Applying AlphaGenome to Autoimmune Disease: Predicting Lupus-Associated Gene Expression Across Immune Cell Types

Aayra Sharma

Abstract

Systemic lupus erythematosus (SLE) is a chronic autoimmune disease where the immune system mistakenly attacks healthy cells and tissues, causing inflammation and organ damage throughout the body. While over 300 different genetic risk loci have been identified through genome-wide association studies (GWAS), cell-type specific patterns of gene expression remain poorly understood. Our work provides a case study on how to identify distinct gene expression signatures across different immune cell types and associated with different genetic variants by using AlphaGenome, an AI tool developed by Google Deepmind. Three lupus associated genes — TNFRSF17, STAT4, and IRF5 — were analyzed across B cells, T cells, and kidney tissues. Our results show that a lupus-associated SNP in TNFRSF17 increases gene expression in B cells, that STAT4 is expressed across kidney, T cells, and B cells, and that IRF5 shows similar high expression in both B and T cells, remaining consistent with their roles in lupus and immune signaling. These findings demonstrate how multiple lupus risk genes affect the same immune cell types, each gene contributing to its own layer of immune dysfunction.

Introduction

Systemic lupus erythematosus (SLE), also known as lupus, is a chronic autoimmune disorder where the immune system attacks healthy tissues in the body by mistake. This can lead to inflammation and damage to the skin, joints, kidneys, and brain. [1] Lupus has been classified as heterogeneous, meaning symptoms and the severity of the disease varies between different people. [2] This diagnosis process and treatment is extremely challenging. Regardless, physicians and researchers have identified a core set of common symptoms.These include extreme fatigue, joint pain, stiffness, fever, and rashes. [1]

Lupus is more common in women compared to men. Studies report that about 9 of 10 people with lupus are women. [3] While there have been treatments to minimize Lupus risk and help manage lupus, there is currently no cure. By understanding what causes lupus at the genetic level, future treatments could be better developed in the future.

While the exact cause of lupus is still unknown, research suggests that the disease comes from genetic predisposition, hormones, and/or environmental factors. [4] Out of these, genetics play a vital role. Studies show that people with a family history of lupus are at a higher risk of developing the disease, demonstrating that individuals can be genetically predisposed. In fact, 20% of people with lupus will have a parent or sibling who already has lupus or may develop lupus and about 5% of the children born to individuals with lupus will develop the illness. [1] Rather than being caused by a single gene, lupus risk mainly comes from multiple genetic variants that each contribute a small effect to developing the disease. Mutations in the *TLR7, IRF5, STAT4* genes cause immune system pathway defects in the synthesis of C1q, C4, and C2 proteins, increasing the chances that an individual is diagnosed with lupus. Widescale research through genome-wide association studies (GWAS) have identified over 300 locations

in the human genome that are associated with lupus risk. Many of these locations or risk-loci affect the amount of the gene that is produced or the cell that it is activated in.

AlphaGenome is an artificial intelligence (AI) tool developed by Google DeepMind. It is able to predict the functional effects of genetic variants using DNA sequence information.. This tool can also read long stretches of DNA, around one million base pairs of genetic code, and predict many regulatory features, including splice sites and splice junctions, chromatin accessibility, transcription factor binding, and 3D chromatin contacts across many cell types and tissues. It can also estimate the functional effect of genetic variants by comparing predictions for a reference DNA sequence with predictions for a mutated sequence. Rather than merely identifying whether a variant is present, AlphaGenome helps predict how a DNA sequence difference may change regulatory activity, gene expression, or other molecular processes, including whether a variant is likely to increase or decrease expression of particular genes in specific cellular contexts. These variants are often referred to as single nucleotide polymorphisms (SNPs).

In this study, AlphaGenome is used to study three genes that have been directly linked to lupus. These genes are *STAT4, IRF5*, and *TNFRSF17* which are all strongly associated with lupus susceptibility. These genes play a crucial role in immune signaling, cell activation, and inflammation, all processes that are uncontrolled in lupus. We use AlphaGenome to examine how the expression patterns of these genes differ across different tissues and cell types, particularly T-cells, B-cells, and the kidneys. By identifying how lupus-associated DNA variants influence gene expression across different genes and several different tissues, this study seeks to demonstrate how computational genomics can help connect genetic variation to lupus disease risk.

This study addresses a gap in current literature by providing one of the first applications of AlphaGenome to lupus-associated genetic variants and genes, establishing a framework for using deep learning models in autoimmune disease research. Beyond demonstrating the tool’s utility, this paper focuses on identifying differences between reference and alternate allele expression and variation in gene expression across different tissues and cells — findings that cannot be obtained through GWAS alone. Notably, AlphaGenome predicted that STAT4 and IRF5 are expressed at meaningful levels across both B cells and T cells simultaneously, suggesting that their risk variants can affect immune signaling across multiple cell types at once, rather than acting through a single pathway.

Methods

Gene Selection

Three genes were selected for this study based on their connections to lupus. STAT4 is a transcription factor that plays an important role in immune system regulation and helps control inflammation and immune responses. [5] IRF5 is another transcription factor which acts as a principal regulator of inflammatory cytokine production. [6] It is involved in the body's defense against viruses and plays a key role in producing type I interferons that help activate the immune system. TNFRSF17, also known as BCMA, is a gene that helps B-cell development, survival, and differentiation. It is an immune cell that makes antibodies and helps them survive and function. [7]

Data Analysis Using AlphaGenome

AlphaGenome was used to help analyze genes and DNA sequences which contain lupus associated variants. For each gene, AlphaGenome was used to predict transcriptional activity across different cell types and tissues relevant to lupus, including T cells, B cells, and kidney tissue. For each of the three genes, the DNA sequence which surrounds the gene was retrieved from the human reference genome. This sequence is about one million base pairs long and was fed into AlphaGenome, which predicts the gene expression levels for that gene across the different tissue/cell types mentioned above.

Specifically for the TNFRSF17 analysis, two sequence conditions were compared. Every human has two copies of each autosomal chromosome, and at any position, there is a reference version of the DNA letter—this is called the reference allele. Sometimes, people can carry the other version of the letter at that position—this is the alternative allele. One with the reference (REF) allele (C) which is the reference DNA sequence found in the reference genome and one with the alternative (ALT) allele (G) which is the sequence containing the lupus associated variant. Both sequences were run through AlphaGenome and the predicted gene expression levels in B-cells were compared.

Results

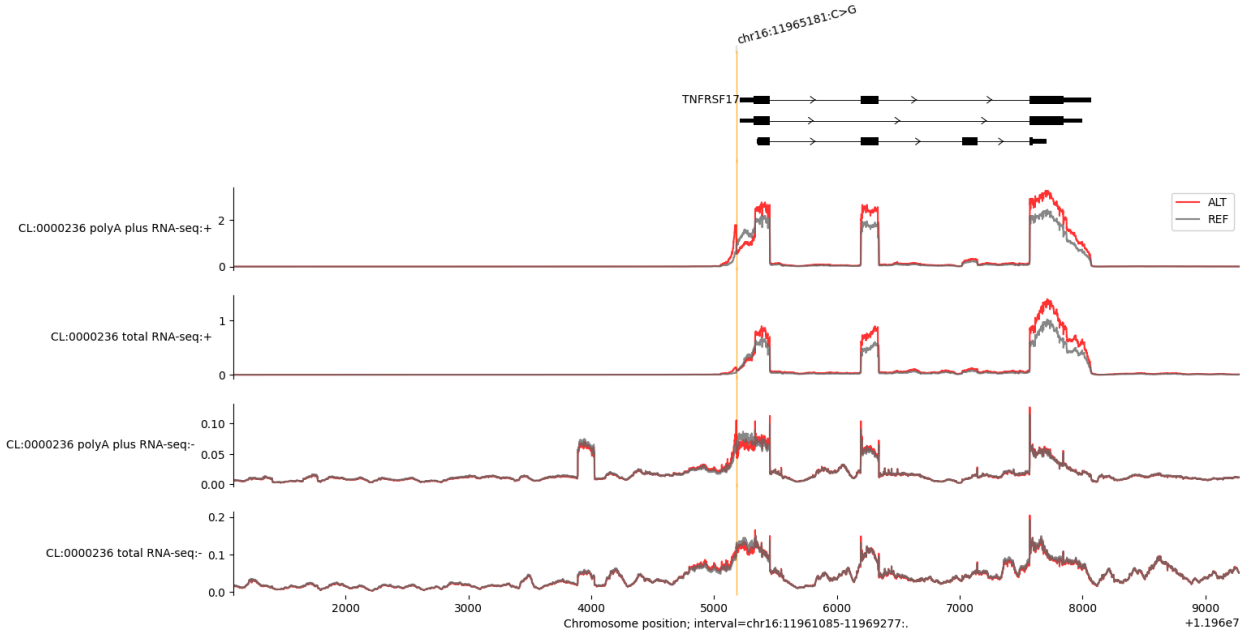


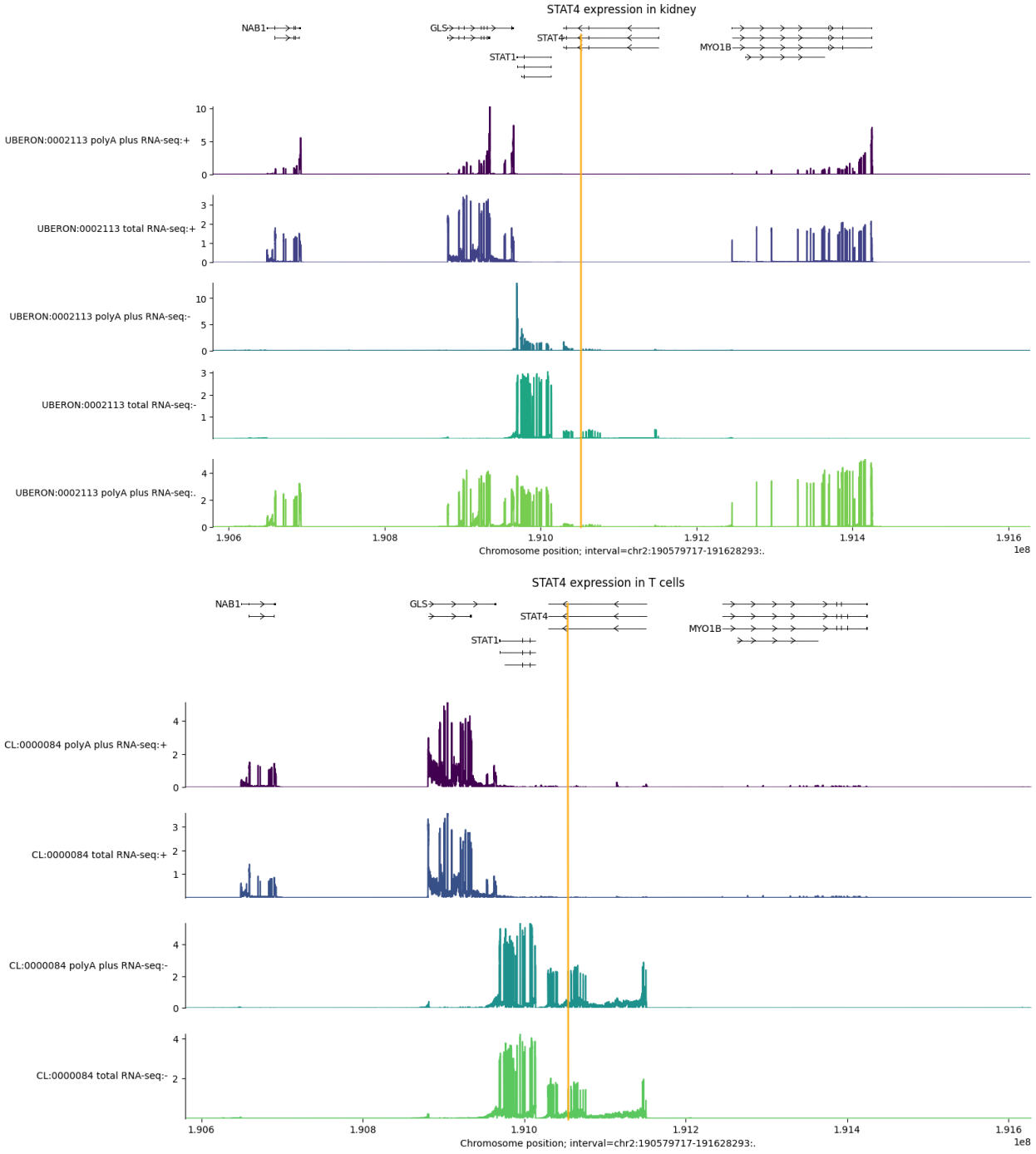
Figure 1. Predicted regulatory effects of the *TNFRSF17* lupus associated variant (chr16:11965181:C>G) generated using AlphaGenome. Tracks show predicted RNA-seq activity in B cells for the reference (C) and alternative (G) alleles across polyA-selected and total RNA-seq, plus and minus strands. The higher values indicate greater predicted gene expression compared to the reference sequence.

Figure 1 shows predicted RNA-seq activity at the TNFRSF17 gene locus in B cells. It specifically shows the results of comparing the reference allele and an alternative allele that contains a SNP. In the case of this figure, the reference allele is C (cytosine), and the alternative allele is G (guanine). TNFRSF17 has been established as a lupus-risk gene in genetic studies, including a paper by Nature identifying it as one of the important genes that associate with Lupus. [8]

In Figure 1, the x-axis shows the position of chromosome 16 and the yellow line marks the location of the SNP. The red represents the predicted gene expression for the alternative allele and the gray represents the predicted gene expression for the reference allele. The y-axis shows how much RNA is predicted to be produced, where the higher the line the higher the gene expression.

The four tracks on the left side of the figure represent different ways to measure RNA. RNA-seq comes in two different types: polyA-selected RNA-seq which only captures mature messenger RNA (mRNA), and total RNA-seq which captures all RNA within the cell. Since DNA is double stranded so genes can be read from either direction, each type of RNA-seq is measured on two different strands: the plus (+) strand and the minus (-) strand. TNFRSF17 is usually more active on the plus strand tracks.

Through analyzing the plus strand tracks in Figure 1, it is apparent that the red line (ALT allele) constantly remains above the gray line (REF allele). This means that AlphaGenome predicts that people who have the alternate allele (G) would have higher TNFRSF17 expression in their B cells compared to people who carry the REF or C allele.



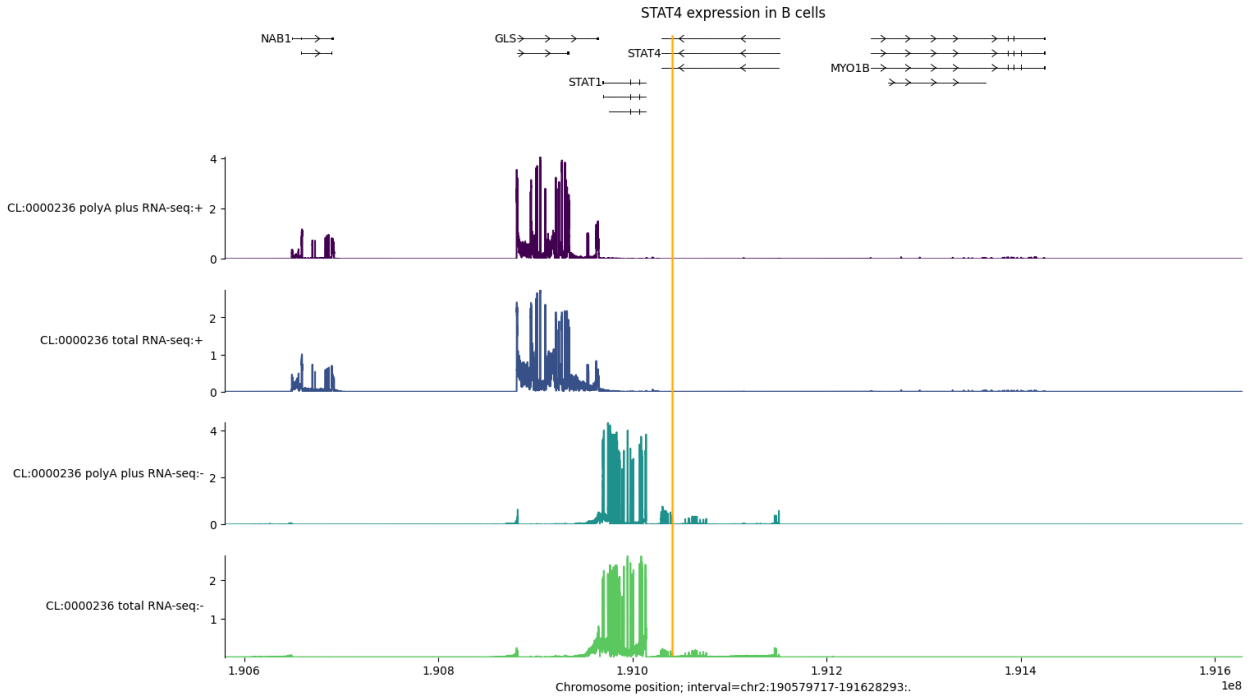


Figure 2. Predicted STAT4 Expression across Kidney Tissue, T cells, and B cells (top, middle, and bottom panels, respectively). Tracks show predicted polyA-selected and total RNA-seq signal on the plus and minus strands across the STAT4 region on chromosome 2.

Figure 2 shows AlphaGenome's predicted STAT4 expression across different tissues and cells (Kidney, T cells, and B cells). STAT4 is a transcription factor involved in cytokine signaling pathways, particularly those associated with T-helper cell differentiation and inflammatory immune responses. STAT4 is one of the most widely known lupus risk genes and certain variants in this gene are more commonly found in people with lupus. [5] It also activates in response to a signaling protein known as interleukin-12 (IL-12) which helps increase interferon-gamma production, which is a strong inflammatory molecule. [9]

In Figure 2, each panel has bumps and peaks on the RNA-seq tracks. These bumps and peaks represent predicted areas of strong gene activity. In areas where the graph is flat or near 0, there is no gene expression. In areas where the graph has peaks, the gene is being transcribed and expressed the most.

STAT4 shows predicted expression across all three tissue/cell types examined: kidney, T cells, and B cells. Of the three, T cells show the highest predicted expression. B cells also show predicted expression, though at a lower level than T cells.

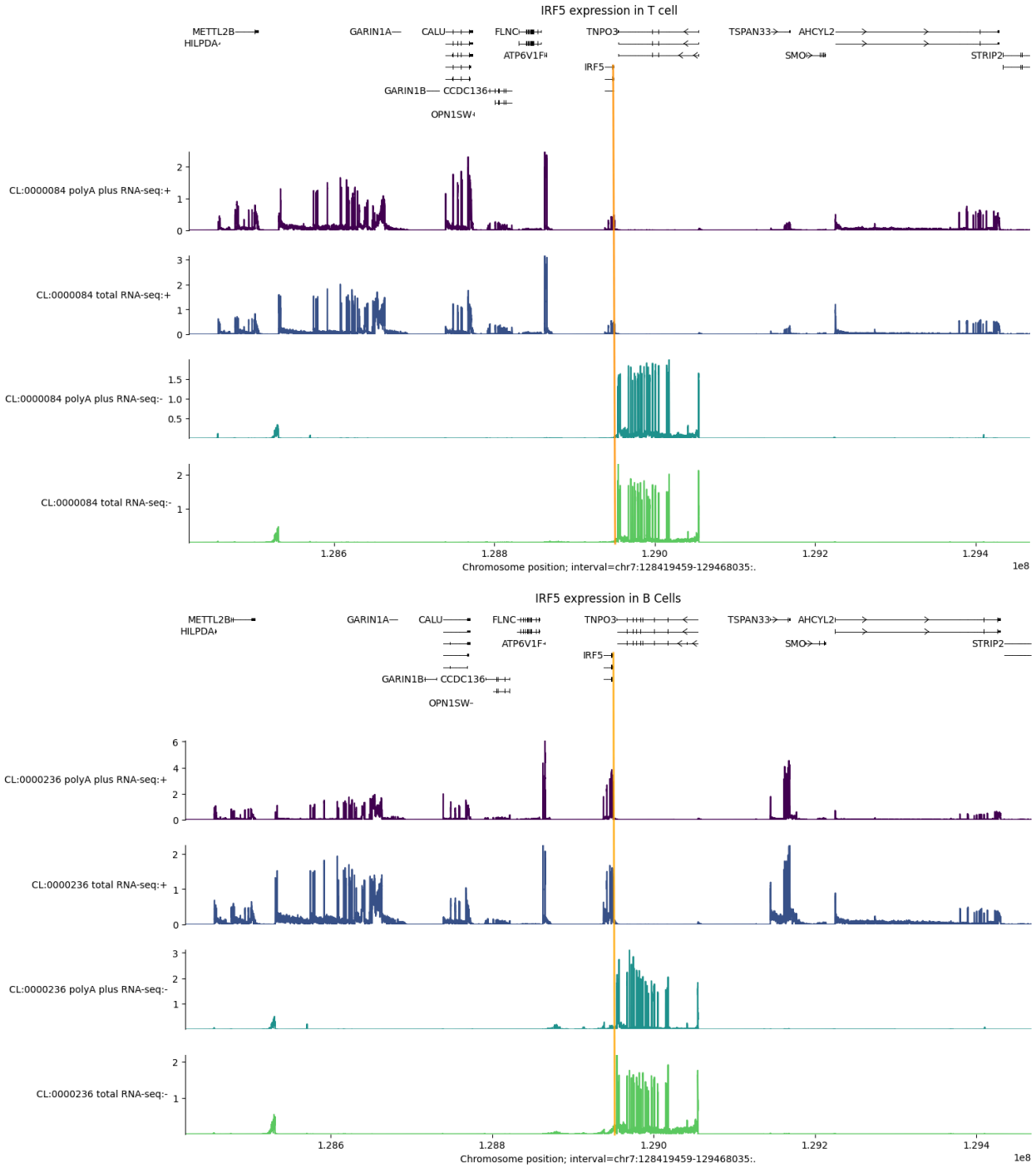


Figure 3. Predicted IRF5 expression across two immune cell types: B cells (top panel) and T cells (bottom panel) Tracks show predicted polyA-selected and total RNA-seq signal on the plus and minus strands across the IRF5 gene region on chromosome 7.

Figure 3 shows AlphaGenome's predicted IRF5 expression across two immune cell types: B cells in the top panel and T cells in the bottom panel. IRF5 is a transcription factor which are

proteins that regulate gene expression by binding to DNA sequences to regulate gene expression. The main job of IRF5 is to regulate inflammatory cytokine production and produce type I interferons like interferon-alpha and interferon-beta. Interferons are very powerful immune signaling proteins that are meant to help the body fight viruses. In lupus, interferon production increases which leads to inflammation and causes the immune system to attack healthy tissues. IRF5 is one of the most strongly associated lupus risk genes identified by GWAS.

In Figure 3, each panel has the same x-axis, which is the IRF5 gene region on chromosome 7. The y-axis shows the gene expression and predicted RNA-seq signal. The taller the peak the more active transcription and higher gene expression.

IRF5 shows high predicted gene expression in both B cells and T cells, with both being quite similar in expression levels, though B cell expression is somewhat higher overall, corroborating previous studies suggesting that IRF5 genes are heavily expressed and associated with B cells. [6]

Discussion and Conclusion

Summary of Findings

This paper used AlphaGenome to help study gene expression patterns of three lupus associated genes — TNFRSF17, STAT4, and IRF5 across multiple different cell and tissue types. We also tested whether a specific SNP in the TNFRSF17 gene would change its expression in B cells. Our analyses demonstrate that multiple lupus risk genes are active in the same cell and tissue types which collectively contributes to the immune system dysfunction that drives the disease. The main findings are summarized below.

First, the alternative allele (G) in the TNFRSF17 variant chr16:11965181:C>G has a higher predicted gene expression in B cells compared to the reference allele (C).

Second, STAT4 shows predicted expression in B cells, T cells, and kidneys, with the highest predicted expression in T cells.

Third, IRF5 is predicted to be expressed at similarly high levels in both B cells and T cells.

Biological Significance

TNFRSF17 helps B cells survive, develop, and produce antibodies, so higher gene expression associated with the alternative allele could potentially lead to overactive antibody production, which is directly connected to lupus. The findings in Figure 1 support the idea that this DNA variant can influence lupus risk via increasing TNFRSF17 expression in B-cells.

Since STAT4 shows predicted expression in the kidney, B cells, and T cells, this suggests it may be linked to lupus across multiple tissues at once. In the kidneys, lupus is particularly prevalent through lupus nephritis, where the immune system attacks kidney nephrons and kidney inflammation occurs; STAT4 expression in kidney tissue could suggest that STAT4 signaling contributes to kidney damage in lupus patients, which can lead to higher risk of getting the disease altogether. In the T cells panel, STAT4 shows the highest expression, which makes

sense because T cells are one of the main cells where STAT4 functions; the higher expression in T cells corroborates the hypothesis that STAT4 expression plays an important role in regulating immune system responses which could potentially support lupus development. In B cells, STAT4 is also expressed, but not as much due to STAT4 being mainly prevalent in T cells; this is still consistent with the idea that higher STAT4 expression leads to more antibody production in lupus.

Since IRF5 shows high predicted gene expression in B cells and T cells, it could possibly be linked to lupus. In B cells, the IRF5 promotes the production of type I interferons and pro-inflammatory cytokines which directly contributes to the overproduction of antibodies in the immune system. In T cells, IRF5 promotes pathogenic Th1 and Th17 differentiation which amplifies tissue inflammation. Together, as shown in the figure, higher expression of IRF5 of both B and T cells leads to higher lupus risk.

Taken together, our results show that lupus is a disease where multiple genetic risk factors on the same immune cell types (B cells and T cells) contribute to cellular dysfunction, which increases the likelihood of contracting the disease. TNFRSF17 keeps autoreactive B cells alive, which can potentially cause the body to attack healthy tissues. STAT4 amplifies inflammatory cytokine signals that push B cells and T cells to act more aggressively. IRF5 keeps interferons constantly switched on, causing the inflammation that comes with having lupus.

Limitations

This paper has multiple limitations that must be acknowledged. First, only three lupus genes were studied out of more than 100 that can be identified by GWAS. These three genes were selected based on available literature and interest for testing in AlphaGenome. There are many other significant genes that were tested during the research stage of this paper using AlphaGenome. Some of these genes include TNFAIP3, BLK, PTPN22, CXCR5, TYK2 and C4A.

Second, AlphaGenome is a genomics computational model so it is trained off of existing genomic data. It is not an experimental measurement. Every prediction that AlphaGenome produces is a model-based estimate that still needs to be validated by real laboratory experiments before the results can be considered confirmed. With that being said, AlphaGenome was trained on human and mouse genomes and matches or exceeds the strongest available external models, so it is still a beneficial tool which provides lots of collective information and insight.

Third, the variant analysis in this paper was limited to TNFRSF17. Many other lupus-associated variants exist across each of the three genes that have been studied.

Future Directions

Future research can apply similar AlphaGenome-based approaches across all lupus-associated risk loci beyond the three genes examined here, and the same workflow could be extended to other autoimmune diseases as well. The variant-level analysis performed here for TNFRSF17 could also be extended to additional lupus-associated SNPs in STAT4 and IRF5, rather than relying on baseline expression predictions alone.

Conclusion

This research paper demonstrates the usage of AlphaGenome as a powerful tool used to investigate the genetic basis of lupus by analyzing gene expression levels across different cell and tissue types. By analyzing three key lupus genes, TNFRSF17, STAT4, and IRF5 across B cells, T cells, and kidney tissue, this paper has demonstrated that multiple lupus risk genes are active in the exact same immune cell types. Together, they combine and contribute to dysfunction in the immune system. Although lupus is a complex disease, genomic analysis represents a meaningful step towards understanding why certain genetic variants and higher gene expression increase disease risk.

References

[1] Pons-Estel, G. J.; Alarcón, G. S.; Scofield, L.; Reinlib, L.; Cooper, G. S. Understanding the Epidemiology and Progression of Systemic Lupus Erythematosus. *Seminars in Arthritis and Rheumatism* 2010, *39* (4), 257–268. https://doi.org/10.1016/j.semarthrit.2008.10.007.

[2] Banchereau, R.; Hong, S.; Cantarel, B.; Baldwin, N.; Baisch, J.; Edens, M.; Cepika, A.-M.; Acs, P.; Turner, J.; Anguiano, E.; Vinod, P.; Khan, S.; Obermoser, G.; Blankenship, D.; Wakeland, E.; Nassi, L.; Gotte, A.; Punaro, M.; Liu, Y.-J.; Banchereau, J.; Rossello-Urgell, J.; Wright, T.; Pascual, V. Personalized Immunomonitoring Uncovers Molecular Networks That Stratify Lupus Patients. *Cell* 2016, *165* (3), 551–565. https://doi.org/10.1016/j.cell.2016.03.008.

[3] Nusbaum, J. S.; Mirza, I.; Shum, J.; Freilich, R. W.; Cohen, R. E.; Pillinger, M. H.; Izmirly, P. M.; Buyon, J. P. Sex Differences in Systemic Lupus Erythematosus. *Mayo Clinic Proceedings* 2020, *95* (2), 384–394. https://doi.org/10.1016/j.mayocp.2019.09.012.

[4] Mok, C. C.; Lau, C. S. Pathogenesis of Systemic Lupus Erythematosus. *J Clin Pathol* 2003, *56* (7), 481–490. https://doi.org/10.1136/jcp.56.7.481.

[5] Fike, A. J.; Chodisetti, S. B.; Bricker, K. N.; Choi, N. M.; Chroneos, Z. C.; Kaplan, M. H.; Rahman, Z. S. M. STAT4 Is Largely Dispensable for Systemic Lupus Erythematosus–like Autoimmune- and Foreign Antigen–Driven Antibody-Forming Cell, Germinal Center, and Follicular Th Cell Responses. *ImmunoHorizons* 2021, *5* (1), 2–15. https://doi.org/10.4049/immunohorizons.2000111.

[6] Cham, C. M.; Ko, K.; Niewold, T. B. Interferon Regulatory Factor 5 in the Pathogenesis of Systemic Lupus Erythematosus. *Clinical and Developmental Immunology* 2012, *2012*, 1–11. https://doi.org/10.1155/2012/780436.

[7] Nobari, S. T.; Nojadeh, J. N.; Talebi, M. B-Cell Maturation Antigen Targeting Strategies in Multiple Myeloma Treatment, Advantages and Disadvantages. *J Transl Med* 2022, *20* (1), 82. https://doi.org/10.1186/s12967-022-03285-y.

[8] Avsec, Ž.; Latysheva, N.; Cheng, J.; Guido Novati; Taylor, K. R.; Ward, T.; Bycroft, C.; Nicolaisen, L.; Arvaniti, E.; Pan, J.; Thomas, R.; Dutordoir, V.; Perino, M.; De, S.; Karollus, A.; Gayoso, A.; Sargeant, T.; Mottram, A.; Wong, L. H.; Drotár, P.; Kosiorek, A.; Senior, A.; Tanburn, R.; Applebaum, T.; Basu, S.; Hassabis, D.; Kohli, P. Advancing Regulatory Variant Effect Prediction with AlphaGenome. *Nature* 2026, *649* (8099), 1206–1218. https://doi.org/10.1038/s41586-025-10014-0.

[9] Yang, C.; Mai, H.; Peng, J.; Zhou, B.; Hou, J.; Jiang, D. STAT4: An Immunoregulator Contributing to Diverse Human Diseases. *Int. J. Biol. Sci.* 2020, *16* (9), 1575–1585. https://doi.org/10.7150/ijbs.41852.