Genome-wide association meta-analysis identifies critical regulators of immune dysfunction and cell stress pathways driving cardiovascular disease and systemic lupus erythematosus

Jessica E. Kain[a,b], Katherine A. Owen[a], Mete Civelek[b], Peter E. Lipsky[a], Amrie C. Grammer[a,1]

Technical Paper
Presented to the Faculty of the Department of Biomedical Engineering

Jessica Kain
Fall 2019, Spring 2020

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____ Jessica Kain

Approved _____ Date _____ Kate Owen Ph.D,
AMPEL BioSolutions

Genome-wide association meta-analysis identifies critical regulators of immune dysfunction and cell stress pathways driving cardiovascular disease and systemic lupus erythematosus

Jessica E. Kain[a,b], Katherine A. Owen[a], Mete Civelek[b], Peter E. Lipsky[a], Amrie C. Grammer[a,1]

[a] AMPEL BioSolutions LLC, Charlottesville, VA
[b] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA
[1] Correspondence: amriegrammer@ampelbiosolutions.com

Abstract

Systemic lupus erythematosus (SLE) is an autoimmune syndrome characterized by multi-organ inflammation and immune dysregulation and is highly associated with the development of cardiovascular disease (CVD). Although studies exploring the association between SLE and premature CVD demonstrate that altered immune function plays a pivotal role in the increased cardiovascular morbidity and mortality observed in SLE patients, additional work is needed to identify critical pathways in SLE and CVD pathogenesis that can be used as novel points of therapeutic interventions. Here, published Immunochip and genome-wide association studies (GWAS) from SLE and coronary artery disease (CAD) were used to identify 96 overlapping SNPs significantly associated with both conditions. Variants were linked to 189 predicted causal genes via expression quantitative trait loci (eQTL) mapping, the identification of functional variants in coding regions and transcription factor binding sites, as well as traditional SNP-gene annotation. The predicted genes were validated using numerous datasets of differential expression in SLE tissue. 118 differentially expressed genes and their upstream regulators were used to predict biological pathways. Dysregulated pathways representative of both SLE and CAD centered around dysfunctional immune function and cell stress. Drug targets identified within the signaling pathways were matched to existing drugs and ranked using the Combined Lupus Treatment Scoring (CoLTs) system. Ultimately, 18 novel drug candidates with CoLTs scores equal to or greater than the current standard-of-care drug, belimumab, were identified, 8 of which are FDA-approved.

Keywords: System lupus erythematosus, cardiovascular disease, GWAS, genetic mapping, drug repurposing

INTRODUCTION

Systemic lupus erythematosus (SLE) is estimated to affect nearly 1.5 million people in the United States alone [1]. SLE is an autoimmune syndrome characterized by multi-organ inflammation and immune dysregulation and is highly associated with the development of cardiovascular disease (CVD). Compared to the general population, patients with SLE have a 2-10 fold increased risk of CVD. The relative risk for women with SLE between the ages of 35 and 45 is increased 50-fold [4] and the occurrence of fatal myocardial infarction has been reported to be 3 times greater in SLE patients [2]. Additionally, many SLE patients who have a myocardial infarction are relatively young, suggesting an increased risk with SLE rather than chance occurrences [2].

The therapeutic challenge presented by SLE is largely due to the extensive heterogeneity of the disease. In general, SLE is associated with hyperactivity of the innate and adaptive immune system such as T and B cell abnormalities, overproduction of autoantibodies and disturbed cytokine balance. Heterogeneity of SLE includes differential expression of these abnormalities and clinical manifestations [8]. Standard-of-care treatments for SLE include glucocorticoids, non-steroidal anti-inflammatory drugs (NSAIDs), antimalarials, and immunosuppressive drugs [7]. These drugs only treat symptoms and control the progression of the disease. Recently, belimumab has been approved for treatment of SLE as well [8]. Belimumab was not only the first new drug approved for SLE in decades, it also is the first biological agent used for treating SLE [8]. Despite the moderate effectiveness, the approval of belimumab is revolutionary as it marks a shift in treatments for SLE away from symptom relieving medicine.

Although mortality from infections and active disease have decreased in SLE patients, CVD-related death rates have not improved [5] and the standardized mortality ratio due to CVD has actually increased [6]. Treatment options remain limited as statins have little effect on cardiovascular outcomes in SLE populations despite their

effective preventative role in non-SLE patients. Recent studies exploring the association between SLE and premature CVD demonstrate that alterations of specific immune functions play a pivotal role in the increased cardiovascular morbidity and mortality observed in SLE patients [3]. Nonetheless, additional studies are needed to identify critical pathways in CVD pathogenesis in lupus that can be used as novel points of therapeutic intervention.

Genetic predispositions are important risk factors for both SLE and CVD. The lack of a correlation between severity of lupus and cardiac outcomes in SLE patients [9] supports the hypothesis that genetic components play a role in lupus patients for developing CVD. Although genome-wide association studies (GWAS) have been successful in mapping disease loci in both autoimmune and cardiovascular disease, these results have failed to impact clinical practice. Understanding the functional mechanisms of causal genetic variants underlying SLE and CVD may provide essential information to identify shared molecular pathways and therapeutic targets relevant to disease mechanisms. Here, we evaluate shared pathways underlying CVD in SLE with a focus on identifying novel therapeutic options. Using a comprehensive bioinformatics approach, existing drugs are matched to the molecular pathways associated with both SLE and CVD. By repurposing FDA-approved drugs, the process of bringing new therapeutic options to the market can be greatly expedited.

RESULTS

Identification of genetic variants linked to SLE and CAD

Genome-Wide Association Study (GWAS) and Immunochip results were used to obtain SNPs associated with each disease (Materials). For CVD, the most recent trans-ancestral meta-analysis of GWAS studies for coronary artery disease (CAD), a large subset of CVD, was used. For SLE, results of multiple GWAS and Immunochip studies were included to account for all ancestries. Using a significance threshold of p-value $< 1 \times 10^{-6}$, 7,222 and 16,163 SNPs are significantly associated with SLE and CAD, respectively. 96 of these SNPs are significantly associated with both conditions (Figure 1.A).

Before making gene predictions with these genetic variants, statistical analysis was performed to ensure that the observed overlap is significant, not simply a result of intersecting datasets of their respective sizes. The Monte Carlo Simulation Method was used to estimate the probability of observing an overlap of at least 96 between unrelated subsets of 7,222 and 16,163 SNPs (Methods). Three versions of simulations were executed to generate the null distributions of overlap size: 1) overlapping the 7,222 significant SLE SNPs with random subsets of 16,163 SNPs (Figure 1.B), 2) overlapping the 16,163 significant CAD SNPs with random subsets of over 7,222 SNPs (Figure 1.C), and 3) overlapping random subsets of 16,163 and over 7,222 SNPs (Figure 1.D). These null distributions were then used to estimate the probability that an overlap of 96 is obtained from intersecting random sets of over 7,222 and 16,163 SNPs. Out of the 10,000 iterations, 0 resulted in an overlap of 96 SNPs or more, thus estimating a p-value less than 1/10,000. This was the case for all three simulation versions. As such, using Monte Carlo Simulations, the

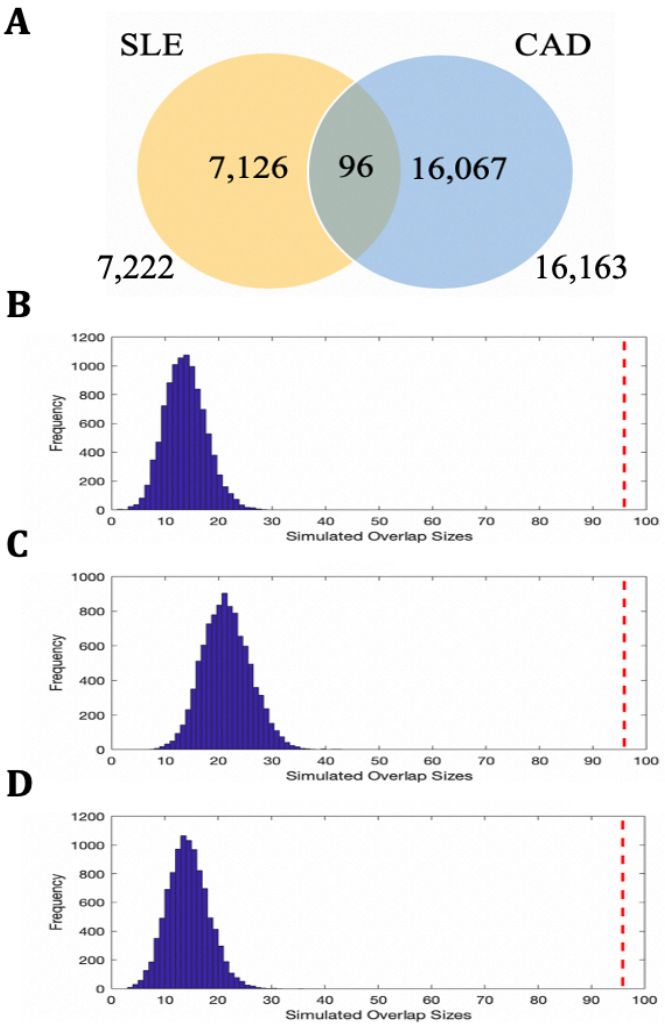


Figure 1. Significant SNPs associated with both SLE and CAD. **(A)** Venn diagram of significant SNPs (p-value $< 1*10^{-6}$) associated with SLE and CAD. (**B-D**) Histograms of the distributions of overlap size for the three Monte Carlo Simulation versions**.** Red dotted line represents the 96-SNP-overlap between SLE and CAD-associated SNPs. **(B)** Distribution of overlap sizes between random subsets of 16,163 SNPs and the 7,222 SLE-associated SNPs. **(C)** Distribution of overlap sizes between random subsets of 7,222-7,335 SNPs and the 16,163 CAD-associated SNPs. **(D)** Distribution of overlap sizes between random subsets of 7,222-7,335 SNPs and random subsets of 16,163 SNPs.

probability of obtaining a 96-SNP-overlap between over 7,222 and 16,163 random SNPs is estimated to be p-value < 0.0001. The histograms further emphasize the unlikelihood that the observed 96-SNP-overlap is a trivial product of intersecting subsets of this size (Figure 1.B-D).

Prediction and validation of genes implicated by genetic variants associated with both SLE and CAD

Multiple bioinformatic-based approaches were used to identify the most plausible gene(s) affected by the each of the genetic variants significantly associated with both SLE and CAD. First, a number of tools were used to identify and classify the genomic locations of the 96 SNPs to determine their functional categories. The 96 SNPs were run through Ensembl's Variant Effect Predictor (VEP) which provides a comprehensive list of genomic functions and consequences by including predictions made by a number of platforms. As VEP yields numerous predicted consequences for each variant, additional tools such as dbSNP and HaploReg were used to confirm variant locations and effects. Additionally, specialized databases were used to determine if SNPs are located within expression quantitative trait loci (eQTL) or regulatory regions, such as enhancers and promoters. The Genotype-Tissue Expression (GTEx) database identified SNPs located in eQTLs and Human Active Enhancers to Interpret Regulatory Variants (HACER) identified SNPs located in regulatory regions.

The genomic databases were then used to predict genes with respect to the functional locations of SNPs (Figure 2A and Methods). SNPs located in coding regions of genes, exons, were mapped to 6 coding (C-) genes using VEP, dbSNP, and HaploReg. SNPs located within eQTLs were mapped to 159 expression (E-) genes using GTEx. SNPs located in distal and cis regulatory regions were mapped to 26 downstream target (T-) genes using HACER. Lastly, 59 proximal (P-) genes located within 5 kb of the SNPs were identified and confirmed using VEP, dbSNP, and Stanford's Genomic Regions Enrichment of Annotations Tool (GREAT). In total, the 96 SNPs mapped to 189 genes, as SNPs mapped to multiple genes and some genes were predicted by various SNPs under multiple functional consequences (Figure 2B). One gene, *MUC22* was shared within all four groups, and limited commonality was observed between T-, P- and E-Genes, with only 5 genes shared among the three groups.

Next, the 189 predicted genes were validated using gene expression data to determine if they exhibited altered expression in SLE. We used a wide range of SLE differential expression datasets in various tissues, including whole blood, PBMCs, B cells, T cells, synovium, skin and kidney. Of the 189 predicted genes, 118 (62%) were determined to be differentially expressed genes (DEGs) in one or more SLE tissues (Figure S1). The 118 validated DEGs are used in further analysis, while the remaining 71 genes were filtered out.

Identification of molecular pathways involving predicted differentially expressed genes

Pathway analysis of the predicted DEGs was performed and upstream regulators (UPRs) were utilized to elucidate the key signaling networks implicated by the genes. UPRs of the 118 predicted DEGs were determined by Ingenuity Pathway Analysis (IPA). Using a significance threshold of p-value < 0.05, 164 genes were determined to be UPRs of the expression networks significantly enriched in the set of predicted DEGs and are included in the construction and analysis of protein networks for additional representation and emphasis of dominant signaling networks. Next, protein-protein interactions were determined by the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). The protein interaction networks were then visualized in Cytoscape and clustered using MCODE to provide an additional level of function annotation (Figure 3A and Methods). The resulting networks were further simplified into metaclusters 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 (Figure 3B).

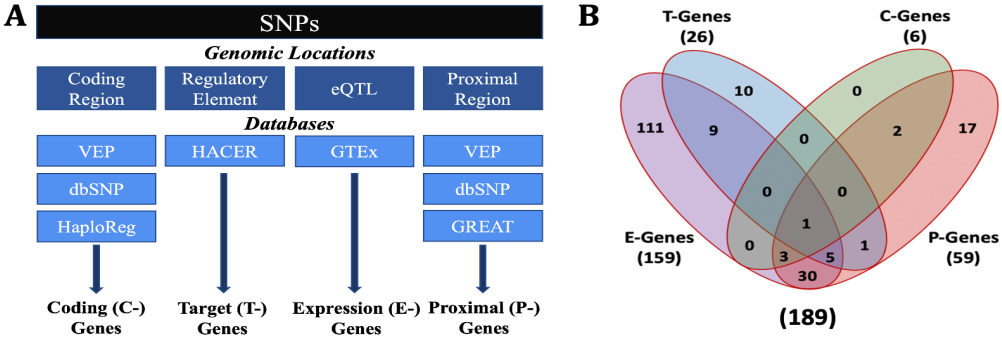


Figure 2. SNPs significantly associated with SLE and CAD predict functional genes associated with both conditions. **(A)** Model of approach of comprehensive SNP to gene predictions with respect to functional genomic mutations location of the variants. **(B)** Venn diagram depicting the overlap between the corresponding SNP-predicted E-, T-, C- and P-Genes.

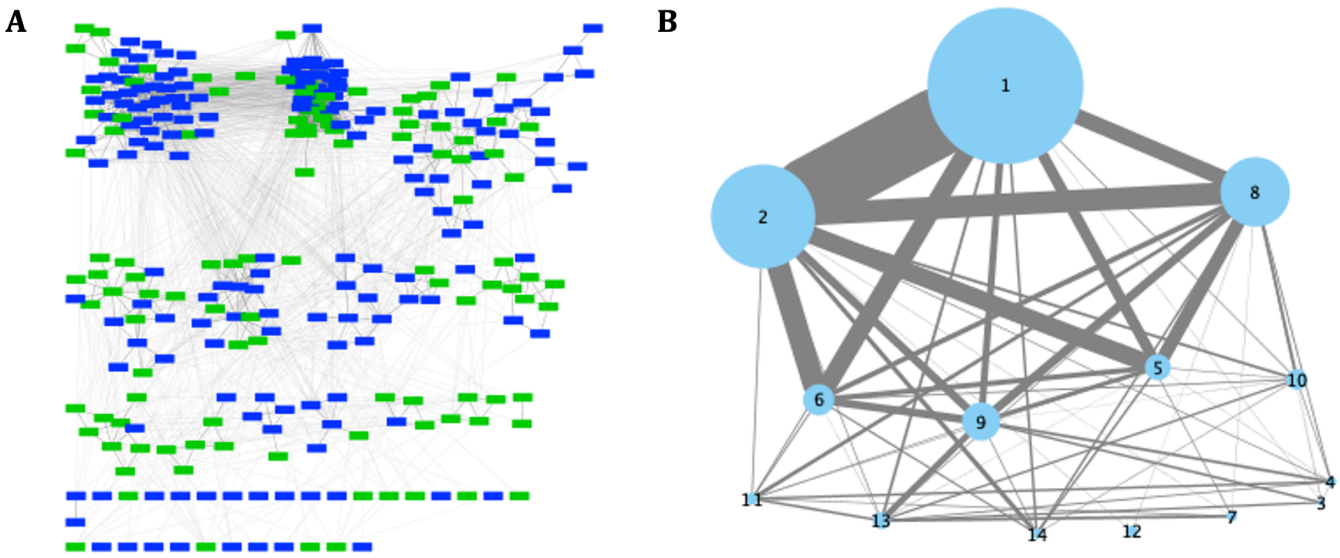


Figure 3. Visualization of protein interaction network and gene clusters associated with SLE and CAD. (**A**) Protein-protein interactions of predicted DE genes and their UPRs were obtained with STRING, visualized with Cytoscape for visualization and clustered using MCODE. Green nodes represent SNP-predicted genes; blue nodes represent UPRs. (**B**) MCODE clusters were further simplified into metaclusters where the size of each cluster represents the number of intra-cluster connections and the edge weight represents the number of inter-cluster connections.

	Big-C Categories	Immuno-Scope	Top Canonical Pathways	P-value
Cluster 1	MHC-Class-I MHC-Class-II Immune-Secreted Pattern-Recognition Receptors	Antigen Presenting Cell Myeloid	T Helper Cell Differentiation	7.11E-39
			Th1 and Th2 Activation Pathway	5.92E-35
			Th1 Pathway	5.53E-34
			Antigen Presentation Pathway	1.07E-31
			T Cell Exhaustion Signaling Pathway	1.42E-30
Cluster 2	Intracellular-Signaling, Immune-Secreted, MHC-Class-I, Nuclear-Receptor-Transcription, Proteasome, Pattern-Recognition-Receptors, Immune-Signaling	NK Cell Myeloid	Cardiac Hyperthrophy Signaling (Enhanced)	5.27E-12
			Systemic Lupus Erythematosus in B Cell Signaling Pathway	4.01E-11
			Crosstalk between Dendritic Cells and Natural Killer Cells	1.69E-09
			Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	2.95E-09
			HMGB1 Signaling	5.08E-09
Cluster 8	MHC-Class-II Anti-Apoptosis General-Transcription Chromatin-Remodeling DNA-Repair	Antigen Presenting Cell Myeloid	Assembly of RNA Polymerase II Complex	2.88E-06
			Primary Immunodeficiency Signaling	1.33E-04
			Estrogen Receptor Signaling	1.55E-04
			VDR/RXR Activation	4.97E-04
			Glucocorticoid Receptor Signaling	5.15E-04
Cluster 6	Immune Signaling	T Cell and Myeloid	Rac Signaling	1.06E-04
			Cdc42 Signaling	3.46E-04
			Acute Phase Response Signaling	4.24E-04
			B Cell Receptor Signaling	4.67E-04
			Signaling by Rho Family GTPases	1.04E-03
Cluster 9	Nuclear-Receptor-Transcription Secreted-and-ECM Immune-Secreted	Monocytes	FXR/RXR Activation	7.11E-12
			LXR/RXR Activation	2.62E-06
			Complement System	3.30E-06
			Atherosclerosis Signaling	1.25E-04
			IL-12 Signaling and Production in Macrophages	1.53E-04
Cluster 5	Unfolded-Protein-and-Stress Chromatin-Remodeling		BAG2 Signaling Pathway	5.98E-06
			Glucocorticoid Receptor Signaling	6.71E-06
			Unfolded protein response	1.33E-05
			Huntington's Disease Signaling	3.94E-05
			Aldosterone Signaling in Epithelial Cells	2.94E-04

Table 1 of Big-C categories, immune cell types, and canonical pathways significantly associated with SLE/CAD genes in MCODE clusters. Significant (p-value < 0.05) Big-C categories and immune cell types were obtained using AMPEL's in-house genomic platform. Top 5 canonical pathways and associated p-values were obtained from IPA variant effect analysis.

Finally, individual gene clusters were analyzed using IPA and AMPEL's in-house genomic platform to characterize the pathways, molecules, and cell types associated with disease. Functional annotation was determined by AMPEL's Biologically Informed Gene Clustering (BIG-C), a functional aggregation tool developed to understand the functional groupings of large gene sets. Similarly, AMPEL's I-Scope was used to detect immune and inflammatory cell type signatures within large gene sets to identify dominant immune cell populations driving disease pathology. Using a significance threshold of p-value < 0.05, canonical pathways, functional annotations, and cell types enriched within the gene sets were identified using IPA, BIG-C, and I-Scope, respectively (Table 1 and methods). Clusters 1, 2, and 8 were heavily dominated by immune-based processes, including the *TH1 and TH2 activation pathway* and *SLE in B cell signaling pathway*, whereas clusters 6 and 9 are enriched in pathways associated with acute inflammation (*acute phase response signaling* and *complement system*) and cluster 5 and 10 with cell stress and repair (*unfolded protein response* and *nuclear excision repair pathway*). I-Scope categories also reveal enrichment in myeloid-lineage cells and/or monocytes, in line with the role of these cells in the development of both SLE and CAD.

Confirmation of immune pathways associated with CAD

To examine the reproducibility of the predicted molecular pathways underlying CAD in SLE, this analysis was repeated using a new set of SNPs. As Illumina's Immunochip contains SNPs highly associated with major autoimmune and inflammatory diseases, such as Crohn's disease and Diabetes, the overlap between the 16,163 CAD-associated SNPs and approximately 250,000 SNPs included on the Immunochip was used.

Of the 16,163 CAD SNPs, 2,467 SNPs (~15%) are included on the Immunochip (Figure 4A). The Monte Carlo Simulation Method was used to estimate the probability of observing an overlap of at least 2,467 between the Immunochip SNPs and an unrelated set of 16,163 SNPs. 10,000 simulations were executed to generate a null distribution of overlap size resulting from intersecting the 252,969 Immunochip SNPs and a randomly generated subset of 16,163 SNPs from the over 7 million SNPs accounted for in the CAD GWAS study (Figure 4B). Out of the 10,000 iterations, 0 resulted in an overlap of 2,467 SNPs or more. As such, using Monte Carlo Simulations, the probability of the observed overlap between the Immunochip SNPs and 16,163 random SNPs is estimated to be p-value < 0.0001.

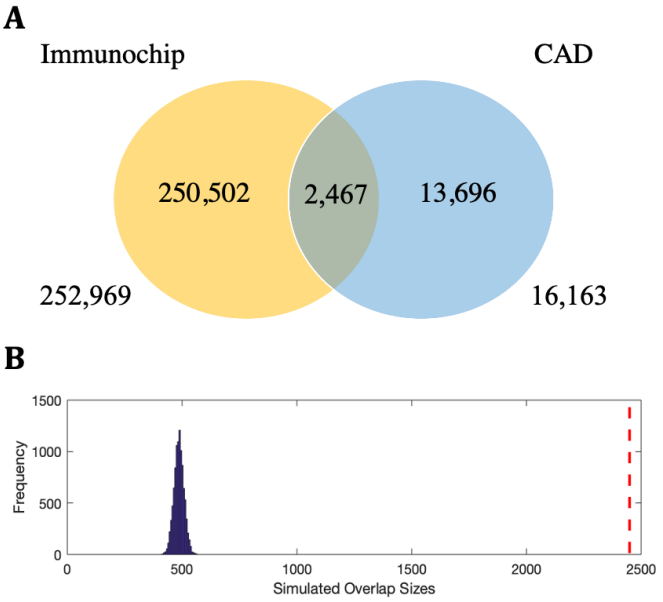


Figure 4. Immunochip SNPs significantly associated with CAD. **(A)** Venn diagram of Immunochip SNPs and SNPs significantly associated with CAD (p-value < $1*10^{-6}$). (**B**) Histograms of the distribution of overlap sizes between the 252,969 SNPs included on the Immunochip and 10,000 random subsets of 16,163 GWAS SNPs.

The 2,467 SNPs mapped to 915 genes total, including coding, expression, target, and proximal genes. Differential gene expression in SLE tissue was not used to filter out genes predicted by Immunochip SNPs associated to CAD to preserve non-SLE-specific genes implicated in CAD. Using a significance threshold of p-value < 0.05, 497 genes were determined to be UPRs of the expression networks significantly enriched in the set of SNP-predicted genes. Interactions between the 915 predicted genes and their 497 UPRs were identified by STRING, clustered using MCODE, and visualized as metaclusters in Cytoscape (Figure 5A). Using a significance threshold of p-value < 0.05, canonical pathways, functional annotations, and cell types significantly enriched within the individual gene clusters were identified using IPA, BIG-C, and I-Scope, respectively (Table 2).

The majority of top canonical pathways, functional annotations, and cell types enriched in the Immunochip/CAD gene clusters correspond to the analysis of SLE/CAD gene clusters. Furthermore, pathway analysis of cluster 2 mirrors that of cluster 1 from the SLE/CAD gene network, as both are dominated by immune-based processes such as *TH1 and TH2 activation pathway, antigen presentation pathway, T helper cell differentiation,* and *T cell exhaustion signaling pathway*. Similarly, cluster 3 parallels cluster 9 from the SLE/CAD gene network, as the genes are both enriched in monocytes and are indicative of

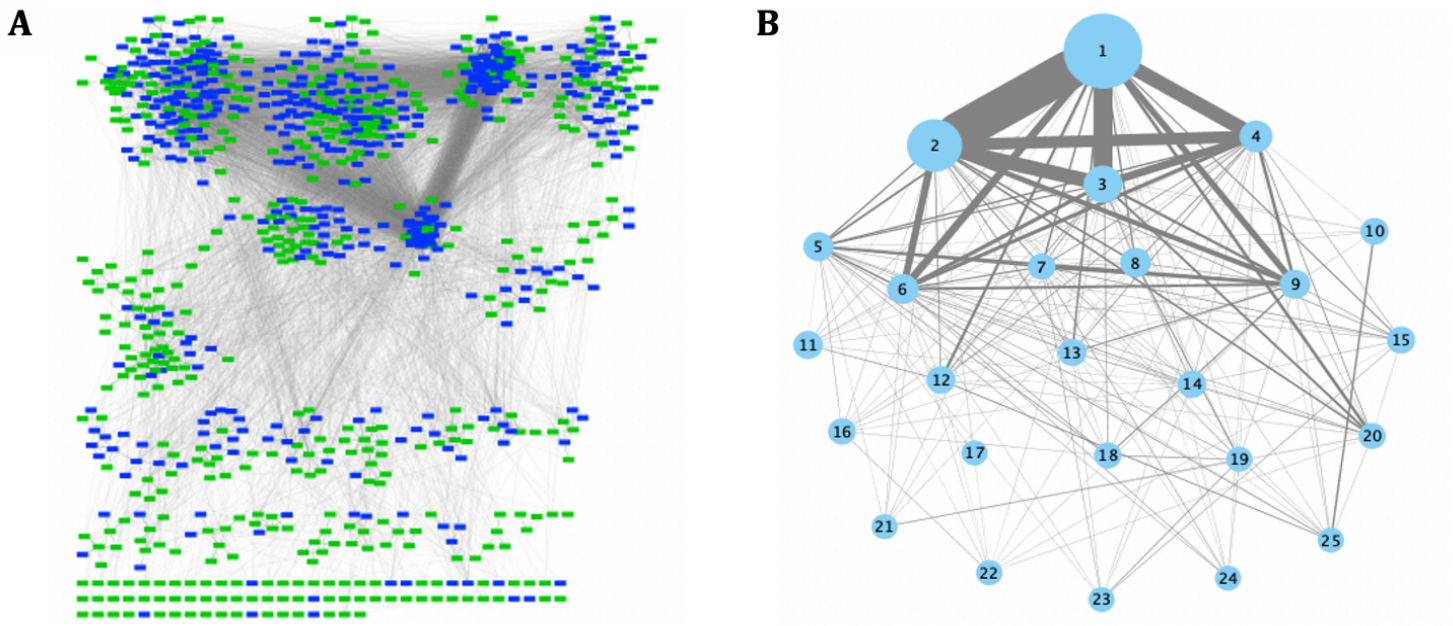


Figure 5. Visualization of protein interaction network and gene clusters associated with CAD and major autoimmune and inflammatory disease. (**A**) Protein-protein interactions of predicted genes and their UPRs were obtained with STRING, visualized with Cytoscape for visualization and clustered using MCODE. Green nodes represent SNP-predicted genes; blue nodes represent UPRs. (**B**) MCODE clusters were further simplified into metaclusters where the size of each cluster represents the number of intra-cluster connections and the edge weight represents the number of inter-cluster connections.

	Big-C Categories	Immuno-Scope	Top Canonical Pathways	P-value
Cluster 1	Intracellular-Signaling, Immune-Secreted, Integrin-Pathway, Nuclear-Receptor-Transcription	B and Myeloid cells, Anergic or Activated T cells, NK or T cells	JAK/Stat Signaling	5.012E-20
			Glucocorticoid Receptor Signaling	3.162E-18
			IL-3 Signaling	5.012E-18
			Systemic Lupus Erythematosus In B Cell Signaling Pathway	6.31E-18
			Role of JAK family kinases in IL-6-type Cytokine Signaling	1.585E-16
Cluster 2	MHC-Class-I MHC-Class-II Nuclear-Receptor-Transcription, Pattern-Recognition-Receptors, Immune-Secreted	Antigen Presenting Cell	Antigen Presentation Pathway	7.943E-19
			Th1 and Th2 Activation Pathway	1E-18
			T Cell Exhaustion Signaling Pathway	1.995E-18
			Systemic Lupus Erythematosus In T Cell Signaling Pathway	6.31E-18
			T Helper Cell Differentiation	2.512E-17
Cluster 3	Immune-Secreted, Reactive-Oxygen-Species-Protection, Secreted-and-ECM, Nuclear-Receptor-Transcription, Ubiquitylation-and-Sumoylation, Autophagy, Mitochondria-General	Monocytes	FXR/RXR Activation	3.162E-17
			LXR/RXR Activation	2.512E-16
			Hepatic Fibrosis Signaling Pathway	6.31E-14
			Cardiac Hypertrophy Signaling (Enhanced)	1E-10
			IL-12 Signaling and Production in Macrophages	2.188E-08
Cluster 4	Intracellular-Signaling, Immune-Secreted, Integrin-Pathway, Nuclear-Receptor-Transcription		Clathrin-mediated Endocytosis Signaling	5.129E-10
			NRF2-mediated Oxidative Stress Response	4.074E-06
			LXR/RXR Activation	2.042E-05
			Huntington's Disease Signaling	3.09E-05
			Unfolded protein response	3.09E-05
Cluster 6	Nuclear-Receptor-Transcription, Chromatin-Remodeling, Transcription-Factors, Mitochondria-TCA-Cycle		Notch Signaling	5.248E-06
			VDR/RXR Activation	0.0001096
			PPARα/RXRα Activation	0.0003631
			Role of Oct4 in Mammalian Embryonic Stem Cell Pluripotency	0.0003981
			RAR Activation	0.0004169

Table 2 of Big-C categories, immune cell types, and canonical pathways significantly associated with Immunochip/CAD genes in MCODE clusters. Significant (p-value < 0.05) Big-C categories and immune cell types were obtained using AMPEL's in-house genomic platform. Top 5 canonical pathways and associated p-values were obtained from IPA variant effect analysis.

nuclear receptor transcription with *FXR/RXR activation* and *LXR/RXR activation* as the most significant canonical pathways.

Potential drug targets in predicted molecular pathways underlying CAD in SLE

Lastly, target identification and drug matching was performed on the molecular pathways associated with CAD in SLE. Potential therapeutic targets and drugs were identified within the predicted gene networks using IPA and AMPEL's in-house genomic platform. The drugs matched to potential therapeutic targets were ranked using the Combined Lupus Treatment Scoring (CoLTs) system, which has been developed to provide a hypothesis-based approach to rank potential therapeutic candidates [16]. This system takes into account scientific rationale, experiments in lupus mice and human cells, and any previous clinical experience in autoimmunity, drug properties, and adverse event profile. FDA-approved drugs are ranked on a scale of -16 to 11 and drugs in development are ranked on a scale of -5 to 8.

The current standard-of-care treatment for SLE is belimumab has a CoLTs score of 5. Including belimumab, 19 candidate drugs with a score greater than or equal to 5 were matched to the gene network associated with CAD in SLE (Figure 6). Of the 19 candidate drugs, 8 are already FDA-approved for treatment of other conditions. Additionally, 2 of the FDA-approved candidate drugs, curcumin and bortezomib, have a higher CoLTs score than belimumab.

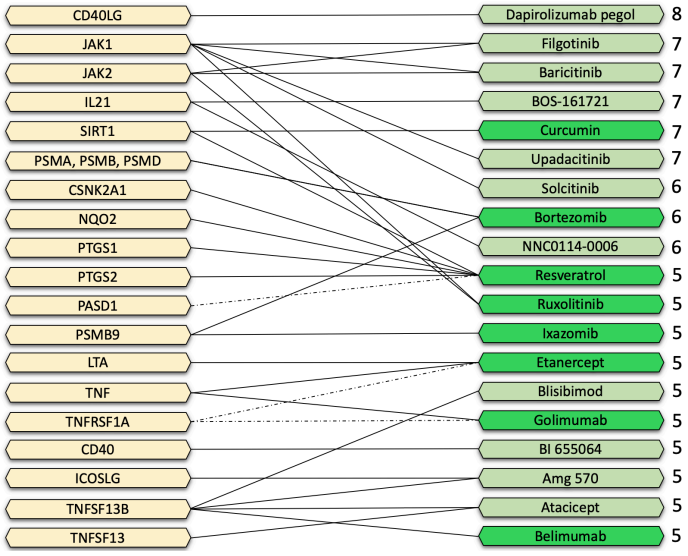


Figure 6. Visualization of existing drugs targeting potential therapeutic targets within SLE/CAD gene networks. Drugs targets (left column, yellow) were identified within the molecular pathways enriched in SLE/CAD genes and matched to existing compounds (right column, green) using AMPEL's in-house genomic platform, including direct targets (solid line) and indirect targets (dashed line). Identified FDA-approved drugs (bright green) and drugs in development (light green) were ranked using the Combined Lupus Treatment Scoring (CoLTs) system (numbers on far right).

DISCUSSION

Our approach comprehensively maps genetic variants to molecular pathways for drug discovery. Here, we mapped 96 SNPs significantly associated with both CAD and SLE to genes and UPRs for protein network construction and drug targeting. While immune dysfunction is not currently considered a CVD risk factor, analysis of the SLE/CAD gene clusters suggest the involvement of immune pathways in the development of CVD. This is also supported by the sizable overlap of Immunochip and CAD-associated SNPs. Ultimately, 18 novel drug candidates with CoLTs scores equal to or greater than the current standard-of-care drug, belimumab, were identified. Furthermore, both curcumin and bortezomib are FDA-approved and scored above belimumab. Curcumin is used as a chemotherapeutic agent in numerous types of cancer and exhibits anti-inflammatory activity [17]. Bortezomib inhibits the proteasome enzyme complex and is also used for treatment of cancers, including myeloma and lymphoma [18]. Additional studies are needed to assess the therapeutic potential of these drugs for SLE.

While this approach has been used to determine ancestry-specific pathways contributing to SLE (Owen et al., submitted manuscript), it has not previously been used to investigate molecular mechanisms shared between associated conditions. Furthermore, our analysis has been improved by testing the statistical significance of the observed SNP overlap via Monte Carlo simulations. Additionally, our results using the CAD-associated Immunochip SNPs demonstrate that the products of this analysis are reproducible. Next steps for this study include further validation of predicted genes, pathways, and drugs, using different datasets. Similarly, performing the full analysis on randomly generated subsets of SLE or CAD-associated SNPs can serve as a control for comparison. Lastly, while our current approach is exceptionally comprehensive in mapping SNPs to genes, genes implicated in both CAD and SLE via distinct genomic variants are not accounted for here. As such, modifying this approach by first mapping all significant SLE and CAD SNPs to genes, then overlapping both gene sets for pathway analysis and drug targeting, may be more robust.

MATERIALS & METHODS

CAD GWAS studies

1. van der Harst, Pim, and Niek Verweij. "Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease." (2018)

SLE Immunochip and GWAS studies

1. Langefeld, Carl D., et al. "Transancestral mapping and genetic load in systemic lupus erythematosus." (2017)
2. Lessard, Christopher J., et al. "Identification of a systemic lupus erythematosus susceptibility locus at 11p13 between PDHX and CD44 in a multiethnic study." (2011)
3. Morris, David L., et al. "Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus." (2016)
4. Sun, Celi, et al. "High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry." (2016)

Statistical analysis of overlap between SNPs associated with both SLE and CAD

Using a significance threshold of p-value < $1*10^{-6}$, 7,222 and 16,163 SNPs significantly associated with SLE and CAD were identified, respectively. 96 of these SNPs were significantly associated with both SLE and CAD. The Monte Carlo Simulation method was used to estimate the probability of an overlap of at least 96 SNPs between 7,222 and 16,163 unrelated SNPs. This method can be used to assess the significance of an outcome by simulating the event many times for a close approximation of the outcome probability.

Implemented in MATLAB, a random subset of equivalent size to the set of significant SLE or CAD associated SNPs was selected from all SNPs tested for in the respective study. The random subset is then intersected with the significant SNPs associated to the other disease or another random subset of that size. This is repeated 10,000 times to generate a null distribution of the number of SNPs occurring in unrelated subsets containing 7,222 and 16,163 SNPs (Figure 1 B-D). The null distributions were then used to estimate the probability that an overlap of 96 SNPs is obtained from intersecting random sets of 7,222 and 16,163 SNPs. The estimated probabilities were determined by calculating the percent of trials resulting in an overlap of 96 or more SNPs.

First, the likelihood of 96 SNPs overlapping the 7,222 significant SLE SNPs and 16,163 unrelated SNPs was estimated by generating random subsets of 16,163 SNPs from the over 7 million SNPs included in the CAD GWAS (Figure 1B). Similarly, 838 SNPs were randomly selected from the Immunochip SNPs and 6,497 SNPs were randomly selected from the GWAS SNPs. Both random subsets were then overlapped with the 16,163 CAD SNPs and the total number of unique SNPs overlapping the CAD SNPs were recorded to generate a null distribution (Figure 1D). There were 113 SNPs determined to be significantly associated with SLE by both the Immunochip and GWAS results, hence 7,222 SNPs total. However, when 838 and 6,497 random SNPs were separately chosen, there was rarely overlap, generating closer to 7,335 SNPs. The simulation airs on the safer side by holding the number of SNPs identified in each study constant as opposed to the total number, thus determining the overlap of CAD SNPs with over 7,222 SNPs. Lastly, a third simulation was performed in which both sets of SNPs were randomly generated as described for the other simulations (Figure 1C).

Identification of SLE-associated SNPs and predicted genes

Expression quantitative trait loci (eQTLs) were identified using GTEx version 68 (GTEXportal.org) ("The Genotype-Tissue Expression (GTEx) Project" n.d.) and mapped to their associated eQTL expression genes (E-Genes). To find SNPs in enhancers and promoters, and their associated transcription factors and downstream target genes (T-Genes), we queried the atlas of Human Active Enhancers to interpret Regulatory variants (HACER, http://bioinfo.vanderbilt.edu/AE/HACER) (Wang et al. 2019)). To find structural SNPs in protein-coding genes (C-Genes), we queried the human Ensembl genome browser (GRCh38.p12; www.ensembl.org) and dbSNP (www.ncbi.nlm.nih.gov/snp). Additional databases were used to generate loss-of-function prediction scores, including SIFT4G (http://sift-dna.org/sift4g (Vaser et al. 2016; Sim et al. 2012)). All other SNPs were linked to the most proximal gene (P-Gene) or gene region as previously detailed(Langefeld et al. 2017). For overlap studies, Venn diagrams were computed and visualized using InteractiVenn (interactivenn.net) (Heberle et al. 2015).

Genomic functional categories

The Variant Effect Predictor (VEP) tool available on the Ensembl genome browser 93 (https://www.ensembl.org) 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, synonymous and nonsynonymous (missense and nonsense) mutations. The online resource tool HaploReg (version 4.1; https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward and Kellis 2016) were also used to identify DNA features, regulatory elements and assess regulatory potential. The dbSNP tool available on the NCBI browser (https://www.ncbi.nlm.nih.gov/snp/) was used to confirm gene predictions.

Functional gene set analysis and identification of upstream regulators (UPRs)

Predicted genes were examined using Biologically Informed Gene Clustering (BIG-C; version 4.4.). BIG-C is a custom functional clustering tool developed to annotate the biological meaning of large lists of genes. 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 (Catalina, Bachali, et al. 2019; Catalina, Owen, et al. 2019).

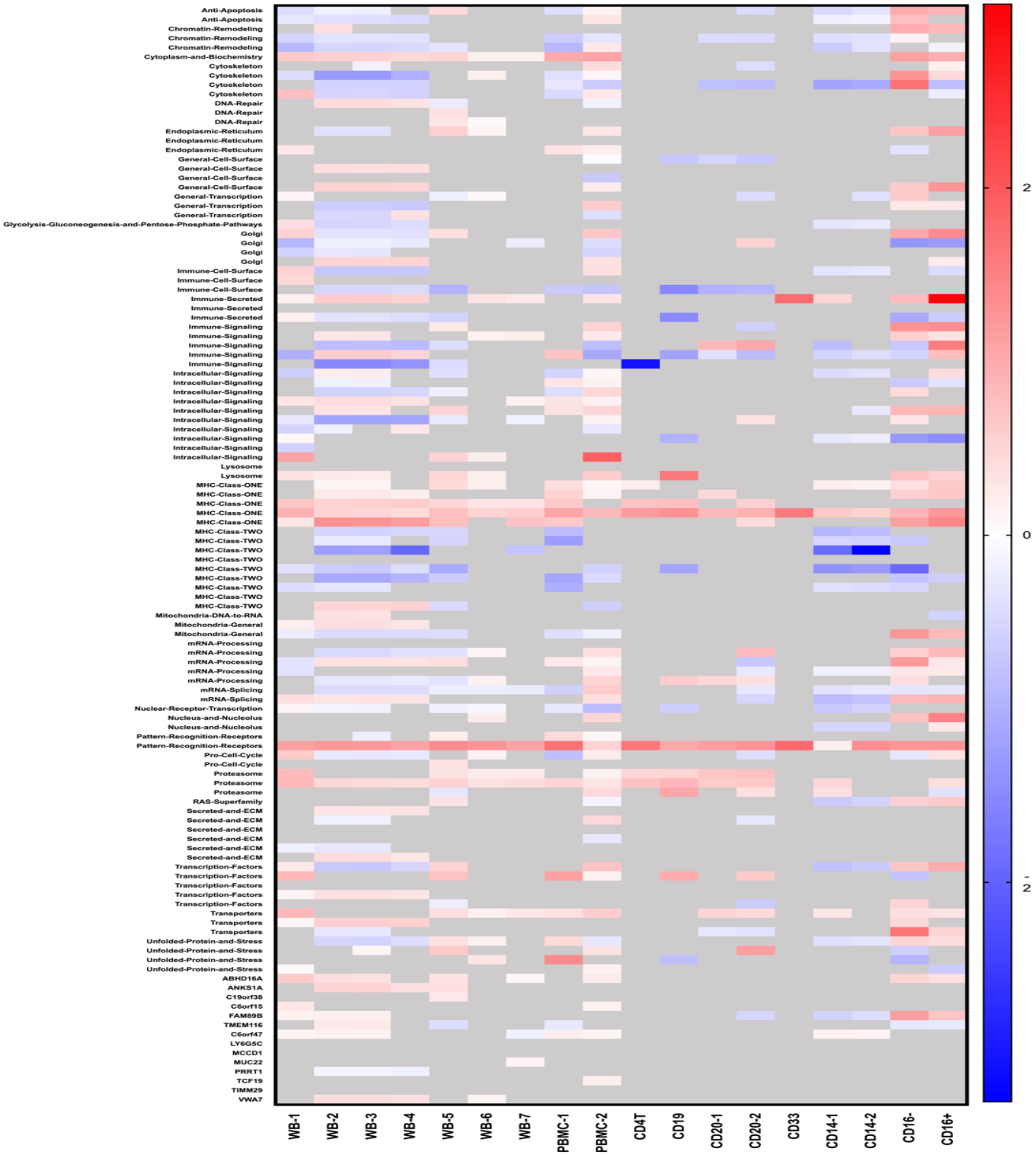
I-Scope is a custom clustering tool used to identify immune infiltrates in large gene datasets, and has been described previously (Ren et al. 2019). Briefly, I-Scope was created through an iterative search of more than 17,000 genes identified in more than 50 microarray datasets. These genes were researched 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 & B cells, B cells, activated B cells, T &B & monocytes, monocytes & B cells, MHC Class II expressing cells, monocyte dendritic cells, dendritic cells, plasmacytoid dendritic cells, Langerhans cells, myeloid cells, plasma cells, erythrocytes, neutrophils, low density granulocytes, granulocytes, platelets, and all hematopoietic stem cells.

Enrichment of GO Biological Processes (BP) using the Database for Annotation, Visualization and Integrated Discovery (DAVID; david.ncifcrf.gov) and the Ingenuity Pathway Analysis (IPA; https://www.qiagenbioinformatics.com) 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 (version 3.6.1) software. Briefly, STRING (version 1.3.2) generated networks were imported into Cytoscape (version 3.6.1) and partitioned with MCODE via the clusterMaker2 (version 1.2.1) plugin.

SUPPLMENTARY FIGURES



Supplementary Figure 1. Comparison of SLE/CAD SNP-associated genes with SLE differential expression datasets. SNP-associated genes were matched with SLE differential expression (DE) data and organized by BIG-C category. The heatmap shows the relative average expression of differentially expressed genes (rows). Columns represent different differential expression datasets and are labeled by their tissue or cell type. Grey boxes represent insignificant differential expression.

REFERENCES

1. Genetics Home Reference, NIH. (2019, September 10). Systemic lupus erythematosus. Retrieved from https://ghr.nlm.nih.gov/condition/systemic-lupus-erythematosus#definition.
2. Zeller, C. B., & Appenzeller, S. (2008). Cardiovascular disease in systemic lupus erythematosus: the role of traditional and lupus related risk factors. Current cardiology reviews, 4(2), 116–122. doi:10.2174/157340308784245775
3. Liu, Y., & Kaplan, M. J. (2018). Cardiovascular disease in systemic lupus erythematosus. Current Opinion in Rheumatology, 30(5), 441–448. doi: 10.1097/bor.0000000000000528
4. Leonard, D., Svenungsson, E., Dahlqvist, J., Alexsson, A., Ärlestig, L., Taylor, K., … Rönnblom, L. (2018). Novel gene variants associated with cardiovascular disease in systemic lupus erythematosus and rheumatoid arthritis. doi: 10.1136/annrheumdis-2017-212614
5. Björnådal L, Yin L, Granath F, et al. Cardiovascular disease a hazard despite improved prognosis in patients with systemic lupus erythematosus: results from a Swedish population based study 1964-95. J Rheumatol 2004;31:713–9.
6. Bernatsky S, Boivin JF, Joseph L, et al. Mortality in systemic lupus erythematosus. Arthritis Rheum2006;54:2550–7. 10.1002/art.21955
7. Nasonov, E., Soloviev, S., Davidson, J. E., Lila, A., Togizbayev, G., Ivanova, R., … Pereira, M. H. (2015). Standard medical care of patients with systemic lupus erythematosus (SLE) in large specialised centres: data from the Russian Federation, Ukraine and Republic of Kazakhstan (ESSENCE). *Lupus science & medicine*, *2*(1), e000060. doi:10.1136/lupus-2014-000060
8. Aringer M, Burkhardt H, Burmester GR et al. Current state of evidence on “off label” therapeutic options for systemic lupus erythematosus, including biological immunosuppressive agents, in Germany, Austria, and Switzerland — a consensus report. Lupus 2012;21:386-401 doi:10.1177/0961203311426569
9. Ciccacci C. (2018). Discovering the genetic contribution to cardiovascular diseases in patients affected by autoimmune diseases. *Annals of translational medicine*, *6*(Suppl 1), S44. doi:10.21037/atm.2018.09.67
10. Alenghat F. J. (2016). The Prevalence of Atherosclerosis in Those with Inflammatory Connective Tissue Disease by Race, Age, and Traditional Risk Factors. *Scientific reports*, *6*, 20303. doi:10.1038/srep20303
11. Langefeld, C. D., Ainsworth, H. C., Cunninghame Graham, D. S., Kelly, J. A., Comeau, M. E., Marion, M. C., … Vyse, T. J. (2017). Transancestral mapping and genetic load in systemic lupus erythematosus. *Nature communications*, *8*,16021.doi:10.1038/ncomms16021
12. van der Harst, P., & Verweij, N. (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation research, 122*(3), 433–443. doi:10.1161/CIRCRESAHA.117.312086
13. Grammer, A. C., & Lipsky, P. E. (2017). Drug repositioning strategies for the identification of novel therapies for rheumatic autoimmune inflammatory diseases. *Rheumatic Disease Clinics*, *43*(3), 467-480.
14. Lipsky, P. E. (2017). SP0156 How big data help us understand new and old therapy targets.
15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 2003 Nov; 13(11):2498-504
16. Grammer, A. C., Ryals, M. M., Heuer, S. E., Robl, R. D., Madamanchi, S., Davis, L. S., … Lipsky, P. E. (2016). Drug repositioning in SLE: crowd-sourcing, literature-mining and Big Data analysis. *Lupus*, *25*(10),1150–1170. https://doi.org/10.1177/ 0961203316657437
17. Marchiani, A., et al. "Curcumin and curcumin-like molecules: from spice to drugs." Current medicinal chemistry 21.2 (2014): 204-222.
18. Cleveland Clinic Cancer. (n.d.). Bortezomib. Retrieved from http://chemocare.com/chemotherapy/drug-info/bortezomib.aspx