Adjusting for Population structure (Demo)

dhriti.sengupta@wits.ac.za

Association testing

```
[dhriti@n07 Day3_Popstructure]$ plink --bfile demo_data --allow-no-sex --assoc  --out assoc_results
PLINK v1.90b6.25 64-bit (5 Mar 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to assoc_results.log.
Options in effect:
  --allow-no-sex
  --assoc
  --bfile demo_data
  --out assoc_results

451213 MB RAM detected; reserving 225606 MB for main workspace.
239972 variants loaded from .bim file.
607 people (0 males, 0 females, 607 ambiguous) loaded from .fam.
Ambiguous sex IDs written to assoc_results.nosex .
607 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 607 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.999986.
239972 variants and 607 people pass filters and QC.
Among remaining phenotypes, 299 are cases and 308 are controls.
Writing C/C --assoc report to assoc_results.assoc ... done.
```

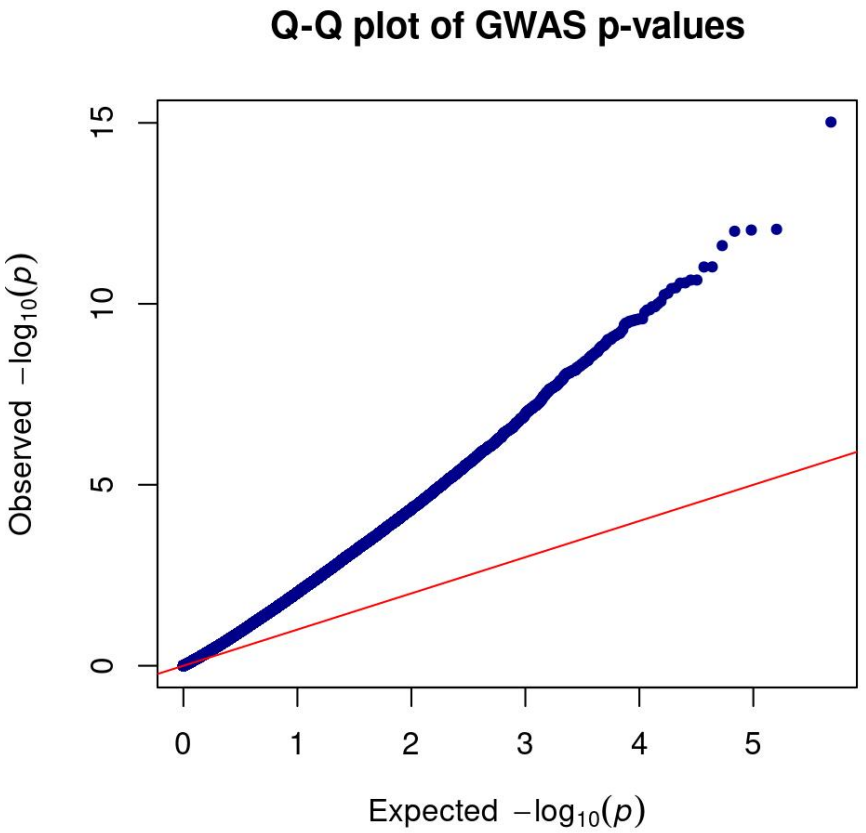
QQ Plot

```
[dhriti@n07 Day3_Popstructure]$ Rscript QQ_plot.R assoc_results.assoc assoc_results.qqplot.jpeg

null device
          1
```

Input file name

Output file name

assoc_results.qqplot.jpeg



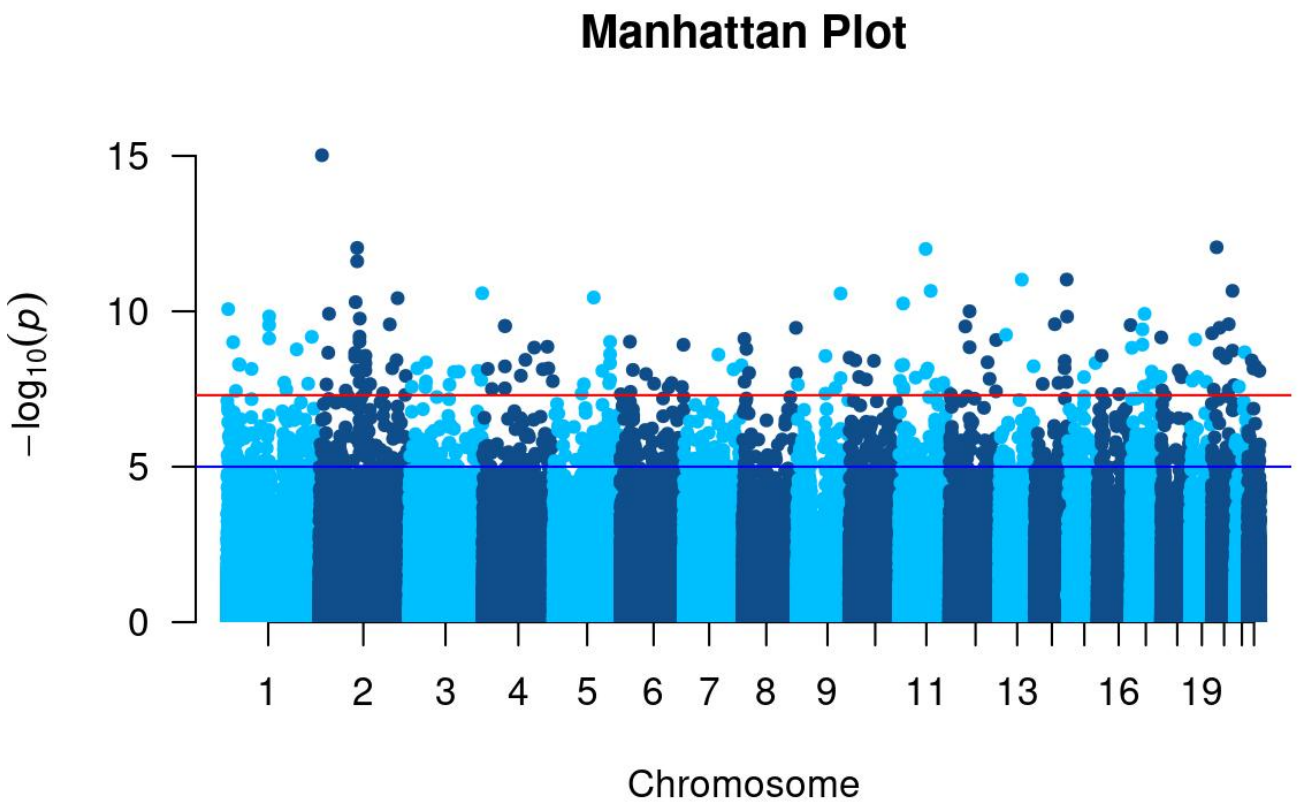
Manhattan Plot

```
[dhriti@n07 Day3_Popstructure]$ Rscript Manhattan_plot.R assoc_results.assoc assoc_results.manhattan.jpeg

null device
          1
```

Input file name

Output file name

assoc_results.manhattan.jpeg



Run association analysis –GC based correction

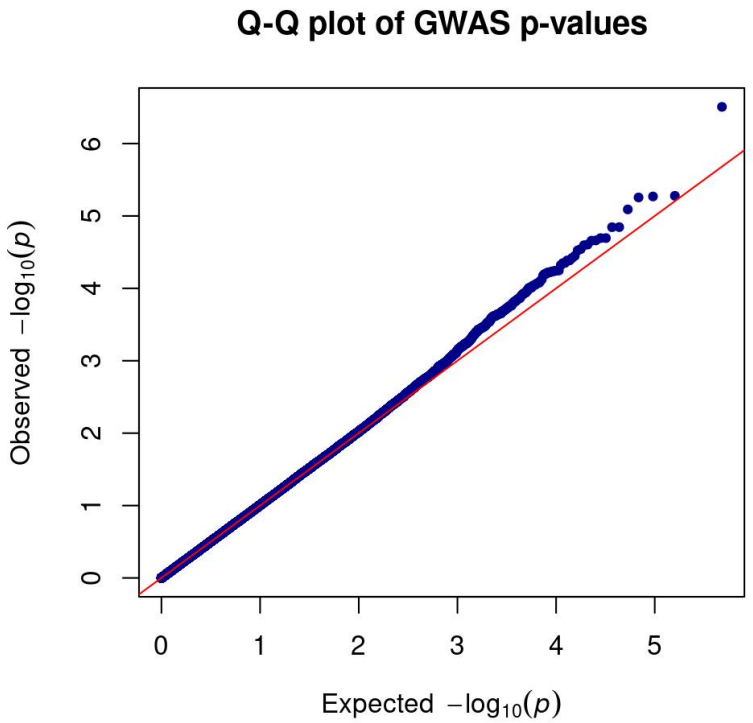
```
[dhriti@n07 Day3_Popstructure]$ plink --bfile demo_data --allow-no-sex --assoc --adjust  --out assoc.results
PLINK v1.90b6.25 64-bit (5 Mar 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to assoc.results.log.
Options in effect:
  --adjust
  --allow-no-sex
  --assoc
  --bfile demo_data
  --out assoc.results

451213 MB RAM detected; reserving 225606 MB for main workspace.
239972 variants loaded from .bim file.
607 people (0 males, 0 females, 607 ambiguous) loaded from .fam.
Ambiguous sex IDs written to assoc.results.nosex .
607 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 607 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.999986.
239972 variants and 607 people pass filters and QC.
Among remaining phenotypes, 299 are cases and 308 are controls.
Writing C/C --assoc report to assoc.results.assoc ... done.
--adjust: Genomic inflation est. lambda (based on median chisq) = 2.46489.
--adjust values (239972 variants) written to assoc.results.assoc.adjusted .
```

Understand the assoc and assoc.adjust files

```
[dhriti@n07 Day3_Popstructure]$ head assoc.results.assoc
 CHR                 SNP         BP   A1      F_A      F_U   A2        CHISQ            P           OR
   1           rs3131972     752721    A   0.2191   0.2305    G       0.2285       0.6327       0.9364
   1           rs4951931     800383    C    0.204   0.2419    T         2.51       0.1132       0.8033
   1          rs72631880     805556    A   0.2843   0.3328    T        3.345      0.06741       0.7963
   1          rs28705752     836896    C   0.4599   0.4042    T         3.83      0.05035        1.255
   1           rs7523690     838931    C   0.2876   0.3003    A       0.2357       0.6273       0.9406
   1          rs11516185     843405    G   0.4013   0.4367    A        1.557       0.2121       0.8648
   1           rs4970333     846489    T   0.3478   0.3847    C        1.781       0.1821       0.8529
   1          rs13303369     852875    C   0.4415   0.3847    T        4.029      0.04472        1.264
   1          rs60837925     860688    A    0.403   0.4642    G        4.613      0.03173       0.7793
[dhriti@n07 Day3_Popstructure]$ head assoc.results.assoc.adjusted
 CHR                 SNP       UNADJ         GC         BONF        HOLM   SIDAK_SS   SIDAK_SD     FDR_BH     FDR_BY
   2           rs7590460  9.517e-16  3.112e-07  2.284e-10  2.284e-10  2.284e-10  2.284e-10  2.284e-10  2.961e-09
  20           rs6078884  8.726e-13  5.272e-06  2.094e-07  2.094e-07  2.094e-07  2.094e-07   5.92e-08  7.675e-07
   2           rs2072205  9.151e-13  5.377e-06  2.196e-07  2.196e-07  2.196e-07  2.196e-07   5.92e-08  7.675e-07
  11           rs1540129  9.868e-13  5.549e-06  2.368e-07  2.368e-07  2.368e-07  2.368e-07   5.92e-08  7.675e-07
   2          rs55677186  2.462e-12  8.117e-06  5.909e-07  5.909e-07  5.909e-07  5.909e-07  1.182e-07  1.532e-06
  14          rs55690407  9.523e-12  1.426e-05  2.285e-06  2.285e-06  2.285e-06  2.285e-06   3.29e-07  4.266e-06
  13           rs348055  9.597e-12   1.43e-05  2.303e-06  2.303e-06  2.303e-06  2.303e-06   3.29e-07  4.266e-06
  20           rs6014945  2.198e-11  2.021e-05  5.274e-06  5.273e-06  5.274e-06  5.273e-06  5.842e-07  7.575e-06
  11           rs7127610  2.203e-11  2.023e-05  5.287e-06  5.287e-06  5.287e-06  5.287e-06  5.842e-07  7.575e-06
```

QQ plot after adjusting for GC

```
[dhriti@n07 Day3_Popstructure]$ sed 's/GC/P/g' assoc.results.assoc.adjusted >
assoc.results.assoc.adjusted1

[dhriti@n07 Day3_Popstructure]$ Rscript QQ_plot.R assoc.results.assoc.adjusted1
assoc_results.GCadjusted.qqplot.jpeg
```



So.. Lets try the PCA-based correction approach to control for population structure!!

Perform PCA on the dataset

Step 1: Prune for SNPs in linkage disequilibrium

```
[dhriti@n07 Day3_Popstructure]$ plink --bfile demo_data --indep-pairwise 50 5 0.5
....
....
Pruning complete.  46990 of 239972 variants removed.
Marker lists written to plink.prune.in and plink.prune.out .
```

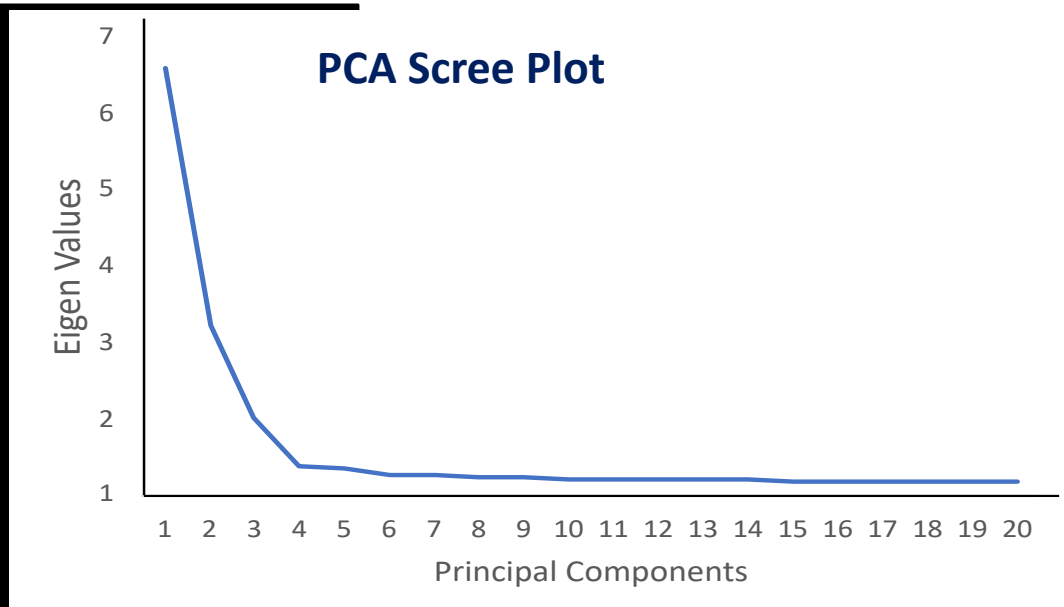
Step 2: Extract the LD pruned SNPs and run PCA

```
[dhriti@n07 Day3_Popstructure]$ plink --bfile demo_data --extract plink.prune.in --pca 'header' --out demo_data.pca
....
....
Calculating allele frequencies... done.
Total genotyping rate is 0.999985.
192982 variants and 607 people pass filters and QC.
Note: No phenotypes present.
Relationship matrix calculation complete.
--pca: Results saved to demo_data.pca.eigenval and demo_data.pca.eigenvec .
```

Quick look at the eigenval and eigenvec files

```
[dhriti@n07 Day3_Popstructure]$ head demo_data.pca.eigenvec | cut -d " " -f 1-7
FID IID PC1 PC2 PC3 PC4 PC5
KHV HG01595 0.0187399 -0.0243556 0.0208827 -0.00650665 0.000110606
KHV HG01596 0.0225297 -0.0204755 0.0253923 -0.00736197 -0.0198281
KHV HG01597 0.0224866 -0.0256135 0.0285914 0.010246 -0.00540424
KHV HG01598 0.0232304 -0.0221139 0.0269933 0.0178236 0.00520078
KHV HG01599 0.0236758 -0.0198614 0.0176917 0.00256827 0.00396549
KHV HG01600 0.0201669 -0.0230967 0.0211958 0.00784369 0.00997305
KHV HG01840 0.0215074 -0.0201304 0.0201439 0.000774127 -0.00786104
```

```
[dhriti@n07 Day3_Popstructure]$ head demo_data.pca.eigenval
6.59159
3.2266
2.01989
1.39515
1.36478
1.28436
1.26394
1.25561
1.23794
1.22799
```



Visualise the PCA plot

Step 1: Generate input file for plotting PCA i.e. merge case control and eigenvec values information for each individual

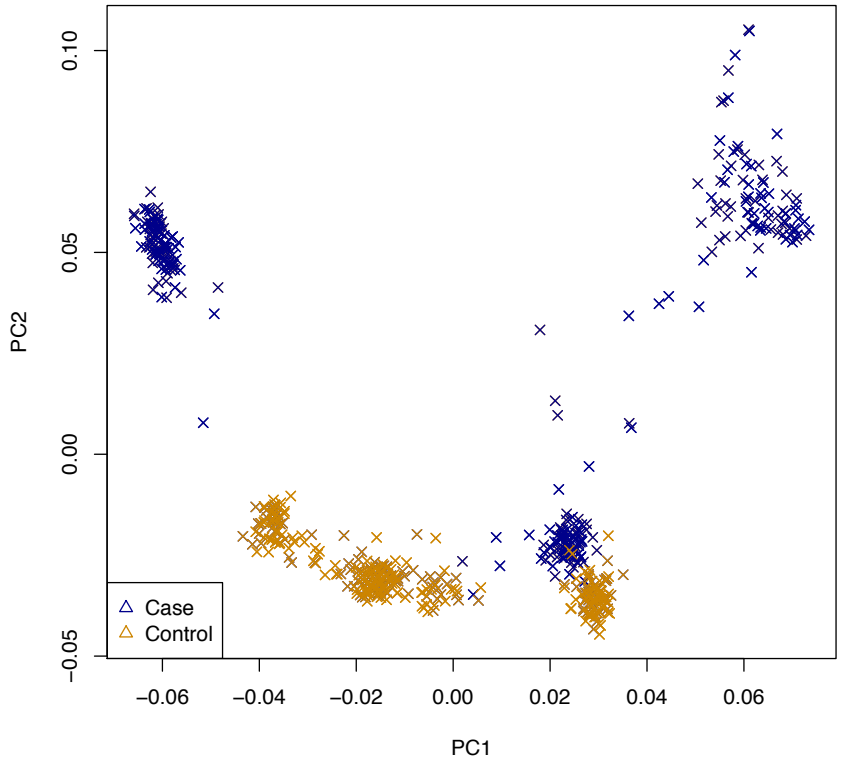
```
[dhriti@n07 Day3_Popstructure]$ echo IID Pheno > demo_data.phe

[dhriti@n07 Day3_Popstructure]$ awk '{print $2,$6}' demo_data.fam  >> demo_data.phe

[dhriti@n07 Day3_Popstructure]$ paste demo_data.pca.eigenvec demo_data.phe | awk '{if ($2==$23)
print $24,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11,$12}' | sed 's/^1/Control/g' | sed 's/^2/Case/g' >
demo_data.pca.input
```

```
[dhriti@n07 Day3_Popstructure]$ head demo_data.pca.input
Pheno IID PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10
Case HG01595 0.0187399 -0.0243556 0.0208827 -0.00650665 0.000110606 0.0568241 0.0220414 -0.0194875 -0.0244423 -0.0021596
Case HG01596 0.0225297 -0.0204755 0.0253923 -0.00736197 -0.0198281 0.0507521 0.0437305 -0.0328907 0.00680026 -0.0152018
Case HG01597 0.0224866 -0.0256135 0.0285914 0.010246 -0.00540424 0.0401487 0.0536894 -0.0338571 0.00725142 0.00400779
Case HG01598 0.0232304 -0.0221139 0.0269933 0.0178236 0.00520078 0.0768699 0.0405253 -0.032455 -0.0157789 -0.00711748
Case HG01599 0.0236758 -0.0198614 0.0176917 0.00256827 0.00396549 0.0467725 0.0298377 -0.0228021 0.0164138 0.0299299
Case HG01600 0.0201669 -0.0230967 0.0211958 0.00784369 0.00997305 0.0280208 0.0166649 -0.035853 0.00596662 0.00773415
Case HG01840 0.0215074 -0.0201304 0.0201439 0.000774127 -0.00786104 0.0455911 -0.00239304 -0.00982353 0.0163127 -0.0024375
Case HG01841 0.0243822 -0.0268529 0.0237612 -0.00104106 0.00241319 0.0608502 0.0372692 0.0135422 0.00730925 7.63257e-05
Case HG01842 0.0253671 -0.0237707 0.0178792 0.0156248 -0.0127382 0.0502243 0.0422749 -0.0746345 -0.0239062 0.0229318
```

Visualise the PCA plot.. contd

Step 2: Run the R Script

```
[dhriti@n07 Day3_Popstructure]$ Rscript plot_PCA.R
null device
          1
```



Run association analysis – adjusting for PCA

```
[dhriti@n07 Day3_Popstructure]$ plink --bfile demo_data --allow-no-sex --logistic hide-covar --covar demo_data.pca.eigenvec
--covar-name PC1,PC2,PC3,PC4,PC5,PC6 --out logistic_results
PLINK v1.90b6.25 64-bit (5 Mar 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to logistic_results.log.
Options in effect:
  --allow-no-sex
  --bfile demo_data
  --covar demo_data.pca.eigenvec
  --covar-name PC1,PC2,PC3,PC4,PC5,PC6
  --logistic hide-covar
  --out logistic_results

451213 MB RAM detected; reserving 225606 MB for main workspace.
239972 variants loaded from .bim file.
607 people (0 males, 0 females, 607 ambiguous) loaded from .fam.
Ambiguous sex IDs written to logistic_results.nosex .
607 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
--covar: 6 out of 20 covariates loaded.
Before main variant filters, 607 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.999986.
239972 variants and 607 people pass filters and QC.
Among remaining phenotypes, 299 are cases and 308 are controls.
Writing logistic model association results to logistic_results.assoc.logistic
... done.
```

Examine the .assoc.logistic file and generate plots

```
[dhriti@n07 Day3_Popstructure]$ head logistic_results.assoc.logistic
 CHR                   SNP         BP   A1       TEST    NMISS          OR          STAT             P
   1             rs3131972     752721    A        ADD      607      0.1631        -1.191        0.2337
   1             rs4951931     800383    C        ADD      607      0.2162         -1.22        0.2226
   1            rs72631880     805556    A        ADD      607       1.014       0.01477        0.9882
   1            rs28705752     836896    C        ADD      607      0.6929       -0.4066        0.6843
   1             rs7523690     838931    C        ADD      607       1.473        0.4407        0.6594
   1            rs11516185     843405    G        ADD      607      0.7538        -0.332        0.7399
   1             rs4970333     846489    T        ADD      607       1.425        0.3521        0.7248
   1            rs13303369     852875    C        ADD      607       1.634        0.5348        0.5928
   1            rs60837925     860688    A        ADD      606       2.069        0.8613        0.3891
```

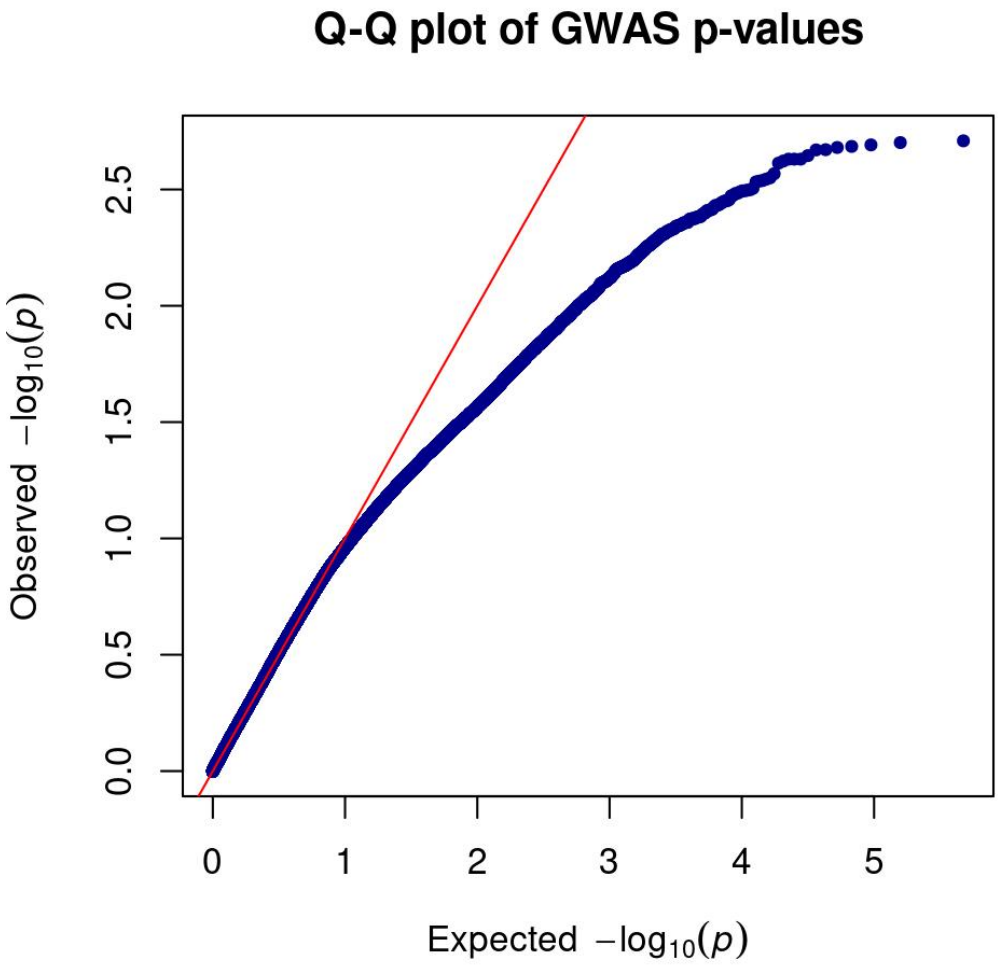
Remove the lines with NA

```
[dhriti@n07 Day3_Popstructure]$ awk '!/'NA'/' logistic_results.assoc.logistic > logistic_results.assoc_2.logistic
```

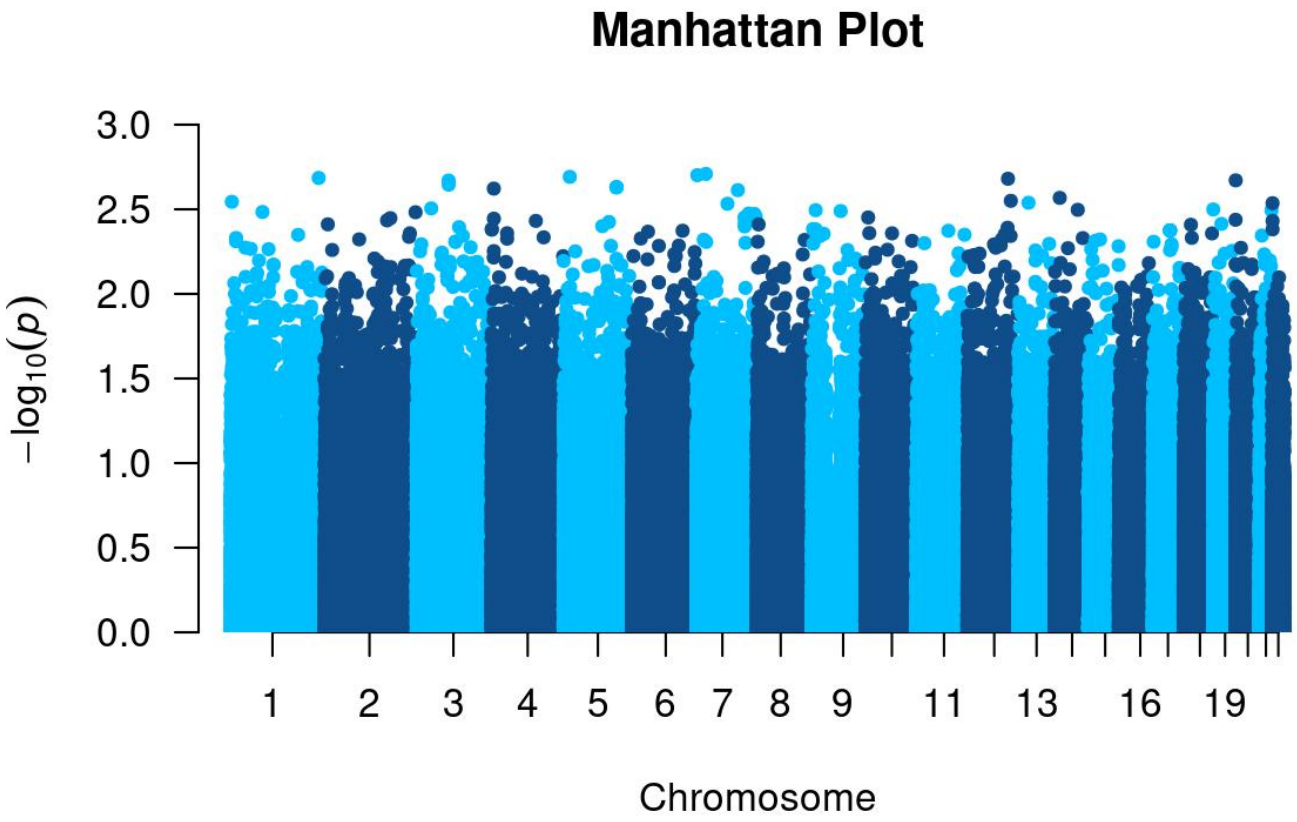
Generate the QQ plot and Manhattan plot using R scripts

```
[dhriti@n07 Day3_Popstructure]$ Rscript QQ_plot.R logistic_results.assoc_2.logistic logistic_results.qqplot.jpeg

[dhriti@n07 Day3_Popstructure]$ Rscript Manhattan_plot.R logistic_results.assoc_2.logistic logistic_results.manhattan.jpeg
```

Manhattan and QQ plot after PCA based correction



logistic_results.qqplot.jpeg



logistic_results.manhattan.jpeg