North American Rheumatoid Arthritis Consortium (NARAC) Rheumatoid arthritis GWAS data

PART 1

Perform genetic data cleaning of the NARAC GWAS data. Then, perform PCA on the data to identify study outliers, and create a set of PCs that can be used in association analyses. In your write up, state, and justify the analyses you did and in what order, and how many individuals and SNPs you removed and retained at each step. Provide your recommendations on which PCs to include in case-control GWAS analyses and explain your choice. Be sure to state the hypotheses you are testing, and the methods you use (including any parameters you must choose). Present the results and your conclusions clearly.

Hypothesis

The hypothesis central to this analysis posits that underlying population substructure within the North American Rheumatoid Arthritis Consortium (NARAC) GWAS dataset significantly influences the rheumatoid arthritis (RA) phenotype expression among the study participants. This substructure, if present, may confound genetic associations unless adequately accounted for, potentially leading to spurious results in genome-wide association studies (GWAS). By identifying and adjusting for these substructures through PCA, we seek to clarify the genuine genetic contributions to RA without the interference of population-based genetic differences.

The first part of this report starts with a detailed account of the quality control (QC) and principal component analysis (PCA) applied to the NARAC GWAS dataset aimed at studying genetic factors in rheumatoid arthritis. Ensuring data integrity through meticulous QC and adjusting for population stratification with PCA are crucial steps in preparing for robust GWAS. PLINK 1.90b6.27 was utilized throughout this project. The dataset comprised 544,276 genetic variants and 2,062 individuals (569 males and 1,493 females).

Quality Control

The first QC step removed variants with a minor allele frequency (MAF) of less than 0.01. This threshold ensures that the analysis includes only variants that are common enough to provide statistically meaningful results and reduces the chance of false-positive associations due to rare variants. The second step excluded variants with more than 5% missing genotype data to minimize genotyping errors that could distort association signals. Individual samples with more than 5% missing genotypes were excluded to ensure high data completeness. The Hardy-Weinberg Equilibrium (HWE) test excluded variants deviating from equilibrium in controls at a p-value threshold of 1e-6, as deviations could indicate technical errors or population stratification.

```
plink --bfile $PLINKFILE --maf 0.01 --geno 0.05 --mind 0.05 --hwe 1e-6 --make-bed --out $OUTPUTDIR/narac_filtered
```

The order of PLINK operation performed is individuals (missing genotypes) -> SNPs (missing genotypes)- > HWE-> MAF. After filtering, 18,398 SNPs were removed due to missing genotype data, 663 SNPs were eliminated for failing the HWE test, and 22,913 SNPs were excluded due to low MAF. One individual was removed for failing the sex check, leaving 502,302 variants and 2,061 individuals (867 cases and 1,194 controls). 0 people removed due to missing genotype data (--mind). The overall genotyping rate among retained individuals was 0.992688, indicating a high level of data completeness. This step ensures that subsequent analyses are based on high-quality data. Further, as maximizing sample size is not important, applying the filter on individuals last is not required to yield a bigger sample. Further since we don't want to keep individuals whose data are not optimal the default order used in this project is simpler to implement.

Linkage Disequilibrium (LD) Pruning

Data cleaning was followed by LD pruning, it reduced redundancy in the dataset by removing correlated SNPs that could artificially inflate association signals. The pruning procedure takes all SNPs within 10,000 kilobases of the start of the chromosome, finds all pairs of variants that are correlated with squared correlation r2>0.2, and removes one from each of those pairs. It then slides over 1 marker and does the same thing. In the end, we are left with a set of markers where none have squared correlation > 0.2. This step is crucial for obtaining accurate PCA results, as correlated SNPs could skew the analysis, leading to misinterpretation of population structure.

```
plink --bfile $OUTPUTDIR/narac_filtered --indep-pairwise 10000kb 1 0.2 --out $OUTPUTDIR/narac_pruned_data
```

Following SNP filtering, LD pruning was performed to reduce redundancy in the genetic data. The goal of LD pruning is to remove correlated SNPs that could inflate the association signals. We used a window size of 10000kb, a step size of 1 SNP, and a variance inflation factor (VIF) threshold of 0.2, which significantly reduced the number of SNPs, ensuring that the retained SNPs are relatively independent. 394617 of 502302 redundant variants were removed, resulting in a set of relatively independent markers. This step is crucial for the PCA to accurately reflect the underlying population structure without interference from LD.

```
plink --bfile $OUTPUTDIR/narac_filtered --extract $OUTPUTDIR/narac_pruned_data.prune.in --make-bed --out $OUTPUTDIR/narac_pca
smartpca -p $OUTPUTDIR/smartpca.par > $OUTPUTDIR/smartpca.log
```

After LD pruning, PCA was conducted to identify the principal components (PCs) that capture the primary axes of genetic variation within the dataset. SMARTPCA from the EIGENSOFT package was used to compute the PCA with the following parameters:

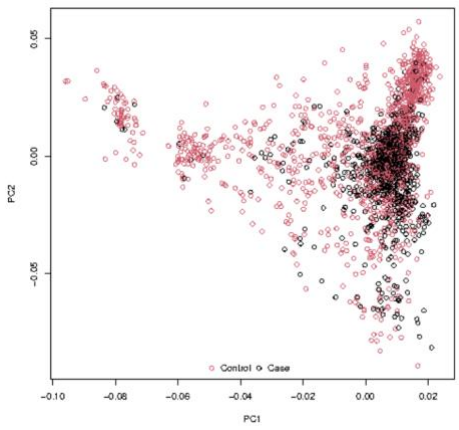
- genotypename: narac_pca.bed
- snpname: narac_pca.bim
- indivname: narac_pca.fam
- evecoutname: narac_pca.evec
- evaloutname: narac_pca.eval
- altnormstyle: NO
- numoutevec: 10
- numoutlieriter: 0

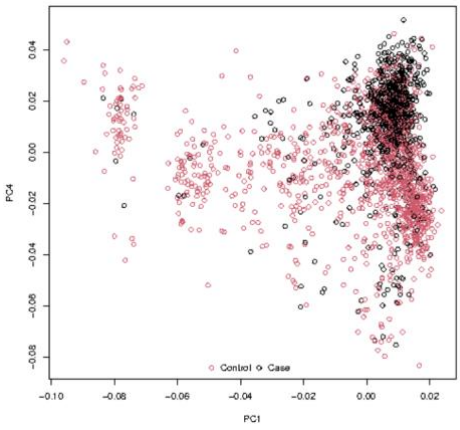
After looking at the log file these are the significant p values (i.e., less than 0.005):

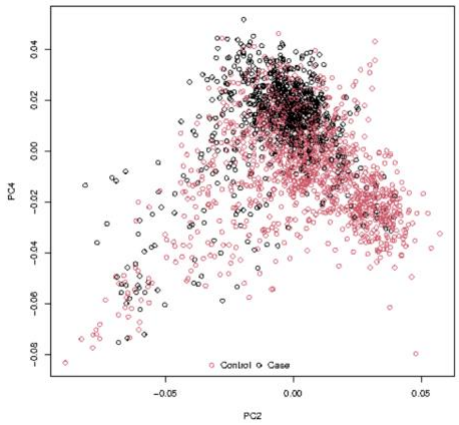
- **PC1**:
  - Mean difference between controls (-0.004) and cases (0.006).
  - ANOVA p-value: ++++ (possibly too small to be presented)
- **PC2**:
  - Mean difference between controls (0.004) and cases (-0.006).
  - ANOVA p-value: 2.22045e-15.
- **PC4**:
  - Mean difference between controls (-0.008) and cases (0.010).
  - ANOVA p-value: 3.33067e-16.

The first, second, and fourth PCs significantly differentiate between RA cases and controls, as their p-values are highly significant. These PCs likely capture the genetic variation associated with RA and/or population stratification. Given these results, PCs 1, 2, and 4 appear to be particularly informative for differentiating between cases and controls. Since these components show significant differences in their means between the groups and have highly significant p-values, plotting these would be insightful for visualizing the population structure and its impact on RA within your dataset.

Plots of PC1 vs PC2, PC1 vs PC4 and PC2 vs PC4 -







After visually comparing the plots there doesn't seem to be any major differences in the clusters of points. The two plots are similar in that they have a large group of points that are bunched together in one corner, and then points especially control i.e. in red that are spread out from the main bunch on either x or y axis. Both also have points that don't bunch with the rest, that have middle values for all the three PCs. There doesn't seem to be any singular significant outliers that deviate significantly from the data points present.

In summary for the first part these analyses were carefully designed to identify genetic variation patterns that may affect RA development or identify subpopulations prone to RA. Removing rare and error-prone variants ensures high data quality, which is essential for GWAS robustness. LD pruning eliminates redundancy, ensuring that PCA accurately identifies genuine population structure without interference from correlated markers. PCA results provided visual confirmation of the differentiation between cases and controls. Plotting combinations of PC1, PC2, and PC4 revealed clustering patterns that suggest genetic stratification associated with the RA phenotype. These PCs, particularly PC1, PC2, and PC4, are recommended for use as covariates in future GWAS to control for population stratification, thereby reducing confounding effects and enabling more accurate identification of genetic associations with RA.

PART 2

It is well known that the HLA region on chromosome 6p21 plays an important role in RA. It is also well known that females are affected by RA much more frequently than males. Your goals are to determine if there are additional genomic regions (in addition to the HLA region on chromosome 6) that are associated with RA in females in this sample, and to determine whether any of the regions are sex specific. For all analyses, be sure to state and justify the significance criteria you use.

Since we will be using or taking into consideration the PCs from question 3 as covariates for future association analyses. We will need to re-format the ".evec" file to be in the format plink

requires, where the first two columns are Family id and individual D, and the rest of the columns are covariates, and all columns are whitespace delimited. The ':' between family ID and IID should be removed, and a header should be added in the ".evec" file so that plink can read it.

a. Perform two genome-wide association analyses for rheumatoid arthritis: one using only female subjects, and one using only male subjects. Explain how you chose covariates, and how you accounted for population structure (or, if you chose not to account for population structure, justify your decision). Present a written summary of your results with appropriate plots and tables that describe your findings.

To identify genomic regions associated with RA in male and female subjects, separate genome-wide association analyses (GWAS) were conducted for each sex. The analyses were tailored to adjust for population structure using PCs identified through previous PCA. The findings from these sex-specific analyses shed light on potential genetic differences in RA susceptibility between males and females.

Before performing genome-wide association analyses narac_pca_covar.txt generated as an output from smartpca was used to determine which PCs are associated with case status.

```
plink --bfile $OUTPUTDIR/narac_filtered --covar $OUTPUTDIR/narac_pca_covar.txt --covar-name PC1-PC10 --out
$OUTPUTDIR/checkPCs --allow-no-sex --logistic no-snp beta
```

Selecting Covariates

The significant principal components (PC1, PC2, and PC4) were identified from the output of **smartpca** and confirmed through logistic regression to be associated with case status (RA vs. control) with p-values below 0.005. These PCs were therefore chosen as covariates for adjusting population structure in subsequent GWAS analyses. Specifically, PC2 and PC4 were included in the logistic regression models for both male and female analyses.

The PCs with significant p values (i.e., less than 0.005) are below:

TEST	NMISS	BETA	STAT	P
PC1	2061	34.46	8.951	3.517e-19
PC2	2061	-28.38	-10.47	1.161e-25
PC4	2061	46.9	16.72	9.046e-63

These p-values confirm that the chosen PCs capture significant population stratification and are therefore essential for inclusion in subsequent analyses.

GWAS for females:

The GWAS conducted for females included only female participants, and significant PCs were used as covariates. The logistic regression model adjusted for PC2 and PC4 to account for population structure.

CHR	SNP	BP	A1	TEST	NMISS	BETA	SE	L95	U95	STAT	P	ALLELE1	ALLELE2	FREQ
1	rs3094315	752566	G	ADD	1484	-0.14	0.1209	-0.3769	0.09685	-1.159	0.2466	G	A	0.1672
1	rs12562034	768448	A	ADD	1455	-0.132	0.1426	-0.4115	0.1474	-0.9261	0.3544	A	G	0.1076
1	rs3934834	1005806	A	ADD	1463	-0.1396	0.1175	-0.3699	0.09058	-1.189	0.2345	A	G	0.1532
1	rs9442372	1018704	A	ADD	1114	-0.6373	0.1069	-0.8468	-0.4279	-5.964	2.463e-09			

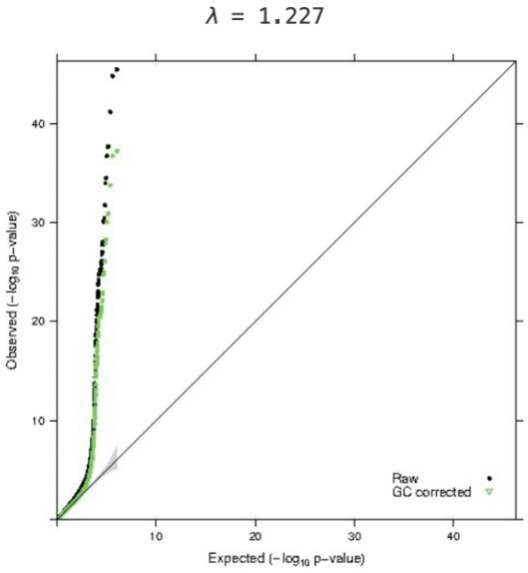
A total of 517068 SNPs have p<0.0001 in female GWAS. Above represents the first few lines of the GWAS_female.assoc.logistic output. Since we might need additional allele and frequency information for further parts such as meta-analysis those columns were also included through awk commands.

```
# Example command for GWAS (female subjects only) using significant PCs as covariates with the addition of
frequency and allele 1 and allele 2 columns
plink --bfile $PLINKFILE --filter-females --covar $OUTPUTDIR/narac_pca_covar.txt --covar-name PC2,PC4 --logistic
beta hide-covar --ci 0.95 --out $OUTPUTDIR/GWAS_female
awk 'BEGIN {FS=" "; OFS=" "}
    NR==FNR {a[$2]=$3; b[$2]=$4; freq[$2]=$5; next}
    FNR==1 {print $0, "ALLELE1", "ALLELE2", "FREQ"}
    FNR>1 {if ($2 in a) print $0, a[$2], b[$2], freq[$2]; else print $0, "NA", "NA", "NA"}' $OUTPUTDIR/allele_freqs.frq
$OUTPUTDIR/GWAS_female.assoc.logistic > $OUTPUTDIR/GWAS_female_ready.assoc.logistic
```

A similar procdedure was followed in the male-only GWAS analysis. PC2 and PC4 were again used as covariates to account for population structure and a total of 515084 SNPs have p<0.0001 in male GWAS.

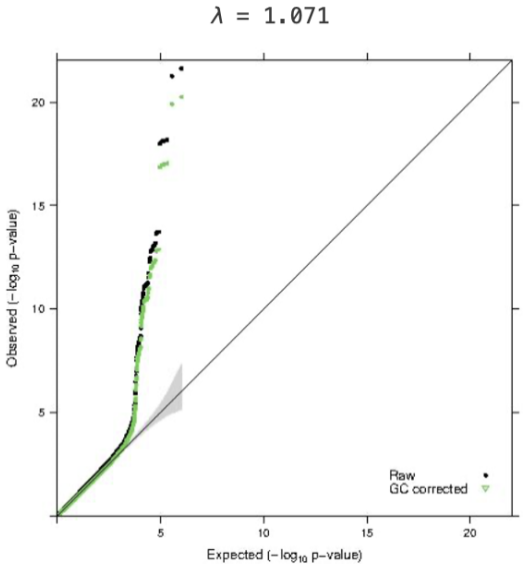
QQ and Manhattan plots

Female QQ plot -



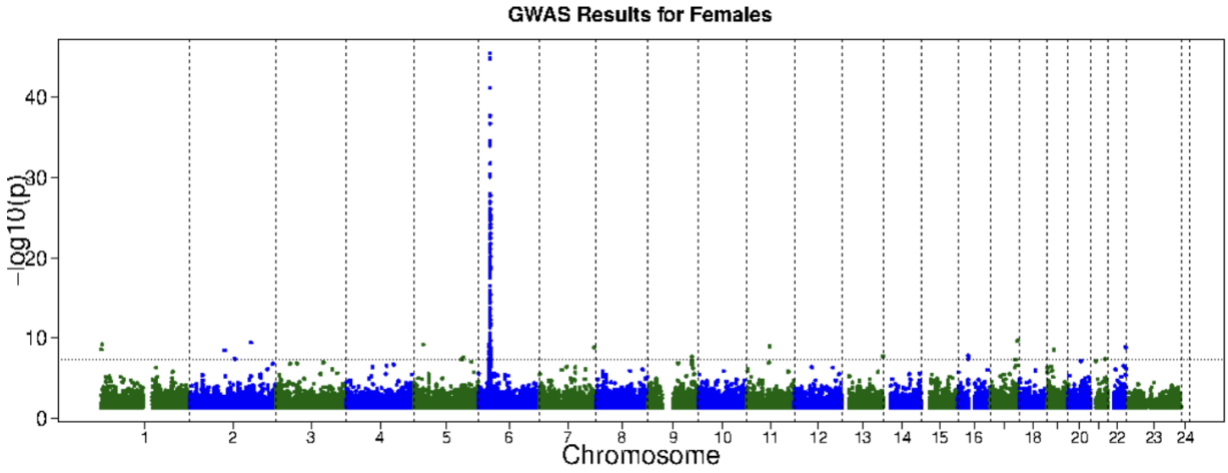
- Lambda (λ) = 1.227: This value indicates some genomic inflation in the female dataset, possibly due to population stratification or other biases. However, the genomic control-corrected p-values closely follow the raw values, suggesting consistency.
- Pattern: Significant p-values deviate from the expected distribution at the tail end supporting the presence of variants whether in chromosome 6 or other regions.

Male QQ plot -



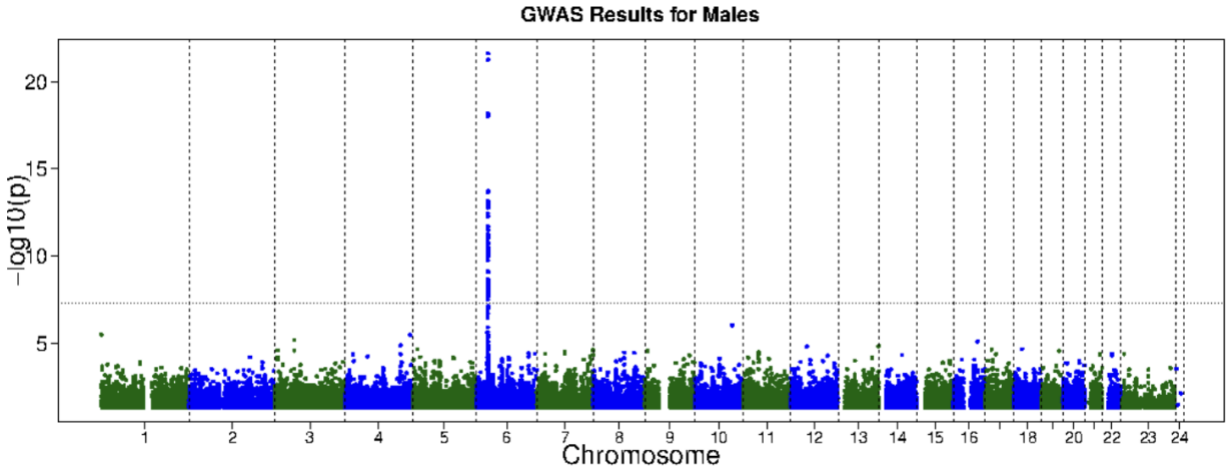
- Lambda (λ) = 1.071: This value is close to 1, indicating minimal inflation and suggesting that population stratification or other biases are not significantly affecting the results.
- Pattern: The observed p-values follow the expected distribution, however like the female plot there are deviations especially at the tail end although not extensive as in the female plot.

Female Manhattan Plot -



There is spike of variants that stack on top of each other for all the chromosomes. As expected, based on the QQ plots there are some points that approach genome wide significance other than at chromosome 6 for female and none for male. The plot visually shows peaks in multiple chromosomes other than chromosome 6, indicating genomic regions that may play a role in RA for females. Particularly chromosomes 1, 2, 5, 7, 9, 11, 13, 16, 17, 19, and 22 all showed significant association with RA, as evidenced by the presence of SNPs above the genome-wide significance threshold.

Male Manhattan Plot –



The plot illustrates a clear, singular peak for chromosome 6, while other chromosomes do not reach genome-wide significance.

The results from these sex-specific GWAS analyses provide insights into potential sex-specific associations in RA. The data indicate that females exhibit multiple genomic regions associated with RA beyond chromosome 6, while males only show significant associations in the HLA region. These findings support the hypothesis that additional genomic regions contribute to RA susceptibility in females specifically, potentially due to biological differences in immune function or other sex-specific factors. However, given that many significant associations in females consist of only a few SNPs above the threshold, further replication and analysis are warranted to confirm these observations. Larger sample sizes could increase the reliability of these findings and aid in understanding the underlying genetic mechanisms in RA.

b. Perform a genome-wide meta-analysis that combines the male-only and female-only results with a test for heterogeneity. Present a written summary of your findings, including appropriate plots and tables, and be sure to address the questions: do males and females show association in the same regions? Do the significantly associated SNPs appear to have similar effects in males as in females? Discuss the limitations of the data and the methods you used.

A genome-wide meta-analysis was conducted to combine the results from separate GWAS for male and female subjects to explore potential sex-specific genetic differences in RA. This meta-analysis aimed to identify genomic regions that are associated in both sexes and to examine whether the effects of significantly associated SNPs are similar across sexes. Using the METAL software, a framework to assess heterogeneity between the studies, which is crucial for identifying SNPs with consistent or disparate effects between genders.

Meta-Analysis

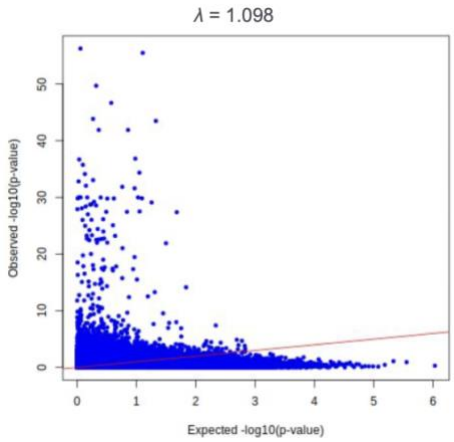
The analysis setup involved aggregating effect sizes based on their standard errors, adjusting p-values to handle potential inflation from population stratification, and calculating allele frequencies. It also tested for differences in effect sizes between genders, indicative of sex-specific genetic effects. The meta-analysis effectively merged data from male and female GWAS, ensuring that markers, alleles, effect sizes, standard errors, and other relevant data were uniformly processed, resulting in a comprehensive dataset of 534,786 markers and the smallest p-value being 5.556e-57 at marker 'rs660895'.

The overall contributions from the studies incorporated into the meta-analysis appear well, as evidenced by the minimal occurrences of missing values in the direction column, typically indicated by a "?". This suggests comprehensive data reporting across studies. However, there are signs of inconsistency in the effect directions (e.g., "+-" symbols in some variants), indicating divergent effects across the studies.

Upon filtering the results for variants with genome-wide significant p-values (i.e., $p < 5e-8$):

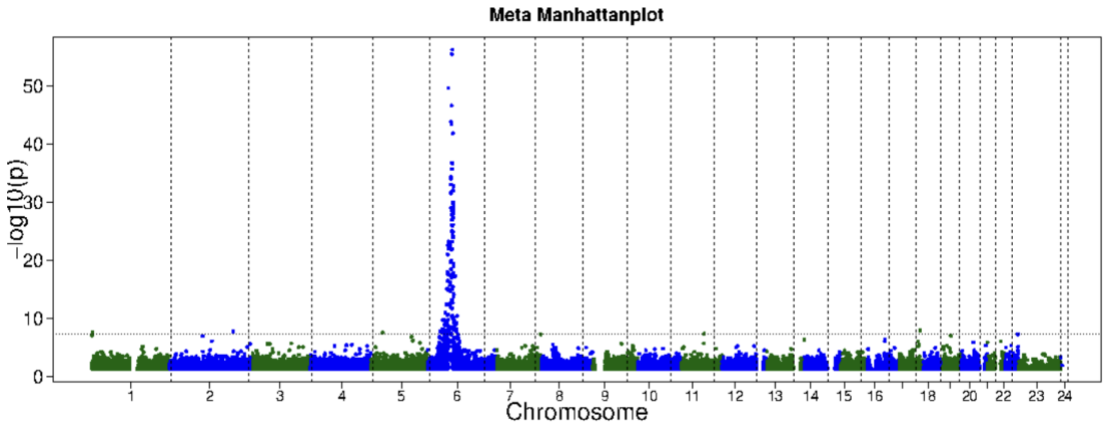
The heterogeneity p-values (HetPVal) for most SNPs do not show significance, suggesting consistent effect sizes across the male and female datasets. Yet, a few SNPs exhibit mild heterogeneity (HetISq values greater than 0), implying some variation in effect sizes between genders. This observation warrants further investigation into potential sex-specific genetic influences on RA. The Effect column generally reflects the beta coefficients, providing insight into the magnitude and direction of the genetic effects across different datasets. The Direction column, which uses symbols like "++" and "--," further confirms whether these effects are consistently positive or negative, respectively, across studies. While chromosome 6, particularly the HLA region, harbors the majority of significant SNPs, notable associations on chromosomes 1, 2, 7, 11, 17, and 22 also emerge. These findings underscore potential novel genomic regions contributing to RA susceptibility beyond the well-characterized HLA locus. Interestingly, SNPs on non-chromosome 6 locations often exhibit a difference in their direction, suggesting possible sex-specific genetic effects. Key examples include rs788160 (CHR 2), rs2493291 (CHR 1), rs10501396 (CHR 11), rs11868709 (CHR 17), rs11704238 (CHR 22), and rs11762043 (CHR 7).

Meta analysis QQ plot -



The QQ Plot assesses inflation or bias in the meta-analysis by comparing observed p-values to expected distributions. Significant p-values deviate from the expected distribution at the tail end supporting the claim from GWAS that there are variants present whether it is in chromosome 6 or other regions.

Meta analysis Manhattan plot -



The Manhattan plot visualizes the results across chromosomes and highlights the loci that show genome-wide significance in the meta-analysis. The Manhattan plot as well proves the hypothesis that there are some significant regions other than chromosome 6 particularly at chromosomes 1,2,7,11,17,22 that cross the threshold line again proving the hypothesis and supporting the results that were obtained after filtering the metal analysis results.

Do Males and Females Show Association in the Same Regions?

The meta-analysis clearly highlighted that both sexes share significant associations in the well-established HLA region on chromosome 6. This shared association reinforces the region's pivotal role in RA across both genders. Beyond chromosome 6, the meta-analysis showed that while certain SNPs on chromosomes 1, 2, 7, 11, 17, and 22 are associated with RA they seem sex specific, the degree and consistency of these associations vary. Some loci displayed strong associations in one sex but not the other, hinting at potential sex-specific genetic influences.

Do Associated SNPs Have Similar Effects in Both Sexes?:

The heterogeneity p-values (HetPVal) indicated that most SNPs show similar effect sizes across males and females, evidenced by non-significant heterogeneity values. However, a few SNPs did exhibit significant heterogeneity, implying that the effect sizes differ between genders. For many SNPs, the direction of effect (represented in the Direction column) remains consistent across both genders, but certain loci show opposing effects (e.g., "+-" or "-+"). These differences in direction may highlight sex-specific genetic interactions or differences in RA risk mechanisms.

Limitations and Future Directions

- Population Stratification: Although genomic control was applied, residual stratification might still influence the results.
- Sample Size and Power: Differences in sample sizes between genders could affect the detection power of associations.
- Functional Validation: Further functional studies are required to understand the biological implications of sex-specific associations.

- Replication and Validation: Independent replication in diverse cohorts is crucial to validate these findings and enhance understanding of the genetic architecture of RA.

In the initial GWAS conducted separately on male and female samples, multiple chromosomes outside of the well-studied chromosome 6 were identified as significant in influencing RA. However, subsequent meta-analyses, which combined and compared results from the male-only and female-only GWAS, highlighted chromosomes 1, 2, 7, 11, 17, and 22 as particularly significant. These findings suggest that the identified genomic regions on these chromosomes may be relevant and potentially associated with RA predominantly in females, indicating a sex-specific genetic influence. This conclusion supports the hypothesis that certain genetic factors contributing to RA may vary between sexes, necessitating further investigation into these chromosomal regions to understand their specific roles in the pathogenesis of RA in females compared to males.

PART 3

Use LD score regression and the Okada et al summary statistics to 1) estimate the heritability of RA, and 2) compare the heritability in the Asian and European populations. Describe your methods (including all assumptions you've made) and present and explain your results.

To evaluate the heritability of RA and compare it between Asian and European populations, LD score regression was applied using summary statistics from the study by Okada et al. The summary statistics were preprocessed with a log transformation to normalize effect sizes, and total sample sizes were calculated by adding cases and controls together for each population.

For the European ancestry analysis, the process involved preparing the summary statistics with the munge_sumstats.py tool, specifying SNPs, alleles, signed summary statistics as log-transformed odds ratios, p-values, and the total sample size, which amounted to 58,284. The LD score regression was then executed using the ldsc.py script, referencing LD scores tailored for the European population. The European data exhibited a heritability estimate on the observed scale of 0.1422, with a standard error of 0.0238. This value reflects a moderate genetic contribution to RA susceptibility within this population. The Lambda GC value close to 1 (1.0466) and the intercept (0.9621) near 1 indicate that the inflation from population stratification or confounding biases is minimal, which supports the robustness of these results. The mean Chi-Square value (1.1361) further suggests that the genetic signals are reliable and not overly inflated by potential confounders.

Similarly, for the Asian ancestry analysis, the same preprocessing steps were followed, but with a total sample size of 22,515. LD score regression analyses were separately conducted for East Asian, Central/South Asian, and Middle Eastern groups using corresponding LD score

references. This approach aimed to capture potential regional genetic heterogeneities within the broader Asian category. Across different Asian subpopulations, the heritability estimates are slightly higher than those observed in Europeans. These differences point to a possibly greater genetic influence on RA within these populations or differences in environmental interactions with genetic factors. The highest observed-scale heritability among the Asian groups is in East Asians ($h^2 = 0.1773$), which could indicate unique genetic risk factors prevalent in this group or a more homogeneous genetic structure that amplifies the detectable genetic signal. Central/South Asian and Middle Eastern show heritability estimates ($h^2 = 0.1627$ and 0.1731, respectively) that are similar to each other but distinctly higher than the European estimate. This might reflect shared genetic backgrounds or environmental factors that are more common in these regions compared to Europe.

From the results of the LD score regression analysis, several key conclusions can be drawn: The variation in heritability estimates across populations underscores the complex interplay between genetic and environmental factors in the pathogenesis of RA. These findings highlight the importance of considering population-specific genetic architectures when studying complex diseases like RA. The differences in heritability suggest that certain risk alleles or genetic factors might be more prevalent or have different effects in specific populations. This necessitates targeted genetic studies within diverse populations to uncover these unique genetic risk factors. Understanding the variability in genetic susceptibility can aid in developing more personalized approaches to treatment and management of RA, potentially leading to more effective interventions based on genetic risk profiles. These findings should motivate expanded genetic research in underrepresented populations to ensure that the genetic data used in research and clinical settings reflect the diversity of the global population. Additionally, healthcare strategies could be optimized by incorporating genetic screening and risk assessment tools that account for population-specific genetic variations in RA.

Assumptions and Considerations:

- Genetic Homogeneity within Groups: This analysis assumes relatively homogeneous genetic backgrounds within the European and Asian groups, despite known sub-population structures, which could influence the results.
- Effect Sizes Translation: The log transformation of odds ratios assumes that the effect sizes are normally distributed, facilitating their use in LD score regression.
- Population-Specific LD Scores: LD scores specific to each population group were used, assuming these scores accurately reflect the linkage disequilibrium structure in each population.

Limitations:

- Population Stratification: While lambda GC values were slightly above 1, indicating minimal inflation, there remains a risk that population stratification could affect the heritability estimates.
- Sample Size Variability: The significantly larger sample size for Europeans compared to Asians could lead to more stable and reliable heritability estimates in the European group.

- Regional Differences within Asians: Combining different Asian subgroups under a broader regional categorization might mask finer-scale genetic differences influencing RA heritability.

Overall, this LD score regression analysis has provided valuable insights into the heritable nature of RA and highlighted slight differences in genetic susceptibility between European and Asian populations. Further studies using larger and more genetically diverse Asian samples would be beneficial to confirm these findings and refine the understanding of RA's genetic underpinnings.

APPENDIX A

Bash script:

```
#!/bin/bash

module load R
module load eigensoft

# Load the necessary PLINK module
module load plink/1.90b6.27

# Define input and output directories
DATADIR="/projectnb/bs859/data/RheumatoidArthritis/final_project"
OUTPUTDIR="/projectnb/bs859/students/bkapalli/A.project"

# List files in the data directory for verification
echo "Listing all files in the data directory:"
ls -lh $DATADIR

# Define PLINK binary file prefix
PLINKFILE=$DATADIR/narac_hg19

wc -l $DATADIR/narac_hg19.bed
# 754504 /projectnb/bs859/data/RheumatoidArthritis/final_project/narac_hg19.bed
wc -l $DATADIR/narac_hg19.bim
# 544276 /projectnb/bs859/data/RheumatoidArthritis/final_project/narac_hg19.bim
wc -l $DATADIR/narac_hg19.fam
# 2062 /projectnb/bs859/data/RheumatoidArthritis/final_project/narac_hg19.fam
```

```
# Check the sex of individuals based on X chromosome genotyping
plink --bfile $PLINKFILE --check-sex --out $OUTPUTDIR/narac_sex_check

# Create a file to exclude the problematic individual
echo "1050200 1050200" > $OUTPUTDIR/exclude.txt

# Step 1: Quality Control with individual exclusion
# Remove SNPs with a high missingness rate (>5%), minor allele frequency (MAF) < 1%, and those not in HWE in controls (p < 1e-6)
plink --bfile $PLINKFILE --remove $OUTPUTDIR/exclude.txt --maf 0.01 --geno 0.05 --mind 0.05 --hwe 1e-6 --make-bed --out $OUTPUTDIR/narac_filtered

# Step 3: LD Pruning
# Prune SNPs to reduce the redundancy in the data
plink --bfile $OUTPUTDIR/narac_filtered --indep-pairwise 10000kb 1 0.2 --out $OUTPUTDIR/narac_pruned_data
# Use the pruned data for PCA
plink --bfile $OUTPUTDIR/narac_filtered --extract $OUTPUTDIR/narac_pruned_data.prune.in --make-bed --out $OUTPUTDIR/narac_pca

#run smartpca
smartpca -p $OUTPUTDIR/smartpca.par > $OUTPUTDIR/smartpca.log

#Plot PCs by case status
# command takes 4 arguments:
# 1) The name of the output file from smartpca
# 2 and 3) the two PCs to plot on x and y axes, respectively
# 4) Number of PCs in the file
# this script assumes the output is from smartpca, so the
# first column is the individual ID and the last column is
# case status (from the plink fam file used to run smartpca)
Rscript --vanilla plotPCs.R narac_pca.evec 1 2 10
Rscript --vanilla plotPCs.R narac_pca.evec 2 4 10
Rscript --vanilla plotPCs.R narac_pca.evec 1 4 10
```

APPENDIX B

Bash script for GWAS:

```
#!/bin/bash

module load R
module load plink/1.90b6.27

# Define input and output directories
DATADIR="/projectnb/bs859/data/RheumatoidArthritis/final_project"
OUTPUTDIR="/projectnb/bs859/students/bkapalli/A.project"

# Define PLINK binary file prefix
PLINKFILE=$DATADIR/narac_hg19

##make a temporary file to work on:
awk 'NR>1 {print $0}' $OUTPUTDIR/narac_pca.evec > $OUTPUTDIR/temp1.evec
##remove the colon from the ID:
sed 's/:/\t/g' $OUTPUTDIR/temp1.evec > $OUTPUTDIR/temp2.evec
#add the header:
echo -e "FID IID PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10" > $OUTPUTDIR/narac_pca_covar.txt
awk '{print $1,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11,$12}' $OUTPUTDIR/temp2.evec >>
$OUTPUTDIR/narac_pca_covar.txt

plink --bfile $OUTPUTDIR/narac_filtered --covar $OUTPUTDIR/narac_pca_covar.txt --covar-name PC1-PC10 --out
$OUTPUTDIR/checkPCs --allow-no-sex --logistic no-snp beta

# Calculate allele frequencies for the filtered dataset
plink --bfile $OUTPUTDIR/narac_filtered --freq --out $OUTPUTDIR/allele_freqs

# Merge allele frequencies into the GWAS output
# For female GWAS results
# Perform GWAS for female subjects only, using significant PCs as covariates
plink --bfile $PLINKFILE --filter-females --covar $OUTPUTDIR/narac_pca_covar.txt --covar-name PC2,PC4 --logistic
beta hide-covar --ci 0.95 --out $OUTPUTDIR/GWAS_female
awk 'BEGIN {FS=" "; OFS=" "}
    NR==FNR {a[$2]=$3; b[$2]=$4; freq[$2]=$5; next}
```

```
    FNR==1 {print $0, "ALLELE1", "ALLELE2", "FREQ"}
    FNR>1 {if ($2 in a) print $0, a[$2], b[$2], freq[$2]; else print $0, "NA", "NA", "NA"}' $OUTPUTDIR/allele_freqs.frq
$OUTPUTDIR/GWAS_female.assoc.logistic > $OUTPUTDIR/GWAS_female_ready.assoc.logistic


# How many SNPs in this GWAS have p-value < 0.0001
awk 'NR==1||$9<0.0001{print $0}' GWAS_female.assoc.logistic|wc



# Repeat for males
plink --bfile $PLINKFILE --filter-males --covar $OUTPUTDIR/narac_pca_covar.txt --covar-name PC2,PC4 --logistic
beta hide-covar --ci 0.95 --out $OUTPUTDIR/GWAS_male
awk 'BEGIN {FS=OFS=" "}
    NR==FNR {allele1[$2]=$3; allele2[$2]=$4; freq[$2]=$5; next}
    FNR==1 {print $0, "ALLELE1", "ALLELE2", "FREQ"}
    FNR>1 {print $0, allele1[$2], allele2[$2], freq[$2]}' $OUTPUTDIR/allele_freqs.frq
$OUTPUTDIR/GWAS_male.assoc.logistic > $OUTPUTDIR/GWAS_male_ready.assoc.logistic


# How many SNPs in this GWAS have p-value < 0.0001
awk 'NR==1||$9<0.0001{print $0}' GWAS_male.assoc.logistic|wc


# QQ and Manhattan plots using custom R scripts
# QQ plots
Rscript --vanilla $OUTPUTDIR/qq_umich_gc.R $OUTPUTDIR/GWAS_female.assoc.logistic "female GWAS" ADD
Rscript --vanilla $OUTPUTDIR/qq_umich_gc.R $OUTPUTDIR/GWAS_male.assoc.logistic "male GWAS" ADD

# Manhattan plots
Rscript --vanilla $OUTPUTDIR/gwaplot.R $OUTPUTDIR/GWAS_female.assoc.logistic "GWAS Results for Females"
female_manhattan
Rscript --vanilla $OUTPUTDIR/gwaplot.R $OUTPUTDIR/GWAS_male.assoc.logistic "GWAS Results for Males"
male_manhattan

# Summary tables can be created from the PLINK output
awk 'NR==1 || $9<0.05 {print}' $OUTPUTDIR/GWAS_female.assoc.logistic >
$OUTPUTDIR/GWAS_female_summary.txt
```

```
awk 'NR==1 || $9<0.05 {print}' $OUTPUTDIR/GWAS_male.assoc.logistic > $OUTPUTDIR/GWAS_male_summary.txt
```

Bash script for Meta analysis:

```
module load metal R
# Define input and output directories

# Define PLINK binary file prefix
PLINKFILE=$DATADIR/narac_hg19

##HW7.2:  run the meta analysis
metal metal.txt > metal.log

#Extract Necessary Data from the .bim File:
awk '{print $2, $1, $4}' $DATADIR/narac_hg19.bim > $OUTPUTDIR/snp_chr_bp.txt

# Filtering values with significant p values
awk 'NR==1||$10<5e-8{print $0}' merged_meta_data.tbl

# Assuming your output file is named ra_meta_analysis1.tbl
awk '{print $1, $9, $14}' ra_meta_analysis1.tbl > extracted_meta_analysis_data.txt


#Merge CHR and BP Information
awk 'BEGIN {OFS="\t"}
    FNR==NR {chr[$1]=$2; pos[$1]=$3; next}
    FNR==1 {print $0, "CHR", "BP"}
    FNR>1 && $1 in chr {print $0, chr[$1], pos[$1]}
    FNR>1 && !($1 in chr) {print $0, "NA", "NA"}' $OUTPUTDIR/snp_chr_bp.txt FS=" " ra_meta_analysis1.tbl FS="\t" >
$OUTPUTDIR/merged_meta_data.tbl

#HW7.2b:  qq plot and Manhattan plot
##first, extract chr, bp, and p value from the METAL output:
awk 'BEGIN {FS=OFS="\t"}
    NR==1 {print "CHR", "BP", "P-value"}  # Print the header for the new file
    NR>1 {print $16, $17, $10}' $OUTPUTDIR/merged_meta_data.tbl > $OUTPUTDIR/toplot.txt


Rscript --vanilla metalqqplot.R toplot.txt "Metal QQ Plot"
```

```
Rscript --vanilla gwaplot.R toplot.txt "MetaManhattan" metaman.png
```

Metal file:

```
# METAL configuration for RA GWAS Meta-Analysis
SCHEME STDERR
GENOMICCONTROL ON
AVERAGEFREQ ON
MINMAXFREQ ON
MARKER SNP
ALLELE1 ALLELE1
ALLELE2 ALLELE2
EFFECT BETA
STDERR SE
PVALUE P
WEIGHT NMISS
FREQLABEL FREQ

# Study-specific settings for Female GWAS
PROCESS /projectnb/bs859/students/bkapalli/A.project/GWAS_female_ready.assoc.logistic

# Study-specific settings for Male GWAS
PROCESS /projectnb/bs859/students/bkapalli/A.project/GWAS_male_ready.assoc.logistic

OUTFILE ra_meta_analysis .tbl
ANALYZE HETEROGENEITY
```

APPENDIX C

Bash script -

```
#!/bin/bash

# Define directories and files
RA_DIR='/projectnb/bs859/students/bkapalli/A.project/Okada_2014'
LDSCORES_DIR='/projectnb/bs859/data/ldscore_files'
OUTPUT_DIR='/projectnb/bs859/students/bkapalli/A.project'

module load R
```

```
module load python2
module load ldsc

# Preprocessing GWAS summary statistics with log transformation

zcat $RA_DIR/RA_GWASmeta_Asian_v2.txt.gz | \
awk 'BEGIN{OFS="\t"} {if(NR==1) print $0, "log_OR_A1"; else print $0, log($6)}' | \
gzip > $RA_DIR/RA_GWASmeta_Asian_v2_with_log.gz

zcat $RA_DIR/RA_GWASmeta_European_v2.txt.gz | \
awk 'BEGIN{OFS="\t"} {if(NR==1) print $0, "log_OR_A1"; else print $0, log($6)}' | \
gzip > $RA_DIR/RA_GWASmeta_European_v2_with_log.gz


# Prepare summary statistics by formatting them correctly with munge_sumstats.py
# European ancestry
# Total sample size = 14361 (cases) + 43923 (controls)
munge_sumstats.py \
  --sumstats $RA_DIR/RA_GWASmeta_European_v2_with_log.gz \
  --snp SNPID \
  --a1 A1 \
  --a2 A2 \
  --signed-sumstats log_OR_A1,0 \
  --p P-val \
  --N 58284 \
  --merge-alleles $LDSCORES_DIR/w_hm3.snplist \
  --out $OUTPUT_DIR/RA_EUR

# Asian ancestry
# Total sample size = 4873 (cases) + 17642 (controls)
munge_sumstats.py \
  --sumstats $RA_DIR/RA_GWASmeta_Asian_v2_with_log.gz \
  --snp SNPID \
  --a1 A1 \
  --a2 A2 \
  --signed-sumstats log_OR_A1,0 \
  --p P-val \
```

```
  --N 22515 \
  --merge-alleles $LDSCORES_DIR/w_hm3.snplist \
  --out $OUTPUT_DIR/RA_ASN

# Estimate heritability using LD score regression
# European ancestry

ldsc.py \
--h2 $OUTPUT_DIR/RA_EUR.sumstats.gz \
--ref-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.EUR.rsid \
--w-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.EUR.rsid \
--out $OUTPUT_DIR/RA_h2_EUR

# East Asian
ldsc.py \
  --h2 $OUTPUT_DIR/RA_ASN.sumstats.gz \
  --ref-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.EAS.rsid \
  --w-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.EAS.rsid \
  --out $OUTPUT_DIR/RA_h2_ASN_EAS

# Central/South Asian
ldsc.py \
  --h2 $OUTPUT_DIR/RA_ASN.sumstats.gz \
  --ref-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.CSA.rsid \
  --w-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.CSA.rsid \
  --out $OUTPUT_DIR/RA_h2_ASN_CSA

# Middle Eastern
ldsc.py \
  --h2 $OUTPUT_DIR/RA_ASN.sumstats.gz \
  --ref-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.MID.rsid \
  --w-ld $LDSCORES_DIR/UKBB.ALL.ldscore/UKBB.MID.rsid \
  --out $OUTPUT_DIR/RA_h2_ASN_MID

done

echo "All analyses are complete."
```