BS859 Final Project: The potential association between Rheumatoid Arthritis and Type 1 Diabetes

Background:

Rheumatoid arthritis (RA) and type 1 diabetes (T1D) are all known as the autoimmune disease, as the fact that many patients were diagnosed with both two traits. Therefore, these two traits may share some common genetic variants or may have any causal relationships. And conducting analysis to evaluate the potential common genetic factors provide an insight into the shared biological mechanisms of them and may contribute to the new therapy approaches for RA or T1D diseases.

Methods:

(1) Explore the relationship between RA and T1D by estimating their genetic correlation (using LD score regression).

(2) Assessing the predictive ability of T1D-associated variants for RA risk (using PRS analysis).

(3) Investigating the potential causal relationships between the two diseases (using the Bi-directional Mendelian randomization analysis).

The three methods could provide different perspectives on the associations between 2 traits.

Data:

(1) GWAS RA, a GWAS of self-reported Rheumatoid Arthritis in the UK Biobank.

(2) GWAS T1D, a genome-wide association study of type 1 diabetes (T1D) using 520,580 European samples.

(3) PLINK format files for NARAC RA (for PRS analysis).

(4) UK Biobank LD scores (for LD score regression).

Step 1 (LD score regression): Estimate the genetic correlation between RA and T1D.

The genetic correlation explores the shared genetic effects between two traits, and many variants would affect both traits if they have a pleiotropic relationship. Therefore, we will use the LD score regression to estimate the genetic correlation between RA and T1D, and what we expect is that they will be positively genetically correlated, meaning there are many variants could affect both traits and all of them would be in the same direction. We used two GWAS summary statistics files for RA and T1D, and the UK Biobank Europe LD scores since both two GWAS analysis mainly focus on European population.

Before we conduct the LD score regression analysis, we need to know the key assumptions: (1) The genetic variants are not associated with any confounding factors (like population stratification) that may bias the estimates of associations. (2) The heritability of the trait

could be captured by the genotyped SNP. (3) The causal genetic variants responsible for the trait are correlated with the genotyped SNPs.

We need to reformat the input files for the *ldsc* tool. But before reformatting it using the code in *munge_sumstats.py*, we notice that the original two GWAS files for RA (which have 1,605 cases and 359,589 controls) are very messy, and the required columns for this analysis are stored in two files 'M13_RHEUMA.gwas.imputed_v3.both_sexes.tsv.gz' and 'variants.tsv.gz', respectively. Therefore, we first sorted and merged the two files by rsID first, and then choose the 'rsID, effect allele, reference allele, the frequency of effect allele, beta, standard error, p-value and sample size' from the merged file and stored in 'RAuk_merged_final.tsv'. For the GWAS file of T1D, we filtered out the SNPs which have non-numeric p-value and those whose p-values exceed the threshold range that *ldsc* can handle (only keep the one with p- value < 1e300), then get the prepared file 'T1D_final_4.txt'.

After reformatted the two files, we conduct LD score regression analysis and the results are shown as follows.

```
Heritability of phenotype 1
---------------------------
Total Observed scale h2: 0.0013 (0.0013)
Lambda GC: 1.0195
Mean Chi^2: 1.0207
Intercept: 1.0112 (0.0075)
Ratio: 0.5415 (0.3596)

Heritability of phenotype 2/2
-----------------------------
Total Observed scale h2: 0.0428 (0.0066)
Lambda GC: 1.2005
Mean Chi^2: 1.3969
Intercept: 1.0844 (0.0241)
Ratio: 0.2126 (0.0607)

Genetic Covariance
------------------
Total Observed scale gencov: 0.003 (0.0015)
Mean z1*z2: 0.0283
Intercept: 0.0067 (0.0061)

Genetic Correlation
-------------------
Genetic Correlation: 0.4029 (0.3125)
Z-score: 1.2892
P: 0.1973

Summary of Genetic Correlation Results

p1              p2      rg      se       z       p  h2_obs  h2_obs_se  h2_int
h2_int_se  gcov_int  gcov_int_se

RA.sumstats.gz  T1D.sumstats.gz  0.4029  0.3125  1.2892  0.1973  0.0428    0.0066
1.0844     0.0241    0.0067       0.0061
```

For RA, approximately 0.13% of the variance in RA can be explained by the additive genetic effects of SNPs that are present in the hapmap phase 3 snp set, suggesting a very small proportion of the phenotypic variance of RA can be explained by the genetic variants included in the GWAS summary statistics. The intercept of 1.0112, which is very close to 1, suggesting most of the inflation can be explained by polygenic architecture rather than population stratification or other confounding factors.

For T1D, approximately 4.28% of the variance in T1D can be explained by the additive genetic effects of SNPs that are present in the hapmap phase 3 snp set, still suggesting a very small proportion of the phenotypic variance of T1D can be explained by the genetic variants included in the GWAS summary statistics. The intercept of 1.0844, which is also very close to 1, suggesting there might be no population stratification or other confounding factors.

The summary results show that the genetic correlation between RA and T1D was estimated to be 0.4029, which is positive and with a standard error of 0.3125, a Z-score of 1.2892, and a p-value of 0.1973. The positive effect estimate suggests a shared genetic architecture and the variants that increase RA also tend to increase the risk of T1D. However, P=0.1973 > 0.05 show the result is not statistically significant.

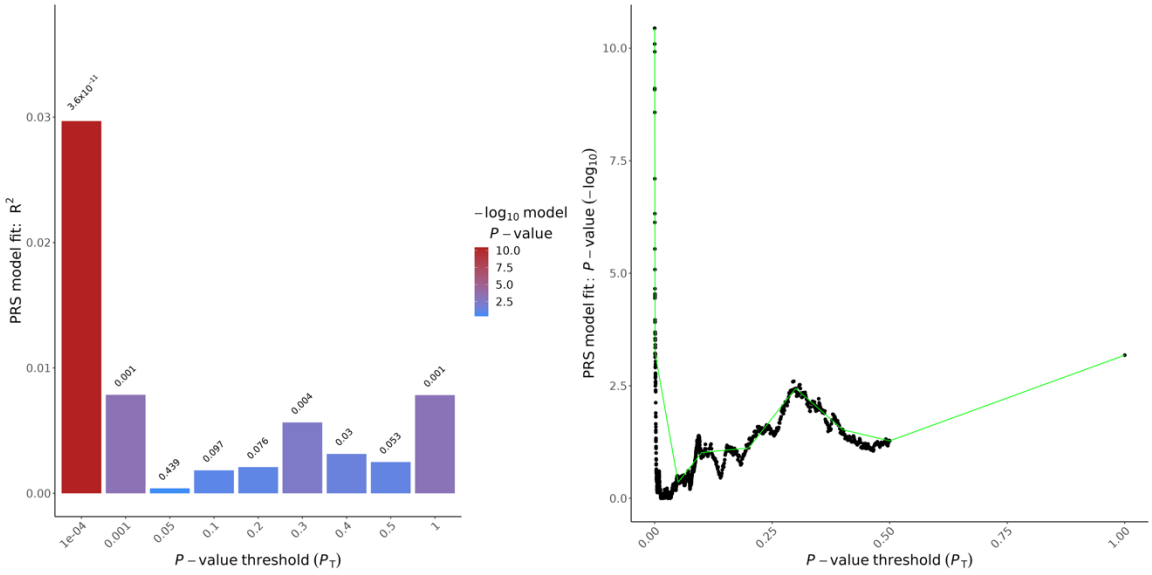
Overall, the result is not statistically significant, suggesting that there might be no genetic relationship between T1D and RA, which is different with our expectations. However, the non-significant result does not mean that there is no genetic correlation between the two traits since we could notice there is extremely small proportion of the phenotypic variance can be explained by GWAS summary statistics for both diseases, which may violate the assumption of LD regression analysis that a substantial proportion of the heritability of the traits is captured by the available genetic markers.

Step 2 (PRS analysis): Using T1D summary statistic data as the base and the NARAC PLINK format files for RA as the target sample to assess how the T1D-associated variants could predict RA risk.

To further estimate the potential associations between RA and T1D, we conduct the polygenic risk score (PRS) analysis to explore how T1D-associated genetic variants can predict the risk of RA in a target sample. Hence, we performed the PRS analysis using T1D summary statistics as the base data, and the NARAC RA files (in PLINK format) as the target sample.

PRS analysis assumes the SNPs included are independent, so it would remove the highly correlated SNPs (in linkage disequilibrium). It also assumes that the effect of each genetic variant on the trait is additive. These assumptions may not hold for some traits when gene-interactions exist. Also, PRS analysis assumes the LD patterns between genetic variants may be same in both the base and target samples.

After conducting the PRS analysis, the results are shown as follows.



# of SNPs after clumping	64,288
# snps in optimal score	368
proportion of variance in RA explained by optimal score	0.0296839 (2.96%)
optimal p-value threshold	0.0001
p-value of best PRS with T1D phenotype in RA	3.60594e-11
Empirical-P	0.000999001
Coefficient	-7.69967
Standard.Error	1.16319

```
Phenotype   Set   Threshold   PRS.R2      Full.R2     Null.R2     Prevalence
      Coefficient Standard.Error    P     Num_SNP     Empirical-P

-     Base  0.0001      0.0296839   0.0296839   0     -     -7.69967
      1.16319     3.60594e-11 368   0.000999001
```

After removing the highly correlated SNPs, 64,288 independent SNPs remains in the PRS analysis, which ensure the risk score is based on independent genetic signals. The optimal polygenic risk score includes 368 SNPs (T1D-associated variants used to predict the risk of developing RA), and it explains approximately 2.96% of the variance in RA risk in our target sample. The optimal p-value threshold of 0.0001 indicating that the optimal PRS includes SNPs with p-values less than or equal to this threshold in the GWAS summary statistics of T1D. The p-value of 3.60594e-11 indicates a statistically significant association between the best PRS and RA risk in the target sample, showing the T1D-associated genetic variants have good power to predict the RA disease. The empirical p-value of 0.000999001 indicates that the significant association between the optimal PRS and RA risk after permutation test. But the negative coefficient of -7.69967 indicates that the direction of the association between polygenic risk score and RA risk is negative, that a higher PRS (higher genetic risk for T1D) is associated with a lower risk of RA. Although

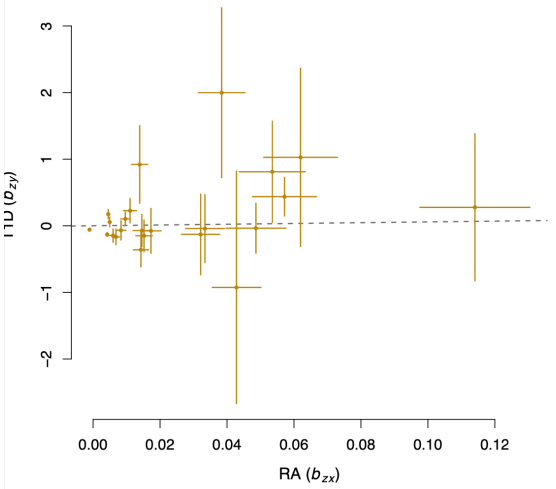
the unexpected direction of effect, it still suggests that some genetic factors contribute to the risk of both RA and T1D.

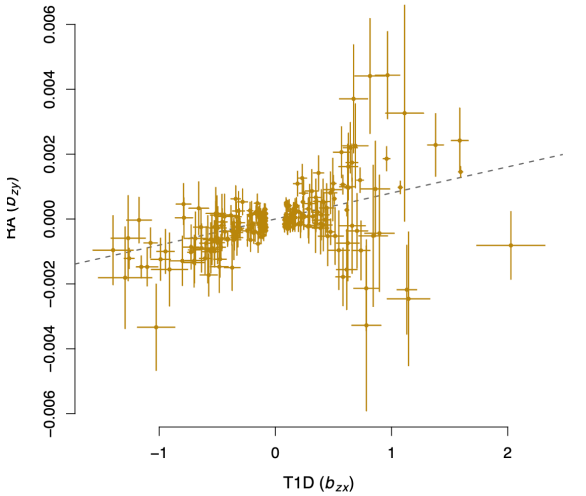
Overall, the genetic variants that associated with T1D are associated with RA in the target sample. The result is significant but show a negative association between PRS and RA risk, which is also very different with our expectation. But the unexpected results may also be due to some bias like the choose of the target sample, the population stratification, or the effect of each genetic variant on the T1D is complex and not additive which may violate the assumption of PRS analysis. These still needs more exploration.

Step 3 (Bi-directional Mendelian randomization): Investigate the potential causal relationship between RA and T1D.

In this step, we conducted a bi-directional Mendelian randomization (MR) analysis to investigate the potential causal relationship between RA and T1D. This method aims to estimate the causal effect of one trait on another.

There are three key assumptions in MR. The genetic variants must be strongly associated with RA or T1D, must be independent of potential confounders that could influence both RA and T1D, and must only affect the outcome (T1D or RA) through their association. And in a bi-directional MR analysis, these assumptions must hold for both directions of the causal relationship (RA to T1D and T1D to RA).





Exposure	Outcome	bxy	se	p	nsnp
RA	T1D	0.582445	2.61013	0.823421	23
T1D	RA	0.000803912	0.000101789	2.83758e-15	259

OR=exp(0.000803912)≈1.0008

exp(0.000803912+1.96*0.000101789)≈1.0001

exp(0.000803912-1.96*0.000101789)≈1.0006

Odds Ratio and its 95% CI (T1D vs. RA) = 1.0008 (1.0001, 1.0006)

In this case, the GSMR analysis provides evidence of a potential causal relationship between T1D and RA (p = 2.83758e-15), but the causal effect of RA on T1D is not statistically significant (p = 0.823421). This suggests that T1D might have a causal influence on RA, but the relationship in the opposite direction is less clear. The result of T1D on RA provides strong evidence of a causal relationship between T1D and RA as the fact that the p-value of 2.83758e-15 is extremely small. But what we need pay attention to is the effect estimates of 0.000803912 is extremely small, suggesting the causal relationship between T1D and RA is strong but very weak. The OR of 1.0008 (1.0001, 1.0006) is very close to 1.0, which means that for each unit increase in T1D, the odds of developing RA would only increase by 0.08%. And 259 SNPs were used as instrumental variables to estimate the causal effect of T1D on RA. Both the effect estimate, and the OR suggest that the casual relationship is weak, not expected for what we know.

The results are not expected for what we know may be because the complex biological mechanisms between the two traits or the true association between two traits violate the assumptions for our analysis, which lead to the bias and unexpected results. For instance, there might be some complex genetic interactions and the genetic variants might be not independent of potential confounders that could influence both RA and T1D.

Conclusions:

In summary, we performed three different analyses to explore the potential genetic associations between the rheumatoid arthritis (RA) and type 1 diabetes (T1D). The LD score regression analysis suggested a positive but not statistically significant genetic correlation between two traits. The PRS analysis showed that 368 T1D-associated genetic variants could predict RA risk in our target sample, but the direction of the association was negative, which is contrary to our expectations. The, the bi-directional Mendelian randomization analysis provided strong evidence of a potential causal relationship between T1D and RA, but the causal effect was weak, with a small odds ratio close to 1.0.

Although some parts of the results were not entirely consistent with our expectations, they still provide evidence of the association between RA and T1D, but we still need more analysis to further explore it as there might be various reasons such as complex biological mechanisms between the two traits, the violation of assumptions in these analyses, or the presence of other confounders. From another perspective, these analyses revealed the limitations of these three methods. They all depend on their strict assumptions and may rely on the selection of the dataset, which may be too idealistic for the real data, especially for those traits have complex genetic mechanisms. Further exploration is needed to better understand the exact relationship between RA and T1D and to assess the findings obtained from our analyses.

Limitations:

The genetic correlation estimates may be extremely biased when there is a complex genetic architecture underlying the traits. And PRS is sensitive to differences in the underlying genetic and population structure between the base sample and target sample, which means the proportion of variance explained by the PRS might be imprecise if the true effect sizes

of genetic risk factors are not captured in the base GWAS. GSMR would easily produce biased estimates of causal effects if the assumptions are violated.

Overall, it is important to consider the assumptions and limitations when interpreting the results from these methods, and further research is needed to confirm the shared genetic mechanisms and causal relationships between RA and T1D.

Future directions:

The multi-trait meta-analysis using MTAG could be performed to see if the integration of the two associated traits could bring any improvements to the results for each individual trait. (i.e., the number of identified variants or any new identified variants). The MTAG would generate the improved Beta, SE, and P-value of each SNP for each trait separately, and then could be used to compare with the original GWAS summary statistics to see if there are any improvements.