Molecular Therapy
Nucleic Acids
Original Article

Identification of Regulatory Modules That Stratify Lupus Disease Mechanism through Integrating Multi-Omics Data

Ting-You Wang,[1] Yong-Fei Wang,[1] Yan Zhang,[1] Jiangshan Jane Shen,[1,2,3] Mengbiao Guo,[1] Jing Yang,[1] Yu Lung Lau,[1] and Wanling Yang[1]

[1]Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong; [2]Collaborative Innovation Center for Birth Defect Research and Transformation of Shandong Province, Jining Medical University, Jining, China; [3]Lupus Research Institute, Affiliated Hospital of Jining Medical University, Jining, China

Although recent advances in genetic studies have shed light on systemic lupus erythematosus (SLE), its detailed mechanisms remain elusive. In this study, using datasets on SLE transcriptomic profiles, we identified 750 differentially expressed genes (DEGs) in T and B lymphocytes and peripheral blood cells. Using transcription factor (TF) binding data derived from chromatin immunoprecipitation sequencing (ChIP-seq) experiments from the Encyclopedia of DNA Elements (ENCODE) project, we inferred networks of co-regulated genes (NcRGs) based on binding profiles of the upregulated DEGs by significantly enriched TFs. Modularization analysis of NcRGs identified co-regulatory modules among the DEGs and master TFs vital for each module. Remarkably, the co-regulatory modules stratified the common SLE interferon (IFN) signature and revealed SLE pathogenesis pathways, including the complement cascade, cell cycle regulation, NETosis, and epigenetic regulation. By integrative analyses of disease-associated genes (DAGs), DEGs, and enriched TFs, as well as proteins interacting with them, we identified a hierarchical regulatory cascade with TFs regulated by DAGs, which in turn regulates gene expression. Integrative analysis of multi-omics data provided valuable molecular insights into the molecular mechanisms of SLE.

INTRODUCTION

Systemic lupus erythematosus (SLE [MIM 152700]) is a chronic autoimmune disease with extreme clinical heterogeneity. In recent years, genome-wide association studies (GWASs) have significantly advanced our understanding of the genetic architecture of SLE, revealing more than 80 susceptibility loci.[1–3] However, the identified variants so far only explain approximately 20% of disease heritability for SLE.[4] It is noteworthy that the majority of identified risk variants are located outside of protein coding regions,[5] which highlights their potential roles in gene expression regulation.

Besides advances in SLE genetics, gene expression as an intermediate phenotype can provide valuable information for understanding the molecular mechanisms of the disease and insights into the effects of genetic variation.[6] So far, the most striking and interesting finding is the dominant pattern of the interferon (IFN) gene expression signature from patients with SLE using high-throughput technologies such as expression microarrays.[7] However, the specific contribution of different IFN families and family members to both the IFN signature and overall SLE pathogenesis is still poorly understood.[8]

Despite the achievements in genetics and transcriptomics for SLE, the existing studies treat them as isolated layers of aberrations that may lead to disease manifestations with little understanding of the interplay of these changes. Fortunately, technological advances have revolutionized the omics field, including a variety of roadmaps of regulatory elements that were revealed by international collaborative projects, such as the Encyclopedia of DNA Elements (ENCODE) project[9] and the Genotype-Tissue Expression (GTEx) project.[10] Thus, integrative analysis of such advances may help us to better explain the complicated disease mechanisms of SLE.

In this study, starting from identifying differentially expressed genes (DEGs) using publicly available data from T cells, B cells, and peripheral blood cells (PBCs) from SLE patients and matched healthy controls, we performed an integrative analysis of various types of biological data for SLE by adapting both data-driven and knowledge-based approaches (Figure 1). The strategies used may provide a novel means for interpretation of large-scale datasets, and the findings may expand our understanding of gene expression regulation and its roles in SLE pathogenesis.

Received 21 November 2018; accepted 11 November 2019;
https://doi.org/10.1016/j.omtn.2019.11.019.

Correspondence: Wanling Yang, Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong.
E-mail: yangwl@hku.hk
Correspondence: Yu Lung Lau, Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong.
E-mail: lauylung@hku.hk

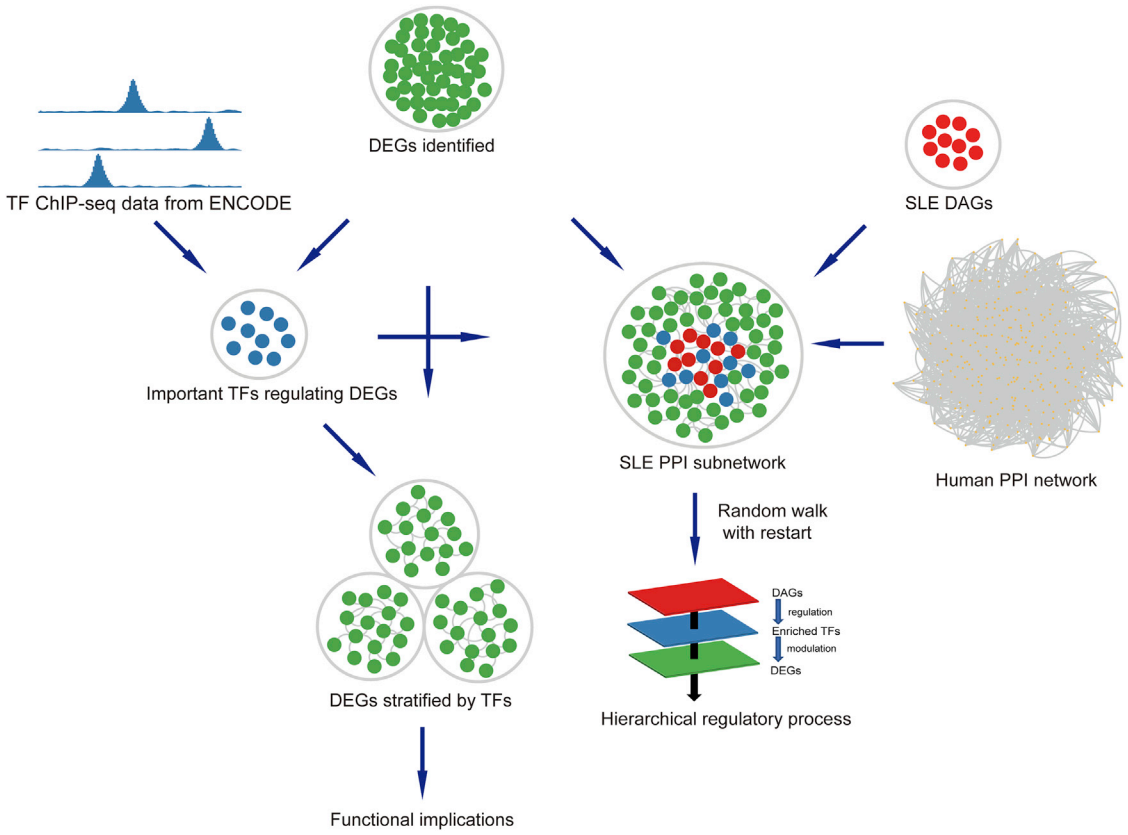

Figure 1. Schematic Overview of This Study

First, we identified DEGs using meta-analysis of datasets on SLE transcriptomic profiles. In the data-driven approach, important TFs regulating DEGs were identified and then they were used to stratify DEGs based on TF binding profiles. In the knowledge-based approach, an SLE PPI sub-network was built by incorporating the information of PPIs. When a random walk with restart algorithm was applied for this network, a hierarchical regulatory process was suggested.

RESULTS

Meta-analysis to Identify DEGs in SLE

The selected gene expression datasets comprised 60 (32 cases and 28 controls), 65 (38 cases and 27 controls), and 132 (65 cases and 67 controls) samples for T cells, B cells, and PBCs, respectively. Meta-analysis was performed in order to combine the summary statistics from different studies to increase power and to minimize potential problems caused by inter-study variation. Comparison between SLE cases and controls identified 215 DEGs (154 upregulated and 61 downregulated) for T cells, 265 DEGs (155 upregulated and 110 downregulated) for B cells, and 378 DEGs (218 upregulated and 160 downregulated) in PBCs. Significant overlap of the upregulated DEGs between the three types of cells was observed, but to a much lesser extent for the downregulated DEGs (Figure 2).

Enriched Gene Ontology (GO) terms represented by the DEGs are shown in Figure S1. Upregulated DEGs were mostly involved in functions such as "response to virus," "type I interferon signaling pathway," and "response to interferon-gamma." Remarkably, besides type I and type II IFN, upregulated DEGs in PBCs were also involved in "neutrophil degranulation," "positive regulation of inflammatory response," and "innate immune response," whereas functions such as "translational initiation" and "ribonucleoprotein complex assembly" were significantly enriched for the downregulated DEGs in PBCs. The functions of "neutrophil degranulation" and "regulation of inflammatory response" are consistent with the involvement of NETosis[11] and inflammatory pathways in SLE. Intriguingly, housekeeping genes are enriched in the downregulated DEGs compared to non-housekeeping genes (chi-square test p value = 7.66×10^{-4}), consistent with suppression of basic cellular functions such as protein synthesization as a defensive mechanism under viral infection or inflammation.

Identification of the Transcription Factors Mediating Differential Gene Expression in SLE

As a key component in gene expression regulation, transcription factors (TFs) play a central role in immune function regulation. Based on ENCODE chromatin immunoprecipitation sequencing (ChIP-seq) datasets, through analysis of TF binding peaks in the transcription start site (TSS) regions of the DEGs in comparison to those from the same number of randomly chosen genes, we identified a number of TFs significantly enriched in regulating the upregulated DEGs

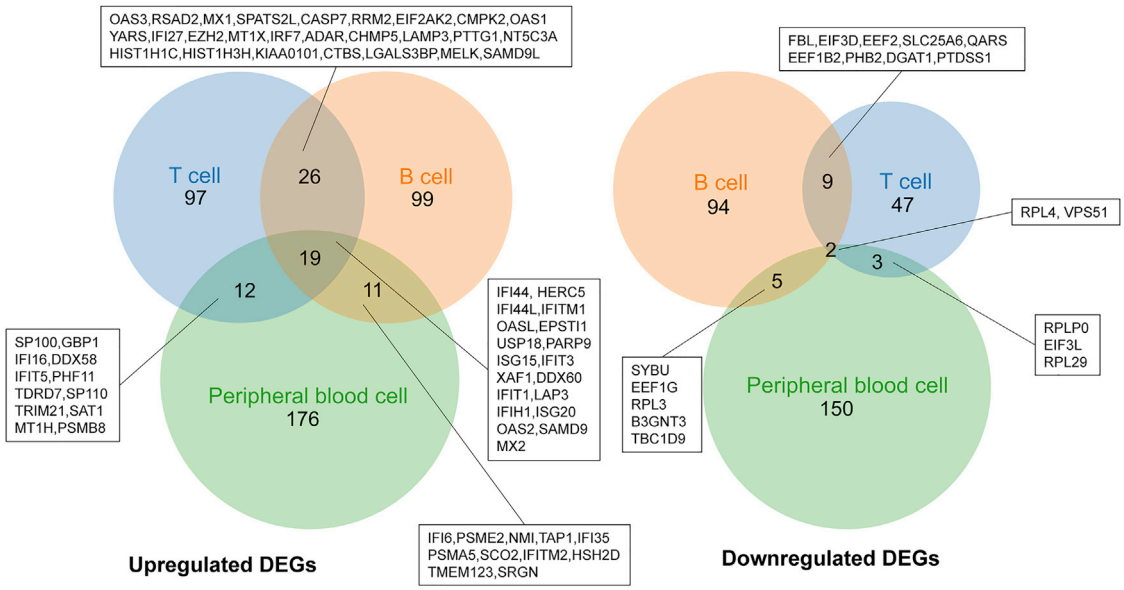

Figure 2. Overlapping of DEGs from T Cells, B Cells, and PBCs

(Figure 3), suggesting key roles of these TFs in initiating and/or maintaining the transcription characteristics of dysregulated gene expression in SLE.

Interestingly, 18 TFs were found significantly enriched for all three types of cells. In addition, the degree of sharing for TFs among the cells is much higher than that for DEGs themselves (26.5% for TFs versus 4.3% for DEGs) (Figure 3), suggesting signal coalescence in gene expression regulation. The most prominent TFs among them are those particularly important for IFN-I signaling, such as *STAT1* and *STAT2*, two essential components of the IFN-stimulated gene (ISG) factor 3 (ISGF3) complex that binds to IFN-stimulated response element (ISRE) in the promoters of ISGs. *IKZF1* and *IKZF2*, two SLE susceptibility genes[2,12] that play a critical role in the pathogenesis of SLE,[13] are also found enriched for all three types of cells in regulating the upregulated genes. Furthermore, the susceptibility variant in *IKZF1* (rs4917014) is found to be a *trans*-expression quantitative trait locus (eQTL), associated with expression of *C1QB* and five ISGs,[14] while four of the five ISGs were also found upregulated in SLE samples in our study. Although no *cis*-eQTL information was available for this SNP, it is very likely that *IKZF1* is responsible for the *trans*-effect of rs4917014, considering the role of *IKZF1* in regulating gene expressions as a TF.

For some of the TF ChIP-seq data, stimulated cell lines were used, which provided us an opportunity to investigate same TFs under different treatments. Remarkably, treatment of IFNs significantly enhanced binding of these enriched TFs to the upregulated DEGs. Upon IFNα treatment for 6 h, signal transducer and activator of transcription 1 (STAT1) and STAT2 were 14-fold more likely to bind to the TSS regions of the upregulated DEGs in SLE than for randomly chosen genes in both B cells and T cells (Figure 3). This observation was in agreement with the chronicity of IFNα production in SLE patients as the most prominent molecular manifestation. Meanwhile, the difference in fold changes for STAT2 between IFNα 0.5-h and 6-h treatments is much bigger than that for STAT1, indicating that the STAT1 effect can reach steady-state sooner than STAT2 upon INFα treatment. This observation also suggested that compared with STAT1, STAT2 may be more sensitive and constitutive for IFN-I-stimulated transcriptional responses.[15]

The difference in fold changes for STAT1/2 and IFN regulatory factor 1 (IRF1) binding to DEGs was observed across all three types of cells (Figure 3). Taking STAT1 as an example, although the fold change for STAT1 with IFNα treatment is higher than that with IFNγ treatment (Figure 3B), a prominent IFNγ response indicated that type II IFN (IFNγ) also plays an important role in SLE pathogenesis,[16] which was consistent with our GO enrichment results for the upregulated genes (Figure S1).

In the K562 cell line, STAT1 or STAT2 binds to a number of ISGs (*OAS3*, *ISG15*, *HERC5*, and *IFI6*) only upon IFNα treatment for 6 h but not after a 30-min treatment (Figure 4), which suggests that sustained IFNα treatment may be required for inducing some of the long-term responses in SLE. Of note, the STAT1 binding profile of the upregulated DEGs showed that the response genes for 6-h IFNα and IFNγ treatment are almost mutually exclusive (Figure 4), which suggests that IFN-I and IFN-II signaling pathways may contribute differently to SLE pathogenicity. Interestingly, it was observed that early response genes of STAT2 upon IFNα treatment were enriched in cell cycle regulation, and late response genes of STAT1 upon IFNγ treatment were enriched in apoptosis (Figure 4), suggesting the important role of these two biological processes in SLE etiology and the different roles of the IFN-I and IFN-II pathways.

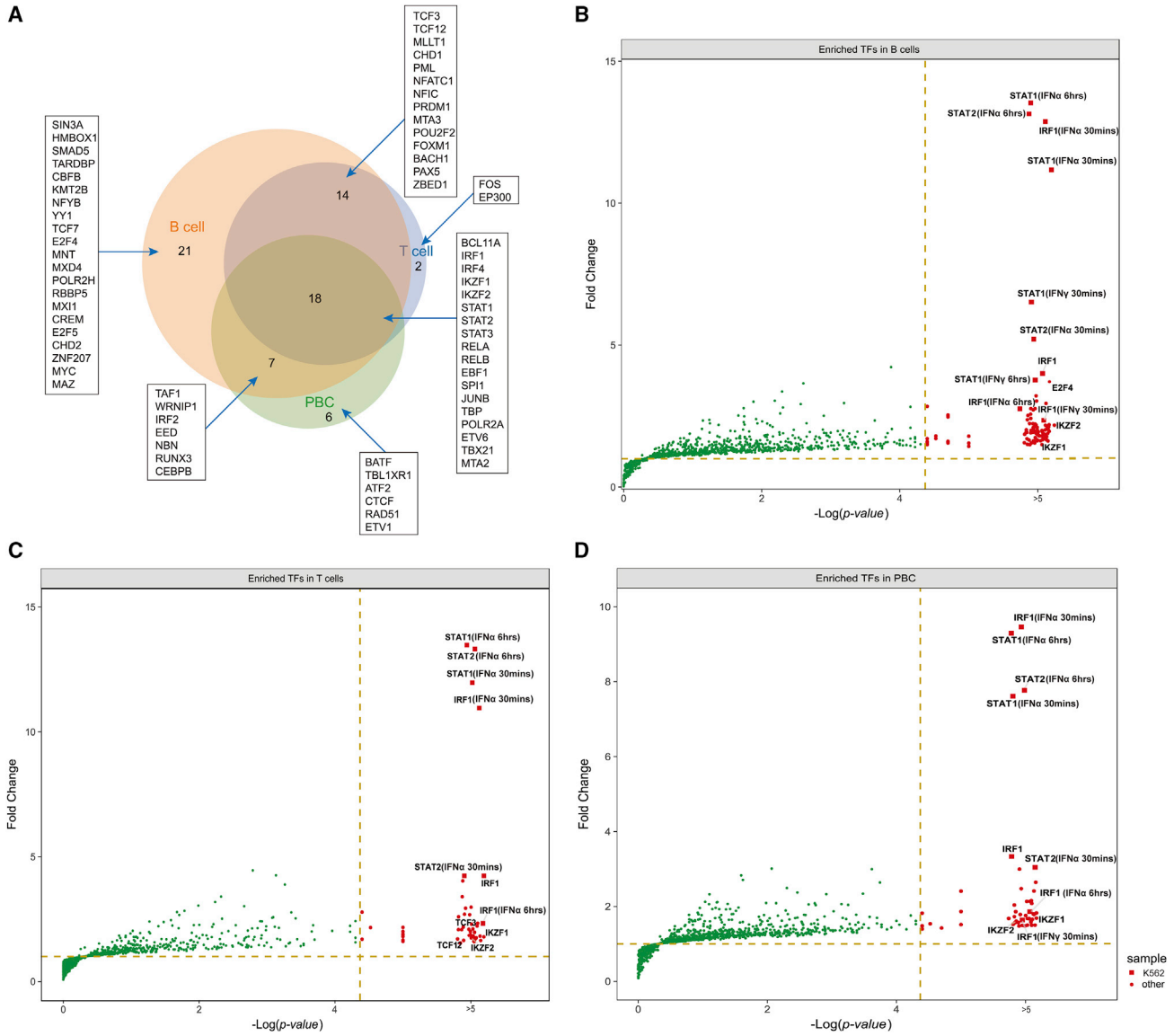

Figure 3. TFs Enriched in the Upregulated DEGs

(A) Comparing the TFs enriched in the TSS regions of the upregulated DEGs in T cells, B cells, and PBCs. (B–D) TFs enriched in the TSS regions of the upregulated DEGs in B cells (B), T cells (C), and PBCs (D). TFs with high fold change or small p values were highlighted. Table S3 showed details of enriched TFs in upregulated DEGs

Identification of NcRTF and NcRG by a Data-Driven Approach

Regulation of transcription in eukaryotes is a complicated process and involves coordination of multiple TFs and cofactors. Using upregulated DEGs and the TFs important in regulating their expression (Figure 3), we tried to infer a network of co-regulated genes (NcRG) and a network of co-regulating TFs (NcRTFs) for SLE. The SLE NcRTF (Figure S2) was composed of 74 TFs with 255 interactions, inferring close interactions and coordination of TFs in regulating gene expression in SLE. TFs that tend to regulate a similar set of genes may act cooperatively. This assumption was exemplified by the module on the right in the NcRTFs, which includes *STAT1* and *STAT2* that are known to form an ISGF3 complex to induce ISGs, and STAT3, which is known to bind histone acetyltransferase EP300 to promote interleukin-10 signaling.[17]

The SLE NcRG (Figure 5) was composed of 358 genes with 6,349 interactions. To evaluate this inferred network, we compared it with 1,000 power law-preserving randomized networks[18] based on protein-protein interactions (PPI) data or gene co-expression data. Interestingly, NcRG tends to be more similar to networks using protein interaction (empirical p value = 0.001) rather than co-expression (empirical p value = 0.163), suggesting that these co-regulatory relationships inferred by TF binding reflect more on the shared functionality at the protein level rather than the expression level.

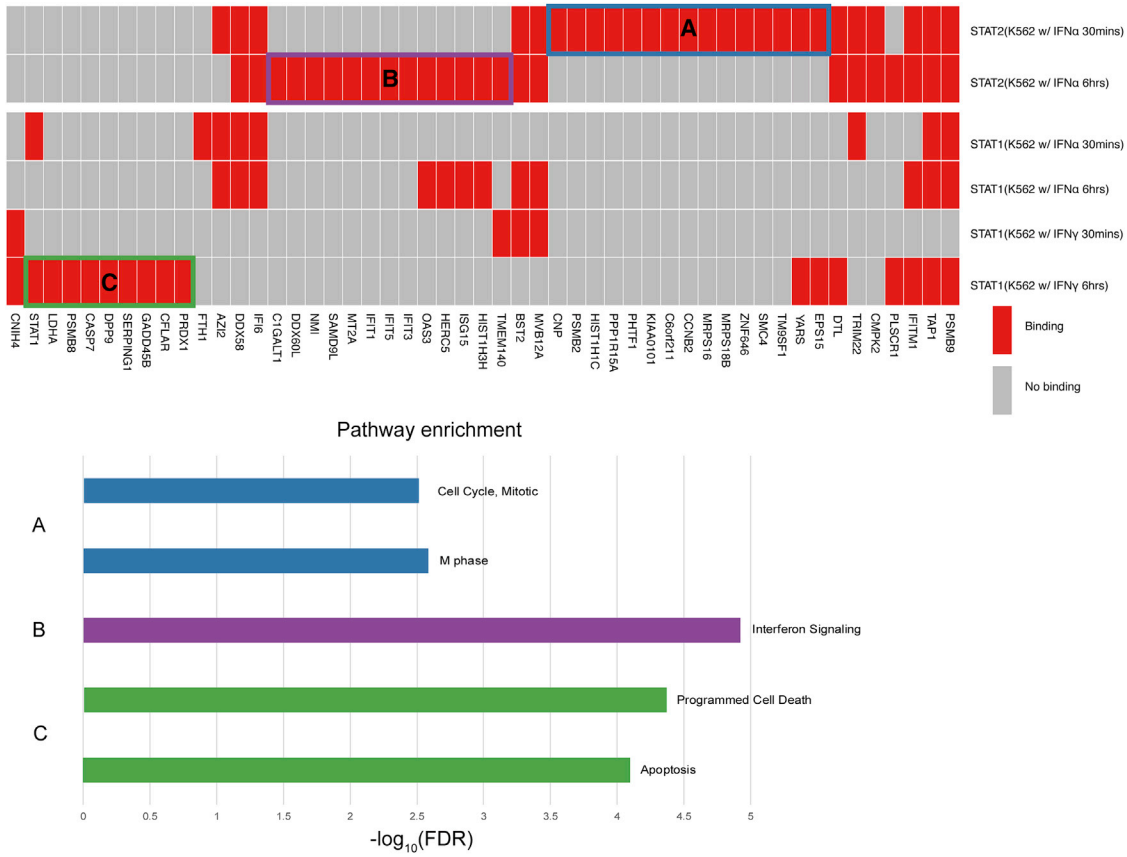

Figure 4. Upregulated DEGs Stratified by STAT1/2 Binding Profiling

(A) Upregulated DEGs only bound by STAT2 with IFNα treatment for 30 min. (B) Upregulated DEGs only bound by STAT2 with IFNα treatment for 6 h. (C) Upregulated DEGs only bound by STAT1 with IFNγ treatment for 6 h.

Modularization of NcRG

The cellular function of a gene cannot be fully understood without understanding its interplay with other genes, and grouping these genes into functional modules may help us better understand the implications of the genes in disease pathogenesis. Within the SLE NcRG, six functional modules (Figure 5) were identified using a community-finding algorithm by maximizing network modularity.[19,20] The modularity score was 0.339, an indication of a moderate community structure in comparison to a random structure for which the modularity score would be equal to 0.

Four of the six modules in the network, with the exception of modules 3 and 6, have enrichment on type I IFN signaling pathway, suggesting functional partitioning of the type I IFN signature. ISGs are also stratified within the network. Module 2 and module 5 had a higher proportion of ISGs, 46% and 19%, respectively (chi-square test p value $< 2.2 \times 10^{-16}$), whereas the remaining modules contain fewer than five ISGs each. Several ISGs that belong to the same gene family also appeared in distinct modules. For example, the positive regulators of oligoadenylate synthetase (OAS), *OAS1* and *OAS2*, are grouped in module 5, whereas *OAS3* is partitioned to module 2, indicating that although they have the same functional domains[21] and similar functions, they may be regulated differently.

Module 3 is the only module involved in the complement cascade, which is known to play an important role in SLE pathogenesis. Two related genes, *C4BPB* and *ELANE*, were found in this module. *C4A/B* and *C1Q* were known contributors for lupus risk, which are involved in immune complex processing and phagocytosis[22,23].

For module 6 (Figure 6), functional enrichment was on cell cycle progression, epigenetic regulation of gene expression, and DNA conformational changes. The major contributing TFs identified for this module included MNT, MYC, E2F4, E2F5, ETV1, PML, CHD2, and SIN3A. The function of these TFs for this module was consistent with that of the genes, even though they themselves are not members of the module. Most of these TFs are regulators of cell cycle, such as E2F4 and E2F5, two essential components of the DREAM (dimerization partner [DP], retinoblastoma [RB]-like, E2F. and multi-vulval class B [MuvB]) complex,[24] and MNT and MYC, which can form a TF network controlling cell cycle progression.[25] It was observed that the vast majority of the genes in this

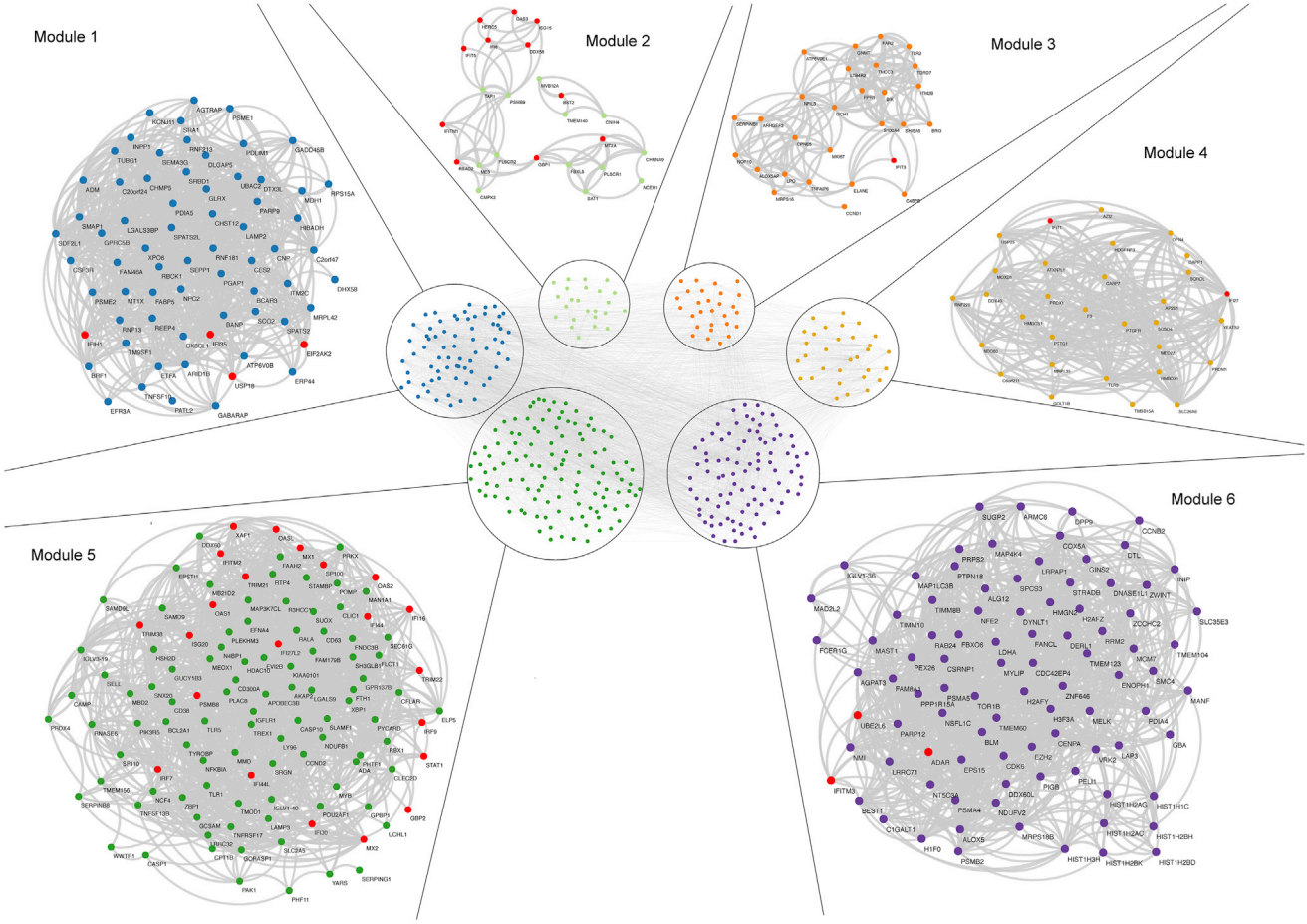

Figure 5. The Modular Repertoire of SLE NcRG

Six modules labeled by different colors were identified using the Louvain algorithm by maximizing network modularity. ISGs are highlighted in red.

module, including two known ISGs, *ADAR* and *UBE2L6*, are regulated by E2F5. Meanwhile, we also found that ETV1, a TF of the ETS family, binds widely to the genes in this module. Remarkably, 35 of the 91 genes in this module were bound by both E2F5 and ETV1, suggesting that E2F5 and ETV1 are upstream regulators of module 6 and play a vital role in dysregulation of cell cycle control and epigenetic regulation in SLE. It is noteworthy that a prominent and upregulated histone cluster was also observed in this module. Upregulation of histone genes may be involved in gene expression dysregulation in SLE and high prevalence of anti-histone antibodies in SLE patients.[26]

Module 2 (Figure 7) had the highest proportion of ISGs (11/24), suggesting that IFN signaling was the major function of this module. STAT1/2/3 were identified as major contributing TFs for the module. *ISG15*, one of the most highly induced ISGs and the main actor of ISGylation,[27] as well as its ligase *HERC5*, a positive regulator of innate antiviral response, belonged to this module, and both genes were bound and likely induced by STAT1 and STAT2 upon 6 h of IFNα treatment. Remarkably, genes in a small cluster, composed of *MVB12A*, *BST2*, *TMEM140*, and *CNIH4*, were all bound by STAT1 upon 30 min of IFNγ treatments, indicating IFNγ involvement in their expression. However, STAT1 upon 6 h of IFNγ treatment is not identified as a major contributing TF in this module, which might suggest that DEGs bound by this TF may contribute to IFN-II signaling pathways in other modules.

Interestingly, genes bound by STATs in this module seem to be sensitive to a different time course in treatment and different interactions among the STATs. For example, upon 6 h of IFNα treatment, STAT1 was found bound to *OAS3*, *HERC5*, *ISG15*, *IFI6*, *DDX58*, *TAP1*, and *PSMB9*, whereas STAT2 was only bound to *OAS3*, *ISG15*, and *HERC5*. Meanwhile, *CNIH4* was bound by STAT1 and STAT3, and *MT2A* was bound by STAT2 and STAT3, whereas *TMEM140* was bound by all three STATs. Further studies are needed to understand the intricate regulation of gene expression in SLE, as hinted by the processes demonstrated by these modules.

Module 5 (Figure S3) is the biggest module in the SLE NcRG and was quite diverse functionally. It includes IFN signaling, regulation

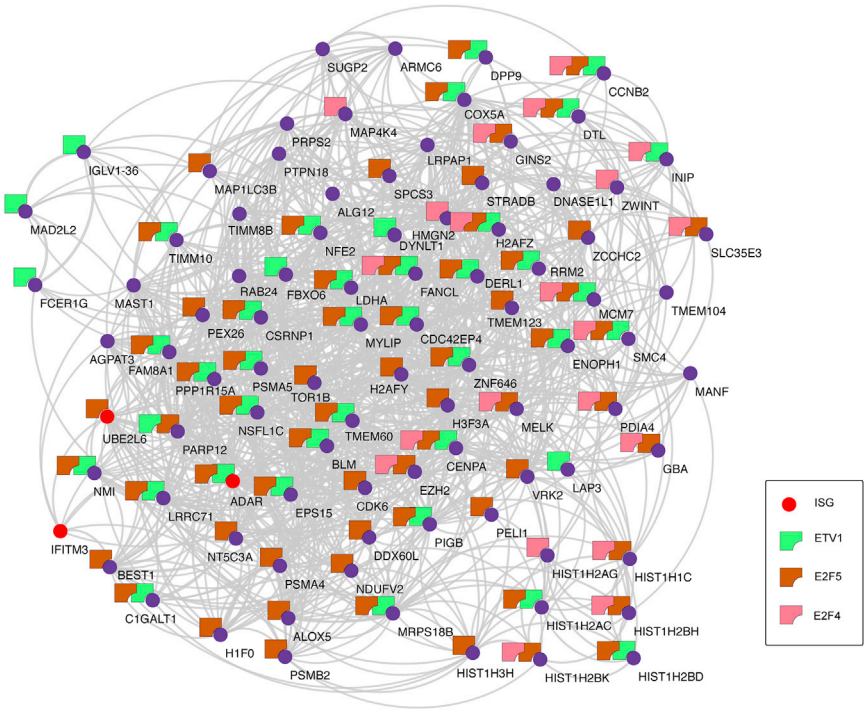

Figure 6. The Regulatory Module 6

This regulatory module is mainly involved in cell cycle progression, epigenetic regulation of gene expression, and DNA conformational changes. E2F4, E2F5, and ETV1 are major contributing TFs.

of cytokine production, Toll-like receptor signaling, and necroptosis pathways. Positive regulators of IFN signaling such as *IRF7*, *IRF9*, and *STAT1* and antiviral effector ISGs such as *MX1/2*, *TRIM21/22/38*, and *IFITM2/21* were observed in this module, supporting the role of IFN in SLE pathogenesis. Interestingly, this module also showed functional enrichment. For instance, neutrophil degranulation, regulation of kidney development, and negative regulation of striated muscle cell differentiation were found enriched in this module, which was in agreement with neutrophil, renal, and heart involvement in SLE pathogenesis and symptoms.[11,28,29] Therefore, this module may represent various phenotypic effects not only for the immune system, but also at the tissue level. The detailed results of GO biological process and pathway enrichment on all the regulatory modules are shown in Tables S1 and S2, respectively.

Identification of a Hierarchical Regulatory System by a Knowledge-Based Approach

A SLE PPI sub-network was constructed, which was composed of a total of 646 DAGs, DEGs, and enriched TFs, with a total of 4,539 interactions based on InWeb_IM. Since DAGs serve as the genetic architecture of SLE pathogenesis, we asked the question of how other genes are ranked as far as their relationships with the DAGs are concerned. To this end, the random walk with restart (RWR) algorithm was used to analyze the proximity of genes to DAGs in the SLE PPI sub-network. Interestingly, the RWR scores of enriched TFs in this study were significantly higher than those of the DEGs (Welch two-sample t test p value = 8.365×10^{-5}), suggesting that these enriched TFs are much closer functionally to DAGs than to DEGs. The DEGs and enriched TFs can be categorized into different layers according to the RWR scores, thus forming a hierarchical regulatory process (Figure 8A).

This layered regulatory model was well exemplified in the pathways involving JAK2 (Figure 8B) and CDKN1B (Figure 8C), respectively. CDKN1B is a cyclin-dependent kinase (CDK) inhibitor and a susceptibility gene reported in one of our previous studies on Asian populations.[3] It plays a critical role in inhibition of cell-cycle progression.[30] Regulators of cell cycle, E2F4 and E2F5 in the DREAM complex, and MYC and MNT in Myc/Max/Mad network are all interacting proteins of CDKN1B. These TFs regulate DEGs in immunity-relevant protein complexes (Figure 8C), such as immunoproteasome, BASC complex (BRCA1-associated genome surveillance complex), and a cluster of different histone proteins. Therefore, it is suggested that CDKN1B, together with other susceptibility genes to be identified, might contribute to cell cycle regulation, DNA repair, and apoptosis in SLE via TFs in the DREAM complex and Myc/Max/Mad network, leading to gene expression aberration.

The hierarchical regulatory system we are proposing in this study was also supported by eQTL data. We surveyed the known SLE susceptibility loci considering the public eQTL data and the upregulated DEGs. Four susceptibility loci (*MIR146A*, *IRF7*, *IKZF1*, and *SH2B3*) were found to be associated with expression changes of DEGs. *SH2B3* (rs10774625) was recently identified as a susceptibility gene for SLE in European populations,[2] and it was associated with expression of three other genes, *STAT1*, *GBP2*, and *UBE2L6*, all of which were found to be upregulated DEGs in this study (Figure S4). These observations suggest potential link between susceptibility genes and DEGs, likely mediated by TFs.

DISCUSSION

Recently, integrative analyses of multi-omics data began to draw attention from the scientific community, aiming to decipher the complexity of disease pathogenesis and molecular mechanisms.[31] In this study, we applied both a data-driven approach and a knowledge-based approach for integrating findings on genetics and transcriptomics, with information on TF binding and PPI, to provide unique insights into the molecular mechanisms of SLE.

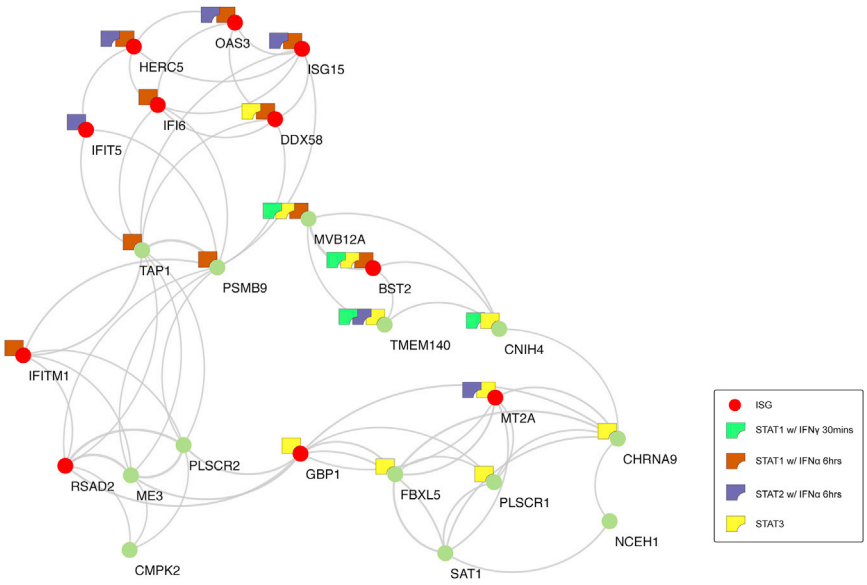

Figure 7. The Regulatory Module 2
This regulatory module is mainly involved in IFN signaling, and STAT1/2/3 are major contributing TFs.

In the data-driven approach, we have made good use of the TF binding profiles to infer a NcRG for SLE. This method provided a novel perspective for understanding gene regulation underlying the disease. Most importantly, upon applying modularization analysis of this inferential network, for the first time the IFN gene expression signature was being stratified, potentially shedding light on IFN signaling in SLE through detailed dissection. Meanwhile, multiple co-regulatory modules and their corresponding upstream regulators (TFs) were identified as well, and they might help us to better understand the functional roles of the DEGs and the regulatory mechanisms involved in SLE. For example, module 2 was the major contributor of IFN signature in SLE pathogenesis. A few TFs were identified as the major regulators in this module, including STAT1 and STAT2, two essential subunits of the ISGF3 complex responsible for the induction of ISGs. They are interacting proteins of a number of susceptibility genes as well, including JAK2 and SOCS1.[32] Therefore, this piece of information suggested that these susceptibility genes may contribute to the IFN signature in SLE via the ISGF3 complex, leading to gene expression aberration shown in module 2.

Additionally, in the data-driven approach, a SLE NcRTF was inferred as well. Besides showing known TF interaction pairs such as STAT1-STAT2 and STAT3-EP300, the NcRTF may also uncover unknown synergistic relationships between the TFs. For example, the key component of the polycomb repressive complex 2 (PRC2), EED, was observed to interact with IKZF1/2 (Figure S2), which are both SLE susceptibility genes, indicating that IKZF1/2 might be involved in the function of EED, leading to epigenetically mediated hypersensitivity and upregulation of ISGs in SLE. It was also observed that EED preferentially binds to ISGs (chi-square test p value = 0.014), consistent with the finding that significant hypomethylation events tend to occur in IFN-related genes.[33]

In the knowledge-based approach, by incorporating the information of PPI, an SLE PPI sub-network was built. It revealed a hierarchical regulatory process consisting of DAGs on the top layer, TFs in the middle layer, and upregulated DEGs in the bottom layer. This regulatory process was also supported by our analysis based on eQTLs data in blood cells. Additionally, Figures 8B and 8C provided two examples illustrating the potential information flow from DAGs to TFs, and then to DEGs. The hierarchical regulatory cascade could be useful in the translation from GWAS findings to clinical utility in the future. Our previous study showed that DAGs tend to interact with SLE drug targets.[34] Thus, based on the regulatory relationship between DAGs and enriched TFs (Table S3), these TFs or genes interacting with them could be promising pharmaceutical targets, providing a new clue to repurposing existing drugs for SLE therapy.

Conclusions

We presented an integrative analysis of DEGs in SLE from T cells, B cells, and PBCs incorporating multi-layer omics data, our results provided a novel way to interpret transcriptomics and also a framework to bridge GWAS findings and gene expression aberrations, and it may provide valuable molecular insights for SLE pathogenesis.

MATERIALS AND METHODS

DEGs

We mined the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database to find publicly available gene expression datasets for SLE. We selected the following datasets for further analysis: T cells, GEO: GSE4588, GSE10325,[35,36] and GSE13887;[37] B cells, GEO: GSE4588, GSE10325, and GSE30153;[38] and PBCs, GEO: GSE12374,[39] GSE20864,[40] and GSE50635.[41] Among them, non-SLE samples and stimulated samples were excluded in our analysis. These datasets were downloaded from the NCBI GEO database using GEOquery[42] R package, and probes were annotated to Entrez Gene identifiers for consistency. Genes with missing values in more than 20% of the samples were excluded from further analysis. Gene expression values were log transformed if necessary and normalized by quantile normalization.[43]

Principal-component analysis (PCA) was employed to overcome hidden confounding factors that may affect gene expression. We included principal components (PCs) that fit the following criteria in our further analysis: (1) they explained more variation than average

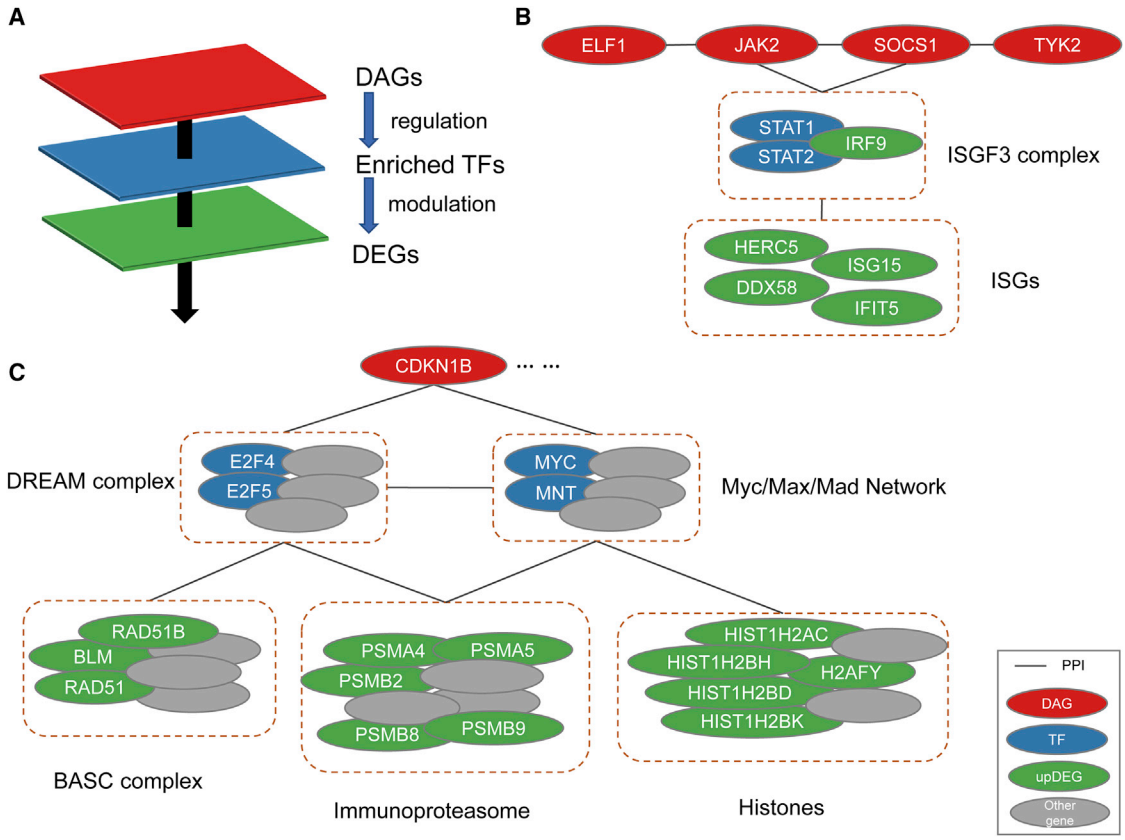

Figure 8. Diagram of the Hierarchical Regulatory Process among DAGs, TFs, and DEGs

(A) DAGs were classified as the top layer, enriched TFs as the middle layer, and upregulated DEGs as the bottom layer in a hierarchical regulatory process for SLE, supporting the notion that susceptibility variants may have contributed to gene expression alteration through TFs, which in turn regulate gene expression aberration in SLE. Two Examples (B (pathway involving JAK2) and C (pathway involving CDKN1B)) showing that DAGs, TFs, and upregulated DEGS are forming regulatory hierarchical networks based on inferred information from protein interaction.

when assuming each variable would contribute equally; and (2) they had no correlation with disease status or other available metadata such as sex and age. A linear regression model with gene expression as the dependent variable, and disease status, selected PCs, and available metadata as independent variables was applied to identify genes that are differentially expressed between cases and controls. A p value and fold change for every gene was generated by linear regression analysis implemented in R.

A weighted Z score approach was applied for meta-analysis across different studies of the same type of cells. Original p values from each study were converted to Z scores, taking into account the sign of the log-transformed fold change as upregulations or downregulations. A weighted sum of Z scores was calculated by weighing each Z score by the square root of the effective sample size for each study. The meta-analysis Z scores were then converted to p values based on a normal distribution. DEGs were determined after correction for multiple testing by Benjamini and Hochberg[44] false discovery rate (FDR) to control the error rate at 0.1 for B cells and T cells. For PBCs, a more stringent cutoff threshold of 1×10^{-3} was used, based on the mixed nature of the cells for peripheral blood and potential variation in their composition.

TF Enrichment Analysis

We collected TF ChIP-seq peak data called using the irreproducible discovery rate (IDR) framework[45] from the ENCODE project[9] (version March 15, 2017, https://www.encodeproject.org/), which includes 1,183 TF-biosample pairs after removing problematic ones with errors in the experiments or unqualified for the consortium's standards. Chromatin accessibility information from DNase peaks can increase reliability of TF binding peaks identified from ChIP-seq data.[46] Thus, we also overlapped TF ChIP-seq data with DNase master peak data from ENCODE project phase 2, which include open chromatin regions from multiple tissues.

To study TFs most relevant in regulating the DEGs, we identified those ChIP-seq peaks that are located within a certain range of the DEGs. NCBI Entrez RefSeq GRCh37 dated in December 2013 was used to define genomic locations and TSSs of the transcripts. The proximal TF binding peaks were assigned to a nearby gene if

they overlapped with the TSS region of the gene, which was defined as the 4-kb region centered on the TSS of the gene. A minimum overlapping size of 100 bp is required, which is the resolution of ChIP-seq technology.[47] For genes with multiple TSSs, the averaged count of binding peaks for different transcripts was used.

Construction of NcRGs and NcRTFs

The TF binding peak profile for each gene was constructed based on the TF ChIP-seq data from ENCODE. Each data point stored the number of binding peaks for a specific TF in the TSS region of the gene. The numbers of TF binding peaks were normalized to a range of 0–1 for comparison purpose using minimum (min)-maximum (max) scaling:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)},$$

where $x = (x_1, \ldots, x_n)$ is an original value denoting the number of TF binding peaks in one ChIP-seq dataset, and z_i is the normalized number of TF binding peaks.

On the basis of the TF binding profiles, Pearson's correlation coefficient (PCC) was employed to measure the correlation between genes or TFs. The gene-gene or TF-TF interaction was determined using PCC on z_i values. Student's t test was used for evaluating the statistical significance of PCC, adopting corrections for multiple testing (at FDR at 0.05 threshold). Thus, NcRGs and NcRTFs were built based on gene and TF interactions, respectively. In this study, based on the TF binding peak profiles of upregulated DEGs and corresponding enriched TFs from T cells, B cells, and PBCs, NcRG and NcRTF for SLE were constructed. In order to illustrate clearer TF co-regulation relationships, in the NcRTF, a PCC cutoff threshold of 0.4 was used[48] in addition to a FDR cutoff.

We utilized a randomized network method[18] to assess the reliability of the NcRG inferred from TF binding peak profiles. The strategy of this approach is to compare this inferred network with 1,000 power law-preserving randomized networks on the basis of external gene interactions. In this study, PPI data from InWeb_IM[49] and gene co-expression data in whole blood from the GTEx project[50] were used as the external gene interactions. In practice, we counted one if the number of gene interactions in the randomized network is bigger than that in the NcRG. The empirical p value was calculated by the count number divided by 1,000.

Modularization of NcRG

Identification of communities and modules within a network improves our understanding of the organization of the biological systems.[51] In order to identify modules in the NcRG, the Louvain algorithm[19,20] was employed to define co-regulatory modules. For every co-regulatory module, the major contributing TFs were identified by L1 regularized logistic regression,[52] which minimizes the classification error while selecting a small number of TFs that have nonzero coefficients.

In order to functionally characterize the co-regulatory modules, GO,[53] Reactome,[54] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[55] pathway annotations were employed to detect overrepresented functions in each co-regulatory module. For enrichment analysis, an FDR corrected p value < 0.05 was considered as significant.

SLE PPI Sub-Network

In order to bridge genetic components and gene expression components of SLE pathogenesis, a SLE PPI sub-network was constructed using SLE DAGs,[1,2] enriched TFs, and identified DEGs on the basis of InWeb_IM,[49] which is so far the most comprehensive protein interaction network stemmed from eight heterogeneous resources. In practice, the SLE PPI sub-network was built using the SLE DAGs, enriched TFs, and DEGs when there are edges between them in the network of InWeb_IM.

Random Walk with Restart to Analyze Relationships among DAGs, DEGs, and TFs

Random walk iteratively is a process that explores the global structure of a network, starting at given source nodes to reach random neighbors in order to estimate the proximity among vertices (genes). As a variant of random walk, the walker may also choose to teleport to the start nodes with a given restart probability r, which controls how far the random walker moves away from the start nodes. In this study, $r = 0.5$ was used, and thus the probability of moving forward and moving backward in every step is equal. The equation for the random walk with restart is defined as:

$$p^{t+1} = (1 - r)Wp^t + rp^0,$$

where r is the restart probability, W is the column-normalized adjacency matrix of the network graph, and p^t is a vector of size equal to the number of nodes in the graph where the ith element holds the probability of being at node i at time step t. The initial probability vector p^0 was constructed such that equal probabilities were assigned to each DAG, while a probability of 0 was given to all other genes in the network. The final score of a gene in the network was defined as the steady-state probability that the random walker would stay at the gene. These final scores can be viewed as the "influential impact" over the network imposed by the start nodes (DAGs). RWR was carried out by NetWalker.[56]

More information is available in Supplemental Materials and Methods.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2019.11.019.

AUTHOR CONTRIBUTIONS

W.Y. and Y.L.L designed the study; T.-Y.W. analyzed the data; Y.-F.W. helped to analyze the data; Y.Z., M.G., and J.Y. provided

expertise on genetics and genomics and contributed to interpretation of the data; and T.-Y.W. wrote the manuscript. J.J.S. contributed to editing the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

W.Y. and Y.-F. W. received grant support from National Key Research and Development Program of China (2017YFC0909001). This work was also supported by the Research Grant Council of Hong Kong (GRF 17146616 and GRF 17125114). The authors thank Hong Kong PhD Fellowship Scheme, HKU Postgraduate Scholarships and the Edward & Yolanda Wong Fund for supporting postgraduate students who participated in this work.

REFERENCES

1. Harley, J.B., Alarcón-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K., et al.; International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN) (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. Nat. Genet. *40*, 204–210.

2. Bentham, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, 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. Yang, W., Tang, H., Zhang, Y., Tang, X., Zhang, J., Sun, L., Yang, J., Cui, Y., Zhang, L., Hirankarn, N., et al. (2013). Meta-analysis followed by replication identifies loci in or near *CDKN1B*, *TET3*, *CD80*, *DRAM1*, and *ARID5B* as associated with systemic lupus erythematosus in Asians. Am. J. Hum. Genet. *92*, 41–51.

4. Yang, W., and Lau, Y.L. (2015). Solving the genetic puzzle of systemic lupus erythematosus. Pediatr. Nephrol. *30*, 1735–1748.

5. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

6. Frangou, E.A., Bertsias, G.K., and Boumpas, D.T. (2013). Gene expression and regulation in systemic lupus erythematosus. Eur. J. Clin. Invest. *43*, 1084–1096.

7. Baechler, E.C., Batliwalla, F.M., Karypis, G., Gaffney, P.M., Ortmann, W.A., Espe, K.J., Shark, K.B., Grande, W.J., Hughes, K.M., Kapur, V., et al. (2003). Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc. Natl. Acad. Sci. USA *100*, 2610–2615.

8. Banchereau, R., Cepika, A.M., Banchereau, J., and Pascual, V. (2017). Understanding human autoimmunity and autoinflammation through transcriptomics. Annu. Rev. Immunol. *35*, 337–370.

9. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

10. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

11. Gupta, S., and Kaplan, M.J. (2016). The role of neutrophils and NETosis in autoimmune and renal diseases. Nat. Rev. Nephrol. *12*, 402–413.

12. Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H., Cai, Z.M., Huang, W., Zhao, G.P., et al. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. *41*, 1234–1237.

13. Hu, S.J., Wen, L.L., Hu, X., Yin, X.Y., Cui, Y., Yang, S., and Zhang, X.J. (2013). IKZF1: a critical role in the pathogenesis of systemic lupus erythematosus? Mod. Rheumatol. *23*, 205–209.

14. 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.

15. Blaszczyk, K., Nowicka, H., Kostyrko, K., Antonczyk, A., Wesoly, J., and Bluyssen, H.A. (2016). The unique role of STAT2 in constitutive and IFN-induced transcription and antiviral responses. Cytokine Growth Factor Rev. *29*, 71–81.

16. Pollard, K.M., Cauvi, D.M., Toomey, C.B., Morris, K.V., and Kono, D.H. (2013). Interferon-γ and systemic autoimmunity. Discov. Med. *16*, 123–131.

17. Hedrich, C.M., Rauen, T., Apostolidis, S.A., Grammatikos, A.P., Rodriguez Rodriguez, N., Ioannidis, C., Kyttaris, V.C., Crispin, J.C., and Tsokos, G.C. (2014). Stat3 promotes IL-10 expression in lupus T cells through *trans*-activation and chromatin remodeling. Proc. Natl. Acad. Sci. USA *111*, 13457–13462.

18. Li, H., and Liang, S. (2009). Local network topology in human protein interaction data predicts functional association. PLoS ONE *4*, e6410.

19. Newman, M.E. (2006). Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA *103*, 8577–8582.

20. Vincent, D.B., Jean-Loup, G., Renaud, L., and Etienne, L. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.

21. Hovanessian, A.G., and Justesen, J. (2007). The human 2′-5′ oligoadenylate synthetase family: unique interferon-inducible enzymes catalyzing 2′-5′ instead of 3′-5′ phosphodiester bond formation. Biochimie *89*, 779–788.

22. Walport, M.J., Davies, K.A., and Botto, M. (1998). C1q and systemic lupus erythematosus. Immunobiology *199*, 265–285.

23. Christiansen, F.T., Dawkins, R.L., Uko, G., McCluskey, J., Kay, P.H., and Zilko, P.J. (1983). Complement allotyping in SLE: association with C4A null. Aust. N. Z. J. Med. *13*, 483–488.

24. Sadasivam, S., and DeCaprio, J.A. (2013). The DREAM complex: master coordinator of cell cycle-dependent gene expression. Nat. Rev. Cancer *13*, 585–595.

25. Adhikary, S., and Eilers, M. (2005). Transcriptional regulation and transformation by Myc proteins. Nat. Rev. Mol. Cell Biol. *6*, 635–645.

26. Rekvig, O.P., van der Vlag, J., and Seredkina, N. (2014). Review: antinucleosome antibodies: a critical reflection on their specificities and diagnostic impact. Arthritis Rheumatol. *66*, 1061–1069.

27. Zhao, C., Collins, M.N., Hsiang, T.Y., and Krug, R.M. (2013). Interferon-induced ISG15 pathway: an ongoing virus-host battle. Trends Microbiol. *21*, 181–186.

28. Tincani, A., Rebaioli, C.B., Taglietti, M., and Shoenfeld, Y. (2006). Heart involvement in systemic lupus erythematosus, anti-phospholipid syndrome and neonatal lupus. Rheumatology (Oxford) *45* (*Suppl 4*), iv8–iv13.

29. Saxena, R., Mahajan, T., and Mohan, C. (2011). Lupus nephritis: current update. Arthritis Res. Ther. *13*, 240.

30. Ophascharoensuk, V., Fero, M.L., Hughes, J., Roberts, J.M., and Shankland, S.J. (1998). The cyclin-dependent kinase inhibitor $p27^{Kip1}$ safeguards against inflammatory injury. Nat. Med. *4*, 575–580.

31. Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol. *18*, 83.

32. Morris, D.L., Sheng, Y., Zhang, Y., Wang, Y.F., Zhu, Z., Tombleson, P., Chen, L., Cunninghame Graham, D.S., Bentham, J., Roberts, A.L., et al. (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.

33. Absher, D.M., Li, X., Waite, L.L., Gibson, A., Roberts, K., Edberg, J., Chatham, W.W., and Kimberly, R.P. (2013). 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. *9*, e1003678.

34. Wang, Y.F., Zhang, Y., Zhu, Z., Wang, T.Y., Morris, D.L., Shen, J.J., Zhang, H., Pan, H.F., Yang, J., Yang, S., et al. (2018). Identification of *ST3AGL4*, *MFHAS1*, *CSNK2A2* and *CD226* as loci associated with systemic lupus erythematosus (SLE) and evaluation of SLE genetics in drug repositioning. Ann. Rheum. Dis. *77*, 1078–1084.

35. Hutcheson, J., Scatizzi, J.C., Siddiqui, A.M., Haines, G.K., 3rd, Wu, T., Li, Q.Z., Davis, L.S., Mohan, C., and Perlman, H. (2008). Combined deficiency of proapoptotic regulators Bim and Fas results in the early onset of systemic autoimmunity. Immunity *28*, 206–217.

36. Becker, A.M., Dao, K.H., Han, B.K., Kornu, R., Lakhanpal, S., Mobley, A.B., Li, Q.Z., Lian, Y., Wu, T., Reimold, A.M., et al. (2013). SLE peripheral blood B cell, T cell and myeloid cell transcriptomes display unique profiles and each subset contributes to the interferon signature. PLoS ONE *8*, e67003.

37. Fernandez, D.R., Telarico, T., Bonilla, E., Li, Q., Banerjee, S., Middleton, F.A., Phillips, P.E., Crow, M.K., Oess, S., Muller-Esterl, W., and Perl, A. (2009). Activation of mammalian target of rapamycin controls the loss of TCRζ in lupus T cells through HRES-1/Rab4-regulated lysosomal degradation. J. Immunol. *182*, 2063–2073.

38. Garaud, J.C., Schickel, J.N., Blaison, G., Knapp, A.M., Dembele, D., Ruer-Laventie, J., Korganow, A.S., Martin, T., Soulas-Sprauel, P., and Pasquali, J.L. (2011). B cell signature during inactive systemic lupus is heterogeneous: toward a biological dissection of lupus. PLoS ONE *6*, e23900.

39. Lee, H.M., Mima, T., Sugino, H., Aoki, C., Adachi, Y., Yoshio-Hoshino, N., Matsubara, K., and Nishimoto, N. (2009). Interactions among type I and type II interferon, tumor necrosis factor, and β-estradiol in the regulation of immune response-related gene expressions in systemic lupus erythematosus. Arthritis Res. Ther. *11*, R1.

40. Lee, H.M., Sugino, H., Aoki, C., and Nishimoto, N. (2011). Underexpression of mitochondrial-DNA encoded ATP synthesis-related genes and DNA repair genes in systemic lupus erythematosus. Arthritis Res. Ther. *13*, R63.

41. Ko, K., Koldobskaya, Y., Rosenzweig, E., and Niewold, T.B. (2013). Activation of the interferon pathway is dependent upon autoantibodies in African-American SLE patients, but not in European-American SLE patients. Front. Immunol. *4*, 309.

42. Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics *23*, 1846–1847.

43. Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

44. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

45. Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. *5*, 1752–1779.

46. Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. *13*, 840–852.

47. Risca, V.I., and Greenleaf, W.J. (2015). Unraveling the 3D genome: genomics tools for multiscale exploration. Trends Genet. *31*, 357–372.

48. Mukaka, M.M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med. J. *24*, 69–71.

49. Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Stærfeldt, H.H., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods *14*, 61–64.

50. Pierson, E., Koller, D., Battle, A., Mostafavi, S., Ardlie, K.G., Getz, G., Wright, F.A., Kellis, M., Volpi, S., and Dermitzakis, E.T.; GTEx Consortium (2015). Sharing and Specificity of co-expression networks across 35 human tissues. PLoS Comput. Biol. *11*, e1004220.

51. Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA *99*, 7821–7826.

52. Lee, S.-I., Lee, H., Abbeel, P., and Ng, A.Y. (2006). Efficient L_1 regularized logistic regression. In Proceedings of the 21st National Conference on Artificial Intelligence, *Volume 1* (AAAI Press), pp. 401–408.

53. Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. *43*, D1049–D1056.

54. Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway Knowledgebase. Nucleic Acids Res. *44* (D1), D481–D487.

55. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44* (D1), D457–D462.

56. Zhang, B., Shi, Z., Duncan, D.T., Prodduturi, N., Marnett, L.J., and Liebler, D.C. (2011). Relating protein adduction to gene expression changes: a systems approach. Mol. Biosyst. *7*, 2118–2127.