North American Rheumatoid Arthritis Consortium (NARAC) GWAS

You will work with the NARAC GWAS data that was provided for Genetic Analysis Workshop 16. The NARAC data come from a case-control study of Rheumatoid Arthritis (RA; N=868 cases, N=1194 controls). Cases were recruited from across the United States and are predominantly of Northern European origin. All met the American College of Rheumatology criteria for RA. The controls are derived from the New York Cancer Project, were enrolled in the New York metropolitan area and are somewhat enriched for individuals of Southern European or Ashkenazi Jewish ancestry compared with cases. All subjects are believed to be unrelated.

QUESTION 01

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.

1.1 Genetic data cleaning of the NARAC GWAS data

1.1.1 Code

```
(miniconda3)[neharao@scc-gf3 final-project]$ plink --bfile
/projectnb/bs859/students/neharao/final-project/RA-fp/narac_hg19 --maf 0.01 --geno 0.05 --hwe 1e-6
--freq --make-bed --out ra_cleaned_1
(miniconda3)[neharao@scc-gf3 final-project]$ plink --bfile ra_cleaned_1 --mind 0.05 --make-bed --out
ra_cleaned
```

1.1.2 Reasoning

- The first stage filters out markers (SNPs) based on minor allele frequency (MAF), genotyping rate, and Hardy-Weinberg Equilibrium (HWE) criteria. This initial filtering ensures that only high-quality markers are retained for subsequent analyses.

- The second stage filters out individuals based on the genotyping rate (mind 0.05). This step is applied after the marker filtering, which is generally preferred because removing individuals first can lead to the inadvertent removal of rare variants or markers with low MAF.
- By applying the individual filtering step last, we maximize the sample size, which is important for maintaining statistical power in downstream association analyses, particularly for a complex disease like Rheumatoid Arthritis.
- Additionally, this allows for a more controlled and systematic approach to data cleaning, where we can assess the impact of each filtering step separately and make informed decisions based on the quality control metrics.

1.1.3 Individuals removed retained in each step

- Step 01 statistics:
    - 544276 variants loaded from .bim file.
    - 2062 people (569 males, 1493 females) loaded from .fam.
    - 2062 phenotype values loaded from .fam.
    - 18402 variants removed due to missing genotype data (--geno).
    - 663 variants removed due to Hardy-Weinberg exact test.
    - 22907 variants removed due to minor allele threshold(s)
    - 502304 variants and 2062 people pass filters and QC.
    - Among remaining phenotypes, 868 are cases and 1194 are controls.
- Step 02 statistics:
    - 502304 variants loaded from .bim file.
    - 2062 people (569 males, 1493 females) loaded from .fam.
    - 2062 phenotype values loaded from .fam.
    - 0 people removed due to missing genotype data (--mind).
    - 502304 variants and 2062 people pass filters and QC.
    - Among remaining phenotypes, 868 are cases and 1194 are controls.

1.2 Pruning of cleaned data

1.2.1 Code

```
(miniconda3)[neharao@scc-gf3 final-project]$ plink --bfile ra_cleaned --geno 0.01 --maf 0.02
--indep-pairwise 10000kb 1 0.15 --out ra
```

```
(miniconda3)[neharao@scc-gf3 final-project]$ plink --bfile ra_cleaned --extract ra.prune.in --make-bed --out ra_cleaned_pruned
```

1.2.2 Reasoning

- A lower missing rate threshold and stringent genotyping rate threshold ensures that only high-quality markers are retained for the pruning step, which is important for accurate ancestry inference using PCA.
- A lower LD threshold helps to remove more highly correlated markers, which can improve the accuracy of PCA by reducing the influence of LD patterns on the principal components.

1.2.3 Individuals removed and retained

- Step 01 statistics:
    - 502304 variants loaded from .bim file.
    - 2062 people (569 males, 1493 females) loaded from .fam.
    - 2062 phenotype values loaded from .fam.
    - 68100 variants removed due to missing genotype data (--geno).
    - 4847 variants removed due to minor allele threshold(s)
    - 429357 variants and 2062 people pass filters and QC.
    - Among remaining phenotypes, 868 are cases and 1194 are controls.
    - Pruning is complete. 349713 of 429357 variants removed.
- Step 02 statistics:
    - 502304 variants loaded from .bim file.
    - 2062 people (569 males, 1493 females) loaded from .fam.
    - 2062 phenotype values loaded from .fam.
    - --extract: 79644 variants remaining.
    - 79644 variants and 2062 people pass filters and QC.
    - Among remaining phenotypes, 868 are cases and 1194 are controls.

1.3 PCA

1.3.1 Code

```
(miniconda3)[neharao@scc-gf3 final-project]$ cat q1.par
genotypename: ra_cleaned_pruned.bed
```

```
snpname: ra_cleaned_pruned.bim
indivname: ra_cleaned_pruned.fam
evecoutname: ra_pruned.evec
evaloutname: ra_pruned.eval
altnormstyle: NO
numoutevec: 10
numoutlieriter: 0
outliersigmathresh: 4
outlieroutname: outliers.removed
(miniconda3)[neharao@scc-gf3 final-project]$ smartpca -p q1.par >q1.out
```

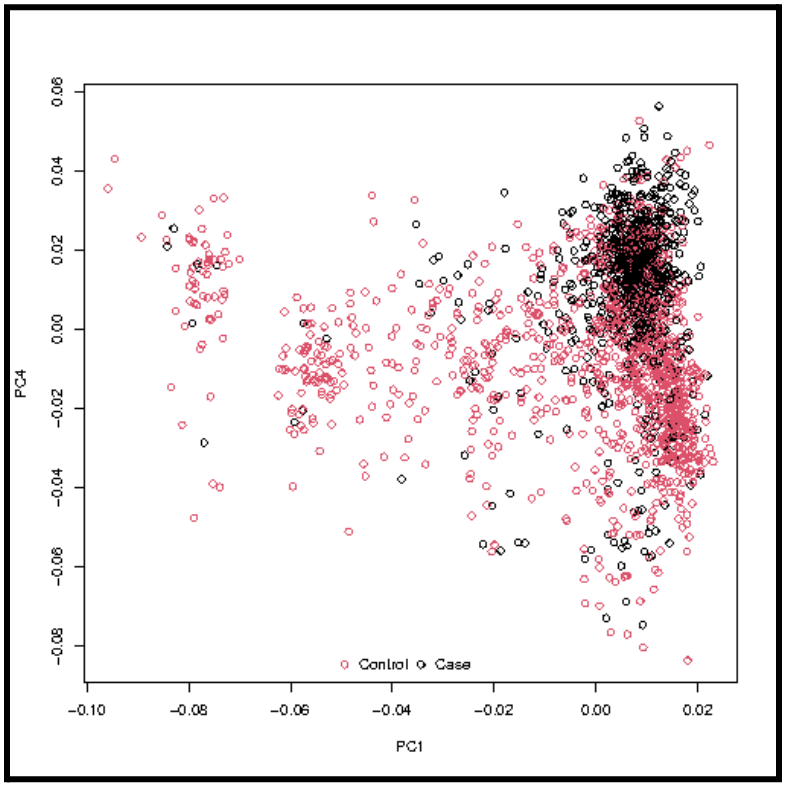
1.3.2 Output

```
eigenvector 1:means
        Control     -0.004
           Case      0.006
## Anova statistics for population differences along each eigenvector:
                                p-value
        eigenvector_1_Control_Case_   4.44089e-16 +++
eigenvector 2:means
           Case     -0.006
        Control      0.004
        eigenvector_2_Control_Case_   1.88738e-15 +++
eigenvector 3:means
           Case     -0.000
        Control      0.000
        eigenvector_3_Control_Case_      0.451357
eigenvector 4:means
        Control     -0.007
           Case      0.010
        eigenvector_4_Control_Case_   6.66134e-16 +++
eigenvector 5:means
           Case     -0.001
        Control      0.001
        eigenvector_5_Control_Case_      0.178092
eigenvector 6:means
           Case     -0.000
        Control      0.000
        eigenvector_6_Control_Case_      0.921495
eigenvector 7:means
        Control     -0.000
           Case      0.000
        eigenvector_7_Control_Case_      0.656963
```

```
eigenvector 8:means
          Control     -0.001
             Case      0.001
          eigenvector_8_Control_Case_     0.0977927
eigenvector 9:means
          Control     -0.000
             Case      0.000
          eigenvector_9_Control_Case_      0.401673
eigenvector 10:means
             Case     -0.000
          Control      0.000
         eigenvector_10_Control_Case_      0.969107
```

1.3.3 Reasoning

Eigenvectors 1 and 4 both exhibit very low p-values, indicating significant population differences between cases and controls. By including these two eigenvectors, we will capture a substantial portion of the population stratification that may confound our GWAS analyses. Additionally, since they represent the largest sources of population variation, they are likely to have a more significant impact on the results compared to higher eigenvectors.



1.4 Usage of the file RA_pcs.txt (RA_pcs.txt is the output from smartpca with a column header added) to determine which PCs are associated with case status.

"--logistic no-snp" produces results for the specified covariates, with no SNPs in the model.

1.4.1 Code

```
(miniconda3)[neharao@scc-gf3 final-project]$ plink --bfile ra_cleaned --covar RA_pcs.txt
--covar-name PC1-PC10 --out checkPCs --logistic no-snp beta
(miniconda3)[neharao@scc-gf3 final-project]$ cat checkPCs.assoc.logistic
    TEST    NMISS       BETA        STAT          P
     PC1     2061       32.6       8.781   1.623e-18
     PC2     2061     -28.33      -10.58    3.56e-26
     PC3     2061     -3.734      -1.048      0.2944
     PC4     2061      45.04       16.53   2.096e-61
     PC5     2061     -4.326      -1.801     0.07167
     PC6     2061     -0.217    -0.09033       0.928
     PC7     2061      1.333      0.5548       0.579
     PC8     2061      3.738       1.568      0.1169
     PC9     2061      2.809       1.164      0.2445
    PC10     2061   -0.08917    -0.03686      0.9706
```

1.4.2 Reasoning

PCs 1 and 4 have a significant association with case status using the logistic regression model. PC1 has a positive beta coefficient, indicating an increased odds of the outcome (case status) for a one-unit increase in the predictor variable (PC1). PC4 has a positive beta coefficient as well, indicating a similar association. PC2 has a negative beta coefficient, but it is still significantly associated with case status.

QUESTION 02

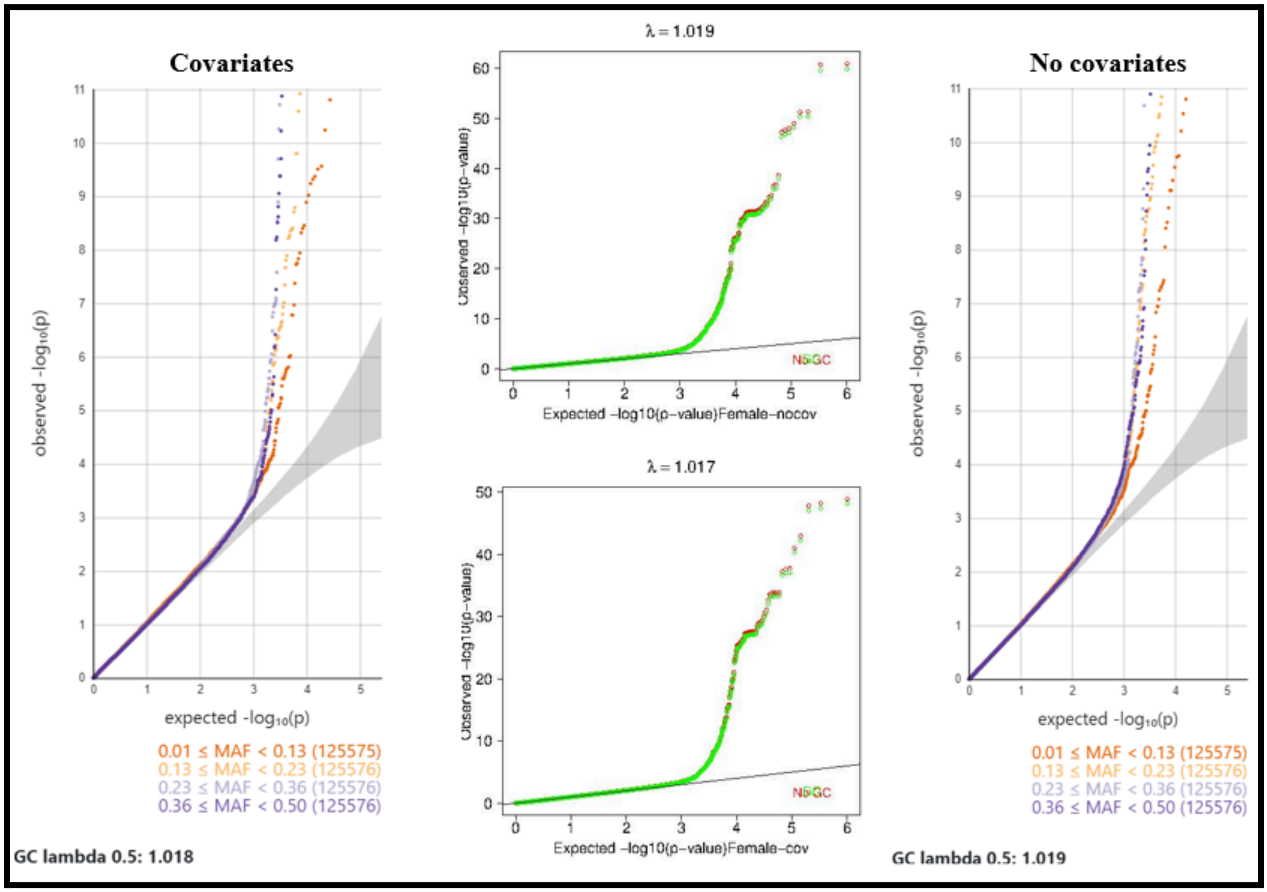
It is well known that the HLA region on chromosome 6p21 plays an important role in RA. It is also well known that females are affected by RA much more frequently than males. Your goals are to determine if there are additional genomic regions (in addition to the HLA region on chromosome 6) that are associated 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.

2.1 Perform two genome-wide association analyses for rheumatoid arthritis: one using only female subjects, and one using only male subjects.

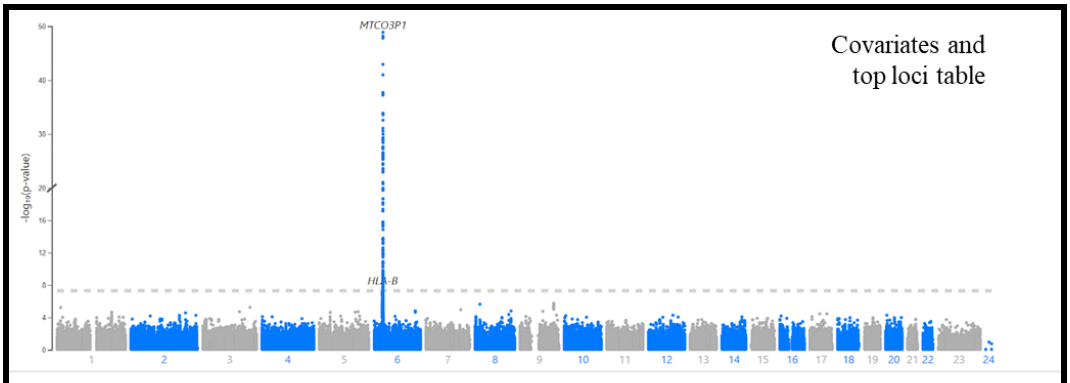
2.1.1 Female GWAS code

```
(miniconda3)[neharao@scc-vf1 fpv2]$ plink --bfile ra_cleaned --filter-females --extract ra.prune.in
--freq --make-rel square --out q2a-grm-female
(miniconda3)[neharao@scc-vf1 fpv2]$ Rscript --vanilla GMMAT.R
#awk commands to replace the "PVAL" column header with "P", "SCORE" with "BETA" and the
"POS" with "BP" to match PLINK output
(miniconda3)[neharao@scc-vf1 fpv2]$ awk 'NR==1{$4="BP";$9="BETA";$11="P"};{print $0}'
test.glmm.score.female.nocov>test.glmm.score.female.nocov.txt
(miniconda3)[neharao@scc-vf1 fpv2]$ awk 'NR==1{$4="BP";$9="BETA";$11="P"};{print $0}'
test.glmm.score.female.covariates>test.glmm.score.female.covariates.txt
```

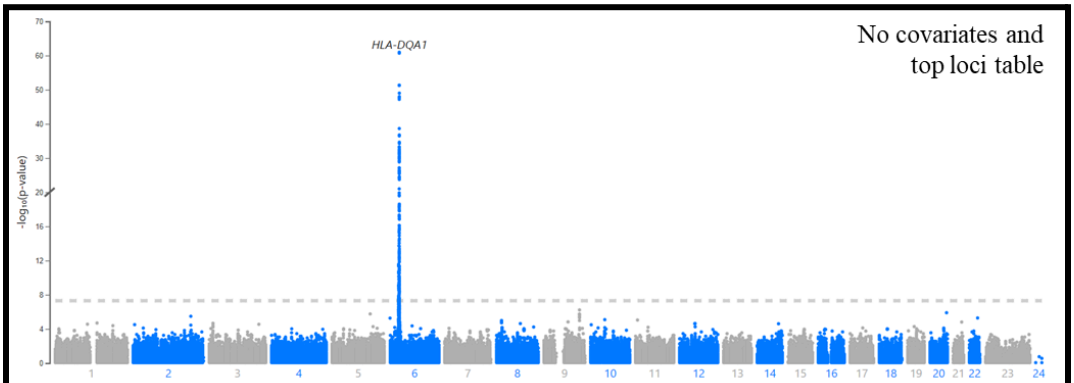
2.1.2 Female QQ plot



2.1.3 Female Manhattan plot



Marker	rsID	Nearest gene(s)	-log10(p)
6: 32,659,878 A/G	rs9275224	MTCO3P1	48.891
6: 31,240,431 A/C		HLA-B	7.671
6: 30,792,235 G/A		LINC00243	6.345
6: 30,114,955 A/G		TRIM40	7.113



Marker	rsID	Nearest gene(s)	-log10(p)
6: 32,577,380 G/A		HLA-DQA1	60.972
9: 123,706,382 A/G		C5	6.224

2.1.4 Females: Plot interpretation

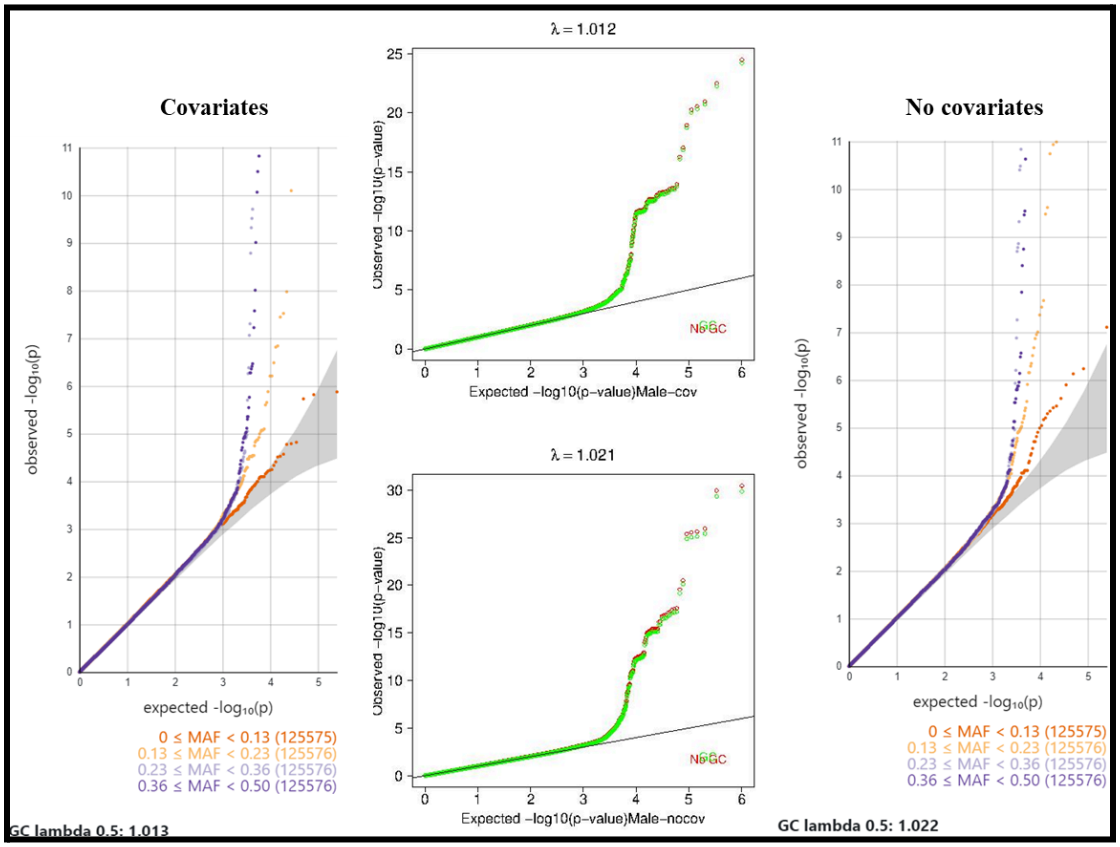
- No covariates QQ plot: Lambda GC of 1.01, which is not of concern and does not indicate genomic inflation. There is some deviation towards the end.
- No covariates Manhattan plot: As expected, significance is solely observed in chromosome 6. Chromosomes 8 and 9 have some samples a little beneath the significance threshold.

- Covariates QQ plot: Lambda GC of 1.01, which is not of concern and does not indicate genomic inflation. There is some deviation towards the end.
- Covariates Manhattan plot: As expected, significance is solely observed in chromosome 6. Chromosomes 9 and 20 have some samples a little beneath the significance threshold.

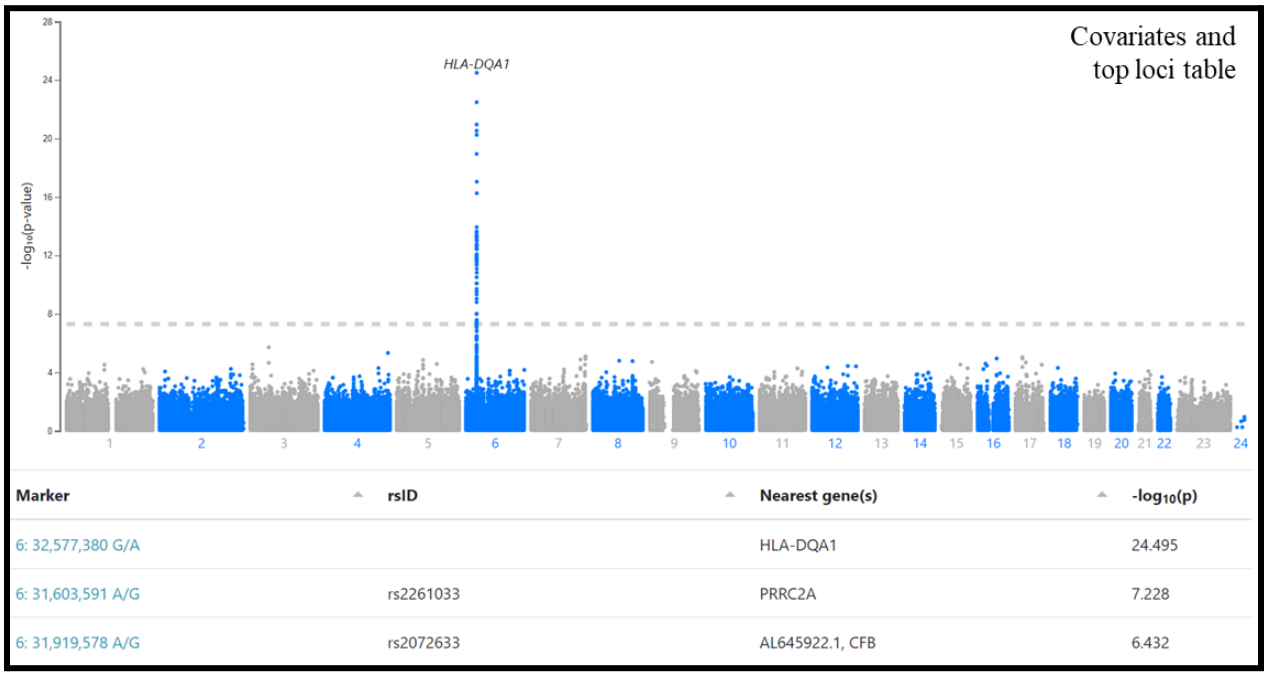
2.1.5 Male GWAS code

```
(miniconda3)[neharao@scc-vf1 fpv2]$ plink --bfile ra_cleaned --filter-males --extract ra.prune.in --freq --make-rel square --out q2a-grm-male
(miniconda3)[neharao@scc-vf1 fpv2]$ Rscript --vanilla GMMAT.R
#awk commands to replace the "PVAL" column header with "P", "SCORE" with "BETA" and the "POS" with "BP" to match PLINK output
(miniconda3)[neharao@scc-vf1 fpv2]$ awk 'NR==1{$4="BP";$9="BETA";$11="P"};{print $0}' test.glmm.score.male.nocov>test.glmm.score.male.nocov.txt
(miniconda3)[neharao@scc-vf1 fpv2]$ awk 'NR==1{$4="BP";$9="BETA";$11="P"};{print $0}' test.glmm.score.male.covariates>test.glmm.score.male.covariates.txt
```

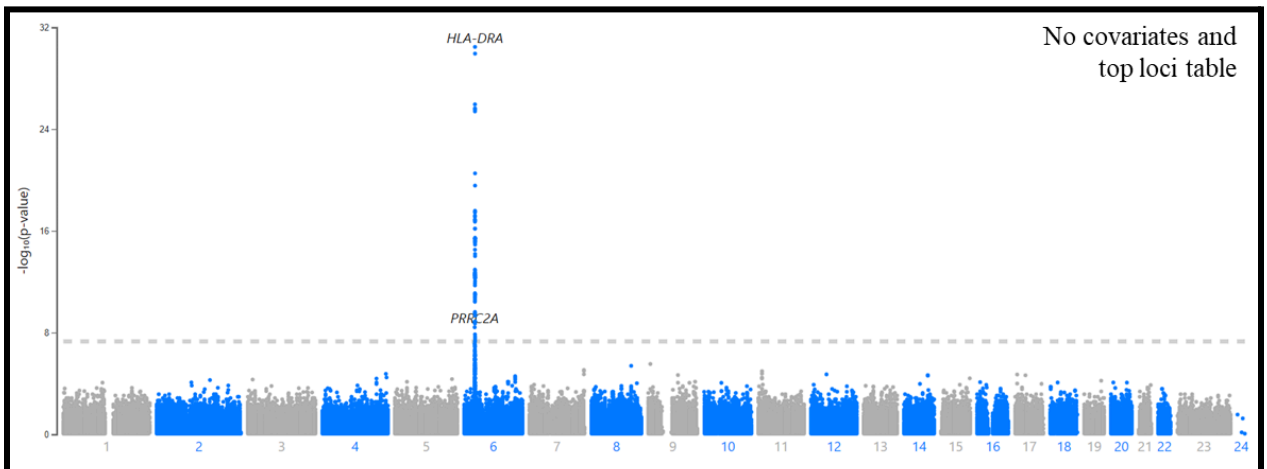
2.1.6 Male QQ plot



2.1.7 Male Manhattan plot



Marker	rsID	Nearest gene(s)	-log10(p)
6: 32,577,380 G/A		HLA-DQA1	24.495
6: 31,603,591 A/G	rs2261033	PRRC2A	7.228
6: 31,919,578 A/G	rs2072633	AL645922.1, CFB	6.432



Marker	rsID	Nearest gene(s)	-log10(p)
6: 32,405,026 A/G		HLA-DRA	30.457
6: 31,603,591 A/G	rs2261033	PRRC2A	8.413
6: 31,392,906 G/A		HCP5	7.106
6: 31,919,578 A/G	rs2072633	AL645922.1, CFB	6.588

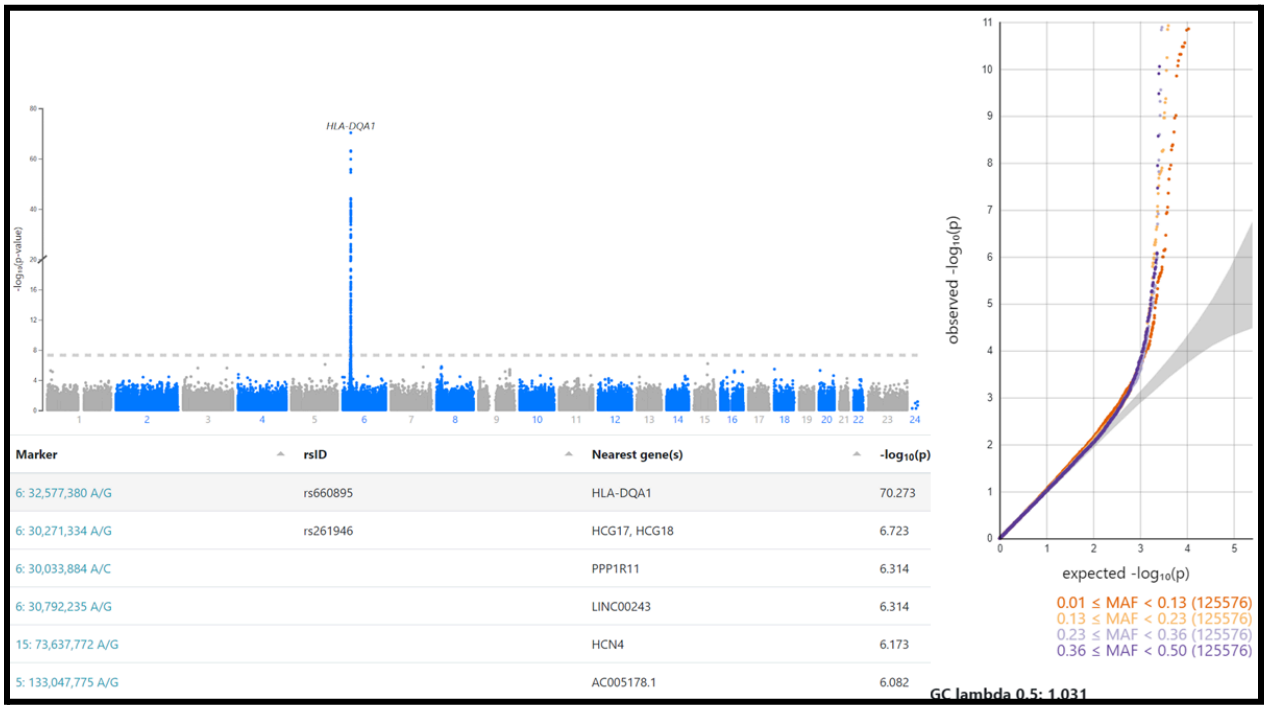
2.1.8 Males: Plot interpretation

- No covariates QQ plot: Lambda GC of 1.02, which is not of concern and does not indicate genomic inflation. There is some deviation towards the end.

- No covariates Manhattan plot: As expected, significance is solely observed in chromosome 6.
- Covariates QQ plot: Lambda GC of 1.01, which is not of concern and does not indicate genomic inflation. There is some deviation towards the end.
- Covariates Manhattan plot: As expected, significance is solely observed in chromosome 6.

2.2 Perform a genome-wide meta-analysis that combines the male-only and female-only results with a test for heterogeneity.

2.2.1 QQ and Manhattan plot



Marker	rsID	Nearest gene(s)	-log10(p)
6: 32,577,380 A/G	rs660895	HLA-DQA1	70.273
6: 30,271,334 A/G	rs261946	HCG17, HCG18	6.723
6: 30,033,884 A/C		PPP1R11	6.314
6: 30,792,235 A/G		LINC00243	6.314
15: 73,637,772 A/G		HCN4	6.173
5: 133,047,775 A/G		AC005178.1	6.082

2.2.2

QUESTION 03

---

a., b., and c. use summary statistics from the study Okada et al. “Genetics of rheumatoid arthritis contributes to biology and drug discovery.” Nature. 2014 Feb 20;506(7488):376-81. (http://www.nature.com.ezproxy.bu.edu/nature/journal/v506/n7488/full/nature12873.html), which includes 14,361 cases and 43,923 controls.

The file RA_GWASmeta_European_v2.txt contains the genome-wide meta-analysis results for the European ancestry samples (14361 RA cases, 43923 controls)
The file RA_GWASmeta_Asian_v2.txt.gz contains the genome-wide meta-analysis results for the Asian ancestry samples from the same study (4873 RA cases, 17642 controls).
The file RA_GWASmeta_TransEthnic_v2.txt.gz is a GWAS meta-analysis of all the European and Asian samples. (19234 RA cases and 61565 controls)

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.