Final Project
Jriya186
North American Rheumatoid Arthritis Consortium (NARAC) Analysis

Question 1: Data Cleaning

The NARAC GWAS dataset initially included 2062 individuals (comprising 1194 controls and 868 cases) and 544,276 SNPs. To prepare the dataset for association analysis, a series of quality control (QC) steps were performed using PLINK and EIGENSOFT tools.

The first step in data cleaning was to verify that all individuals had valid sex information. A check of the .fam file confirmed that every individual had valid sex coding (1 for male, 2 for female), and no inconsistencies were found. Next, overall genotyping rate was assessed using PLINK's --freq command, which showed a high genotyping call rate of 99.27%, indicating good-quality data.

Variants were further filtered based on multiple quality metrics. SNPs with a minor allele frequency (MAF) below 1%, more than 5% missing genotype data, or violating Hardy-Weinberg equilibrium (HWE) at $p < 1e-6$ were excluded. Individuals with >5% missing data were also excluded, although in this dataset, all 2062 individuals passed the filter. After this step, 502,304 SNPs remained for analysis.

To avoid potential bias due to genotype missingness correlated with case-control status, PLINK's --test-missing function was used. SNPs with a significant difference in missingness between cases and controls ($p < 0.0001$) were identified and removed. This resulted in the exclusion of 6,504 SNPs, leaving 495,800 variants in the cleaned dataset.

To reduce redundancy in the genotype data and prepare for PCA, linkage disequilibrium (LD) pruning was performed using a sliding window of 10,000 kb, step size 1, and an r^2 threshold of 0.2. The pruning was restricted to autosomal chromosomes (1–22). After pruning, 104,505 SNPs remained for further analysis.

To identify any cryptic relatedness or duplicated samples, Identity-by-Descent (IBD) estimates were computed using PLINK's --genome command. None of the individual pairs had a PI_HAT value greater than 0.25, suggesting there were no duplicate or closely related samples in the dataset.

PCA Analysis

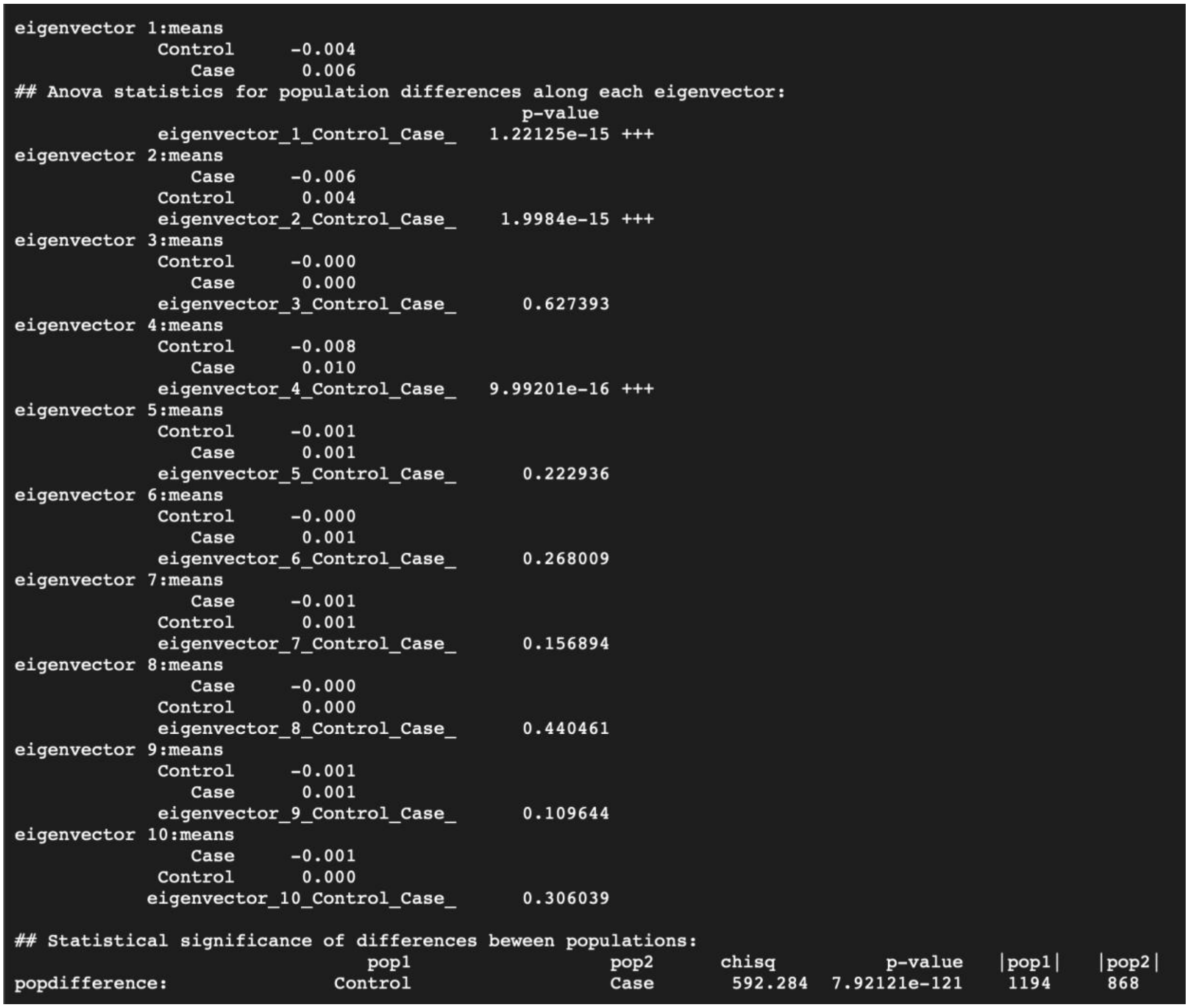
Principal component analysis (PCA) was conducted using the EIGENSOFT smartpca tool. The analysis showed that PC1, PC2, and PC4 were significantly different between cases and controls ($p < 1e-15$), indicating population structure or potential stratification. These components likely capture systematic ancestry-related differences that could confound association results.

To confirm the relevance of the principal components, a logistic regression analysis was performed using PLINK with the top 10 PCs as covariates. This confirmed that PC1, PC2, and

PC4 were significantly associated with case-control status (p-values: 3.04e-19, 2.63e-25, and 1.4e-63, respectively). Other PCs (PC3, PC5–PC10) did not show significant associations and are therefore unlikely to reflect meaningful structure in the context of this analysis.

In conclusion, PC1, PC2, and PC4 should be included as covariates in downstream genome-wide association analyses to control for population structure. Including additional PCs that are not associated with population structure may introduce noise rather than improve model fit. After filtering and pruning, the dataset contained 104,505 high-quality, independent SNPs and 2062 individuals, ready for association analysis.

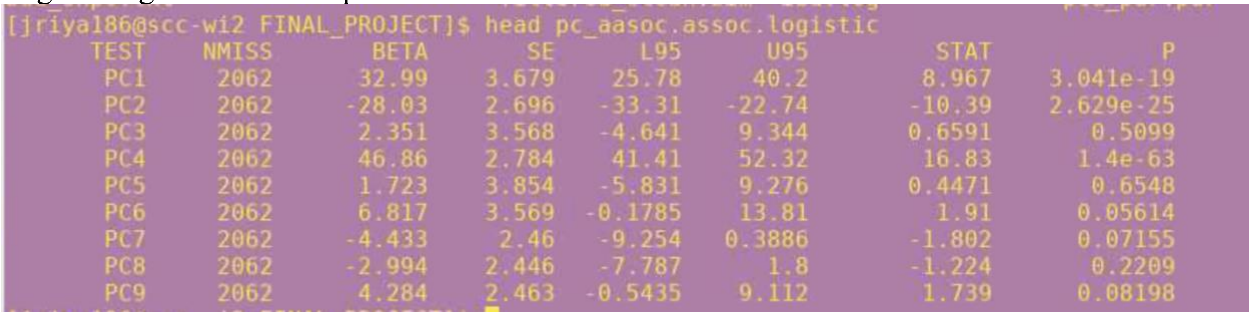
ANOVA output:



```
eigenvector 1:means
             Control     -0.004
                Case      0.006
## Anova statistics for population differences along each eigenvector:
                                              p-value
             eigenvector_1_Control_Case_   1.22125e-15 +++
eigenvector 2:means
                Case     -0.006
             Control      0.004
             eigenvector_2_Control_Case_    1.9984e-15 +++
eigenvector 3:means
             Control     -0.000
                Case      0.000
             eigenvector_3_Control_Case_      0.627393
eigenvector 4:means
             Control     -0.008
                Case      0.010
             eigenvector_4_Control_Case_   9.99201e-16 +++
eigenvector 5:means
             Control     -0.001
                Case      0.001
             eigenvector_5_Control_Case_      0.222936
eigenvector 6:means
             Control     -0.000
                Case      0.001
             eigenvector_6_Control_Case_      0.268009
eigenvector 7:means
                Case     -0.001
             Control      0.001
             eigenvector_7_Control_Case_      0.156894
eigenvector 8:means
                Case     -0.000
             Control      0.000
             eigenvector_8_Control_Case_      0.440461
eigenvector 9:means
             Control     -0.001
                Case      0.001
             eigenvector_9_Control_Case_      0.109644
eigenvector 10:means
                Case     -0.001
             Control      0.000
            eigenvector_10_Control_Case_      0.306039

## Statistical significance of differences beween populations:
                                pop1                  pop2      chisq          p-value    |pop1|   |pop2|
popdifference:               Control                  Case       592.284  7.92121e-121     1194     868
```

Logistic regression with top 10 PCs as covariates

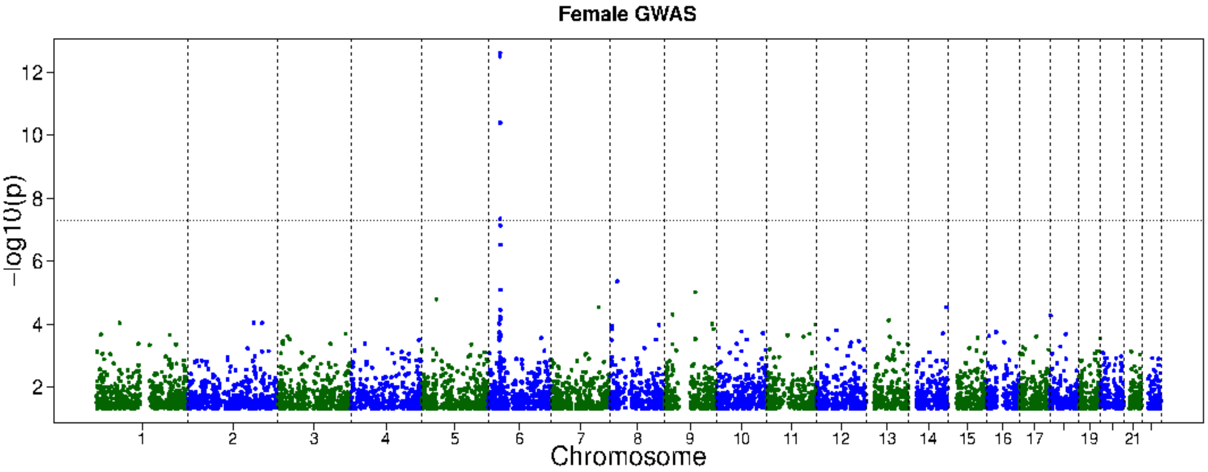


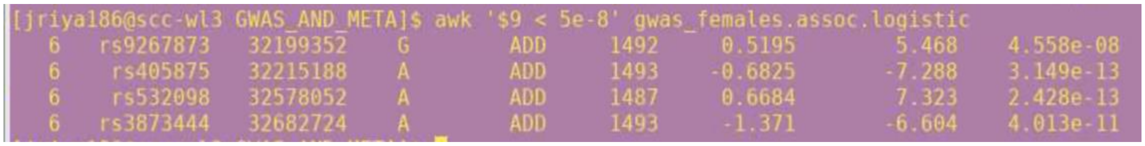
```
[jriyal86@scc-wi2 FINAL_PROJECT]$ head pc_aasoc.assoc.logistic
      TEST    NMISS       BETA       SE        L95       U95          STAT           P
       PC1     2062      32.99    3.679      25.78      40.2         8.967   3.041e-19
       PC2     2062     -28.03    2.696     -33.31    -22.74        -10.39   2.629e-25
       PC3     2062      2.351    3.568     -4.641     9.344        0.6591      0.5099
       PC4     2062      46.86    2.784      41.41     52.32         16.83     1.4e-63
       PC5     2062      1.723    3.854     -5.831     9.276        0.4471      0.6548
       PC6     2062      6.817    3.569    -0.1785     13.81          1.91     0.05614
       PC7     2062     -4.433     2.46     -9.254    0.3886        -1.802     0.07155
       PC8     2062     -2.994    2.446     -7.787       1.8        -1.224      0.2209
       PC9     2062      4.284    2.463    -0.5435     9.112         1.739     0.08198
```

Question 2: GWAS

a. Sex differentiated Genome Wide Analyses

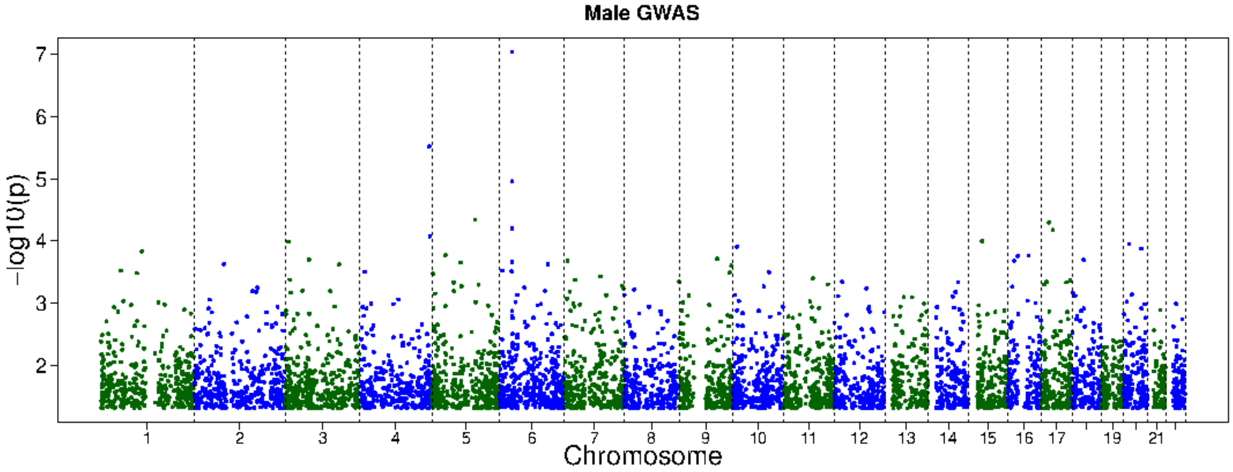
I divided the cohort into male and female subsets using the sex information from the narac.cov file, classifying individuals with a sex label of "2" as female and those with "1" as male. I conducted genome-wide association analyses separately for each sex using logistic regression, incorporating PC1, PC2, and PC4 as covariates based on their significant association with case status. These principal components were selected to account for population structure and reduce confounding in the analysis.



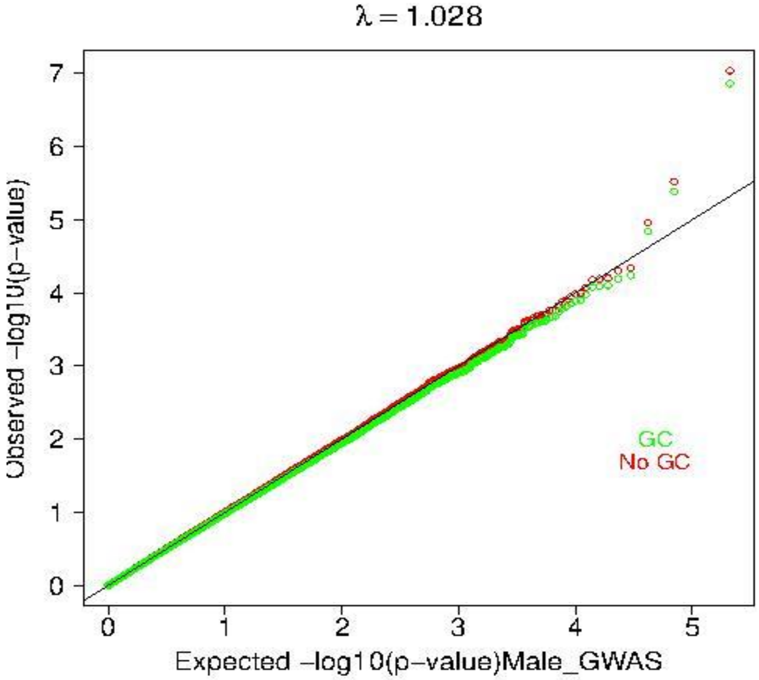


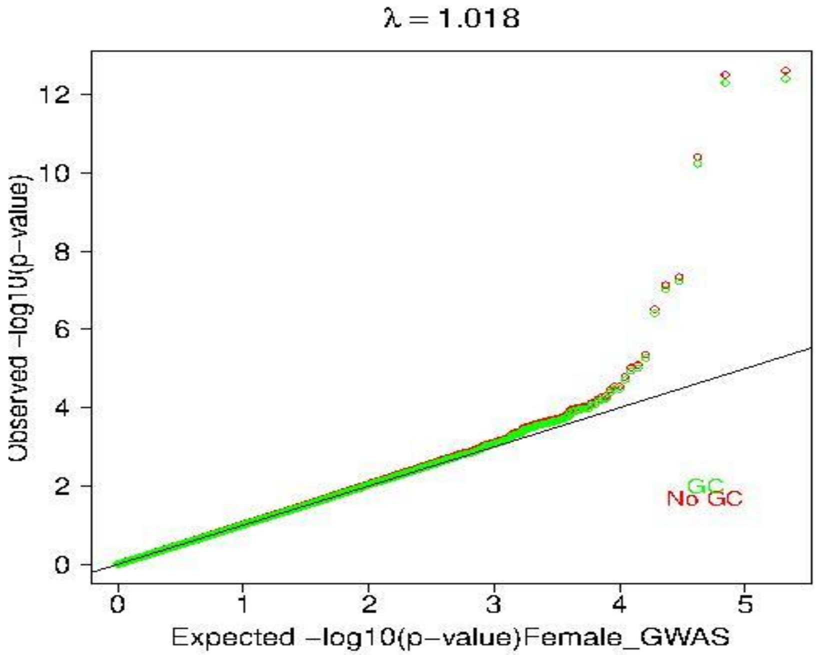
```
[jriyal86@scc-wl3 GWAS_AND_META]$ awk '$9 < 5e-8' gwas_females.assoc.logistic
   6   rs9267873   32199352    G          ADD       1492     0.5195          5.468   4.558e-08
   6    rs405875   32215188    A          ADD       1493    -0.6825         -7.288   3.149e-13
   6    rs532098   32578052    A          ADD       1487     0.6684          7.323   2.428e-13
   6   rs3873444   32682724    A          ADD       1493     -1.371         -6.604   4.013e-11
```

As seen in the Manhattan plot and terminal output, I identified four SNPs that surpassed the genome-wide significance threshold ($p < 5 \times 10^{-8}$). These SNPs showed moderate to large effect sizes, and the Manhattan plot revealed distinct peaks corresponding to their loci.



In contrast, the male-only GWAS did not yield any genome-wide significant SNPs. The top-ranked SNPs in males had weaker statistical signals, with p-values exceeding the genome-wide threshold. This likely reflects reduced power due to smaller sample size or potential sex-specific differences in genetic architecture.



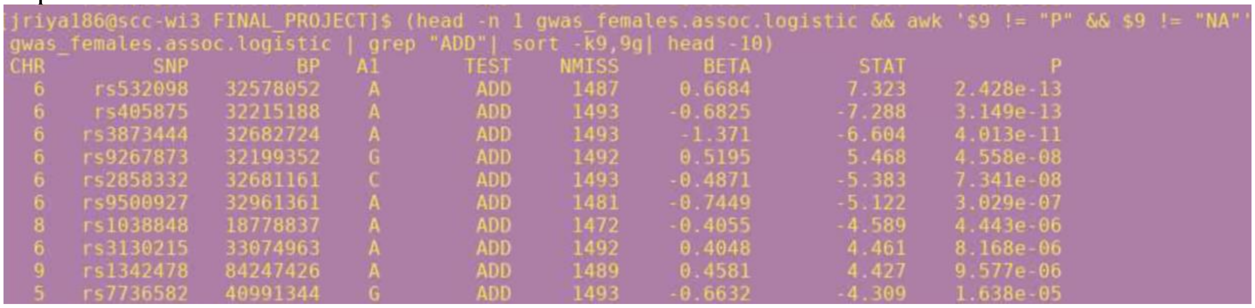


In the male-only GWAS, the QQ plot shows that both the unadjusted and GC-adjusted p-values closely follow the expected null distribution, with points aligning tightly along the diagonal. The genomic inflation factor (λ = 1.028) is very close to 1, indicating that population stratification and other confounders have been effectively controlled through the inclusion of principal components. Only a few SNPs deviate from the null expectation at the tail, which is expected in a well-calibrated GWAS with limited statistical power or few strong associations. Overall, the male analysis appears robust but underpowered, yielding no genome-wide significant hits.

In contrast, the female-only GWAS QQ plot reveals a clear upward deviation at the tail, highlighting the presence of truly associated SNPs. Despite this deviation, the λ value remains very close to 1 (λ = 1.018), suggesting that the analysis is well-calibrated and free from significant inflation. Multiple SNPs exhibit very small p-values, consistent with the detection of genome-wide significant variants and potentially stronger genetic signals in females. These findings align with the higher prevalence of rheumatoid arthritis in females and suggest the possibility of female-specific genetic contributions to disease susceptibility.

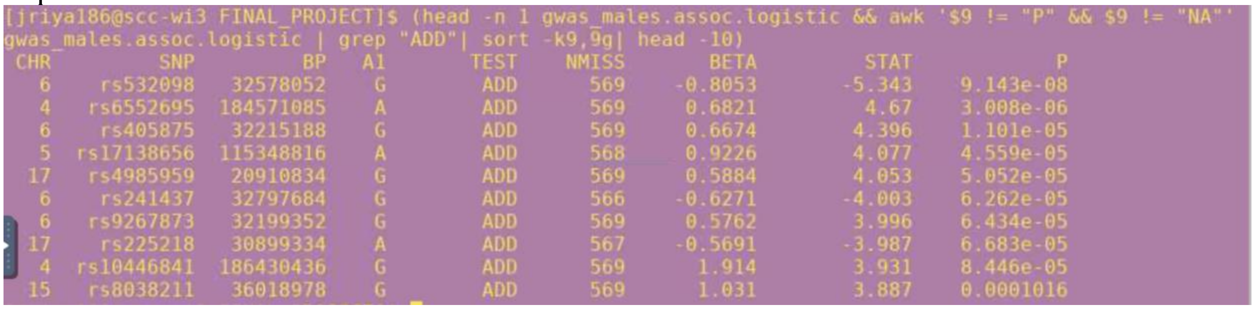
Below are 2 tables presenting the SNPs with the highest P- values in the male and female GWAS:

Top SNPs Females



```
[jriya186@scc-wi3 FINAL_PROJECT]$ (head -n 1 gwas_females.assoc.logistic && awk '$9 != "P" && $9 != "NA"'
gwas_females.assoc.logistic | grep "ADD"| sort -k9,9g| head -10)
 CHR         SNP         BP   A1       TEST    NMISS       BETA         STAT            P
   6    rs532098   32578052    A        ADD     1487     0.6684        7.323    2.428e-13
   6    rs405875   32215188    A        ADD     1493    -0.6825       -7.288    3.149e-13
   6   rs3873444   32682724    A        ADD     1493     -1.371       -6.604    4.013e-11
   6   rs9267873   32199352    G        ADD     1492     0.5195        5.468    4.558e-08
   6   rs2858332   32681161    C        ADD     1493    -0.4871       -5.383    7.341e-08
   6   rs9500927   32961361    A        ADD     1481    -0.7449       -5.122    3.029e-07
   8   rs1038848   18778837    A        ADD     1472    -0.4055       -4.589    4.443e-06
   6   rs3130215   33074963    A        ADD     1492     0.4048        4.461    8.168e-06
   9   rs1342478   84247426    A        ADD     1489     0.4581        4.427    9.577e-06
   5   rs7736582   40991344    G        ADD     1493    -0.6632       -4.309    1.638e-05
```

Top SNPs Males

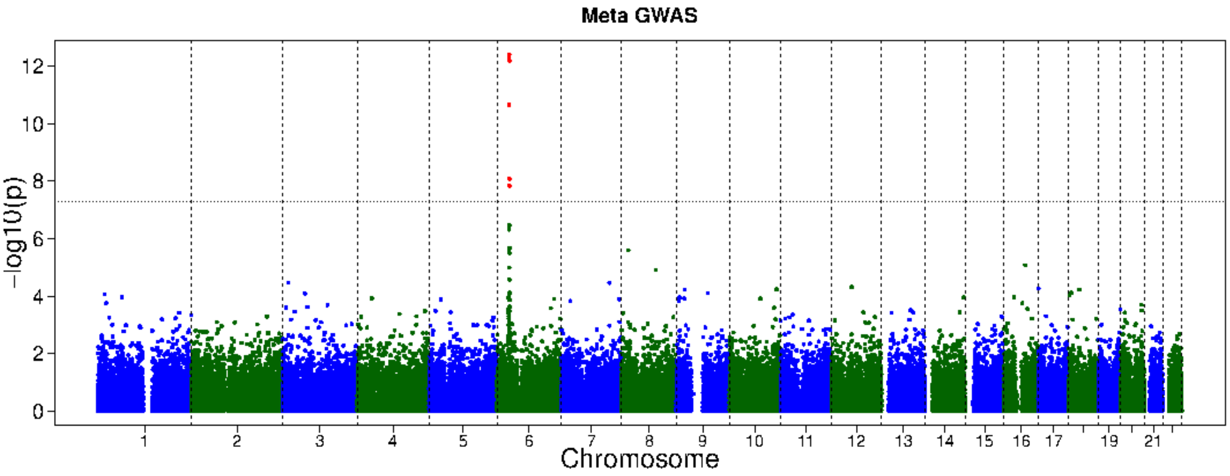


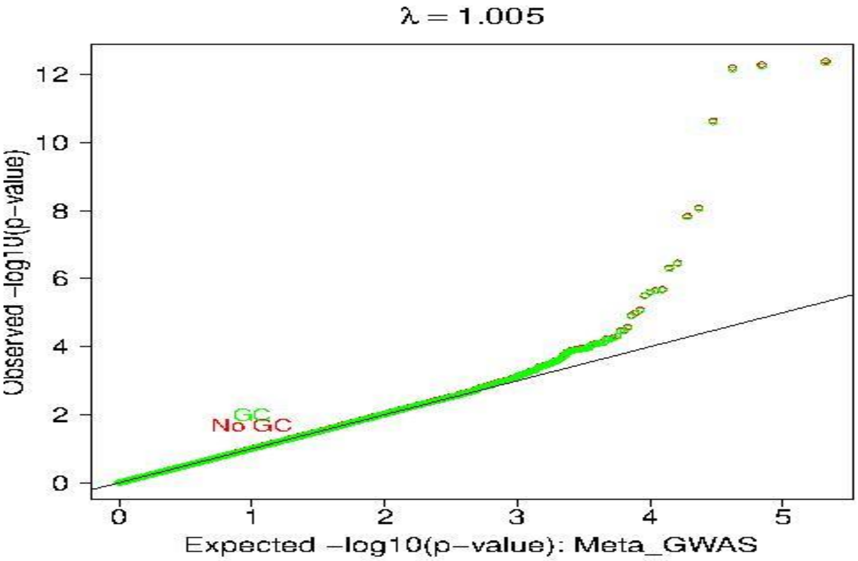
```
[jriya186@scc-wi3 FINAL_PROJECT]$ (head -n 1 gwas_males.assoc.logistic && awk '$9 != "P" && $9 != "NA"'
gwas_males.assoc.logistic | grep "ADD"| sort -k9,9g| head -10)
 CHR         SNP         BP   A1       TEST    NMISS       BETA         STAT            P
   6    rs532098   32578052    G        ADD      569    -0.8053       -5.343    9.143e-08
   4   rs6552695  184571085    A        ADD      569     0.6821         4.67    3.008e-06
   6    rs405875   32215188    G        ADD      569     0.6674        4.396    1.101e-05
   5  rs17138656  115348816    A        ADD      568     0.9226        4.077    4.559e-05
  17   rs4985959   20910834    G        ADD      569     0.5884        4.053    5.052e-05
   6    rs241437   32797684    G        ADD      566    -0.6271       -4.003    6.262e-05
   6   rs9267873   32199352    G        ADD      569     0.5762        3.996    6.434e-05
  17    rs225218   30899334    A        ADD      567    -0.5691       -3.987    6.683e-05
   4  rs10446841  186430436    G        ADD      569      1.914        3.931    8.446e-05
  15   rs8038211   36018978    G        ADD      569      1.031        3.887    0.0001016
```

Certain SNPs on chromosome 6 appear among the top-ranked associations in both male and female GWAS results, with rs532098 and rs405875 being particularly notable. In the female-only analysis, both SNPs reached genome-wide significance, while in the male-only analysis, they showed suggestive associations. Interestingly, both SNPs display opposite directions of effect between sexes, for instance, rs405875 showed a negative effect size in females but a positive one in males. This directional discordance may suggest sex-specific genetic effects, although it could also arise from strand inconsistencies or allele flipping. These findings underscore the importance of performing sex-stratified analyses, as some associations may be masked or misrepresented in pooled data.

b. Genome-wide meta-analysis

To prepare the METAL input files, I used an R script that merged the GWAS association results, BIM file, and allele frequency data. I filtered the PLINK results to include only additive tests with valid p-values, then calculated standard errors as BETA divided by the test statistic. I matched SNPs with their allele information from the BIM file and minor allele frequencies from the frequency file. After formatting the data to match METAL's required column names, I exported clean input files for each sex.
In METAL, I used the inverse variance-weighted scheme with standard errors (SCHEME STDERR), enabled genomic control, and ran ANALYZE HETEROGENEITY on the female and male input files. This produced a meta-analysis output file with combined effect sizes, p-values, and heterogeneity statistics.





The Manhattan plot of the meta-analysis reveals a strong peak on chromosome 6, where several SNPs surpass the genome-wide significance threshold ($p < 5 \times 10^{-8}$). This region is known to harbor immune-related genes frequently associated with RA. The accompanying QQ plot shows minimal genomic inflation ($\lambda = 1.005$), indicating that population stratification and confounding were well controlled. The upward deviation at the tail of the QQ plot suggests the presence of truly associated SNPs.

Concordance between Males and Females

Out of the six top significant SNPs identified on chromosome 6 in the meta-analysis, two SNPs (rs532098 and rs405875) were also among the top SNPs in both male and female GWAS. However, in both cases, the direction of effect differs between sexes with positive effect sizes in one sex and negative in the other. For instance, rs405875 showed a protective effect in females ($\beta = –0.68$) and a risk effect in males ($\beta = +0.67$). The Direction column in the meta-analysis table reflects this discordance, with entries like +? and -?, indicating missing or opposite directionality.

Despite the stronger associations observed in females, some overlap in genomic regions (especially on chromosome 6) was evident. The heterogeneity p-values for the top SNPs were all non-significant ($p > 0.28$), suggesting that the differences in effect size between sexes were not statistically significant, though biologically meaningful discordance remains possible.

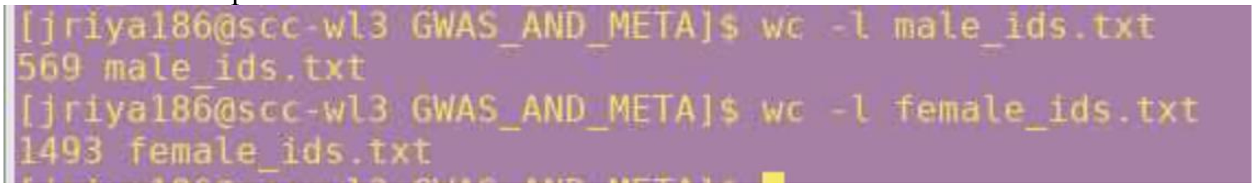
SNPs that cross the genome wide threshold in the meta-analysis:

Chromosome	**Position**	**MarkerName**	**Allele1**	**Allele2**	**Freq1**	**FreqSE**	**MinFreq**	**MaxFreq**	**Effect**	**StdErr**	**P**	**Direction**	**HetISq**	**HetChiSq**	**HetDf**	**HetPVal**
6	32681161	rs2858332	a	c	0.5443	0.0106	0.5281	0.5512	0.4342	0.0766	1.423E-08	++	11.5	1.130	1	0.2878
6	32578052	rs532098	a	g	0.4970	0.0000	0.4970	0.4970	0.6684	0.0921	3.95E-13	+?	0.0	0.000	0	1
6	32199352	rs9267873	a	g	0.5383	0.0100	0.5318	0.5536	-0.5365	0.0802	2.243E-11	--	0.0	0.105	1	0.746
6	32682724	rs3873444	a	g	0.0806	0.0039	0.0780	0.0864	-1.2497	0.1738	6.381E-13	--	7.0	1.075	1	0.2997
6	32961361	rs9500927	a	g	0.1278	0.0023	0.1241	0.1293	-0.7161	0.1242	8.242E-09	--	0.0	0.136	1	0.7123
6	32215188	rs405875	a	a	0.4916	0.0000	0.4916	0.4916	-0.6825	0.0945	5.101E-13	-?	0.0	0.000	0	1

Limitations

There are a few limitations to this analysis. First, the male GWAS had a smaller sample size, which likely reduced its power to detect significant associations and contributed to some missing effect directions in the meta-analysis. Second, although genomic control and principal component adjustment were applied, population structure or technical artifacts could still influence effect estimates. Third, the meta-analysis used a fixed-effects model, which assumes a shared effect across sexes and may underestimate true heterogeneity. Finally, discordant alleles (e.g., same SNPs with different coded alleles) across sexes could also impact directionality if not harmonized correctly.

Difference in sample size:



```
[jriya186@scc-wl3 GWAS_AND_META]$ wc -l male_ids.txt
569 male_ids.txt
[jriya186@scc-wl3 GWAS_AND_META]$ wc -l female_ids.txt
1493 female_ids.txt
```

Question 3: b. PRS analysis

Polygenic risk scores (PRS) were calculated using PRSice, a software tool designed for high-throughput PRS calculation and clumping. The base summary statistics were derived from a RA genome-wide association study conducted in a European cohort by Okada et al. This file contained odds ratios (ORs) as effect sizes, and p-values labeled under the column header P-val. Since PRSice requires the p-value column name to be free of special characters such as hyphens, the column header P-val was renamed to P using the sed command in Bash. This ensured compatibility with the --pvalue flag in PRSice. The edited file was saved as a new version to preserve the original data.

The target genotype data consisted of LDpruned PLINK binary files, and phenotypic covariates were supplied via a separate file (narac_pcs.txt). Covariates PC1, PC2, and PC4 were included in the analysis. The phenotype was binary, indicating case-control status, and this was specified using the --binary-target T flag.

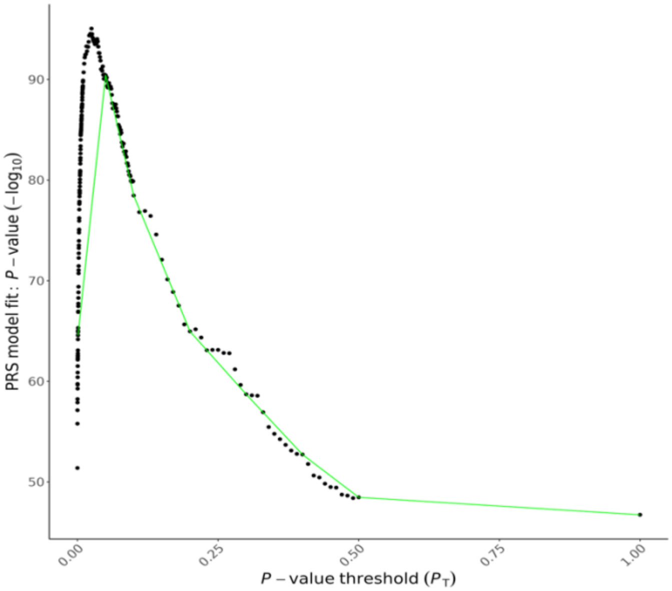
Summary Table:

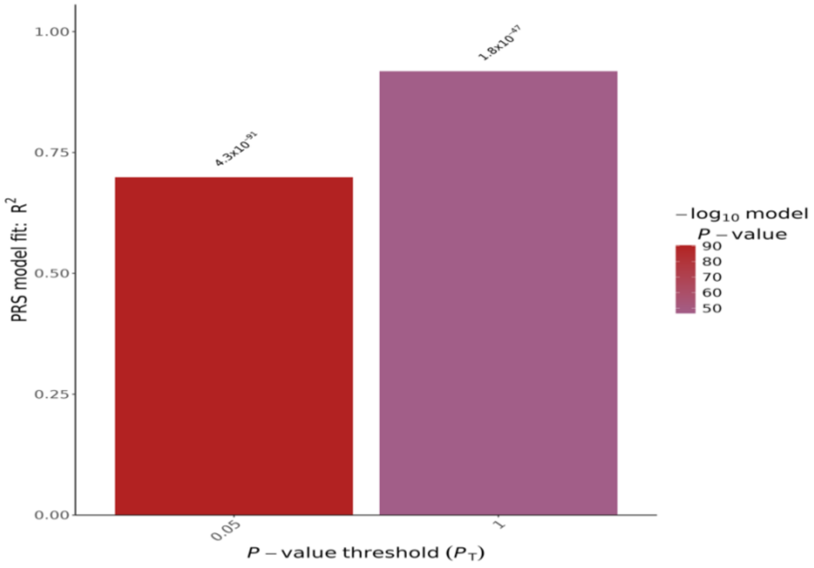
Phenotype: -
Set: Base
Threshold: 1
PRS R: 0.918221
Full R^2: 0.947002
Null R^2: 0.351941
Prevalence: -
Coefficient : 87,846.2
Standard Error: 6,069.99
P-value : 1.82×10^{-47}
Number of SNPs: 69,776
Empirical P-value :0.000999001

The polygenic risk score (PRS) was computed using all SNPs with p-values below the threshold of 1 from the GWAS summary statistics. At this threshold, 69,776 SNPs were included in the score calculation. The PRS alone explained 91.8% of the variation in the phenotype (PRS.R^2 = 0.918), and the full model including both the PRS and covariates explained 94.7% of the variance (Full $R^2 = 0.947$). In contrast, the covariates alone (Null model) explained only 35.2% (Null $R^2 = 0.352$). The association was statistically significant with a p-value of 1.82×10^{-47} and an empirical permutation-based p-value of 0.000999, confirming that the result is unlikely due to chance. The regression coefficient was extremely large (87,846.2) with a relatively small standard error (6,069.99), suggesting a strong linear relationship between the PRS and the phenotype.

While the statistical output appears extremely significant, the PRS.R^2 value of 91.8% is suspiciously high and raises concerns about the validity of the result. In most real-world polygenic risk score studies, especially those involving complex traits, PRS typically explains only a modest portion of the variance, often below 10%, and rarely above 30% even for highly heritable traits like height. A variance explained greater than 90% is biologically implausible in most human phenotypes and indicates potential technical or methodological issues.

A possibility is that the base and target datasets are not independent, meaning there could be sample overlap leading to inflated predictive accuracy due to overfitting. Additionally, including all SNPs at a threshold of 1 may incorporate substantial noise and bias into the PRS model, which can artificially boost R^2 values if not properly regularized. Lastly, improper control for population stratification, even with principal components, can sometimes leave residual structure that inflates associations.





1.00
0.75
0.50
0.25
0.00
PRS model fit: R^2
$4.3x10^{-91}$
$1.8x10^{-47}$
0.05
1
$P-$value threshold (P_T)
$-\log_{10}$ model $P-$value
90
80
70
60
50