GWAS Analysis & Interactive plots

Bruno Ambrozio

11/10/2019

Description

- A GWAS (https://en.wikipedia.org/wiki/Genome-wide_association_study) analysis, with post-analytic visualization and interrogation.
    - Analysis steps derived from the paper: “A guide to genome-wide association analysis and post-analytic interrogation” (doi: 10.1002/sim.6605) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5019244/) (Reed et al., 2015).
    - Dataset from the PennCATH study of coronary artery disease (CAD). Paper: “Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies.” (doi: 10.1016/S0140-6736(10)61996-4) (http://www.ncbi.nlm.nih.gov/pubmed/21239051) ((Reilly et al., 2011))
    - Detailed tutorial can be found here (http://www.stat-gen.org/tut/tut_intro.html)

Summary

1. Download necessary R packages and setting global parameters to save progress while working through the GWA analysis.
2. Include quality control steps for both SNP and sample level filtering. * PCA (https://en.wikipedia.org/wiki/Principal_component_analysis) for population stratification (https://en.wikipedia.org/wiki/Population_stratification) in statistical modeling, as well as imputation of non-typed SNPs using 1000 Genomes reference genotype data.
3. GWAS analysis’ strategies.
    - Basic linear modeling functionality.
    - Imputed data using functionality contained.
4. Post-analytic interrogation.
    - Performance of statistical models.
    - Visualization of the global and subsetted GWAS output.

Preliminary

- Install packages:
    - Bioconductor (http://www.bioconductor.org/).
        - snpStats (http://www.bioconductor.org/packages/release/bioc/html/snpStats.html).
            - Read in various formats of genotype data.
            - Qality control.
            - Imputation and association analysis.
        - SNPRelate (http://master.bioconductor.org/packages/release/bioc/html/SNPRelate.html).
        - Sample level quality control.
        - Cmputationally efficient principal component calculation.
    - LDheatmap (http://cran.r-project.org/web/packages/LDheatmap/index.html) and postgwas (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071775).
        - Data visualization.
    - plyr (http://plyr.had.co.nz/).
        - Data manipulation.
    - genABEL (http://www.genabel.org/).
        - Statistical calculation.
    - doParallel (http://cran.r-project.org/web/packages/doParallel/index.html).
        - parallel processing.

Files and variables of interest

- Files formatted for PLINK (http://zzz.bwh.harvard.edu/plink/).
  - .bed: binary genotype information
  - .bim: SNP's data.
    - Columns: chromosome, SNP name, genetic distance, chromosomal position, identity of allele 1 (pertains to the minor, or less common allele), identity of allele 2.
  - .fam: samples' information.
    - Columns: family ID, individual ID, paternal ID, maternal ID, sex (1 = male, 2 = female), and phenotype.
  - .CSV: A supplemental clinical file for outcome variables and additional covariates.
    - Columns: sample ID (Family ID in the .fam file), coronary artery disease status (0 = control, 1 = affected), sex (1 = male, 2 = female), age (years), triglyceride level (mg/dL), high-density lipoprotein level (mg/dL), low-density lipoprotein level (mg/dL).

```
## Loading required package: survival
```

```
## Loading required package: Matrix
```

Data pre-processing

- The `geno` object contains a `genotype` member of type `SnpMatrix` where each column is a SNP and each row is a sample. For convenience, we assign that to the object, `genotype` .
- Filter the `genotype` data to only include samples with corresponding clinical data.

```
## A SnpMatrix with  1401 rows and  861473 columns
## Row names:  10002 ... 11596
## Col names:  rs10458597 ... rs5970564
```

```
##            chr        SNP gen.dist position   A1 A2
## rs10458597   1 rs10458597        0   564621 <NA>  C
## rs12565286   1 rs12565286        0   721290    G  C
## rs12082473   1 rs12082473        0   740857    T  C
## rs3094315    1  rs3094315        0   752566    C  T
## rs2286139    1  rs2286139        0   761732    C  T
## rs11240776   1 rs11240776        0   765269    G  A
```

```
##       FamID CAD sex age  tg hdl ldl
## 10002 10002   1   1  60  NA  NA  NA
## 10004 10004   1   2  50  55  23  75
## 10005 10005   1   1  55 105  37  69
## 10007 10007   1   1  52 314  54 108
## 10008 10008   1   1  58 161  40  94
## 10009 10009   1   1  59 171  46  92
```

```
## A SnpMatrix with  1401 rows and  861473 columns
## Row names:  10002 ... 11596
## Col names:  rs10458597 ... rs5970564
```

SNP level filtering

- Remove SNPs that fail to meet minimum criteria due to missing data, low variability or genotyping errors.

```
##            Calls Call.rate Certain.calls       RAF         MAF       P.AA
## rs10458597  1398 0.9978587             1 1.0000000 0.000000000 0.00000000
## rs12565286  1384 0.9878658             1 0.9483382 0.051661850 0.00433526
## rs12082473  1369 0.9771592             1 0.9985391 0.001460920 0.00000000
## rs3094315   1386 0.9892934             1 0.8217893 0.178210678 0.04761905
## rs2286139   1364 0.9735903             1 0.8621701 0.137829912 0.02199413
## rs11240776  1269 0.9057816             1 0.9988180 0.001182033 0.00000000
##                   P.AB      P.BB       z.HWE
## rs10458597 0.000000000 1.0000000          NA
## rs12565286 0.094653179 0.9010116 -1.26529432
## rs12082473 0.002921841 0.9970782  0.05413314
## rs3094315  0.261183261 0.6911977 -4.03172248
## rs2286139  0.231671554 0.7463343 -0.93146122
## rs11240776 0.002364066 0.9976359  0.04215743
```

- Keep the subset of SNPs that meet minimum call rate and MAF (https://en.wikipedia.org/wiki/Minor_allele_frequency) criterias.

```
## 203287 SNPs will be removed due to low MAF or call rate.
```

```
## A SnpMatrix with  1401 rows and  658186 columns
## Row names:  10002 ... 11596
## Col names:  rs12565286 ... rs5970564
```

Basic sample filtering

- `row.summary` for sample level quality control for missing data and heterozygosity
  - Additional heterozygosity F statistic:
    - `|F|=(1-O/E)` , where:
      - O = observed proportion of heterozygous genotypes for a given sample.
      - E = expected proportion of heterozygous genotypes for a given sample, based on the MAF across all non-missing SNPs for a given sample.

```
## Loading required package: gdsfmt
```

```
## SNPRelate -- supported by Streaming SIMD Extensions 2 (SSE2)
```

```
##       Call.rate Certain.calls Heterozygosity          hetF
## 10002 0.9826554             1      0.3289825 -0.0247708291
## 10004 0.9891581             1      0.3242931 -0.0103236529
## 10005 0.9918427             1      0.3231825 -0.0062550972
## 10007 0.9861027             1      0.3241469 -0.0098475016
## 10008 0.9823333             1      0.3228218 -0.0075941985
## 10009 0.9913034             1      0.3213658 -0.0002633189
```

- Apply filtering on call rate and heterozygosity, selecting only those samples that meet the criteria.

```
## 0 subjects will be removed due to low sample call rate or inbreeding coefficient.
```

Identity-by-descent (https://en.wikipedia.org/wiki/Identity_by_descent) analysis

- Filter on relatedness criteria (demands GDS file format)
- `SNPRelate` package to perform IBD analysis on a subset of SNPs that are in linkage equilibrium by iteratively removing adjacent SNPs that exceed an LD (https://en.wikipedia.org/wiki/Linkage_disequilibrium) threshold in a sliding window (https://stackoverflow.com/a/8269948/7224879) using the `snpgdsLDpruning` function.

```
## Start snpgdsBED2GDS ...
##  BED file: "/Users/bambrozi/Downloads/gwas_cad_data/GWAStutorial.bed" in the SNP-major mod
e (Sample X SNP)
##  FAM file: "/Users/bambrozi/Downloads/gwas_cad_data/GWAStutorial.fam", DONE.
##  BIM file: "/Users/bambrozi/Downloads/gwas_cad_data/GWAStutorial.bim", DONE.
## Tue Oct 15 14:01:14 2019     store sample id, snp id, position, and chromosome.
##  start writing: 1401 samples, 861473 SNPs ...
##      Tue Oct 15 14:01:14 2019    0%
##      Tue Oct 15 14:01:19 2019    100%
## Tue Oct 15 14:01:20 2019     Done.
## Optimize the access efficiency ...
## Clean up the fragments of GDS file:
##     open the file '/Users/bambrozi/Downloads/gwas_cad_data/GWAStutorial.gds' (292.4M)
##     # of fragments: 39
##     save to '/Users/bambrozi/Downloads/gwas_cad_data/GWAStutorial.gds.tmp'
##     rename '/Users/bambrozi/Downloads/gwas_cad_data/GWAStutorial.gds.tmp' (292.4M, reduce
d: 252B)
##     # of fragments: 18
```

```
## Hint: it is suggested to call `snpgdsOpen' to open a SNP GDS file instead of `openfn.gds'.
```

```
## SNP pruning based on LD:
## Excluding 203,287 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
## Working space: 1,401 samples, 658,186 SNPs
##     using 1 (CPU) core
##     sliding window: 500,000 basepairs, Inf SNPs
##     |LD| threshold: 0.2
##     method: composite
## Chromosome 1: 8.25%, 5,863/71,038
## Chromosome 3: 8.10%, 4,906/60,565
## Chromosome 6: 8.06%, 4,364/54,176
## Chromosome 12: 8.59%, 3,619/42,124
## Chromosome 21: 9.40%, 1,171/12,463
## Chromosome 2: 7.67%, 5,655/73,717
## Chromosome 4: 8.23%, 4,582/55,675
## Chromosome 7: 8.51%, 3,947/46,391
## Chromosome 11: 7.90%, 3,495/44,213
## Chromosome 10: 8.01%, 3,837/47,930
## Chromosome 8: 7.68%, 3,709/48,299
## Chromosome 5: 8.08%, 4,537/56,178
## Chromosome 14: 8.79%, 2,467/28,054
## Chromosome 9: 8.25%, 3,392/41,110
## Chromosome 17: 11.17%, 2,227/19,939
## Chromosome 13: 8.36%, 2,863/34,262
## Chromosome 20: 9.40%, 2,139/22,753
## Chromosome 15: 9.25%, 2,396/25,900
## Chromosome 16: 9.30%, 2,566/27,591
## Chromosome 18: 8.90%, 2,335/26,231
## Chromosome 19: 13.01%, 1,494/11,482
## Chromosome 22: 10.96%, 1,248/11,382
## 72,812 markers are selected in total.
```

```
## 72812 will be used in IBD analysis
```

- `snpgdsIBDMoM` function computes the IBD coefficients using method of moments. The result is a table indicating kinship among pairs of samples.

```
## Hint: it is suggested to call `snpgdsOpen' to open a SNP GDS file instead of `openfn.gds'.
```

```
## IBD analysis (PLINK method of moment) on genotypes:
## Excluding 788,661 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
## Working space: 1,401 samples, 72,812 SNPs
##     using 1 (CPU) core
## PLINK IBD:    the sum of all selected genotypes (0,1,2) = 32757268
## Tue Oct 15 14:01:49 2019    (internal increment: 23040)
##
[..................................................]  0%, ETC: ---
[==================================================] 100%, completed in 7s
## Tue Oct 15 14:01:56 2019    Done.
```

```
##     ID1   ID2        k0         k1    kinship
## 1 10002 10004 0.9201072 0.07989281 0.01997320
## 2 10002 10005 0.9478000 0.05220002 0.01305001
## 3 10002 10007 0.9209875 0.07901253 0.01975313
## 4 10002 10008 0.9312527 0.06874726 0.01718682
## 5 10002 10009 0.9386937 0.06130626 0.01532656
## 6 10002 10010 0.9146065 0.08539354 0.02134839
```

- Using the IBD pairwise sample relatedness measure, iteratively remove samples that are too similar using a greedy strategy in which the sample with the largest number of related samples is removed. The process is repeated until there are no more pairs of samples with kinship coefficients above the cut-off.

```
## 0 similar samples removed due to correlation coefficient >= 0.1
```

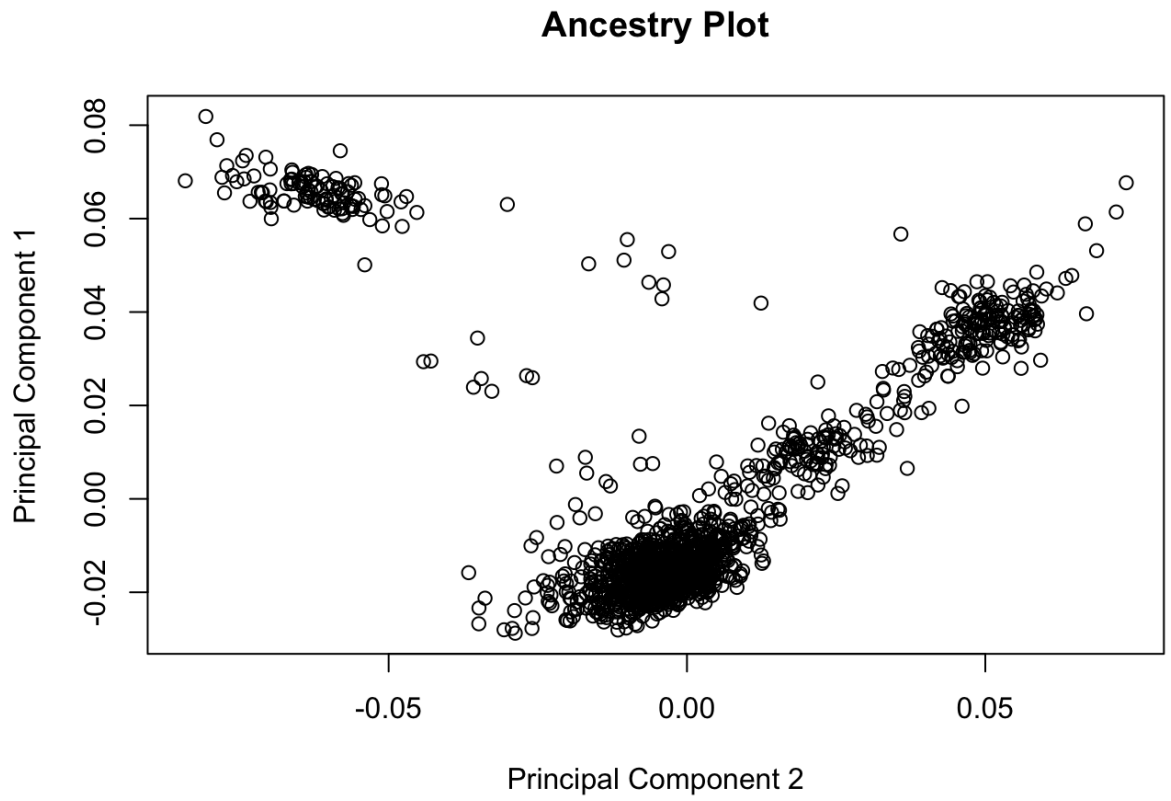
```
## A SnpMatrix with  1401 rows and  658186 columns
## Row names:  10002 ... 11596
## Col names:  rs12565286 ... rs5970564
```

Ancestry

- Plot the first two principal components of the genotype data. (Achieved from `snpgdsPCA` function from `SNPRelate`).
    - Assuming homogeneity of the sample (other datasets might have to test it first), coming from european ancestry. (This is why 0 SNP will be excluded as result)

```
## Hint: it is suggested to call `snpgdsOpen' to open a SNP GDS file instead of `openfn.gds'.
```

```
## Principal Component Analysis (PCA) on genotypes:
## Excluding 788,661 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
## Working space: 1,401 samples, 72,812 SNPs
##     using 1 (CPU) core
## PCA:    the sum of all selected genotypes (0,1,2) = 32757268
## CPU capabilities: Double-Precision SSE2
## Tue Oct 15 14:01:59 2019    (internal increment: 720)
##
[..................................................]  0%, ETC: ---
[==================================================] 100%, completed in 28s
## Tue Oct 15 14:02:27 2019    Begin (eigenvalues and eigenvectors)
## Tue Oct 15 14:02:28 2019    Done.
```



SNP Filtering - HWE (https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle) filtering on control samples

- Rejection of Hardy-Weinberg equilibrium can be an indication of population substructure or genotyping errors.
  - Remove SNPs with p-values, corresponding to the HWE test statistic on CAD controls, of less than 1×10^{-6}.
    - HWE on CAD controls due to possible violation of HWE caused by disease association.

```
## 1296 SNPs will be removed due to high HWE.
```

```
## A SnpMatrix with  1401 rows and  656890 columns
## Row names:  10002 ... 11596
## Col names:  rs12565286 ... rs28729663
```

New data generation

Re-compute PCA

- Calculate principal components to be included as covariates in the GWA models.
  - Adjust for remaining substructure that may confound SNP level association.
  - LD pruning, then ancestry filtering to calculate PCs using the `snpgdsPCA` function on the filtered genotype data set. (using the first 10 principal components in GWA models).

```
## Hint: it is suggested to call `snpgdsOpen' to open a SNP GDS file instead of `openfn.gds'.
```

```
## SNP pruning based on LD:
## Excluding 204,583 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
## Working space: 1,401 samples, 656,890 SNPs
##     using 1 (CPU) core
##     sliding window: 500,000 basepairs, Inf SNPs
##     |LD| threshold: 0.2
##     method: composite
## Chromosome 1: 8.23%, 5,845/71,038
## Chromosome 3: 8.08%, 4,893/60,565
## Chromosome 6: 8.03%, 4,352/54,176
## Chromosome 12: 8.56%, 3,606/42,124
## Chromosome 21: 9.41%, 1,173/12,463
## Chromosome 2: 7.66%, 5,647/73,717
## Chromosome 4: 8.20%, 4,567/55,675
## Chromosome 7: 8.49%, 3,939/46,391
## Chromosome 11: 7.89%, 3,489/44,213
## Chromosome 10: 7.96%, 3,814/47,930
## Chromosome 8: 7.65%, 3,694/48,299
## Chromosome 5: 8.04%, 4,514/56,178
## Chromosome 14: 8.77%, 2,460/28,054
## Chromosome 9: 8.21%, 3,374/41,110
## Chromosome 17: 11.14%, 2,222/19,939
## Chromosome 13: 8.30%, 2,843/34,262
## Chromosome 20: 9.39%, 2,137/22,753
## Chromosome 15: 9.23%, 2,390/25,900
## Chromosome 16: 9.27%, 2,558/27,591
## Chromosome 18: 8.87%, 2,327/26,231
## Chromosome 19: 12.99%, 1,491/11,482
## Chromosome 22: 10.92%, 1,243/11,382
## 72,578 markers are selected in total.
```

```
## 72578
```

```
## Hint: it is suggested to call `snpgdsOpen' to open a SNP GDS file instead of `openfn.gds'.
```

```
## Principal Component Analysis (PCA) on genotypes:
## Excluding 788,895 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
## Working space: 1,401 samples, 72,578 SNPs
##     using 1 (CPU) core
## PCA:    the sum of all selected genotypes (0,1,2) = 32714193
## CPU capabilities: Double-Precision SSE2
## Tue Oct 15 14:06:21 2019    (internal increment: 720)
##
[..................................................]  0%, ETC: ---
[==================================================] 100%, completed in 30s
## Tue Oct 15 14:06:51 2019    Begin (eigenvalues and eigenvectors)
## Tue Oct 15 14:06:52 2019    Done.
```

```
##   FamID          pc1          pc2           pc3           pc4
## 1 10002  0.007764870  0.014480384 -0.0006315881  0.0028664643
## 2 10004 -0.012045108 -0.007231015 -0.0030012896 -0.0107972693
## 3 10005 -0.016702930 -0.005347697  0.0144498361 -0.0006151058
## 4 10007 -0.009537235  0.004556977  0.0026835662  0.0166255657
## 5 10008 -0.015392106 -0.002446933  0.0205087909 -0.0057241772
## 6 10009 -0.015123858 -0.002353917  0.0213604518  0.0069156529
##             pc5          pc6           pc7          pc8          pc9
## 1 -0.0188391406  0.009680646  0.0276468057 -0.006645818 -0.023429747
## 2 -0.0077705400 -0.004645751  0.0018061075 -0.003087891 -0.001833242
## 3  0.0345170160  0.038708551  0.0205790788 -0.012265508  0.003592690
## 4 -0.0002363142  0.005514627  0.0159588869  0.027975455  0.029777180
## 5 -0.0039696226  0.005354244 -0.0007269312  0.027014714  0.010672162
## 6  0.0400677558  0.023222478  0.0152485234  0.013296852  0.022746352
##           pc10
## 1  0.010492314
## 2 -0.004538746
## 3 -0.002287043
## 4 -0.007461255
## 5 -0.003352997
## 6  0.013143889
```

Imputation of SNPs

1. In addition to the genotyped SNPs from the study, impute SNPs on chromosome 16. (as it is useful to extend the analysis to other known SNPs)
2. Performance of genotype imputation requires reference data. Using the HapMap 1000 Genomes data (https://www.internationalgenome.org/category/hapmap/).
    - Derive imputation “rules” for the additional SNPs that were not typed in the study using `snp.imputation` based on the genotypes from the 1000 Genomes data. Each rule represents a predictive model for genotypes of untyped SNPs associated with near-by typed SNPs. Using these rules, calculate the expected posterior value of the non-typed SNPs using the impute function from SNPRelate.
3. Remove un-typed SNPs in which it fails to derive imputation “rules”.
4. Filter out SNPs that have low estimated MAF, and low imputation accuracy.
    - The latter is based on the R2 value of the model estimated by the `snp.imputation` function.

```
##           SNP position A1 A2
## 1 rs140769322    60180  3  2
## 2 rs188810967    60288  2  1
## 3  rs76368850    60291  2  4
## 4 rs185537431    60778  3  1
## 5 rs542544747    60842  2  1
## 6   rs4021615    61349  1  3
```

```
## A SnpMatrix with  99 rows and  377819 columns
## Row names:  CEU_1 ... CEU_99
## Col names:  rs140769322 ... rs111706106
```

```
## A SnpMatrix with  99 rows and  20632 columns
## Row names:  CEU_1 ... CEU_99
## Col names:  rs41340949 ... rs4785775
```

```
## SNPs tagged by a single SNP: 82119
## SNPs tagged by multiple tag haplotypes (saturated model): 115769
```

```
## Imputation rules for 197888 SNPs were estimated
```

```
## 162565 imputation rules remain after imputations with low certainty were removed
```

```
## 162565 imputation rules remain after MAF filtering
```

```
## A SnpMatrix with  1401 rows and  162565 columns
## Row names:  10002 ... 11596
## Col names:  rs560777354;rs80001234 ... rs62053708
```

Genome-wide association analysis

- Regressing each SNP separately on a given trait, adjusted for sample level clinical, environmental, and demographic factors.
    - single additive model employed (Due the large number of SNPs).
- Value of the genotype should reflect the number of minor alleles. However, following conversion of the values will reflect the opposite. To fix this a `flip.matrix` procedure is included in the `GWAA` function, which can be turned on or off using the flip argument.
- Due to the large number of models that require fitting, the GWA analysis can be deployed in parallel across multiple processors or machines to reduce the running time. Here we demonstrate two basic methods for performing parallel processing using the `doParallel` package.

Phenotype data preparation

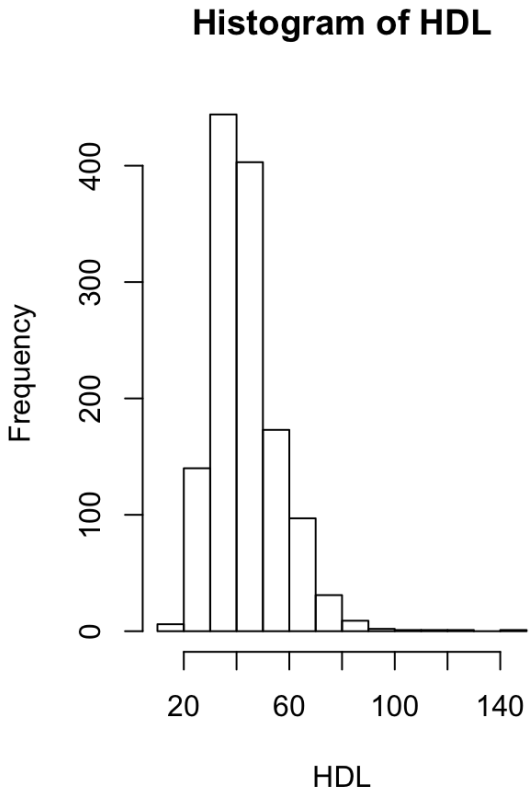
1. Create a data frame of phenotype features
    - clinical features + first ten principal components.
2. The HDL feature is normalized using a rank-based inverse normal transform.
3. Remove unneded variables for the GWA analysis.
4. Remove samples with missing normalized HDL data.

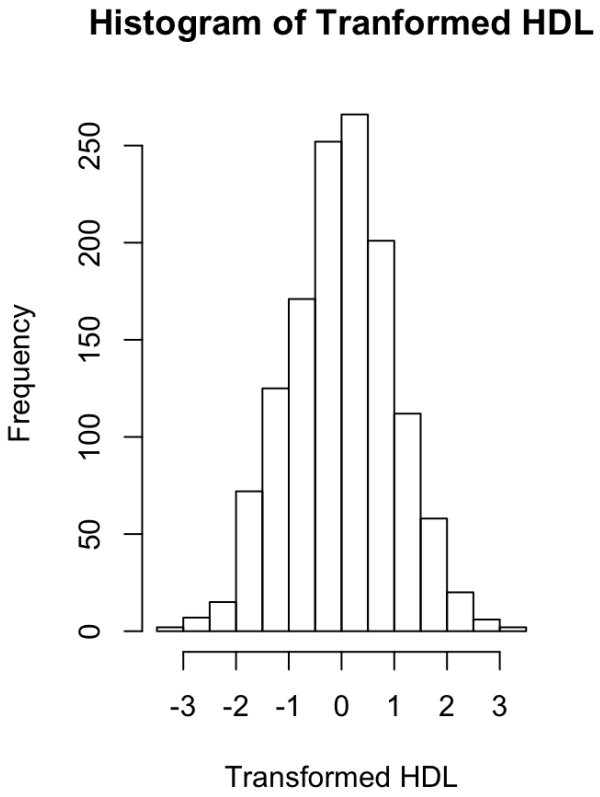
```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     genotype
```

```
## Loading required package: GenABEL.data
```






```
##      id sex age          pc1          pc2          pc3          pc4
## 2 10004   2  50 -0.012045108 -0.007231015 -0.003001290 -0.0107972693
## 3 10005   1  55 -0.016702930 -0.005347697  0.014449836 -0.0006151058
## 4 10007   1  52 -0.009537235  0.004556977  0.002683566  0.0166255657
## 5 10008   1  58 -0.015392106 -0.002446933  0.020508791 -0.0057241772
## 6 10009   1  59 -0.015123858 -0.002353917  0.021360452  0.0069156529
## 7 10010   1  54 -0.012816157  0.005126124  0.014654847 -0.0147082270
##             pc5           pc6           pc7          pc8          pc9
## 2 -0.0077705400 -0.0046457510  0.0018061075 -0.003087891 -0.001833242
## 3  0.0345170160  0.0387085513  0.0205790788 -0.012265508  0.003592690
## 4 -0.0002363142  0.0055146271  0.0159588869  0.027975455  0.029777180
## 5 -0.0039696226  0.0053542437 -0.0007269312  0.027014714  0.010672162
## 6  0.0400677558  0.0232224781  0.0152485234  0.013296852  0.022746352
## 7 -0.0008190769 -0.0003831342 -0.0131606658 -0.013647709 -0.008912913
##           pc10  phenotype
## 2 -0.004538746 -2.2877117
## 3 -0.002287043 -0.4749316
## 4 -0.007461255  0.8855512
## 5 -0.003352997 -0.1644639
## 6  0.013143889  0.3938940
## 7 -0.056187339  1.7109552
```

Parallel model fitting

- Perform model fitting on each of the typed SNPs in the genotype object and write the results to a .txt file.

```
## Loading required package: doParallel
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

```
## Loading required package: parallel
```

```
## 656890  SNPs included in analysis.
## 1309 samples included in analysis.
## socket cluster with 8 nodes on host 'localhost'
## GWAS SNPs 1-65689 (10% finished)
## GWAS SNPs 65690-131378 (20% finished)
## GWAS SNPs 131379-197067 (30% finished)
## GWAS SNPs 197068-262756 (40% finished)
## GWAS SNPs 262757-328445 (50% finished)
## GWAS SNPs 328446-394134 (60% finished)
## GWAS SNPs 394135-459823 (70% finished)
## GWAS SNPs 459824-525512 (80% finished)
## GWAS SNPs 525513-591201 (90% finished)
## GWAS SNPs 591202-656890 (100% finished)
## [1] "Done."
```

```
## Time difference of 1.065868 hours
```

Model fitting of non-typed SNPs

- Association testing on additional SNPs from genotype imputation.
- Perform the analysis based on the imputation “rules” calculated previously.
- The resulting SNPs are combined with the chromosome position information to create a table of SNPs, location and p-value.
- Take –log10 of the p-value for plotting.

```
##          SNP      p.value position chr    type Neg_logP
## 1  rs1532624 9.805683e-08 57005479  16 imputed 7.008522
## 2  rs7205804 9.805683e-08 57004889  16 imputed 7.008522
## 3 rs12446515 1.430239e-07 56987015  16 imputed 6.844591
## 4 rs17231506 1.430239e-07 56994528  16 imputed 6.844591
## 5   rs173539 1.430239e-07 56988044  16 imputed 6.844591
## 6   rs183130 1.430239e-07 56991363  16 imputed 6.844591
```

Mapping associated SNPs to genes

- Using a separate data file containing the chromosome and coordinate locations of each protein coding gene, locate coincident genes and SNPs.
- Use a function to extract the subset of SNPs that are near a gene of interest.
- The SNP with the lowest p-value in both the typed and imputed SNP analysis lies within the boundaries of the cholesteryl ester transfer protein gene, CETP.
- Call the `map2gene` function for “CETP” to filter the imputed genotypes and extract only those SNPs that are near CETP. This will be used for post-analytic interrogation to follow.

Post-analytic visualization and genomic interrogation

- Combine the results, and isolate just those SNPs in the region of interest.
- Following similar steps as for imputed SNPs, the typed SNPs are loaded from a file generated by the `GWAA` function.
- Attach chromosome and position to each SNP, order by significance, and take –log10 of the p-value.

```
##          SNP   Estimate  Std.Error   t.value      p.value Neg_logP chr
## 1  rs1532625  0.2024060 0.03756207  5.388575 8.452365e-08 7.073022  16
## 2   rs247617  0.2119357 0.03985979  5.317030 1.243480e-07 6.905361  16
## 3 rs10945761  0.1856564 0.04093602  4.535282 6.285358e-06 5.201670   6
## 4  rs3803768 -0.3060086 0.06755628 -4.529685 6.451945e-06 5.190309  17
## 5  rs4821708 -0.1816673 0.04020915 -4.518058 6.825085e-06 5.165892  22
## 6  rs9647610  0.1830434 0.04072772  4.494320 7.607161e-06 5.118777   6
##    position
## 1  57005301
## 2  56990716
## 3 162065367
## 4  80872028
## 5  38164106
## 6 162066421
```

- Isolate CETP (https://en.wikipedia.org/wiki/Cholesterylester_transfer_protein)-specific SNPs
- The two tables of typed and imputed genotypes are combined into a single table.
- Concatenate just the SNPs near CETP and display them.

```
##          SNP   Estimate  Std.Error   t.value      p.value Neg_logP chr
## 1  rs1532625  0.2024060 0.03756207  5.388575 8.452365e-08 7.073022  16
## 2   rs247617  0.2119357 0.03985979  5.317030 1.243480e-07 6.905361  16
## 3 rs10945761  0.1856564 0.04093602  4.535282 6.285358e-06 5.201670   6
## 4  rs3803768 -0.3060086 0.06755628 -4.529685 6.451945e-06 5.190309  17
## 5  rs4821708 -0.1816673 0.04020915 -4.518058 6.825085e-06 5.165892  22
## 6  rs9647610  0.1830434 0.04072772  4.494320 7.607161e-06 5.118777   6
##    position  type
## 1  57005301 typed
## 2  56990716 typed
## 3 162065367 typed
## 4  80872028 typed
## 5  38164106 typed
## 6 162066421 typed
```

```
##              SNP Estimate Std.Error t.value   p.value     Neg_logP chr
## 818521 rs62048372       NA        NA      NA 0.9999838 7.048600e-06  16
## 818522  rs8056666       NA        NA      NA 0.9999838 7.048600e-06  16
## 818523  rs8057313       NA        NA      NA 0.9999838 7.048600e-06  16
## 818524  rs8061812       NA        NA      NA 0.9999838 7.048600e-06  16
## 818525  rs9940700       NA        NA      NA 0.9999838 7.048600e-06  16
## 818526 rs13334556       NA        NA      NA 0.9999843 6.825503e-06  16
##        position    type
## 818521 53775940 imputed
## 818522 53794830 imputed
## 818523 53794855 imputed
## 818524 53794856 imputed
## 818525 53795409 imputed
## 818526  5463800 imputed
```

```
##            SNP      p.value   Neg_logP chr position    type gene
## 1    rs1532625 8.452365e-08 7.07302173  16 57005301   typed CETP
## 2     rs289742 3.788738e-04 3.42150548  16 57017762   typed CETP
## 3     rs289715 4.299934e-03 2.36653823  16 57008508   typed CETP
## 4    rs6499863 1.382602e-02 1.85930275  16 56992017   typed CETP
## 5    rs1800777 8.833782e-02 1.05385333  16 57017319   typed CETP
## 6    rs4783962 1.039467e-01 0.98318933  16 56995038   typed CETP
## 7   rs12708980 6.375740e-01 0.19546941  16 57012379   typed CETP
## 8    rs1532624 9.805683e-08 7.00852215  16 57005479 imputed CETP
## 9    rs7205804 9.805683e-08 7.00852215  16 57004889 imputed CETP
## 10  rs17231506 1.430239e-07 6.84459142  16 56994528 imputed CETP
## 11    rs183130 1.430239e-07 6.84459142  16 56991363 imputed CETP
## 12   rs3764261 1.430239e-07 6.84459142  16 56993324 imputed CETP
## 13    rs821840 1.430239e-07 6.84459142  16 56993886 imputed CETP
## 14  rs11508026 1.151771e-06 5.93863373  16 56999328 imputed CETP
## 15  rs12444012 1.151771e-06 5.93863373  16 57001438 imputed CETP
## 16  rs12720926 1.151771e-06 5.93863373  16 56998918 imputed CETP
## 17   rs4784741 1.151771e-06 5.93863373  16 57001216 imputed CETP
## 18  rs34620476 1.155266e-06 5.93731819  16 56996649 imputed CETP
## 19    rs708272 1.155266e-06 5.93731819  16 56996288 imputed CETP
## 20    rs711752 1.155266e-06 5.93731819  16 56996211 imputed CETP
## 21  rs12720922 3.238664e-06 5.48963411  16 57000885 imputed CETP
## 22   rs8045855 3.238664e-06 5.48963411  16 57000696 imputed CETP
## 23  rs12149545 3.245934e-06 5.48866029  16 56993161 imputed CETP
## 24  rs11076175 1.400697e-05 4.85365587  16 57006378 imputed CETP
## 25   rs7499892 1.400697e-05 4.85365587  16 57006590 imputed CETP
## 26   rs1800775 1.747444e-05 4.75759678  16 56995236 imputed CETP
## 27   rs3816117 1.747444e-05 4.75759678  16 56996158 imputed CETP
## 28  rs11076176 1.089765e-04 3.96266723  16 57007446 imputed CETP
## 29    rs289714 1.121002e-04 3.95039374  16 57007451 imputed CETP
## 30    rs158478 2.513994e-04 3.59963575  16 57007734 imputed CETP
## 31   rs9939224 2.868544e-04 3.54233851  16 57002732 imputed CETP
## 32  rs12447620 3.868267e-04 3.41248361  16 57014319 imputed CETP
## 33    rs158480 3.868267e-04 3.41248361  16 57008227 imputed CETP
## 34    rs158617 3.868267e-04 3.41248361  16 57008287 imputed CETP
## 35 rs112039804 4.305196e-03 2.36600705  16 57018856 imputed CETP
## 36  rs12708985 4.305196e-03 2.36600705  16 57014610 imputed CETP
## 37    rs736274 4.305196e-03 2.36600705  16 57009769 imputed CETP
## 38  rs11076174 4.439341e-03 2.35268153  16 57003146 imputed CETP
## 39    rs158479 1.358926e-02 1.86680426  16 57008048 imputed CETP
## 40 rs201825234 1.392675e-02 1.85615030  16 56991948 imputed CETP
## 41   rs2115429 1.392675e-02 1.85615030  16 56992842 imputed CETP
## 42   rs6499861 1.392675e-02 1.85615030  16 56991495 imputed CETP
## 43   rs6499862 1.392675e-02 1.85615030  16 56991524 imputed CETP
## 44    rs289713 1.902194e-02 1.72074521  16 57006829 imputed CETP
## 45  rs12720918 2.238286e-02 1.65008448  16 56994212 imputed CETP
## 46  rs12920974 2.238286e-02 1.65008448  16 56993025 imputed CETP
## 47  rs36229787 3.026885e-02 1.51900413  16 56993897 imputed CETP
## 48    rs820299 4.470355e-02 1.34965802  16 57000284 imputed CETP
## 49    rs289712 4.529779e-02 1.34392301  16 57006305 imputed CETP
## 50  rs34946873 5.624406e-02 1.24992336  16 56991143 imputed CETP
## 51  rs12597002 6.153983e-02 1.21084368  16 57002404 imputed CETP
## 52  rs60545348 6.153983e-02 1.21084368  16 57001985 imputed CETP
## 53    rs708273 6.153983e-02 1.21084368  16 56999949 imputed CETP
## 54   rs4369653 6.333149e-02 1.19838029  16 56997551 imputed CETP
## 55      rs5880 7.129792e-02 1.14692314  16 57015091 imputed CETP
## 56   rs4587963 8.354674e-02 1.07807049  16 56997369 imputed CETP
## 57   rs1800776 9.239564e-02 1.03434852  16 56995234 imputed CETP
## 58    rs289746 9.693910e-02 1.01350102  16 57020205 imputed CETP
## 59  rs12447839 1.042017e-01 0.98212538  16 56993935 imputed CETP
## 60  rs12447924 1.042017e-01 0.98212538  16 56994192 imputed CETP
## 61    rs158477 1.519849e-01 0.81819960  16 57007610 imputed CETP
```

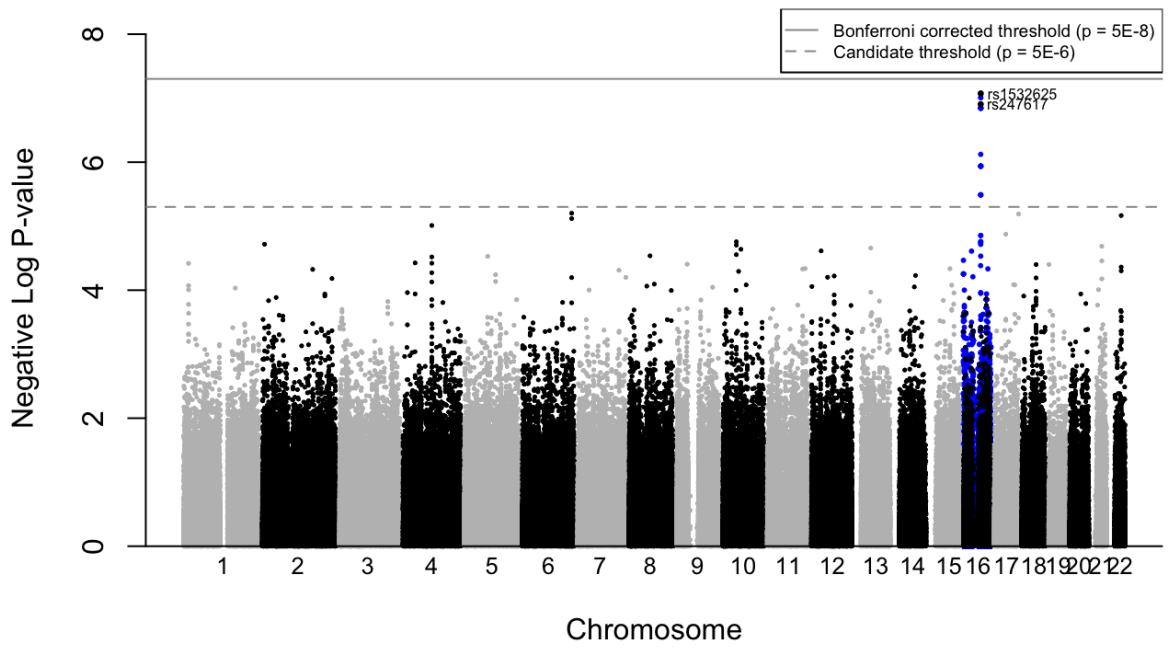
```
## 62  rs12720889 2.755963e-01 0.55972661  16 57012563 imputed CETP
## 63  rs12708983 2.772136e-01 0.55718551  16 57014411 imputed CETP
## 64  rs66495554 2.790835e-01 0.55426586  16 57018636 imputed CETP
## 65  rs12934552 3.156022e-01 0.50085994  16 57021433 imputed CETP
## 66  rs12708968 3.597273e-01 0.44402664  16 56994819 imputed CETP
## 67  rs17245715 3.597273e-01 0.44402664  16 56994990 imputed CETP
## 68   rs4783961 4.335221e-01 0.36298880  16 56994894 imputed CETP
## 69  rs12598522 5.138788e-01 0.28913932  16 57022352 imputed CETP
## 70  rs56315364 5.138788e-01 0.28913932  16 57021524 imputed CETP
## 71 rs117427818 5.582634e-01 0.25316088  16 57010486 imputed CETP
## 72  rs36229786 5.721591e-01 0.24248319  16 56993901 imputed CETP
## 73  rs11860407 6.108898e-01 0.21403710  16 57010828 imputed CETP
## 74   rs2033254 6.108898e-01 0.21403710  16 57009985 imputed CETP
## 75   rs1800774 6.293251e-01 0.20112492  16 57015545 imputed CETP
## 76   rs7405284 6.519531e-01 0.18578366  16 57001275 imputed CETP
## 77  rs12708974 9.096021e-01 0.04114853  16 57005550 imputed CETP
```

Visualization and QC

- Visualize the GWA analysis findings while performing quality control checks.
- Identifying data inconsistencies and potential systemic biases.

Manhattan plot

- plot –log10 of the p-value against SNP position across the entire set of typed and imputed SNPs.
- The plot will show two horizontal lines. The higher of the two is the commonly used “Bonferroni” adjusted significance cut-off of $-\log10(5\times10^{-8})$, while the lower is less stringent (“Candidate”) cut-off of $-\log10(5\times10^{-6})$. Typed and imputed SNPs will be represented by black and blue, respectively. We label the typed SNPs with signals that have surpassed the less stringent cutoff.



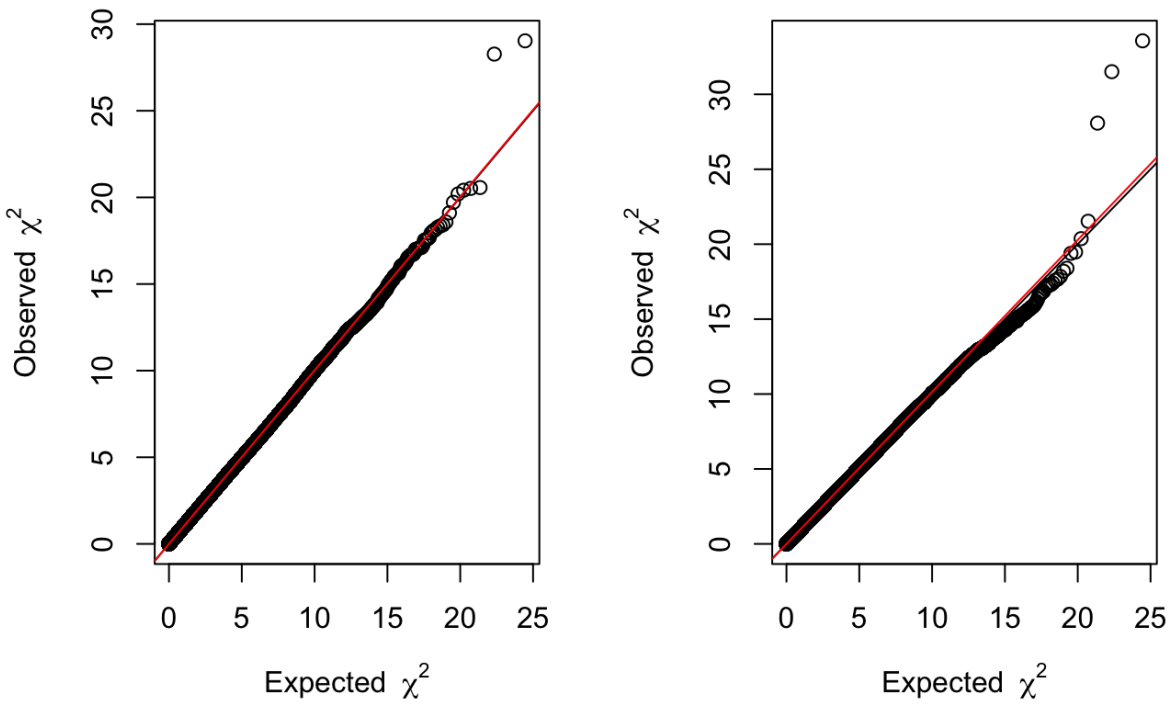
Quantile-quantile plots and the λ-statistic

- Relationship between the expected and observed distributions of SNP level test statistics.

- Compare statistics for the unadjusted model (left) compared with the model adjusted for confounders by incorporating the first ten principal components along with clinical covariates.
- A new set of models is generated with only the phenotype (HDL) and no additional factors.

```
## 656890  SNPs included in analysis.
## 1309 samples included in analysis.
## socket cluster with 8 nodes on host 'localhost'
## GWAS SNPs 1-65689 (10% finished)
## GWAS SNPs 65690-131378 (20% finished)
## GWAS SNPs 131379-197067 (30% finished)
## GWAS SNPs 197068-262756 (40% finished)
## GWAS SNPs 262757-328445 (50% finished)
## GWAS SNPs 328446-394134 (60% finished)
## GWAS SNPs 394135-459823 (70% finished)
## GWAS SNPs 459824-525512 (80% finished)
## GWAS SNPs 525513-591201 (90% finished)
## GWAS SNPs 591202-656890 (100% finished)
## [1] "Done."
```

```
## Time difference of 55.2289 mins
```




```
## Unadjusted lambda: 1.01417377078806
## Adjusted lambda: 1.00214021515844
```

```
## Standardized unadjusted lambda: 1.0108279379588
## Standardized adjusted lambda: 1.00163500012104
```

QA points of attention

- The tail of the distribution is brought closer to the y=x line after accounting for confounding by race/ethnicity in the modeling framework.

- If the data in this figure were shifted up or down from the y=x line, then we would want to investigate some form of systemic bias.
- The degree of deviation from this line is measured formally by the λ-statistic, where a value close to 1 suggests appropriate adjustment for the potential admixture.
- A slight deviation in the upper right tail from the y=x line suggests crudely that some form of association is present in the data. There is only a slight improvement in λ between the unadjusted model and the model with PCs indicating that the population is relatively homogenous.

Heatmap

- Visualize the Linkage Disequilibrium (https://en.wikipedia.org/wiki/Linkage_disequilibrium) pattern between significant SNPs other SNPs in nearby regions.
- Include the most significant SNP from the analysis and other SNPs near CETP.
- The darker shading indicates higher LD. The plot also includes –log10(p) values to illustrate their connection with physical location and LD.

```
## Loading required package: GenomicRanges
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:GenABEL':
##
##     annotation, strand, strand<-
```

```
## The following objects are masked from 'package:Matrix':
##
##     colMeans, colSums, rowMeans, rowSums, which
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##     Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which, which.max,
##     which.min
```

```
## Loading required package: S4Vectors
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:plyr':
##
##     rename
```

```
## The following object is masked from 'package:Matrix':
##
##     expand
```

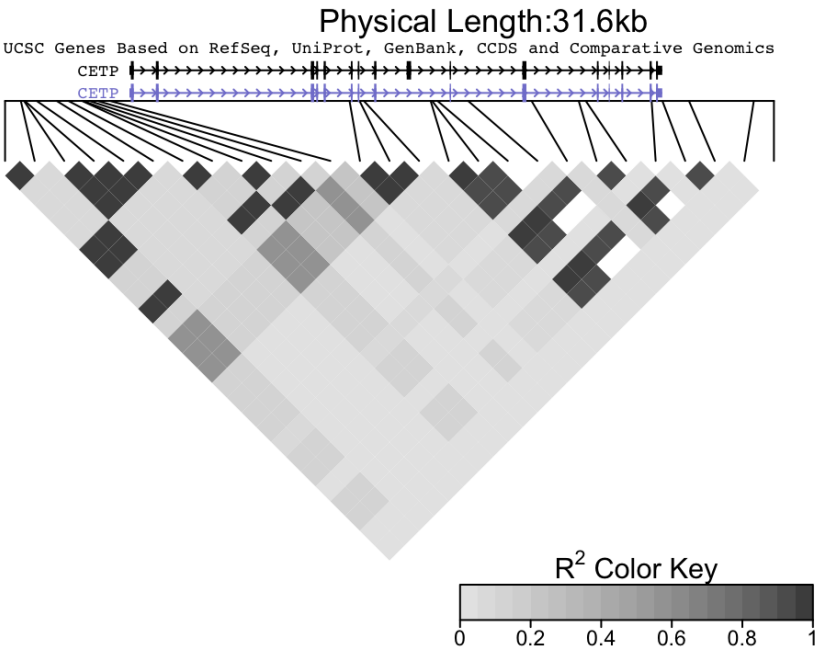
```
## The following object is masked from 'package:base':
##
##     expand.grid
```

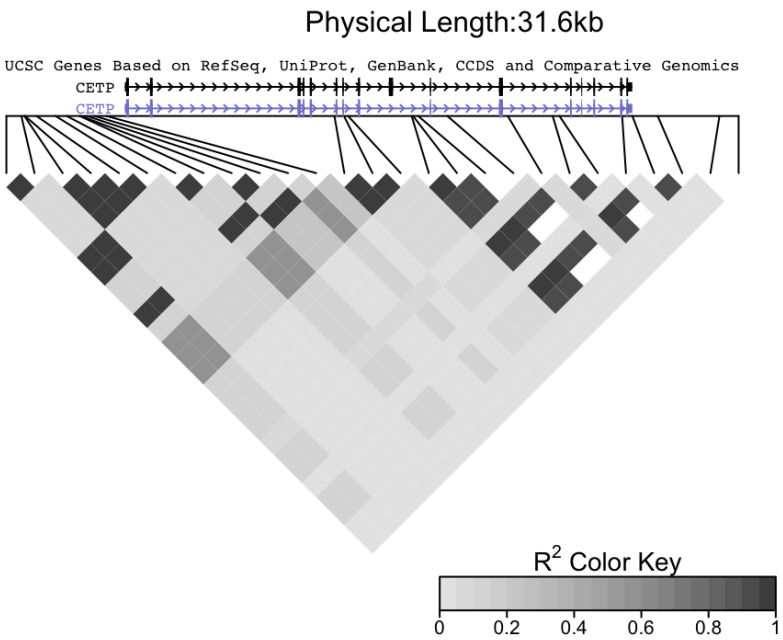
```
## Loading required package: IRanges
```

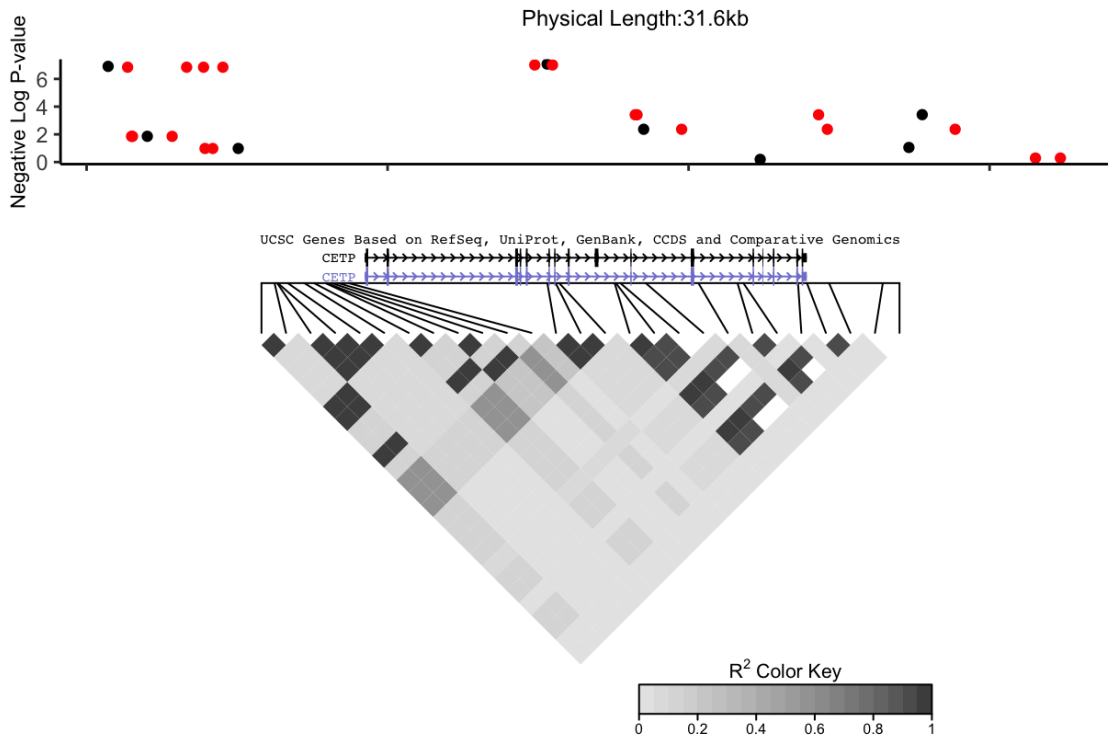
```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:plyr':
##
##     desc
```

```
## Loading required package: GenomeInfoDb
```


Physical Length:31.6kb
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CETP
CETP
R2 Color Key
0
0.2
0.4
0.6
0.8
1


Physical Length:31.6kb
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CETP
CETP
R2 Color Key
0
0.2
0.4
0.6
0.8
1



Regional Association

- Visualization of SNP-wise signal accross a segment of a particular chromsome with the pairwise correlation between SNPs.
- By default it will use the most recent Genome Reference Consortium human genome build.

```
# # Commented out due the issue: Unexpected format to the list of available marts.
# # Bug: https://github.com/merns/postgwas/issues/1
```