Investigations into the Prediction of Patient Response to Rheumatoid Arthritis Treatment

Sophia Ju SOPHIA.JU@MAIL.MCGILL.CA
COMP 401
McGill University

Professor Yue Li YUELI@CS.MCGILL.CA
School of Computer Science
McGill University

1. Introduction

The autoimmune disease Rheumatoid arthritis (RA) affects millions of people worldwide and is an exceedingly common inflammatory arthritis (1). RA patients often suffer from damage to the synovial joint lining, which not only causes pain but also leads to social and financial difficulties (2). Throughout a patient's lifetime, severe instances of RA can also affect other organs such as the eyes, skin, or lungs (3). Treatment of RA typically begins with disease-modifying anti-rheumatic drugs, or DMARDs, such as methotrexate or anti-TNF therapies (4). However, these medications do not always work and may come with adverse side effects. At first, RA patients are prescribed methotrexate due to its cheaper cost. If their condition does not improve or if they cannot endure the side effects, they are then prescribed anti-TNF treatments (5). Still, around 30% of patients do not respond well to anti-TNF therapy (6). The failure of these treatments may not only increase the risk of infection but also puts additional economic burden on patients. Therefore, the limitations of current therapies for RA emphasize the need for response prediction for these medications as well as personalized treatment plans.

The Dialogue for Reverse Engineering Assessments and Methods (DREAM) RA Responder Challenge invited research teams to create models for patient response to anti-TNF therapy (7). The winning model relied heavily on limited genetic input and was unable to correctly predict responses in a large number of subjects. Using both the clinical and single nucleotide polymorphism (SNP) data provided, we built upon the results of this competition to create a predictive model that performed better than the winning model. We also had access to a separate, smaller dataset of RA patients that included the RNAseq counts of patients that were treated with methotrexate and clinical notes of their response. Using these counts, we could do statistical analysis and find genes that were differentially expressed (DE-genes) in responders and non-responders. After finding the most significant SNPs and DE-genes from each dataset respectively, we were able to investigate the genes that overlapped in location with significant SNPs and visualize their expression. From this, we identified 5 genes in the T-cells of patients undergoing treatment that are expressed differently in responders and nonresponders. These results will hopefully serve as a baseline for future explorations into the prediction of response to RA treatments.

2. Background

2.1 Methotrexate vs. Anti-TNF Therapies

This report investigates patient responses to multiple different treatments against RA. For the sake of exploring new datasets, we assumed that a patient will respond to methotrexate treatment if and only if they respond to anti-TNF treatment. However, this is a rather bold assumption to make as methotrexate and anti-TNFs have their differences. Methotrexate is a conventional synthetic DMARD (csDMARD) and has been used to treat rheumatoid arthritis since the 1980s. Although the full mechanisms by which it works at low doses are not yet fully understood, one of the proposed theories for its anti-inflammatory properties is folate antagonism. This reduces cellular proliferation by inhibiting purine and pyrimidine synthesis and works to lessen synovial inflammation by suppressing cells such as T lymphocytes (8). On the other hand, anti-TNF treatment was optimized to target the tumor necrosis factor (TNF) cytokine. The body produces the TNF cytokine to mediate inflammation in response to injury or stress, but in active RA patients, TNF is continuously present in synovial tissue (9). Monoclonal antibodies against TNF, or anti-TNF, work by inhibiting the TNF cytokine as well as other pro-inflammatory cytokines (9). There are a few different types of anti-TNF therapies available that are commonly used when a patient does not respond well to methotrexate therapy alone; these include adalimumab, etanercept, and infliximab (10). In this report, we will be using data for RA patients who are receiving solely methotrexate therapy, solely anti-TNF therapy, or a combination of both.

2.2 DREAM Challenge Dataset

In the original DREAM challenge, teams were provided with the clinical and whole-genome SNP data of 2706 patients with at least moderate disease activity according to their composite disease activity scores for 28 joints (DAS28). The datasets were combined from 13 different cohorts of European ancestry and all patients were diagnosed by a board-certified rheumatologist. The binary responder status of a patient was determined by the EULAR criteria in addition to the patient's change in DAS28 (ΔDAS28) from before starting treatment to 3-12 months after initiation of anti-TNF therapy. Teams were invited to take part in the open challenge to create the best model for classification of response to anti-TNF therapy as well as for predicting the ΔDAS28 as a continuous feature. This challenge ended in 2014, and despite a significant heritability estimate for the trait of treatment non-response, a review of the challenge concluded that the results of the competition supported the idea that prediction accuracy does not significantly benefit from genetic contributions (7).

We hoped to improve upon these results by using different techniques to filter and choose the genomic data to be used in training the model. Analogously to the DREAM challenge, we were given access to the DREAM data for all 2706 patients through Synapse (synapse.sagebase.org). Clinical data consisted of the patient's ΔDAS28, response/non-response, EULAR response category, genotyping batch, cohort, drug (adalimumab, etanercept, or infliximab), baseline DAS, age, gender, and methotrexate co-therapy. Genomic data was given in the form of single nucleotide polymorphism data or SNP data. An SNP is a variation of a single nucleotide in a particular spot in our DNA. They occur across the genome and can be used for association studies for complex genetic traits (11). In the DREAM challenge, data was imputed across all cohorts to a common dataset of about 2.5 million SNPs. We were provided with both SNP dosage data, or estimated counts of the reference allele, as well as genotype probabilities for each variant in chromosomes 1 through 22. This meant that in the dosage dataset each SNP had one dosage value per individual, and in the genotype probability dataset each SNP had the probabilities of the AA, BB, and AB genotype for each individual.

2.3 RNAseq Dataset

In addition to the DREAM dataset, we also were given access to the clinical and RNAseq data for 6 RA patients that were treated with methotrexate. RNA sequencing, or RNAseq, uses high throughput sequencing to quantify expression levels of genes under varying conditions (12). To obtain the count data for a cell, RNA from the cell is converted to cDNA fragments and then sequenced to get short sequence reads. These reads are then mapped to a reference genome and the number of read counts is recorded for each gene (12). These counts can be used to understand the differential expression of genes in different conditions. In our case, we were provided with the count data of RNA from T-cells from our 6 RA patients for both pre- and post-treatment. 3 of the patients were classified as responders, 2 as partial responders, and 1 as a nonresponder after 6-12 months of methotrexate therapy. All genes from RNAseq were identified and named using their unique HGNC gene symbol.

3. Methods

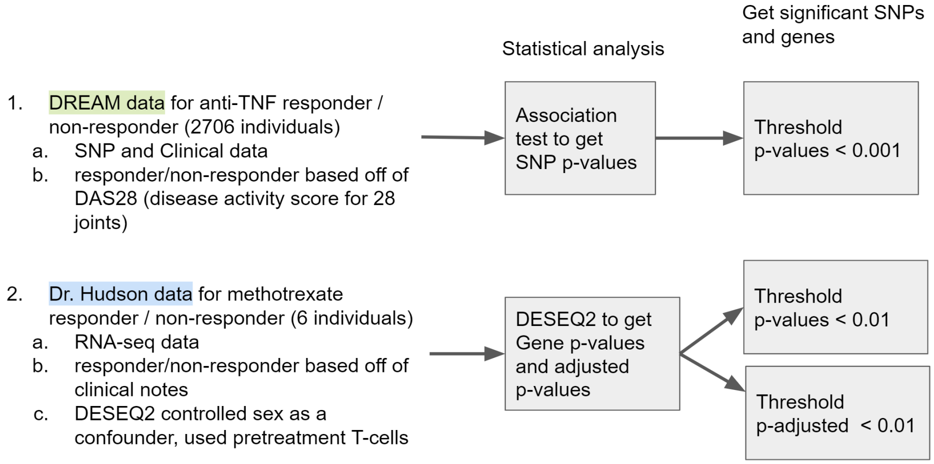
3.1 Classification of responders/non-responders for the DREAM dataset

We split the patients into the same training (N=2031) and testing (N=675) subsets that the teams had access to during the training period of the DREAM competition. In contrast to the winning team from the DREAM challenge that selected SNPs based on existing literature on RA (13), we utilized the full genome-wide dataset of 2.5 million SNPs to choose the most significant ones for model training. When considering a genome-wide map of SNPs, we must also consider linkage disequilibrium (LD), or the tendency for neighboring alleles to correlate (14). For this reason, a single causal mutation could generate statistical significance in other nearby variants (15). To handle this issue we turned to PLINK, an open source whole genome association analysis toolset (16). Given the genomic probability data (.gen files) and phenotype data (.sample file) of the training set, PLINK produced a pruned subset of SNPs that were in approximate linkage equilibrium with a window size of 100kb, step size of 5, and r2 threshold of 0.2. An association analysis could then be run on this subset to calculate the marginal p-value of each SNP based on whether or not it predicted response. We could then perform LD clumping to remove SNPs with borderline significant p-values if they were in linkage disequilibrium with significant SNPs at a p-value threshold of 0.001. This process left us with 284 SNPs for which we had to extract the dosage data to be able to train a model.

Our next step was to train and test different models, which we did using python and the sklearn library. Instead of predicting a binary outcome of responder or non-responder, we used regression to predict a continuous outcome, the change in DAS28 from baseline to 3-12 months after initiation of anti-TNF therapy. When training the model, the labels for each patient were the ΔDAS28 values provided in the clinical data. By taking the predicted response scores and evaluating against the true binary response labels for each patient, we could assess performance using AUROC. We trained both linear and non-linear models (support vector regression, random forest regressor, and least absolute shrinkage and selection operator, or LASSO) on the clinical data alone, SNP data alone, and combined SNP and clinical data. We also plotted the feature importance of the top 30 SNPs and clinical covariates.

3.2 Finding DE genes and aligning with significant SNPs

After the process outlined above was performed, our project took an alternative direction as a different RA dataset became available to us. This new dataset allowed us to see if significant SNPs from the DREAM dataset aligned with significant genes from new RA patients. Pre-treatment transcriptomic data of RNA from our 6 RA patients were used to find genes that were differentially expressed in the T-cells of methotrexate responders and non-responders. To do the differential expression analysis, the DESeq2 R/bioconductor package was used. DESeq2 takes the raw count data and uses an empirical Bayes approach to shrinkage for dispersion estimates to improve the stability of analysis results for small numbers of samples (17). Partial responders considered as non-responders and with this modification, we had 3 responders (2 male, 1 female) and 3 non-responders (2 male, 1 female). Differential expression analysis of the raw count data with sex as a confounder and response as the condition was performed using an R script.


1. DREAM data for anti-TNF responder / non-responder (2706 individuals)
   a. SNP and Clinical data
   b. responder/non-responder based off of DAS28 (disease activity score for 28 joints)
2. Dr. Hudson data for methotrexate responder / non-responder (6 individuals)
   a. RNA-seq data
   b. responder/non-responder based off of clinical notes
   c. DESEQ2 controlled sex as a confounder, used pretreatment T-cells

Statistical analysis
Get significant SNPs and genes
Association test to get SNP p-values
Threshold p-values < 0.001
DESEQ2 to get Gene p-values and adjusted p-values
Threshold p-values < 0.01
Threshold p-adjusted < 0.01

Figure 1: Overview of datasets and workflow for finding significant SNPs and DE genes with respect to responder versus non-responder phenotype

Both empirical and Benjamini-Hochberg adjusted p-values were calculated for each gene. Genes with p-values < 0.01 and genes with p-adjusted < 0.01 were selected for further investigation. We also took another pass at the SNP data from the DREAM challenge and performed association tests without LD-pruning, in preparation for finding overlapping significant regions. We ended up with 4264 SNPs that had marginal p-values < .001.

HGNC gene and SNP locations were found using Ensembl's human genes and human short variants datasets respectively. Only the gene locations of those on chromosomes 1-22 were searched for, as we only had access to SNPs from those chromosomes. After gene locations were found, we wanted to check if significant SNPs overlapped in location with significantly DE genes. To do this, we extended the SNP locations by 1Mb (1 million bp) to the left and the right. If the SNP location crossed a gene location at any point, both the SNP and the DE gene were recorded as overlapping. Each chromosome was checked for these overlaps and the results were recorded. These particular recorded SNPs and genes were then plotted and highlighted in Manhattan plots using R. Additionally, the expressions of the identified genes were plotted in a count heatmap as well as in strip plots that indicate the differences in expression between responders and non-responders.

4. Results

4.1 Model performance on classification of responders/non-responders

The best performing method was Random Forest (RF), a non-linear model that uses decision trees to progressively separate subjects into groups based on the most predictive features. Support Vector Regression (SVR) also outperformed linear methods. Compared to only clinical covariates such as age and sex, adding SNPs improved the prediction from an area under the receiver operating curve (AUROC) of 0.63 to 0.67, i.e., 0.04 improvement. The AUROC of 0.67 was also 0.046 greater than the DREAM challenge winner. Between the different training datasets, ie clinical, SNP, or both, the combined dataset performed the best with AUROC of 0.67, followed by clinical data and then SNP data with AUROCs of 0.63 and 0.62 respectively.

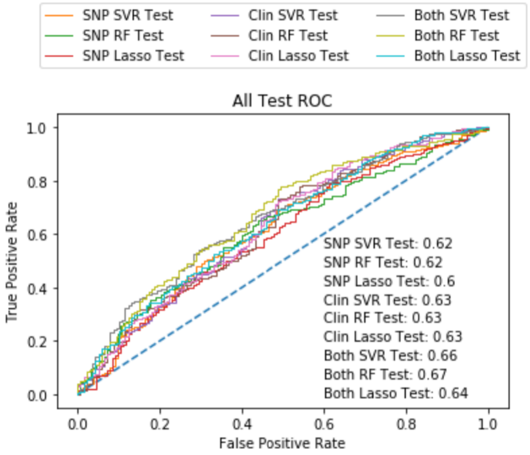


Figure 2: Plot of the test AUROCs for LASSO, SVR, and RF models using the clinical-only, SNP-only, and clinical + SNP datasets.

When plotting the top feature importance that was obtained when training the RF model we observed that among clinical data and SNp data, baseline DAS was the most significant clinical covariate followed by age. We also checked for the top 30 features that were found using just the SNP data and saw that many overlapped between the two datasets (RF trained on SNP and RF trained on combined).

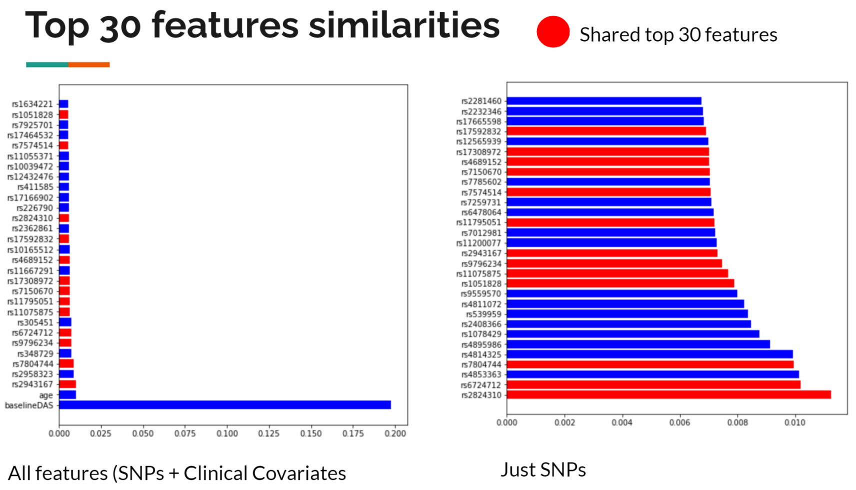
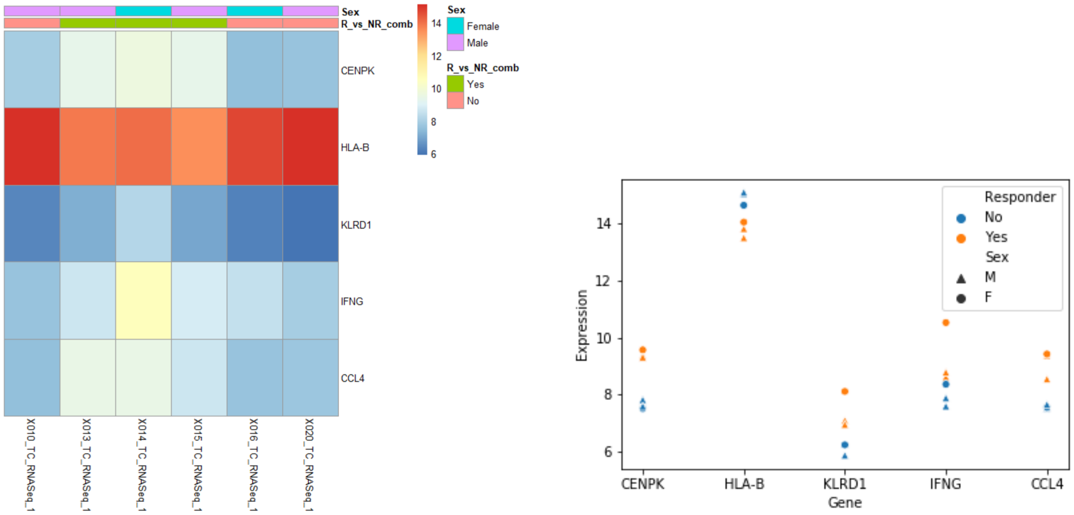


Figure 3: Plot of the top 30 feature importances calculated during training. On the left is the top 30 features for the combined model, and on the right is the top 30 features for the SNP-only model. Bars highlighed in red indicate the features that are in both groups.

4.2 Aligning DE genes with SNPs and plotting expression

After running the association tests without LD-pruning on the SNP training data, we were left with 4264 SNPs with p-values below .001. DESeq2 on the RNAseq dataset resulted in 73 genes with marginal p-values less than .01 and 7 genes with adjusted p-values less than .01. When these genes were put through Ensembl to find their locations, we were left with 60 and 6 genes respectively, as some were not in the database or on chromosomes 1-22. The next step was to find the SNPs overlapping with DE genes. 201 unique SNPs overlapped in location with any one of 44 unique DE genes in the marginal p-value dataset, and 22 unique SNPs that overlapped with any one of 5 DE genes in the adjusted p-value dataset. Many SNPs overlapped in location with the same DE gene. These SNPs and DE genes were labeled in the Manhattan plots (6), where we can visualize the importance of the genes and SNPs and their locations in the genome. Most notably, the 5 DE genes from the adjusted p-value dataset that aligned with significant SNPs were the genes *CENPK, HLA-B, KLRD1, IFNG,* and *CCL4*. After plotting the heatmaps and strip plots of the expressions of these 5 genes, we see that *CENPK, KLRD1, IFNG,* and *CCL4* are expressed more in responders and *HLA-B* is expressed more in non-responders.



(a) Heatmap of gene expression, each column represents a patient

(b) Strip plot of the DE genes and their levels of expression, blue points are non-responders and orange points responders

Figure 4: Plotting gene expression of *CENPK, HLA-B, KLRD1, IFNG,* and *CCL4*

5. Discussion

This report is part of the growing number of studies into the development of methods to personalize RA treatment by predicting response to treatment. Overall, the improvement in AUROC performance of our model using LD-pruning and p-value thresholding on SNPs suggests that using genome-wide statistical analysis on SNP data for feature selection does support the need for further investigation. Moreover, non-linear methods such as RF and SVR gave larger predictive improvements compared to linear methods. This may imply some interaction between SNPs and clinical covariates could serve as potential predictors of response to anti-TNF therapy in RA. Although the models trained exclusively on the SNPs did not perform very well, the increase in performance of the model when trained on both the SNP and clinical data versus the clinical data alone indicates that genomic information provides additional information to the clinical data.

Overfitting of the data also posed a challenge for our model and is something that affects how well the model is able to generalize. Combined with the fact that all patients in this dataset are of European ancestry, this presents a difficult problem for the use of this model with the genetic data of other populations. More samples from a diverse population would be necessary to more universally predict response and non-response. Although we had access to data from the UK Biobank, the patients were self-diagnosed with RA and the individuals' treatment was not specified. Additionally, while this dataset lacked response or non-response labels, there is still an opportunity to make use of this data in the future by potentially pretraining linear weights using the label of RA patient versus non-RA patient before applying the model to the DREAM dataset.

The aligning of DE genes with SNPs and plotting expression yields some interesting results in terms of identifying important gene regions associated with RA and feature selection. By plotting and highlighting the significant SNPs and DE genes on the Manhattan plot we can identify the regions of interest as well as their significance. Moreover, we can see which of these genes are expressed in pretreatment T cells of responders (*CENPK, KLRD1, IFNG, CCL4*) and of non-responders (*HLA-B*). This gives us an idea of which genes are relevant in both the DREAM and RNAseq datasets. In turn, it is possible to verify in existing literature whether the given genes belong to a known pathway or protein related to RA response to treatment. A similar study by Tao et al. used RNAseq to find genes that differ in expression and methylation with respect to the anti-TNF treatments of adalimumab and etanercept (18); however, none of our 5 DE genes correlated with the ones that they had determined to be significant. Furthermore, we found the genes reported in their paper to be associated with response to adalimumab overlapped with our significant SNPs in a Manhattan plot (9) but few overlapped with our most significant SNPs.

6. Conclusion

Ultimately, this report serves as a baseline for the development of using machine learning models to predict responders and non-responders to RA treatment using SNP and clinical data, as well as an introduction into studying the different expressions of the associated genes. Moreover, the methods outlined in the report can function as a template for subsequent studies with a similar goal of using machine learning to personalize RA therapies for individuals. In the future, we hope to incorporate both a larger quantity and a greater variety of data into both the genome-wide and transcriptional data analyses and research. Additionally, we hope to consider various biological explanations for RA-associated genes.

Appendix A.

.1 Manhattan plots for SNPs that overlap with p-value less than .01 DE genes

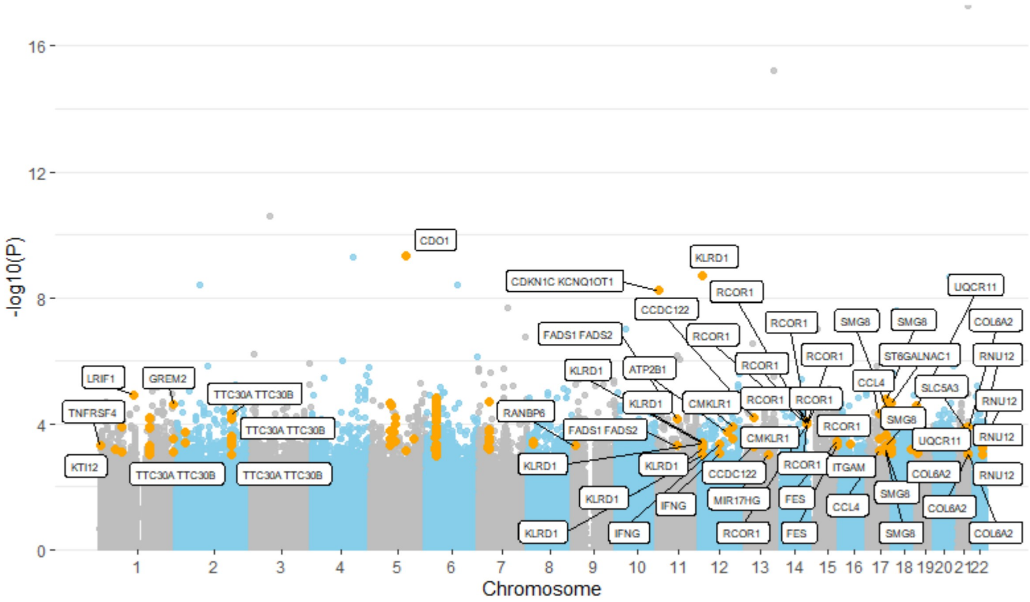


Figure 5: SNP significances (-log10(p-values)) by location in the chromosome. SNPs highlighted in orange and labelled with the DE genes that they overlap with (could be multiple). Not all are labelled due to space constraints.

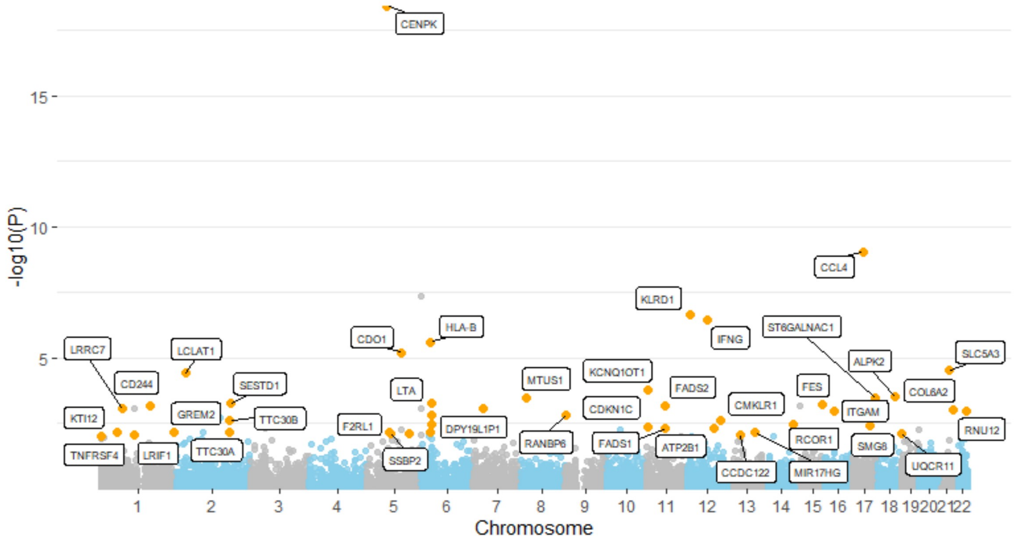


Figure 6: Gene significances (-log10(p-values)) by location in the chromosome. DE genes that overlap with significant SNPs are highlighted in orange and labelled.

.2 Manhattan plots for SNPs that overlap with p-adj less than .01 DE genes

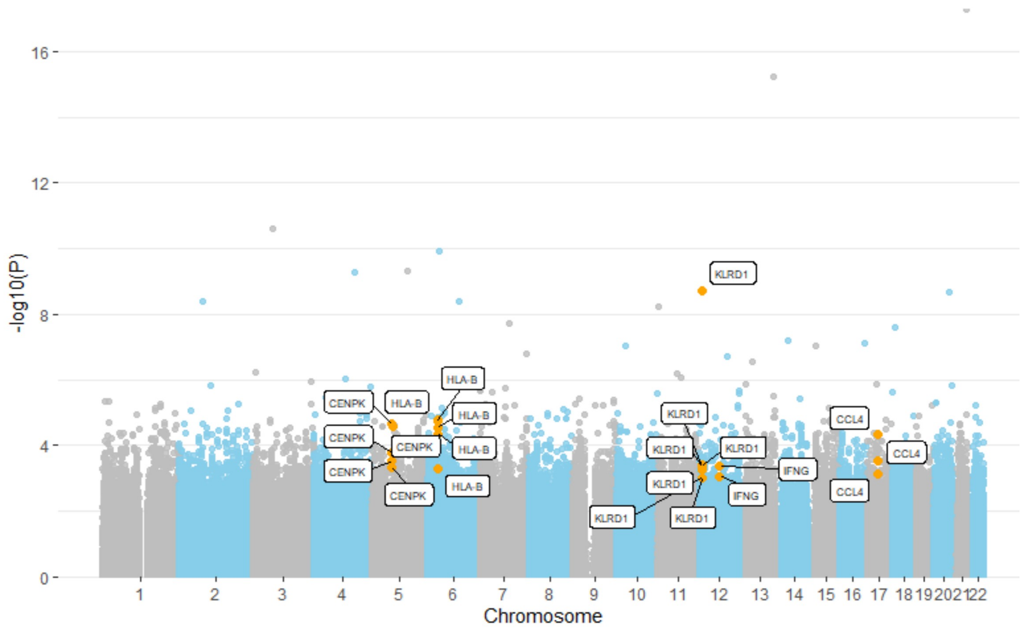


Figure 7: SNP significances (-log10(p-values)) by location in the chromosome. SNPs highlighted in orange and labelled with the DE genes that they overlap with. The DE genes here are the ones with adjusted p-value less than .01.

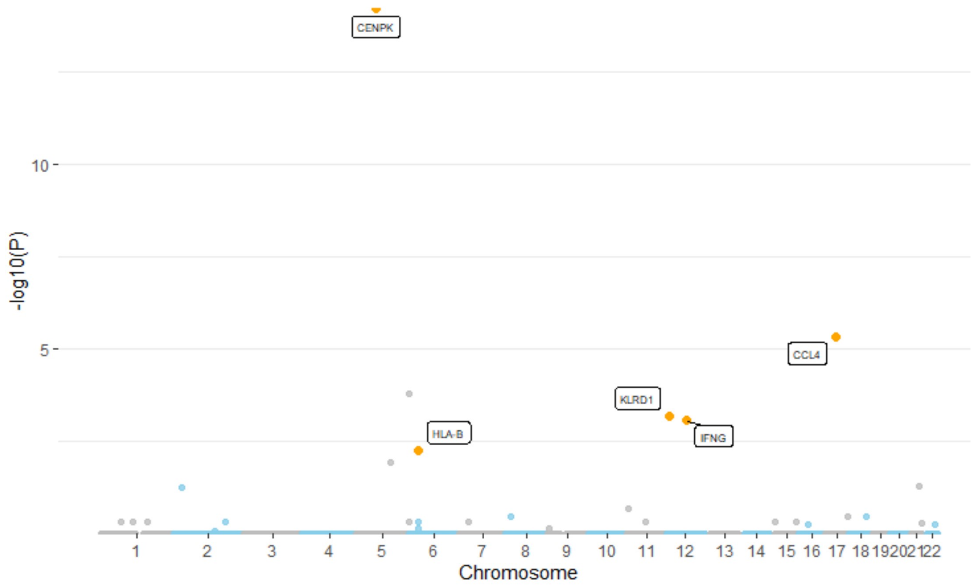


Figure 8: Gene significances (-log10(p-adj)) by location in the chromosome. DE genes that overlap with significant SNPs are highlighted in orange and labelled.

.3 Manhattan plot for SNPs that overlap with genes found by Tao et al.

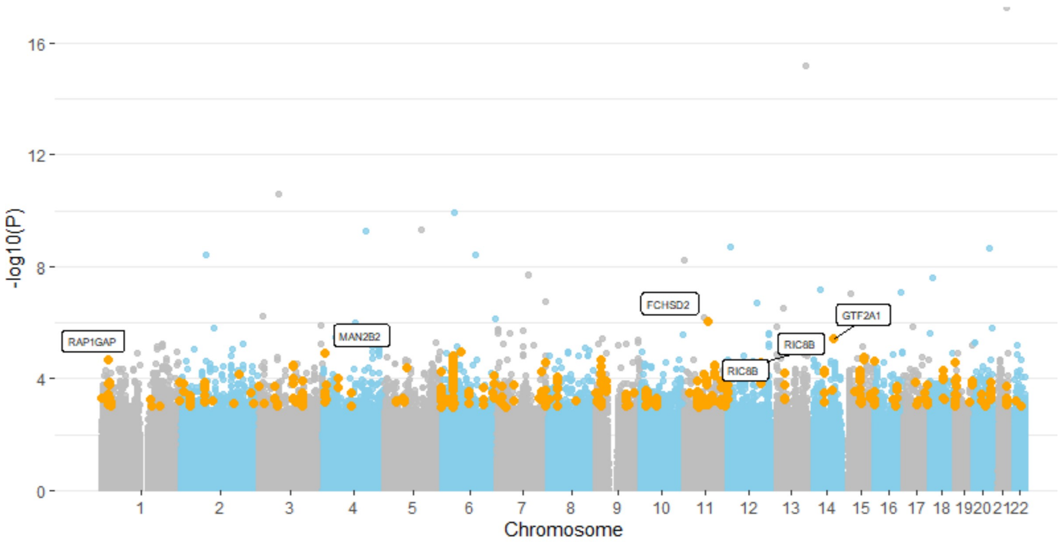


Figure 9: SNP significances (-log10(p-values)) by location in the chromosome. SNPs highlighted in orange and labelled with the DE genes that they overlap with. The DE genes here are the ones identified by Tao et al. (18) Many labels are not shown due to space constraints.

References

[1] Firestein, G. S. (2003). Evolving concepts of rheumatoid arthritis. *Nature*, 423(6937), 356-361.

[2] Gibofsky, A. (2012). Overview of epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis. *The American journal of managed care*, 18(13 Suppl), S295-302.

[3] Matteson, E., and Davis, J. (2012). Overview of the systemic and nonarticular manifestations of rheumatoid arthritis. *In Mayo Clinic Proceedings* (Vol. 87, No. 7, pp. 659-673).

[4] Chen, Y. F., Jobanputra, P., Barton, P., Jowett, S., Bryan, S., Clark, W., ... and Burls, A. (2006). A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. *Health technology assessment (Winchester, England), 10*(42), iii-iv.

[5] Ling, S., Bluett, J., and Barton, A. (2018). Prediction of response to methotrexate in rheumatoid arthritis. *Expert review of clinical immunology, 14*(5), 419-429.

[6] Callaghan, C. A., Boyter, A. C., Mullen, A. B., and McRorie, E. R. (2014). Biological therapy for rheumatoid arthritis: is personalised medicine possible?. *European Journal of Hospital Pharmacy, 21*(4), 229-237.

[7] Sieberts, S. K., Zhu, F., García-García, J., Stahl, E., Pratap, A., Pandey, G., ... and Mangravite, L. M. (2016). Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. *Nature communications, 7*(1), 1-10.

[8] Chan, E. S., and Cronstein, B. N. (2010). Methotrexate—how does it really work?. *Nature Reviews Rheumatology, 6*(3), 175-178.

[9] Monaco, C., Nanchahal, J., Taylor, P., and Feldmann, M. (2015). Anti-TNF therapy: past, present and future. *International immunology, 27*(1), 55-62.

[10] Schmitz, S., Adams, R., Walsh, C. D., Barry, M., and FitzGerald, O. (2012). A mixed treatment comparison of the efficacy of anti-TNF agents in rheumatoid arthritis for methotrexate non-responders demonstrates differences between treatments: a Bayesian approach. *Annals of the rheumatic diseases, 71*(2), 225-230.

[11] Kwok, P. Y. (2001). Methods for genotyping single nucleotide polymorphisms. *Annual review of genomics and human genetics, 2*(1), 235-258.

[12] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), 57-63.

[13] Guan, Y., Zhang, H., Quang, D., Wang, Z., Parker, S. C., Pappas, D. A., ... and Zhu, F. (2019). Machine learning to predict Anti–Tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers. *Arthritis & Rheumatology, 71*(12), 1987-1996.

[14] Pritchard, J. K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics, 69*(1), 1-14.

[15] Berisa, T., and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics, 32*(2), 283.

[16] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics, 81*(3), 559-575.

[17] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology, 15*(12), 1-21.

[18] Tao, W., Concepcion, A. N., Vianen, M., Marijnissen, A. C., Lafeber, F. P., Radstake, T. R., and Pandit, A. (2021). Multiomics and Machine Learning Accurately Predict Clinical Response to Adalimumab and Etanercept Therapy in Patients With Rheumatoid Arthritis. *Arthritis & Rheumatology, 73*(2), 212-222.