Understanding the Mechanisms Behind Multiple Autoimmune Syndrome (MAS)

Sarah Cross, Sunny Kim, and Allison Ma *

May 1, 2024

Abstract

Autoimmune diseases (ADs) are estimated to affect 12% of Americans, and pose a significant challenge to the healthcare system. To better understand the mechanisms governing AD pathogenesis, the Regulatory Element Locus Intersection (RELI) algorithm was re-implemented and employed to identify significant intersections between transcription factor binding sites (TFBS) and single nucleotide polymorphisms (SNPs) associated with a given AD; we focused, in this research paper, on Systemic Lupus Erythematosus (SLE), vitiligo, primary biliary cirrhosis (PBC), and rheumatoid arthritis (RA), although this algorithm may be applied to an arbitrary disease with genetic components. Analysis revealed key transcription factors intersecting with SLE that corroborate with and extend off existing research, namely EBNA2, HMGN1, IRF1, and NFE2; these TFs intersect significantly with SLE loci, and therefore likely play an integral role in disease pathogenesis. Both SLE and vitiligo intersect significantly with similar TFs, namely POLR2A and CEPBA, which may suggest common pathways underlying their co-occurrence; all TFs found have been implicated in autoimmunity or are part of pathways associated with autoimmunity. Cross-referencing TFs found with the genes they encode for highlights the downstream effects of TFBS alterations on gene expression, typically resulting in the downregulation of the gene in question. This study underscores the importance of understanding transcriptional regulation in ADs, providing insight into why specific genes are downregulated in those with a particular AD and highlighting potential therapeutic targets.

1 Introduction

12% of Americans are estimated to have at least one autoimmune disease (AD), and, of those, 25% are more likely to have another autoimmune disease (AD). It is estimated that over $168 billion is spent annually by patients to mitigate AD symptoms.

ADs are chronic, and often significantly impact patients for the entirety of their lives, with treatments remaining costly, invasive, and frequently uncovered or underfunded by healthcare providers [1]. Fundamental challenges in the healthcare system result in ADs being under-prioritized, as research on ADs receives less funding than "killer diseases" such as cancer (when interviewing Tiphaine Martin, an autoimmune disease expert, as part of a related project, Martin, during the interview, admitted that she had recently switched to a cancer lab) [14]. Since the 1950s, clinical literature has been unable to follow the increasing trends of chronic and autoimmune diseases, producing a healthcare crisis by which the healthcare system neglects patients with chronic diseases. It is also difficult to discern the cause of autoimmune diseases, as a combination of environmental and genetic factors can cause the onset of illness [20]. The recent uptick in those testing positive for Antinuclear Antibodies (ANAs, a common risk factor for autoimmune disease) suggests that environmental factors may better explain this sudden growth.

* sarahcross@cmu.edu minjeon2@andrew.cmu.edu ama4@andrew.cmu.edu

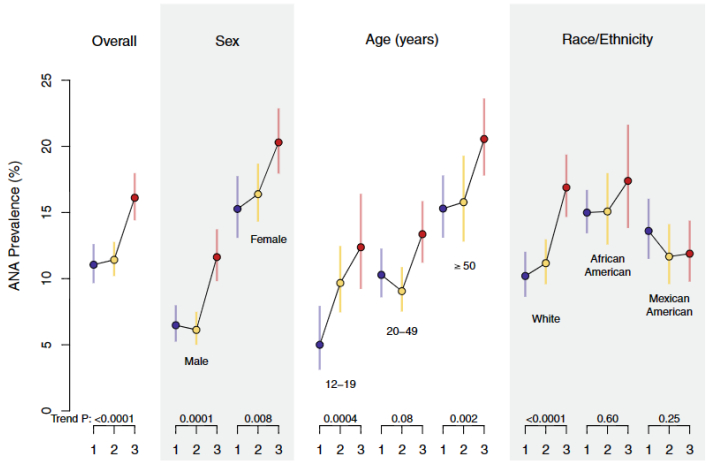

Figure 1: Increasing ANA prevalence, across three periods of time (blue for Period 1 (1988-1991), yellow for Period 2 (1999-2004), and red for Period 3 (2011-2012)) [6].

Transcription factors are proteins that turn genes "on" and "off" by binding to nearby DNA, boosting or reducing the rate at which a gene is produced. Mutations within these binding locations can significantly affect the regulatory mechanisms of gene expression. Existing research indicates that insertion mutations where a binding site is typically alter the location the transcription factor binds [13]. A single modification from one nucleotide to another can also affect the strength at which the transcription factor binds, leading to additional down or upregulation of the gene. For example, the SNP rs13239597, associated with systemic lupus erythematosus (SLE) and systemic sclerosis, lead to increased EVI1 binding and an increase in IRF5 expression [21]. Further studies of rs13239597 demonstrated that its risk A allele increased EVI1 binding and acted as an allele-specific enhancer to regulate IRF5 expression [16].

Multiple Autoimmune Syndrome (MAS) describes the occurrence of three or more ADs within a person or family. These autoimmune diseases additionally often occur in groups, named type 1, type 2, or type 3. Those with MAS are additionally highly likely to have at least one dermatological condition such as vitiligo or alopecia areata [4]. Analyzing MAS not only aids in better understanding its pathogenesis but also the pathways linking similar autoimmune diseases.

The primary approach used to highlight specific transcription factors of interest within this project is particularly useful for low-data scenarios when it is not possible to obtain enough data surrounding the diseases researched–in this case, lupus and vitiligo. There currently exists no database publicly available on patients with multiple autoimmune diseases, facilitating a need for combining multiple sources of information to better understand the interplay between genes, transcription factors, and genetic mutations.

2 Data Used

We largely used supplemental data included within the paper whose algorithm was reproduced [11]. Most ChIP-seq files were obtained from ENCODE; a full list of ChIP-seq sources and null model data is available within the supplementary dataset attached [12]. Linkage disequilibrium block data was sourced from the 1000 Genomes dataset [2].

To understand the relationship between transcription factors and diseases, gene data available through Autoimmune Diseases Explorer (ADeX) [15], and gene relationships available through GeneMANIA [22] were used.

The model, as well as the figures and data shown, are publicly available on GitHub [5].

3 Methods

3.1 Overview

We implemented the RELI algorithm, described below, and then applied it to single nucleotide polymorphisms (SNPs) correlated to SLE, vitiligo, primary biliary cirrhosis (PBC), and rheumatoid arthritis (RA).

3.2 Regulatory Element Locus Intersection (RELI) Algorithm

3.2.1 Computational Problem

Existing research indicates that SNPs in TFBS can drastically change the regulation of a given gene [16]. Using the location a TF binds to, in conjunction with a list of relevant SNPs, to find intersections between the two of them can be formalized into the following computational problem:

TFBS-SNP Intersection Problem:
Input: Locations a TF binds to in a given genome and a list of SNPs and their corresponding locations.
Output: The TF and SNP intersection significance, quantified using Relative Risk (RR), mean, standard deviation, and p-value.

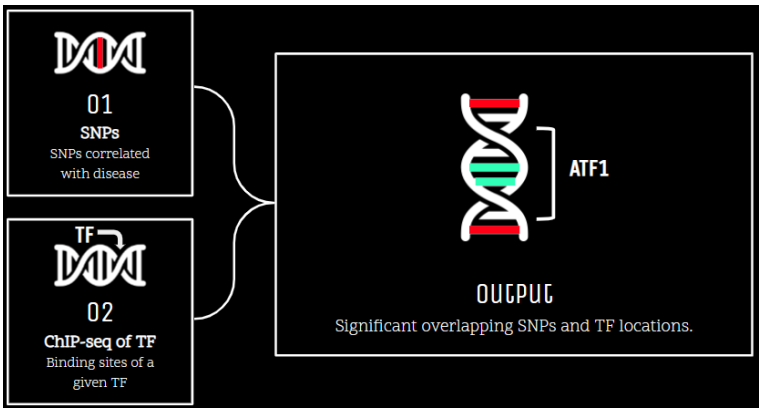

Figure 2: Summary of RELI Function.

3.2.2 Implementation

We first implemented the Regulatory Element Locus Intersection (RELI) algorithm as created by the Weirauch Lab [11]. The algorithm works in the following way:

1. **Load Data:** The following information is necessary to run RELI: [1]

   (a) The ChIP-seq index file for a particular TF. The file should contain a list of all locations a given TF binds to in 4-column Browser Extensible Data (BED) format, such as the following example:

Chromosome	Start	End	TF
chr1	10433275	10433576	EBNA1
chr1	22065630	22065931	EBNA1

Table 1: An example ChIP-seq file (sample data truncated from hg19_0001)

   (b) Null MAF Model; this is a null model containing a list of all known SNP locations within the human genome and their length. This is a list of entries spaced apart by newlines, where each entry lists a chromosome location (added up cumulatively across chromosomes) and a length separated by a tab. For example:

[1] The location of each chromosome is an additional optional field; by default, RELI uses the hg19 human genome.

```
3095334357      9
3095334364      4
3095334380      4
3095334388      2
3095334416      8
3095334418      2
3095334532      8
```

The above is a small sample of entries in the null model used by RELI within the Y chromosome.

(c) Information about linkage disequilibrium (LD) blocks; each line indicates a given reference SNP and its corresponding tab-separated correlated SNPs. The reference SNP, which is the most strongly associated variant in the set, is separated from its correlated SNPs by a colon, then a space. For example:

```
rs10516487:     rs10516487      rs1125271       rs13106926 \
rs13107572      rs13107612      rs13129744      rs13135381 \
rs13136796      rs1421627       rs17200824      rs17266594 \
rs2052445       rs2080820       rs34029191      rs34749007 \
rs35388091      rs4637409       rs55657829      rs55768089 \
rs5860695       rs66976837      rs71597109
```

The entry above indicates that the SNP with ID `rs10516487` tends to occur in conjunction with SNPs such as `rs13106926` and `rs1125271`.

2. **Calculate observed intersections:** This is done by iterating over all LD blocks and, for each LD block, determining if an intersection exists with the list of locations the transcription factor binds as supplied in the ChIP-seq dataset. If an intersection exists, the RSID of the reference SNP is added to a list.

3. **Calculate expected intersections:** The expected number of intersections is calculated by taking the reference SNP for each LD block, then calculating the distance from each SNP in the LD block with the reference, measured in bases. A null SNP with a length randomly chosen from the null MAF model and in the same chromosome is then formed and reference SNPs are generated using the distance vector for the LD block. The null SNP reference SNPs are then counted up; this is the number of expected intersections due to random chance. The RELI model allowed for additionally matching random genomic variants by their allele frequencies; however, because this requires additional data not available for the other phenotypes tested, we opted not to reimplement the SNP matching option. Push the number of random SNPs that overlap into a list.

4. **Repeat simulation:** Perform steps 3 and 4 rep_max times. In experiments, we opted for using $rep\_max = 2000$, as this appeared to be enough to determine if a statistically significant difference existed between the simulated and observed data.

5. **Calculate statistics:** Using the list of null overlapping SNPs as a null distribution, calculate the mean and standard deviation of the null distribution, and use them to determine if the difference between the number of intersections observed is statistically significant when compared to the null distribution. The mean is calculated using:

$$\mu = \frac{\sum_{x_i \in X}(x_i)}{|X|} \qquad \text{(Average Expected Intersections)}$$

Where X is the set of all expected intersections.

Standard deviation is calculated with:

$$\sigma = \sqrt{\frac{\sum_{x_i \in X}(x_i - \mu)^2}{|X|}} \qquad \text{(Std of Expected Intersections)}$$

Z-score is calculated using:

$$z = \frac{d - \mu}{\sigma} \qquad \text{(Z-score of Expected Intersections)}$$

In which d is the number of observed intersections.

The p-value is found using the cumulative distribution function (the area under the curve of the normal distribution function):

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}e^{-\frac{x-\mu}{2\sigma^2}}} \qquad \text{(Normal Distribution Function)}$$

$$F(z) = \int_{-\infty}^{x} N(z)\ \mathrm{d}z \qquad \text{(Cumulative distribution function)}$$

$$= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \qquad \text{(See [18])}$$

Where *erf* is the following function:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_{0}^{x} \exp(-t^2)\ \mathrm{d}t$$

$$p = 1 - F(z) \qquad \text{(P-Value of Observed Intersections)}$$

We implemented RELI using Python and modified it as described below in order to best fit the data it was applied to:

1. **Cross-platform:** The original paper used C++ GSL libraries specific to Unix platforms; the paper's implementation required using Windows Subsystem for Linux (WSL) to run on Windows. This implementation uses Python and standard data processing cross-platform libraries (numpy and scipy), enabling this implementation of RELI to be usable on most modern systems [7].
2. **Split loading and processing data:** When running RELI across different TFs for the same disease, all of the data is loaded individually every time. This is expensive, slow, and prevents running multiple instances at once due to memory constraints; to resolve this issue, we created a separate class containing LD blocks, ChIP-seq indexing information, chromosome information, and the null model information passed into the RELI simulation. This allows for instances running on different transcription factors to use the same set of shared information.
3. **Multi-threaded:** After sharing common data across different TFBS simulation instances, we implemented multithreading using the `multiprocessing` package [8], separating the TFBS data into N bins, then assigning each bin to a thread running in a detached state. Although this speeds up the model on machines with multiple processors, it may produce a slowdown on machines with a single-processor due to frequent context switching [17].
4. **Relative Risk:** Relative Risk (RR) is calculated by dividing the number of observed intersections by the number of expected intersections; however, if no intersections are expected, then the paper's model previously would set RR to 0. We replaced this value with an Inf value, and dealt with these points manually, as the relative risk is much *higher* than expected, rather than lower, as the value 0 captures.

3.3 Using RELI to Identify MAS TFBS

Several transcription factors and their intersections with SNP loci were calculated for SLE, vitiligo, rheumatoid arthritis (RA), and primary biliary cirrhosis (PBC) using RELI. The diseases above were chosen for the following reasons:

- There exists a large amount of data surrounding SNPs associated with each disease (¿ 1000 SNPs across the human genome have been found to be associated with each disease, all which come from high-caliber peer-reviewed papers).
- Because MAS tends to occur in clusters, two ADs from Type 2 (PBC and RA) and two ADs from Type 3 (SLE and vitiligo) were chosen to observe differences between the clusters. Finding transcription factors associated with Type 2, for example, and not Type 3, may indicate that the TF may result in the regulation of a gene in a way that is disease-promoting.

3.4 Analyzing TF Pathways

Upon finding transcription factors significantly associated with SNP disease loci, we looked to understand the interplay between transcription factor and disease through cross-referencing TFs with existing gene data available through Autoimmune Diseases Explorer (ADeX) [15] and GeneMANIA [22].

ADeX is a database consisting of a collection of datasets performing gene expression profiling on healthy patients and those with an autoimmune disease; the ADs covered within the database are limited to Sjodren's syndrome, Type 1 diabetes, lupus, and rheumatoid arthritis.
For ADeX, of the transcription factors of interest, the corresponding gene encoded by the TF was found using the "Gene Query" function, and each dataset corresponding to the disease analyzed was searched to find values for which Wilcoxon's p-value was smaller than 0.1 [19]. This was done to find how SNP modifications to a TFBS contributed to the up or downregulation of its corresponding gene.

GeneMANIA is a web platform that visualizes interactions between different genes through cross-referencing existing papers, displaying genes as nodes and the type and strength of their interaction by the color and thickness of edges. Each edge may be the following:

- **Physical interactions (red):** Two gene nodes are connected with this edge if they were found to interact in a protein-protein interaction study; that is, the genes produce proteins that regulate one another.
- **Co-expression (purple):** When, across differing conditions elicited by a gene expression study, two genes are found to have similar expression levels.
- **Pathways (blue):** Two genes are connected by this edge if they participate in the same reaction within a known pathway.
- **Genetic interactions (green):** If perturbing (modifying the expression) of one gene is found to occur as a result of the perturbation of another gene, the two are functionally associated and connected through a genetic interaction.
- **Shared protein domains (yellow):** When two genes have the same protein domain.
- **Co-localization (light blue):** This edge connects genes that are expressed within the same tissue.

The genes corresponding to the significant (later expanded to all genes of note) transcription factors identified to intersect with disease loci were placed into GeneMANIA and analyzed for interactions between each other, as well as with other genes.

4 Results

The significance of an intersection calculation was measured using p-value (calculated using the formulas described in the Methods section). A p-value < 0.05 is considered "significant," and < 0.3 to note (although a larger p-value may indicate the intersection occurred due to random chance, this, in combination with additional information such as high intersection with a similar disease, is more likely to be a correct measurement).

Quantifying the extent to which two values intersect was measured using Relative Risk (RR). Relative Risk is a ratio of the number of intersections seen divided by the number of expected intersections: ($\text{RR} = \frac{\text{number of intersections}}{\text{number of expected intersections}}$). A RR of 4.12, for example, indicates that observed intersections occurred 4.12x more often than what would be expected due to random chance.

4.1 Replicating RELI

After implementing the RELI algorithm with the changes described in the Methods section, we confirmed our results matched that produced by the original RELI implementation using three tests highlighted in the paper:

Test Comparison	RELI P-Value (Original)	RELI P-Value (Implementation)
Lupus and EBNA2	$2.69019e^{-30}$	$4.18906e^{-29}$
Lupus and BATF	0.5	0.5
Vitiligo and HSF1	0.5834720306469277	0.5834330764460361

Table 2: Comparing original RELI algorithm to its Python implementation

RELI additionally found that vitiligo was heavily related to the transcription factor FOS, with a Relative Risk of 8.14. Furthermore, we were able to see the relationship between Lupus and EBNA2 (Epstein-Barr virus), with a relative risk of 5.97; we received the same values as the original model for these values.

All calculations were within at least 0.0001 of one another, beyond which it is likely for random variations to occur due to random chance. The number of observed intersections and relative risk were identical across the two models.

4.2 Lupus (SLE) - High-Scoring TFBS Intersections

The pathway the RELI paper explored the most in-depth was the Epstein-Barr virus (EBV); contracting EBV as a child makes a patient up to 50x more likely to contract SLE later in life. The paper suggests that the reason why this likely occurs is because the EBNA2 transcription factor intersects highly with SLE loci, with 26 intersections occurring (over 5x more intersections occurring than what would be expected due to random chance). Our observations corroborate with this information, as shown below.

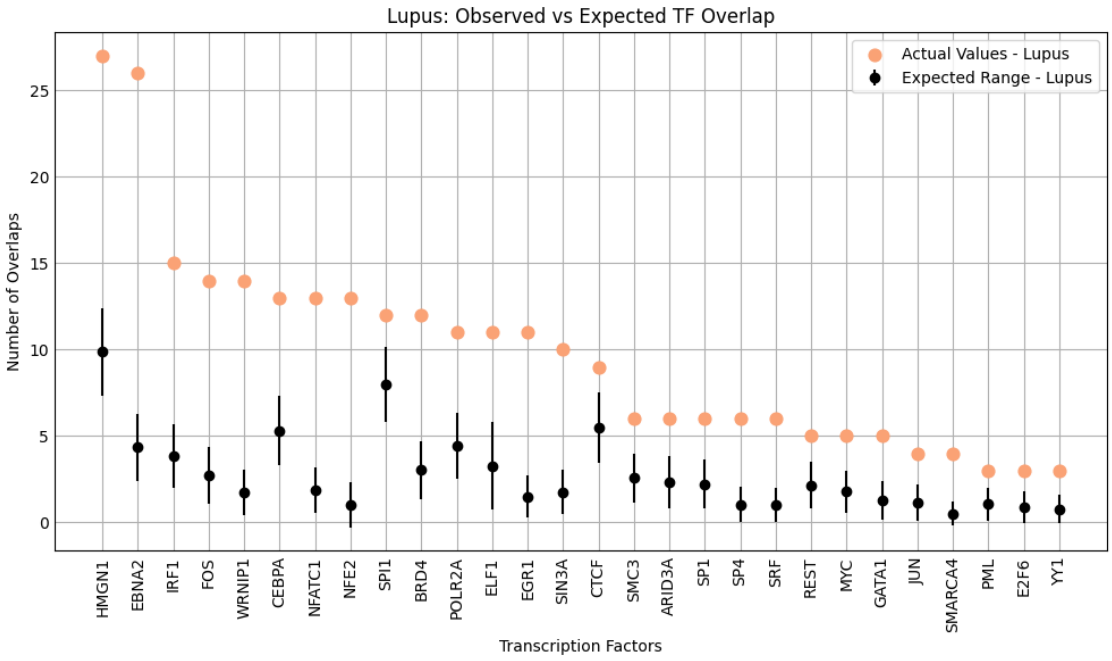

Figure 3: Significant Transcription Factors for SLE (p-value < 0.05).

It was additionally found that the TFs HMGN1, IRF1, and FOS have binding sites that significantly intersect with SNP loci. IRF1 (interferon regulatory factor) has been implicated in autoimmunity, inducing the transcription of pro-inflammatory cytokines and posed as a potential drug target for SLE [3]. The GeneMANIA output, shown below, indicates that HMGN1 holds a strong physical interaction with NFE2, which also strongly intersects with SLE loci. NFE2 has also been found to regulate interferon receptor expression, altering the regular function of macrophages in patients with lupus [9]. Interferon regulatory factor genes regulate the transcription of type-1 interferons (IFNs),

which promote inflammation, typically in response to a virus or pathogen. Within ADs, inflammation tends to be regulated incorrectly and occurs more often than necessary to fight off a perceived pathogen.

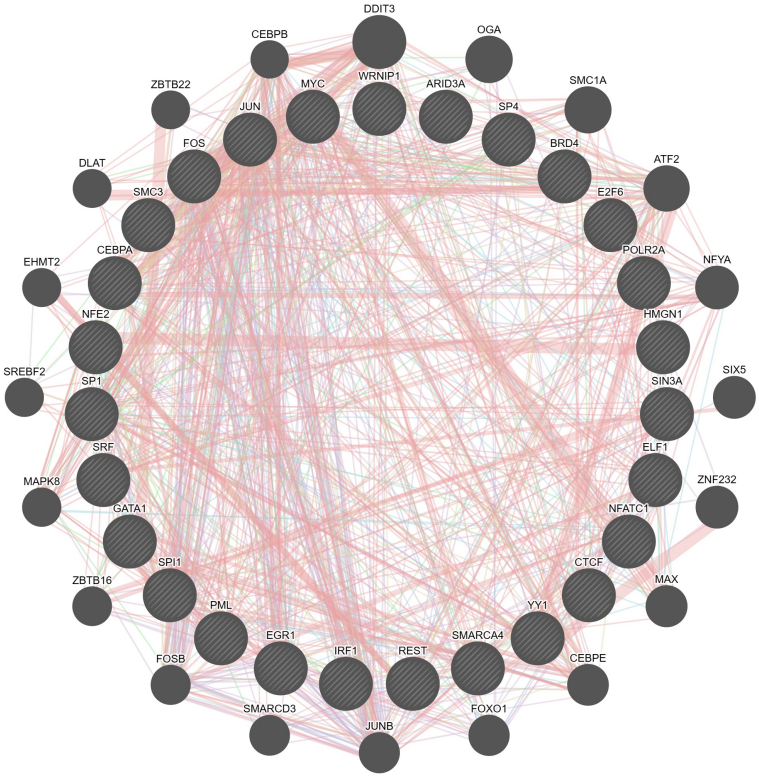

Figure 4: Strongly Connected SLE Gene Network (p-value < 0.05).

Within SLE patients, SNPs occurring within these TFBS locations alters the location the transcription factor binds, and therefore less of the intended gene is produced, resulting in lower gene expression. For the transcription factors highlighted, IRF1 and NFE2, the corresponding genes are likely down-regulated, resulting in inflammatory IFNs being less regulated and therefore promoting inflammation. Preliminary analysis of gene expression data for IRF1 from ADeX corroborates with this theorized interaction, with IRF1 being downregulated in patients with SLE:

Figure 5: IRF Gene in SLE vs Healthy Patients

4.3 Lupus (SLE) and Vitiligo

Comparing TFBS with lupus loci to vitiligo loci yields further insights into what makes these diseases tend to cluster together.

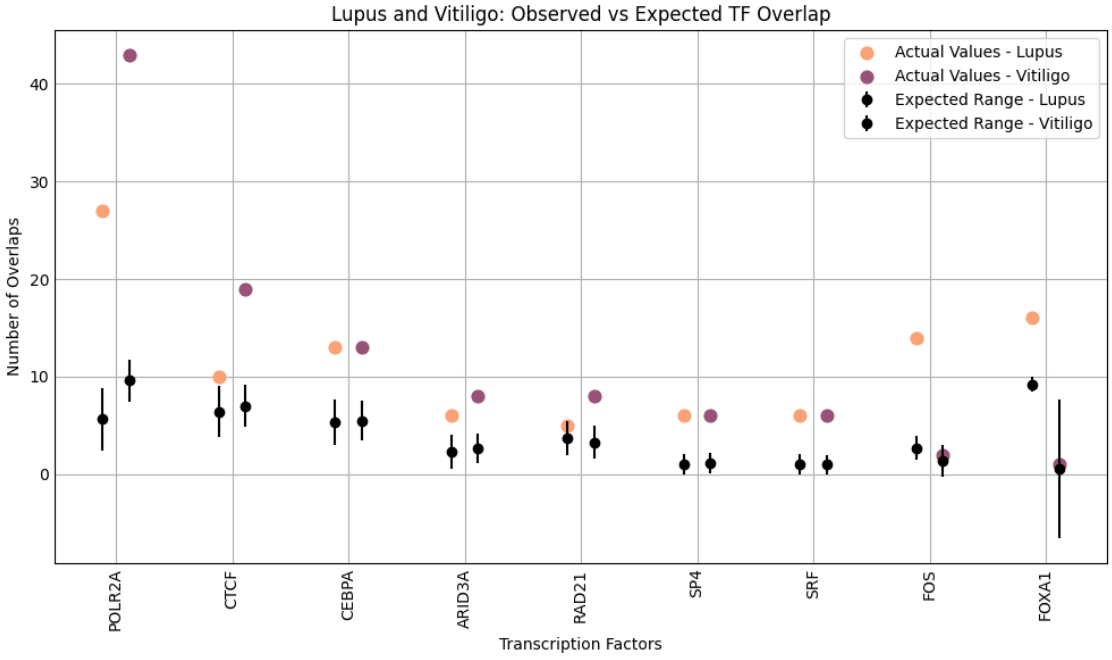

Figure 6: Transcription Factors of Note for SLE and Vitiligo (p-value < 0.3).

TFs	Overlaps (Vitiligo)	Overlaps (Lupus)	P Value (Vitiligo)	P Value (Lupus)
POLR2A - hg19_1540	43.0	27.0	5.255647e-26	7.738376e-23
CTCF - hg19_0233	19.0	10.0	2.338125e-06	4.743428e-02
CEBPA - hg19_0134	13.0	13.0	5.597509e-04	8.027643e-05
ARID3A - hg19_0021	8.0	6.0	1.159125e-03	7.860840e-03
RAD21 - hg19_1037	8.0	5.0	4.252969e-03	2.230640e-01
SP4 - hg19_1245	6.0	6.0	4.384681e-06	4.738247e-07
SRF - hg19_1276	6.0	6.0	3.430541e-07	3.107924e-07
FOS - hg19_1543	2.0	14.0	2.952545e-01	5.207750e-12
FOXA1 - hg19_0481	1.0	16.0	2.712008e-01	1.697701e-01

Figure 7: Transcription Factors of Note for SLE and Vitiligo (p-value < 0.3).

Figure 6 indicates transcription factors of note for both SLE and vitiligo; it was found that both diseases shared several strong transcription factor intersections in common, namely POLR2A and CEBPA. The GeneMANIA graph network, shown below, indicates that CEBPA is co-expressed with both POLR2A and ARID3A; CEBPA plays an integral role in differentiating immature granulocytes, which are found in much higher levels within lupus and vitiligo patients. This indicates that the downregulation of CEBPA likely results in granulocytes being incompletely differentiated and left in an immature state.

As shown in the graph, there are several other transcription factors also overlaps with SLE and vitiligo, such as CTCF, FOS, and SP4; these all have been found to be implicated in autoimmunity.

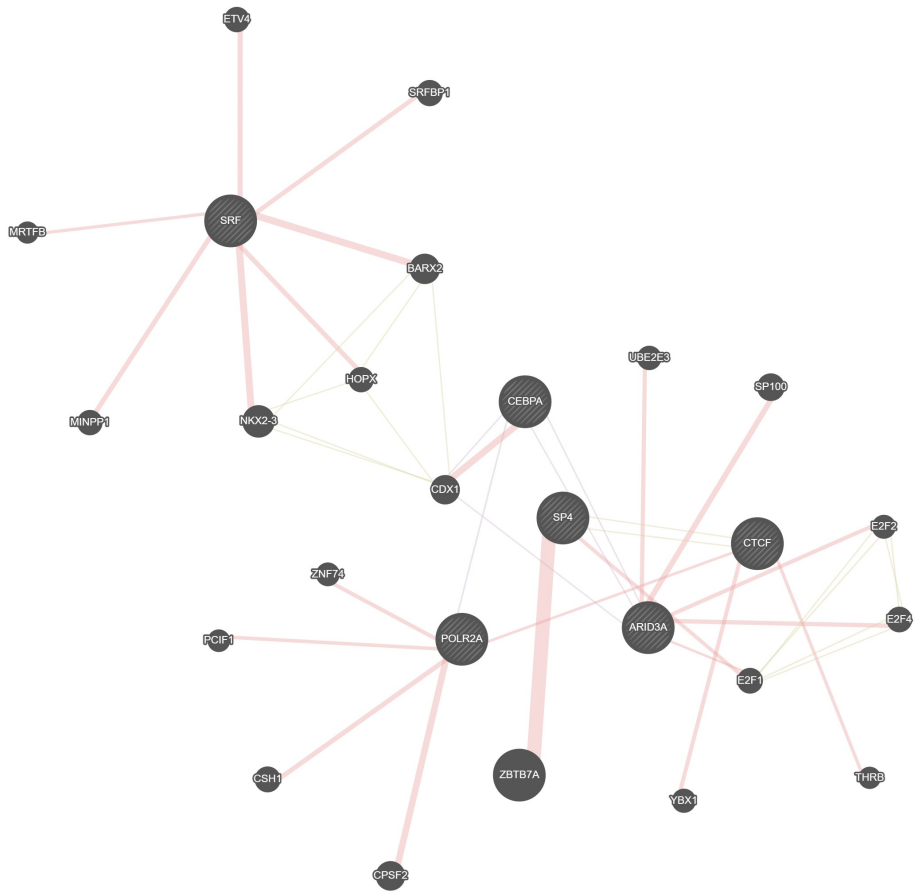

Figure 8: GeneMANIA Network for Vitiligo and SLE Genes

4.4 Lupus: TFBS Gene Expression Analysis

Performing a RELI analysis identified six transcription factors that very significantly intersected SNPs associated with lupus ($p < 0.01$): CTCF, CEBPA, ARID3A, SP4, SRF, and HMGN1. Research indicates that mutations in TFBS alters the binding site and therefore reduces expression of the gene encoded by the particular TF [13]; these findings were confirmed for these transcription factors, with every gene with a p-value below 0.1 having lower than normal expression in those with SLE.

Figure 9: CTCF for Lupus

The graph above shows the distribution of the prevalence of the transcription factor CTCF. For patients with lupus, the median value was under 9.5, whereas regular patients had a median higher than 9.5. After testing the significance of this difference, the p-value was very small and well under 0.05, which indicates that this difference is pronounced and may display a potential relationship.

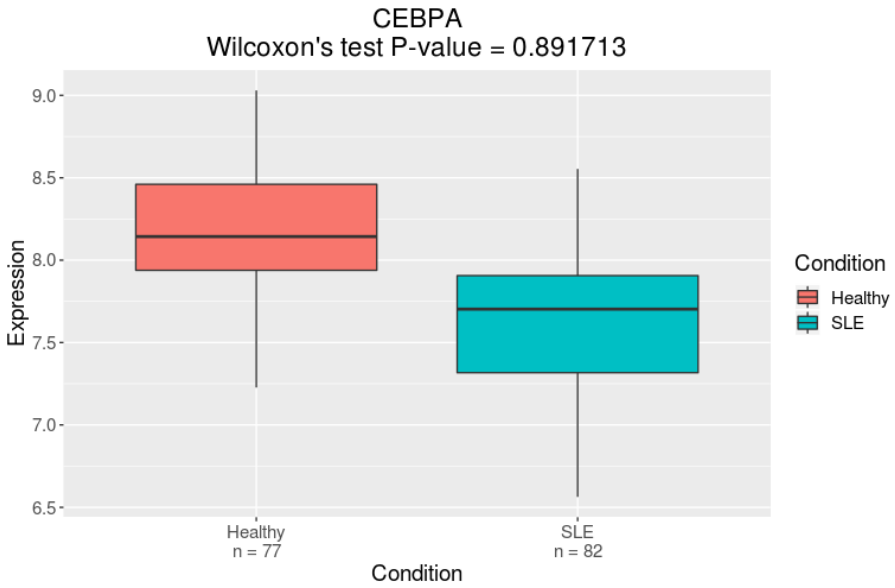

Figure 10: CEBPA for Lupus

For CEBPA, we found that the median of prevalence for lupus patients was less than 7.75, while the median for patients without lupus was greater than 8. While this may seem like a significant difference, the p-value was pretty large at just under 1, indicating that there is no significant relationship between this transcription factor and lupus.

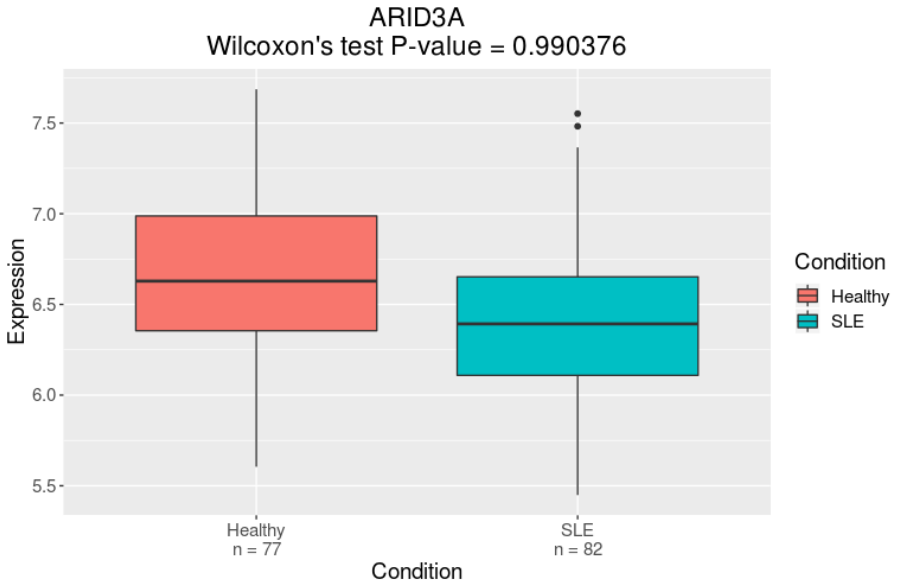

Figure 11: ARID3A for Lupus

For ARID3A, we found around a 0.25 difference between the median of prevalence for patients with lupus and the median for patients without lupus, where the median for healthy patients was greater. However, the p-value of this difference was pretty large at just under 1, indicating that this difference is not statistically significant and thus there is no significant relationship between ARID3A and lupus.

Figure 12: SP4 for Lupus

For SP4, we found around a 0.25 difference between the median of prevalence for patients with lupus and the median for patients without lupus, where the median for patients with lupus was greater. The p-value of this difference was 0.11, indicating that this difference is statistically significant and thus there is a significant relationship between SP4 and lupus.

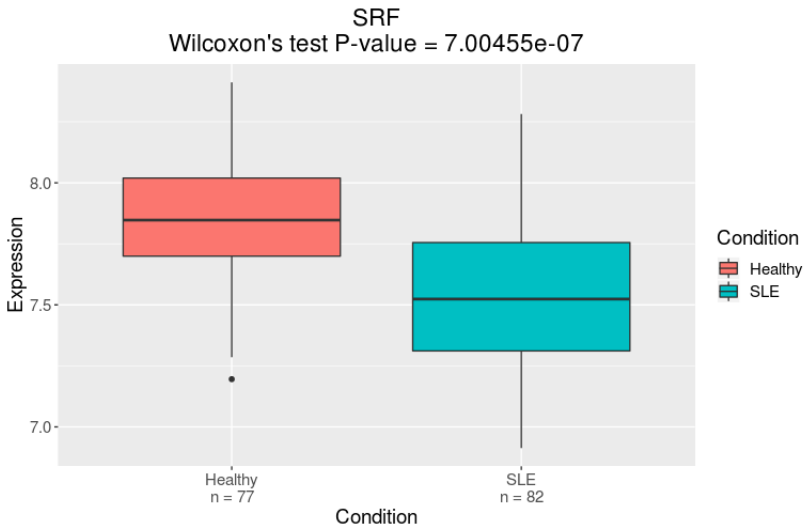

Figure 13: SRF for Lupus

For SRF, we found around a 0.325 difference between the median of prevalence for patients with lupus and the median for patients without lupus, where the median for healthy patients was greater. The p-value of this difference was very small, indicating that this difference is statistically significant and thus there is a significant relationship between SRF and lupus.

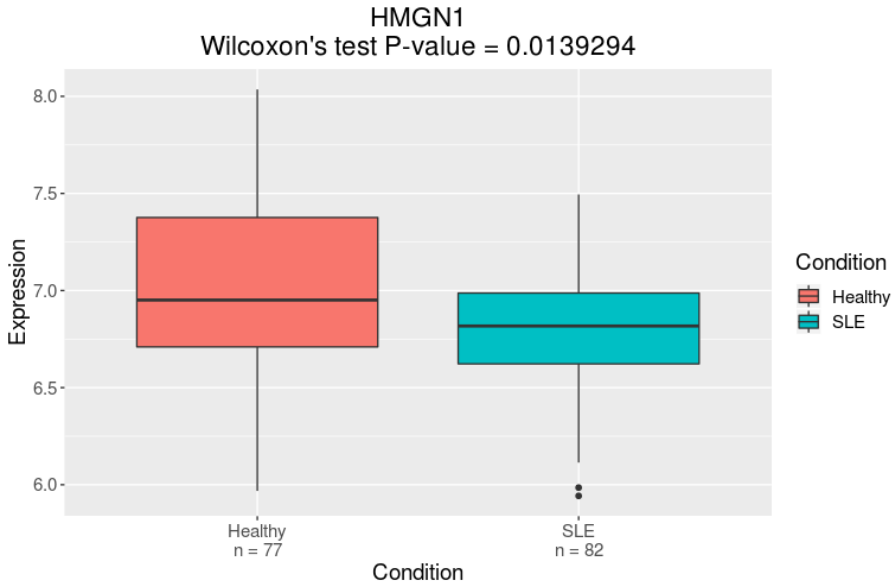

Figure 14: HMGN1 for Lupus

For HMGN1, we found less than a 0.25 difference between the median of prevalence for patients with lupus and the median for patients without lupus, where the median for healthy patients was greater. The distribution of prevalence for healthy patients, however, was much greater, with the data being skewed towards the right. The p-value of the difference in medians was around 0.01, indicating that this difference is statistically significant and thus there is a significant relationship between HMGN1 and lupus.

4.5 Lupus: Gene Network Analysis

When the genes that these transcription factors code for were analyzed together, POLR2A, CTCF, SP4, ARID3A, and CEBPA were found to be related to each other. SRF was only loosely related to other significant genes through a physical interaction with genes that share a protein domain with POLR2A.

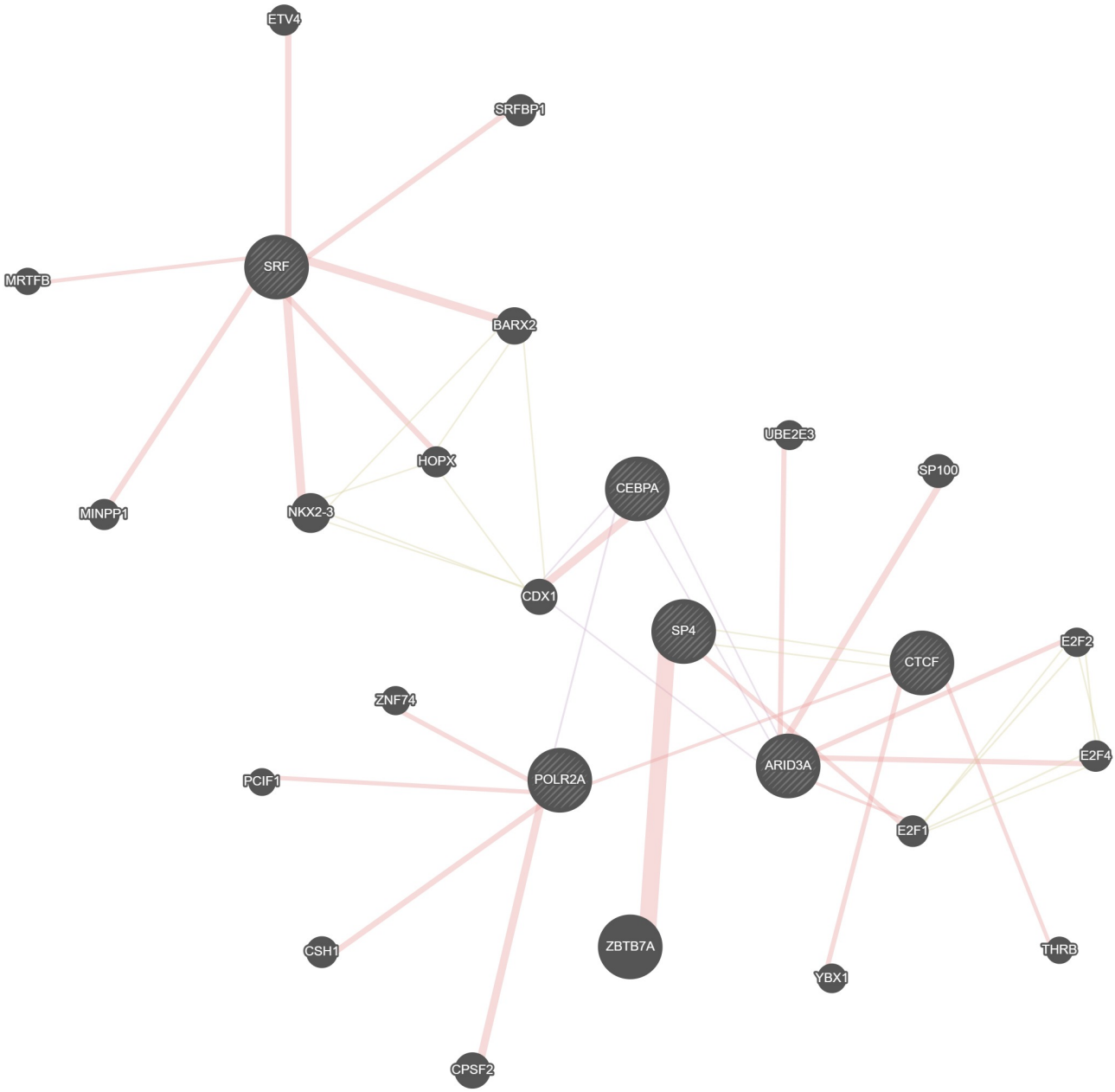

Figure 15: Transcription Factor Gene Network

Each of the transcription factor genes are connected to a network of other genes that each have their own functions. POLR2A, CEBPA, and ARID3A are connected through co-expression. POLR2A is also associated with CTCF through physical interactions, which is connected to SP4 through shared protein domains. These five genes form a network, and within the networks are other smaller genes. These genes connect with genes connecting to SRF through shared protein domains.
Given the fact that these corresponding transcription factors were highly significant for both lupus and vitiligo, these results are not surprising.

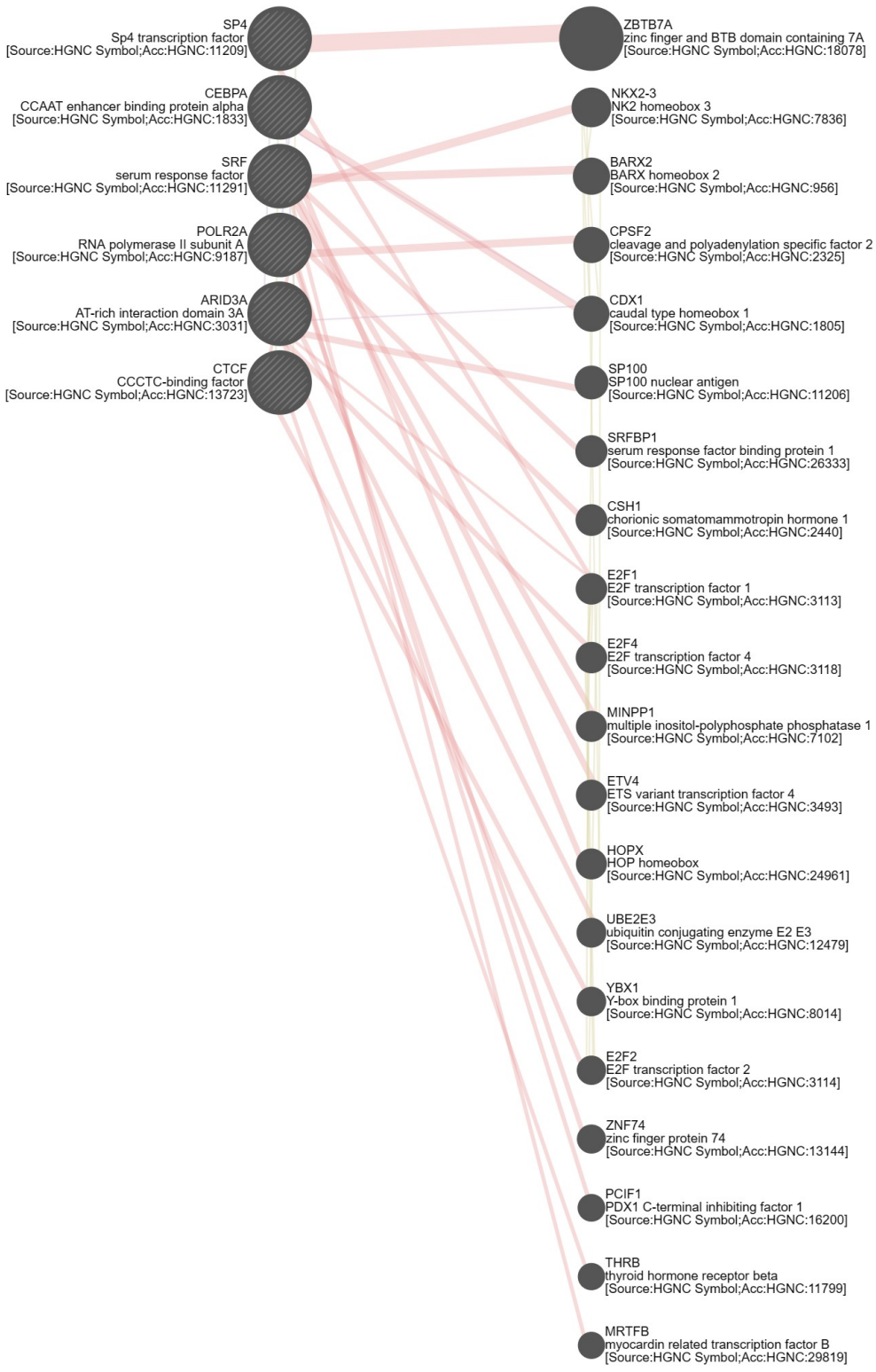

Figure 16: Transcription Factor Gene Network with Descriptions

This figure above more clearly describes each gene, especially the ones we didn't originally set out to explore but found connections to as well.

4.6 Type 2 vs Type 3 MAS

Lupus and vitiligo are both in the Type 3 for Multiple Autoimmune Syndrome. We also wanted to see if the type of MAS played a role in which TFs were significant. Two of the Type 2 diseases we looked at were Rheumatoid Arthritis (RA) and Primary Biliary Cirrhosis (PBC).

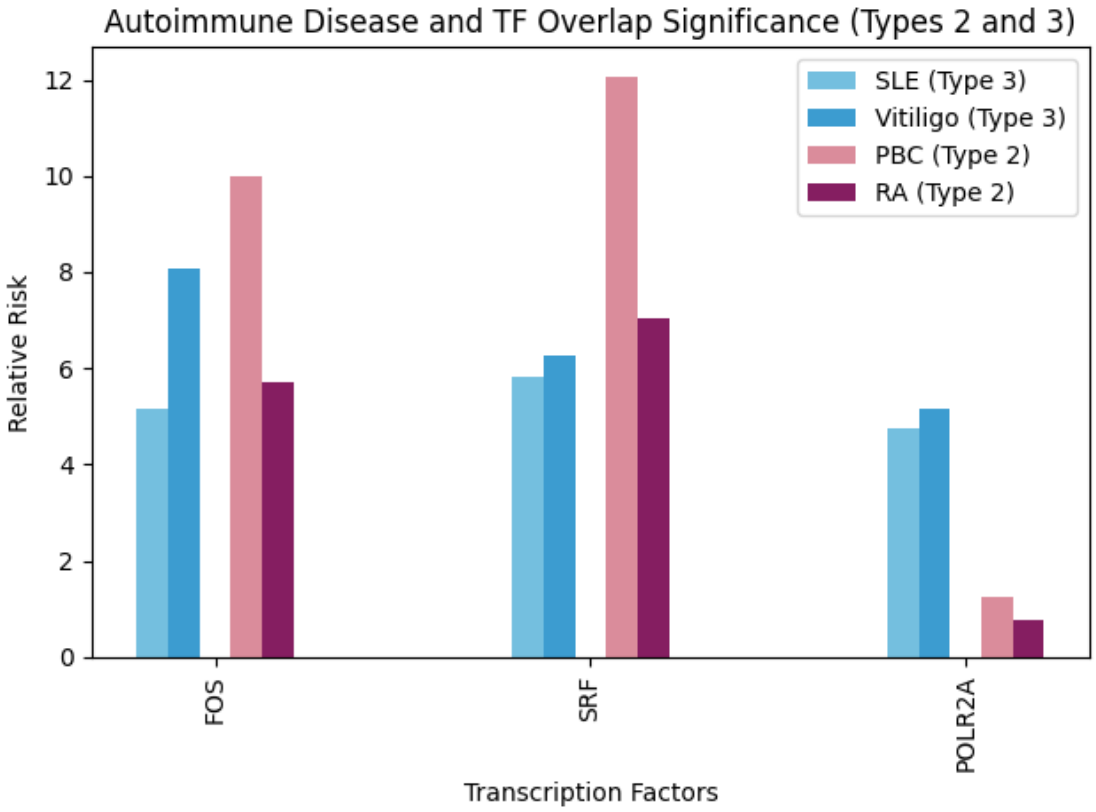

Figure 17: Type 2 and Type 3 Transcription Factors.

We found that all four autoimmune diseases had SNPs that strongly intersected with similar transcription factors. Interestingly, there were several TFs that only overlapped strongly with SLE and vitiligo, both type 3 autoimmune diseases. Among those were FOS, SRF, and POLR2A.
Interestingly enough, the type three diseases had high relative risk for POLR2A (x for lupus and y for vitiligo), but was much lower for the type two diseases (x for RA and y for PBC). POLR2A encodes for the subunit of RNA Pol II responsible for synthesizing mRNA. Although mutations in POLR2A are largely tied to neurodevelopmental disorders, it is a wide-ranging gene with mutations that have large consequences on a variety of functions within the body, including the immune system [10].
Analyzing why these transcription factors only intersected with autoimmune diseases correlating to a specific type may aid in understanding why these diseases tend to be contracted together.

5 Further Extension

There are several potential directions off which to continue this research:

- **Include TFs specific to ADs** - Several TFs, such as XBP1, have been already identified as correlated with the pathogenesis of ADs such as vitiligo; however, data on many of these specific TFs was not included in this study and may be valuable to compare to the theorized TFs highlighted here.

- **Taking into account gene upregulation** - Because SNPs in TFBS leads to downregulation within the gene correlated with the TF, this study largely focused on situations where genes were downregulated as a result of having a given AD; however, further looking into what causes upregulation may aid in creating a better picture of disease versus healthy gene expression.

References

[1] National Coalition of Autoimmune Patient Groups American Autoimmune Related Disease Association. *The Cost Burden of Autoimmune Disease: The Latest Front in the War on Healthcare Spending*. 2011. URL: http://www.diabetesed.net/page/_files/autoimmune-diseases.pdf (visited on 04/22/2024).

[2] "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489 (2012), pp. 57–74. URL: https://www.nature.com/articles/nature11247.

[3] Tatsuma Ban et al. "Genetic and chemical inhibition of IRF5 suppresses pre-existing mouse lupus-like disease". In: (). URL: https://www.nature.com/articles/s41467-021-24609-4.

[4] M Cojocaru and Inimioara Cojocaru. "Multiple Autoimmune Syndrome". In: *MaeDica* 5.2 (2010), pp. 132–134. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150011/.

[5] Sarah Cross. May 1, 2024. URL: https://github.com/html1101/RELI-Analysis.

[6] Gregg Dinse et al. "Increasing Prevalence of Antinuclear Antibodies in the United States". In: *Arthritis Rheumatology* 72(6) (2020), pp. 1026–1035. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7255943/.

[7] Python Software Foundation. URL: https://www.python.org/.

[8] Python Software Foundation. *multiprocessing Package*. URL: https://docs.python.org/3/library/multiprocessing.html.

[9] Shuhong Han, Pui Y Zhuang Haoyang Lee, and Mingjia Li. "NF-E2-Related Factor 2 Regulates Interferon Receptor Expression and Alters Macrophage Polarization in Lupus". In: *Arthritis Rheumatology* 72(10) (2020), pp. 1707–1720. URL: https://pubmed.ncbi.nlm.nih.gov/32500632/.

[10] Adam W. Hansen et al. "Germline Mutation in POLR2A: a Heterogeneous, Multi-Systemic Developmental Disorder Characterized by Transcriptional Dysregulation". In: 2 (Jan. 14, 2021). URL: https://www.cell.com/hgg-advances/fulltext/S2666-2477(20)30014-2.

[11] John B. Harley et al. "Transcription Factors Operate Across Disease Loci, with EBNA2 Implicated in Autoimmunity". In: *Nature Genetics* 50 (2018), pp. 699–707. URL: https://www.nature.com/articles/s41588-018-0102-3.

[12] John B. Harley et al. *Transcription Factors Operate Across Disease Loci, with EBNA2 Implicated in Autoimmunity - Supplementary Dataset 2*. URL: https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-018-0102-3/MediaObjects/41588_2018_102_MOESM4_ESM.xlsx.

[13] Frederick Kamanu and Yulia Medvedeva. "Mutations and Binding Sites of Human Transcription Factors". In: *Frontiers in Genetics* 3 (2012), p. 100. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3365286/.

[14] Tiphaine Martin. "Autoimmune Diseases". Interview on Autoimmune Diseases. Feb. 25, 2022.

[15] Jordi Martorell-Marugán, Raúl López-Dominguez, Adrián García-Moreno, et al. "A Comprehensive Database for Integrated Analysis of Omics Data in Autoimmune Diseases". In: *BMC Bioinformatics* 22 (2021), p. 343. URL: https://doi.org/10.1186/s12859-021-04268-4.

[16] "Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases". In: (). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8235176/.

[17] Robbert van Renesse. *Context Switching and Threads*. 2020. URL: https://www.cs.cornell.edu/courses/cs4411/2020fa/schedule/slides/week3_contextswitch_threads.pdf.

[18] Joram Soch. Mar. 20, 2020. URL: https://statproofbook.github.io/P/norm-cdf.html.

[19] Penn State Department of Statistics. URL: https://online.stat.psu.edu/stat415/book/export/html/836.

[20] "The Relation of the Chronic Disease Epidemic to the Health Care Crisis". In: *2(3)* (2020), pp. 167–173. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7077778/.

[21] Hlaing Thynn et al. "An Allele-Specific Functional SNP Associated with Two Systemic Autoimmune Diseases Modulates IRF5 Expression by Long-Range Chromatin Loop Formation". In: *J Invest Dermatology* 140(2) (2020), pp. 348–360. URL: https://pubmed.ncbi.nlm.nih.gov/31421124/.

[22] Donaldson Warde-Farley D et al. "The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function". In: *38* (July 1, 2010), pp. 214–220. URL: https://genemania.org/.