

AWARD NUMBER: W81XWH-20-1-0686

TITLE: Multi-Ancestral Genomic Approach to SLE-Precision Medicine

PRINCIPAL INVESTIGATOR: Carl D. Langefeld

CONTRACTING ORGANIZATION: Wake Forest University Health Sciences
Winston-Salem, NC

REPORT DATE: October 2021

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 2021		2. REPORT TYPE Annual		3. DATES COVERED 30Sep2020-29Sep2021	
4. TITLE AND SUBTITLE Multi-Ancestral 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 three ancestral groups (European, 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					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Unclassified	18. NUMBER OF PAGES 78	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code)

TABLE OF CONTENTS

	<u>Page</u>
1. Introduction	4
2. Keywords	4
3. Accomplishments	5
4. Impact	18
5. Changes/Problems	19
6. Products	20
7. Participants & Other Collaborating Organizations	21
8. Special Reporting Requirements	26
9. Appendices	26

1. Introduction

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

2. Keywords

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

3. Accomplishments

What were the major goals of the project?

Within the first 12 months, this project has shown significant progress across all three specific Aims. We believe we are on schedule and even slightly farther along in progress than anticipated. In part, this is the result of a highly interactive and positive collaboration among investigators. One of the major accomplishments includes the successful identification and compilation of SLE risk SNPs datasets that are linked to their respective plausible gene-targets, across multiple ancestries. We have also developed and **successfully implemented** the programming pipelines necessary to complete high-throughput analysis of *in silico* binding for drug targets with FDA-approved small molecules.

We have provided two tables outlining the project's accomplishments and progress, to-date. **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 1-12).

Major Accomplishment/Progress Category	Descriptor
Manuscripts and Presentations	• Two manuscripts accepted and published • One manuscript submitted • Two manuscripts in preparation (Submission within next 2-6 months) • Two presentations at the Lupus 21st Century 2021 Conference
Data processing and Analytic Pipelines	• Drug target identification and processing for molecular docking • Summarization of molecular docking, summaries include generation of $\Delta\Delta G$ for paired docking comparisons (e.g., risk versus non-risk protein variants for drug target proteins) and graphical illustrations.
Created Datasets	• SLE risk SNP datasets linked to their most plausible-implicated gene targets (e.g., via transcription factor binding, tGenes; gene expression, eGenes; and protein coding variants cGenes. If SLE-risk SNPs do not link to any gene by the above process, then the SNP was linked to proximal genes. • ZINC 15 (FDA-approved small molecules) IDs linked to public repositories via identifiers (e.g., SMILES, pubchemID, and common drug names). • Database of ΔG and $\Delta\Delta G$ binding energies for ZINC 15 molecules across 9 drug targets (2 structures per drug target, corresponding to the risk and non-risk amino acid linked to the SLE risk SNP). ○ <i>This represents over 38,000 in silico binding experiments</i> (2,114 small molecules x 9 drug targets x 2 isoforms (per target)).

	○ This dataset will continue to grow with the analysis of additional SLE drug targets.
System biology analyses	• Pathways and system biology analysis for SLE-associated SNPs in three ancestries (European, African, and Asian Ancestries). • Pathways and systems biology analyses for genes implicated by methylation study in female MZ twins discordant for SLE.

Table 2: Project completion and status (Months 1-12) as related to proposal's SOW.

Goals and Milestones as listed in the original SOW.		Progress Report 10-2021 Update. Completion Status
Specific Aims (specified in proposal)	Timeline	
Specific Aim 1: Identify SLE-risk Genes	Months	
Subtask 1: Identify SLE-risk single nucleotide polymorphisms (SNPs) in women	1-24	• SNP associations have been compiled across the SLE Immunochip studies (female cohorts) which includes four ancestries: European, African, Hispanic, and Asian. • SNP associations have been compiled across the SLE Korean GWAS and SLE GWAS in patients of Amerindian Ancestry. • The African American GWAS study is still in progress. Completion Progress ~70%
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. • SLE-risk SNPs linked to genes via transcription factor binding, eQTL, and proximity within European, African, and Asian SLE studies. Completion Progress ~60% 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 (one published, one under construction) • Differentially expressed genes identified by annotation from DNA methylation study in MZ female twins discordant for SLE summarized in submitted manuscript. Completion Progress ~40% 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. • 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 (submitted to Genes). Completion progress ~30%
Subtask 5: Identify and write potential manuscripts on multi-omic analysis of SLE-risk associated variants and genes.	6-36	• Multi-omic manuscript submitted and published in the American Journal of Human Genetics. Manuscript focused on pathway analysis for ancestry-dependent and ancestry-independent SLE associations in patients of African and European ancestry. • Manuscript submitted exploring the differential analysis of methylation in twins who are discordant for SLE. • Manuscript in preparation which explores biological pathways for genetic markers of an SLE-subtype with clinically severe outcomes (enriched within patients of Asian ancestry).
Milestone(s) Achieved: Lists of SLE-risk associated genes informed by ancestry, tissue, and female sex	3-30	• Lists of SLE-risk genes informed by ancestry and female-sex. Completion progress ~40%
Local IRB/IACUC Approval	Completed	
Specific Aim 2: For genes and gene lists discovered in Specific Aim 1, complete systems biology and pathway analysis		
Subtask 1: Identify drug targets: Process lists of genes in Aim 1 into one of four Target Groups based on functional criteria, including pathway analyses	3-30	• Group 3 targets (TBFS or Protein Coding variants) have been identified across four ancestries: European, African, Hispanic, and Asian Ancestries. • Identified targets have been analyzed via Biological Pathways

		analysis to identify relevant (enriched) pathways within SLE (Group 4 targets). Completion progress ~30%
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	- All Group 3 targets derived from nonsynonymous SLE risk SNPs have been prioritized based on the availability for high-quality protein structures. - Availability of high-quality structures includes the protein databank (PDB) and now, AlphaFold (available as of July 2021). ~100 drug targets identified via process outlined in Aims 1 and 2 have been submitted and reviewed via CoLTS scoring and prioritization algorithm. Targets span 10 cGenes. Completion progress ~20%
Subtask 3: Identify and write potential manuscripts incorporating systems biology and drug target prioritization to evaluate genetic architecture of SLE.	12-36	Manuscript in preparation for systems biology and drug target evaluation of Group 3 targets (based on cGenes data). This manuscript also relates to Aim 3 "Identify and Prioritize Drugs". Completion progress ~20%
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). - Identified optimal Protein Data Bank structures - Identified alternative source for missing PDB structures (AlphaFold AI-predicted structures) Completion progress ~20%
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	- For current list of genes in Aim 1, completed and summarized in the three manuscripts (one published, one submitted, one under construction). - Expand analyses with additional genes identified in the evolving analyses in Aims 1 and 2. Completion progress ~20%
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	- Successfully developed python pipeline to annotate binding data of FDA-approved small molecules with common-names and common-database identifiers (e.g., SMILES, pubChemID) - Completed analysis of all c-Gene based targets with existing PDB structures. - Prepared data for c-Gene based targets with AlphaFold structural information (e.g., those without PDB structures). - Generated quantitative metrics to summarize differential binding between risk and non-risk SLE-protein structures. Completion Status: ~25%
Subtask 3: Prioritize drugs from Subtasks 1 and 2 using CoLTS scoring algorithm	6-36	Developing strategies to enable efficient analysis of drugs and small molecules not previously analyzed by CoLTS scoring algorithm. Completion Status: ~10%
Milestone(s) Achieved: Lists of genetically-informed FDA-approved drugs and small molecules, novel to treatment of SLE.	12-36	- List of molecular docking results for over 38,000 analyses. These results represent 2,114 docking experiments for each of the cGene drug targets that had high-quality PDB structures (as of September 2021). All results summarized and linked to common drug names for next stage of drug prioritization. - Current on list of cGenes, expanding to other target genes (eGenes, tGenes, pGenes) coming year. Completion status: ~20%

What was accomplished under these goals?

For each of the three specific aims in the proposal, we have outlined and described the 1) major activities, 2) specific objectives, and 3) significant results and key outcomes.

Specific Aim 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*).

Specific Objectives: Leveraging a multi-omics approach, we are investigating and compiling multiple sources of SLE-associations (SNP associations, differential gene expression, and differential methylation) to delineate a list of genes that are plausibly implicated in SLE risk.

Major activities:

Major progress in this aim centers on:

- 1) Examination of all published GWAS and ImmunoChip studies to identify associated regions and corresponding SNPs.
- 2) Link identified SLE-risk SNPs to genes via protein coding changes (cGenes); transcription factor binding (tGenes); and gene expression (eGenes). If none of these links exist, then assign SLE-risk SNPs to proximal genes consistent with linkage disequilibrium (pGenes).
- 3) Link via Illumina annotation CpG sites showing differential methylation in MZ female twins discordant for SLE.
- 4) Result of the above analyses is the current list of genes with ancestry-specific or trans-ancestral evidence linked to SLE risk in women.

Significant Results:

The full list of cGenes, tGenes, eGenes and pGenes is too long to include in this report. As an important example, please see Table 3, under summary of Aim 2, for list of cGenes. The relevant lists of genes (cGenes, tGenes, eGenes, tGenes) exist as supplemental tables in our three manuscripts. See Appendix for published (American Journal of Human Genetics) and submitted (Genes) manuscripts.

Here, we summarize some of the primary results of the analyses.

We published the manuscript *Analysis of Single Nucleotide Polymorphisms Associated with Systemic Lupus Erythematosus Provides Insights into Molecular Pathways Operative in Ancestral Groups* in the American Journal of Human Genetics (Owen Am J Hum Gene. 2020 Nov 5;107(5):864-881). We hypothesize that the disparities in disease prevalence, activity, and response to standard medications between African American and European American populations is partially conferred by genomic influences on biological pathways. To test this hypothesis, we completed a comprehensive bioinformatic analyses of all genes predicted from SNP-associated risk loci detected with the ImmunoChip. By combining genes predicted via eQTL analysis, as well as those predicted from base-pair changes in intergenic enhancer sites, coding-region variants, and SNP-gene proximity, we were able to identify 1,731 potential ancestry-specific and trans-ancestry genetic drivers of SLE. Gene associations were linked to upstream and downstream regulators using connectivity mapping, and predicted biological pathways were mined for candidate drug targets. Examination of trans-ancestral pathways reflect the well-defined role for interferons in SLE and revealed pathways associated with tissue repair and remodeling. EA-dominant genetic drivers were more often associated with innate immune and myeloid cell function pathways, whereas AA-dominant pathways mirror clinical findings in AA subjects, suggesting disease progression is driven by aberrant B cell activity accompanied by ER stress and metabolic dysfunction. Finally, potential ancestry-specific and non-specific drug candidates were identified. The integration of all SLE SNP-predicted genes into functional pathways revealed critical molecular pathways representative of each population, underscoring the influence of ancestry on disease mechanism and also providing key insight for therapeutic selection.

Motivated by the results from the direct comparison of ancestral groups, we are currently completing analyses

and writing a manuscript (*Molecular Pathways Identified from Single Nucleotide Polymorphisms Demonstrate Mechanistic Differences in Systemic Lupus Erythematosus Patients of Asian and European Ancestry*) focused the observed differences in pathways based on genes derived from SLE-risk SNPs identified in Asian ancestral

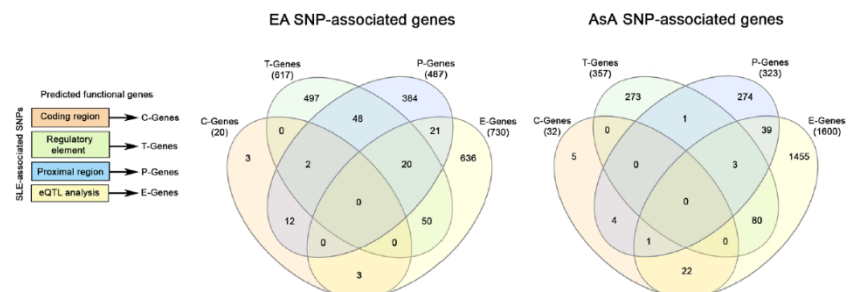


Figure 1. Venn diagram of SLE-risk SNPs functional overlap

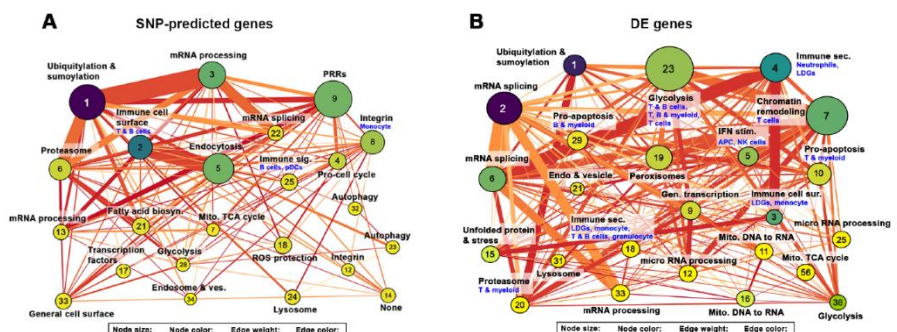


Figure 2. Network analysis of genes associated with risk of SLE

populations compared to those identified European ancestry. Here we note that there is convincing evidence that individuals of Asian-Ancestry (AsA) tend to experience lupus more severely, exhibiting increased renal involvement and tissue damage compared to European-Ancestry (EA) populations. However, the mechanisms underlying elevated severity in this population remain unclear. We again applied a comprehensive bioinformatic and systems biology approach based on all SNP associations found in published Asian ancestry and European ancestry lupus studies detected with the ImmunoChip. We identified 3450 ancestry-specific and trans-ancestry genetic drivers of SLE. We linked the SLE-risk SNPs to genes via coding changes (cGenes), transcription factor binding (tGenes), eQTLs (eGenes), and proximity (pGenes) (**Figure 1**). Functional and canonical pathway analyses via Ingenuity Pathway Analysis, protein-protein interactions networks and cluster analyses (MCODE, Cytoscape) helped to elucidate the distinctions in primary pathways between ancestral groups (**Figure 2**). Genetic associations were examined using connectivity mapping, and gene signatures based on predicted biological pathways were used to interrogate SLE microarray datasets. In summary, dominant pathways in a large proportion of AsA patients reflect elevated oxidative stress, altered metabolism and mitochondrial dysfunction, whereas lupus pathways associated within EA data were more often driven by a robust interferon response (type I and II) related to enhanced cytosolic nucleic acid sensing and signaling. To validate our findings, we are analyzing an independent dataset derived from a published genome-wide association data in SLE patients of Asian ancestry; and to-date, we have identified similar molecular pathways. Finally, gene expression data from AsA SLE patients confirmed the molecular pathways predicted by SNP associations. Together, these analyses suggest fundamental differences in SLE ancestry-related genetic risk predicted molecular pathways differentially impacting AsA and EA individuals with SLE. Importantly, these distinct pathways likely form clinical subtypes of lupus that are present in *any* ancestry, but may be more enriched within some ancestries. The elucidation of these distinct pathways will be informative for drug repurposing efforts that can focus on these pathways.

We completed the analysis and manuscript preparation of *Nucleic acid sensing and interferon-inducible pathways show differential methylation in MZ twins discordant for lupus and overexpression in independent lupus samples: Implications for pathogenic mechanism and drug targeting*, currently under review at Genes.

We hypothesize that epigenetic factors are a significant contributor to SLE risk and may be informative for identifying pathogenic mechanisms and therapeutic targets. To test this hypothesis while controlling for genetic background, we performed an epigenome-wide analysis of DNA methylation in genomic DNA from whole blood in three pairs of female monozygotic (MZ) twins of European-ancestry, discordant for SLE. Results were replicated on the same array in four cell types from a set of four Danish female MZ twin pairs discordant for SLE. Genes implicated by the epigenic analyses were then evaluated in 10 independent SLE gene expression datasets from the Gene Expression Omnibus (GEO).

There were 59 differentially methylated loci between unaffected and affected MZ twins in whole blood, including 11 novel loci. All but two of these loci were hypomethylated in the SLE twins relative to the unaffected twins. The genes harboring these hypomethylated loci exhibited increased expression in multiple independent datasets of SLE patients. This pattern was largely consistent regardless of disease activity, cell type, or renal tissue type. The genes proximal to CpGs exhibiting differential methylation (DM) in the SLE-discordant MZ twins and exhibiting differential expression (DE) in independent SLE GEO cohorts (DM-DE genes), clustered into two pathways: the nucleic acid sensing pathway and the type I interferon pathway (**Figure 3; Table 3**).

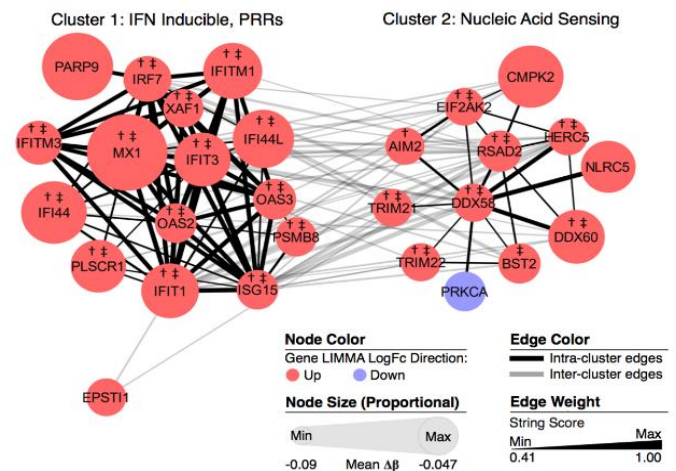


Figure 3. Dominant networks derived from genes linked to SLE-risk differential methylation in female MZ twins discordant for lupus.

Table 3. Differentially methylated probes from three monozygotic twin pairs discordant for SLE.

CpG*	Chr	Pos (bp) [†]	Gene	$\Delta\beta$				P-value	Interferon-Regulated [‡]	Relation to CpG ^{††}
				pair1	pair2	pair3	mean			
cg13304609	1	79085162	IFI44L	-0.24	-0.27	-0.37	-0.29	1.58x10 ⁻¹⁴	IRG	
cg06872964	1	79085250	IFI44L		-0.26	-0.21	-0.24	1.05x10 ⁻⁷¹	IRG	
cg03607951	1	79085586	IFI44L	-0.27	-0.3	-0.21	-0.26	7.23x10 ⁻²²	IRG	
cg17515347	1	159047163	AIM2	-0.09	-0.11	-0.07	-0.09	3.01x10 ⁻¹²	IRG	
cg08272268	1	200380059	ZNF281	-0.08	-0.07	-0.11	-0.09	4.33x10 ⁻¹⁵		S_Shore
cg01028142	2	7004578	CMPK2	-0.22	-0.36	-0.43	-0.33	7.98x10 ⁻⁸	IRG	N_Shore
cg10959651	2	7018020	RSAD2	-0.13	-0.1	-0.16	-0.13	3.14x10 ⁻¹⁴	IRG	
cg10549986	2	7018153	RSAD2	-0.08	-0.09	-0.1	-0.09	1.95x10 ⁻⁹¹	IRG	
cg14126601	2	37384708	EIF2AK2	-0.08	-0.1	-0.12	-0.1	5.55x10 ⁻¹⁶	IRG	S_Shore
cg26337070	2	85999873	ATOH8	-0.06	-0.12	-0.11	-0.1	7.55x10 ⁻⁹		
cg04781494	2	202047246	CASP10	-0.07	-0.13	-0.08	-0.09	8.39x10 ⁻⁸	IRG	
cg15768138	2	219030752	CXCR1	-0.09	-0.12	-0.11	-0.11	7.38x10 ⁻²⁷		
cg13411554	3	53700276	CACNA1D	-0.06	-0.12	-0.09	-0.09	8.66x10 ⁻⁸		
cg22930808	3	122281881	PARP9-DTX3L	-0.36	-0.34	-0.4	-0.37	6.74x10 ⁻¹²⁶	IRG	N_Shore
cg08122652	3	122281939	PARP9-DTX3L	-0.34	-0.31	-0.51	-0.38	1.11x10 ⁻⁹	IRG	N_Shore
cg00959259	3	122281975	PARP9-DTX3L	-0.37	-0.3	-0.34	-0.34	1.32x10 ⁻⁵⁶	IRG	N_Shore
cg06981309	3	146260954	PLSCR1	-0.24	-0.28	-0.21	-0.24	6.41x10 ⁻³¹	IRG	N_Shore
cg02556393	3	168866705	MECOM	-0.08	-0.09	-0.1	-0.09	3.14x10 ⁻⁹⁵		N_Shore
cg07809027	4	15007205	CPEB2	-0.07	-0.1	-0.12	-0.1	2.08x10 ⁻¹⁴		S_Shore
cg02215171	4	89379156	HERC5	-0.08	-0.09	-0.11	-0.09	4.48x10 ⁻¹⁸	IRG	S_Shore
cg17786255	4	108814389	SGMS2	-0.07	-0.09	-0.11	-0.09	2.01x10 ⁻¹⁶	IRG	
cg21873524	4	190942744		-0.1	-0.1	-0.12	-0.11	1.03x10 ⁻⁵⁵		Island
cg24740632	5	134486678		-0.11	-0.12	-0.14	-0.12	2.26x10 ⁻⁶⁰		
cg06012695	6	28770593			-0.1	-0.13	-0.11	3.59x10 ⁻¹⁶		
cg25138053	6	31368016		-0.11	-0.09	-0.07	-0.09	3.67x10 ⁻¹⁵		S_Shore
cg22708150	6	31649619	LY6G5C	-0.12	-0.14	-0.17	-0.14	1.05x10 ⁻¹⁹		N_Shore
cg07292773	6	156718177		0.07	0.1	0.11	0.1	2.22x10 ⁻¹⁷		Island
cg12013713	7	139760671	PARP12	-0.12	-0.14	-0.09	-0.12	1.44x10 ⁻¹⁶	IRG	N_Shore
cg20190772	8	48572496	KIAA0146	-0.08	-0.07	-0.13	-0.09	1.40x10 ⁻⁸		
cg14864167	8	66751182	PDE7A	-0.25	-0.35	-0.45	-0.35	1.21x10 ⁻⁹		N_Shelf
cg06102678	8	81491328		-0.08	-0.12	-0.07	-0.09	1.00x10 ⁻⁸		Island
cg12110437	8	144098888	LY6E	-0.16	-0.17	-0.27	-0.2	3.14x10 ⁻⁹	IRG	N_Shore
cg17555806	10	74448117		-0.08	-0.12	-0.07	-0.09	1.51x10 ⁻⁸		N_Shelf
cg02314339	10	91020653		-0.08	-0.14	-0.11	-0.11	1.72x10 ⁻⁰⁸		
cg06188083	10	91093005	IFIT3	-0.29	-0.16	-0.31	-0.25	6.18x10 ⁻⁸	IRG	
cg05552874	10	91153143	IFIT1	-0.2	-0.28	-0.3	-0.26	6.01x10 ⁻¹⁶	IRG	
cg14910175	10	131840954		-0.07	-0.11	-0.08	-0.09	1.56x10 ⁻¹¹		N_Shelf
cg10552523	11	313478	IFITM1	-0.14	-0.12	-0.14	-0.13	5.90x10 ⁻¹¹⁵	IRG	N_Shelf
cg20566897	11	313527	IFITM1	-0.11	-0.11	-0.09	-0.1	7.00x10 ⁻⁶²	IRG	N_Shelf
cg23570810	11	315102	IFITM1	-0.24	-0.25	-0.34	-0.27	1.43x10 ⁻¹⁸	IRG	N_Shore
cg03038262	11	315262	IFITM1	-0.24	-0.22	-0.29	-0.25	4.41x10 ⁻⁴⁰	IRG	N_Shore
cg20045320	11	319555		-0.19	-0.13	-0.2	-0.18	4.85x10 ⁻¹⁷		S_Shore
cg17990365	11	319718	IFITM3	-0.16	-0.15	-0.15	-0.16	8.78x10 ⁻²⁹⁵	IRG	S_Shore
cg08926253	11	614761	IRF7	-0.15	-0.14	-0.23	-0.17	2.01x10 ⁻⁹	IRG	Island
cg12461141	11	5710654	TRIM22	-0.1	-0.08	-0.12	-0.1	6.35x10 ⁻²⁵	IRG	
cg23571857	17	6658898	XAF1	-0.07	-0.13	-0.11	-0.1	1.46x10 ⁻⁸	IRG	
cg04927537	17	76976091	LGALS3BP	-0.14	-0.11	-0.2	-0.15	2.77x10 ⁻¹⁰	IRG	
cg25178683	17	76976267	LGALS3BP	-0.15	-0.11	-0.21	-0.16	2.01x10 ⁻⁸	IRG	
cg16503797	18	19476805		-0.08	-0.12	-0.08	-0.09	5.39x10 ⁻¹²		N_Shore
cg15871086	18	56526595		-0.07	-0.11	-0.08	-0.09	2.08x10 ⁻¹¹		N_Shelf
cg23352030	20	62198469	PRIC285	0.13	0.19	0.11	0.14	2.36x10 ⁻¹¹		Island
cg16785077	21	42791867	MX1	-0.11	-0.09	-0.12	-0.11	8.45x10 ⁻²⁷	IRG	N_Shore
cg22862003	21	42797588	MX1	-0.31	-0.25	-0.35	-0.31	1.62x10 ⁻²⁵	IRG	N_Shore
cg26312951	21	42797847	MX1	-0.26	-0.17	-0.2	-0.21	6.28x10 ⁻¹⁵	IRG	N_Shore
cg21549285	21	42799141	MX1	-0.5	-0.35	-0.57	-0.47	6.59x10 ⁻¹³	IRG	S_Shore
cg05543864	22	24979755	GGT1	-0.08	-0.08	-0.1	-0.09	1.44x10 ⁻⁴⁵		
cg20098015	22	50971140	ODF3B	-0.19	-0.22	-0.21	-0.21	9.88x10 ⁻⁸³	IRG	S_Shore
cg05523603	22	50973101		-0.17	-0.23	-0.27	-0.22	5.51x10 ⁻¹⁴		S_Shelf
cg02247863	22	50983415		-0.07	-0.1	-0.11	-0.09	2.51x10 ⁻¹³		N_Shore

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

Specific Aim 2: For genes and gene lists discovered in Specific Aim 1, complete systems biology and pathway analysis.

Specific Objectives: Complete systems biology analyses for genes identified in Specific Aim 1, as this will help identify biological systems implicated in SLE and SLE subtypes. For the Aim 1 gene lists, categorize genes into relevant drug targets lists (Figure 4). Drug targets will then be prioritized based on availability of high-quality protein structures.

Major activities:

Major progress in this aim centers on:

- 1) System biology analyses on SNPs identified as ancestry-dependent and ancestry-independent across three SLE ImmunoChip studies (European, African, and Asian Ancestry). To date, this work has yielded a manuscript published in the American Journal of Human Genetics, a submitted manuscript to *Genes*, and two other manuscripts in preparation for submission.
- 2) Prioritization of drug targets from Group 3 (Figure 4) which correspond to protein-coding SLE risk variants. High-quality PDB structure have been identified; and we have also identified an additional source of protein structure (AlphaFold –released July 2021) which will greatly expand docking experiments. This work has enabled development and optimization of processing targets (from Specific Aim 1) which will be beneficial towards ongoing work during the next reporting session.

Significant Results:

Systems biology analysis of SLE risk SNPs across three ancestries

We refer the reader to the previously described (in progress for Specific Aim 1) three manuscripts.

Target Identification from C-Genes (Drug Target Group 3).

SLE-associated SNPs with protein-altering (e.g., nonsynonymous, splice-site, stop-loss) status were collected from previously published SLE ImmunoChip studies, spanning four ancestries (European, African, Hispanic, and Asian). These SLE SNPs map to our Group 2 Targets (**Figure 4**). For each drug target, we generated the protein structure that corresponded to the risk and non-risk alleles. This was accomplished by first evaluating each drug target (gene) for additional protein-altering SNPs in high linkage disequilibrium with the SLE risk SNP using LDlink.[Machiela et al., 2015. Bioinformatics. 2015 Nov 1;31(21):3555-7] and the appropriate reference population based on ancestry. Protein alterations were confirmed with dbSNP and each variant's flanking amino acids were obtained using the UCSC Genome Browser (GRCh37).

Protein structures were retrieved from one of two resources: 1) experimentally derived structures from the Protein Data Bank (PDB) and 2) neural network

derived (predicted) structures from AlphaFold (alphafold.ebi.ac.uk). Computationally derived structures were only used if no suitable experimental structure existed (as described below).

Experimentally derived PDB structures were identified using NCBI for the protein product of the gene of interest. The presence of the relevant amino acid region (i.e., the region containing the SLE-associated variant) within the PDB structures were confirmed by aligning the protein sequences from NCBI and PDB structures, using Clustal Omega [Sievers et al., Mol Syst Biol. 2011 Oct 11;7:539]. For protein products with multiple PDB entries, the PDB model with the best combination of resolution and best representative of biological assembly (i.e., naturally occurring complexed proteins) were selected for drug-modelling (example, **Figure 5**). AlphaFold structures were identified based on Gene accession numbers. Again, the presence of

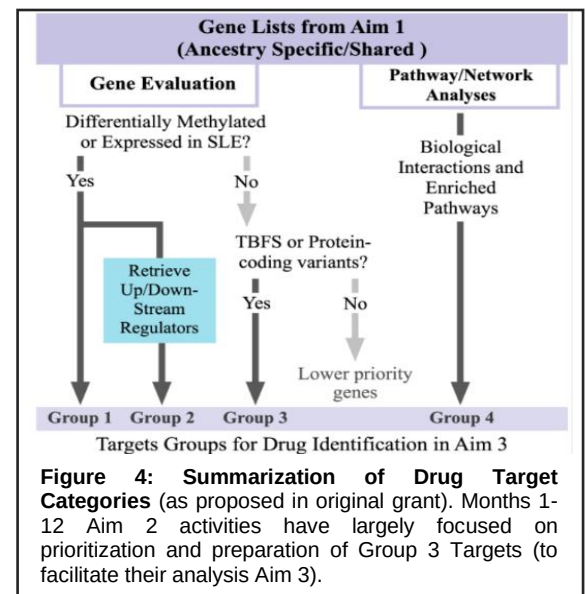


Figure 4: Summarization of Drug Target Categories (as proposed in original grant). Months 1-12 Aim 2 activities have largely focused on prioritization and preparation of Group 3 Targets (to facilitate their analysis Aim 3).

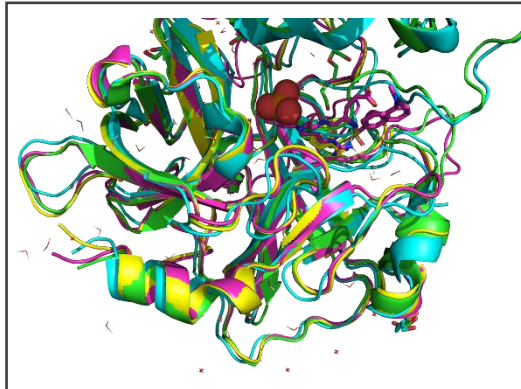


Figure 5. Alignment of C-Terminus for four PLAT PDB structures. When multiple protein structures are available with comparable resolution, the structures are compared for variability. Here, four PDB structures (5ZLZ, 5BRR, 1RTF, 1BDA) for PLAT were aligned and compared. These four structures had differences less than 2 angstroms, from one another, indicating any of these structures would be suitable for molecular docking.

the amino acid region of interest was verified based on alignment with NCBI protein sequence. Confidence measures for the amino acid region of interest were retained for reporting (Example **Figure 6**).

Protein structures (drug targets) were prepared (for docking in Aim 3) by removing the heteroatoms, including water molecules and any co-crystallized ligands/proteins, addition of polar hydrogens and Kollman charges using AutoDock Tools version 1.5.7. (Morris, Huey et al. 2009). The structures were saved as pdbqt files for subsequent docking simulations. Risk variants (alleles) were constructed via *in silico* substitution of amino acid residues using Schrodinger's PyMOL version 2.3.2. (PyMOL | pymol.org). Target-ligand interactions for select docking results were identified and visualized using the ligand site tools in PyMOL.

Summary: All targets related to SLE C-genes (Group 3) have been successfully mapped to their corresponding risk and non-risk amino acids (**Table 4**). All C-gene related targets have had appropriate protein structures identified either by experimentally derived structures (PDB) or AI-predicted structures (AlphaFold –available as of July 2021). Each structure was then optimized for molecular docking (Aim 3) and the corresponding risk structure were also generated (e.g., but substituting the risk allele in the model). *Accomplishments for Molecular Docking are discussed in Aim 3.*

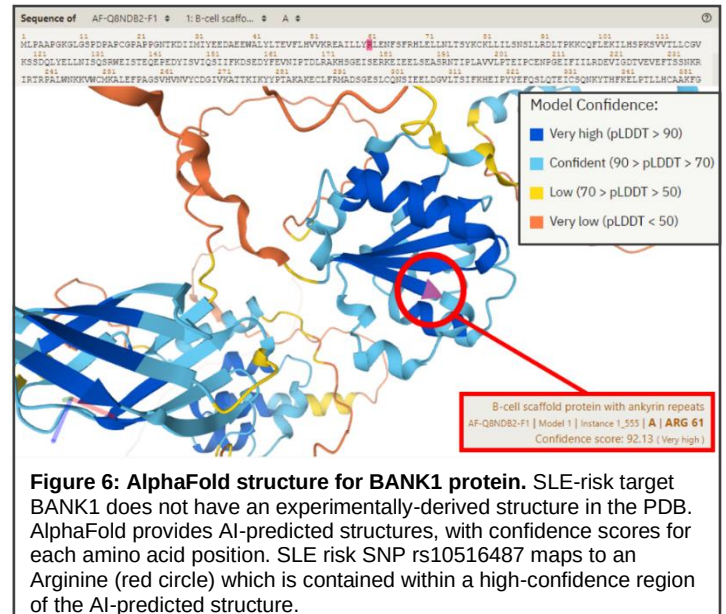


Figure 6: AlphaFold structure for BANK1 protein. SLE-risk target BANK1 does not have an experimentally-derived structure in the PDB. AlphaFold provides AI-predicted structures, with confidence scores for each amino acid position. SLE risk SNP rs10516487 maps to an Arginine (red circle) which is contained within a high-confidence region of the AI-predicted structure.

Table 4: Summary of C-gene based SLE targets for molecular modeling.

cGene	SNP	Chr.	Position (hg38)	Ref	Other	Ref-allele Frequency (cases)	Ref-allele Frequency (controls)	Risk ALLELE (reported)	Risk/Non-risk - AMINO ACID	Ancestry	PDB Structure	AlphaFold ID (UNIPROT)	Molecular Docking Complete
FCGR2A	rs1801274	1	161509955	C	T	0.542	0.518	C	Arg/His	EA	3RY5	P12318	Y
TNFAIP3 (ALIAS=A20)	rs2230926	6	137874929	G	A	0.061	0.033	C	Ser/Phe, Cys/Phe	EA	3DKB	P21580	Y
TYK2	rs34536443	19	10352442	C	A	0.022	0.042	C	Pro/Ala	EA	5WAL	Q9Y2R2	N
PTPN22	rs2476601	1	113834946	A	G	0.127	0.101	A	Arg/Trp	EA & AA		Q8NDB2	N
BANK1	c	4	101829919	T	C	0.199	0.237	C	Arg/His	EA & AA		P11215	N
ITGAM	rs1143679	16	31265490	A	G	0.18	0.108	A	His/Arg	EA & AA		Q9NV64	N
IFIH1	rs35667974	2	162268127	C	T	0.012	0.021	T	Ile/Val	EA	3GA3	Q6BDS2	N
TMEM39A	rs1132200	3	119431989	A	G	0.138	0.159	G	Ala/Thr	EA & AA		Q6BDS2	N
UHRF1BP1	rs11755393	6	34856859	G	A	0.38	0.346	G	Arg/Gln	EA & AA		Q6BDS2	N
UHRF1BP1	rs13205210	6	34864079	C	T	0.129	0.106	C	Thr/Met	EA & AA			N
NT5E (ALIAS=CD73)	rs2229524	6	85489525	G	A	0.069	0.059	C	Thr/Met	EA	7JV8		Y
CR1	rs4844600	1	207505962	A	G	0.203	0.19	A	Glu/Asp	EA	2MCZ		Y
RCAN3	rs196432	1	24535214	A	G	0.455	0.483	G	Arg/Gln	EA & AA		Q9UKA8	N
CDH17	rs1051624	8	94130944	T	G	0.432	0.44	G	Gly/Glu	EA		Q12864	N
QARS	rs11539148	3	49101377	C	T	0.053	0.065	T	Asn/Ser	EA	4YE9		Y
PLAT	rs1804182	8	42176001	A	G	0.042	0.022	A	Ter/Arg	AA	5ZLZ		Y
EML2	rs7252175	19	45629998	T	C	0.135	0.121	T	Phe/Leu	EA		Q95834	N
OTUD3	rs10916668	1	19906593	A	T	0.069	0.065	T	Thr/Ala	EA		Q512D3	N
IDNK	rs1052690	9	83643770	G	T	0.179	0.187	A	Lys/Thr	EA		Q516J7	N
AKAP3	rs1990313	12	4626920	G	A	0.036	0.03	G	Thr/Ile	EA		O75969	N
CR2	rs17617 (note: reported as rs6540433 in iCHIP paper)	1	207480050	C	A	0.168	0.149	C	Gly/Ala	EA		P20023	N
AGBL2	rs7939069	11	47666907	A	G	0.01	0.014	G	---	EA		Q5U5Z8	N
BYSL	rs3828855	6	41932668	A	G	0.077	0.06	A	Ser/Pro	HA		Q13895	N
CCDC187	rs34291323	9	136254577	G	A	0.525	0.481	T	Gly/Arg	HA			N
PRR14	rs3747481	16	30655046	T	C	0.236	0.214	T	Leu/Pro	HA		Q9BWN1	N
ZACN	rs2257020	17	76080334	C	T	0.507	0.533	T	Thr/Ala	HA		Q401N2	N
LRRC34	rs10936600	3	169796797	A	T	0.426 (KR) 0.486 (HC) 0.455 (MC) 0.067 (KR)	0.386 (KR) 0.455 (HC) 0.483 (MC) 0.099 (KR)	A	Ile/Leu	Asian		Q8IZ02	N
ATG16L2	rs11235604	11	72822491	C	T	0.092 (HC) 0.093 (MC)	0.100 (HC) 0.119 (MC)	T	Trp/Arg	Asian		Q8NAA4	N
AP5B1	rs12146493	11	65779862	A	G	0.299 (KR) 0.311 (HC) 0.29 (MC)	0.329 (KR) 0.317 (HC) 0.348 (MC)	G	Leu/Phe	Asian		Q2VPB7	N
AP5B1	rs12362011	11	65779984	A	C	0.208 (KR) 0.216 (HC) 0.192 (MC) 0.195 (KR)	0.246 (KR) 0.219 (HC) 0.254 (MC) 0.262 (KR)	C	Gly/Val	Asian		Q2VPB7	N
IRAK1	rs1059702	X	154018741	G	A	0.169 (HC) 0.168 (MC) 0.421 (KR)	0.217 (HC) 0.211 (MC) 0.361 (KR)	A	Phe/Ser	Asian	6BFN		Y
WDFY4	rs7097397	10	48817351	G	A	0.401 (HC) 0.349 (MC)	0.318 (HC) 0.321 (MC)	G	Arg/Gln	Asian		A0A2R8RXW6	N

Specific Aim 3: Identify and prioritize drugs

Specific Objectives: For identified targets (e.g., proteins translated from SLE risk genes) and biological pathways (e.g., upregulated within SLE-patients), identify FDA-approved drugs and small molecules that show potential evidence of relevant activity via molecular docking.

Major activities: Over months 1-12, activities have focused on developing and optimizing pipelines related to molecular docking and drug identification. We have successfully implemented these pipelines for all cGene targets (Group 3, **Figure 4**) with available (as of September 2021) high quality Protein Databank structures. To date, we have successfully completed over 38,000 in silico docking experiments (11 drug targets (with 2 structures each), each docked with 2,114 FDA-small molecules). These results (drug binding to SLE targets) have been processed into tabular and graphical results summaries, in preparation for activities in the next reporting period (drug prioritization).

Significant Results:

Curation of FDA approved drugs dataset for in silico molecular docking:

The FDA approved drug subset was downloaded from the ZINC 15 database in SDF format (Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model*, 2012). This subset contained 1615 individual molecules, totaling 2114 unique conformers. The SDF files were split and converted to 3D pdbqt structure files using Open Babel version 3.1.1 software (O'Boyle, N. M., et al. (2011). *J. Cheminform* 3: 33.). A database was created which linked the 2,114 ZINC 15 identifiers with other chemical repositories (e.g., SMILES and CIDs) This was implemented via a custom Python 3 script which utilized the pubchempy package [O'Boyle, N. M., et al. (2011). *J. Cheminform* 3: 33]. This database also automatically curated common drug names for each ZINC 15 ID via information from <https://pubchem.ncbi.nlm.nih.gov/> (Example shown in Table 5).

Table 5: Four entries (of 2,114) from the automated ZINC 15 curated database. To facilitate drug assessment (post-molecular docking), the 2,114 FDA-approved small molecules within the ZINC 15 database were linked to common drug identifiers via pubchem.

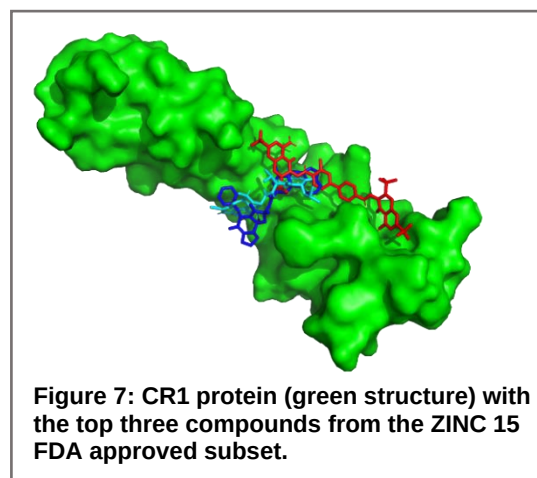
ZINC_ID	SMILES (simplified molecular-input line-entry system)	PUBCHEM_ID	ALT_NAME_1	ALT_NAME_2	ALT_NAME_3	ALT_NAME_4
ZINC000000001382	<chem>C=C(CC)C(=O)c1ccc(OCC(=O)O)c(Cl)c1Cl</chem>	3278	ETHACRYNIC ACID	Etacrynic Acid	58-54-8	ethacrynate
ZINC000005179119	<chem>C/C(O)=N/S(=O)(=O)c1ccc(N)cc1</chem>	5320	sulfacetamide	144-80-9	Sulphacetamide	Acetosulfamine
ZINC00000538273	<chem>C[C@H]1COc2c(N3CCN(C)CC3)c(F)cc3c(=O)c(C(=O)O)cn1c23</chem>	149096	LEVOFLOXACIN	100986-85-4	Levaquin	Quixin
ZINC000003830339	<chem>CC(C)(C)NC[C@H](O)COc1cccc2c1CCCC2=O</chem>	39468	LEVOBUNOLOL	47141-42-4	Levobunololum	Betagan

Docking and predicted binding affinity evaluation:

Ridged docking simulations of the FDA library were performed using Autodock Vina (Morris, Huey et al. 2009) via a Strawberry Perl ([Strawberry Perl for Windows](#)) script and the resulting output log files were appended to a compiled CSV file for analysis using a BASH ([Bash - GNU Project - Free Software Foundation](#)) script in Ubuntu ([Enterprise Open Source and Linux | Ubuntu](#)). The ligand libraries were ranked by the predicted affinity (ΔG_p), and to identify compounds with a greater affinity for the risk allele, compared using the following formula:

$$\Delta\Delta G_p = |\Delta G_{Risk}| - |\Delta G_{non-Risk}|$$

Predicted affinities were calculated on a high-throughput basis with custom R scripts and merged with the previously described database of common drug name identifiers. Output summaries included four metrics of information related to the *in silico* molecular docking (per FDA-approved small molecule): 1) ΔG_p for the protein structure containing the SLE-risk amino acid; 2) ΔG_p for the protein structure containing the non-risk SLE-risk amino acid; 3) $\Delta\Delta G_p$ to capture docking affinity between the two protein structures; 4) $|\Delta\Delta G_p|$ to capture the magnitude of change between predicted affinities for the risk and non-risk protein structures.



Molecular Docking Results:

We have completed all the cGenes that had available high-quality PDB structures (**See Table 4**). With the release of the alpha-fold database, we will continue to evaluate the remaining genes identified from the genomic analysis in Target Group 3 (**See Figure 4**) and other target groups as they progress.

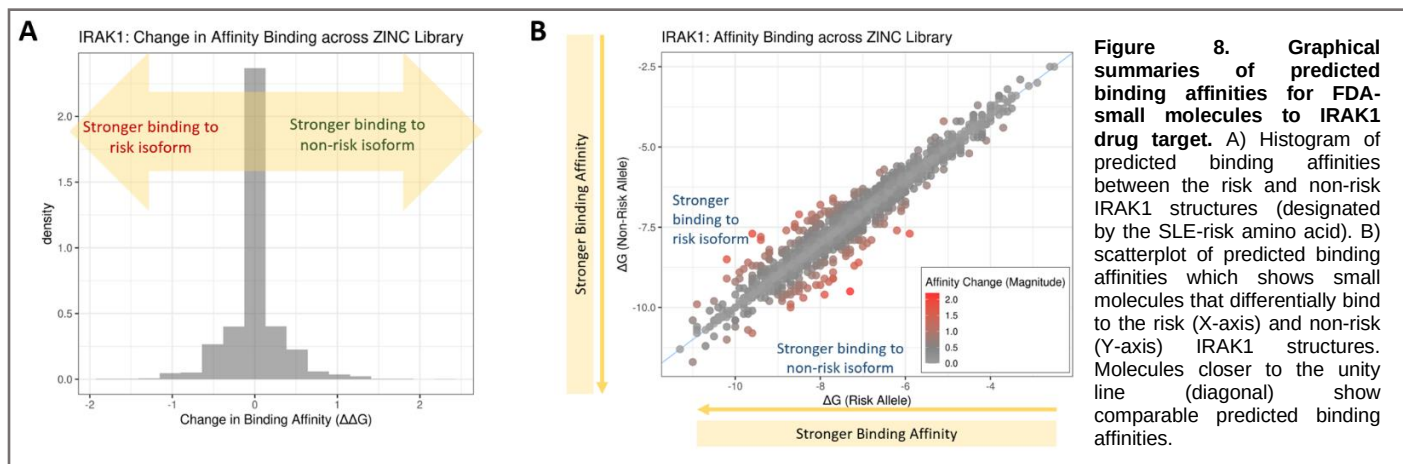
For each drug target that completed the molecular docking pipeline, we have generated the following results:

- 1) Binding affinities for the risk and non-risk protein structures (ΔG_p).
- 2) Changes in predicted binding affinity between the risk and non-risk structures (summarized as $\Delta\Delta G_p$ and $|\Delta\Delta G_p|$).
- 3) Common names and identifiers for each ZINC 15 conformer.
- 4) Binding locale of small molecules to the drug target. (**Example shown in Figure 7**).

These results have been summarized in both tabular (**Example in Table 6**) and graphical formats (**Figure 8**).

Table 6: Summaries of molecular docking predicted affinities for 8 (out of 2,114) conformers from the ZINC 15 database with IRAK1 drug target. Subset of results from IRAK1 molecular docking analyses. Two (one risk, one non-risk) protein structures of IRAK1 were used for molecular docking, predicted binding affinities and changes in binding affinity (based on the substitution for the risk amino acid) are shown. Common identifiers for the ZINC 15 IDs are also shown.

Drug Target	ZINC 15 ID	Predicted Affinity (ΔG_p)		Change in Affinity		Alternate Identifiers linked to ZINC 15 ID			
		Risk	Non-Risk	$\Delta\Delta G_p$	$ \Delta\Delta G_p $	SMILES* (Simplified Molecular-Input Line-Entry System)	PUBCHEM ID	Alternate Name 1	Alternate Name 2
IRAK1	ZINC000005844792	-7.3	-9.5	2.2	2.2	<chem>O[C@@H](CNC[C@H](O)[C@@H]1CCc2cc(F)ccc2O1)[C@@H]1CCc2cc(F)ccc2O1</chem>	189563	(-)-Nebivolol	118457-16-2
IRAK1	ZINC000003993846	-9.6	-7.7	-1.9	1.9	<chem>C/C(=C/c1csc(C)n1)[C@@H]1C[C@H]2O[C@]2(C)CCC[C@H](C)[C@H](O)[C@@H](C)C(=O)C(C)[C@@H](O)CC(=O)N1</chem>	6445540	Ixabepilone	219989-84-1
IRAK1	ZINC000038945666	-5.9	-7.7	1.8	1.8	<chem>O=C(CCCc1ccccc1)OCC(COC(=O)CCc1ccccc1)OC(=O)CCc1ccccc1</chem>	10482134	GLYCEROL PHENYLBU TYRATE	611168-24-2
IRAK1	ZINC000003989268	-7.9	-9.6	1.7	1.7	<chem>CCO/N=C(\C(=O)N[C@@H]1C(=O)N2C(C(=O)O)=C(Sc3nc(-c4cc[n+](C)cc4)cs3)CS[C@H]12)c1nsc(NP(=O)(O)O)n1</chem>	10033059	SCHEMBL1 7508872	ZINC3989268
IRAK1	ZINC000053683151	-10.2	-8.5	-1.7	1.7	<chem>CC(C)[C@H]1C(=O)N2CCC[C@H]2[C@]2(O)O[C@](NC(=O)[C@@H]3C=C4C5cccc6[nH]c(Br)c(c56)C[C@H]4N(C)C3)(C(C)C(=O)N12</chem>	31101	bromocriptine	Bromocryptine
IRAK1	ZINC000003914596	-9.4	-7.8	-1.6	1.6	<chem>CC(C)(C)NC(=O)[C@@H]1C[C@H]2CCCC[C@H]2CN1C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)[C@@H](CC(N)=O)NC(=O)c1ccc2ccccc2n1</chem>	441243	SAQUINAVIR	127779-20-8
IRAK1	ZINC000004097383	-9.4	-7.9	-1.5	1.5	<chem>CC(=O)O[C@H]1C[C@@H]2CC[C@H]3[C@H](C)[C@@H]4(C)[C@H]3C[C@H](N+3(C)CCCC3)[C@@H]4OC(C=O)[C@@]2(C)C[C@H]1N+1(C)CCCC1</chem>	441289	pancuronium	Pavulon
IRAK1	ZINC000004097404	-7.1	-8.6	1.5	1.5	<chem>CC(=O)O[C@H]1C[C@@H]2CC[C@H]3[C@H](C)[C@@H]4(C)[C@H]3C[C@H](N+3(C)CCCC3)[C@@H]4OC(C=O)[C@@]2(C)C[C@H]1N1CCCC1</chem>	39765	Vecuronium	86029-43-8



Optimized analysis and summarization of in silico molecular docking has generated lists of drug binding affinities with SLE drug targets. During the next reporting stage (13-24 months), we will continue to evaluate drug targets from the other target groups and develop high-throughput prioritization of identified drugs (e.g., CoLTs score criteria).

Returning to the submitted paper, *Nucleic acid sensing and interferon-inducible pathways show differential methylation in MZ twins discordant for lupus and overexpression in independent lupus samples: Implications*

for pathogenic mechanism and drug targeting. The differentially methylated – differentially expressed (DM-DE) genes identified in Table 3 above were also informatically queried for potential gene-drug interactions, yielding a list of 41 drugs including a known SLE therapy (Table 7). The DM-DE genes delineate two important biologic pathways which are not only reflective of the heterogeneity of SLE but may also correlate with distinct IFN responses that depend on the source, type, and location of nucleic acid molecules and the activated receptors in individual patients. Cell- and tissue-specific analyses will be critical to the understanding of genetic factors dysregulating the nucleic acid sensing and IFN pathways, and whether those factors could be appropriate targets for therapeutic intervention.

Table 7. Predicted drugs targeting hypomethylated genes and associated pathways with $\Delta\beta < -0.085$.

CpG*	Chr	Pos(bp) [†]	Gene	Mean $\Delta\beta$	P-value	STITCH[43]	IPA [‡]	DGIdb[44]
cg16526047	1	949893	ISG15	-0.11	1.28x10 ⁻⁴			Irinotecan ^F
cg10959651	2	7018020	RSAD2	-0.13	3.14x10 ⁻¹⁴	Fludarabine ^F		
cg14126601	2	37384708	EIF2AK2	-0.1	5.55x10 ⁻¹⁶			Indirubin derivative E804
cg15768138	2	219030752	CXCR1	-0.11	7.38x10 ⁻²⁷	Reparixin ^D	Reparixin ^D	SCH-527123, Ketoprofen ^F
cg06981309	3	146260954	PLSCR1	-0.24	6.41x10 ⁻³¹	Wogonin ^G		
cg15065340	3	195632915	TNK2	-0.16	4.04x10 ⁻³	Dasatinib ^{1 F}	Osimertinib ^F , Vemurafenib ^F	Debromohymenialdisine
cg08099136	6	32811251	PSMB8	-0.11	1.43x10 ⁻⁴	Carfilzomib ⁴ , Oprozomib ^D , Bortezomib ^{6 F}	Carfilzomib ^{4F}	Carfilzomib ^{4 F} ,
cg00052684	6	35694245	FKBP5	-0.16	1.65x10 ⁻³	Rapamycin/Sirolimus ^{2 F} , Tacrolimus ^{5 F}		Venlafaxine ^F , Clomipramine ^F
cg14864167	8	66751182	PDE7A	-0.35	1.21x10 ⁻⁹			
cg12110437	8	144098888	LY6E	-0.2	3.14x10 ⁻⁹		DLYE5953A ^D	Ketotifen ^F , Dyphylline ^F
cg06188083	10	91093005	IFIT3	-0.25	6.18x10 ⁻⁸	Imidazoles ^D		
cg08926253	11	614761	IRF7	-0.17	2.01x10 ⁻⁹	Hesperidin ^D		
cg03753191	13	43566902	EPSTI1	-0.1	9.23x10 ⁻⁵	Methotrexate ^{F T} , Vinblastine ^F , Doxorubicin ^F , Cisplatin ^F		
cg00246969	13	99159656	STK24	-0.11	6.26x10 ⁻⁶	Staurosporine ^D		
cg23378941	17	64361956	PRKCA	-0.11	6.89x10 ⁻⁵	Staurosporine ^D	Aprinocarsen	Midostaurin ^F , Enzastaurin ^D , Quercetin ^{D G} , Aprinocarsen, Ruboxistaurin ^D , Ingenol Mebutate ^{FW} , Bryostatins ^D , Sotrastaurin Acetate ^D , Tamoxifen ^{2 F}
cg07839313	19	17514600	BST2	-0.12	3.48x10 ⁻³	Resveratrol ^{6 D G}		
cg21549285	21	42799141	MX1	-0.47	6.59x10 ⁻¹³	Mitomycin C ^F , Colchicine ^F		
cg19460508	22	44422195	PARVB	-0.1	1.64x10 ⁻³	Lovastatin ^{3 F}		
Bortezomib ^{6 F}								

*CpGs with $P < 1 \times 10^{-3}$ and $\Delta\beta < -0.085$. [†]Positions are from Build 37.

[‡] Qiagen Bioinformatics: ingenuity.com ^F: FDA approved.

^D: Ongoing clinical trial or DiD ^G: GRAS. ^T: Known utility in lupus therapy. ^{FW}: Ingenol mebutate is FDA-approved in the US but withdrawn in the EU. CoLTS scores in superscript (range: -16 to +11)

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

Mr. Haoqing Wu, an undergraduate research student, completed an internship under the direction of Dr. Reeves where he completed the molecular docking analysis of FDR-approved drugs for Complement Receptor Protein 1 (CR1). Analysis included preparation of drug derivatives using ZINC 15 database as SMILES string; preparation of protein target from the Protein Data Bank; compute Gibbs Free Energy for each docking pose and conformation; and summarize and interpret these results. Mr. Wu is completing his undergraduate work and plans to pursue graduate studies in chemistry.

Mr. Nolan H. Hamilton just completed his undergraduate degree (BA mathematics) and started working in Dr. Langefeld's lab on a one-year study in preparation for graduate school; he has been accepted by University of North Carolina Chapel Hill Bioinformatics graduate program. For this grant, Mr. Nolan has assisted in the preparation of drug derivatives from the ZINC 15 database as SMILES strings for all of the cGenes. This has allowed Drs. Ainsworth and Langefeld to mentor him in learning basic concepts of biostatistics, statistical genetics, and genetics.

How were the results disseminated to communities of interest?

We have published two papers, submitted a third paper, and are in the process on completing two more papers.

Published (See Appendix):

- Owen KA, Price A, Ainsworth H, Aidukaitis BN, Bachali P, Catalina MD, Dittman JM, Howard TD, Kingsmore KM, Labonte AC, Marion MC, Robl RD, Zimmerman KD, Langefeld CD, Grammer AC, Lipsky PE. Analysis of Trans-Ancestral SLE Risk Loci Identifies Unique Biologic Networks and Drug Targets in African and European Ancestries. *Am J Hum Genet.* 2020 Nov 5;107(5):864-881. PMID: PMC7675009.
- 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.

Submitted to Genes (See Appendix)

- Marion MC, Ramos PS, Bachali P, Labonte AC, Zimmerman KD, Ainsworth AC, Heuer SE, Robl RD, Catalina MD, Kelly JA, Howard TD, Lipsky PE, Grammar AC, Langefeld CD. *Nucleic acid sensing and interferon-inducible pathways show differential methylation in MZ twins discordant for lupus and overexpression in independent lupus samples: Implications for pathogenic mechanism and drug targeting.*

We had two presentations at the LUPUS 2021 meeting

- Kate A. Owen, *Understanding the Influence of Ancestry on SLE Disease Mechanisms.*
- Carl D. Langefeld, *Ancestry's Impact on SLE Phenotypes*

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

During the next reporting period (months 13-24), we will build upon the first year's success by continuing progress towards milestones in each Aim. We will continue to use the pipelines developed and optimized in months 1-12 (e.g., summarization of ΔG and $\Delta\Delta G$ from molecular docking); and we will also work to develop a high-throughput method for drug prioritization (Aim 3, subtask 3). **Table 8** outlines planned steps and work for the next reporting period. Planned activities are shown by each specific task and subtask as provided in the Project's original SOW.

Table 8. Goals and milestones

Goals and Milestones as listed in the original SOW.		Progress Report 10-2021 Update. Planned Activities for next Period
Specific Aims (specified in proposal)	Timeline	
Specific Aim 1: Identify SLE-risk Genes	Months	
Subtask 1: Identify SLE-risk single nucleotide polymorphisms (SNPs) in women	1-24	• Continue curation of SLE risk SNPs across relevant sources of data as they become available.
Subtask 2: Link SLE-risk SNPs to genes via eQTL, proximity, transcription factor binding, protein coding, gene-based testing	3-27	• Evaluate gene-linking for SNPs associations unique to SLE studies within patients of Hispanic ancestry. • Evaluate additional links to genes with updated ENCODE regulatory information.(e.g., ensuring most plausible gene link for TFBS SNPs). • Gene-based testing
Subtask 3: Transcriptomic analysis, differential expression of genes identified in subtask 2	3-30	• Continue differential expression analysis for genes not yet analyzed from Subtask 2.
Subtask 4: DNA methylation analyses, differential methylation of genes identified in subtask 2	1-24	• Analysis of differential methylation (in publically available datasets) for genes identified in Subtask 2.
Subtask 5: Identify and write potential manuscripts on multi-omic analysis of SLE-risk associated variants and genes.	6-36	• Submit 2nd manuscript on differential pathways by ancestry. • Submit differential analysis paper on Methylation within discordant twins.

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

4. Impact

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

Analytic and programming pipelines have established methods that can be applied to other studies which are exploring precision medicine therapeutics based on disease-associated risk loci. For example, the ZINC 15 database is an established resource of FDA-approved small molecules. The development of efficient annotation of ZINC 15 compounds with publicly available datasets (e.g., SMILES, PubChem) enables efficient application of these methods, representing a reduction of preparatory data steps needed. Furthermore, many of the identified drug targets originate from genes that are implicated in other diseases (e.g., autoimmune diseases). Thus, molecular docking of FDA-approved compounds to these targets will be valuable datasets for similar studies in other diseases. Further, these pipelines will enable exploration of precision medicine for other lupus phenotypes and other ancestral groups.

What was the impact on other disciplines?

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

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

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

In identifying suitable drug targets for molecular docking, we encountered a challenge due to the limitation of high-quality experimentally derived protein structure (e.g., via the Protein Databank). For example, of the 25 cGENE-based drug targets, only 9 had suitable PDB structures available. As of July 2021, AlphaFold was officially released which contains neural network derived (predicted) structures. Inclusion of this database increased our coverage of drug targets to 100%. We will continue to use experimentally-derived structures when available; but the inclusion of AlphaFold presents an additional opportunity to explore more of our identified targets. We have also updated our protocols to report the prediction confidence of the protein structure at the site of the SLE-risk amino acid (within the AlphaFold predicted structure).

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

The Covid-19 pandemic has caused multiple problems, including the restriction from meeting in person and the inability to recruit potential interns or PhD students to the lab. The Wake Forest investigators have now largely converted to laptops, and we have been effective in using Zoom, Microsoft Teams and WebEx. We have only successfully recruited one intern and no PhD. students. Assuming the pandemic eases, we anticipate recruiting an intern and potentially a student this summer.

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, Price A, Ainsworth H, Aidukaitis BN, Bachali P, Catalina MD, Dittman JM, Howard TD, Kingsmore KM, Labonte AC, Marion MC, Robl RD, Zimmerman KD, Langefeld CD, Grammer AC, Lipsky PE. Analysis of Trans-Ancestral SLE Risk Loci Identifies Unique Biologic Networks and Drug Targets in African and European Ancestries. *Am J Hum Genet.* 2020 Nov 5;107(5):864-881. PMID: PMC7675009.
- 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.

Submitted to the journal *Genes* (See Appendix)

- Marion MC, Ramos PS, Bachali P, Labonte AC, Zimmerman KD, Ainsworth AC, Heuer SE, Robl RD, Catalina MD, Kelly JA, Howard TD, Lipsky PE, Grammar AC, Langefeld CD. *Nucleic acid sensing and interferon-inducible pathways show differential methylation in MZ twins discordant for lupus and overexpression in independent lupus samples: Implications for pathogenic mechanism and drug targeting.*

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

Nothing to Report.

Other publications, conference papers, and presentations.

We had two presentations at the LUPUS 2021 meeting

- Kate A. Owen, *Understanding the Influence of Ancestry on SLE Disease Mechanisms.*
- Carl D. Langefeld, *Ancestry's Impact on SLE Phenotypes*

Website(s) or other Internet site(s)

Nothing to Report.

Technologies or techniques

Please see:

- 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.

Inventions, patent applications, and/or licenses

Nothing to Report.

Other Products

Other products produced by this work in the first year include:

1. Database of ZINC 15 IDs matched to common database (e.g. pubCHEM) identifiers and common drug names.
2. Growing database of gene-linked SLE risk SNPs by ancestry (e.g., linked to most plausible functional gene via gene expression or nonsynonymous variants).
3. Growing database of target-specific molecular docking for FDA-approved small molecules. Currently this dataset will feed into drug prioritization for SLE therapeutics, but since many of the identified drug targets are relevant in other autoimmune diseases, this could eventually be evaluated in the context of other diseases.
4. Hierarchicell: an R-package for estimating power for tests of differential expression with single-cell data.

7. Participants & Other Collaborating Organizations

What individuals have worked on the project?

Wake Forest investigative team

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

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

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

	binding experiments.
--	----------------------

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

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

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

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

Name:	Jessica Kane
Project Role:	Data analyst
Researcher Identifier (e.g. ORCID ID):	NA
Nearest person month worked:	4 calendar months (33% effort)
Contribution to Project:	Data analyst working primarily on genetic basis of the diversity of lupus as manifested in specific ancestries and individual patient phenotypes
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, 2020 through September 30, 2021

Carl D. Langefeld, PhD

ACTIVE

Grant #: W81XWH-21-1-0075

Grant Title: Epigenetic modulation of SOS1 gene promotes breast cancer disparity by altering immune landscape

Effort: 4%

Grant #: R01 DK128871

Grant Title: Identification and Characterization of Loci Associate with Non-alcoholic Fatty Liver Disease

Effort: 5%

COMPLETED**Grant #:** R01 NS093870**Grant Title:** Recurrent Hemorrhagic Stroke in Minority Populations**Effort:** 5%**Grant #:** W81XWH-18-1-0050**Grant Title:** Modulation of Epithelial Biomarkers of Breast Cancer Risk by a Community-Based Intervention for Diabetes Prevention**Effort:** 2%**Grant #:** U01NS036695**Grant Title:** Genetic and Environmental Risk Factors for Hemorrhagic Stroke**Effort:** 7%**Grant #:** W81XWH-19-1-0062**Grant Title:** Effects of Ionizing Radiation (IR) on Myocardial DNA Methylation Profiles in Relation to Cardiomyopathy in a Non-Human Primate Model**Effort:** 4%**Grant #:** R01 AI139126**Grant Title:** Early Life Factors, Gene-Environment Interaction and Eosinophilic Esophagitis**Effort:** 5%**Grant #:** R01 DK118243**Grant Title:** Identifying Genetic Drivers of Expression Networks Causing Insulin Resistance**Effort:** 4%**Grant #:** 1-18-ICTS-113**Grant Title:** Altered DNA methylation modulates adipose tissue transcript expression and insulin resistance in African Americans**Effort:** 2%**Grant #:** No Number (Funded by the Alliance for Lupus Research)**Grant Title:** Determinants and Predictors of Lupus Patients Engaging in Clinical Trials**Effort:** 5%**Tony Reeves, PhD****ACTIVE****Grant #:** National Science Foundation EIR CHE-2101221**Grant Title:** Development of Novel PET Tracers**Effort:** 0.6 CM**Grant #:** Wake Forest/Atrium CTSI Pilot Award**Grant Title:** Development of FABP4 Inhibitors**Effort:** 0.6 CM**COMPLETED****Grant #:** R21NS084899**Grant Title:** Formulation and Encapsulation of Enzymic Countermeasures against Organophosphorus Nerve Agents**Effort:** 4.0 CM

ACTIVE

Grant #: R21 AG073733

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

Effort: 10%

Grant #: 1R01 ES031940-01A1

Grant Title: Epidemiologic Identification and Mechanistic Investigation of Early Life Environmental Risk Factors for Eosinophilic Esophagitis

Effort: 3%

COMPLETED

Grant #: 1R01HL122393-01A1

Grant Title: Social Stress, Diet, and Primate Monocyte Programming in Cardiovascular Risk

Effort: 5%

Grant #: W81XWH-19-1-0062

Grant Title: Effects of Ionizing Radiation (IR) on Myocardial and Peripheral DNA Methylation Profiles in Relation to Cardiomyopathy in Non-Human Primate Model

Effort: 5%

Grant #: 1R21 DK122297-01A1

Grant Title: Molecular and Epigenetic Predictors and Mechanisms of Treatment Response to Topical Steroids in Eosinophilic Esophagitis

Effort: 2%

Grant #: American Diabetes Association

Grant Title: Altered DNA Methylation Modulates Adipose Tissue Transcript Expression and Insulin Resistance in African Americans

Effort: 4%

Grant #: R01 HL135009-01

Grant Title: A longitudinal Epigenetic Study of Atherosclerosis

Effort: 2%

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

- I. Owen KA, Price A, Ainsworth H, Aidukaitis BN, Bachali P, Catalina MD, Dittman JM, Howard TD, Kingsmore KM, Labonte AC, Marion MC, Robl RD, Zimmerman KD, Langefeld CD, Grammer AC, Lipsky PE. Analysis of Trans-Ancestral SLE Risk Loci Identifies Unique Biologic Networks and Drug Targets in African and European Ancestries. *Am J Hum Genet.* 2020 Nov 5;107(5):864-881. PMID: PMC7675009.
- II. 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.
- III. Marion MC, Ramos PS, Bachali P, Labonte AC, Zimmerman KD, Ainsworth AC, Heuer SE, Robl RD, Catalina MD, Kelly JA, Howard TD, Lipsky PE, Grammar AC, Langefeld CD. *Nucleic acid sensing and interferon-inducible pathways show differential methylation in MZ twins discordant for lupus and overexpression in independent lupus samples: Implications for pathogenic mechanism and drug targeting.* Submitted Genes.

Analysis of Trans-Ancestral SLE Risk Loci Identifies Unique Biologic Networks and Drug Targets in African and European Ancestries

Katherine A. Owen,^{1,*} Andrew Price,¹ Hannah Ainsworth,² Bryce N. Aidukaitis,¹ Prathyusha Bachali,¹ Michelle D. Catalina,¹ James M. Dittman,¹ Timothy D. Howard,² Kathryn M. Kingsmore,¹ Adam C. Labonte,¹ Miranda C. Marion,² Robert D. Robl,¹ Kip D. Zimmerman,² Carl D. Langefeld,² Amrie C. Grammer,^{1,3} and Peter E. Lipsky^{1,3}

Summary

Systemic lupus erythematosus (SLE) is a multi-organ autoimmune disorder with a prominent genetic component. Individuals of African ancestry (AA) experience the disease more severely and with an increased co-morbidity burden compared to European ancestry (EA) populations. We hypothesize that the disparities in disease prevalence, activity, and response to standard medications between AA and EA populations is partially conferred by genomic influences on biological pathways. To address this, we applied a comprehensive approach to identify all genes predicted from SNP-associated risk loci detected with the Immunochip. By combining genes predicted via eQTL analysis, as well as those predicted from base-pair changes in intergenic enhancer sites, coding-region variants, and SNP-gene proximity, we were able to identify 1,731 potential ancestry-specific and trans-ancestry genetic drivers of SLE. Gene associations were linked to upstream and downstream regulators using connectivity mapping, and predicted biological pathways were mined for candidate drug targets. Examination of trans-ancestral pathways reflect the well-defined role for interferons in SLE and revealed pathways associated with tissue repair and remodeling. EA-dominant genetic drivers were more often associated with innate immune and myeloid cell function pathways, whereas AA-dominant pathways mirror clinical findings in AA subjects, suggesting disease progression is driven by aberrant B cell activity accompanied by ER stress and metabolic dysfunction. Finally, potential ancestry-specific and non-specific drug candidates were identified. The integration of all SLE SNP-predicted genes into functional pathways revealed critical molecular pathways representative of each population, underscoring the influence of ancestry on disease mechanism and also providing key insight for therapeutic selection.

Introduction

Systemic lupus erythematosus (SLE) (MIM: 152700) is a multi-organ autoimmune disorder associated with significant morbidity and mortality. SLE is strongly influenced by genetic factors and recent candidate gene, Immunochip, and genome-wide association studies (GWASs) have identified more than 100 SLE susceptibility loci.^{1–6} However, disease development is complex and unpredictable, with considerable clinical heterogeneity among ancestral groups. Specifically, individuals of African ancestry (AA) experience more severe disease and more co-morbidities compared to European ancestry (EA) populations.^{7–9} Furthermore, there seems to be variability in the response of individuals within different ancestral groups to standard medications, including cyclophosphamide, mycophenolate, rituximab, and belimumab. For example, belimumab, a monoclonal antibody directed to TNFSF13B exhibiting clinical benefit in moderately active SLE, was reported to be less effective in treating AA populations.^{10–13}

Although GWASs have achieved great success in mapping disease loci in polygenic autoimmune diseases, the majority of GWAS findings have failed to impact clinical

practice.¹⁴ Moreover, for many single-nucleotide polymorphisms (SNPs), the biologic implications have usually not been identified. Here we hypothesize that using the more global approach of identifying all of the genes implicated by GWASs and modeling them into biologic pathways might provide a broader view of the impact of genetics on SLE, and, also, an indication of the disparate genetic influences manifest in affected individuals of different ancestries. Utilization of expression quantitative trait loci (eQTL) mapping^{15–17} as well as identification of variants disrupting transcription factor binding site (TFBS) occupancy in active regulatory regions resulted in the identification of genes implicated by ancestry-specific and trans-ancestral SNP associations. We then applied a comprehensive systems biology approach to predict SLE-associated biological pathways. Putative pathways were validated by connectivity mapping to differentially expressed genes (DEGs) in SLE and candidate treatments for each ancestral group were identified. Together, these genetic and gene expression analyses have identified biological pathways common to both EA and AA, as well as those implicated by the differential strength of the ancestry-specific association as

¹AMPEL BioSolutions LLC, Charlottesville, VA 22902, USA; ²Wake Forest School of Medicine, Winston-Salem, NC 27109, USA

³These authors contributed equally

*Correspondence: kate.owen@ampelbiosolutions.com

<https://doi.org/10.1016/j.ajhg.2020.09.007>.

© 2020 American Society of Human Genetics.



more dominant in either ancestry, and have helped identify novel drug candidates that might uniquely impact EA and AA SLE.

Material and Methods

Identification of SLE-Associated SNPs and Predicted Genes

The SLE Immunochip study¹⁸ identified single-nucleotide polymorphisms (SNPs) significantly associated with SLE in AA (2,970 case subjects, 2,452 control subjects) and EA (6,748 case subjects, 11,516 control subjects) cohorts. SNP proxies, identified via rAggr (see [Web Resources](#)) in linkage disequilibrium (LD) ($r^2 > 0.5$) with these SLE-associated SNPs, were then determined, using the Central European Utah (CEU) population as background for EA SNPs and the Yoruban (YRI) population for AA SNPs. Expression quantitative trait loci (eQTLs) were then identified as previously described¹⁸ using GTEx v.6¹⁹ and the Blood eQTL browser database²⁰ and mapped to their associated eQTL expression genes (E-Genes). In parallel, random E-Gene datasets were generated from randomly selected SLE Immunochip SNPs using the same methodology. SNP proxies were then queried by GTEx to generate eQTLs and matched to ENSEMBL gene IDs. To find SNPs in enhancers and promoters and their associated transcription factors and downstream target genes (T-Genes), we queried the atlas of Human Active Enhancers to interpret Regulatory variants (HACER)²¹ and the GeneHancer database.²² 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 SIFT4G,^{23,24} PolyPhen-2,²⁵ PROVEAN,²⁶ and PANTHER.²⁷ All other SNPs were linked to the most proximal gene (P-Gene) or gene region as previously detailed.¹⁸ For overlap studies, Venn diagrams were computed and visualized using InteractiVenn.²⁸ All predicted genes were divided into an AA, EA, or shared group depending on the ancestral designation of the original SLE-associated SNP.

Statistical Analysis

The single-locus and multi-locus ancestry-specific tests of association within each ancestral group have been previously reported.¹⁸ Specifically, to test for an association between a SNP and case/control status separately for the AA and EA ancestries, logistic regression models were computed adjusting for population substructure using admixture factors as covariates. The Benjamini-Hochberg false discovery rate adjusted p values (p_{FDR}) were computed, and SNPs were considered for subsequent analyses in this manuscript if they met a $p_{FDR} < 0.05$ threshold. The two ancestry-specific analyses (i.e., AA and EA) were meta-analyzed using the weighted inverse normal (weighted by sample size) method and tested for heterogeneity also as previously described (p_{HET})¹⁸. The following algorithm was used to classify significant associations in either ancestral group ($p_{FDR} < 0.05$) as shared or ancestry specific (i.e., primarily driven by the EA or AA ancestry subpopulations). First, if the $p_{HET} > 0.01$ then the association was considered common (shared) across the EA and AA ancestries. If the $p_{HET} < 0.01$, then we considered the direction (odds ratio: $OR > 1$, $OR < 1$) and the ancestry-specific p values. If $p_{HET} < 0.01$ and the OR was in the same direction with suggestive evidence of association ($p < 0.05$; not FDR adjusted), then the association was considered

shared. If $p_{HET} < 0.01$ and the OR was in the same or opposite directions without at least suggestive evidence of association in both populations ($p < 0.05$), then the association was considered ancestry specific and driven by the ancestry with the significant association ($p_{FDR} < 0.05$). Finally, if $p_{HET} < 0.01$ but the associations were significant and in opposite directions, the association was considered shared (noting the ancestry-specific direction of the associations). Graphpad PRISM 8.0 was used to perform mean, 95% confidence intervals, and unpaired t test with Welch's correction.

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, introns, and intergenic regions. Regulatory regions include transcription factor binding sites (TFBS), promoters, enhancers, repressors, promoter flanking regions, and open chromatin. Coding regions were broken down further and include 5' UTRs, 3' UTRs, and synonymous and non-synonymous (missense and nonsense) mutations. The online resource tool HaploReg (v.4.1)²⁹ was also used to identify DNA features and regulatory elements and to assess regulatory potential.

Differential Expression Analysis of E-Genes

Predicted genes were compared to multiple differential expression datasets, as summarized in [Table S1](#). These datasets include the log-fold changes of all genes with significant ($FDR < 0.2$) differential expression in whole blood (WB), peripheral blood mononuclear cells (PBMC), B cells, T cells, myeloid cells, synovium, skin, kidney glomerulus (G), and kidney tubulointerstitium (TI). The FDR was selected *a priori* to avoid excluding false negatives from the analysis. Cohorts are SLE versus control (CTL) unless noted otherwise. Additional cohorts include SLE synovium versus osteoarthritis (OA) synovium, discoid lupus erythematosus (DLE) skin versus CTL skin, and subacute cutaneous lupus erythematosus (CLE) skin versus CTL skin. Datasets include GEO: GSE88884 (Illuminate 1 and 2), GSE49454, GSE22908, GSE61635, GSE29536, GSE39088, GSE50772, FDABMC3, EMTAB2713, GSE10325, GSE4588, GSE38351, GSE36700, GSE52471, GSE72535, GSE81071, and GSE32591.

Functional Gene Set Analysis and Identification of Upstream Regulators (UPRs)

For both ancestral groups, predicted gene lists were examined using Biologically Informed Gene Clustering (BIG-C; v.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 pathways, NCBI, PubMed, and the Interactome and has been previously described.^{30,31}

I-Scope is a custom clustering tool used to identify immune infiltrates in large gene datasets and has been described previously.³² 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 cells, CD4 T cells, CD8 T cells, gamma-delta T cells, NK/NKT cells, T and B cells, B cells, activated B cells, T

and B and monocytes, monocytes and 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 the Database for Annotation, Visualization and Integrated Discovery (DAVID) and the Ingenuity Pathway Analysis (IPA) platform provided additional genetic pathway identification. IPA upstream regulator (UPR) analysis was also used to identify potential transcription factors, cytokines, chemokines, etc. that can contribute to the observed gene expression pattern in the input dataset.

Network Analysis and Visualization

Visualization of protein-protein interaction and relationships between genes within datasets was done using Cytoscape (v.3.6.1) software. Briefly, STRING (v.1.3.2)-generated networks were imported into Cytoscape (v.3.8.1) and partitioned with MCODE via the clusterMaker2 (v.1.2.1) plugin. The resulting networks were further simplified into metastructures defined by the number of genes in each cluster, the number of significant intra-cluster connections predicted by MCODE, and the strength of associations connecting members of different clusters to each other.

Gene Set Variation Analysis (GSVA)

The GSVA³³ (v.1.25.0) software package for R/Bioconductor and has been described previously.^{30,34} Briefly, GSVA is a non-parametric, unsupervised method for estimating the variation of pre-defined gene sets in case 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-Smirnoff (KS)-like random walk statistic and a negative value for each gene set. EA and AA SNP-predicted genes were used to create GSVA gene signatures (official gene symbols for each signature are listed in Table S2). In the case of leukotriene biosynthesis, cell cycle, ubiquitylation and sumoylation, apoptosis signaling, nuclear receptor signaling, and PKA signaling, genes were initially identified following protein-protein interaction network construction and MCODE clustering. Cluster identity was determined by BIG-C and/or IPA canonical pathway analysis, where each cluster was used as a GSVA probe. Gene signatures for diapedesis, TH1 activation pathway, unfolded protein and stress, T cell exhaustion, and SLE in B cell signaling were all informed by established IPA canonical pathways. The signature for lysosome was derived from the Lysosome BIG-C category. All interferon and cytokine signatures (core IFN, IFNB1, IFNA2, IFNW, IFNG, IL12, and TNF) have been described previously.³⁰ Metabolic signatures for oxidative phosphorylation and glycolysis were based on literature mining and established IPA canonical pathways. Enrichment of each signature was examined in EA and AA SLE subjects and healthy control subject whole blood from GEO: GSE88884. Differences between control subjects and SLE-affected subject GSVA enrichment scores were determined using the Welch's *t* test for unequal variances in Graphpad PRISM 8.0.

Drug Candidate Identification and CoLT Scoring

Drug candidates were identified using LINCX, STITCH (v.5.0), and IPA. Each of these 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. In addition to iden-

tifying drugs targeting predicted genes directly, these tools were also used to identify drugs targeting select upstream regulators. Where information was available, drugs were assessed by CoLT scoring to rank potential drug candidates for repositioning in SLE as previously described.³⁵

Results

SLE-Associated Variants Predict Downstream Target Genes

We examined the distribution (e.g., coding, non-coding) of 834 non-HLA single-nucleotide polymorphisms (SNP) reported as significantly associated with SLE in a large trans-ancestral genetic association study¹⁸ (Figure 1A; Table S3). The majority of SNPs were found in intronic (43.1%) or intergenic (24.4%) regions. Approximately 26% of SLE-associated SNPs mapped to regions encoding transcripts (7.8% exons, 5' UTRs, 3' UTRs) or known regulatory regions (18.8%; TFBS, promoters, enhancers, etc.). 5.7% of SNPs mapped to regions containing long non-coding (lnc)RNAs or micro (mi)RNAs.

The reported test of SNP-by-ancestry (i.e., heterogeneity) from the meta-analysis and ancestry-specific effect size allowed determination of whether the SNP was more associated with only one ancestry (e.g., AA specific) or was a trans-ancestry association (i.e., comparable effect size across AA and EA)¹⁸ (Table S4). Next, we attempted to identify the most plausible gene(s) affected by the SLE-SNP association (Figure 1B). Using the GTEx and Blood eQTL browser databases,^{19,20} we identified 78 EA and 21 AA-related eQTLs linked to 207 and 29 expression genes (E-Genes) primarily associated with EA and AA, respectively. A total of 148 eQTL were common to both ancestries and were associated with 523 shared E-Genes (Table S5). Interestingly, we observed that E-Genes predicted by a single SNP tended to be enriched for a common molecular function (Figure S1; Table S6). The remaining 587 SNPs that were not associated with eQTLs were assigned to the closest proximal gene (P-Gene), revealing 520 P-Genes (465 EA, 34 AA, and 21 shared) (Figures 1C and 1D; Table S5).

Since variants that alter or disrupt transcription factor binding are also known to dysregulate gene expression,³⁶ we sought to identify SNPs within regulatory elements (e.g., enhancers and promoters). Computational tools, including GeneHancer and HACER that connect putative regulatory SNPs with transcription factors and downstream target genes (T-Genes),^{21,22} were used to identify 64 SNPs overlapping regulatory elements or promoters predicted to impact the expression of 627 T-Genes (475 EA, 9 AA, and 143 shared) and the action of 95 transcription factors (Tables S5 and S7). For variants located in coding regions, 23 SNPs (14 EA, 2 AA, 7 shared) were associated with either non-synonymous or nonsense changes, affecting 22 genes (C-Genes; 14 EA, 2 AA, and 6 shared) (Table S8).

Figure 1C depicts the overlap between SNPs based on discovery method, whereas Figure 1D shows the overlap

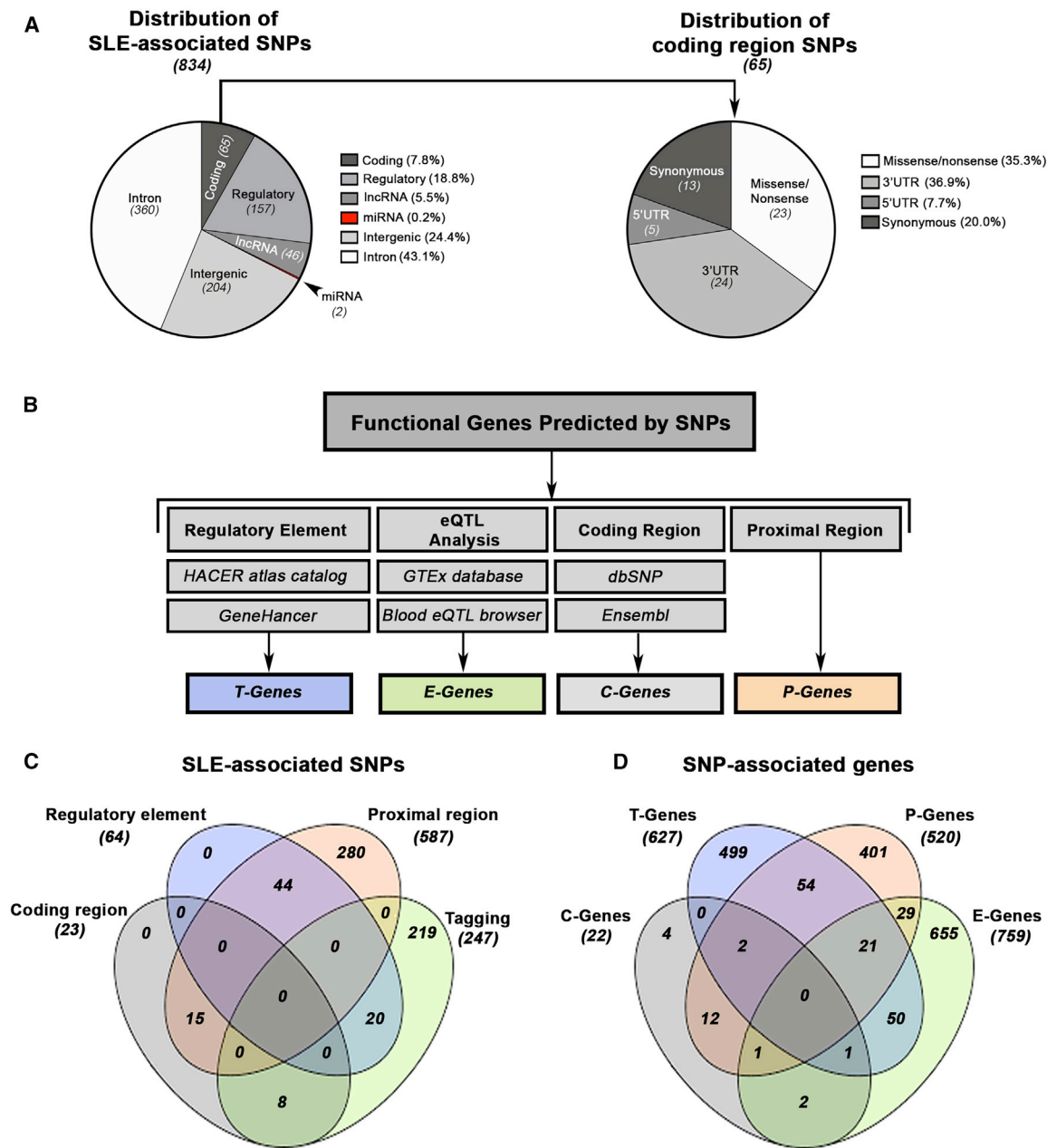


Figure 1. Mapping the Functional Genes Predicted by SLE-Associated SNPs

(A) Distribution of genomic functional categories for all ancestry-specific non-HLA-associated SLE SNPs (tiers 1–3). Non-coding regions include micro (mi)RNAs, long non-coding (lnc)RNAs, introns, and intergenic regions. Regulatory regions include transcription factor binding sites (TFBS), promoters, enhancers, repressors, promoter flanking regions, and open chromatin. Coding regions were broken down further and include 5' UTRs, 3' UTRs, and synonymous and nonsynonymous (missense and nonsense) mutations.

(B) Functional genes predicted by SNPs are derived from four sources including regulatory elements (T-Genes), eQTL analysis (E-Genes), coding regions (C-Genes), and proximal gene-SNP annotation (P-Genes).

(C and D) Venn diagram depicting the overlap of all SLE-associated SNPs (C) and all predicted E-, T-, P-, and C- Genes (D).

between the corresponding SNP-predicted E-, T-, C-, and P-Genes. No genes were shared within all four groups, and we observed limited commonality between T-, P-, and E-Genes, with only 21 genes shared among the three groups. Despite the overall diversity of genes observed in each list, significant overlap was observed in the number of genes shared between ancestries (Figure S2, Table S5).

Characterization of Gene Signatures

We next carried out a series of bioinformatics analyses to determine the biological function of the full array of EA (1,676) and AA (725) SNP-predicted genes (Figure S2). P-Genes were analyzed separately from E-T-C-Genes to avoid overrepresentation of immune-based processes because of the ImmunoChip design bias.³⁷ Both the EA and AA gene sets exhibited similarity in biological function reflected

in the number of matching functional pathways (immune signaling), IPA canonical pathways (*TH1 pathway*, *IFN signaling pathway*, *glucocorticoid receptor signaling*, etc.), and GO terms (epidermis development [GO: 0008544], immune response [GO: 0006955]), but also a number of clear differences. In addition, EA and AA gene sets were examined using a clustering program (I-Scope, see [Material and Methods](#)) that detects immune and inflammatory cell type signatures within large gene lists to identify dominant immune cell populations contributing to disease pathology within each ancestry.³² Both AA and EA exhibited enrichment of B cell genes, with AA exhibiting a significantly greater B cell signal. As a control, genes derived from non-associated Immunochip SNPs (“random”) were examined and found to be modestly enriched in some biosynthesis and metabolic pathways but did not overlap with any EA or AA pathways, nor did they exhibit enrichment in cellular categories.

Parallel analyses were also conducted examining associations that were more strongly linked to one ancestry over the other (because of higher allelic frequency), independent of the shared gene cohort. Analysis of representative EA genes revealed enrichment in processes related to innate immune function, including the functional category for pattern recognition receptors, GO terms for the cellular response to LPS (GO:007122), canonical pathways for *JAK/STAT signaling* and *agranulocyte adherence and diapedesis*, and cellular categories for myeloid cells and T and B cells ([Figures 2A–2D](#) and [S3](#)). AA-associated genes were differentially enriched in the adaptive immune response reflected by the cellular categories for anergic or activated T cells and GO terms for B cell activation (GO:0042113) and T cell co-stimulation (GO: 0031295), with additional enrichment in biological processes for degradation, including the functional category lysosome, and IPA pathways for *autophagy and phagosome maturation* ([Figures 2A–2D](#) and [S3](#)). To establish that ancestry-driven differences were not related to inequality in the number of SNPs used to predict genes for analysis (604 EA SNPs versus 77 AA SNPs), we directly compared all genes associated with AA SNPs to those predicted by the 77 most significant EA SNPs to generate an approximately equal cohort of genes. Overall, limited overlap was observed in the number of genes (3 total; *c7orf72*, *FIGNL1*, and *IKZF1*) and enriched pathways shared between ancestries ([Figure S4](#)).

Protein Interaction-Based Clustering of SNP-Associated Genes

We next sought to assess the relationship between all SNP-associated genes regardless of ancestral origin to obtain information about the likely nature of the molecular pathways driving SLE. Overall, 52.7% of E-T-C-Genes (701/1,330) and 53.2% of P-Genes (272/520) were incorporated into protein-protein interaction (PPI) networks with the majority of E-T-C-Genes coalescing into several large, multi-functional clusters. The integration of all EA and AA genes into the same network highlights the overall

commonality in gene function and connectivity observed among the ancestries, including cluster 7 enriched in molecules associated with immune signaling and Golgi function, cluster 15 dominated by metabolic processes, and cluster 3 containing a robust interferon signature ([Figures 3A and S5](#)). Networks constructed of all P-Genes revealed the predominance of immune function with seven out of ten of the largest, intra-connected clusters enriched in immune activity ([Figure 3B](#)).

To confirm that the predicted molecular pathways derived from SNP-predicted genes were specific for SLE and not related by chance, we carried out a parallel analysis examining PPI networks composed of genes derived from non-associated (random) Immunochip SNPs ([Figure 3C](#); [Table S9](#)). Examination of meta-structures reveals that random gene networks incorporated fewer genes overall (258/1,033; 24.9%) and exhibited significantly fewer inter-cluster connections and fewer intra-cluster connections, appearing as independent entities lacking robust functional relationships with neighboring clusters ([Figure 3D](#)).

Predicted Functional Genes Are Associated with Altered Expression in SLE and Are Enriched in Differential Expression Datasets

To determine whether genes more frequently linked to specific ancestries exhibited altered expression in SLE, ancestry-driven SNP-predicted genes were matched to differentially expressed gene (DEGs) in independent SLE datasets from various tissues, including whole blood, PBMCs, B cells, T cells, synovium, skin, and kidney ([Table S10](#)). We observed that 63% of EA and 70% of AA SNP-predicted genes were identified as DEGs across all datasets ([Figures 4A and 4B](#); [Tables S10 and S11](#)). Of the 661 genes shared between ancestries, 441 genes (66.7%) were identified as DEGs, with the interferon-stimulated genes (*HERC5*, *IFI35*, *IFI44L*, *IFI6*, *IFIT1*, *MX1*, and *SPATSL2L*), interferon-regulatory factors (*IRF4*, *IRF5*, and *IRF7*), and PRRs (*OAS1*, *OAS2*, *OAS3*, and *SLC15A4*) differentially expressed across all datasets ([Figure 4C](#); [Table S12](#)).

Relationship of SNP-Predicted Genes and Upstream Regulators of Gene Expression Profiles

Using IPA, DEGs were used to identify potential biologic upstream regulators (UPRs) with the goal of determining whether UPRs of the altered gene expression profile in SLE were SNP-predicted genes. Overall, 141 UPRs predicted from the altered SLE gene expression profiles were also SNP-associated genes, including surface receptors, signaling molecules, cytokines, and transcription factors, many with known roles in SLE such as *IRF7*, *ITGAM*, *IFNG*, *IKZF1*, and *CD40* ([Table S13](#)). In addition, 41 UPRs were transcription factors predicted to interact with transcription factor binding sites altered by an SLE-associated SNP, including *MYC*, *EZH2*, *NFATC1*, *STAT3*, *STAT5a*, *Fos*, *JunB*, and *RelA*. Thus, of the 1,238 UPRs predicted to drive SLE gene expression, 181 were SNP-predicted genes.

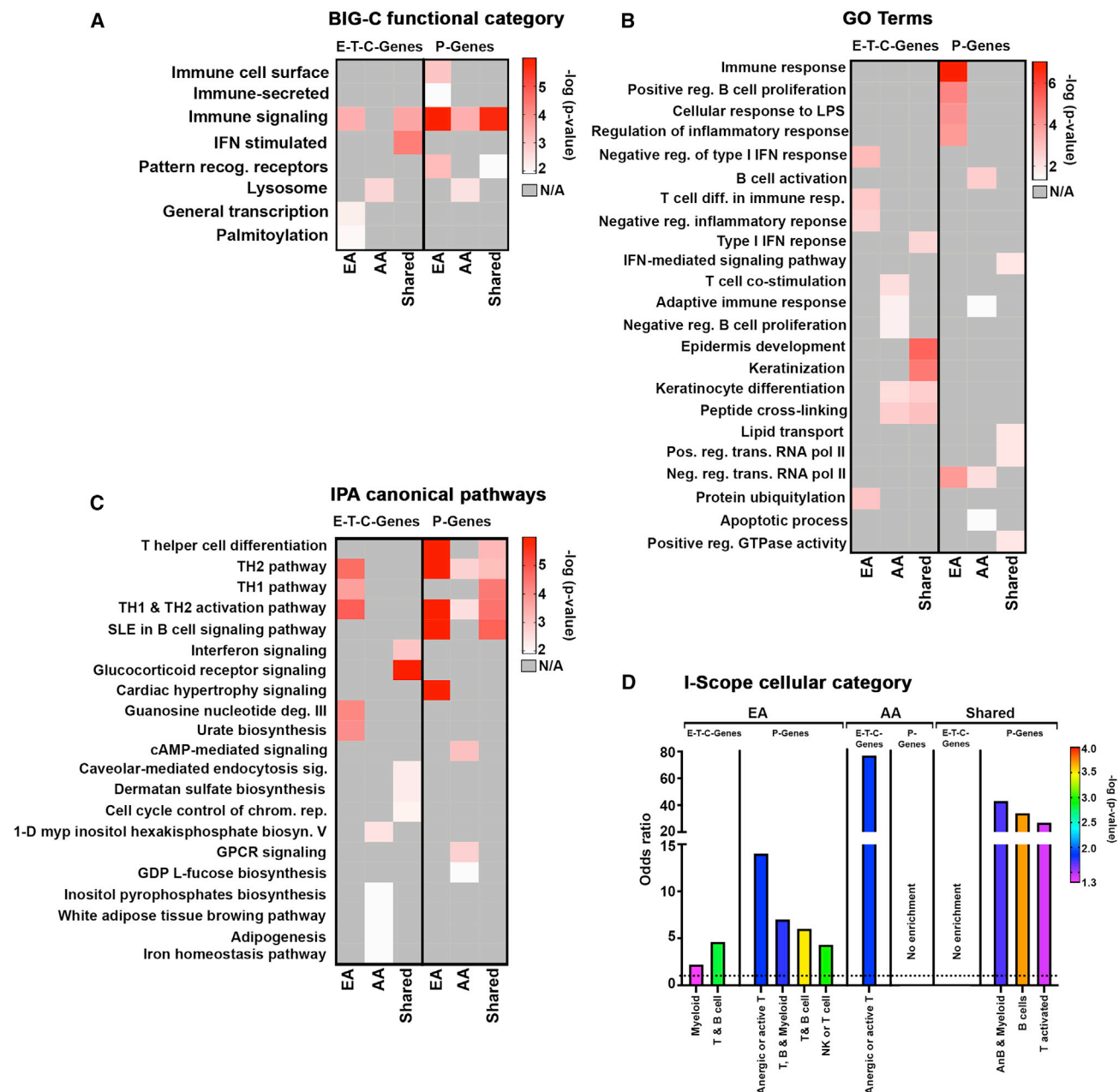


Figure 2. Functional Characterization of Predicted Genes

(A) Ancestry-dependent and -independent SNP-predicted genes were analyzed to determine enrichment using functional definitions from the BIG-C (Biologically Informed Gene Clustering) annotation library. E-, T-, and C-Genes were analyzed together; P-Genes were examined separately. Enrichment was defined as any category with an odds ratio (OR) > 1 and $-\log_{10}(p\text{ value}) > 1.33$.

(B and C) Heatmap visualization of the top five significant IPA canonical pathways and gene ontology (GO) terms for each gene list (E-T-C-Genes and P-Genes) organized by ancestry. Top pathways with $-\log_{10}(p\text{ value}) > 1.33$ are listed.

(D) I-Scope hematopoietic cell enrichment defined as any category with an OR > 1, indicated by the dotted line, and $-\log_{10}(p\text{ value}) > 1.33$.

Delineation of Signaling Pathways Identified by Ancestry Specific SNP-Associated Genes and UPRs

Connectivity mapping of SNP-associated genes and all UPRs predicted from SLE gene expression profiles were used as input to build more complete PPI networks, and individual gene clusters were analyzed by BIG-C and IPA to identify those molecules and pathways highly associated with disease. A total of 45 pathways were representative of EA genes

and UPRs, with the largest clusters (1 and 3) heavily involved in pattern recognition receptor signaling (Figures 5A and 5B; Table S14). Clusters 4 and 5 revealed enrichment in lymphocyte activation and differentiation pathways that were also common to both the AA and shared gene networks. Twenty pathways were detected only in EA, including several involved in cellular communication, cytokine signaling, and migration. The AA network was smaller

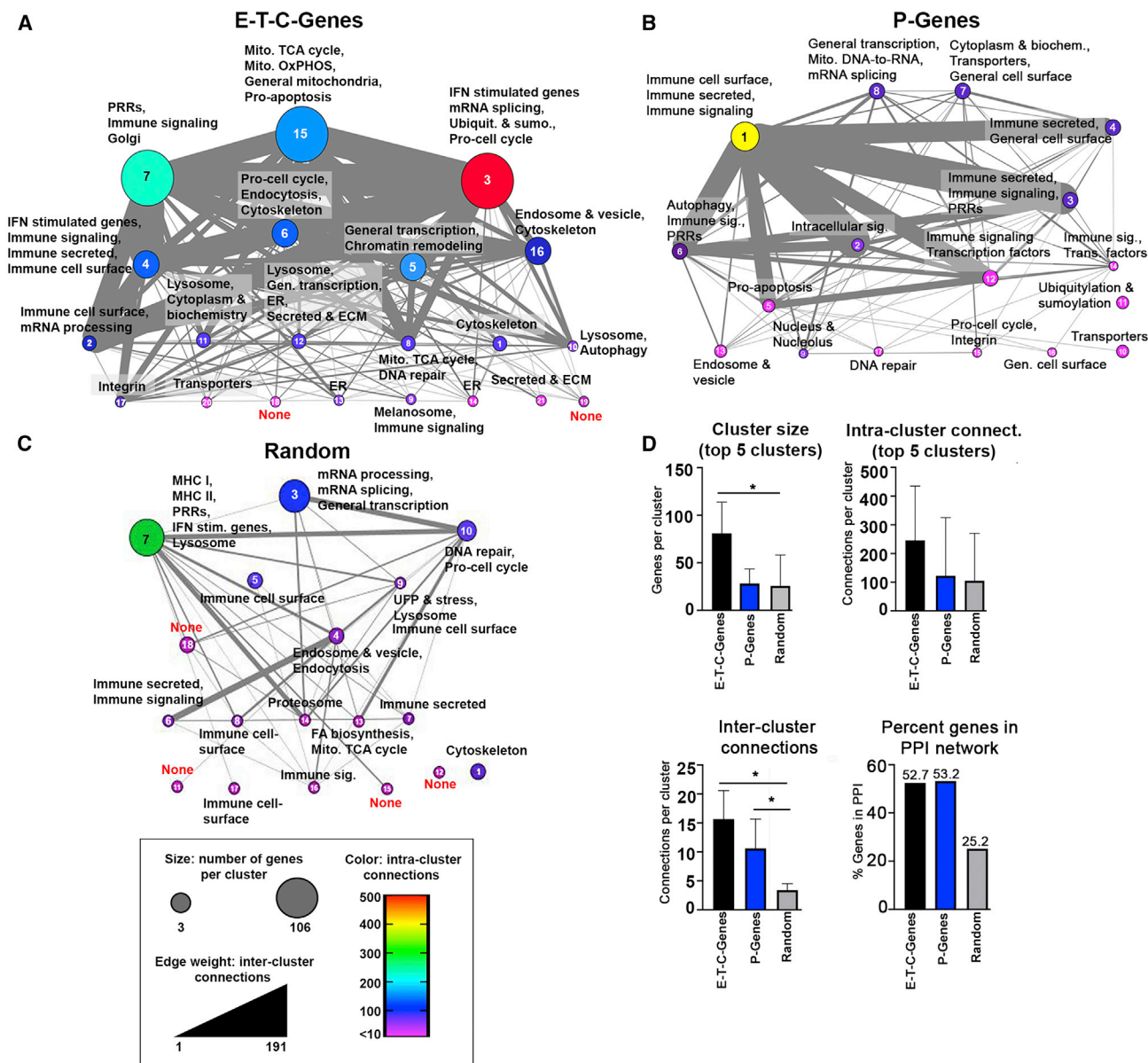


Figure 3. Cluster Metastructures for SLE-Predicted and Randomly Generated Genes

(A–C) Cluster metastructures were generated based on PPI networks, clustered using MCODE, and visualized in CytoScape. 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. Cluster number is indicated within each metacluster. A random gene network (1,033 genes) was clustered along-side networks for E-T-C-Genes and P-Genes. Functional enrichment for each cluster was determined using BIG-C.

(D) Quantitation of cluster size, intra-cluster connections, inter-cluster connections, and the percent of genes incorporated into each network is displayed. Error bars represent the 95% confidence interval; asterisks (*) indicate a p value < 0.05 using Welch's t test.

(Figure 6; Table S14), with fewer SNP-predicted genes and associated UPRs, yet contained genes critical for T cell responses such as *PRDM1* and *IKZF1*. Pathways differentially associated with AA were overwhelmingly represented by processes related to degradation and cellular stress, found in clusters 5, 3, and 11, as well as metabolic processes in cluster 6 (Figure 6B). In general, we observed good agreement between enrichment tools, with BIG-C functional categories similar or complementary to the listed canonical pathways. Discrepancies were usually the result of the wider range of pathways described by IPA utilizing a subset of

genes captured by BIG-C. This was apparent in cluster 2 of the EA network (Figure 5B), where top-ranked pathways for *Glucocorticoid signaling* and *Hepatic fibrosis* include numerous genes related to the immune-secreted and immune cell surface BIG-C categories (i.e., *CXCL8*, *CCL5*, *IL6*, etc.). This was also the case for cluster 11 in the AA network (Figure 6B) where functional enrichment in autophagy is driven by *GABARAP*, a multifunctional ATG-8 subfamily member involved in both autophagosome formation as well as functioning upstream of OPRK1 (Opioid Receptor Kappa 1) within the *Opioid signaling pathway*.

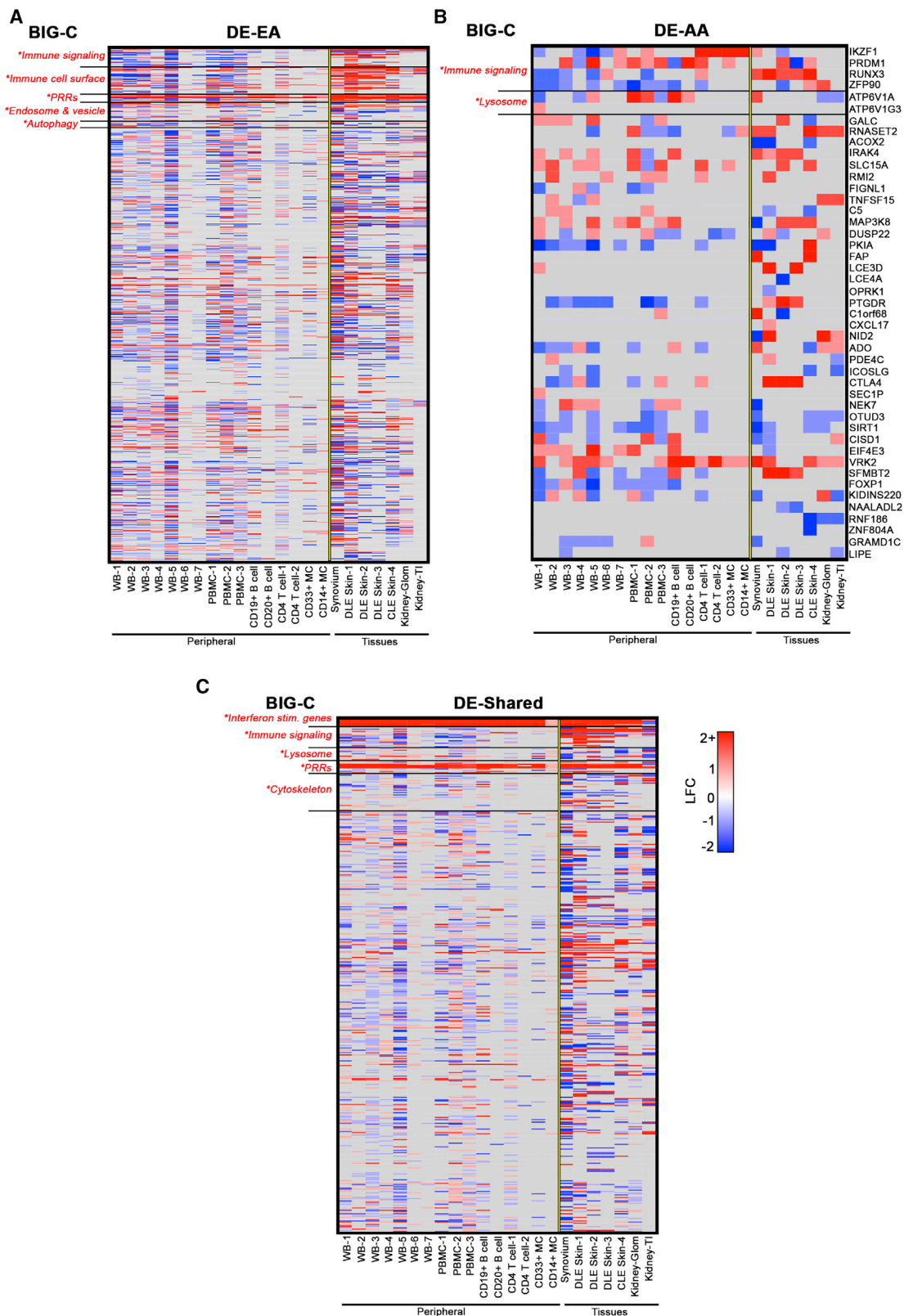


Figure 4. Comparison of EA, AA, and Shared SNP-Predicted Genes with SLE Differential Expression Datasets

SNP-predicted genes were matched with SLE differential expression (DE) data and organized by ancestry. The fold-change variation of EA, AA, and shared genes is shown. Heatmaps are organized by BIG-C category. Enriched categories indicated with an asterisk. Enrichment was defined as any category with an OR > 1 and $-\log_{10}(p \text{ value}) > 1.33$.

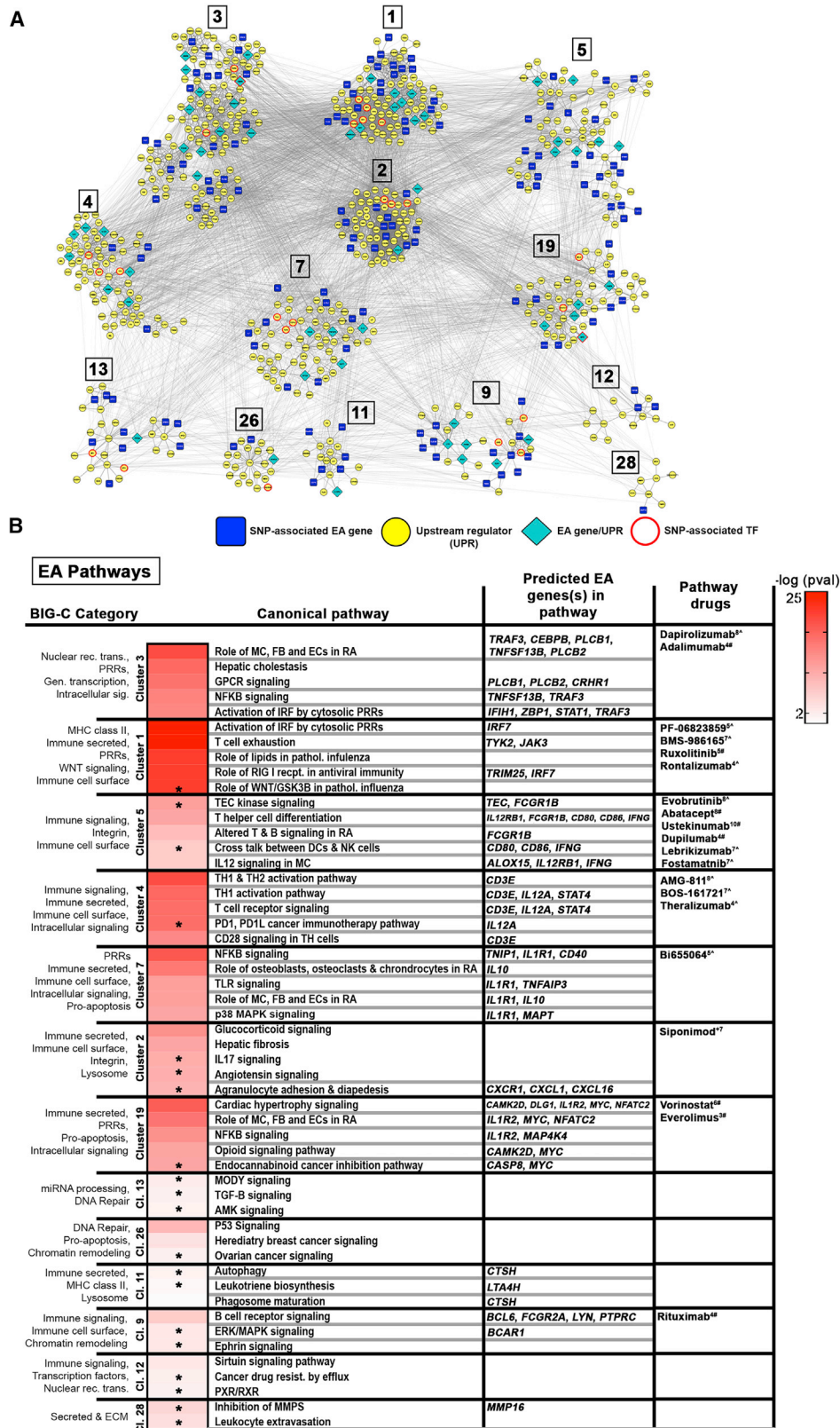


Figure 5. Key Pathways Determined by EA Genes and Upstream Regulators

(A) Differentially expressed EA genes and their upstream regulators (UPRs) were used to create STRING-based PPI networks. DE EA genes identified as UPRs and SNP-predicted TFs are indicated. Clusters were generated via CytoScope using the MCODE plugin.

(legend continued on next page)

Pathways that were enriched in both EA and AA (shared) included *IL12 signaling and production by macrophages*, *TLR signaling*, and *activation of IRFs by cytosolic PRRs*, as well as *PRRs in the recognition of bacteria and virus* (Figure 7). Figure 8A depicts both the unique and overlapping canonical pathways predicted by the EA and AA gene sets, whereas Figure 8B shows the broad overall pathway categories shared between EA and AA and those that are differentially enriched within each ancestral group.

To support our pathway predictions, Gene Set Variation Analysis (GSVA)³³ was applied to identify differentially enriched gene signatures in whole blood (WB) samples from SLE-affected subjects (EA and AA) and control subjects (Table S15). SNP-associated genes were used to create a collection of signatures informed by protein-protein interaction networks and IPA canonical pathways, or were previously defined as SLE associated (Table S2).³⁰ As shown in Figure 8C (and Figure S6), enrichment of a number of pathways in both EA and AA SLE was noted, including TH1 activation pathway, cell cycle, and lysosome, as well as cytokine-based signatures for core interferon (IFN), IFNG, IL12, and the IFN subtypes IFNA2, IFNB1, and IFNW. GSVA enrichment scores using signatures for leukotriene biosynthesis and diapedesis were able to separate EA SLE-affected individuals, but not AA subjects, from healthy control subjects (Figure 8D), whereas signatures for unfolded protein response, T cell exhaustion, and B cell signaling were specifically enriched in AA SLE (Figure 8E). A number of pathways including ubiquitylation and sumoylation, apoptosis signaling, nuclear receptor signaling, and TNF separated SLE-affected individuals from control subjects with additional enrichment observed in AA (versus EA) SLE (Figure 8F).

Gene signatures for metabolic pathways, including mitochondrial oxidative phosphorylation and glycolysis, were also investigated but did not demonstrate any significant change between SLE-affected individuals and control subjects or between ancestries, although there was a trend toward enrichment in SLE-affected individuals (Figure S6). It is possible that the relatively few control samples available in this dataset may have masked subtle differences between disease and control. In contrast, GSVA scores for the PKA signaling gene signature were significantly lower in SLE-affected individuals compared to control subjects.

Pathway Analysis Facilitates Drug Prediction

Identified pathways were employed to facilitate identification of possible new therapeutic interventions and numerous drug candidates were predicted. Canonical

pathways related to T cell function are shared among ancestries, as are many predicted drugs targeting T cell activity including abatacept, theralizumab, and AMG-811 (Figure 8B). Broader analysis of common pathway categories also suggested the utility of targeting T cell signaling, as well as cytokine pathways such as IL12/23 signaling with ustekinumab and/or interferon signaling with anifrolimab (Figure 8B). Drugs targeting pathways associated with EA include BMS-986165, a small molecular inhibitor of TYK2 (Figure 5B). Therapeutic candidates targeting pathways representative of AA include the proteasome inhibitor bortezomib, as well as PF-06650833, an IRAK4-specific inhibitor (Figure 6B). Unique pathway categories identified for EA and AA suggest additional ancestry-driven interventions, such as the small molecule inhibitor of sphingosine-1-phosphate receptor 1 (S1PR1) siponimod for EA (prevents leukocyte egress) and the HDAC inhibitor vorinostat for AA SLE (Figure 8B).

Discussion

SLE is a chronic autoimmune disease with a strong genetic component.³⁸ Although socioeconomic disparities are also known to impact disease development,^{39,40} genetic heterogeneity between ancestral populations is widely acknowledged to be important in SLE risk, since affected individuals of African descent have a higher prevalence of lupus and experience the disease more severely than those of European ancestry.^{7,8} Despite improved understanding of how inherited genetic variation impacts disease risk, genetic analyses to date have failed to provide a clear path toward novel therapeutic development. This is of particular concern with respect to AA populations where the control of disease activity remains suboptimal.^{10–12} Furthermore, most GWASs only consider the most significant variants (lead SNP, marker with the lowest p value), typically following-up on those that are novel.² Here, we propose a novel strategy employing integration of all SLE SNP association-predicted genes into functional pathways to identify the genetic contributions to SLE disease pathogenesis and possible differences contributed by ancestry. To accomplish this, we employed statistical and computational analyses along with data acquired from functional genomic assays and differential gene expression studies to map the global gene expression landscape of SLE and further define the disease-associated pathways responsible for the inherent disparities influencing SLE progression.

It is important to reiterate that the SNPs examined here are generally present in both ancestral populations and are associated with a significant cohort of common genes.

(B) Top IPA canonical pathways representing individual clusters and enriched ($OR > 1$, $p \text{ value} < 0.05$) BIG-C categories are listed; heatmap depicts the $-\log(p \text{ value})$ for significant IPA pathways. Unique pathways are indicated by asterisks. Predicted EA genes and select drugs acting on direct gene targets or on any of the pathways are listed. ColT scores (-16 – $+11$) are in superscript; # denotes FDA-approved drugs, ^ denotes drugs in development. SOC, standard of care.

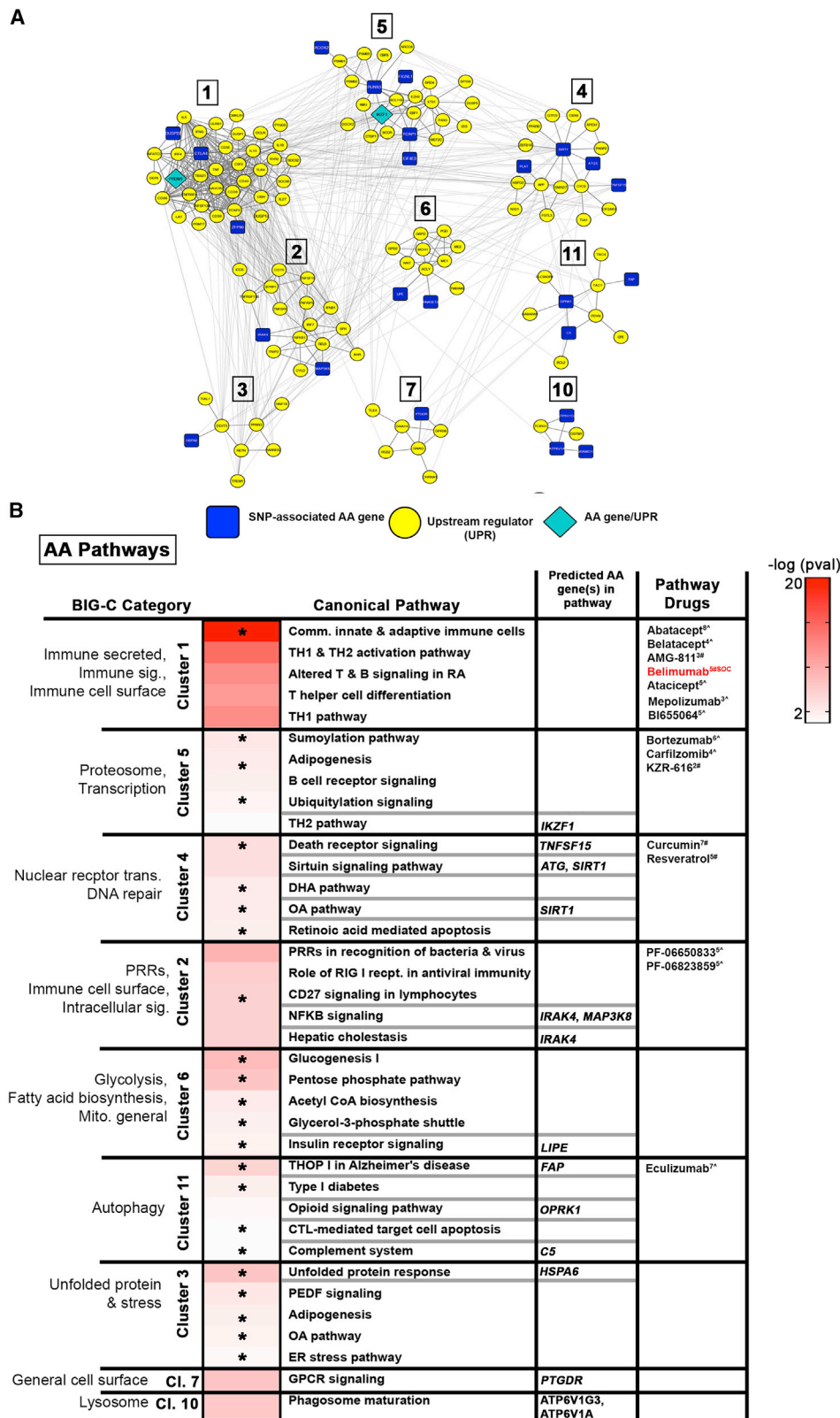


Figure 6. Key Pathways Determined by AA Genes and Upstream Regulators

(A) Differentially expressed AA genes and their upstream regulators (UPRs) were used to create STRING-based PPI networks. DE AA genes identified as UPRs are indicated. Clusters were generated via CytoScape using the MCODE plugin.

(legend continued on next page)

However, differences in allele frequencies suggest that some SNP associations and pathways may be more representative of one ancestry over another. Here, eQTL analysis identified 247 tagging SNPs associated with 759 E-Genes (77 EA, 21 AA, 523 shared). Given that the majority of eQTLs identified here map to multiple E-Genes (many within the same functional network), eQTL-based gene prediction may be particularly valuable for network modeling and disease analysis. Recent studies have also shown that disease-susceptibility variants frequently lie in regulatory enhancer elements.⁴¹ However, only 20% (157) of SNPs analyzed here were located in known regulatory regions. While other methodologies, such as promoter capture Hi-C, link regulatory SNPs to putative target genes,⁴² we were particularly interested in predicting the effect of variants in non-coding regions, especially since the majority of SNPs examined here were intergenic. Using computational gene prediction algorithms that incorporate chromatin interaction data and intergenic enhancer annotation from several hundred cell lines,²¹ additional regulatory SNPs were identified that were predicted to change transcription factor binding and were associated with 627 downstream targets (T-Genes; 472 EA, 9 AA, 143 shared). Although some regulatory SNPs also exhibit eQTL effects, we nonetheless uncovered 496 unique T-Genes enriched in a diverse array of functional categories. Finally, we identified 23 variants resulting in nonsense or non-synonymous amino acid changes affecting 22 genes (C-Genes), 12 of which were predicted to negatively impact protein function. The remaining 587 risk SNPs were mapped to the nearest, most proximal gene, resulting in 520 P-Genes (465 EA, 34 AA, 21 shared).

One major limitation to the current study is that all computational and experimental approaches outlined here are inferential; by attempting to provide a more comprehensive translation of GWAS findings, a major challenge remains in determining those genes that are causative. To address this, PPI networks and clustering based on interaction strength helped exclude those genes lacking strong connections to molecules within or between similarly functioning clusters. Compared to SNP-predicted E-, T-, C-, and P-Genes where we observed large, highly connected clusters, randomly generated genes generally formed smaller clusters, exhibited fewer intra- and inter-cluster connections, and ultimately appeared as independent entities. In addition, SNP-predicted genes were compared to SLE datasets (SLE versus control) to determine those genes that were differentially expressed in active disease. This was important to amplify the GTEx predictions that were based on data from a collection of cell lines and not primary SLE cells.¹⁹ We observed a high percentage of SNP-predicted genes differentially expressed across all datasets, although

expression was not always in the same direction (i.e., all up-regulated or all downregulated). In EA, for example, functionally enriched categories included immune signaling and immune cell surface whereas DE genes falling into these categories were more upregulated in the tissues (synovium, skin, and kidney) and predominantly downregulated in the periphery (WB, PBMC, T cells, B cells, and myeloid cells). This could reflect a cellular response to different inflammatory stimuli, such as the enhanced recruitment of myeloid cells to tissue in response to immune complex deposition. All DE genes were then used as input into IPA to generate upstream and downstream regulators and combined for further network and clustering analysis. This allowed us to identify biologically relevant pathways unique to each ancestry, a strategy that revealed essential differences between EA and AA SLE, as well as many pathways that were shared.

A second caveat to the current study is the use of the Immunochip which was constructed to cover multiple major autoimmune diseases and enable the identification of top-ranked SNPs associated with disease.³⁷ This platform was designed for use in EA populations and is therefore less informative for other ancestral groups, especially in non-HLA-associated regions. Furthermore, as chip coverage was confined to autoimmune and inflammatory diseases, SNPs affecting non-immune-related processes could be under-represented.

Despite these drawbacks, the pathway-based analysis of predicted genes and their upstream regulators presented here helps clarify the complex polygenic risk associated with SLE in multiple ancestries. We observed key dysregulated EA pathways centered around innate immune function and the response to inflammation, including cell movement, cytokine signaling, and cell-cell communication. Remarkably, GSVA gene signatures for leukotriene biosynthesis and diapedesis were sufficiently selective to separate EA SLE-affected individuals from control subjects, providing additional evidence for these pathways in SLE pathogenesis. Furthermore, SNP-predicted EA genes were enriched in myeloid and NK cell signatures, along with T and B cell signatures, findings that are consistent with previous reports showing increased myeloid lineage cell modules in EA subjects.^{43,44}

Pathways identified by AA-associated SNPs included those linked to protein degradation, such as the sumoylation pathway, ubiquitylation signaling, the ER stress pathway, unfolded protein response, osteoarthritis pathway (cell stress), and the neuroprotective role of THOP1 in Alzheimer disease, a pathway involved in the presentation of antigen generated by the proteasome. To ensure that the differences in ancestrally related genes were not related to imbalances in the number of SNPs

(B) Top IPA canonical pathways representing individual clusters and enriched ($OR > 1$, p value < 0.05) BIG-C categories are listed; heat-map depicts the $-\log(p$ value) for significant IPA pathways. Unique pathways are indicated by asterisks. Predicted AA genes and select drugs acting on direct gene targets or on any of the pathways are listed. CoLT scores (-16 to $+11$) are in superscript; # denotes FDA-approved drugs; ^ denotes drugs in development. SOC, standard of care.

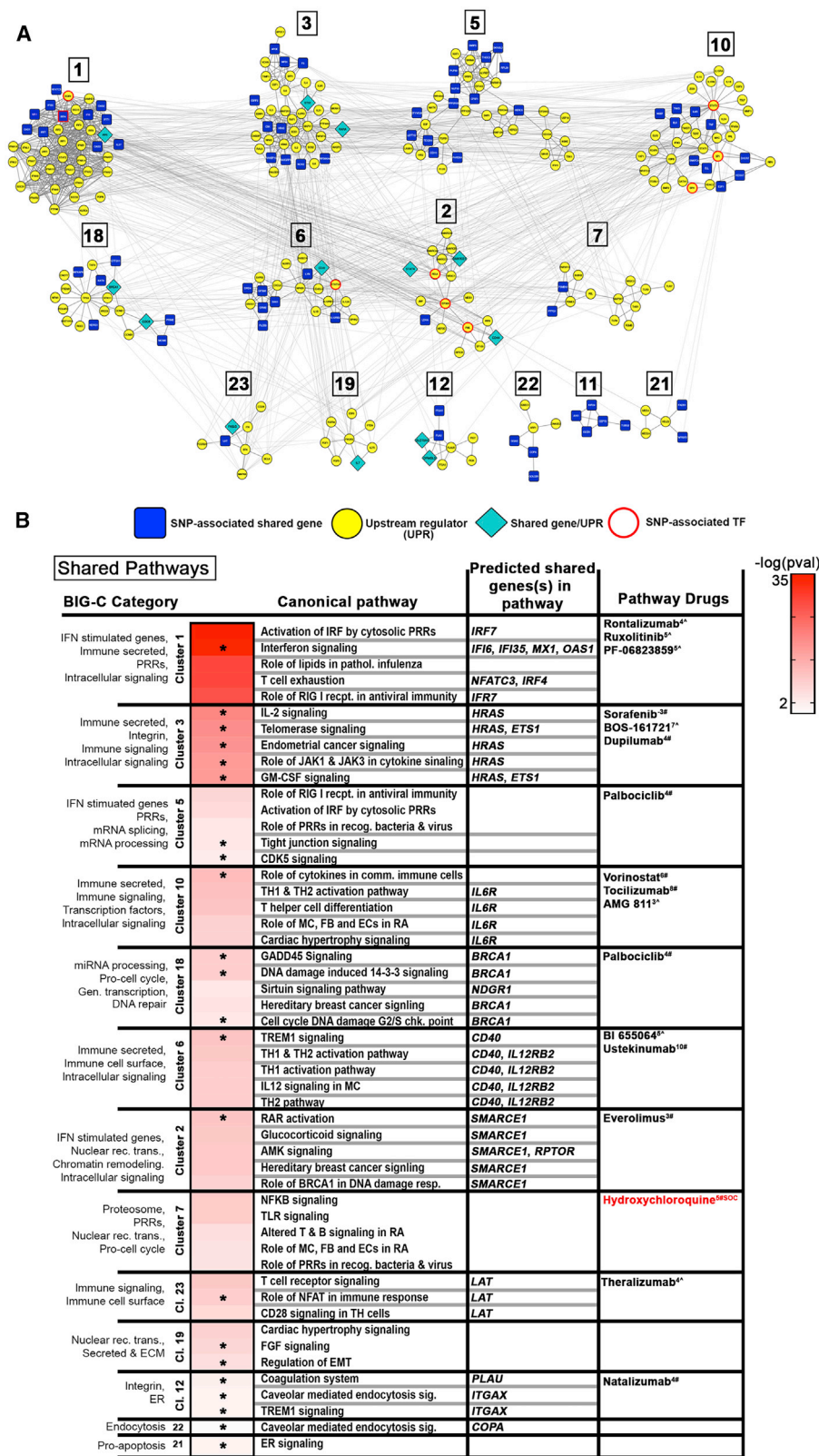


Figure 7. Key Pathways Determined by Shared Genes and Upstream Regulators

(A) Differentially expressed shared genes and their upstream regulators (UPRs) were used to create STRING-based PPI networks. DE shared genes identified as UPRs and SNP-predicted TFs are indicated. Clusters were generated via CytoScape using the MCODE plugin.

(legend continued on next page)

assessed, we separately looked only at the most highly associated genes in AA and EA populations and still were able to detect ancestry-related pathway differences. The importance of these pathways was confirmed by GSVA which demonstrated the unique enrichment of cellular stress mechanisms and B cell signaling in AA SLE-affected individuals. These observations are in line with reports showing increased B cell activation and plasma cells in AA case subjects.^{43–45} In fact, a recent comprehensive analysis of gene expression profiling in whole blood⁴⁴ showed that a major AA ancestral difference is related to the propensity to form autoantibodies to RNP and Sm and to have more types of autoantibodies in general. With these results together with the data presented in the current study demonstrating an underlying genetic predisposition toward B cell-driven disease, dysregulated B cell responses in AA case subjects likely contribute to the overall higher prevalence of tissue damage in this ancestral population. The increase in B cell-driven responses in AA subjects manifested by an increase in circulating B lineage cells was determined by the use of I-Scope, a validated analytic tool that converts gene expression profiles from bulk mRNA into the most likely cellular constituents.⁴⁴ Finally, it is increasingly recognized that ER stress and the UPR signaling pathway in dysregulated immune responses is closely tied to aberrant B cell activity in SLE.^{46,47}

Gene signatures representing cellular processes shared between ancestries provide further validation for our comprehensive pathway analysis. GSVA enrichment scores for the interferon response (IFN core, IFNA2, IFNB1, IFNW1, and IFNG) and inflammatory cytokines (IL-12 and TNF) exhibited the greatest difference between SLE and control, independent of ancestry. In line with this, work by Catalina et al.³⁰ showed that multiple IFN signatures are operative in an array of SLE subject samples from whole blood and tissues. Despite the report that metabolic abnormalities, including heightened glycolysis and mitochondrial glucose oxidation,⁴⁸ are associated with SLE, metabolic gene signatures for OXPHOS and glycolysis did not distinguish SLE-affected individuals from control subjects. However, examination of protein kinase A (PKA) signaling, a pathway that participates in the regulation of immune effector functions in T cells,⁴⁹ demonstrated significantly lower GSVA scores for the PKA signaling signature in both EA and AA SLE-affected individuals compared to control subjects. Consistent with these findings, previous reports have shown that T cells from SLE-affected individuals have a metabolic disorder of the PKA pathway characterized by markedly diminished PKA activity and dysfunctional T cell activity.^{50,51}

By focusing on pathways instead of individual genes, the current approach identified “actionable” points of thera-

peutic intervention with the potential to impact EA and AA SLE specifically. Thus, EA subjects may derive particular benefit from treatments that prevent leukocyte or lymphocyte infiltration into tissues, highlighting drugs that modulate this process. For example, sphingosine-1 phosphate receptor (S1PR) is a pleiotropic lipid mediator involved in the regulation of many cellular functions, including proliferation, survival, and cell motility.^{52,53} Siponimod, an FDA-approved treatment for multiple sclerosis, promotes the internalization of S1PR expressed on lymphocytes preventing cell migration to sites of inflammation.^{54–56} Given its high combined lupus treatment score (CoLTS),^{34,35,57} siponimod represents a high-priority small-molecule drug with potential for repurposing in SLE.

Similarly, given the enrichment of pathways linked to proteasomal and degradative processes in AA, therapeutic intervention may include proteasome inhibitors, such as bortezomib (BZ), an FDA-approved drug for mantle cell lymphoma and multiple myeloma,⁵⁸ as well as B cell-targeting therapies. Despite reports of adverse events, BZ was claimed to have efficacy in treating refractory SLE in a small, uncontrolled clinical trial and was shown to deplete bone marrow and peripheral plasma cells.⁵⁹ In addition, AA SLE-affected individuals showed a better response to rituximab (RTX), an anti-CD20 inhibitor,⁶⁰ and a trend toward better response in AA subjects with lupus nephritis.⁶¹ However, the anti-BAFF inhibitor belimumab (BEL) failed to demonstrate a positive effect among AA subjects. The reasons for this discrepancy are unclear but may be related to the fact that anti-CD20 therapies like RTX induce a broad and deep B cell depletion, whereas BEL has a significantly more restricted and/or attenuated B cell effect.^{62,63} Together, these data indicate the possibility that BZ (and/or more selective immunoproteasome inhibitors), and therapies broadly targeting B cell may hold promise for AA subjects who respond poorly to conventional therapies.

In conclusion, our study demonstrates that multilevel analysis is capable of defining gene regulatory pathways which not only reflect differences in EA and AA populations but also represents candidate pathways that may be the target of ancestry-specific therapies. Indeed, the ancestral SNP-associated predicted genes and gene expression profiles outlined here illustrate fundamental differences in lupus molecular pathways between ancestries and indicate that unique sets of drugs may be particularly effective at treating lupus within each ancestral group.

Data and Code Availability

All microarray datasets listed in this publication are available on the NCBI's database Gene Expression Omnibus (GEO).

(B) Top IPA canonical pathways representing individual clusters and enriched (OR > 1, p value < 0.05) BIG-C categories are listed; heat-map depicts the $-\log(p \text{ value})$ for significant IPA pathways. Unique pathways are indicated by asterisks. Predicted shared genes and select drugs acting on direct gene targets or on any of the pathways are listed. CoLT scores (−16 to +11) are in superscript; # denotes FDA-approved drugs; ^ denotes drugs in development. SOC, standard of care.

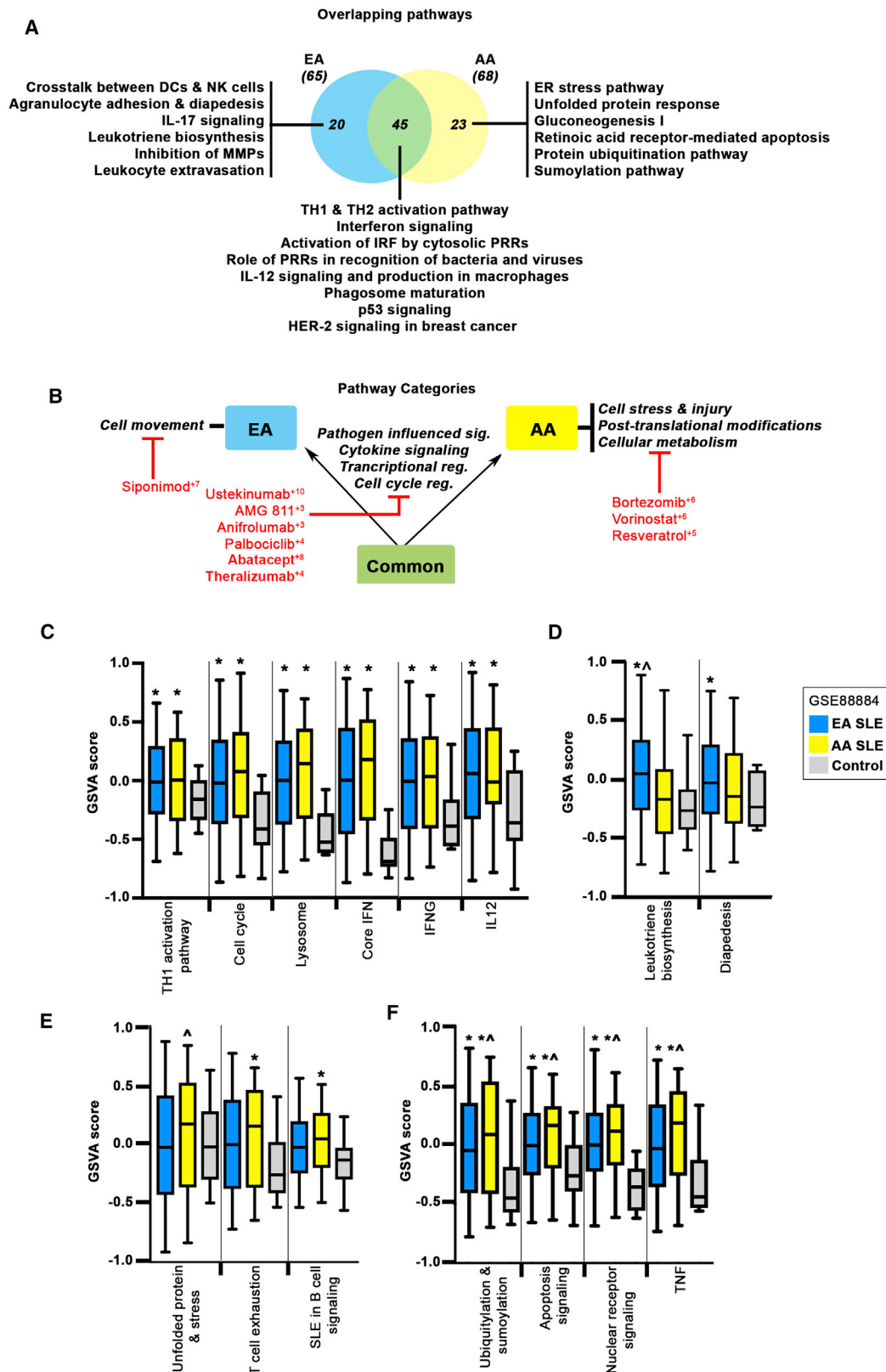


Figure 8. Overlapping Pathways and Categories Defining the EA and AA Predicted Gene Sets

(A) Venn diagram showing the number of overlapping pathways between EA and AA predicted genes and their UPRs. Representative IPA canonical pathways are indicated.

(B) Overall pathway categories are defined; shared categories are between the arrows, EA-specific (left) and AA-specific categories (right) are indicated. Select drugs at points of intervention are noted. Superscript denotes CoLT score.

(legend continued on next page)

Supplemental Data

Supplemental Data can be found online at <https://doi.org/10.1016/j.ajhg.2020.09.007>.

Acknowledgments

The work presented in this manuscript was funded by a grant awarded to P.E.L. and A.C.G. of the RILITE Research Institute by the John and Marcia Goldman Foundation. The funder provided support in the form of salaries for authors (K.A.O., A.P., B.N.A., P.B., M.D.C., J.M.D., K.M.K., A.C.L., R.D.R.). The authors gratefully acknowledge additional support from the Lupus Research Alliance (LRA) and the Department of the Army (W81XWH-20-1-0686).

Declaration of Interests

The authors declare no competing interests.

Received: June 19, 2020

Accepted: September 16, 2020

Published: October 7, 2020

Web Resources

CytoScape, <https://cytoscape.org>
DAVID, <https://david.ncifcrf.gov/>
dbSNP, <https://www.ncbi.nlm.nih.gov/snp/>
Ensembl, <http://www.ensembl.org>
GeneHancer, <https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=geneHancer>
GeneCards, <https://www.genecards.org>
Genotype-Tissue Expression (GTEx) project portal, <http://www.gtexportal.org>
Gene Expression Omnibus (GEO), <https://www.ncbi.nlm.nih.gov/geo/>
Human Active Enhancer to Interpret Regulatory variants (HACER), <http://bioinfo.vanderbilt.edu/AE/HACER>
HaploReg, <https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>
InteractiVenn, <http://www.interactivenn.net>
Ingenuity Pathway Analysis (IPA), <https://www.qiagenbioinformatics.com>
John and Marcia Goldman Foundation, jmgoldmanfoundation.org
Library of Integrated Network-based Cellular Signatures (LINCS), <http://www.lincsproject.org/>
Lupus Research Alliance, lupusresearch.org
OMIM, <https://www.omim.org/>
Protein Analysis Through Evolutionary Relationships (PANTHER), <http://www.pantherdb.org>
Polymorphisms Phenotyping v2 (PolyPhen-2), genetics.bwh.harvard.edu
PROVEAN, <http://provean.jcvi.org>

rAggr (site no longer live; please use SNIIPA instead), <http://raggr.usc.edu>

SNIIPA, https://snipa.helmholtz-muenchen.de/snipa3/index.php?task=proxy_search

Sorting Intolerant From Tolerant (SIFT-4G), <http://sift-dna.org/sift4g>

STITCH, <http://stitch.embl.de>

Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), <https://string-db.org>

Variant Effect Predictor (VEP), <http://www.ensembl.org>

References

1. Rullo, O.J., and Tsao, B.P. (2013). Recent insights into the genetic basis of systemic lupus erythematosus. *Ann. Rheum. Dis.* 72 (Suppl 2), ii56–ii61.
2. Benthall, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tomblinson, P., Behrens, T.W., Martín, 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.
3. Morris, D.L., Sheng, Y., Zhang, Y., Wang, Y.F., Zhu, Z., Tomblinson, P., Chen, L., Cunningham, D.S., Benthall, 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.
4. 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.
5. Alarcón-Riquelme, M.E., Ziegler, J.T., Molineros, J., Howard, T.D., Moreno-Estrada, A., Sánchez-Rodríguez, E., Ainsworth, H.C., Ortiz-Tello, P., Comeau, M.E., Rasmussen, A., et al. (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.
6. Hanscombe, K.B., Morris, D.L., Noble, J.A., Dilthey, A.T., Tomblinson, P., Kaufman, K.M., Comeau, M., Langefeld, C.D., Alarcón-Riquelme, M.E., Gaffney, P.M., et al. (2018). Genetic fine mapping of systemic lupus erythematosus MHC associations in Europeans and African Americans. *Hum. Mol. Genet.* 27, 3813–3824.
7. Williams, E.M., Bruner, L., Adkins, A., Vrana, C., Logan, A., Kamen, D., and Oates, J.C. (2016). I too, am America: a review of research on systemic lupus erythematosus in African-Americans. *Lupus Sci. Med.* 3, e000144.
8. Barnado, A., Carroll, R.J., Casey, C., Wheless, L., Denny, J.C., and Crofford, 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.

(C–F) GSA enrichment scores were calculated for ancestry-specific and independent gene signatures in patient WB (GEO: GSE88885). GSA signature scores (C) separating SLE-affected individuals (EA and AA) from control subjects, signature scores (D) distinguishing EA SLE-affected individuals from AA subjects and/or healthy control subjects, signature scores (E) distinguishing AA SLE-affected individuals from EA subjects or control subjects, and signature scores (F) separating SLE-affected individuals (EA and AA) from control subjects and that are additionally elevated in AA subjects compared to EA subjects. Error bars indicate 95% confidence interval. Asterisks (*) indicate a p value < 0.05 using Welch's t test comparing SLE to control; ^ indicates a p value < 0.05 using Welch's t test comparing EA to AA.

9. Goulielmos, G.N., Zervou, M.I., Vazgiourakis, V.M., Ghodke-Puranik, Y., Garyfallos, A., and Niewold, T.B. (2018). The genetics and molecular pathogenesis of systemic lupus erythematosus (SLE) in populations of different ancestry. *Gene* 668, 59–72.
10. Navarra, S.V., Guzmán, R.M., Gallacher, A.E., Hall, S., Levy, R.A., Jimenez, R.E., Li, E.K., Thomas, M., Kim, H.Y., León, M.G., et al.; BLISS-52 Study Group (2011). Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial. *Lancet* 377, 721–731.
11. Furie, R., Petri, M., Zamani, O., Cervera, R., Wallace, D.J., Tegzová, D., Sanchez-Guerrero, J., Schwarting, A., Merrill, J.T., Chatham, W.W., et al.; BLISS-76 Study Group (2011). A phase III, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits B lymphocyte stimulator, in patients with systemic lupus erythematosus. *Arthritis Rheum.* 63, 3918–3930.
12. Lamore, R., 3rd, Parmar, S., Patel, K., and Hilas, O. (2012). Belimumab (benlysta): a breakthrough therapy for systemic lupus erythematosus. *P&T* 37, 212–226.
13. Stohl, W. (2017). Inhibition of B cell activating factor (BAFF) in the management of systemic lupus erythematosus (SLE). *Expert Rev. Clin. Immunol.* 13, 623–633.
14. Fike, A.J., Elcheva, I., and Rahman, Z.S.M. (2019). The Post-GWAS Era: How to Validate the Contribution of Gene Variants in Lupus. *Curr. Rheumatol. Rep.* 21, 3.
15. Schadt, E.E., Monks, S.A., Drake, T.A., Lusk, A.J., Che, N., Colina, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302.
16. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdóttir, S., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423–428.
17. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 8, e1002639.
18. 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., Croker, J.A., Morris, D.L., et al. (2017). Trans-ancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* 8, 16021.
19. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585.
20. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243.
21. 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 (D1), D106–D112.
22. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017.
23. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9.
24. Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40 (Web Server issue, W1), W452–7.
25. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet. Chapter 7 (SUPPL.76)*, 20.
26. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 7, e46688.
27. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 45 (D1), D183–D189.
28. Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* 16, 169.
29. Ward, L.D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 44 (D1), D877–D881.
30. 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.
31. 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.
32. 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.
33. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14, 7.
34. 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.
35. Grammer, A.C., Ryals, M.M., Heuer, S.E., Robl, R.D., Madamanchi, S., Davis, L.S., Lauwerys, B., Catalina, M.D., and Lipsky, P.E. (2016). Drug repositioning in SLE: crowd-sourcing, literature-mining and Big Data analysis. *Lupus* 25, 1150–1170.
36. Chorley, B.N., Wang, X., Campbell, M.R., Pittman, G.S., Nour-edine, M.A., and Bell, D.A. (2008). Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat. Res.* 659, 147–157.
37. Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* 13, 101.
38. Alarcón-Segovia, D., Alarcón-Riquelme, M.E., Cardiel, M.H., Caeiro, F., Massardo, L., Villa, A.R., Pons-Estel, B.A.; and Grupo Latinoamericano de Estudio del Lupus Eritematoso (GLADEL) (2005). Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort. *Arthritis Rheum.* 52, 1138–1147.

39. Barber, M.R.W., and Clarke, A.E. (2017). Socioeconomic consequences of systemic lupus erythematosus. *Curr. Opin. Rheumatol.* **29**, 480–485.
40. Carter, E.E., Barr, S.G., and Clarke, A.E. (2016). The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat. Rev. Rheumatol.* **12**, 605–620.
41. van Arensbergen, J., Pagie, L., FitzPatrick, V.D., de Haas, M., Baltissen, M.P., Comoglio, F., van der Weide, R.H., Teunissen, H., Vösa, U., Franke, L., et al. (2019). High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169.
42. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al.; BLUEPRINT Consortium (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369–1384.e19.
43. Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A.M., Acs, P., Turner, J., Anguiano, E., et al. (2016). Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 1548–1550.
44. Catalina, M., Bachali, P., Yeo, A., Geraci, N., Petri, M., Grammer, A., and Lipsky, P. (2020). The Influence of Ancestry, Serology and Treatment on the Transcriptomic Fingerprint of Systemic Lupus Erythematosus. *JCI Insight* **5**, e140380.
45. Menard, L.C., Habte, S., Gonsiorek, W., Lee, D., Banas, D., Holloway, D.A., Manjarrez-Orduno, N., Cunningham, M., Stetsko, D., Casano, F., et al. (2016). B cells from African American lupus patients exhibit an activated phenotype. *JCI Insight* **1**, e87310.
46. Navid, F., and Colbert, R.A. (2017). Causes and consequences of endoplasmic reticulum stress in rheumatic disease. *Nat. Rev. Rheumatol.* **13**, 25–40.
47. Lam, W.Y., and Bhattacharya, D. (2018). Metabolic Links between Plasma Cell Survival, Secretion, and Stress. *Trends Immunol.* **39**, 19–27.
48. Lightfoot, Y.L., Blanco, L.P., and Kaplan, M.J. (2017). Metabolic abnormalities and oxidative stress in lupus. *Curr. Opin. Rheumatol.* **29**, 442–449.
49. Wehbi, V.L., and Taskén, K. (2016). Molecular mechanisms for cAMP-mediated immunoregulation in T cells - role of anchored protein kinase a signaling units. *Front. Immunol.* **7**, 222.
50. Kammer, G.M., Khan, I.U., and Malemud, C.J. (1994). Deficient type I protein kinase A isozyme activity in systemic lupus erythematosus T lymphocytes. *J. Clin. Invest.* **94**, 422–430.
51. Kammer, G.M. (2002). Deficient protein kinase a in systemic lupus erythematosus: a disorder of T lymphocyte signal transduction. *Ann. N Y Acad. Sci.* **968**, 96–105.
52. Cartier, A., and Hla, T. (2019). Sphingosine 1-phosphate: Lipid signaling in pathology and therapy. *Science* **366**, eaar5551.
53. Aoki, M., Aoki, H., Ramanathan, R., Hait, N.C., and Takabe, K. (2016). Sphingosine-1-Phosphate Signaling in Immune Cells and Inflammation: Roles and Therapeutic Potential. *Mediators Inflamm.* **2016**, 8606878.
54. Kappos, L., Bar-Or, A., Cree, B.A.C., Fox, R.J., Giovannoni, G., Gold, R., Vermersch, P., Arnold, D.L., Arnould, S., Scherz, T., et al.; EXPAND Clinical Investigators (2018). Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet* **391**, 1263–1273.
55. Faissner, S., and Gold, R. (2019). Progressive multiple sclerosis: latest therapeutic developments and future directions. *Ther. Adv. Neurol. Disorder.* **12**, 1756286419878323.
56. Gajofatto, A. (2017). Spotlight on siponimod and its potential in the treatment of secondary progressive multiple sclerosis: the evidence to date. *Drug Des. Devel. Ther.* **11**, 3153–3157.
57. Kingsmore, K., Grammer, A., and Lipsky, P. (2020). Drug repurposing to improve treatment of rheumatic autoimmune inflammatory diseases. *Nature Review Rheumatology* **16**, 32–52.
58. Robak, P., and Robak, T. (2019). Bortezomib for the Treatment of Hematologic Malignancies: 15 Years Later. *Drugs R D.* **19**, 73–92.
59. Alexander, T., Sarfert, R., Klotsche, J., Köhl, A.A., Rubbert-Roth, A., Lorenz, H.M., Rech, J., Hoyer, B.F., Cheng, Q., Waka, A., et al. (2015). The proteasome inhibitor bortezomib depletes plasma cells and ameliorates clinical manifestations of refractory systemic lupus erythematosus. *Ann. Rheum. Dis.* **74**, 1474–1478.
60. Merrill, J.T., Newwelt, C.M., Wallace, D.J., Shanahan, J.C., Latinis, K.M., Oates, J.C., Utset, T.O., Gordon, C., Isenberg, D.A., Hsieh, H.J., et al. (2010). Efficacy and safety of rituximab in moderately-to-severely active systemic lupus erythematosus: the randomized, double-blind, phase II/III systemic lupus erythematosus evaluation of rituximab trial. *Arthritis Rheum.* **62**, 222–233.
61. Rovin, B.H., Furie, R., Latinis, K., Looney, R.J., Fervenza, F.C., Sanchez-Guerrero, J., Maciuga, R., Zhang, D., Garg, J.P., Brunetta, P., Appel, G.; and LUNAR Investigator Group (2012). Efficacy and safety of rituximab in patients with active proliferative lupus nephritis: the Lupus Nephritis Assessment with Rituximab study. *Arthritis Rheum.* **64**, 1215–1226.
62. Wallace, D.J., Stohl, W., Furie, R.A., Lisse, J.R., McKay, J.D., Merrill, J.T., Petri, M.A., Ginzler, E.M., Chatham, W.W., McCune, W.J., et al. (2009). A phase II, randomized, double-blind, placebo-controlled, dose-ranging study of belimumab in patients with active systemic lupus erythematosus. *Arthritis Rheum.* **61**, 1168–1178.
63. Jacobi, A.M., Huang, W., Wang, T., Freimuth, W., Sanz, I., Furie, R., Mackay, M., Aranow, C., Diamond, B., and Davidson, A. (2010). Effect of long-term belimumab treatment on B cells in systemic lupus erythematosus: extension of a phase II, double-blind, placebo-controlled, dose-ranging study. *Arthritis Rheum.* **62**, 201–210.

SOFTWARE

Open Access



Hierarchicell: an R-package for estimating power for tests of differential expression with single-cell data

Kip D. Zimmerman^{1*}  and Carl D. Langefeld^{1,2,3*}

Abstract

Background: Study design is a critical aspect of any experiment, and sample size calculations for statistical power that are consistent with that study design are central to robust and reproducible results. However, the existing power calculators for tests of differential expression in single-cell RNA-seq data focus on the total number of cells and not the number of independent experimental units, the true unit of interest for power. Thus, current methods grossly overestimate the power.

Results: *Hierarchicell* is the first single-cell power calculator to explicitly simulate and account for the hierarchical correlation structure (i.e., within sample correlation) that exists in single-cell RNA-seq data. *Hierarchicell*, an R-package available on GitHub, estimates the within sample correlation structure from real data to simulate hierarchical single-cell RNA-seq data and estimate power for tests of differential expression. This multi-stage approach models gene dropout rates, intra-individual dispersion, inter-individual variation, variable or fixed number of cells per individual, and the correlation among cells within an individual. Without modeling the within sample correlation structure and without properly accounting for the correlation in downstream analysis, we demonstrate that estimates of power are falsely inflated. *Hierarchicell* can be used to estimate power for binary and continuous phenotypes based on user-specified number of independent experimental units (e.g., individuals) and cells within the experimental unit.

Conclusions: *Hierarchicell* is a user-friendly R-package that provides accurate estimates of power for testing hypotheses of differential expression in single-cell RNA-seq data. This R-package represents an important addition to single-cell RNA analytic tools and will help researchers design experiments with appropriate and accurate power, increasing discovery and improving robustness and reproducibility.

Keywords: Hierarchical data, Single-cell, RNA-sequencing, Power calculator, Simulation, R-package, Mixed-effects models

Background

Robust and reproducible science depends on the quality of the experimental design. High quality experimental design revolves around focused research questions or hypotheses, appropriate and valid measures of the central variables related to these hypotheses, statistically sound analysis plans, and properly computed power

analysis [1]. While power analyses for genetic association studies and bulk RNA-seq approaches are well-established [2–6], such analyses remain a challenge in single-cell RNA-seq studies due to intra-sample correlation inherent in these data [7]. Such within sample correlations exist because cells from the same individual share a common genetic and environmental background that often leads to greater similarity in gene expression among cells in the same sample. Therefore, gene expression measures among cells from the same sample have a hierarchical correlation structure where cells nested within an individual are not

* Correspondence: kdzimmer@wakehealth.edu; clange@wakehealth.edu

¹Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **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 Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

independent units. At present, the correlative (hierarchical) nature of these data is often neglected, in both power analyses and tests of hypotheses (e.g., differential expression) [7]. This was recently highlighted in a valuable paper by Andrews et al. which states that “current single-cell differential expression tests treat each individual cell as a biological replicate and cannot account for shared genetic backgrounds or disease state”. [8] Ignoring the hierarchical nature of single-cell RNA-seq data leads to studies that are under powered and inappropriately analyzed [7], leading to incorrect inference, poor reproducibility, and financial investments in those errors. A contributor to these flawed practices is the void of single-cell specific methods and literature that properly account for this hierarchical structure. However, just as Andrews et al. pointed out, we – too - expect that, “as scRNA-seq is applied to larger cohorts and comparison studies, [there will be] further developments that lead to more accurate statistical models for more complex experimental designs.” An excellent starting place for more accurate statistical models are accurate power calculations for improved study design.

Besides the classic, closed form, normal theory power calculations (e.g., ANOVA) that make too many overly simplistic assumptions (e.g., normality, independence), the power calculators for testing single-cell RNA differential expression all simulate cells independently, without the within-subject correlation structure [9–11]. Previously, we documented that in tests of differential expression in single-cell RNA-seq data one needs to account for the within experimental unit (e.g., individual) and showed that mixed-effects models with subject/individual as a random effect is a practical and statistically sound approach for these hypotheses [7]. Here, we present an R-package, *Hierarchicell*, with two purposes: 1) it is a simulator of hierarchical single-cell RNA-seq data, and 2) it computes power estimates using a mixed-effects models for testing hypotheses of differential gene expression in single-cell RNA-seq data. *Hierarchicell* simulates single-cell RNA-seq data with a hierarchical structure that closely resembles that of real data and can be used by researchers to make informed choices on experimental design while balancing the trade-off between cost and power. Our R-package is user friendly and flexible to a variety of scenarios. It incorporates estimates from real data [12] or allows users to input data (e.g., either Fluidigm C1 or 10x Chromium technology, user's own pilot data) to obtain highly translatable and accurate estimates of power tailored to their technology. Within a well-characterized set of parameters that are modeled from either a user-defined or the default single-cell RNA-seq data, the tool provides users with estimates of power relative to a given fold-change, significance threshold, number of independent samples, and number of cells per independent sample. In addition, the calculator

allows for the simulation of either continuous or binary phenotypes of interest. For binary case-control analyses, the user specifies the fold-change they desire to detect. For continuous phenotypes, the user specifies the mean and standard deviation of the phenotype and the degree of correlation with expression the user desires to detect with significance. Currently, most single-cell power calculators only provide estimates for the required number of cells rather than the required number of independent experimental units (e.g. individuals) or are not designed for computing power to detect differences in expression [13–15]. Other power calculations for single-cell RNA-seq are based on bulk RNA-seq methods to estimate the required number of samples [2, 3]. Estimating power for a single-cell RNA-seq study using bulk RNA-seq power calculators is a reasonable solution, but will underestimate the study's power by not incorporating the additional power gained by sequencing numerous cells per individual. This tool provides a valuable resource in an area of critical need for researchers looking to optimize their study's power and experimental design relative to the hierarchical nature that exists in all single-cell data.

Implementation

Overview of the *Hierarchicell* simulation engine

A step-by-step overview of the simulation procedure is provided with R-code examples and detailed explanations in *Hierarchicell*'s accompanying vignette. We encourage users to review this vignette (available on GitHub and in the [Supplementary Materials](#)) before beginning to work with *Hierarchicell*. The single-cell data in that example are used to estimate default simulation parameters for our simulation engine. These data were downloaded from the public accession number E-MTAB-5061 [12]. These data were sequenced using the Smart-Seq2 protocol and they include sequence data from 3514 cells from 10 different individuals [12]. Genes were previously normalized to account for the differences in library size [12]. After filtering down to high quality alpha cells, our dataset contained gene expression values for 22,983 genes and 886 cells (across 10 individuals). This dataset is included as part of the R-package for a number of reasons. Primarily, these data demonstrate the general intra- and inter-individual correlation patterns seen across a variety of single-cell data of different cell types generated by different platforms [7]. In addition, these data are not too large, allowing for the rapid estimation of simulation parameters while also minimizing the size of the R-package.

The simulation procedure was designed to simulate independent genes in a way that approximates the hierarchical structure of real data by empirically estimating the range of parameters (i.e., grand mean of the transcript-

per-million (TPM) values, within sample variance, between sample variance, relationship between the grand mean and dispersion, dropout) that define the observed distribution of TPM values for a gene. To estimate these parameters, genes were pruned to a set of uncorrelated genes that captured the most representative patterns of detectable TPM values, without the resulting parameter estimates being primarily driven by dropout. Specifically, genes were sequentially sampled one at a time and any other gene having transcript abundances correlated (Spearman's correlation coefficient > 0.25) with the gene were removed. To estimate the grand means independently from the hierarchical correlation structure, the grand means were estimated by sampling one cell from each individual and computing the mean TPM value 1000 times. The mean of each of those means was used to approximate the grand mean. To approximate the variance of the within-sample means (inter-individual variance), the means of all non-zero TPM values were computed across all cells within each individual and the variance between those values was subsequently computed. To estimate the within-sample dispersion values, the non-zero TPM values were first used to compute a within-sample variance and within-sample mean. Consistent with the classical definition of the Negative Binomial distribution's dispersion parameter, the within-sample dispersion parameter was then computed as:

$$\alpha_{ij} = \frac{\mu_{ij}^2}{\sigma_{ij}^2} - \mu_{ij} \quad (1)$$

where α_{ij} represents the dispersion parameter, μ_{ij} represents the within-sample mean, and σ_{ij}^2 represents the within-sample variance for gene i and individual j . The grand means and variances were computed empirically from the TPM values previously reported in six different cell types across three different single-cell studies [12, 16, 17]. Once consistent patterns were identified across cell types, alpha cells from the pancreas dataset, were used as the model data for our simulation. A gamma distribution was fit to the global mean of the TPM values of each gene using maximum-likelihood estimation. For each independently simulated gene i , a random value was sampled from this gamma distribution to obtain a grand mean, μ_i . The variance of the within-sample means (inter-individual variance) was modeled as a linear function of the grand means, $f_1(\mu_i)$ and the within-sample dispersion (intra-individual variance) was estimated as a logarithmic function of the within-sample means, $f_2(\mu_i)$, and the probability of dropout was estimated independently as a bounded gamma distribution (Fig. 1). Using a normal distribution with an expected value of zero and a variance computed by the first linear relationship, $f_1(\mu_i)$, a difference in means was

drawn for each of the individuals j in the simulation. This difference was summed with the grand mean to obtain an individual mean, μ_{ij} . Three different methods were used to simulate the number of cells per individual. To simulate scenarios where each of the individuals had the exact same number of cells, the number of cells desired for each individual was fixed at a constant value. In order to simulate scenarios where the number of cells per individual demonstrated slight imbalance, a Poisson distribution with a λ equal to the expected number of cells desired for each individual was then used to obtain the count of cells for each individual. To simulate scenarios where the number of cells per individual demonstrated greater imbalance, the number of cells per individual were modeled as a Negative Binomial random variable with a mean equal to the expected number of cells and a dispersion parameter of one. For each gene i and cell k assigned to an individual j , a read count value, Y_{ijk} , was drawn from a Negative Binomial distribution with an expected value equal to the individual's assigned read count value, μ_{ij} , and a dispersion parameter, α_{ij} , computed by the logarithmic function of the grand mean $f_2(\mu_i)$.

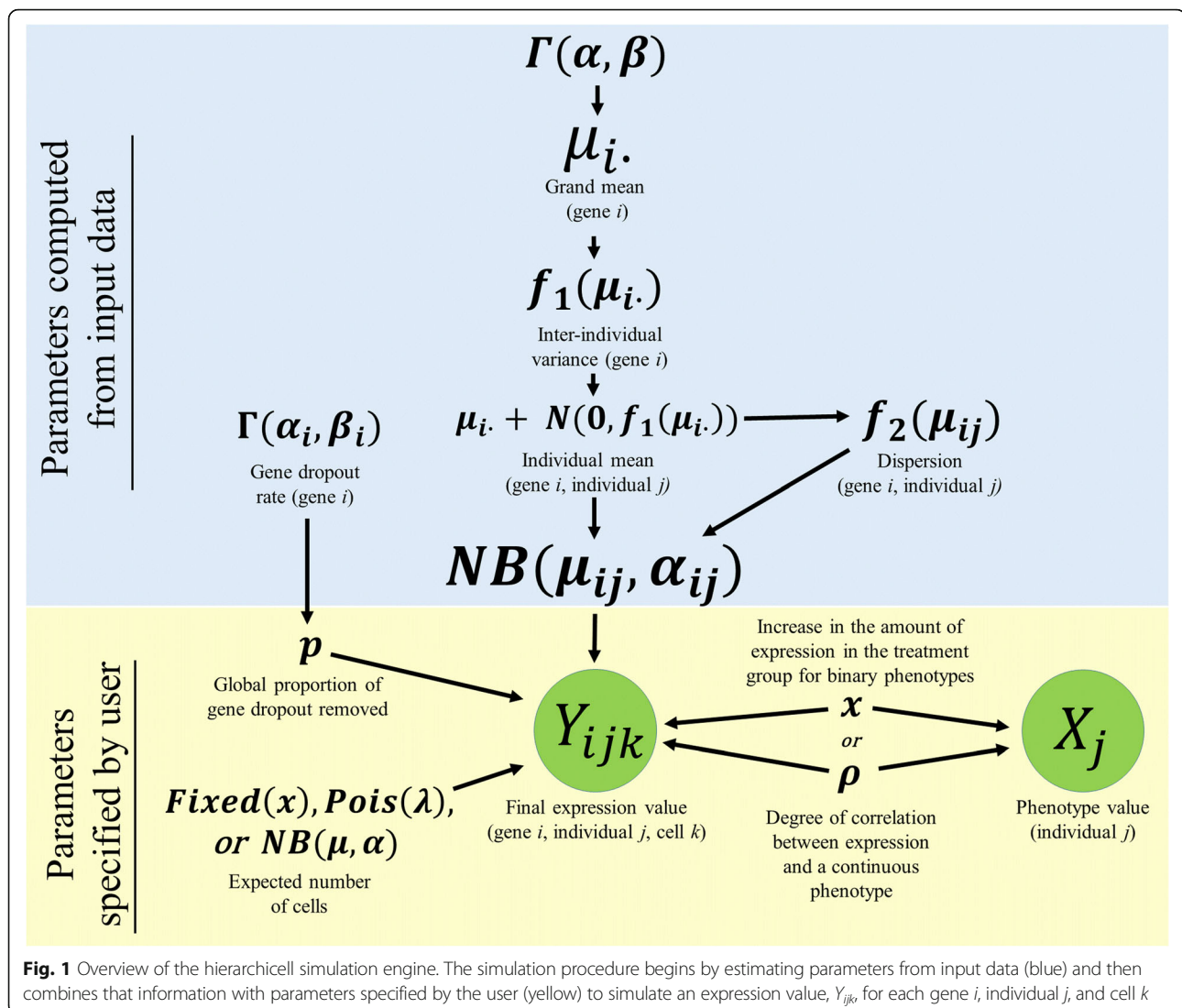
Overview of *Hierarchicell* power calculations

To compute power, transcripts-per-million (TPM) values were simulated for each gene with the user-specified fold-change or ρ parameter (Fig. 1). Fold-change should be specified where users are interested in computing power for two distinct groups. The ρ parameter, which represents the degree of correlation between gene expression and a simulated continuous phenotype, should be specified where users are interested in computing power for association analysis with a continuous trait. Here, fold-change is a constant that was multiplied by the global mean gene expression values to spike the expression of those genes in the desired treatment group. The direction of the fold-change was simulated with a Bernoulli distribution with a probability of 0.5 to allow the direction of effect to vary equally between groups.

We applied a two-part hurdle model with a random effect for individual as directed in MAST's reference manual (7,18). Specifically, a $\log(x + 1)$ transformation of the data was applied and the hurdle model computed to find genes exhibiting differences in expression. Using their same notation, the addition of random effects for differences among persons is as follows:

$$\begin{aligned} \text{logit}(\Pr(Z_{ig} = 1|X_i)) &= X_i\beta_g + W_i\gamma_j \\ \Pr(Y_{ig} = y|Z_{ig} = 1) &= N(X_i\beta_g + W_i\gamma_j, \sigma_g^2) \end{aligned} \quad (2)$$

where Y_{ig} is the expression level for gene g and cell i , Z_{ig} is an indicator for whether gene g is expressed in cell



i , X_i contains the predictor variables for each cell i , and W_i is the design matrix for the random effects of each cell i belonging to each individual j (i.e., the random complement to the fixed X_i). β_g represents the vector of fixed-effects regression coefficients and γ_j represents the vector of random effects (i.e., the random complement to the fixed β_g). γ_j is distributed normally with a mean of zero and variance σ_g^2 . To obtain a single result for each gene, likelihood ratio or Wald test results from each of the two components are summed and the corresponding degrees of freedom for each component are added. These tests have asymptotic χ^2 null distributions; these statistics can be summed and remain asymptotically χ^2 because Z_g and Y_g are defined conditionally independent for each gene. When summed together, these tests provide a single test for the two-part hurdle model. Our package also offers the ability to compute type 1 error rates (and thereby power) for a variety of different

single-cell analysis approaches. New methods that properly handle within sample correlation will be integrated as they become available.

Software implementation

All simulations and data were compiled in RStudio using R-3.6.2 and is freely available on GitHub. The supplementing dataset that is included to run the R-package without user input data was significantly downsized by removing all of the genes correlated with a Spearman's correlation coefficient > 0.25 . This filtering is one of the first steps in our simulation procedure and doing so greatly reduced the size of the source package to 475 KB as well as the data structures held in memory during use. Currently, the simulation typically completes in less than 5 seconds, depending on user specifications. The simulation-based power calculations, however, can take much longer depending on the model that is used. For

the recommended two-part hurdle mixed model (MAST with a random effect for individual), this can range anywhere from 1 to 20 min per simulation-based estimate of power for a given fold-change on a 64-bit Operating system with 8 CPUs and 16 GB of RAM. We note that these run times are heavily dependent on the number of genes, the sample size, the number of cells per individual specified, and the number of CPUs available.

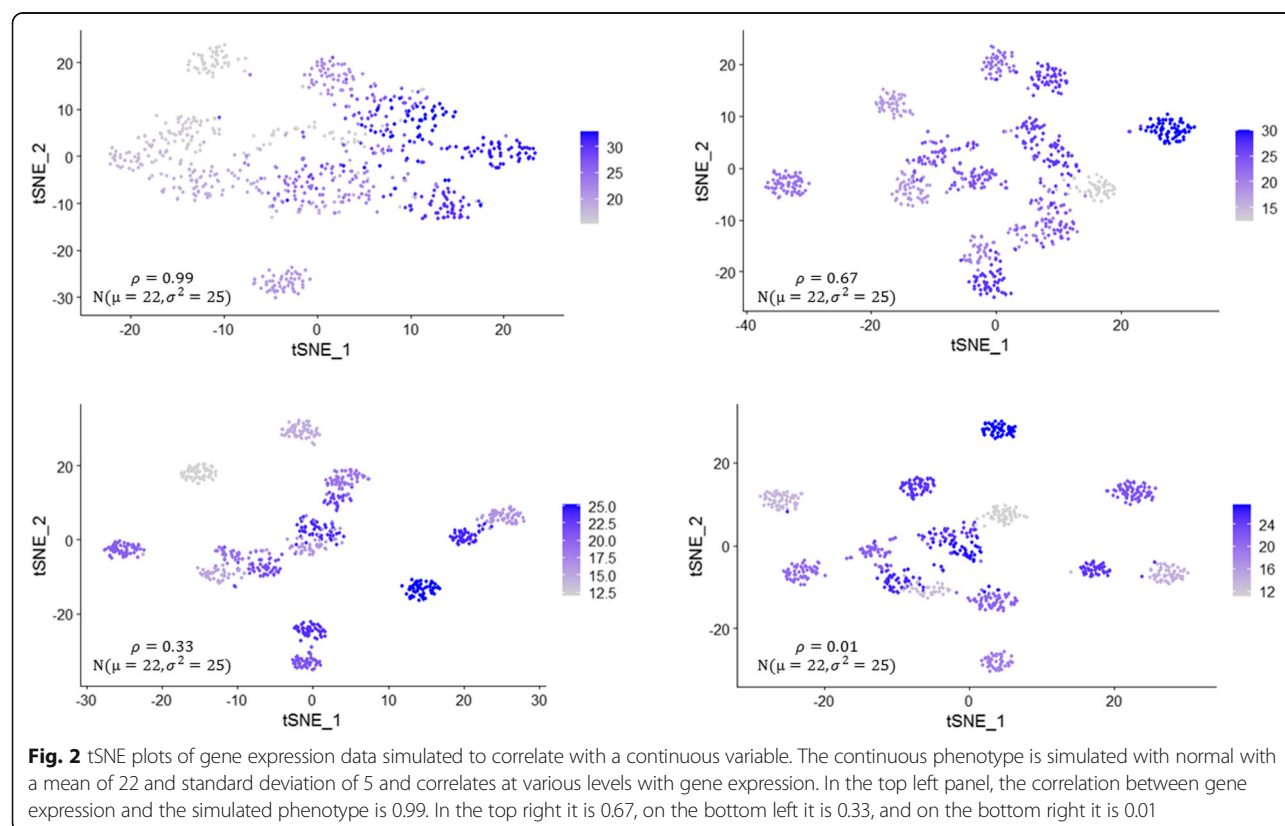
Results and discussion

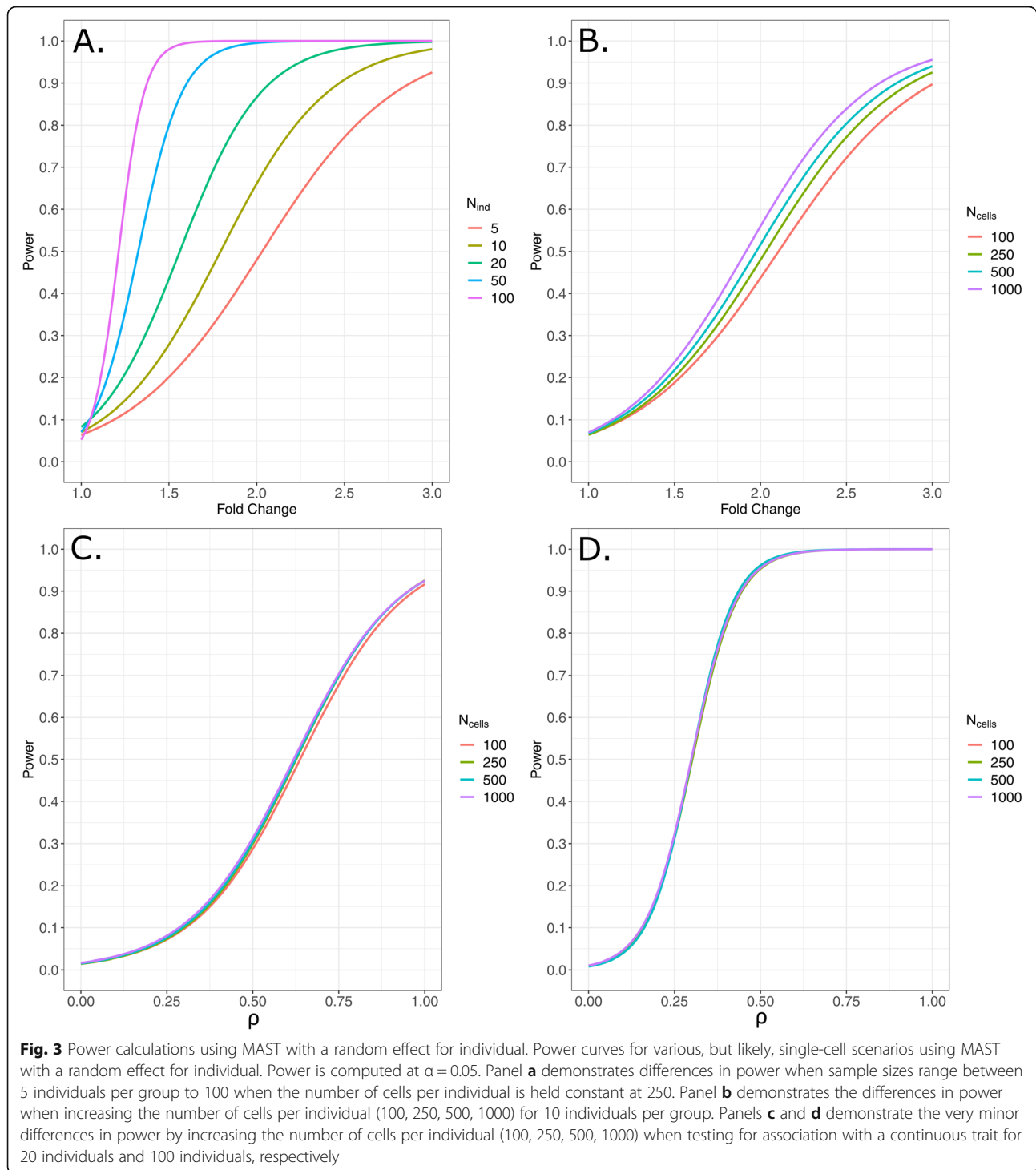
Previously, we demonstrated that our simulation recapitulates the most important aspects of single-cell gene expression data, particularly the hierarchical structure of single-cell RNA-seq data which is rarely accounted for in differential expression analysis [7]. We also applied our simulation engine to demonstrate that mixed models are a statistically sound method that accounts for the within sample correlation and has appropriate type 1 error control without sacrificing power [7]. In addition, we provided power estimates for binary outcomes across a range of experimental conditions to assist researchers in designing appropriately powered studies [7].

We previously reported power calculations for tests of differential expression in single-cell RNA-seq studies for binary phenotypes (i.e., case/control treatment groups) [7]. *Hierarchicell* now also allows users to estimate power for detecting associations between continuous

traits and single-cell gene expression (Fig. 2). The simulated expression data can be computed over a range of correlations with the magnitude of expression in each individual's cells, while accounting for the hierarchical structure of these data.

As expected, increasing the number of independent experimental units (e.g., individuals) in a study is the best way to increase power to detect true differences between traits measured at the individual and not individual cell level (Fig. 3a). Power calculations for binary phenotypes consisting of 10 individuals per treatment group reveal that there are only marginal gains in power when more than 100 cells per individual are sampled for a particular analysis unit (Fig. 3b). We also note that methods that do not account for within person correlation grossly overestimate power. For example, when estimating power with an approach that estimates the power for cells as independent units (assuming a type 1 error rate of $\alpha = 0.05$, a fold change of 1.3, 10 individuals per treatment and 100 cells per individual), the power is overestimated as 0.93 instead of 0.71 when appropriately accounting for the within person correlation. Power calculations for continuous phenotypes, with the same sample sizes and constant within-person correlations among cells, demonstrate even smaller gains in power when more than 100 cells per individual are sampled for a particular analysis unit (Fig. 3c). The gains in power





from sampling more cells per individual will decrease as the numbers of independent experimental units increase (Fig. 3d). This is true for both types of analyses. As the degree of correlation among cells within a person decreases and approaches zero, rarely observed, the value tradeoff between independent experimental units and individual cells will vary. Further, we note that if the

cell-type of interest has much more or much less zero-inflation (i.e., less information), then the gains in power from sampling more cells may be greater or smaller, respectively. This is why estimating the data structure of the cell types of interest from preliminary data is a critically important feature of our *Hierarchicell* package. To consistently identify fold-change differences of at

least 1.2 as statistically significant (power > 0.80), we approximate that researchers will need a minimum of 40 samples per group and 100 cells per sample in a classical case/control design. To consistently identify genes correlated with a correlation coefficient of 0.4 with a phenotype (power > 0.80), we approximate that researchers will need a minimum of 100 samples and 100 cells per sample.

As experiments get larger, computational time will increase. Future work will parallelize the code. To more rapidly close in on plausible sample size options, a researcher can apply the aggregate (“pseudo-bulk”) methods power estimates and as one approaches feasible design shift to refining the estimates using the two-part hurdle mixed model employed here. However, it is important to do this refining step given the differences between these two approaches and the types of scenarios where aggregation methods will be underpowered [7, 18].

Future iterations of this package will incorporate any novel single-cell RNA-seq differential expression methods that properly account for the within sample correlation. In addition, we will parallelize the code and improve the speed of software by building components of the software in other languages (such as C++ via *rcpp*), and/or storing results of a large number of scenarios for quick and easy access to the necessary information. Future developments would be to incorporate the relationship between power and the variance explained by an effect, not simply fold-change between treatments. In real data, the variances explained by an effect fluctuate greatly among genes and cell types. While the simulated expression data herein have variances that are modeled after real data and are allowed to fluctuate by genes, simulating a direct relationship between the variance and an effect will be a meaningful addition to this work.

Conclusions

To date, none of the primary power calculation methods are directly applicable for differential expression analysis with single-cell RNA-seq data. Here, we present an R-package, *Hierarchicell*, with two purposes: 1) simulation of hierarchical single-cell RNA-seq data, and 2) computation of power estimates using a mixed-effects models for testing hypotheses of differential gene expression in single-cell RNA-seq data. *Hierarchicell* allows for a range of inputs and parameter settings and even the evaluation of various single-cell specific methods, but encourages using linear mixed models with individual as a random effect for both binary and continuous outcomes, as implemented in MAST [7, 19]. We recommend these mixed effects models because they retain appropriate type 1 error rates while maintaining power. Proper calculation of statistical power coupled with proper analysis methods that account for the correlation among cells

from the same individual will increase robustness and reproducibility of single-cell studies, thereby reducing the cost while accelerating the rate of scientific discovery.

Availability and requirements

- Project name: *hierarchicell*
- Project home page: <https://github.com/kdzimm/hierarchicell>
- Operating system(s): Linux, Mac, and PC
- Programming language: R
- Other requirements: No
- License: CC0
- Any restrictions to use by non-academics: No

Abbreviations

RNA-Seq: Sequencing technique which uses next-generation sequencing to reveal the presence and quantity of RNA; MAST: Model-based Analysis of Single-cell Transcriptomics

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-07635-w>.

Additional file 1.

Acknowledgments

Not applicable.

Authors' contributions

CDL and KDZ conceived the study. KDZ constructed the R-package and implemented all simulations and analyses with guidance from CDL. KDZ wrote the original draft and reviewed and edited it with CDL. All authors approved the final version of the manuscript.

Funding

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

Availability of data and materials

The data used for the simulated data herein and in the long-form documentation (R vignette) are available under the accession number E-MTAB-5061 at <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5061/>. The R-package is freely available on GitHub at <https://github.com/kdzimm/hierarchicell> (DOI: <https://doi.org/10.5281/zenodo.4608738>).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Authors declare no competing interests.

Author details

¹Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA. ²Department of Biostatistics and Data Science,

Wake Forest School of Medicine, Winston-Salem, NC, USA. ³Comprehensive Cancer Center, Wake Forest Baptist Medical Center, Winston-Salem, NC, USA.

Received: 11 February 2021 Accepted: 21 April 2021

Published online: 01 May 2021

References

- Nayak BK. Understanding the relevance of sample size calculation. *Indian J Ophthalmol*. 2010;58(6):469–70. <https://doi.org/10.4103/0301-4738.71673>.
- Guo Y, Zhao S, Li C-I, Sheng Q, Shyr Y. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Cancer Inform*. 2014; 13(Suppl 6):1–5. <https://doi.org/10.4137/CIN.S17688>.
- Zhao S, Li C-I, Guo Y, Sheng Q, Shyr Y. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. *BMC Bioinformatics*. 2018;19(1): 191. <https://doi.org/10.1186/s12859-018-2191-5>.
- Wang M, Xu S. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity*. 2019;123(3):287–306. <https://doi.org/10.1038/s41437-019-0205-3>.
- Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inf*. 2012;10(2):117–22. <https://doi.org/10.5808/GI.2012.10.2.117>.
- Johnson JL, Abecasis GR. GAS Power Calculator: web-based power calculator for genetic association studies. *bioRxiv*. 2017;17:164343.
- Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun*. 2021;12(1):738. <https://doi.org/10.1038/s41467-021-21038-1>.
- Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc*. 2021;16(1):1–9. <https://doi.org/10.1038/s41596-020-00409-w>.
- Baruzzo G, Patuzzi I, Di Camillo B. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics*. 2020;36(5):1468–75. <https://doi.org/10.1093/bioinformatics/btz752>.
- Li WW, Li JJ. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics*. 2019;35(14):i41–50. <https://doi.org/10.1093/bioinformatics/btz321>.
- Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;18(1). [cited 2019 Jan 23]. Available from: <https://doi.org/10.1186/s13059-017-1305-0>
- Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, et al. Single-cell Transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab*. 2016;24(4):593–607. <https://doi.org/10.1016/j.cmet.2016.08.020>.
- Kim KI, Youn A, Bolisetty M, Palucka AK, George J. Calculating sample size for identifying cell subpopulation in single-cell RNA-seq experiments. *bioRxiv*. 2019;18:706481.
- Davis A, Gao R, Navin NE. SCOPIT: sample size calculations for single-cell sequencing experiments. *BMC Bioinformatics*. 2019;20(1):566. <https://doi.org/10.1186/s12859-019-3167-9>.
- Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*. 2017; 33(21):3486–8. <https://doi.org/10.1093/bioinformatics/btx435>.
- Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189–96. <https://doi.org/10.1126/science.aad0501>.
- Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JLL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*. 2017;49(5):708–18. <https://doi.org/10.1038/ng.3818>.
- Tirrell TF, Rademaker AW, Lieber RL. Analysis of hierarchical biomechanical data structures using mixed-effects models. *J Biomech*. 2018;69:34–9.
- Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015;16(1) Available from: <http://genomebiology.com/2015/16/1/278>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Article

Nucleic acid sensing and interferon-inducible pathways show differential methylation in MZ twins discordant for lupus and overexpression in independent lupus samples: Implications for pathogenic mechanism and drug targeting

Miranda C. Marion^{1,5}, Paula S. Ramos², Prathyusha Bachali³, Adam C. Labonte³, Kip D. Zimmerman⁵, Hannah C. Ainsworth^{1,5}, Sarah E. Heuer^{3,4}, Robert D. Robl³, Michelle D. Catalina³, Jennifer A. Kelly⁶, Timothy D. Howard⁷, Peter E. Lipsky³, Amrie C. Grammer³, Carl D. Langefeld^{1,5*}

¹ Department of Biostatistics and Data Science, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA

² Division of Rheumatology and Immunology, Department of Medicine, Medical University of South Carolina, Charleston, SC, USA

³ AMPEL BioSolutions, LLC and RILITE Research Institute, Charlottesville, VA, USA

⁴ The Jackson Laboratory, Tufts Graduate School of Biomedical Sciences, 600 Main Street Bar, Harbor, ME, 04609, USA.

⁵ Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA

⁶ Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA

⁷ Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA.

Abstract: Systemic lupus erythematosus (SLE) is a chronic, multisystem, autoimmune inflammatory disease with genomic and non-genomic contributions to risk. We hypothesize that epigenetic factors are a significant contributor to SLE risk and may be informative for identifying pathogenic mechanisms and therapeutic targets. To test this hypothesis while controlling for genetic background, we performed an epigenome-wide analysis of DNA methylation in genomic DNA from whole blood in three pairs of female monozygotic (MZ) twins of European-ancestry, discordant for SLE. Results were replicated on the same array in four cell types from a set of four Danish female MZ twin pairs discordant for SLE. Genes implicated by the epigenetic analyses were then evaluated in 10 independent SLE gene expression datasets from the Gene Expression Omnibus (GEO).

There were 59 differentially methylated loci between unaffected and affected MZ twins in whole blood, including 11 novel loci. All but two of these loci were hypomethylated in the SLE twins relative to the unaffected twins. The genes harboring these hypomethylated loci exhibited increased expression in multiple independent datasets of SLE patients. This pattern was largely consistent regardless of disease activity, cell type, or renal tissue type. The genes proximal to CpGs exhibiting differential methylation (DM) in the SLE-discordant MZ twins and exhibiting differential expression (DE) in independent SLE GEO cohorts (DM-DE genes), clustered into two pathways: the nucleic acid sensing pathway and the type I interferon pathway. The DM-DE genes were also informatically queried for potential gene-drug interactions, yielding a list of 41 drugs including a known SLE therapy.

The DM-DE genes delineate two important biologic pathways which are not only reflective of the heterogeneity of SLE but may also correlate with distinct IFN responses that depend on the source, type, and location of nucleic acid molecules and the activated receptors in individual patients. Cell- and tissue-specific analyses will be critical to the understanding of genetic factors dysregulating the nucleic acid sensing and IFN pathways, and whether those factors could be appropriate targets for therapeutic intervention.

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

Citation:

Academic Editor:

Received:

Accepted:

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

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

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

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

2. Materials and Methods

2.1. Discovery Cohort

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

2.2 Replication cohort

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

2.3 Genome-wide DNA methylation assay and array validation in LFRR twins

Genomic DNA (1 µg) from each individual was treated with sodium bisulfite using the EZ 96-DNA methylation kit (Zymo Research, CA, USA), following the manufacturer's standard protocol. Genome-wide DNA methylation was assessed using the Illumina Infinium HumanMethylation450 BeadChip (Illumina, Inc., CA, USA), which interrogates over 485,500 CpG sites that cover 99% of RefSeq genes (including the promoter, 5'UTR, first exon, gene body, and 3'UTR), as well as 96% of CpG islands and island shores. Arrays were processed using the manufacturer's standard protocol, with both members of each twin pair being hybridized to the same row on the microarray to minimize batch effects. GenomeStudio software (Illumina, Inc.) was used to perform initial quality control and to calculate the relative methylation level of each interrogated cytosine, which is reported as a β -value given by the ratio of the normalized signal from the methylated probe to the sum of the normalized signals of the methylated and unmethylated probes. This β -value for each CpG site ranges from 0 (unmethylated) to 1 (fully methylated). CpG loci with a stringent detection P-value $> 1.0 \times 10^{-5}$ in any of the samples were excluded ($n = 2,118$ probes) to control for poor-quality assays. Validation of the array data in the LFRR twins was performed by pyrosequencing two of the most significant CpGs probes: cg13304609 (in *IFI44L*) and cg23570810 (in *IFITM1*). The correlations between the methylation proportions from the array and pyrosequencing for these two probes were $r^2 = 0.98$ and $r^2 = 0.99$, respectively.

2.4 Collection of gene expression experiments from SLE patient datasets

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

2.5 Data analysis

To identify differentially methylated genes between unaffected and SLE-affected twins, a paired t-test on the probe-specific β -values was computed separately for the discovery and replication twin datasets. For the discovery set, CpG sites meeting 1) the Benjamini-Hochberg false discovery rate (FDR) [36] threshold $P_{FDR} < 0.05$ (equivalent to $P < 1.06 \times 10^{-7}$) and 2) a mean DNA methylation difference of $(\Delta\beta) > |0.085|$ were considered statistically significant; the mean methylation difference threshold was obtained by maximizing the area under the receiver operator characteristic curve (AUC) as a function of the β -value (described below). The genes related to the differentially methylated CpG sites (as annotated by Illumina for the HumanMethylation450) were queried in the Interferome online database to identify interferon regulated genes [37]. In addition, significant CpG sites were investigated for evidence of association between DNA methylation pattern and

gene expression (mQTL) using the iMETHYL genome browser.[38] These results are based on 100 healthy subjects with RNA-seq data and DNA methylation data in CD4T cells, monocytes and PBMC.

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

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

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

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

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

3. Results

3.1 Characteristics of the MZ twins

The LFRR MZ twins were all females of European ancestry, and the SLE-diagnosed twins exhibited a range of SLE clinical conditions (Supplemental Table S2). The Danish

MZ twins were also all females of European ancestry. Clinical characteristics such as number of ACR criteria, SLEDAI score, autoantibodies and medications are described in Ulff-Møller et. al., but were not available in GEO [14].

3.2. Identification of differentially methylated regions in twins discordant for SLE

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

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

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

243

CpG*	Chr	Pos (bp) [†]	Gene	$\Delta\beta$				P-value	Interferon-Regulated [‡]	Relation to CpG ^{††}
				pair1	pair2	pair3	mean			
cg13304609	1	79085162	<i>IFI44L</i>	-0.24	-0.27	-0.37	-0.29	1.58x10 ⁻¹⁴	IRG	
cg06872964	1	79085250	<i>IFI44L</i>		-0.26	-0.21	-0.24	1.05x10 ⁻⁷¹	IRG	
cg03607951	1	79085586	<i>IFI44L</i>	-0.27	-0.3	-0.21	-0.26	7.23x10 ⁻²²	IRG	
cg17515347	1	159047163	<i>AIM2</i>	-0.09	-0.11	-0.07	-0.09	3.01x10 ⁻¹²	IRG	
cg08272268	1	200380059	<i>ZNF281</i>	-0.08	-0.07	-0.11	-0.09	4.33x10 ⁻¹⁵		S_Shore
cg01028142	2	7004578	<i>CMPK2</i>	-0.22	-0.36	-0.43	-0.33	7.98x10 ⁻⁸	IRG	N_Shore
cg10959651	2	7018020	<i>RSAD2</i>	-0.13	-0.1	-0.16	-0.13	3.14x10 ⁻¹⁴	IRG	
cg10549986	2	7018153	<i>RSAD2</i>	-0.08	-0.09	-0.1	-0.09	1.95x10 ⁻⁹¹	IRG	
cg14126601	2	37384708	<i>EIF2AK2</i>	-0.08	-0.1	-0.12	-0.1	5.55x10 ⁻¹⁶	IRG	S_Shore
cg26337070	2	85999873	<i>ATOH8</i>	-0.06	-0.12	-0.11	-0.1	7.55x10 ⁻⁹		
cg04781494	2	202047246	<i>CASP10</i>	-0.07	-0.13	-0.08	-0.09	8.39x10 ⁻⁸	IRG	
cg15768138	2	219030752	<i>CXCR1</i>	-0.09	-0.12	-0.11	-0.11	7.38x10 ⁻²⁷		
cg13411554	3	53700276	<i>CACNA1D</i>	-0.06	-0.12	-0.09	-0.09	8.66x10 ⁻⁸		
cg22930808	3	122281881	<i>PARP9-DTX3L</i>	-0.36	-0.34	-0.4	-0.37	6.74x10 ⁻¹²⁶	IRG	N_Shore
cg08122652	3	122281939	<i>PARP9-DTX3L</i>	-0.34	-0.31	-0.51	-0.38	1.11x10 ⁻⁹	IRG	N_Shore
cg00959259	3	122281975	<i>PARP9-DTX3L</i>	-0.37	-0.3	-0.34	-0.34	1.32x10 ⁻⁵⁶	IRG	N_Shore
cg06981309	3	146260954	<i>PLSCR1</i>	-0.24	-0.28	-0.21	-0.24	6.41x10 ⁻³¹	IRG	N_Shore
cg02556393	3	168866705	<i>MECOM</i>	-0.08	-0.09	-0.1	-0.09	3.14x10 ⁻⁹⁵		N_Shore
cg07809027	4	15007205	<i>CPEB2</i>	-0.07	-0.1	-0.12	-0.1	2.08x10 ⁻¹⁴		S_Shore
cg02215171	4	89379156	<i>HERC5</i>	-0.08	-0.09	-0.11	-0.09	4.48x10 ⁻¹⁸	IRG	S_Shore
cg17786255	4	108814389	<i>SGMS2</i>	-0.07	-0.09	-0.11	-0.09	2.01x10 ⁻¹⁶	IRG	
cg21873524	4	190942744		-0.1	-0.1	-0.12	-0.11	1.03x10 ⁻⁵⁵		Island
cg24740632	5	134486678		-0.11	-0.12	-0.14	-0.12	2.26x10 ⁻⁶⁰		
cg06012695	6	28770593			-0.1	-0.13	-0.11	3.59x10 ⁻¹⁶		
cg25138053	6	31368016		-0.11	-0.09	-0.07	-0.09	3.67x10 ⁻¹⁵		S_Shore
cg22708150	6	31649619	<i>LY6G5C</i>	-0.12	-0.14	-0.17	-0.14	1.05x10 ⁻¹⁹		N_Shore

cg07292773	6	156718177		0.07	0.1	0.11	0.1	2.22x10 ⁻¹⁷		Island
cg12013713	7	139760671	<i>PARP12</i>	-0.12	-0.14	-0.09	-0.12	1.44x10 ⁻¹⁶	IRG	N_Shore
cg20190772	8	48572496	<i>KIAA0146</i>	-0.08	-0.07	-0.13	-0.09	1.40x10 ⁻⁸		
cg14864167	8	66751182	<i>PDE7A</i>	-0.25	-0.35	-0.45	-0.35	1.21x10 ⁻⁹		N_Shelf
cg06102678	8	81491328		-0.08	-0.12	-0.07	-0.09	1.00x10 ⁻⁸		Island
cg12110437	8	144098888	<i>LY6E</i>	-0.16	-0.17	-0.27	-0.2	3.14x10 ⁻⁹	IRG	N_Shore
cg17555806	10	74448117		-0.08	-0.12	-0.07	-0.09	1.51x10 ⁻⁸		N_Shelf
cg02314339	10	91020653		-0.08	-0.14	-0.11	-0.11	1.72x10 ⁻⁰⁸		
cg06188083	10	91093005	<i>IFIT3</i>	-0.29	-0.16	-0.31	-0.25	6.18x10 ⁻⁸	IRG	
cg05552874	10	91153143	<i>IFIT1</i>	-0.2	-0.28	-0.3	-0.26	6.01x10 ⁻¹⁶	IRG	
cg14910175	10	131840954		-0.07	-0.11	-0.08	-0.09	1.56x10 ⁻¹¹		N_Shelf
cg10552523	11	313478	<i>IFITM1</i>	-0.14	-0.12	-0.14	-0.13	5.90x10 ⁻¹¹⁵	IRG	N_Shelf
cg20566897	11	313527	<i>IFITM1</i>	-0.11	-0.11	-0.09	-0.1	7.00x10 ⁻⁶²	IRG	N_Shelf
cg23570810	11	315102	<i>IFITM1</i>	-0.24	-0.25	-0.34	-0.27	1.43x10 ⁻¹⁸	IRG	N_Shore
cg03038262	11	315262	<i>IFITM1</i>	-0.24	-0.22	-0.29	-0.25	4.41x10 ⁻⁴⁰	IRG	N_Shore
cg20045320	11	319555		-0.19	-0.13	-0.2	-0.18	4.85x10 ⁻¹⁷		S_Shore
cg17990365	11	319718	<i>IFITM3</i>	-0.16	-0.15	-0.15	-0.16	8.78x10 ⁻²⁹⁵	IRG	S_Shore
cg08926253	11	614761	<i>IRF7</i>	-0.15	-0.14	-0.23	-0.17	2.01x10 ⁻⁹	IRG	Island
cg12461141	11	5710654	<i>TRIM22</i>	-0.1	-0.08	-0.12	-0.1	6.35x10 ⁻²⁵	IRG	
cg23571857	17	6658898	<i>XAF1</i>	-0.07	-0.13	-0.11	-0.1	1.46x10 ⁻⁸	IRG	
cg04927537	17	76976091	<i>LGALS3BP</i>	-0.14	-0.11	-0.2	-0.15	2.77x10 ⁻¹⁰	IRG	
cg25178683	17	76976267	<i>LGALS3BP</i>	-0.15	-0.11	-0.21	-0.16	2.01x10 ⁻⁸	IRG	
cg16503797	18	19476805		-0.08	-0.12	-0.08	-0.09	5.39x10 ⁻¹²		N_Shore
cg15871086	18	56526595		-0.07	-0.11	-0.08	-0.09	2.08x10 ⁻¹¹		N_Shelf
cg23352030	20	62198469	<i>PRIC285</i>	0.13	0.19	0.11	0.14	2.36x10 ⁻¹¹		Island
cg16785077	21	42791867	<i>MX1</i>	-0.11	-0.09	-0.12	-0.11	8.45x10 ⁻²⁷	IRG	N_Shore
cg22862003	21	42797588	<i>MX1</i>	-0.31	-0.25	-0.35	-0.31	1.62x10 ⁻²⁵	IRG	N_Shore
cg26312951	21	42797847	<i>MX1</i>	-0.26	-0.17	-0.2	-0.21	6.28x10 ⁻¹⁵	IRG	N_Shore
cg21549285	21	42799141	<i>MX1</i>	-0.5	-0.35	-0.57	-0.47	6.59x10 ⁻¹³	IRG	S_Shore

cg05543864	22	24979755	<i>GGT1</i>	-0.08	-0.08	-0.1	-0.09	1.44x10 ⁻⁴⁵		
cg20098015	22	50971140	<i>ODF3B</i>	-0.19	-0.22	-0.21	-0.21	9.88x10 ⁻⁸³	IRG	S_Shore
cg05523603	22	50973101		-0.17	-0.23	-0.27	-0.22	5.51x10 ⁻¹⁴		S_Shelf
cg02247863	22	50983415		-0.07	-0.1	-0.11	-0.09	2.51x10 ⁻¹³		N_Shore

*CpGs meeting the $P_{FDR} < 0.05$ threshold (equivalent to $P < 1.06 \times 10^{-7}$) and having $|\Delta\beta| > 0.085$. [†]Positions are from Build 37. 244

[‡]IRG as defined by Interferome.[37] ^{††} Island: CpG sites >200 bp, with GC content > 55% and observed to expected ratio >0.6. N_shore: 0-2 kb 245

upstream from island; S-shore 0-2 kb downstream from island; N_shelf 2-4 kb upstream from island; S_shelf 2-4 kb downstream from island. 246

247

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

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

3.3 Hypomethylated genes are overexpressed in independent cohorts

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

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

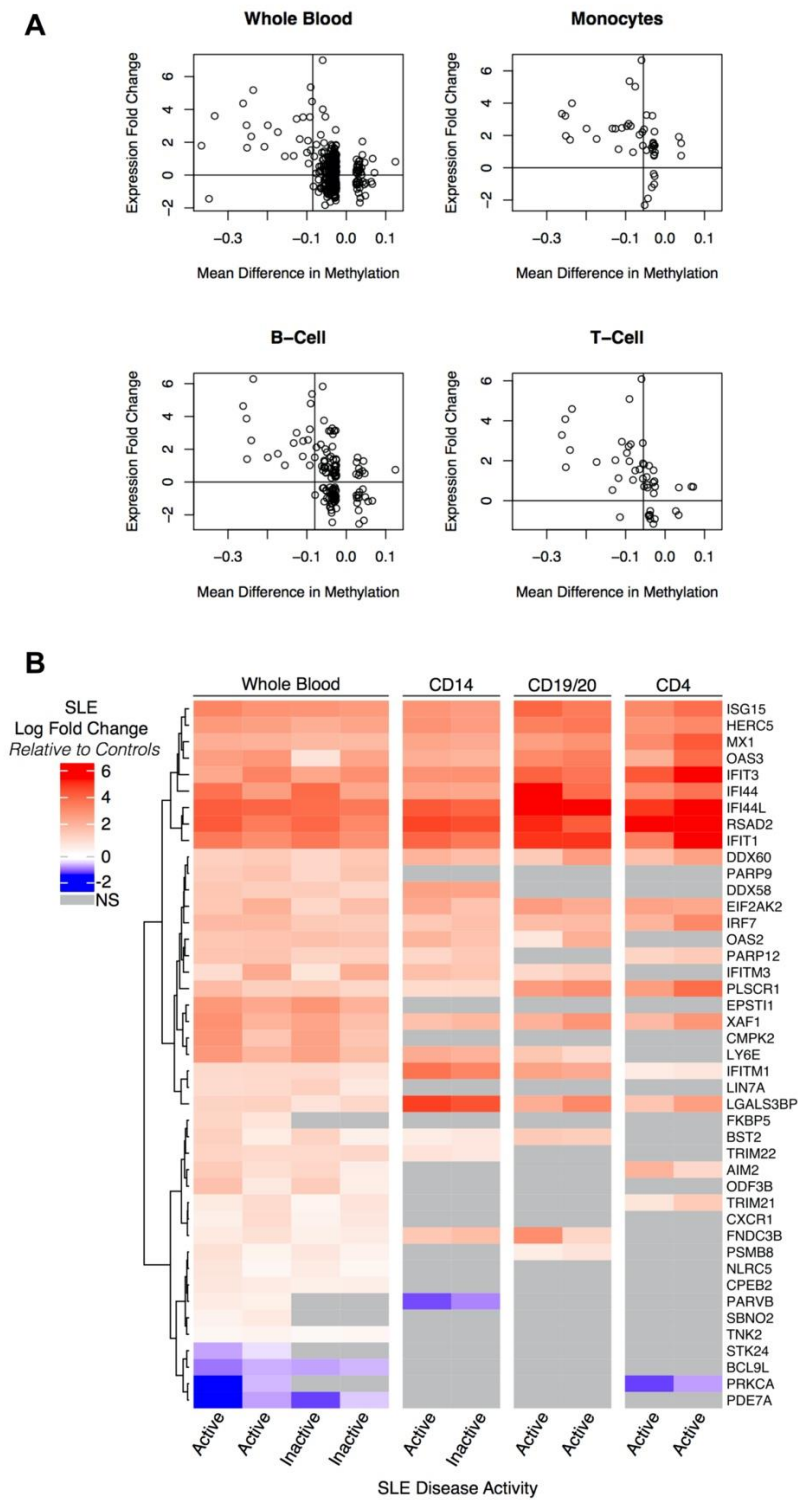


Figure 1. Hypomethylated genes showing differential expression in independent SLE cohorts. (A) Specific thresholds for the difference in the β -value (from the discordant twin methylation experiment in whole blood) that maximize the area under the ROC curve predicting increased differential expression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

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

303

CpG*	Chr	Pos (bp) [†]	Gene	Mean $\Delta\beta$	Methylation P-value	Interferon- Regulated [‡]	Active SLE [§]		Inactive SLE [§]	
							Log FC Expt 1	Log FC Expt 2	Log FC Expt 1	Log FC Expt 2
cg16526047	1	949893	<i>ISG15</i>	-0.11	1.28x10 ⁻⁴	IRG	3.1	2.77	2.74	2.59
cg05696877	1	79088769	<i>IFI44L</i>	-0.3	6.60x10 ⁻⁶	IRG	3.98	3.8	3.64	3.4
cg01079652	1	79118191	<i>IFI44</i>	-0.34	5.34x10 ⁻⁴	IRG	3.54	2.53	3.7	2.33
cg17515347	1	159047163	<i>AIM2</i>	-0.09	3.01x10 ⁻¹²	IRG	1.39	0.86	1.08	0.49
cg01028142	2	7004578	<i>CMPK2</i>	-0.33	7.98x10 ⁻⁸	IRG	2.76	1.5	2.43	1.51
cg10959651	2	7018020	<i>RSAD2</i>	-0.13	3.14x10 ⁻¹⁴	IRG	4.04	3.32	3.76	3.04
cg14126601	2	37384708	<i>EIF2AK2</i>	-0.1	5.55x10 ⁻¹⁶	IRG	1.47	2.02	1.08	1.68
cg15768138	2	219030752	<i>CXCR1</i>	-0.11	7.38x10 ⁻²⁷		0.43	0.96	0.38	0.66
cg08122652	3	122281939	<i>PARP9-DTX3L</i>	-0.38	1.11x10 ⁻⁹	IRG	1.36	1.56	1.07	1.55
cg06981309	3	146260954	<i>PLSCR1</i>	-0.24	6.41x10 ⁻³¹	IRG	1.77	1.25	1.38	1.07
cg02694620	3	172109284	<i>FNDC3B</i>	-0.11	3.80x10 ⁻³		0.57	0.82	0.41	0.52
cg15065340	3	195632915	<i>TNK2</i>	-0.16	4.04x10 ⁻³		0.22	0.31	0.2	0.25
cg07809027	4	15007205	<i>CPEB2</i>	-0.1	2.08x10 ⁻¹⁴		0.66	0.52	0.42	0.45
cg02215171	4	89379156	<i>HERC5</i>	-0.09	4.48x10 ⁻¹⁸	IRG	2.62	2.48	2.14	2.36
cg05883128	4	169239131	<i>DDX60</i>	-0.25	2.13x10 ⁻⁵	IRG	1.24	1.38	1.06	1.46
cg08099136	6	32811251	<i>PSMB8</i>	-0.11	1.43x10 ⁻⁴	IRG	-0.39	-0.13	NS	NS
cg00052684	6	35694245	<i>FKBP5</i>	-0.16	1.65x10 ⁻³		1.11	0.71	NS	NS
cg05994974	7	139761087	<i>PARP12</i>	-0.15	6.89x10 ⁻⁵	IRG	1.52	1.57	1.14	1.25
cg14864167	8	66751182	<i>PDE7A</i>	-0.35	1.21x10 ⁻⁹		-1.24	-0.41	-0.82	-0.23
cg12110437	8	144098888	<i>LY6E</i>	-0.2	3.14x10 ⁻⁹	IRG	2.66	1.92	2.43	1.7
cg03848588	9	32525008	<i>DDX58</i>	-0.1	4.34x10 ⁻⁴	IRG	1.48	1.3	1.32	1.07
cg06188083	10	91093005	<i>IFIT3</i>	-0.25	6.18x10 ⁻⁸	IRG	2.25	3.15	2.3	2.87
cg05552874	10	91153143	<i>IFIT1</i>	-0.26	6.01x10 ⁻¹⁶	IRG	3.39	2.94	3.42	2.81

cg23570810	11	315102	<i>IFITM1</i>	-0.27	1.43×10^{-18}	IRG	1	1.03	1.03	0.81
cg17990365	11	319718	<i>IFITM3</i>	-0.16	8.78×10^{-295}	IRG	0.92	2.23	0.71	2.13
cg08926253	11	614761	<i>IRF7</i>	-0.17	2.01×10^{-9}	IRG	1.84	1.79	1.4	1.37
cg08577913	11	4415193	<i>TRIM21</i>	-0.1	1.74×10^{-3}	IRG	0.56	0.93	0.28	0.75
cg12461141	11	5710654	<i>TRIM22</i>	-0.1	6.35×10^{-25}	IRG	1.14	1	0.99	1.05
cg26811705	11	118781408	<i>BCL9L</i>	-0.09	1.64×10^{-3}		-0.6	-0.35	-0.41	-0.32
cg19347790	12	81332050	<i>LIN7A</i>	-0.09	1.87×10^{-4}		0.93	0.99	1.24	0.61
cg25800166	12	113375896	<i>OAS3</i>	-0.13	5.36×10^{-5}	IRG	2.52	2.69	0.73	2.35
cg19371652	12	113415883	<i>OAS2</i>	-0.11	2.24×10^{-5}	IRG	1.48	1.56	1.64	1.53
cg03753191	13	43566902	<i>EPSTI1</i>	-0.1	9.23×10^{-5}	IRG	2.65	2.26	2.71	2.02
cg00246969	13	99159656	<i>STK24</i>	-0.11	6.26×10^{-6}		0.81	0.32	0.66	0.36
cg07839457	16	57023022	<i>NLRC5</i>	-0.23	6.10×10^{-6}	IRG	0.7	0.23	0.53	0.27
cg23571857	17	6658898	<i>XAF1</i>	-0.1	1.46×10^{-8}	IRG	2.85	1.96	2.35	1.68
cg23378941	17	64361956	<i>PRKCA</i>	-0.11	6.89×10^{-5}	IRG	-1.11	-0.3	NS	NS
cg25178683	17	76976267	<i>LGALS3BP</i>	-0.16	2.0×10^{-8}	IRG	1.16	1.21	0.72	1.05
cg07573872	19	1126342	<i>SBNO2</i>	-0.15	2.77×10^{-3}	IRG	0.38	0.58	NS	NS
cg07839313	19	17514600	<i>BST2</i>	-0.12	3.48×10^{-3}	IRG	1.24	0.49	1.17	0.41
cg21549285	21	42799141	<i>MX1</i>	-0.47	6.59×10^{-13}	IRG	2.12	2	1.86	1.79
cg19460508	22	44422195	<i>PARVB</i>	-0.1	1.64×10^{-3}		0.54	0.39	NS	NS
cg20098015	22	50971140	<i>ODF3B</i>	-0.21	9.88×10^{-83}	IRG	1.61	0.61	1.36	0.47

Differential gene expression values come from GSE39088 (Expt 1) and GSE49454 (Expt 2) in whole blood of lupus patients compared with controls. *CpGs with $P < 0.01$ and $|\Delta\beta| > 0.085$. †Positions are from Build 37. ‡As defined by Interferome[37]

§ Active disease is defined as ≥ 6 on the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI).[35]

304

305

306

307

Hierarchical clustering (Euclidean distance, complete linkage) of the DM-DE genes using the log fold change (LFC) identified a cluster of 9 genes with markedly higher LFC (Figure 1B). This cluster shows a consistent pattern across whole blood, monocytes, B cells and T cells, as well as in both active and inactive SLE disease. In fact, the LFC remained largely consistent between active and inactive disease across all DM-DE genes. Exceptions to this pattern include FK506 binding protein 5 (*FKBP5*), parvin beta (*PARVB*), and strawberry notch homolog 2 (*SBNO2*) in whole blood, where there is upregulation in active patients and non-significant change in inactive patients. This pattern was not replicated in any of the individual cell types.

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

Significant DNA methylation sites were further investigated for evidence of association between DNA methylation at a specific CpG site and gene expression (eQTM) using the iMETHYL genome browser with data on 100 healthy Japanese subjects with RNA-seq data and DNA methylation data in CD4T cells, monocytes and PBMC [38](Supplemental Table S4). Most of the CpGs from Table 1 that are identified in iMETHYL are eQTMs for the gene in which they reside. In contrast, some are eQTMs for additional genes of interest. For example, cg17515347 is in physical proximity to *AIM1*, which has an important role in T cell regulation in autoimmune diseases. However, this CpG site is also an eQTM for five other genes in CD4+ T cells (*TAGLN2*, *SLMF8*, *DUSP23*, *PHYIN1* *FCRL6*), several of which have established autoimmune disease connections. Transgelin-2 may help regulate activation and migration of B cells in lymph node follicles, exhibits increased expression in B cells from lymph nodes in SLE patients, and appears important in host defense.[53,54] SLAM family member 8 (*SLAMF8*) is a member of the SLAM family of genes of which several members have been associated with multiple autoimmune diseases.[55] FcR-like 6 (*FCRL6*), a receptor that binds to major histocompatibility complex (MHC) class II HLA-DR, is expressed in B cells and has a tyrosine-based immunoregulatory function.[56,57] Dual-specificity protein phosphatase 23 (*DUSP23*) expression is reportedly higher in CD4+ T cells from SLE patients compared to healthy controls.[58] Thus, DNA methylation in these regions, and potentially others, may have a complex and multifaceted impact on autoimmunity. Annotation of cg20098015 on chromosome 22 is linked to Outer Dense Fiber of Sperm Tails 3 (*ODF3B*). However, this CpG is an eQTM for *SCO2* homolog, mitochondrial and *SCO* cytochrome oxidase deficient homolog 2 (*SCO2*) and Thymidine Phosphorylase (*TYMP*), both involved in mitochondrial functions.

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

349

CpG*	Chr	Pos (bp) [†]	Gene	Methylation		Interferon-Regulated [‡]	Log FC Glomerulus	Log FC Tubulointerstitium
				Mean $\Delta\beta$	P-value			
cg16526047	1	949893	<i>ISG15</i> ^l	-0.11	1.28x10 ⁻⁴	IRG	3.32	4.7
cg05696877	1	79088769	<i>IFI44L</i> ^Δ	-0.3	6.60x10 ⁻⁶	IRG	5.14	5.94
cg01079652	1	79118191	<i>IFI44</i> ^l	-0.34	5.34x10 ⁻⁴	IRG	3.94	4.76
cg17515347	1	159047163	<i>AIM2</i>	-0.09	3.01x10 ⁻¹²	IRG	0.58	NS
cg01028142	2	7004578	<i>CMPK2</i> ^Δ	-0.33	7.98x10 ⁻⁸	IRG	NS	NS
cg10959651	2	7018020	<i>RSAD2</i>	-0.13	3.14x10 ⁻¹⁴	IRG	4.31	3.36
cg14126601	2	37384708	<i>EIF2AK2</i> ^Δ	-0.1	5.55x10 ⁻¹⁶	IRG	1.54	1.72
cg15768138	2	219030752	<i>CXCR1</i>	-0.11	7.38x10 ⁻²⁷		0.68	-0.17
cg08122652	3	122281939	<i>PARP9-DTX3L</i> ^Δ	-0.38	1.11x10 ⁻⁹	IRG	NS	NS
cg06981309	3	146260954	<i>PLSCR1</i> ^Δ	-0.24	6.41x10 ⁻³¹	IRG	1.92	2.07
cg02694620	3	172109284	<i>FNDC3B</i>	-0.11	3.80x10 ⁻³		NS	0.47
cg15065340	3	195632915	<i>TNK2</i> ^l	-0.16	4.04x10 ⁻³		0.38	-0.4
cg07809027	4	15007205	<i>CPEB2</i>	-0.1	2.08x10 ⁻¹⁴		NS	NS
cg02215171	4	89379156	<i>HERC5</i> [¶]	-0.09	4.48x10 ⁻¹⁸	IRG	3.16	1.96
cg05883128	4	169239131	<i>DDX60</i>	-0.25	2.13x10 ⁻⁵	IRG	1.11	2.31
cg08099136	6	32811251	<i>PSMB8</i>	-0.11	1.43x10 ⁻⁴	IRG	0.76	2.51
cg00052684	6	35694245	<i>FKBP5</i> [§]	-0.16	1.65x10 ⁻³		-1.27	-2.77
cg05994974	7	139761087	<i>PARP12</i> ^l	-0.15	6.89x10 ⁻⁵	IRG	2.26	1.86
cg14864167	8	66751182	<i>PDE7A</i>	-0.35	1.21x10 ⁻⁹		NS	NS
cg12110437	8	144098888	<i>LY6E</i> [◇]	-0.2	3.14x10 ⁻⁹	IRG	1.28	1.23
cg03848588	9	32525008	<i>DDX58</i> [◇]	-0.1	4.34x10 ⁻⁴	IRG	2.89	2.59
cg06188083	10	91093005	<i>IFIT3</i> [◇]	-0.25	6.18x10 ⁻⁸	IRG	2.59	3.14
cg05552874	10	91153143	<i>IFIT1</i> [◇]	-0.26	6.01x10 ⁻¹⁶	IRG	2.24	2.77
cg23570810	11	315102	<i>IFITM1</i>	-0.27	1.43x10 ⁻¹⁸	IRG	2.24	3.29

cg17990365	11	319718	<i>IFITM3</i>	-0.16	8.78×10^{-295}	IRG	2.24	2
cg08926253	11	614761	<i>IRF7</i> [†]	-0.17	2.01×10^{-9}	IRG	2.8	1
cg08577913	11	4415193	<i>TRIM21</i>	-0.1	1.74×10^{-3}	IRG	1.35	0.77
cg12461141	11	5710654	<i>TRIM22</i>	-0.1	6.35×10^{-25}	IRG	1.73	2.86
cg26811705	11	118781408	<i>BCL9L</i>	-0.09	1.64×10^{-3}		NS	NS
cg19347790	12	81332050	<i>LIN7A</i>	-0.09	1.87×10^{-4}		NS	-0.57
cg25800166	12	113375896	<i>OAS3</i>	-0.13	5.36×10^{-5}	IRG	3.77	1.1
cg19371652	12	113415883	<i>OAS2</i>	-0.11	2.24×10^{-5}	IRG	4.86	1.74
cg03753191	13	43566902	<i>EPSTI1</i> [¶]	-0.1	9.23×10^{-5}	IRG	NS	NS
cg00246969	13	99159656	<i>STK24</i>	-0.11	6.26×10^{-6}		NS	0.28
cg07839457	16	57023022	<i>NLRC5</i>	-0.23	6.10×10^{-6}	IRG	NS	NS
cg23571857	17	6658898	<i>XAF1</i>	-0.1	1.46×10^{-8}	IRG	3.14	3.05
cg23378941	17	64361956	<i>PRKCA</i>	-0.11	6.89×10^{-5}	IRG	-0.48	-0.08
cg25178683	17	76976267	<i>LGALS3BP</i>	-0.16	2.0×10^{-8}	IRG	0.57	1.49
cg07573872	19	1126342	<i>SBNO2</i>	-0.15	2.77×10^{-3}	IRG	NS	NS
cg07839313	19	17514600	<i>BST2</i> [‡]	-0.12	3.48×10^{-3}	IRG	NS	2.91
cg21549285	21	42799141	<i>MX1</i> ^Δ	-0.47	6.59×10^{-13}	IRG	4.05	4.64
cg19460508	22	44422195	<i>PARVB</i>	-0.1	1.64×10^{-3}		0.28	NS
cg20098015	22	50971140	<i>ODF3B</i>	-0.21	9.88×10^{-83}	IRG	NS	NS

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

3.4 Pathway analysis of DM-DE genes

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

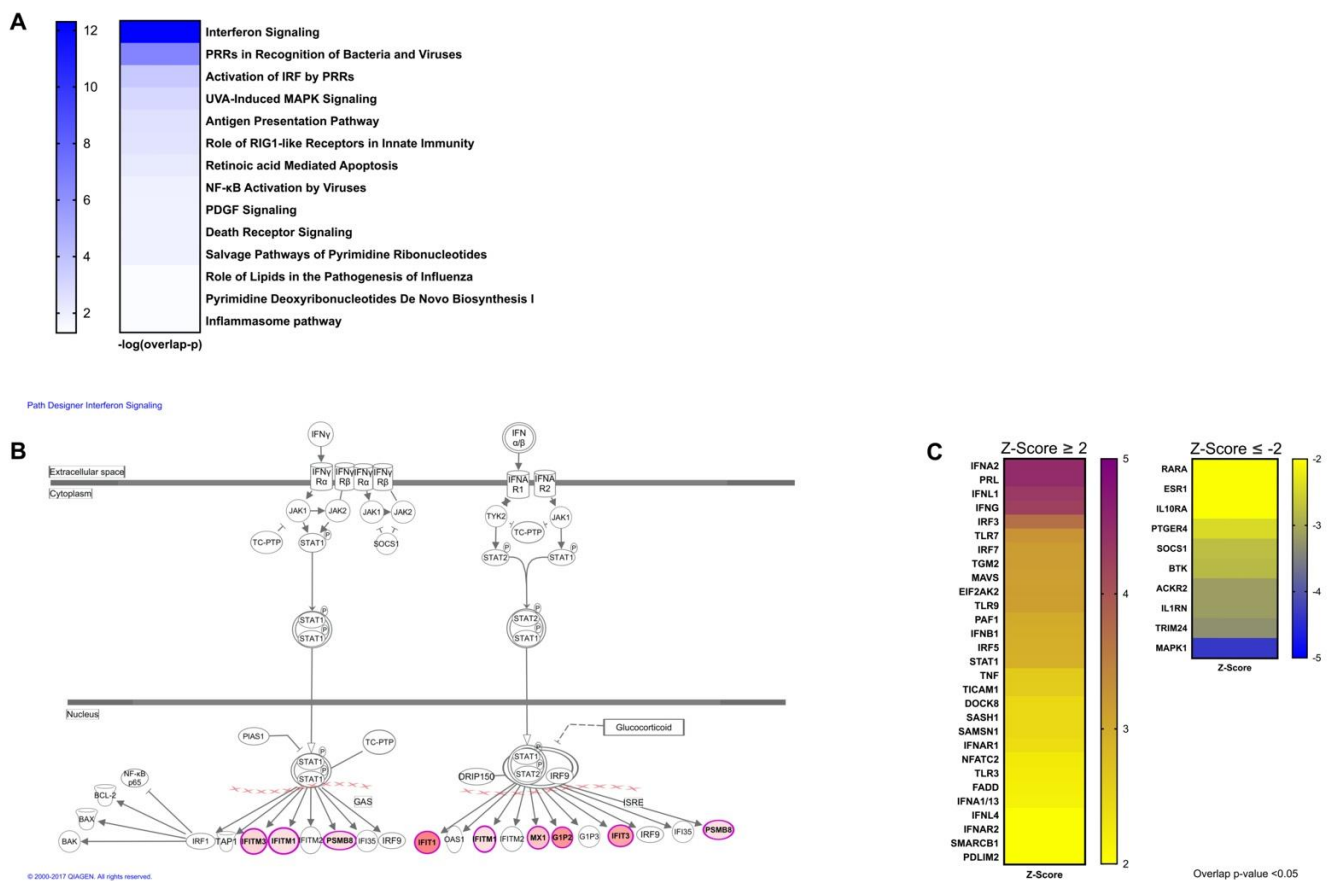


Figure 2. Pathway analyses of hypomethylated genes showing differential expression in independent SLE cohorts. (A) List and statistical significance of the overlap of the IPA canonical pathways comprised of hypomethylated genes showing differential expression in whole blood of independent SLE patients.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

The DM-DE genes were further analyzed in an additional pathway analysis using the MCODE clustering algorithm and STRING networking scores. Two distinct yet related clusters emerged (Figure 3). As expected, there was an enrichment of genes in the

IFN-inducible/pattern-recognition receptor pathway. As visually represented by the colors of the nodes and node outlines in Figure 3, all genes in this cluster were upregulated in both active and inactive SLE patients; all of these except *PARP9* were overexpressed in both kidney tissues.

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

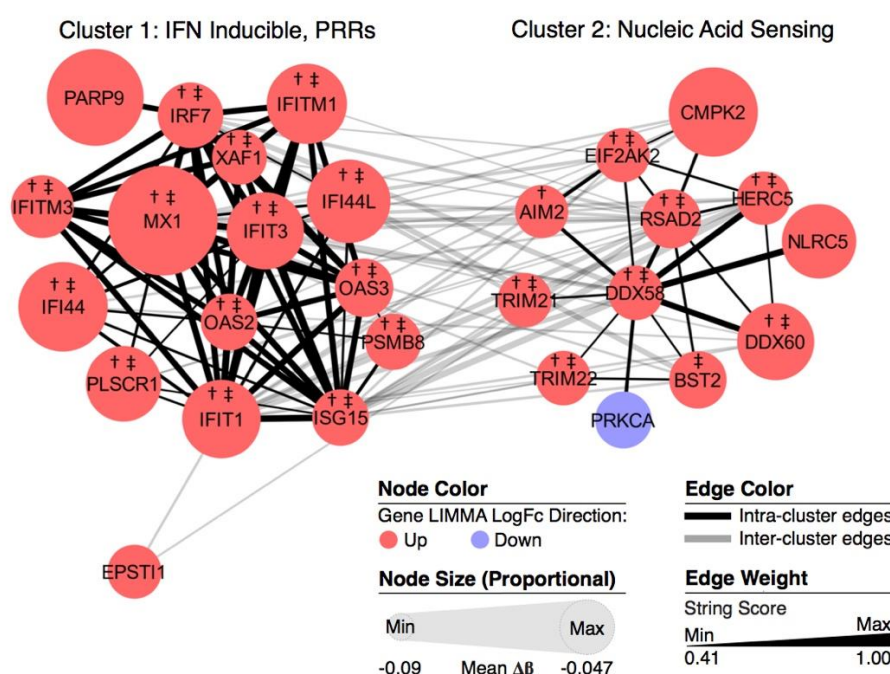


Figure 3. MCODE clustering of hypomethylated genes showing differential expression in independent SLE cohorts. A network scoring degree cutoff of 2, node score cutoff of 0.2, k-Core of 2, and a max depth of 100 were applied. Node color indicates log₂(FC) direction and node size is inversely scaled with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

3.5 Potential drug targets

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

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

CpG*	Chr	Pos(bp) [†]	Gene	Mean $\Delta\beta$	P-value	STITCH[43]	IPA [‡]	DGIdb[44]
cg1652604 7	1	949893	<i>ISG15</i>	-0.11	1.28x10 ⁻⁴			Irinotecan ^F
cg1095965 1	2	7018020	<i>RSAD2</i>	-0.13	3.14x10 ⁻¹⁴	Fludarabine ^F		
cg1412660 1	2	3738470 8	<i>EIF2AK2</i>	-0.1	5.55x10 ⁻¹⁶			Indirubin derivative E804
cg1576813 8	2	2190307 52	<i>CXCR1</i>	-0.11	7.38x10 ⁻²⁷	Reparixin ^D	Reparixin ^D	SCH-527123, Keto- profen ^F
cg0698130 9	3	1462609 54	<i>PLSCR1</i>	-0.24	6.41x10 ⁻³¹	Wogonin ^G		
cg1506534 0	3	1956329 15	<i>TNK2</i>	-0.16	4.04x10 ⁻³	Dasatinib ^{-1 F}	Osimertinib ^F , Vemurafenib ^F	Debromohymenialdisine
cg0809913 6	6	3281125 1	<i>PSMB8</i>	-0.11	1.43x10 ⁻⁴	Carfilzomib ^{4 F} , Opro- zomib ^D , Bortezomib ^{6 F}	Carfilzomib ^{4F}	Carfilzomib ^{4 F} ,
cg0005268 4	6	3569424 5	<i>FKBP5</i>	-0.16	1.65x10 ⁻³	Rapamycin/Sirolimus ^{2 F} , Tacrolimus ^{5 F}		Venlafaxine ^F , Clomipra- mine ^F
cg1486416 7	8	6675118 2	<i>PDE7A</i>	-0.35	1.21x10 ⁻⁹			Ketotifen ^F , Dyphylline ^F
cg1211043 7	8	1440988 88	<i>LY6E</i>	-0.2	3.14x10 ⁻⁹		DLYE5953A ^D	
cg0618808 3	10	9109300 5	<i>IFIT3</i>	-0.25	6.18x10 ⁻⁸	Imidazoles ^D		
cg0892625 3	11	614761	<i>IRF7</i>	-0.17	2.01x10 ⁻⁹	Hesperidin ^D		

414

415

416

cg0375319 1	13	4356690 2	<i>EPSTI1</i>	-0.1	9.23x10 ⁻⁵	Methotrexate ^{F,T} , Vin- blastine ^F , Doxorubicin ^F , Cisplatin ^F		
cg0024696 9	13	9915965 6	<i>STK24</i>	-0.11	6.26x10 ⁻⁶	Staurosporine ^D		
cg2337894 1	17	6436195 6	<i>PRKCA</i>	-0.11	6.89x10 ⁻⁵	Staurosporine ^D	Aprinocarsen	Midostaurin ^F , Enzastau- rin ^D , Quercetin ^{D,G} , Aprinocarsen, Rubox- istaurin ^D , Ingenol Mebu- tate ^{FW} , Bryostatin ^D , Sotrastaurin Acetate ^D , Tamoxifen ^{2,F}
cg0783931 3	19	1751460 0	<i>BST2</i>	-0.12	3.48x10 ⁻³	Resveratrol ^{6,D,G}		
cg2154928 5	21	4279914 1	<i>MX1</i>	-0.47	6.59x10 ⁻¹³	Mitomycin C ^F , Colchi- cine ^F		
cg1946050 8	22	4442219 5	<i>PARVB</i>	-0.1	1.64x10 ⁻³	Lovastatin ^{3,F}		
								Bortezomib ^{6,F}

*CpGs with $P < 1 \times 10^{-3}$ and $\Delta\beta < -0.085$. [†]Positions are from Build 37. [‡]Qiagen Bioinformatics: ingenuity.com ^F: FDA approved.

^D: Ongoing clinical trial or DiD ^G: GRAS. ^T: Known utility in lupus therapy. ^{FW}: Ingenol mebutate is FDA-approved in the US but withdrawn in the EU. CoLTS scores in superscript (range: -16 to +11)

417
418
419
420

4. Discussion

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

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

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

Three recent studies have observed aberrant methylation of IFN genes in SLE patients with renal involvement [12,19,22]. A summary of the literature (Additional File 2) shows our study’s consistencies with these published findings. While hypomethylation of these genes has been confirmed in CD4+ T cells and peripheral blood, no SLE study to

date has examined genome-wide DNA methylation in kidney biopsies. By considering differential gene expression derived from the micro-dissected glomerulus and tubulointerstitium kidneys in an independent cohort of SLE patients, in conjunction with the significance of aberrant methylation in the MZ twin data, this study corroborates many of the loci previously published as being hypomethylated in lupus nephritis patients.

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

Within this study, the genomic locations of hypomethylated CpG sites were highly skewed toward CpG shores (0-2 kb from island) and shelves (2-4 kb from island) instead of islands. Here, only 5 of 59 CpG sites were in a CpG island despite nearly 1/3 of the CpG sites on the Illumina HumanMethylation450 BeadChip are in a CpG island (Supplemental Figure S2). Our findings are consistent with those of Yeung et al, who demonstrated that most CpG sites hypomethylated in their lupus patients, when compared to controls, were located in CpG shores [21]. These data corroborate the hypothesis that CpG islands tend to have lower methylation rates than less dense CpG regions (e.g., shores and shelves) and that lower density allows for greater methylation autonomy in response to the environment, leading to increases in potential functional significance of the shores and shelves.

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

5. Conclusions

The intersection of hypomethylated genes from MZ twins and upregulated genes from multiple independent cohorts and cell types were attributed to two distinct but integrated biologic pathways: the nucleic acid sensing pathway and IFN-inducing pathway. The source, type, and location of nucleic acids found in an SLE patient determine how and by which receptor the NA is recognized, and ultimately which IRF is stimulated. A

multi-omics approach could allow classification of patients into different endotypes, and possible treatment groups. Informatically linking the DM-DE genes to drug therapies identified a list of compounds that could be critically evaluated as potential candidates for future trials, either broadly for SLE or for individuals with specific hypomethylation signatures.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1: Table S1, Table S2, Table S3, Table S4, Figure S1, Figure S2.

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

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

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

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

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

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

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

References

1. International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN), Harley JB, Alarcón-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PTK, KIAA1542 and other loci. *Nat Genet.* 2008 Feb;40(2):204–10.
2. Morris DL, Sheng Y, Zhang Y, Wang Y-F, Zhu Z, Tomblinson P, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat Genet.* 2016 Aug;48(8):940–6.
3. Lessard CJ, Sajuthi S, Zhao J, Kim K, Ice JA, Li H, et al. Identification of a Systemic Lupus Erythematosus Risk Locus Spanning ATG16L2, FCHSD2, and P2RY2 in Koreans. *Arthritis Rheumatol Hoboken NJ.* 2016 May;68(5):1197–209.
4. Sun C, Molineros JE, Looger LL, Zhou X-J, Kim K, Okada Y, et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat Genet.* 2016 Mar;48(3):323–30.
5. Alarcón-Riquelme ME, Ziegler JT, Molineros J, Howard TD, Moreno-Estrada A, Sánchez-Rodríguez E, et al. Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. *Arthritis Rheumatol Hoboken NJ.* 2016 Apr;68(4):932–43.

6. Langefeld CD, Ainsworth HC, Cunninghame Graham DS, Kelly JA, Comeau ME, Marion MC, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat Commun.* 2017 17;8:16021. 578 579
7. Deng Y, Tsao BP. Updates in Lupus Genetics. *Curr Rheumatol Rep.* 2017 Oct 5;19(11):68. 580
8. Chen L, Morris DL, Vyse TJ. Genetic advances in systemic lupus erythematosus: an update. *Curr Opin Rheumatol.* 2017 Sep;29(5):423–33. 581 582
9. Deapen D, Escalante A, Weinrib L, Horwitz D, Bachman B, Roy-Burman P, et al. A revised estimate of twin concordance in systemic lupus erythematosus. *Arthritis Rheum.* 1992 Mar;35(3):311–8. 583 584
10. Hughes T, Sawalha AH. The role of epigenetic variation in the pathogenesis of systemic lupus erythematosus. *Arthritis Res Ther.* 2011;13(5):245. 585 586
11. Jeffries MA, Dozmorov M, Tang Y, Merrill JT, Wren JD, Sawalha AH. Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. *Epigenetics.* 2011 May;6(5):593–601. 587 588
12. Coit P, Renauer P, Jeffries MA, Merrill JT, McCune WJ, Maksimowicz-McKinnon K, et al. Renal involvement in lupus is characterized by unique DNA methylation changes in naïve CD4+ T cells. *J Autoimmun.* 2015 Jul;61:29–35. 589 590
13. Renauer P, Coit P, Jeffries MA, Merrill JT, McCune WJ, Maksimowicz-McKinnon K, et al. DNA methylation patterns in naïve CD4+ T cells identify epigenetic susceptibility loci for malar rash and discoid rash in systemic lupus erythematosus. *Lupus Sci Med.* 2015;2(1):e000101. 591 592 593
14. Ulf-Møller CJ, Asmar F, Liu Y, Svendsen AJ, Busato F, Grønbaek K, et al. Twin DNA Methylation Profiling Reveals Flare-Dependent Interferon Signature and B Cell Promoter Hypermethylation in Systemic Lupus Erythematosus. *Arthritis Rheumatol Hoboken NJ.* 2018 Jun;70(6):878–90. 594 595 596
15. Coit P, Jeffries M, Altorok N, Dozmorov MG, Koelsch KA, Wren JD, et al. Genome-wide DNA methylation study suggests epigenetic accessibility and transcriptional poising of interferon-regulated genes in naïve CD4+ T cells from lupus patients. *J Autoimmun.* 2013 Jun;43:78–84. 597 598 599
16. Coit P, Oggenovski M, Gensterblum E, Maksimowicz-McKinnon K, Wren JD, Sawalha AH. Ethnicity-specific epigenetic variation in naïve CD4+ T cells and the susceptibility to autoimmunity. *Epigenetics Chromatin.* 2015;8:49. 600 601
17. Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, et al. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS Genet.* 2013;9(8):e1003678. 602 603 604
18. Coit P, Yalavarthi S, Oggenovski M, Zhao W, Hasni S, Wren JD, et al. Epigenome profiling reveals significant DNA demethylation of interferon signature genes in lupus neutrophils. *J Autoimmun.* 2015 Apr;58:59–66. 605 606
19. Mok A, Solomon O, Nayak RR, Coit P, Quach HL, Nititham J, et al. Genome-wide profiling identifies associations between lupus nephritis and differential methylation of genes regulating tissue hypoxia and type 1 interferon responses. *Lupus Sci Med.* 2016;3(1):e000183. 607 608 609
20. Chung SA, Nititham J, Elboudwarej E, Quach HL, Taylor KE, Barcellos LF, et al. Genome-Wide Assessment of Differential DNA Methylation Associated with Autoantibody Production in Systemic Lupus Erythematosus. *PloS One.* 2015;10(7):e0129813. 610 611
21. Yeung KS, Chung BH-Y, Choufani S, Mok MY, Wong WL, Mak CCY, et al. Genome-Wide DNA Methylation Analysis of Chinese Patients with Systemic Lupus Erythematosus Identified Hypomethylation in Genes Related to the Type I Interferon Pathway. *PloS One.* 2017;12(1):e0169553. 612 613 614
22. Zhu H, Mi W, Luo H, Chen T, Liu S, Raman I, et al. Whole-genome transcription and DNA methylation analysis of peripheral blood mononuclear cells identified aberrant gene regulation pathways in systemic lupus erythematosus. *Arthritis Res Ther.* 2016 Jul 13;18:162. 615 616 617
23. Imgenberg-Kreuz J, Carlsson Almlöf J, Leonard D, Alexsson A, Nordmark G, Eloranta M-L, et al. DNA methylation mapping identifies gene regulatory effects in patients with systemic lupus erythematosus. *Ann Rheum Dis.* 2018 May;77(5):736–43. 618 619

24. Joseph S, George NI, Green-Knox B, Treadwell EL, Word B, Yim S, et al. Epigenome-wide association study of peripheral blood mononuclear cells in systemic lupus erythematosus: Identifying DNA methylation signatures associated with interferon-related genes based on ethnicity and SLEDAI. *J Autoimmun.* 2019 Jan;96:147–57.
25. Breitbach ME, Ramaker RC, Roberts K, Kimberly RP, Absher D. Population-Specific Patterns of Epigenetic Defects in the B Cell Lineage in Patients With Systemic Lupus Erythematosus. *Arthritis Rheumatol Hoboken NJ.* 2020 Feb;72(2):282–91.
26. Yeung KS, Lee TL, Mok MY, Mak CCY, Yang W, Chong PCY, et al. Cell Lineage-specific Genome-wide DNA Methylation Analysis of Patients with Paediatric-onset Systemic Lupus Erythematosus. *Epigenetics.* 2019 Feb 26;15592294.2019.1585176.
27. Weeding E, Sawalha AH. Deoxyribonucleic Acid Methylation in Systemic Lupus Erythematosus: Implications for Future Clinical Practice. *Front Immunol.* 2018;9:875.
28. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. *Genome Med.* 2014;6(7):60.
29. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubrea J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* 2010 Feb;20(2):170–9.
30. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet.* 2011 Sep;7(9):e1002300.
31. Gervin K, Vigeland MD, Mattingsdal M, Hammerø M, Nygård H, Olsen AO, et al. DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. *PLoS Genet.* 2012 Jan;8(1):e1002454.
32. Häslér R, Feng Z, Bäckdahl L, Spehlmann ME, Franke A, Teschendorff A, et al. A functional methylome map of ulcerative colitis. *Genome Res.* 2012 Nov;22(11):2130–7.
33. Rasmussen A, Sevier S, Kelly JA, Glenn SB, Aberle T, Cooney CM, et al. The lupus family registry and repository. *Rheumatol Oxf Engl.* 2011 Jan;50(1):47–59.
34. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* 1997 Sep;40(9):1725.
35. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. *Arthritis Rheum.* 1992 Jun;35(6):630–40.
36. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika.* 1988;75(4):800–2.
37. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, et al. Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D1040–1046.
38. Hachiya T, Furukawa R, Shiwa Y, Ohmomo H, Ono K, Katsuoka F, et al. Genome-wide identification of inter-individually variable DNA methylation sites improves the efficacy of epigenetic association studies. *NPJ Genomic Med.* 2017;2:11.
39. Jean Wu, Rafael Irizarry, James MacDonald, Jeff Gentry. Background adjustment using sequence information. *Bioconductor;* 2017.
40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015 Apr 20;43(7):e47.
41. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics.* 2003 Jan 13;4:2.
42. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* 2000 Sep 15;28(18):3442–4.
43. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015 Jan;43(Database issue):D447–452.
44. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, et al. DGIdb: mining the druggable genome. *Nat Methods.* 2013 Dec;10(12):1209–10.

45. Grammer AC, Ryals MM, Heuer SE, Robl RD, Madamanchi S, Davis LS, et al. Drug repositioning in SLE: crowd-sourcing, literature-mining and Big Data analysis. *Lupus*. 2016 Sep;25(10):1150–70.
46. Mallya M, Campbell RD, Aguado B. Characterization of the five novel Ly-6 superfamily members encoded in the MHC, and detection of cells expressing their potential ligands. *Protein Sci Publ Protein Soc*. 2006 Oct;15(10):2244–56.
47. Alaridah N, Winqvist N, Håkansson G, Tenland E, Rönnholm A, Sturegård E, et al. Impaired CXCR1-dependent oxidative defence in active tuberculosis patients. *Tuberc Edinb Scotl*. 2015 Dec;95(6):744–50.
48. Xu R, Bao C, Huang H, Lin F, Yuan Y, Wang S, et al. Low expression of CXCR1/2 on neutrophils predicts poor survival in patients with hepatitis B virus-related acute-on-chronic liver failure. *Sci Rep*. 2016 Dec 15;6:38714.
49. Almajhdi FN, Al-Ahdal M, Abdo AA, Sanai FM, Al-Anazi M, Khalaf N, et al. Single nucleotide polymorphisms in CXCR1 gene and its association with hepatitis B infected patients in Saudi Arabia. *Ann Hepatol*. 2013 Apr;12(2):220–7.
50. Swamydas M, Gao J-L, Break TJ, Johnson MD, Jaeger M, Rodriguez CA, et al. CXCR1-mediated neutrophil degranulation and fungal killing promote *Candida* clearance and host survival. *Sci Transl Med*. 2016 Jan 20;8(322):322ra10.
51. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*. 1987 Jul 20;196(2):261–82.
52. Zhao M, Zhou Y, Zhu B, Wan M, Jiang T, Tan Q, et al. IFI44L promoter methylation as a blood biomarker for systemic lupus erythematosus. *Ann Rheum Dis*. 2016 Nov;75(11):1998–2006.
53. Kiso K, Yoshifuji H, Oku T, Hikida M, Kitagori K, Hirayama Y, et al. Transgelin-2 is upregulated on activated B-cells and expressed in hyperplastic follicles in lupus erythematosus patients. *PloS One*. 2017;12(9):e0184738.
54. Kim H-R, Lee H-S, Lee K-S, Jung ID, Kwon M-S, Kim C-H, et al. An Essential Role for TAGLN2 in Phagocytosis of Lipopolysaccharide-activated Macrophages. *Sci Rep*. 2017 Aug 18;7(1):8731.
55. Dragovich MA, Mor A. The SLAM family receptors: Potential therapeutic targets for inflammatory and autoimmune diseases. *Autoimmun Rev*. 2018 Jul;17(7):674–82.
56. Li FJ, Won WJ, Becker EJ, Easlick JL, Tabengwa EM, Li R, et al. Emerging roles for the FCRL family members in lymphocyte biology and disease. *Curr Top Microbiol Immunol*. 2014;382:29–50.
57. Schreeder DM, Cannon JP, Wu J, Li R, Shakhmatov MA, Davis RS. Cutting edge: FcR-like 6 is an MHC class II receptor. *J Immunol Baltim Md 1950*. 2010 Jul 1;185(1):23–7.
58. Balada E, Felip L, Ordi-Ros J, Vilardell-Tarrés M. DUSP23 is over-expressed and linked to the expression of DNMTs in CD4+ T cells from systemic lupus erythematosus patients. *Clin Exp Immunol*. 2017 Feb;187(2):242–50.
59. Cavlar T, Ablasser A, Hornung V. Induction of type I IFNs by intracellular DNA-sensing pathways. *Immunol Cell Biol*. 2012 May;90(5):474–82.
60. Suárez-Calvet X, Gallardo E, Nogales-Gadea G, Querol L, Navas M, Díaz-Manera J, et al. Altered RIG-I/DDX58-mediated innate immunity in dermatomyositis. *J Pathol*. 2014 Jul;233(3):258–68.
61. Crow MK. Collaboration, genetic associations, and lupus erythematosus. *N Engl J Med*. 2008 Feb 28;358(9):956–61.
62. Ambrosi C, Manzo M, Baubec T. Dynamics and Context-Dependent Roles of DNA Methylation. *J Mol Biol*. 2017 May 19;429(10):1459–75.
63. Crowl JT, Gray EE, Pestal K, Volkman HE, Stetson DB. Intracellular Nucleic Acid Detection in Autoimmunity. *Annu Rev Immunol*. 2017 Apr 26;35:313–36.
64. Schübeler D. Function and information content of DNA methylation. *Nature*. 2015 Jan 15;517(7534):321–6.
65. Shrivastav M, Niewold TB. Nucleic Acid sensors and type I interferon production in systemic lupus erythematosus. *Front Immunol*. 2013 Oct 7;4:319.
66. Rathinam VAK, Jiang Z, Waggoner SN, Sharma S, Cole LE, Waggoner L, et al. The AIM2 inflammasome is essential for host defense against cytosolic bacteria and DNA viruses. *Nat Immunol*. 2010 May;11(5):395–402.
67. Man SM, Karki R, Kanneganti T-D. AIM2 inflammasome in infection, cancer, and autoimmunity: Role in DNA sensing, inflammation, and innate immunity. *Eur J Immunol*. 2016 Feb;46(2):269–80.

68. Mustelin T, Lood C, Giltiay NV. Sources of Pathogenic Nucleic Acids in Systemic Lupus Erythematosus. *Front Immunol.* 2019;10:1028. 706
707
69. Christensen K, Kristiansen M, Hagen-Larsen H, Skytthe A, Bathum L, Jeune B, et al. X-linked genetic factors regulate 708
hematopoietic stem-cell kinetics in females. *Blood.* 2000 Apr 1;95(7):2449–51. 709
70. Huang Q, Parfitt A, Grennan DM, Manolios N. X-chromosome inactivation in monozygotic twins with systemic lupus 710
erythematosus. *Autoimmunity.* 1997;26(2):85–93. 711
71. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 712
6, 20 and 22. *Nat Genet.* 2006 Dec;38(12):1378–85. 713
714