Undergraduate Capstone

Transcriptomic Comparison of Lupus Murine Models and SLE Patients to Assist Drug Trial Design
(technical research project in Biomedical Engineering)

by

Christopher Puglisi

April 30, 2020

Technical project collaborators:
Amrie Grammer, PhD
Prathyusha Bachali
Bryan Chun

Words: 5832
Figures and tables: 8
Supplements: 1
References: 25

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 Capstone-Related Assignments.

signed: _____ date: _____

approved: _____ date: _____
Amrie Grammer, PhD

Transcriptomic Comparison of Lupus Murine Models and SLE Patients to Assist Drug Trial Design

Capstone Student: Christopher E. Puglisi [a,1]

Advisors: Prathyusha Bachali [b], Bryan Chun [b], Amrie Grammer [b], Peter Lipsky [b]

[a] Capstone Student
[b] AMPEL BioSolutions
[1] Correspondence: cep4un@virginia.edu, 617-784-4047

Abstract

Systemic lupus erythematosus (SLE) is a multisystem autoimmune disease that causes significant morbidity and mortality, especially in women of child-bearing age. Only one new drug has been approved by the FDA in 60 years for treatment of SLE. The heterogeneity of the disease induces varied symptoms and laboratory abnormalities that occur in different combinations and points in time. To investigate these phenotypic variations, we performed a genome wide mRNA expression comparison between nine human whole blood datasets and three lupus mice splenic datasets. We aim to clarify what aspects of the major preclinical lupus mice used by Pharma for drug studies are relevant to signaling pathways abnormal in human lupus patients that Pharma has developed drugs to treat. Differential expression analysis identified the differentially expressed (DE) genes shared between seven human datasets and three mice strains, BXSB.yaa, NZB/W and MRL/lpr. Through cell type and functional enrichment of the DE genes shared between species, the three strains, BXSB.yaa, NZB/W and MRL/lpr were found to uniquely model the interferon signature, germinal center B cell signature and autophagy, respectively. To characterize the mechanisms of human disease modeled by each strain in greater detail, mouse and human weighted gene co-expression network analysis (WGCNA) modules of genes preserved between species were identified. Cell type and functional enrichment of preserved modules identified that the BXSB.yaa, NZB/W and MRL/lpr mice uniquely modeled cell cycle, the innate immune response and transcription regulation in human disease, respectively. To deconstruct manifestation heterogeneity, a variational autoencoder (VAE) with Gaussian priors was employed in an unsupervised approach to cluster human lupus patients based on clinical characteristics. The VAE clusters identified as the least and most active disease states were used as labels for supervised classification. The supervised models input Log2 gene expression values perturbed with Gaussian noise. The results suggest WGCNA serves as a more robust approach than differential expression for understanding interspecies relationships on the transcriptomic level. Further, WGCNA and module preservation revealed the specific gene modules in mice that regulate immune pathways and model human disease. The results also suggest that unsupervised clustering serves as a novel tool to deconstruct patient heterogeneity by means of clinical expression and that ensemble methods prove most useful to link genetic expression to clinical traits after clusters had been established.

Keywords: Systemic Lupus Erythematosus, Differential Expression, Weighted Gene Co-expression Network Analysis, Gaussian Mixture Variational Autoencoder, Ridge, Random Forest, Gradient Boosted, K-Nearest Neighbors

Introduction

SLE is an autoimmune disease inducing a variety of symptoms including severe fatigue, joint pain, rash and anemia (Schur & Hahn, 2019). The reported prevalence of SLE in the United States is 20 to 150 cases per 100,000, climbing to 406 per 100,000 in African American women (Chakravarty et al., 2007). The heterogeneity of the disease induces varied symptoms and laboratory abnormalities that occur in different combinations and points in time for each patient (Fritzler et al., 2018). The complexity of the disease has challenged immunologists for years, leading to only one new drug to be FDA approved for SLE treatment in 60 years. Clarifying the aspects of the major preclinical lupus mice used by Pharma for drug studies that are relevant to signaling pathways abnormal in human lupus patients would help clinicians in designing preclinical drug trials. No literature currently exists comparing the gene expression of the three main models to be investigated, BXSB.yaa, NZB/W and, MRL/lpr which engender lupus in three distinct manners. All of these mice model limited aspects of lupus, including the mechanism bearing disease progression.

The current state of literature reports the immunological characteristics of each mouse model, but there is a strong overlap in biomarkers and symptom manifestation between mice. All three murine models exhibit glomerulus nephritis, splenomegaly, increased antinuclear antibody expression, increased anti-dsDNA antibody expression and weakened IFN signature (Andrews et al., 1978; Deane et al., 2007; Du, Sanam, Kate, & Mohan, 2015). All three mice imitate immune dysregulation and kidney disease in humans (Gómez-Guzmán et al., 2014; Santiago-Raber et al., 2008). There are apparent differences as well, such as MRL/lpr mice model arthritis and neurological manifestations better than its counterparts, NZB/W and BXSB.yaa, which excel in modelling the endothelial effects of SLE and acute kidney disease, respectively (Santiago-Raber et al., 2008; Virdis et al., 2015). However, the strong overlap in immunological characteristics call for a more robust

understanding of differences in disease manifestation due to genomic variation between species.

Some genetic associations have been made between the models and humans. BXSB.yaa strains double the expression of TLR7 (Santiago-Raber et al., 2008). This gene is crucial for the pathogenesis of the disease in both this strain and human patients as it correlates with IFNα expression (Celhar, Magalhães, & Fairhurst, 2012; Deane et al., 2007). NZB/W mice have the Sle1 and Nba2 loci, which are known to have a syntenic parallel region in humans on chromosome 1, 1q21-44 (Morel et al., 1997). These regions are known to be associated with lupus in both species, specifically to autoantibody production (Shai et al., 1999; Tsao et al., 1997). MRL/lpr mice and lupus patients are both associated with Fas and Fas ligand polymorphisms, which increase susceptibility to lupus (Lee, Choi, Ji, & Song, 2016; Moudi, Salimi, Mashhadi, Sandoughi, & Zakeri, 2013). As this review nearly exhausts the known genetic associations, and considering the heterogeneity of SLE, there remains much to uncover between these models and humans.

The field of computational pharmacology currently does not present a model to highlight the similarities between lupus mice and SLE humans. Further, the field does not indicate ideal treatment for cohorts of lupus patients based on gene expression analysis. However, bioinformaticians are aiming to improve personalized medicine by predicting activity state based on gene expression. In 2018, Kegerreis et al. employed generalized linear models and RF classifiers on microarray data to predict active states of lupus with a peak of 81% accuracy (Kegerreis et al., 2019). This research aims to continue the computationally driven research aspects of personalized medicine in lupus through gene expression analysis.

Through a bioinformatics driven approach, we perform genome wide mRNA expression comparison between nine human whole blood datasets and three lupus mice splenic datasets. Whole blood samples are the best proxy for the splenic samples taken in mice we have access to as both serve to model the innate and adaptive immune response in the disease state (Nikpour et al., 2007). We utilized differential expression techniques to identify the DE genes between healthy humans and SLE patients and mice with and without treatment. By employing an orthogonal analysis to differential expression, we identify gene modules for each data set, grouped by similarity in a scale-free network. We compare the enrichment of genomic signatures of the DE genes shared and the modules preserved between species. We also deconstruct human lupus patient heterogeneity by employing unsupervised clustering and validate the clusters with supervised classification.

Materials and Methods

Selection, QC, and normalization of raw data files

Raw data files for human whole blood samples from SLE patients and healthy controls (HC) were obtained from the publicly accessible Gene Expression Omnibus (GEO) repository. Data from splenic lupus mice were accessed from Roopenian, Reilly and Scholmchik. Accession numbers and descriptions for all datasets used are summarized in S1 Table.

Processing of raw datafiles from microarray samples was conducted with Bioconductor packages GEOquery, affy, affycoretools, and simpleaffy in R. Dataset were assessed for quality control through visual artifacts or poor RNA hybridization using Affymetrix QC plots, and were normalized by RMA, GCRMA or NEQC where appropriate. Log2 intensity values were procured from the normalized expression data and formatted as expression set objects (E-sets). Principal component analysis (PCA) plots were generated for all cell types in each experiment to inspect for batch effects and remove outliers.

The raw data was annotated to the matching Affymetrix product using chip definition files (CDF) and any unidentified genes by Affymetrix CDFs were annotated using the custom definitions from the BrainArray CDF. Alternatively, the raw data was genotyped on the matching Illumina Immunochip. Unnormalized raw data counts collected from RNA-Seq were trimmed for both adaptor sequences and quality. These reads were then aligned to the genome (Ensembl.org) using Bowtie2/Tophat2 and counted via HTSeq.

The chips and genome browsers used for the microarray and RNAseq are summarized in S1 Table. At this point, differential expression analysis and Weighted Gene Co-expression Network Analysis were carried out on data sets.

Differential gene expression analysis

The E-sets were passed through a filter, removing genes or probes that had low intensity or were unannotated. For datasets collected by microarray, the remaining expression values were corrected for variance using local empirical Bayesian shrinkage using the ebayes function in the Bioconductor LIMMA package. RNASeq data was corrected using the lfcShrink function in the DESeq2 package. Benjamini-Hochberg false discovery rate correction was applied. DE genes within each study were filtered to retain DE genes with an FDR < 0.2, which were considered statistically significant. The FDR filter was applied to diminish the number of false negative results.

Weighted gene co-expression network analysis (WGCNA)

Log2 normalized microarray or RNASeq expression values were used as inputs to WGCNA to conduct an unsupervised clustering analysis. This approach resulted in groups of densely interconnected genes we call co-expression modules. For each dataset, the network strength between probe or gene expression was calculated in a scale-free topology matrix (TOM). The probe or gene was clustered into WGCNA modules based on TOM distances. The resultant dendrograms were trimmed to isolate modular groups, cut on a detection height of 1 and merging height of 0.2. The modules were labeled using semi-random color assignments.

Module Preservation

Modules preserved between species were determined by passing three selection criteria. First, we characterized each module by its eigengene, or the module's first principal components. If Pearson correlation by point-biserial correlation found the module eigengene (ME) to be significantly correlated ($p < 0.05$) to the disease state, it was kept for further analysis. Second, we calculated the preservation score given each mouse module to every human module. Modules are preserved if the preservation score, as quantified by an average of co-expression density within modules and connectivity between modules, is over 2. Third, the modules must be correlated to the disease state in the same direction.

Cell type and functional gene characterization

The CellScan and Biologically Informed Gene Clustering (BIG-C) (Grammer et al., 2016) tools characterize genes into cell types and functional groups, respectively. The tools are curated by utilizing publicly available information from online tools and databases including UniProtKB/Swiss-Prot, GO Terms, KEGG pathways, NCBI PubMed, and the Interactome. The functional enrichment of the DE genes was calculated through a Fischer's exact test, referenced against the a priori

tools. Big-C V4.4, human CellScan V1.6 and mouse CellScan V1.0 were used.

Variational Autoencoder (VAE) with Gaussian mixture prior distributions

VAEs are unsupervised models built learn meaningful latent spaces. The VAE is based on an autoencoding framework which can discover linear and nonlinear features through data compression paired with ReLU, Softplus and linear activation functions. Sixteen clinical characteristics from GSE88884 were input to the model and were represented as binary inputs representing yes or no. The traits included human ancestry, lupus manifestations and treatments of drugs. By adapting the generalized architecture of a variational autoencoder (Shu, 2016), we were able to construct an identity function, or the simplest possible representation of the input data.

The standard autoencoder is composed of two sequential neural nets, an encoder and a decoder, and is trained by minimizing reconstruction error. However, the VAE learns the distribution of the explanatory features over samples, through both a mean and standard deviation vector encoding (Way & Greene, 2018).

The goal of an encoder is to create a low-dimensional representation of the input data that represents the prior probability distribution. This prior probability distribution is normally isotropic over the latent variables. However, as lupus patients are greatly heterogenous, we represent the prior distribution as a Gaussian mixture in our model. The goal of a decoder is to reconstruct the original input data from the prior probability distribution in the simplest representation possible.

The network architecture is summarized in figure 4. The encoder consists of three hidden layers, the first two containing eight nodes each and the third containing four nodes. By inputting the clinical characteristics of for each patient individually, we were able to transform the data in the first two layers and use the third layer to summarize the mean and standard deviation of the transformed input data. The decoder is a reflection of the encoder without associated biases for each layer, as input data should model bias, while the simplest reconstruction of that data should not (Dilokthanakul et al., 2017). The network was constructed using Python's TensorFlow (V 1.14) and trained under the AdamOptimizer to minimize cross entropy loss over 500 epochs.

Supervised machine learning algorithm and validation

We employed four machine learning models from three distinct families of classifiers to validate the results of the VAE. The Ridge classifier, k-nearest neighbors classifier (KNN), random forest (RF) classifier, and a gradient boosted (GB) classifier were deployed. The Ridge, RF and GB, and KNN models were deployed using the sklearn .linear model, .ensemble, and .neighbors packages.

The purple and the crimson clusters were determined to be least and most active, respectively. This classification was decided through visual examination of proportionalities of traits with in clusters. The normalized Log2 gene expression from the 430 inactive and 466 active patients was used as inputs for the classifiers. The inputs were filtered to 126 genes, after removing all genes that did not correlate to the disease state (Correlation under 0.25). The models were set to classify patients' activity state under a random 10-fold cross-validation that was carried out by randomly assigning each patient to one of ten groups. One of the groups was used as the test set and the classifier was trained on the remaining data.

To validate the supervised models, gaussian perturbations were added to the genes. A Gaussian distribution was created for the variation each gene in each class. A noise value was selected from the gene's variation distribution and assigned to be added or subtracted from the original value. The same models were rerun with the new noisy input. The receiver operating characteristic (ROC) curves were generated using the sklearn.metrics package.

Results

Differential expression of genes in lupus mice and lupus patients

To assess the genes dysregulated in the disease state, we analyzed gene expression profiles of splenic tissue from lupus mice with and without treatment and compared the results to gene expression profiles of whole blood samples from lupus patients and healthy controls (HC).

Before comparing interspecies gene profiles, we examined the DE genes shared between human whole blood data. For seven of the human datasets (GSE22908, GSE29536, GSE39088, GSE45921, GSE49545, GSE61635 and GSE88884), only 54 statistically significant (FDR < 0.2) DE genes were shared between species (Fig. 1.). The lack of overlapping genes is likely attributed the vast heterogeneity between lupus patients.

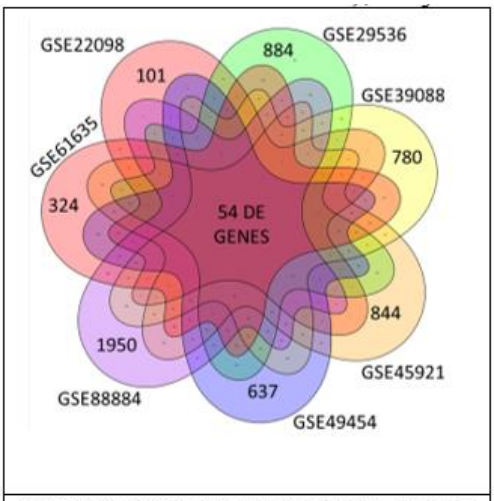


Figure 1. Venn Diagram of DE Human Genes: This figure shows the overlapping DE genes of seven human whole blood datasets.

The heterogeneity is validated by perimeter of figure 1 showing that the majority of DE genes are unique to each dataset. The three mice datasets contained only 64 overlapping genes. This result was expected considering the different clinical manifestations of each strain (Andrews et al., 1978; Theofilopoulos & Dixon, 1985).

Cell type and functional characterization of communal DE gene signatures between species

The significant enrichments ($p < 0.05$, $OR > 1$) of the interspecies cell type and functional changes represented by the divergent gene signatures in SLE patients and lupus mice are summarized in figure 2. The heatmap indicates the significant enrichment of the monocyte, myeloid, B and T cell types between each human dataset and the BXSB.yaa strain. This enrichment validates that the BXSB.yaa strain models both the innate and adaptive immune response of human lupus (Bubier et al., 2009; Herrada et al., 2019). The BXSB.yaa strain models interferon stimulation and pattern recognition receptor pathway, as all human datasets showed communal enrichment of these functions with this strain.

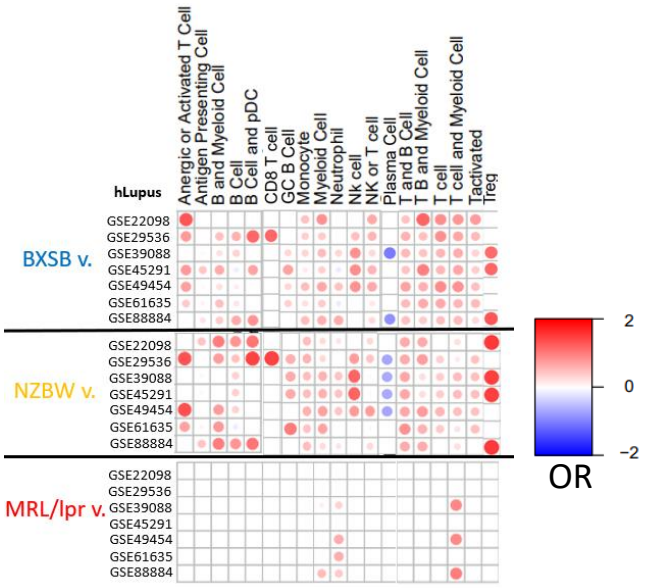


Figure 2. Cell type enrichment heatmap of shared DE genes: The enrichment heatmap shows cell type categories as columns. The rows are for the overlapping DE genes between each mouse and each human dataset. OR is odds ratio.

The DE genes of the NZB/W strain share enrichment of germinal center B and T Regulatory cell types, as well as the anti-proliferation and fatty acid biosynthesis pathways, with the DE genes of four human datasets. The cell type enrichment indicates this strain closely models results expected from spontaneous germinal center activity arising from many human autoimmune disorders (Domeier, Schell, & Rahman, 2017).

The overlapping DE genes between the MRL/lpr models and humans found autophagy and proteasome enrichment in seven human datasets. As perturbations in autophagy lead autoimmune disorders, including lupus, we can conclude that the MRL/lpr models dysregulation of the degradation of cytoplasmic constituents (Pierdominici et al., 2012). However, the MRL/lpr model failed to find significant enrichment of any cell types shared with the human datasets. This was not expected considering the BXSB.yaa and NZB/W strains shared the adaptive immune signatures with humans, and that we expected to see B cell should promote spontaneous T cell activation in the MRL/lpr model (Chan & Shlomchik, 1998).

Further, interspecies comparison showed no enrichments or de-enrichment of plasma cells, which was unexpected considering other studies have measured increased differentiation of B lymphocytes into plasma in all strains and human lupus. (Hutloff et al., 2004; Iii et al., 2004; Terzoglou, Routsias, Moutsopoulos, & Tzioufas, 2004; Yan, Deng, Wang, Sun, & Wei, 2015). These inconsistencies in results suggest that functional enrichment of differential expression analysis fails to characterize all mechanisms driving lupus.

Enrichment of lupus mice gene signatures of disease-correlated WGCNA modules

The inconsistencies from the differential expression analysis were investigated by employing an orthogonal approach, WGCNA, for each mouse dataset. WGCNA generates gene co-expression modules connected in a scale-free network. The modules were prioritized to those significantly correlated to the disease state. WGCNA found four BXSB.yaa, eight NZB/W and thirteen MRL/lpr modules correlated to the disease state.

The majority of cell types excluding plasmacytoid dendritic cells and activated T cells were enriched in at least one module between all three mice. Further, the BXSB.yaa strain uniquely expressed a CD8 T cell signature. This results aligns with established research indicating IL-2 as a potent inducer of CD8+ T cells in BXSB.yaa mice (Bubier et al., 2009). Antigen presenting cells were uniquely enriched in the NZB/W strain. This strain showed a strong enrichment of low-density granulocytes (LDG). The LDG signature is of interest as a target for therapy of lupus humans and has not been identified to be enriched in any mouse model (*LDG*, 2019). The MRL/lpr mouse shows the significant enrichment of the monocyte cell type and the T regulatory cell type.

This analysis serves primarily as a proof of concept for cell type enrichment of the genes within selected module tested, giving the authors confidence in uncovering the enrichment of genes within modules preserved between species.

Strain	Enrichment	Unique Across Strains	Shared Across Strains
BXSB.Yaa	BigC Pathways	Cell Cycle	Inflammatory Signaling, Weak Cell Cycle
	CellScan Cell Types	Interferon Producing Cells	Adaptive Immune Response
NZBW	BigC Pathways	None	Weak Cell Cycle
	CellScan Cell Types	Innate Immune Response	Adaptive Immune Response
MRL/lpr	BigC Pathways	Transcription Regulation	Inflammatory Signaling, Weak Cell Cycle
	CellScan Cell Types	NK/ T Cells	Adaptive Immune Response

Table 1. Preservation Summary: Summarized in the table are the pathways and cell types enriched for the modules preserved between mice and humans. The enrichments are either completely unique to that strain or shared across strains.

Interspecies module preservation enrichment analysis

Functional and cell type enrichment analysis was performed on the genes within the preserved modules between each mouse model and human dataset. In figure 3, the genes within the magenta module of the BXSB.yaa strain and the human modules preserved to the magenta module are analyzed for functional enrichment. The interferon, pattern recognition receptor and pro apoptosis signatures in human lupus were modeled by the magenta module. This comparison was extended between the mouse modules correlated to the disease state and their respective human modules preserved. The results are summarized in table 1.

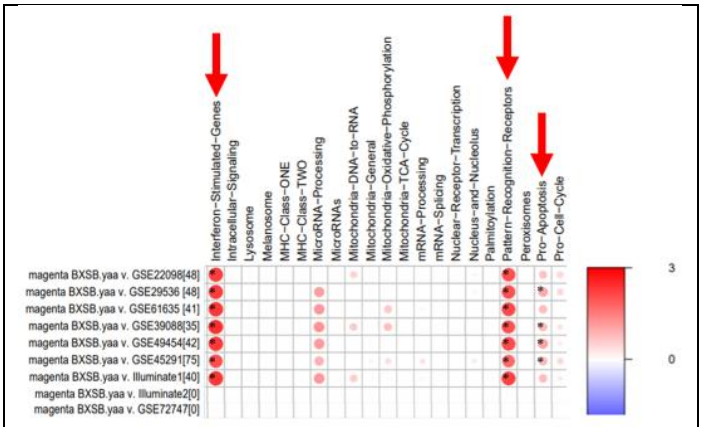


Figure 3. Functional enrichment heatmap of preserved modules: The enrichment overlaps of genes between the BXSB.yaa magenta module and its preserved human modules. Biological functions are columns. The rows are for the overlapping preserved genes between one mouse module and each human dataset. The arrows identify three signatures of interest, interferon, pattern recognition receptor and pro-apoptosis. The scale is odds ratio.

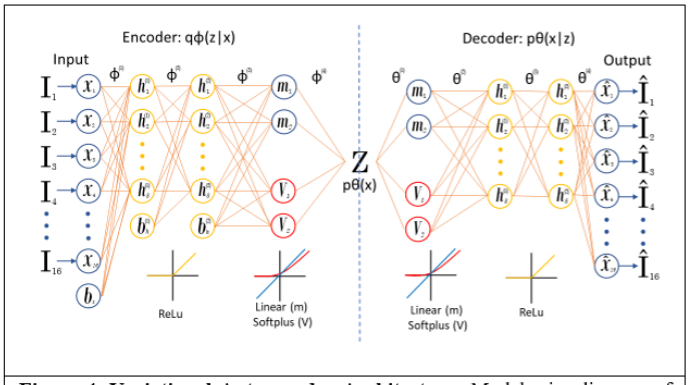


Figure 4. Variational Autoencoder Architecture: Model wire diagram of encoding (the posterior) clinical traits into a mean and standard deviation, reparametrized to Z (the prior) over a Gaussian mixture which is then reconstructed (the likelihood) back to the input. The activation function for the yellow hidden layers use ReLu and the hidden layers for the mean and standard deviation use the linear and Softplus activation functions, respectively.

Based on the significant enrichment of the anti-proliferation, pro-apoptosis and reactive oxygen species protection pathways between a BXSB.yaa module and 8 of the human datasets, this mouse seems uniquely to model the dysregulation of cell cycle in human lupus. This result aligns with the current state of literature (Otani et al., 2020). The significant enrichment of the plasmacytoid dendritic cell over three human datasets indicates the BXSB.yaa strain uniquely models the overexpression of this interferon producing cells in human patients. This result aligns with literature mirroring the analysis regarding the *Yaa* locus and interferon alpha tuning germinal center B cell selection (Lesser et al., 2020). The NZB/W strain was characterized to uniquely model the innate immune response as a myeloid signature was enriched over three human datasets. This novel result calls for experimental validation as it has not been published in literature. The MRL/lpr mouse highlighted functional enrichment of transcription factors, mRNA processing, mRNA splicing and chromatin remodeling in four human datasets, indicating that this strain appropriately model's transcription regulation. This results calls for further investigation as to the specific cell types enriched, specifically splenocytes, B cells and T cells (Liu et al., 2006). Module preservation indicated that all three strains model inflammatory signaling, cell cycle dysregulation and activated adaptive immune response, as expected (Gómez-Guzmán et al., 2014; Santiago-Raber et al., 2008; Virdis et al., 2015). The result summary highlighted the signatures we saw from the differential expression analysis, excluding the autophagy and proteasome enrichment the MRL/lpr mouse modeled. The inconsistencies in results can likely be explained by the modules constructed to include genes irrelevant to the disease state.

Unsupervised clustering of lupus patients characterized by clinical traits

The model was asked to characterized five clusters from the GSE88884 dataset. Five clusters were chosen to represent a tied gaussian mixture model as that combination of mixture type and components minimized the Bayesian Information Criteria, a method structured to identify the true model amongst several candidates.

The model identified five distinct groups- purple (430 patients), blue (52 patients), green (125 patients), orange (130 patients) and crimson (466 patients). The proportions of the patients with a certain clinical manifestation with each cluster are summarized in table _. Notably, double stranded DNA antibodies, a key biomarker in lupus diagnostics, are identified in only 10% of the purple cluster, and in 96% of the crimson cluster (Wu et al., 2017). Further, increased complement c3, which promotes inflammation, was identified in only 7% of the purple cluster and 92% of the crimson cluster. We labeled lupus patients as either in the active group (crimson) or inactive (purple) group. Unsupervised clustering serves as a novel tool to deconstruct patient heterogeneity by means of clinical characteristics.

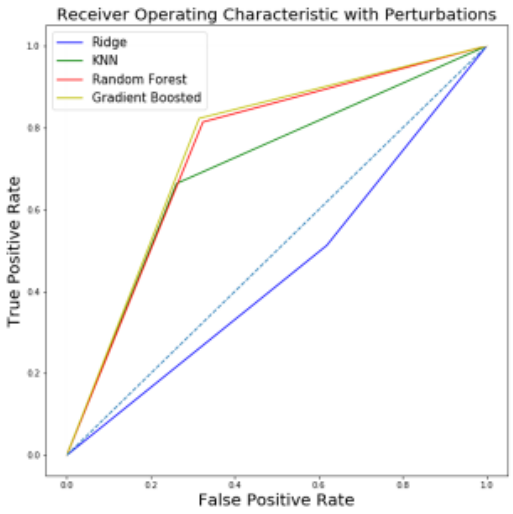
Transcript expression validates unsupervised clusters in supervised machine learning approaches

To validate the clusters created by clinical traits, we employed supervised machine learning models to classify patients labeled to the active and inactive groups (crimson and purple) using their gene expression as inputs. Two ensemble methods, one non-parametric model and one linear regularization model were used. Model performance was assessed by area under curve (AUC), sensitivity, the proportion of true positives to all true positives and false negatives, and specificity, the proportion to true negatives over all true negatives and false positives. The results are averaged over 10-fold cross validations. The non-parametric model (K-Nearest Neighbor) performed the worst of the four models, resulting in an AUC of 0.70, a sensitivity of 0.66 and specificity of 0.74. The ensemble methods (Gradient Boosted and Random Forest) were nearly identical as both classified the active and inactive groups with an AUC of 0.74, a sensitivity of 0.73 and a specificity of 0.75. Lastly, the Ridge model was the most successful classifier, returning and AUC of 0.80, a sensitivity of 0.78 and a specificity of 0.811.

However, when random gaussian perturbations were added to the gene expression data, and used to classify patients, we found that the Ridge model fails completely, and performs similarly to guessing. The ensemble methods maintain a similar performance, but sacrifice sensitivity for specificity. The Random Forest model performs best with an AUC of 0.77, a sensitivity of 0.84 and a specificity of 0.7, concluding the ensemble methods prove most useful to link genetic expression to clinical traits after clusters have been established. The results are summarized by the Receiver Operating Characteristic Curves in figure 5.

VAE Cluster	AA	EA	NAT	ALOPECIA	ANTI DSDNA	ARTHRITIS	LEUKOPENIA	LOW COMPLEMENT	ULCERS	PLEURISY	RASH	VASCULITIS
Purple	0.14	0.75	0.11	0.65	0.10	0.99	0.03	0.07	0.39	0.06	0.75	0.02
Blue	0.17	0.75	0.08	0.60	0.65	0.92	0.10	0.19	0.29	0.10	0.79	0.12
Green	0.12	0.72	0.16	0.50	0.74	0.90	0.11	0.37	0.34	0.03	0.62	0.11
Orange	0.12	0.68	0.20	0.73	0.69	0.95	0.05	0.29	0.28	0.05	0.74	0.05
Crimson	0.09	0.72	0.19	0.56	0.96	0.74	0.11	0.92	0.34	0.07	0.69	0.13

Table 2. VAE Cluster Distribution: Summarized in the table are the proportions of the patients in each cluster with the particular clinical traits of each cluster. Twelve of the final sixteen traits are listed here. Abbreviations include AA -African American, EA – European American, NAT- Native American, ANTI DSDNA – Anti-double stranded DNA.



Model Type	AUC	Sensitivity	Specificity
Gradient Boosted	0.737	0.817	0.658
Random Forest	**0.767**	**0.838**	**0.695**
K-Nearest Neighbor	0.701	0.663	0.739
Ridge	0.481	0.511	0.450

Figure 5. Receiver Operating Characteristic Curves of gene expression perturbed with Gaussian noise: The X axis is the false positive rate, the Y axis the true positive rate. Performance for Ridge, KNN, RF, and GB are summarized in blue, green, red, and yellow, respectively. The table summarizes the AUC, sensitivity and specificity of each model.

Discussion

By using gene expression to find the signaling pathways and cell types of each preclinical lupus mouse model that are significantly enriched and preserved in humans, we aim to construct an enhanced view of the levels of immunological characteristic expression.

We can conclude that WGCNA serves as a more robust approach than DE for understanding interspecies relationships on the transcriptomic level. Further, WGCNA and module preservation revealed the specific gene modules in mice that regulate immune pathways and model human disease. We also learned that unsupervised clustering serves as a novel tool to deconstruct patient heterogeneity by means of clinical expression and that ensemble methods prove most useful to link genetic expression to clinical traits after clusters had been established.

Differentiating the strength of shared characteristics, such as immune dysregulation, will be useful in planning drug trials tailored to improving the condition of a particular clinical manifestation. Progressing the current understanding of genetic associations between mouse models and human lupus patients will lead to the reality of personalized medicine and eventually improve clinical practices. Considering the overlap in immunological characteristics, mouse models are usually selected for preclinical trials based on literature reviews and recommendations from peers. This project fundamentally shifts the approach in selecting mouse models by creating a streamlined guide to select a preclinical mouse for trials.

The novel methodology lies within the statistical filtering pipeline designed to reach said goal. Although the statistical tools to determine differential expression and modules are established in immunology, the scale of the application is novel. Comparing differential expression data between two tissue types between three strains of mice is original, and extending those modules to find similarities in signaling pathways between humans and mice is unprecedented. Further, by applying an orthogonal approach to ask how genes interact with each dataset provides a necessary validation and improvement of differential expression analysis. Gaussian mixture modelling with variational auto-encoders, a deep unsupervised clustering model, were developed recently and given their complex nature, have not reached applications in lupus. This deep model is advantageous over other unsupervised clustering algorithms due to its ability to reduce the importance of noisy data when structuring the observations (Dilokthanakul et al., 2017). The model is also advantageous over more user-dependent approaches, such as a literature-based review of the module's contents, as it may uncover patterns undiscernible under human scrutiny.

By applying the most effective models currently established as well as a few novel categorization models not yet tested in determining activity state, this project improved upon prediction of disease activity status in unrelated gene expression data sets by employing supervised model to link clinical traits and genetic expression. A complete literature review on the state of ML applications in SLE is summarized in Table huh, as to indicate to novelty of this approach.

Treatment will improve for patients in the future, given more drugs pass FDA standards. If drugs are approved by the FDA for lupus or trials show positive results for drugs repositioned into lupus from other conditions, mouse models may indicate which disease manifestations of lupus are ideally treated by a certain drug in people. As there is no current way to prescribe drugs on a patient by patient basis, our unsupervised clustering deconstructs heterogeneity in lupus patients. Although not tailored to the individual yet, this may be the next step in achieving personalized medicine in the field. This research could be used to allow clinicians to make decisions for patients on a genomic level. Those treatments can then be applied at the optimal time by administering the treatment before or at onset of SLE flairs. Therefore, the supervised model designed to distinguish active and inactive states of lupus could then be applied in real time to indicate flares and recommend action.

End Matter

All data sets used in these analyses may be downloaded from GEO using the accession numbers provided in the methods.

Author Contributions and Notes

C.E.P, A.C.G. and P.E.L. designed research, C.E.P., B.C., P.B., performed research, C.E.P and B.C. wrote software, C.E.P., A.C.G., P.B., P.E.L., analyzed data; and C.E.P wrote the paper.

The authors declare no conflict of interest.

This article contains supporting information online.

Acknowledgments

1. I will acknowledge all non-listed authors who supplied the data.

References

Andrews, B. S., Eisenberg, R. A., Theofilopoulos, A. N., Izui, S., Wilson, C. B., Meconahey, P. J., … Dixon, F. J. (n.d.). *SPONTANEOUS MURINE LUPUS-LIKE SYNDROMES Clinical and Immunopathological Manifestations in Several Strains**. Retrieved from https://rupress.org/jem/article-pdf/148/5/1198/478778/1198.pdf

Bubier, J. A., Sproule, T. J., Foreman, O., Spolski, R., Shaffer, D. J., Morse, H. C., … Roopenian, D. C. (2009). A critical role for IL-21 receptor signaling in the pathogenesis of systemic lupus erythematosus in BXSB-Yaa mice. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(5), 1518–1523. https://doi.org/10.1073/pnas.0807309106

Chan, O., & Shlomchik, M. J. (1998). A new role for B cells in systemic autoimmunity: B cells promote spontaneous T cell activation in MRL-lpr/lpr mice. *Journal of Immunology (Baltimore, Md. : 1950)*, *160*(1), 51–59. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9551955

Dilokthanakul, N., Mediano, P. A. M., Garnelo, M., Lee, M. C. H., Salimbeni, H., Arulkumaran, K., & Shanahan, M. (n.d.). *DEEP UNSUPERVISED CLUSTERING WITH GAUSSIAN MIXTURE VARIATIONAL AUTOENCODERS*.

Domeier, P. P., Schell, S. L., & Rahman, Z. S. M. (2017, January 2). Spontaneous germinal centers and autoimmunity. *Autoimmunity*, Vol. 50, pp. 4–18. https://doi.org/10.1080/08916934.2017.1280671

Gómez-Guzmán, M., Jiménez, R., Romero, M., Sánchez, M., Zarzuelo, M. J., Gómez-Morales, M., … Duarte, J. (2014). Chronic hydroxychloroquine improves endothelial dysfunction and protects kidney in a mouse model of systemic lupus erythematosus. *Hypertension*, *64*(2), 330–337. https://doi.org/10.1161/HYPERTENSIONAHA.114.03587

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

Herrada, A. A., Escobedo, N., Iruretagoyena, M., Valenzuela, R. A., Burgos, P. I., Cuitino, L., & Llanos, C. (2019). Innate immune cells' contribution to systemic lupus erythematosus. *Frontiers in Immunology*, Vol. 10. https://doi.org/10.3389/fimmu.2019.00772

Hutloff, A., Büchner, K., Reiter, K., Baelde, H. J., Odendahl, M., Jacobi, A., … Kroczek, R. A. (2004). Involvement of inducible costimulator in the exaggerated memory B cell and plasma cell generation in systemic lupus erythematosus. *Arthritis & Rheumatism*, *50*(10), 3211–3220. https://doi.org/10.1002/art.20519

Iii, H. C. M., Lipsky, P. E., Leonard, W. J., Shaffer, D. J., Akilesh, S., Roopenian, D. C., … Ettinger, R. (2004). Inducer of Blimp-1 and Bcl-6 Plasma Cell Generation by IL-21, a Novel Regulation of B Cell Differentiation and. *J Immunol*, *173*, 5361–5371. https://doi.org/10.4049/jimmunol.173.9.5361

Lesser, M., Davidson Ranjit Sahu, A., Ricketts, P.-G., Akerman, M., Ioana Moisini, T., Huang, W., & Bethunaickan, R. (2020). Systemic Lupus Erythematosus Germinal Center B Cell Selection in Murine Fine-Tune α Locus and IFN-Yaa The. *J Immunol*, *189*, 4305–4312. https://doi.org/10.4049/jimmunol.1200745

Liu, J., Karypis, G., Hippen, K. L., Vegoe, A. L., Ruiz, P., Gilkeson, G. S., & Behrens, T. W. (2006). Genomic view of systemic autoimmunity in MRLlpr mice. *Genes and Immunity*, *7*(2), 156–168. https://doi.org/10.1038/sj.gene.6364286

LOW DENSITY GRANULOCYTES (LDG) IN THE PATHOGENESIS OF SYSTEMIC LUPUS ERYTHEMATOSUS (SLE): TARGETS FOR THERAPY? (n.d.). Retrieved from https://www.huidziekten.nl/zakboek/dermatosen/mtxt/malar-rash.htm,

Nikpour, M., Dempsey, A. A., Urowitz, M. B., Gladman, D. D., & Barnes, D. A. (n.d.). *Association of a gene expression profile from whole blood with disease activity in systemic lupus erythaematosus*. https://doi.org/10.1136/ard.2007.074765

Otani, Y., Ichii, O., Masum, M. A., Kimura, J., Nakamura, T., Hosny, Y., … Kon, Y. (n.d.). *BXSB/MpJ-Yaa mouse model of systemic autoimmune disease shows increased apoptotic germ cells in stage XII of the seminiferous epithelial cycle*. https://doi.org/10.1007/s00441-020-03190-0

Pierdominici, M., Vomero, M., Policlinico, U. I., Roma, D., Cristiana, B., & Maselli, A. (2012). Role of autophagy in immunity and autoimmunity, with a special focus on systemic lupus erythematosus. *Article in The FASEB Journal*. https://doi.org/10.1096/fj.11-194175

RuiShu/vae-clustering: Unsupervised clustering with (Gaussian mixture) VAEs. (n.d.). Retrieved April 30, 2020, from https://github.com/RuiShu/vae-clustering

Santiago-Raber, M.-L., Kikuchi, S., Borel, P., Uematsu, S., Akira, S., Kotzin, B. L., & Izui, S. (2008). Evidence for genes in addition to Tlr7 in the Yaa translocation linked with acceleration of systemic lupus erythematosus. *Journal of Immunology (Baltimore, Md. : 1950)*, *181*(2), 1556–1562. https://doi.org/10.4049/jimmunol.181.2.1556

Terzoglou, A. G., Routsias, J. G., Moutsopoulos, H. M., & Tzioufas, A. G. (2004). *Autoantibodies and autoantigens 1 Differential antibody recognition of the 349-364aa B-cell epitope of human La/SSB protein and its phosphorylated analogue*. https://doi.org/10.1186/ar1043

Theofilopoulos, A. N., & Dixon, F. J. (1985). Murine Models of Systemic Lupus Erythematosus. *Advances in Immunology*, *37*(C), 269–390. https://doi.org/10.1016/S0065-2776(08)60342-9

Virdis, A., Tani, C., Duranti, E., Vagnani, S., Carli, L., Kühl, A. A., … Mosca, M. (2015). Early treatment with hydroxychloroquine prevents the development of endothelial dysfunction in a murine model of systemic lupus erythematosus. *Arthritis Research & Therapy*, *17*, 277. https://doi.org/10.1186/s13075-015-0790-3

Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing*, *0*(212669), 80–95. https://doi.org/10.1142/9789813235533_0008

Wu, H., Zeng, J., Yin, J., Peng, Q., Zhao, M., & Lu, Q. (2017, April 1). Organ-specific biomarkers in lupus. *Autoimmunity Reviews*, Vol. 16, pp. 391–397. https://doi.org/10.1016/j.autrev.2017.02.011

Yan, S. X., Deng, X. M., Wang, Q. T., Sun, X. J., & Wei, W. (2015). Prednisone treatment inhibits the differentiation of B lymphocytes into plasma cells in MRL/MpSlac-lpr mice. *Acta Pharmacologica Sinica*, *36*(11), 1367–1376. https://doi.org/10.1038/aps.2015.76

Data Set	Species	Comparison	HTP Sequencing Method	Chip
GSE22908	Human	Healthy Control to SLE	Microarray	Illumina HumanHT-12 V3.0 expression bead chip
GSE29536	Human	Healthy Control to SLE	Microarray	Illumina human-6 v2.0 expression bead chip
GSE39088	Human	Healthy Control to SLE	Microarray	HG-U133_plus_2
GSE45291	Human	Healthy Control to SLE	Microarray	HT_HG-U133_Plus_PM
GSE49454	Human	Healthy Control to SLE	Microarray	Illumina HumanHT-12 V4.0 expression beadchip
GSE61635	Human	Healthy Control to SLE	Microarray	HG-U133_Plus_2
GSE72747	Human	SLE with suppressive to SLE	Microarray	Illumina HiScanSQ
GSE88884	Human	Healthy Control to SLE	Microarray	HTA-2_0] Affymetrix Human Transcriptome Array 2.0
GSE94750	Mouse	Lupus mouse +/- Bortezomib	Microarray	Agilent-028005 SurePrint G3 Mouse GE 8x60K Microarra
NZB/W	Mouse	Lupus mouse +/- HDAC6 inhibition	RNASeq	Ensemble.org (38.74)
BXSB.yaa	Mouse	Lupus mouse +/- immunosuppressive	RNASeq	Ensemble.org (38.74)

Supplementary Table 1. Datasets Summary: The table summarizes the dataset, species, comparison of subjects in study, high throughput sequencing method and chip used to annotate reads.