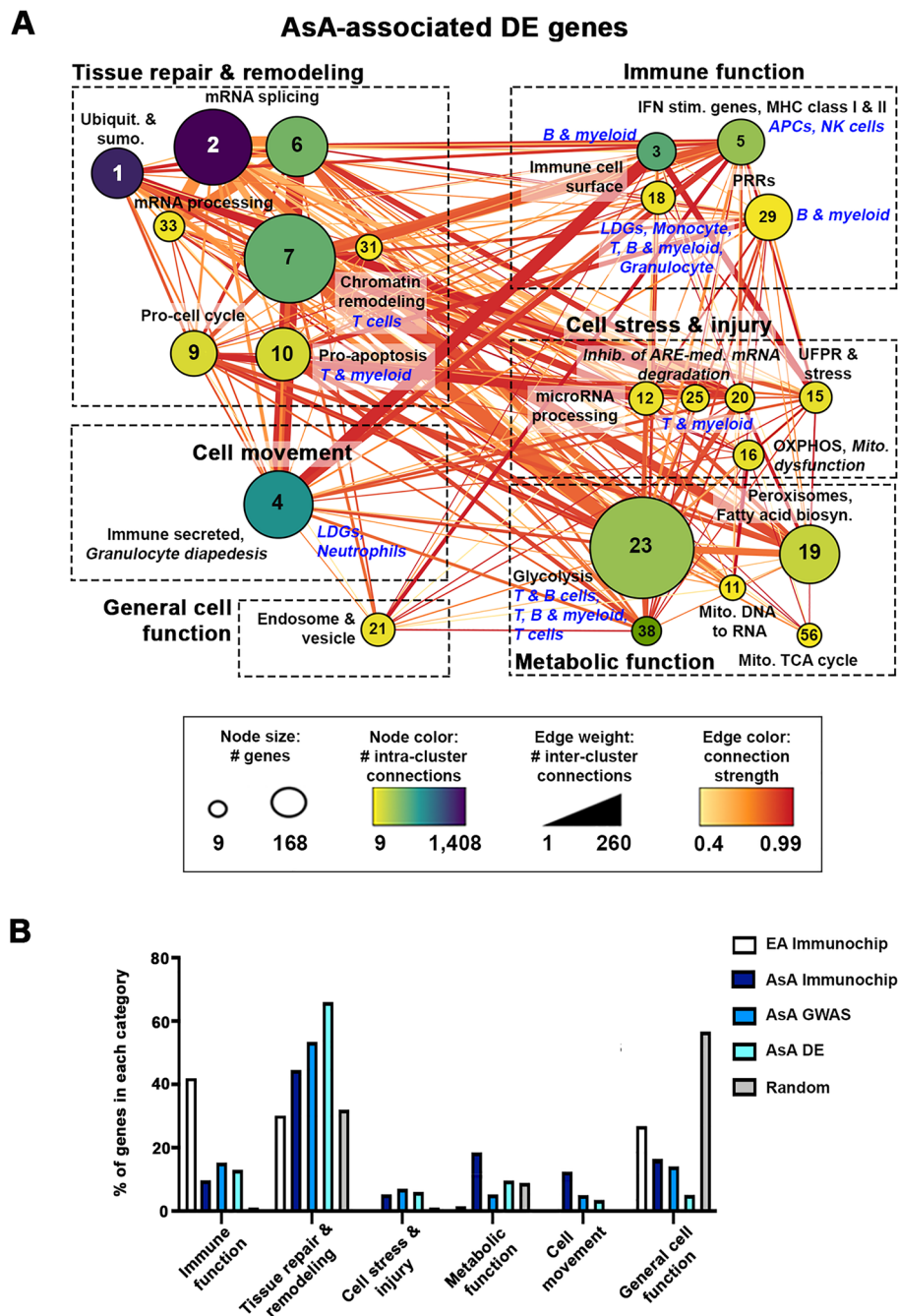


Figure 5. Asian-associated pathways are validated with gene expression data from AsA SLE patients. (A) Using differentially expressed (DE) genes from AsA whole blood samples (E-MTAB-11191), metastructures were generated based on PPI networks, clustered using MCODE and visualized in Cytoscape. Cluster size indicates the number of genes per cluster, edge weight indicates the number of inter-cluster connections and color indicates the number of intra-cluster connections. Enrichment for each cluster was determined by BIG-C and IPA; clusters were then grouped and categorized according to overall function (immune, tissue repair, metabolic, motility [Continued from previous page] or general cell function). (B) Bar graph showing the percent of associated (EA/AsA immunochip and AsA GWAS), differentially expressed and random genes falling into each overall functional category.

T cells or NK cells from EA patients, whereas the B cell category exhibited significant negative correlation ($R^2=0.26$, $p=0.03$). The gene signature for oxidative phosphorylation lacked positive correlation with all cell types in AsA patients, but did exhibit significant correlation ($R^2=0.6$, $p=0.0003$) with T cells in individuals of EA ancestry (Fig. 7A).

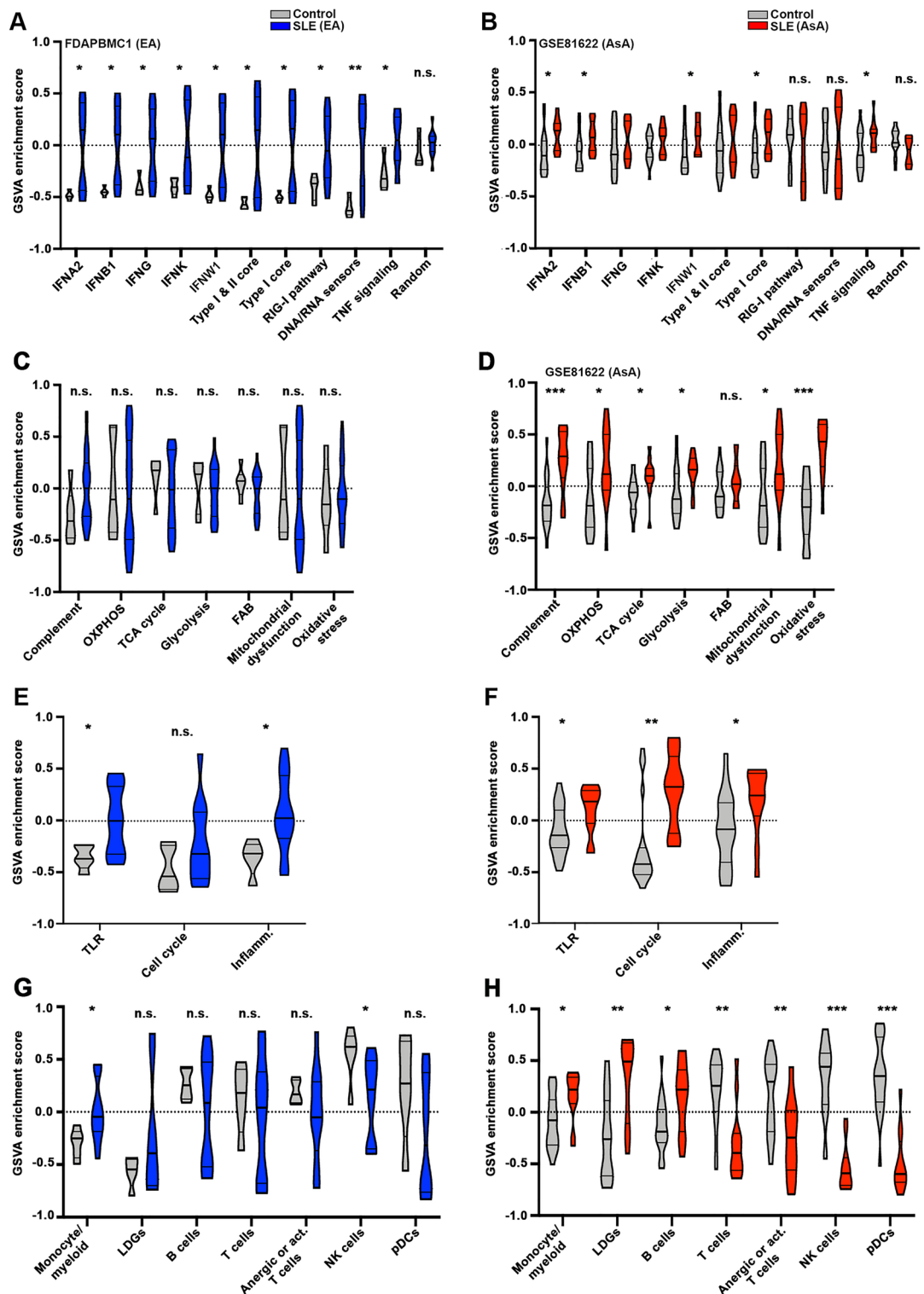


Figure 6. SNP-associated pathways inform gene signatures for GSVA analysis in patient PBMC datasets. GSVA enrichment scores were generated for PBMCs in EA and AsA SLE patients and healthy controls from FDAPBMC1 (EA-only patients and controls) and GSE81622 (AsA-only patients and controls). GSVA scores for type I and type II interferon-based gene signatures (A, B), metabolic gene signatures (C, D), cellular processes (E, F) and individual cell type signatures (G, H) are shown. Asterisks (*) indicate a p-value < 0.05 using Welch's t-test comparing SLE to control.

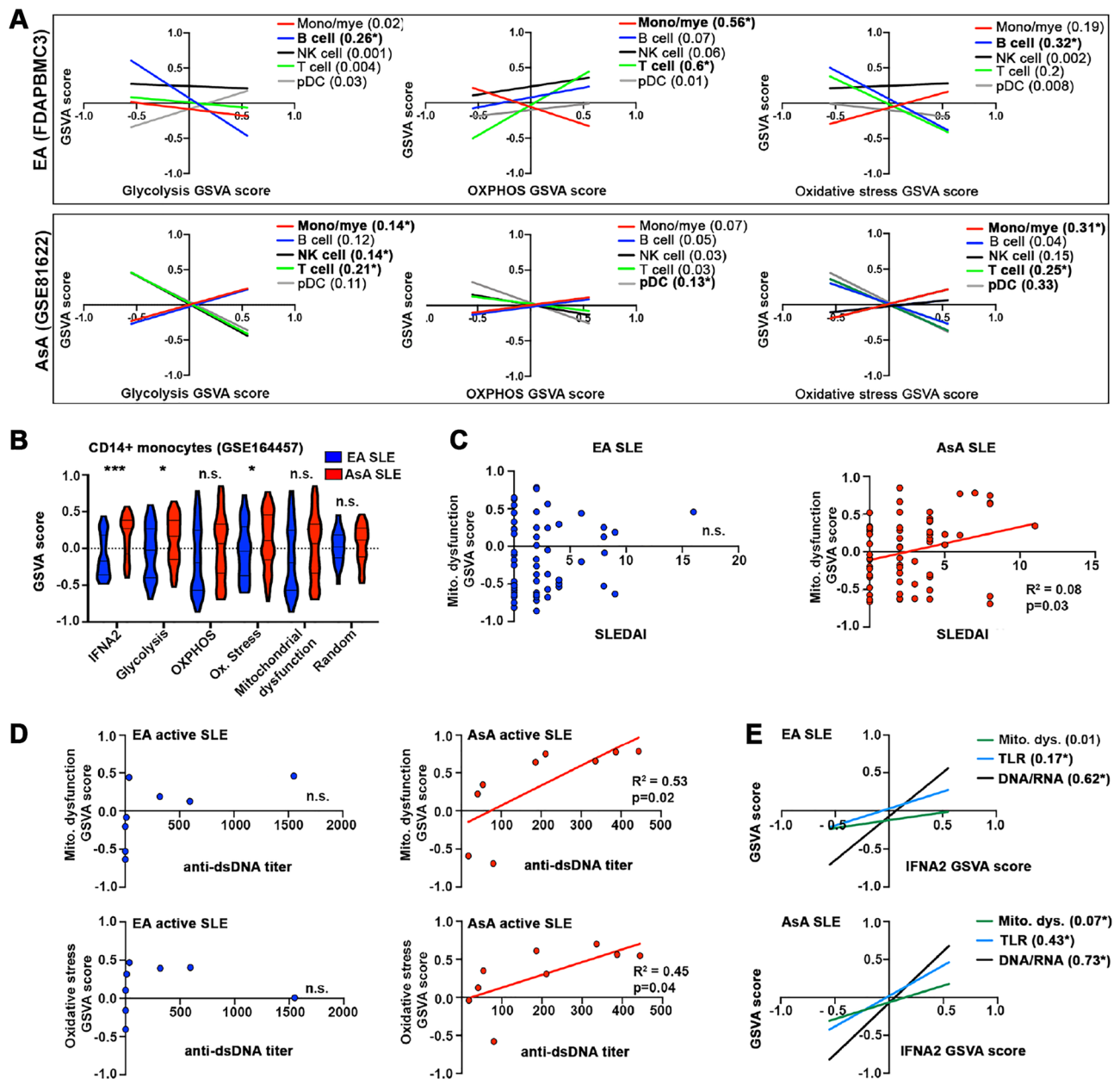


Figure 7. Linear regression to examine the relationship between cell types, biological processes and inflammatory cytokines. **(A)** Linear regression analysis showing the relationship between GSVA scores for glycolysis, oxidative phosphorylation or oxidative stress and individual cell types (pDCs, monocyte/myeloid, B cells, T cells and NK cells) for FDAPBMC1 (EA, upper panels) and GSE81622 (AsA, lower panels). In **(B)**, GSVA enrichment scores for the indicated cellular processes were generated for purified CD14+ monocytes from EA and AsA SLE patients (GSE164457). Using GSE164457, linear regression was used to examine the relationship between cellular processes and SLEDAI **(C)**, anti-dsDNA titers in active patients (SLEDAI ≥ 6) **(D)** and GSVA scores for IFNA2 **(E)**. Categories with linear regression p values < 0.05 are in bold; R² predictive values are listed after the GSVA enrichment category. *Asterisks indicate significant relationship between functional categories. N.s., not significant.

Since altered immune cell metabolism is associated with heightened oxidative cell stress responses in SLE²⁰, we used the same linear regression approach to test the relationship between different immune cell types and GSVA enrichment scores for oxidative stress (Fig. 7A). Monocyte/myeloid cells from AsA patients were the only cell type demonstrating a significant positive relationship with the gene signature for oxidative stress (R² = 0.3, p = 0.0008). These results were confirmed in purified CD14+ monocytes isolated from EA and AsA SLE patients (California Lupus Epidemiology Study, CLUES; GSE164457²¹) (Supplemental Table S11) showing significant enrichment of gene signatures for oxidative stress, glycolysis and IFNA2, along with a trend toward elevated OXPHOS and mitochondrial dysfunction, specifically in AsA SLE patients (Fig. 7B). Given that enhanced oxidative stress and concomitant mitochondrial dysfunction have been shown to correlate with disease activity, DNA

damage, and autoantibody production^{22–24}, we next examined the potential link between enrichment scores for mitochondrial dysfunction and SLE disease activity index (SLEDAI) score in these patients. GSVA scores for mitochondrial dysfunction demonstrated a significant positive relationship with SLEDAI in AsA but not EA SLE monocytes (Fig. 7C). Similarly, enrichment scores for oxidative stress and mitochondrial dysfunction showed a significant positive correlation with anti-dsDNA titers among AsA SLE patients with active disease (SLEDAI ≥ 6) compared to EA patients (Fig. 7D). In addition, overall complement C3 levels were lower in AsA patients and demonstrated a significant negative correlation with both anti-dsDNA and SLEDAI (Supplemental Fig. S7) in accordance with the observation that complement depletion and anti-dsDNA antibodies are often associated with elevated disease activity²⁵.

Finally, given that mitochondrial dysfunction and nucleic acid sensing are potent inducers of interferons²⁶, we tested for a relationship between IFNA2 enrichment scores and enrichment scores for the mitochondrial dysfunction, DNA/RNA sensors and TLR pathways (Fig. 7E). In EA patients, GSVA scores for the DNA/RNA sensors ($R^2 = 0.62$, $p < 0.0001$) and TLR pathway ($R^2 = 0.17$, $p = 0.0013$) signatures, but not mitochondrial dysfunction, were associated with IFNA2. In contrast, all three signatures exhibited a positive, significant correlation with IFNA2 in AsA patients.

Discussion

SLE is a multisystem autoimmune disorder with a strong genetic contribution. The incidence of SLE varies widely across populations, with individuals of Asian, Hispanic and African ancestry demonstrating a three- to four-fold increase in disease prevalence compared to their European counterparts²⁷. The advent of candidate gene, Immunochip and genome wide association studies (GWAS) has transformed our understanding of SLE genetics. However, it remains unclear how genetic ancestry contributes to the incidence, clinical heterogeneity and variation in disease outcomes among SLE patients. Specifically, AsA patients develop SLE at a younger age and with more severe manifestations, including lupus nephritis (LN)²⁸. Whereas increased genetic risk burden in AsA individuals has been hypothesized to account for the increased prevalence of SLE in this population⁹, it does not provide adequate explanation for accelerated disease progression, variation in treatment response, or more extensive organ damage, especially with regard to the development of LN. Our observations suggest that, in addition to higher risk load, underlying differences at the genetic level may significantly influence the dominant biological pathways operative within each ancestry. Here, we show that by identifying a comprehensive list of genes implicated by GWAS and linking them to biologic pathways, we can provide a broader perspective on the genetic influences of SLE. It may also inform on health disparities particular to specific ancestries.

To accomplish this, we adopted a similar strategy to that recently described¹¹, employing statistical and computational analyses along with data acquired from functional genomic assays to map the overall gene expression landscape of SLE and further define the disease-associated pathways. It is important to highlight that the SNPs examined here are generally present in both EA and AsA populations, but differences in allele frequencies suggest some SNP-predicted genes and pathways may be more influential in one ancestry over another. Rather than focusing on only a very limited number of top associations, this approach used a less stringent significance threshold to allow for the integration of more SNPs that could ultimately be pruned and mapped to a larger cohort of genes and pathways. Using SNPs discovered via Immunochip-based association studies, eQTL analysis identified 631 SNPs associated with 1955 E-Genes (730 EA, 1225 AsA). In the Asian ancestry studies, nearly 60% of SNPs were eQTLs, compared to 29% in EA. While eQTLs represented a higher proportion of SNPs associated with SLE in AsA across all genomic functional categories, it was of note that eQTLs linked to ncRNAs were nearly 3 times as frequent in AsA compared to EA patients. The disparity in distribution likely represents the heterogeneous genetic disposition uniquely affecting EA and AsA SLE patients. ncRNAs are a class of mRNA-like transcripts, typically > 200 nucleotides in length, that lack protein coding potential and serve as important regulators of gene expression by actions at the transcriptional, post-transcriptional and post-translational levels²⁹. Several ncRNA eQTLs identified here were associated with anti-sense RNA E-Genes, including *IFNG-AS1* and *IL12A-AS1*, both of which are involved in the regulation of their cognate sense protein-coding genes^{30,31}. Increasing evidence points to an important role for ncRNAs in the differentiation, polarization and activation of both myeloid and lymphoid lineage immune cell types³². Furthermore, abnormal ncRNA expression is associated with mitochondrial dysfunction-induced oxidative stress in a number of pathological conditions, including SLE^{33–35}.

Overall, genes linked to SNPs associated with SLE in AsA cohorts were enriched in processes related to leukocyte migration, PRR signaling and RNA processing, and further detail provided by protein–protein interaction network and pathway analysis revealed multiple clusters enriched in translation/RNA processing, metabolic function, chromatin remodeling, cell stress and mitochondrial dysfunction. In contrast, these pathways, particularly mitochondrial dysfunction and cellular stress responses, were absent from the network analysis of EA SNP-associated genes. Instead, SLE-associated genes in EA data tended to be heavily influenced by immune processes, including the *Role of RIG-I in antiviral innate immunity*, *Antigen presentation*, and the *SLE in T cell signaling pathway*, as well as the functional category for interferon stimulated genes. Cellular enrichment categories were overwhelmingly dominated by T cells, B cells and myeloid cells, and is consistent with previous findings showing increased myeloid/monocyte gene signatures in EA ancestry independent of medication usage (i.e. SLE standard of care drugs including) and autoantibody production³⁶. SNP-predicted pathways common to both EA and AsA were also enriched for myeloid and lymphoid lineage signaling (i.e. *TH1/TH2 activation signaling* and *IL12 signaling and production in macrophages*), along with the well described role of interferons in SLE.

To test our Immunochip SNP-predicted pathway-based findings, we turned to summary GWAS data from AsA SLE patients to create a second “validation” cohort of genes. While little numerical overlap between Immunochip SNPs and those derived from the AsA validation meta-analysis GWAS was observed, network-based analysis of SNP-predicted genes demonstrated a striking resemblance between pathways predicted by the Immunochip and

those from the validation gene set, with multiple large clusters enriched in mRNA processing, degradation and nucleic acid sensing, along with smaller clusters enriched in metabolic activity and mitochondrial dysfunction. Conversely, network generation from an equivalently-sized random gene cohort contained smaller clusters, fewer intra- and inter cluster connections and exhibited little functional similarity with pathways predicted by AsA- or EA-associated genes. Nonetheless, consistent with our AsA Immunochip findings, SNPs within or proximal to ncRNAs were also highly prevalent in the validation dataset, accounting for 22% (299/1350) of total SNPs.

Recently, Wang and colleagues used ancestry-dependent and trans-ancestry meta-analyses to identify 38 novel non-HLA SLE loci³⁷. Six disease variants were associated with SLE exclusively in East Asian populations of which 5 (*HIP1*, *TNFRSF13B*, *PRKCB*, *DSE* and *PLD4*) were linked to genes also identified here. Two of these genes are involved in antibody production, including *TNFRSF13B* that encodes the receptor for BAFF and plays a critical role in B cell development and survival, and *PRKCB*, a protein kinase C family member that regulates B cell activation via BCR-induced NF- κ B activation³⁸. These data suggest that different mechanisms may exist for antibody regulation between AsA and EA populations, whereby pathogenic antibody levels observed more frequently in non-Europeans may contribute to the higher prevalence of SLE in these populations. In fact, data presented here showing enrichment of B cell signatures and positive correlation between B cells and glycolysis may be indicative of increased B cell activation specifically in SLE patients of Asian ancestry. Nonetheless, by integrating all SLE SNP association-predicted genes into functional pathways, we have been able to detect additional differences contributed by ancestry such as those related to stress responses and metabolic dysfunction.

Pathway observations based on genetic findings were also tested in differential gene expression data. The biological pathways determined by AsA-associated DEGs from whole blood samples exhibited impressive similarity to those coupled to AsA SNP-associated genes and included evidence of metabolic and mitochondrial dysfunction, oxidative stress, cell death and gene expression/chromatin remodeling related processes. Furthermore, these results were validated via GSVA analysis in a second expression dataset from PBMCs showing that gene signatures for these processes were specifically enriched in samples from AsA, but not EA SLE patients. Despite the differences in EA and AsA revealed by network pathway analysis, SNP-predicted cellular enrichment common to both ancestral backgrounds included a wide variety of immune cell types, including APCs, T cells, B cell, myeloid, monocytes, LDGs and neutrophils. Additionally, while B cells and LDGs were enriched specifically in AsA SLE patients, the GSVA-based expression pattern demonstrating enrichment in monocyte/myeloid cells and reduced expression of T/NK and APCs was similar between EA and AsA SLE patients and is consistent with the involvement of these cell types in lupus pathogenesis.

Studies examining whether ancestry affects the clinical phenotype of SLE are complicated by the overwhelming heterogeneity of disease manifestations, especially with respect to organ involvement. Nonetheless, many genetic polymorphisms have been associated with specific phenotypes, autoantibody profiles, and/or clinical outcomes in SLE. For example, Fc γ receptor subtypes, such as FCGR3B have been significantly associated with LN and the presence of pathogenic autoantibodies, although it remains unclear whether there is a genetic basis for end-organ involvement based on ancestry³⁹. FCGR3B is almost exclusively expressed by neutrophils and low copy number is associated with glomerulonephritis³⁹. The SNP-predicted pathways described here suggest the presence of different biological mechanisms driving SLE. Importantly, we observed differential enrichment of these pathways in EA and AsA SLE data and thus these pathways may help explain some of the heterogeneity in SLE prevalence and severity across ancestral populations. This is not without precedent, as SNP-predicted genes from African Ancestry (AA) SLE patients identified pathways related to B cell signaling¹¹ and these findings are consistent with reports showing that AA patients experience increased B cell activation and plasma cells compared to their EA counterparts^{36,40,41}. Dominant pathways identified by AsA-associated SNPs and verified by gene expression data include those linked to mitochondrial and/or metabolic dysfunction and subsequent oxidative injury. Mitochondrial components can directly stimulate immune receptors by acting as damage-associated molecular patterns (DAMPs)⁴². They also function as an important source for reactive oxygen species (ROS), as well as facilitate inflammasome activation⁴². In SLE patients, defective mitochondrial function can increase oxidative stresses characterized by increased lipid peroxidation, elevated ROS production and decreased levels of antioxidant enzymes, such as superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPx)²². In fact, enhanced mitochondrial ROS synthesis in SLE has been shown to directly contribute to type I IFN responses⁴³, findings that are consistent with our results showing a significant correlation between the gene signatures for mitochondrial dysfunction and IFNA2 driven by the AsA analyses.

Importantly, mitochondrial dysfunction and associated oxidative damage has also been shown to be a leading factor in the development of chronic and acute renal diseases⁴⁴. Lupus nephritis (LN) is a major risk factor for overall morbidity and mortality in SLE. Previous work has shown that SLE patients of Asian descent are at significantly higher risk for the development of LN⁴⁵, whereas European genetic ancestry was found to be protective against renal disease⁴⁶. In accordance with these findings, enrichment scores for mitochondrial dysfunction and oxidative stress significantly correlated with anti-dsDNA titers in AsA SLE patients with active disease compared to EA patients. These observations are intriguing in light of recent findings by Tsai and colleagues indicating cross-talk between oxidative stress and ncRNAs can trigger autoimmune reactions and perpetuate tissue damage in SLE patients³³. Although the reason for the disparity in renal involvement between ancestral backgrounds has not been fully delineated, we could posit that differential expression of ncRNAs in kidney tissues may contribute to the significant immune dysregulation affecting Asian SLE patients. Whether this might be related to expression of specific regulatory ncRNAs or is a consequence of overall ncRNA burden will require further examination. However, in support of the former, several groups have reported aberrant cell type specific activation linked to the altered expression of non-coding microRNAs, including miR-31, miR145 and miR224 involved in T cell activation, that may be participating in LN pathophysiology⁴⁷. Although these and other studies examining ncRNAs were conducted using samples from patients primarily of Asian ancestry, their results do not obviate the possibility that similar effects occur in EA SLE patients. This highlights a major limitation of

the current study: there is no direct evidence that Asian patients experience a higher degree of mitochondrial dysfunction and oxidative stress-induced tissue damage compared to other ancestral backgrounds. As such, the relationship between altered ncRNA expression, immune dysfunction and tissue injury remains unclear. Nonetheless, genetic and gene expression evidence provided here suggests that AsA ancestral populations may be predisposed to altered cell stress responses where excessive mitochondrial oxidative stress derived ROS and RNS may trigger autoimmune reactivity, increasing cell death and promoting destructive inflammatory reactions in susceptible individuals.

Our genetic and gene expression analyses suggest that metabolic dysfunction is a key feature more prevalent in individuals of Asian compared to European ancestry. Reprogramming of immune cell metabolism is required to sustain the energy demands of effector functions, such as differentiation, clonal expansion, secretion of proinflammatory mediators, phagocytosis, and chemotaxis⁴⁸. Metabolic dysfunction is common in kidney disease and recent work by our group has demonstrated that altered metabolic function in lupus-affected tissues (kidneys and skin) reflect damage induced by myeloid cell infiltration¹⁶. In myeloid lineage cells (monocyte/macrophages), enhanced glucose metabolism, either via glycolysis (characteristic of M1 macrophages) or OXPHOS (characteristic of M2 macrophages) is essential for cell survival, proliferation and to sustain various effector responses⁴⁹. Regression analysis using PBMC and purified CD14+ monocytes isolated from SLE patients revealed a significant positive correlation between monocyte signatures from AsA subjects and glycolysis, but not OXPHOS, suggesting they are likely to be metabolically M1 in nature. Glycolysis was also correlated with B cells in AsA individuals suggesting that B cells, along with monocyte/myeloid cells in this patient population, maintain an activated phenotype.

A limitation of the current study is that the computational and experimental approaches are inferential. By attempting to provide a comprehensive translation of all GWAS findings, it remains challenging to determine those genes that are causal and those that may be considered false-positive associations. To address this, PPI networks and unsupervised MCODE clustering based on interaction strength help exclude those genes lacking strong connections within or between similarly functioning clusters to uncover biologically-relevant pathways. In comparison to networks generated from random SNPs and genes, the densely connected PPIs and highly significant gene-set annotation enrichments displayed in the current manuscript are clearly distinct from background noise and often implicate unexpected molecular pathways that may be ancestrally-motivated.

Other limitations of this study include those related to the data integrated in our pipeline. Initial genetic findings were based on the Immunochip which was specifically constructed for deep replication of major autoimmune and inflammatory diseases and fine-mapping of established GWAS loci¹⁷. As this platform was designed for use in European populations, it may have introduced considerable bias, especially with respect to the dominance of immune and interferon-mediated signaling pathways observed in our EA analyses. Given that type I IFN is a primary pathogenic factor in SLE regardless of ancestral background, additional analyses, including an examination of pathways predicted from summary AsA GWAS, differential gene expression from AsA whole-blood samples and GSVA enrichment of interferon gene signatures in AsA patients confirmed the well-established association between SLE and IFN across ancestral populations. Nonetheless, the specific triggers of IFN production and the mechanisms by which interferon signaling perpetuate the cycle of autoreactive cells and autoreactive antibody production are not completely clear, and evidence presented here suggest that different mechanisms exist for autoantibody and IFN regulation between EA and AsA populations. Finally, the ability to map SNPs 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. However, of all SNP-predicted genes, only those encoding proteins included on the STRINGdb platform were incorporated into PPI interaction networks. While this is useful for filtering out unrelated genes, it excludes the large number of non-coding genes implicated in our SNP-to-gene predictions, an important consideration given the growing evidence highlighting the contributions of long non-coding RNAs and microRNAs in SLE^{34,47}. 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. It is for this reason that multiple platforms were used, including Ingenuity Pathway Analysis (IPA), EnrichR, which pools a myriad of public databases, and cell (IScope) and functional (BIG-C) analytic tools, to obtain orthogonal and reproducible annotations.

In conclusion, the SNP-associated predicted genes and gene expression profiles outlined here implicate fundamental differences in lupus molecular pathways enriched in EA and AsA ancestral populations. Our findings suggest that while certain pathways may be enriched in one ancestral population over another, it is important to note that those pathways may not be active within every patient of a given ancestry and may be active in a patient from different ancestry. Systems bioinformatics and assessment using gene signature enrichment analyses revealed alterations in cellular metabolism and cell stress signatures that may be more prevalent in patients of Asian ancestry. Whereas treatment strategies aimed at restoration of metabolic and/or antioxidant pathways are not straightforward, the current findings suggest that targeting metabolic dysfunction may hold promise for AsA patients who respond poorly to conventional therapies. Indeed, mTOR pathway modulators such as N-acetyl cysteine and rapamycin appear to be viable therapies for reducing disease activity^{50,51}. Recently, pioglitazone, a peroxisome proliferator-activated receptor (PPAR γ) agonist, was found to ameliorate nephritis symptoms in lupus-prone animals⁵². Together, our study demonstrates that multilevel analysis is capable of defining gene regulation that not only reflects differences in EA and AsA populations, but also represents candidate pathways that may be the target of ancestry-specific therapies.

Material and methods

Genomic functional categories. The Variant Effect Predictor (VEP) tool available on the Ensembl genome browser 93 was used for annotation information to specify SNPs located within non-coding regions, including micro (mi)RNAs, long non-coding (lnc)RNAs, splice region variants, non-coding transcript exon variants, introns and intergenic regions. Regulatory regions include transcription factor binding sites (TFBS), promoters, enhancers, repressors, promoter flanking regions (PFRs) and open chromatin (OCRs). Coding regions were broken down further and include 5'UTRs, 3'UTRs, synonymous and nonsynonymous (missense and nonsense) mutations. The online resource tool HaploReg (version 4.1⁵³; was also used to identify DNA features, regulatory elements and assess regulatory potential.

Identification of SLE-associated SNPs and predicted genes. SLE Immunochip studies identified single nucleotide polymorphisms (SNPs) significantly associated with SLE in EA (6748 cases; 11,516 controls, $p < 1 \times 10^{-6}$)⁸. Because of the lower power of the East Asian Immunochip analysis reported in Sun et al.⁶ (2485 cases and 3947 controls from Koreans (KR), Han Chinese (HC) and Malaysian Chinese (MC)), we identified 700 SNPs from 578 associated regions using a significance threshold of $p < 5 \times 10^{-3}$. Because of the extensive linkage disequilibrium in the HLA region, SNPs in the region spanning chr6:28014374-33683352 were omitted from the analysis. Asian validation SNPs were previously described^{7,9}. Expression quantitative trait loci (eQTLs) were then identified using GTEx version 8 (GTExportal.org⁵⁴) and the Blood eQTL browser database¹³ and mapped to their associated eQTL expression genes (E-Genes). To find SNPs in enhancers and promoters, especially in intergenic regions, and their associated transcription factors and downstream target genes (T-Genes), we queried the atlas of Human Active Enhancers to interpret Regulatory variants (HACER)¹⁵ and the GeneHancer database¹⁴. To find structural SNPs in protein-coding genes (C-Genes), we queried the human Ensembl genome browser (GRCh38.p12) and dbSNP. Several additional databases were used to generate loss-of-function prediction scores, including SIFT4C^{55,56} and PolyPhen-2⁵⁷. All other SNPs were linked to the most proximal gene (P-Gene) or gene region as previously detailed⁸. Predicted genes were examined as equal entities; no gene, regardless of provenance, was given more weight or importance over another type. For overlap studies, Venn diagrams were computed and visualized using InteractiVenn⁵⁸. All predicted genes were divided into an EA, AsA or shared group depending on the ancestral designation of the original SLE-associated SNP. The random gene cohort was generated using the Random Gene Set Generator.

Differential gene expression analysis. For DEG analysis of E-MTAB-11191, affymetrix probes were annotated with custom BrainArray (BA) chip definition files (CDFs)⁵⁹ as previously described⁶⁰. Any probes with different Affymetrix and BA gene annotations were excluded. GCRMA-normalized expression values were variance corrected using local empirical Bayesian shrinkage before calculation of differentially expressed genes (DEGs) using the eBayes function in the BioConductor LIMMA package. P-values were adjusted for multiple hypothesis testing using the Benjamini–Hochberg False discovery rate (FDR). Significant Affymetrix and BA probes within each study were merged and filtered to retain probes with a pre-set FDR < 0.2 which were considered statistically significant. This FDR threshold was employed to avoid falsely excluding genes of interest. This list was further filtered to retain only the most significant probe per gene in order to remove duplicate genes.

Functional gene set analysis. For both ancestral groups, predicted gene lists 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. Genes are sorted into 54 categories based on their most likely biological function and/or cellular localization based on information from multiple online tools and databases including UniProtKB/Swiss-Prot, gene ontology (GO) Terms, MGI database, KEGG pathways⁶¹, NCBI, PubMed, and the Interactome, and has been previously described^{62,63}. I-Scope is a custom clustering tool used to identify immune infiltrates in large gene datasets, and has been described previously⁶⁴. Briefly, I-Scope was created through an iterative search of more than 17,000 genes identified in more than 50 microarray datasets. These genes were researched for immune cell specific expression in 30 hematopoietic sub-categories: T cells, regulatory T cells, activated T cells, anergic T cells, CD4 T cells, CD8 T cells, gamma-delta T cells, NK/NKT cells, T & B cells, B cells, activated B cells, T, B & myeloid, monocytes, monocytes & B cells, MHC Class II expressing cells, monocyte dendritic cells, dendritic cells, plasmacytoid dendritic cells, Langerhans cells, myeloid cells, plasma cells, erythrocytes, neutrophils, low density granulocytes, granulocytes, platelets, and all hematopoietic stem cells. Enrichment of GO Biological Processes (BP) using Enrichr⁶⁵ and the Ingenuity Pathway Analysis (IPA) platform provided additional genetic pathway identification.

Network analysis and visualization. For SNP-predicted genes, only those encoding proteins included on the STRINGdb (version 1.3.2) platform were incorporated into protein–protein interaction (PPI) networks. To provide a quality control checkpoint, the significance of network connectivity was assessed, and we only moved forward with pathway analysis if the PPI enrichment p-value was $< 1 \times 10^{-16}$ (lowest p-value, highest possible probability estimate). Visualization of protein–protein interaction and relationships between genes within datasets was done using Cytoscape (version 3.6.1) software. Briefly, STRINGdb generated networks were imported into Cytoscape and partitioned with MCODE via the clusterMaker2 (version 1.2.1) plugin.

Gene set variation analysis (GSVA). The GSVA (V1.25.0) software package for R/Bioconductor and has been described previously. Briefly, GSVA is a nonparametric, unsupervised method for estimating the variation of pre-defined gene sets in patient and control samples of microarray expression datasets. The input for

the GSVA algorithm was a gene expression matrix of log2 microarray of expression values and a collection of pre-defined gene signatures. Enrichment scores (GSVA scores) were calculated non-parametrically using a Kolmogorov-Smirnov (KS)-like random walk statistic and a negative value for each gene set. GSVA gene signatures using official gene symbols are listed in Table S7). All interferon and cytokine signatures (core IFN, IFNB1, IFNA2, IFNW, IFNG and TNF) have been described previously. Metabolic signatures were based on literature mining and established IPA canonical pathways. Enrichment of each signature was examined in EA and AsA SLE patients and healthy control PBMCs from FDAPBMC1 for EA or GSE81622 for AsA. Differences between controls and SLE patient GSVA enrichment scores were determined using the Welch's t-test for unequal variances in Graphpad PRISM 8.0.

Regression models. For all linear models, GSVA scores for cell type and/or pathway in each sample were used as input. Simple linear regression was performed in Graphpad PRISM 8.0.

Data availability

All microarray and RNA-seq datasets listed in this publication (GSE164457, GSE81622, FDAPBMC1 and EMTAB11191) are publicly available on the NCBI's database Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>).

Received: 2 September 2022; Accepted: 29 March 2023

Published online: 01 April 2023

References

- Li, P. H. & Lau, C. S. Lupus in the far East: A modern epidemic. *Int. J. Rheum. Dis.* **20**, 523–525 (2017).
- Li, M. *et al.* Chinese SLE Treatment and Research group (CSTAR) registry: I. Major clinical characteristics of Chinese patients with systemic lupus erythematosus. *Lupus* **22**, 1192–1199 (2013).
- Yap, D. Y. H. & Chan, T. M. Lupus nephritis in Asia: Clinical features and management. *Kidney Dis.* **1**, 100–109 (2015).
- Lanata, C. M. *et al.* Genetic contributions to lupus nephritis in a multi-ethnic cohort of systemic lupus erythematosus patients. *PLoS ONE* **13**, e0199003 (2018).
- Mok, M. Y. & Li, W. L. Do Asian patients have worse lupus?. *Lupus* **19**, 1384–1390 (2010).
- Sun, C. *et al.* High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat. Genet.* **48**, 323–330 (2016).
- Lessard, C. J. *et al.* Identification of a systemic lupus erythematosus risk locus spanning ATG16L2, FCHSD2, and P2RY2 in Koreans. *Arthritis Rheumatol.* **68**, 1197–1209 (2016).
- Langeveld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
- Morris, D. L. *et al.* Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat. Genet.* **48**, 940–946 (2016).
- 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.* <https://doi.org/10.1038/ng.3434> (2015).
- Owen, K. A. *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.* **107**, 864–881 (2020).
- Kain, J. *et al.* Mendelian randomization and pathway analysis demonstrate shared genetic associations between lupus and coronary artery disease. *Cell Rep. Med.* **3**, 100805 (2022).
- Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
- Fishilevich, S. *et al.* GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*. **2017**, bax028 (2017).
- Wang, J. *et al.* HACER: An atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019).
- Kingsmore, K. M. *et al.* Altered expression of genes controlling metabolism characterizes the tissue response to immune injury in lupus. *Sci. Rep.* **11**, 14789 (2021).
- Cortes, A. & Brown, M. A. Promise and pitfalls of the ImmunoChip. *Arthritis Res. Ther.* **13**, 101 (2011).
- Li, P. *et al.* Glutathione peroxidase 4-regulated neutrophil ferroptosis induces systemic autoimmunity. *Nat. Immunol.* **22**, 1107–1117 (2021).
- Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinform.* **14**, 7 (2013).
- Lightfoot, Y. L., Blanco, L. P. & Kaplan, M. J. Metabolic abnormalities and oxidative stress in lupus. *Curr. Opin. Rheumatol.* **29**, 442–449 (2017).
- Andreoletti, G. *et al.* Transcriptomic analysis of immune cells in a multi-ethnic cohort of systemic lupus erythematosus patients identifies ethnicity- and disease-specific expression signatures. *Commun. Biol.* **4**, 488 (2021).
- Shah, D., Mahajan, N., Sah, S., Nath, S. K. & Paudyal, B. Oxidative stress and its biomarkers in systemic lupus erythematosus. *J. Biomed. Sci.* **21**, 23 (2014).
- Namas, R., Renauer, P., Ognenovski, M., Tsou, P. S. & Sawalha, A. H. Histone H2AX phosphorylation as a measure of DNA double-strand breaks and a marker of environmental stress and disease activity in lupus. *Lupus Sci. Med.* **3**, e000148 (2016).
- Zhang, H. *et al.* Anti-dsDNA antibodies bind to TLR4 and activate NLRP3 inflammasome in lupus monocytes/macrophages. *J. Transl. Med.* **14**, 156 (2016).
- Giles, B. M. & Boackle, S. A. Linking complement and anti-dsDNA antibodies in the pathogenesis of systemic lupus erythematosus. *Immunol. Res.* **55**, 10 (2013).
- Caielli, S. *et al.* Erythroid mitochondrial retention triggers myeloid-dependent type I interferon in human SLE. *Cell* **184**, 4464–4479.e19 (2021).
- Lewis, M. J. & Jawad, A. S. The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. *Rheumatology (Oxford)* **56**, i67–i77 (2017).
- Feldman, C. H. *et al.* Epidemiology and sociodemographics of systemic lupus erythematosus and lupus nephritis among US adults with Medicaid coverage, 2000–2004. *Arthritis Rheum.* **65**, 753–763 (2013).
- Robinson, E. K., Covarrubias, S. & Carpenter, S. The how and why of lncRNA function: An innate immune perspective. *Biochimica et Biophysica Acta Gene Regulat. Mech.* **1863**, 194419 (2020).

30. Riege, K. *et al.* Massive effect on lncRNAs in human monocytes during fungal and bacterial infections and in response to vitamins A and D. *Sci. Rep.* **7**, 40598 (2017).
31. Stein, N. *et al.* IFNG-AS1 enhances interferon gamma production in human natural killer cells. *iScience* **11**, 466–473 (2019).
32. Dai, R. & Ahmed, S. A. microRNA, a new paradigm for understanding immunoregulation, inflammation, and autoimmune diseases. *Transl. Res.* **157**, 163 (2011).
33. Tsai, C.-Y. *et al.* Cross-talk between mitochondrial dysfunction-provoked oxidative stress and aberrant noncoding RNA expression in the pathogenesis and pathophysiology of SLE. *Int. J. Mol. Sci.* **20**, 5183 (2019).
34. Tsai, C.-Y. *et al.* Aberrant non-coding RNA expression in patients with systemic lupus erythematosus: Consequences for immune dysfunctions and tissue damage. *Biomolecules* **10**, 1–24 (2020).
35. Wang, X. *et al.* Long noncoding RNAs in the regulation of oxidative stress. *Oxid. Med. Cell. Longev.* **2019**, 1318795 (2019).
36. Catalina, M. D. *et al.* Patient ancestry significantly contributes to molecular heterogeneity of systemic lupus erythematosus. *JCI Insight* **5**, e140380 (2020).
37. Wang, Y.-F. *et al.* Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.* **12**, 772 (2021).
38. Griss, J. *et al.* B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. *Nat. Commun.* **10**, 4186 (2019).
39. Nossent, J. C., Becker-Merok, A., Rischmueller, M. & Lester, S. Susceptibility for lupus nephritis by low copy number of the FCGR3B gene is related to increased levels of pathogenic autoantibodies. *Autoimmune Dis.* **2013**, 750814 (2013).
40. Banchereau, R. *et al.* Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* **165**, 551–565 (2016).
41. Menard, L. C. *et al.* B cells from African American lupus patients exhibit an activated phenotype. *JCI Insight* **1**, e87310 (2016).
42. Cruz, C. S. D. & Kang, M.-J. Mitochondrial dysfunction and damage associated molecular patterns (DAMPs) in chronic inflammatory diseases. *Mitochondrion* **41**, 37 (2018).
43. Buskiewicz, I. A. *et al.* Reactive oxygen species induce virus-independent MAVS oligomerization in systemic lupus erythematosus. *Sci. Signal.* **9**, ra115 (2016).
44. Ratliff, B. B., Abdulmahdi, W., Pawar, R. & Wolin, M. S. Oxidant mechanisms in renal injury and disease. *Antioxid. Redox Signal.* **25**, 119 (2016).
45. Jakes, R. W. *et al.* Systematic review of the epidemiology of systemic lupus erythematosus in the Asia-Pacific region: Prevalence, incidence, clinical features, and mortality. *Arthritis Care Res.* **64**, 159–168 (2012).
46. Richman, I. B. *et al.* European genetic ancestry is associated with a decreased risk of lupus nephritis. *Arthritis Rheum.* **64**, 3374–3382 (2012).
47. Honarpisheh, M., Köhler, P., Rauchhaupt, E. von & Lech, M. The involvement of microRNAs in modulation of innate and adaptive immunity in systemic lupus erythematosus and lupus nephritis. *J. Immunol. Res.* **2018**, 4126106 (2018).
48. Zhang, C., Wang, H., Yin, L., Mao, Y. & Zhou, W. Immunometabolism in the pathogenesis of systemic lupus erythematosus. *J. Transl. Autoimmun.* **3**, 100046 (2020).
49. Grayson, P. C. *et al.* Metabolic pathways and immunometabolism in rare kidney diseases. *Ann. Rheum. Dis.* **77**, 1226 (2018).
50. Fernandez, D., Bonilla, E., Mirza, N., Niland, B. & Perl, A. Rapamycin reduces disease activity and normalizes T cell activation-induced calcium fluxing in patients with systemic lupus erythematosus. *Arthritis Rheum.* **54**, 2983–2988 (2006).
51. Lai, Z.-W. *et al.* N-Acetylcysteine reduces disease activity by blocking mammalian target of rapamycin in T cells from systemic lupus erythematosus patients: A randomized, double-blind, placebo-controlled trial. *Arthritis Rheum.* **64**, 2937–2946 (2012).
52. Liu, D. & Zhang, W. Pioglitazone attenuates lupus nephritis symptoms in mice by modulating miR-21-5p/TIMP3 Axis: The key role of the activation of peroxisome proliferator-activated receptor-γ. *Inflammation* **44**, 1416–1425 (2021).
53. Ward, L. D. & Kellis, M. HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877–D881 (2016).
54. The Genotype-Tissue Expression (GTEx) project. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010069/>. (2013).
55. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
56. Sim, N. L. *et al.* SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–7 (2012).
57. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* <https://doi.org/10.1002/0471142905.hg0720s76> (2013).
58. Heberle, H., Meirelles, V. G., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinform.* **16**, 169 (2015).
59. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
60. Labonte, A. C. *et al.* Identification of alterations in macrophage activation associated with disease activity in systemic lupus erythematosus. *PLoS ONE* **13**, e0208132 (2018).
61. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
62. Catalina, M. D., Bachali, P., Geraci, N. S., Grammer, A. C. & Lipsky, P. E. Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. *Commun. Biol.* **2**, 140 (2019).
63. Catalina, M. D., Owen, K. A., Labonte, A. C., Grammer, A. C. & Lipsky, P. E. The pathogenesis of systemic lupus erythematosus: Harnessing big data to understand the molecular basis of lupus. *J. Autoimmun.* <https://doi.org/10.1016/j.jaut.2019.102359> (2019).
64. Ren, J. *et al.* Selective histone deacetylase 6 inhibition normalizes B cell activation and Germinal Center Formation in a model of systemic lupus erythematosus. *Front. Immunol.* **10**, 2512 (2019).
65. Xie, Z. *et al.* Gene set knowledge discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).

Acknowledgements

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 [K.A.O., K.A.B., A.P., P.B.]. Additional support to all authors was provided by The Alliance for Lupus Research (lupusresearch.org) and the Department of the Army (W81XWH-20-1-0686) to C.D.L. Analysis of data included in GSE164457 was supported by the Centers for Disease Control (5U01DP006486). The authors would like to thank members of the AMPEL/RILITE, Wake Forest and UCSF scientific teams for their critical reading of the manuscript.

Author contributions

Conceptualization, P.E.L., A.C.G. and K.A.O.; Methodology, P.E.L., K.A.B. and K.A.O.; Software, Formal Analysis and Investigation, K.A.O., K.A.B., A.P., P.B., H.A., M.C.M., M.D., J.Y. and N.S.; Writing—Original Draft, P.E.L.

and K.A.O.; Writing—Review & Editing, C.D.L., M.C.M., T.D.H., M.D., J.Y., N.S., P.E.L. and K.A.O.; Funding Acquisition, P.E.L., A.C.G., M.D., J.Y., and C.D.L.; Supervision, P.E.L. and A.C.G.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-32569-6>.

Correspondence and requests for materials should be addressed to K.A.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Reply to: A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis

Received: 19 April 2022

Kip D. Zimmerman¹✉, Ciaran Evans² & Carl D. Langefeld³✉

Accepted: 8 December 2022

Published online: 22 December 2022

Check for updates

 REPLYING TO Murphy, A. E. & Skene, N. G. *Nature Communications* <https://doi.org/10.1038/s41467-022-35519-4> (2022)

The purpose of our publication “A practical solution to pseudoreplication bias in single-cell studies” was to address a need to reduce the number of studies that do not account for within-individual correlation and report astronomically false associations¹. The issue is a statistical hypothesis testing question, and our perspective is influenced by the theory underlying the analysis of correlated data^{2–4}. Both pseudobulk approaches and mixed models account for the within-sample correlation. Murphy and Skene⁵ focus on a minor point in our paper: the relative performance of the pseudobulk approach in comparison to the two-part hurdle mixed model. We concluded that pseudobulk methods “are only slightly underpowered relative to mixed effects models when there is balance in the numbers of cells per individual, [but] not as well powered as mixed effects models when the numbers of cells per individual grow increasingly imbalanced.” While recognizing that pseudobulk approaches are easy to compute and perform well for basic hypotheses, we noted that they are less flexible and slightly conservative. By definition, a statistical test is conservative if the type 1 error rate is less than the nominal level⁶. Mixed models directly model within-individual correlation, are slightly more powerful with unbalanced samples, and test more complicated hypotheses, but are more computationally intensive. We previously recommended the two-part hurdle mixed model because it estimates the proportion of expressing cells as well as the difference in magnitude conditional on expression⁷. Our recommendation was thus not solely based on power and type 1 error rate. Rather, our intent was to provide the community with an immediately actionable method, consistent with robust statistical theory and with great flexibility in modeling a range of hypotheses. Our recommendation, although not the only reasonable approach, is consistent with the established standard approaches of repeated measures and clustered data analysis.

When comparing statistical tests, it is necessary to establish the size of the test (type 1 error rate) before determining the comparative power relative to those tests with appropriate size. Thus, type 1 and type 2 error rates are not treated equally. Neither

the Matthew’s Correlation Coefficient (MCC) nor the receiver operating characteristic (ROC) curves use this classic statistical paradigm of maximizing power after first controlling type 1 error. The MCC is simply the correlation coefficient for a 2 × 2 table and is commonly used to summarize the performance of binary classifiers. The MCC can be expressed as:

$$MCC = \frac{\text{power} - \text{type 1 error rate}}{\sqrt{\frac{1}{\pi(1-\pi)}(\pi(\text{type 1 error rate}) + (1-\pi)\text{power})(\pi(1 - \text{type 1 error rate}) + (1-\pi)(1 - \text{power}))}} \quad (1)$$

where π represents the proportion of hypotheses expected to be rejected under the null. Clearly, if the type 1 error rates are the same, then the test with greater power would have a larger MCC and the MCC agrees with the classic hypothesis testing paradigm. If two tests have different type 1 error rates, the MCC can favor a test which fails to control the type 1 error rate. In fact, as illustrated in Figure 1 of Murphy and Skene’s manuscript, failing to account for the within-sample correlation causes very high type 1 error rates (>0.50) and yet yields high MCC values⁵. Conversely, as observed from Eq. 1, a smaller type 1 error rate can also lead to a larger MCC. On the surface, a smaller type 1 error rate seems desirable. In classical statistical reasoning, such a test should be recalibrated to have the appropriate size in order to (1) understand the expected number of false rejections and apply appropriate multiple comparison methods, and (2) balance type 1 error rate and power relative to the specifics of the application (e.g., clinical trial, scRNA-seq). Importantly, ranking different methods by MCC means we are ranking as a function of the unknown proportion under the alternative hypothesis, power, and type 1 error rate jointly. However, the proportion and power are dependent on the specifics of the two-parameter alternative hypothesis, and the resulting relationship need not be monotonic. Although exceptions may exist, we generally prefer tests with type 1 error as close to the stated nominal level (e.g., 0.05, 0.001) and do not endorse the application of MCC in this context.

¹Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA. ²Department of Statistical Sciences, Wake Forest University, Winston-Salem, NC, USA. ³Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA.

✉ e-mail: kdzimmer@wakehealth.edu; clangefe@wakehealth.edu

Murphy and Skene also use ROC curves to compare the power of different tests at the same type 1 error rate. Unlike the MCC, ROC curves can be interpreted in the classic hypothesis testing framework: a method with a higher AUC tends to be more powerful at any given level of type 1 error and would be preferred if the rejection threshold has a known error rate. However, the threshold for a conservative test to achieve a type 1 error rate of 0.05 may not be known, and this information is hidden by the ROC curve. In the absence of a systematic method for selecting the rejection threshold, we prefer tests with known error.

Importantly, mathematically, the mixed model parameter estimate has variance less than or equal to that of the pseudobulk estimate, with exact equality only if there is an equal number of cells from each individual. As a result, the power of the mixed model test is no smaller than that of the pseudobulk test and is slightly greater in the unbalanced setting. In reality, we never see a perfectly balanced design (though in practice the difference in power is small). Our original simulations and statements are consistent with these theoretical conclusions¹.

In addition to the limitations of MCC and ROC curves for comparing hypothesis tests, there are several misleading components we would like to address⁵. First, in Supplementary Figure 1 of Murphy and Skene, the two-part hurdle model is missing and adding it would reveal how much closer the two-part hurdle model is to the nominal *p*-value than the pseudobulk methods. Second, what Murphy and Skene state as a “flaw” in our simulation regarding normalization is not a flaw. Specifically, we explained in our Methods section that this workflow was implemented to prevent DESeq2’s normalization method from removing the simulated effects. Contrary to Murphy and Skene⁵, we also note that the pseudobulk test, which has a type 1 error rate less than nominal level, is by definition a conservative test. Lastly, Murphy and Skene rightfully point out that our simulations to obtain type 1 errors for each method were not completed on the exact same datasets. However, this is again a very minor issue as both of our simulations are based on random draws from the identical distributions. The type 1 errors presented in Table 1 of our manuscript are based on 250,000 iterations and are consistent with Murphy and Skene’s results in Supplementary Figure 1 which are based on 20,000 iterations. Thus, our estimates go the right place, but the bound on error of our estimates is markedly smaller than Murphy and Skene’s bound on error.

In conclusion, it is of critical importance to account for the within-individual correlation in scRNA-seq data, which produces grossly inflated false positives when ignored. The two-part hurdle mixed model is one powerful and flexible option which controls type 1 error, is easily implemented with existing software, and is consistent with statistical literature on the analysis of clustered data. While aggregate methods are often a suitable alternative, we do not endorse the use of MCC as the metric for comparing hypothesis testing methods and we do not believe the differences observed by Murphy and Skene are sufficient to warrant ignoring mixed models.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

References

1. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
2. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. *Applied Longitudinal Analysis* (Wiley, 2011).

3. Hardin, J. W. *Generalized Estimating Equations (GEE)* (John Wiley & Sons, Ltd, 2005).
4. Millar, R. B. & Anderson, M. J. Remedies for pseudoreplication. *Fish. Res.* **70**, 397–407 (2004).
5. Murphy, A. E. & Skene, N. G. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat. Commun.* <https://doi.org/10.1038/s41467-022-35519-4> (2022).
6. Lehmann, E. L. & Romano, J. P. *Testing Statistical Hypotheses* (Springer, 2005).
7. Finak, G. et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

Acknowledgements

This work was supported by the Wake Forest Center for Public Health Genomics and grant U01 NS036695 (Co-PI Langefeld) from NIH, the Department of Defense W81XWH-20-1-0686 (PI Langefeld), and by the Cancer Center Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30CA012197).

Author contributions

KDZ, CDL, and CE met and discussed the statistical theory underlying this response. KDZ, CDL, and CE all wrote components of the manuscript. KDZ assembled a final version of the manuscript and edited it together with significant input from CDL and CE.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-35520-x>.

Correspondence and requests for materials should be addressed to Kip D. Zimmerman or Carl D. Langefeld.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022