BICF Nano Course: GWAS
GWAS Workflow Development using PLINK

Julia Kozlitina
Julia.Kozlitina@UTSouthwestern.edu
April 28, 2017

##

Getting started

##

Open the Terminal (Search -> Applications -> Terminal), and go to your home directory:

```
cd $HOME
```

You can always find which directory you are in by typing:

```
pwd
```

This should say: /home/student

Let's make a directory for writing the output:

```
mkdir gwas
ls          # list directory contents
```

##

I. Software and Datasets

##

This tutorial will use the following software:

PLINK	Command-line genetic analysis toolset
Haploview	Graphical tool for viewing *PLINK* results and SNP analysis
Locuszoom	Graphical tool for visualizing regional association results

The data used in this exercise are from 661 African and 503 European ancestry individuals from the 1000 Genomes project (http://www.internationalgenome.org). From whole genome sequence data, ~60,000 autosomal SNPs were extracted. The genotypes along with a simulated disease status, quantitative phenotype and some covariates are contained in the following files.

`1kg_data.bed`	genotype data for 59217 SNPs and 1164 individuals
`1kg_data.bim`	chromosomal map file for these SNPs
`1kg_data.fam`	family pedigree information
`1kg_data.covar`	additional covariates to be used in analysis
`1kg_data.pheno`	alternative phenotype files

Go to the course data directory by typing at the command prompt:

```
cd ~/course_files/data/
```

To let the machine know where to find PLINK, we need to add it to the current PATH. Type:

```
export PATH=$PATH:/course_files/seqprg/bin
```

Check that PLINK is working by typing:

```
plink --bfile 1kg_data
```

This should start *PLINK* and generate some output describing the options available with the --bfile command. If you get an error message, you are in the wrong directory.

###

II. Some pointers to working with PLINK

###

- PLINK always generates a LOG file, which includes the details of the implemented commands, and any warning messages. It is very useful for checking if the software is successfully completing commands.

- Exact syntax and spelling is **very important**
  e.g., "–-bfile" is not the same as "--bfile"

- PLINK has excellent web documentation (http://pngu.mgh.harvard.edu/purcell/plink/)

###

III. Exploring the BED/BIM/FAM Data Format

###

(a) **BED** is a binary file that contains the genotype information, similar to a standard **PED** file, but in machine-readable format (it takes much less storage space (10%), and allows for faster processing in PLINK). If we could read it, it would contain the genotype data with 1 line per individual and 1 column for each SNP:

```
A A   A A   C C   C C
A A   A A   C C   C C
A A   A A   C C   C C
A A   A A   T C   C C
…
```

Note: PLINK 1.9 autoconverts PED file sets to binary format.

(b) **BIM** file is similar to a **MAP** file, and contains information on the SNPs included in the data set. The first 6 columns are CHR, SNP, cM, Position, Allele 1 (minor), Allele 2 (major). To view the first few lines of the BIM file, type:

```
head 1kg_data.bim
```

which should produce the following output:

```
1      rs75333668    0      762320        T      C
1      rs148711625   0      865584        A      G
1      rs9988179     0      865694        T      C
1      rs116362966   0      879381        T      C
1      rs113383096   0      879481        C      G
1      rs35471880    0      881918        A      G
1      rs3748597     0      888659        T      C
1      rs3828049     0      889238        A      G
…
```

(c) **FAM** file contains the pedigree information, the same as the first 6 columns of a standard **PED** file. It has 6 columns: FID, IID, Paternal ID, Maternal ID, Sex (1 = Male, 2 = Female), and phenotype (1=unaffected (control), 2=affected (case), 0 or -9 = missing).

```
head 1kg_data.fam

HG00096       HG00096       0      0      1      2
HG00097       HG00097       0      0      2      1
HG00099       HG00099       0      0      2      1
HG00100       HG00100       0      0      2      2
```

(d) **Phenotype file.** Instead of the phenotype in the 6th column of FAM file, it is possible to load a different phenotype to the binary file set from a white-space- or tab-delimited file, with at least three columns: FID, IID, Phenotype value, using the option --pheno (additional columns will be ignored unless --pheno-name is specified):

```
plink --bfile 1kg_data --pheno 1kg_data.pheno
```

To view the file, type:

```
head 1kg_data.pheno

FID          IID          Pheno         lPheno
HG00096      HG00096      54.82         4
HG00097      HG00097      57.4          4.05
HG00099      HG00099      24.79         3.21
HG00100      HG00100      31.89         3.46
HG00101      HG00101      17.17         2.84
..
```

(e) **Covariate file.** Covariate files are similar to phenotype files, and contain additional covariates that will be used in analysis. To load the covariates, use the option --covar.

```
plink --bfile 1kg_data --covar 1kg_data.covar
```

```
FID          IID   Sex AGE        PC1        PC2
HG00096      HG00096 1 55 -0.0136039 -0.0147257
HG00097      HG00097 2 63 -0.0131045 -0.0141718
HG00099      HG00099 2 52 -0.0136478 -0.0128483
HG00100      HG00100 2 52 -0.0130089 -0.0139981
HG00101      HG00101 1 37 -0.0130738 -0.0130549
```

(f) To load the BED/BIM/FAM file set, type:

```
plink --bfile 1kg_data
```

Note: In this exercise, it is assumed that these files have passed standard quality control filters: all individuals with a missing data rate of >5%, gender discordance, duplicates, discordant duplicate pairs, and SNPs with a missing rate of >5%, a MAF <0.1% or an HWE *P* value <1 × 10^{-6} have already been excluded. For more details on how to implement these steps, see Anderson et al., 2010.

##

IV. Binary trait (case/control status)

##

Let's change directory to where you want to write the output:

```
cd ~/gwas
```

1. **Basic association test (allelic).** To perform a basic χ^2 test, which compares frequencies of alleles in cases versus controls, type:

```
plink --bfile ~/course_files/data/1kg_data --assoc --out data
```

This will create an output file 'data.assoc'. Open the output file. It has one row per SNP containing the chromosome [CHR], the SNP identifier [SNP], the base-pair location [BP], the minor allele [A1], the frequency of the minor allele in the affected/cases [F_A] and unaffected/controls [F_U], the major allele [A2] and statistical data for an allelic association test including the χ^2 test statistic [CHISQ], the asymptotic *P*-value [P] and the estimated OR for association between the minor allele and disease [OR].

```
CHR         SNP         BP   A1       F_A       F_U   A2      CHISQ          P         OR
1   rs75333668      762320    T   0.06117   0.05626    C     0.2529      0.615      1.093
1  rs148711625      865584    A   0.02039    0.0245    G     0.4488     0.5029     0.8288
1    rs9988179      865694    T   0.01631   0.01089    C      1.259     0.2618      1.506
1  rs116362966      879381    T   0.01713   0.01633    C    0.02226     0.8814       1.05
1  rs113383096      879481    C   0.04323   0.03811    G     0.3883     0.5332       1.14
1   rs35471880      881918    A   0.02529   0.02813    G     0.1812     0.6703     0.8962
…
```

Note: this test assumes HWE, and may not work optimally when genotype frequencies deviate from HWE in cases or controls. Use only as a descriptive summary.

2. **Association between genotype frequencies and disease status.** When there are no covariates to consider, carry out a simple χ^2 test of association which compares genotype frequencies in cases versus controls, by using the --model option:

```
plink --bfile ~/course_files/data/1kg_data --model --out data
```

This command will perform the test of association under several genetic models:

- Genotypic (2 df) test
- Cochran-Armitage trend test (additive model)
- Allelic test (1df)
- Dominant gene action (1df) test
- Recessive gene action (1df) test

Open the output file ‘data.model’. It contains five rows per SNP, one for each of the association tests described in **table 2**. Each row contains the chromosome [CHR], the SNP identifier [SNP], the minor allele [A1], the major allele [A2], the test performed [TEST: GENO (genotypic association); TREND (Cochran-Armitage trend); ALLELIC (allelic association); DOM (dominant model); and REC (recessive model)], the cell frequency counts for cases [AFF] and controls [UNAFF], the χ^2 test statistic [CHISQ], the degrees of freedom for the test [DF] and the asymptotic *P* value [*P*].

```
 CHR          SNP   A1   A2     TEST          AFF       UNAFF      CHISQ   DF          P
   1   rs75333668    T    C     GENO     3/69/541    2/58/491         NA   NA         NA
   1   rs75333668    T    C    TREND      75/1151     62/1040     0.2492    1     0.6176
   1   rs75333668    T    C  ALLELIC      75/1151     62/1040     0.2529    1      0.615
   1   rs75333668    T    C      DOM       72/541      60/491         NA   NA         NA
   1   rs75333668    T    C      REC        3/610       2/549         NA   NA         NA
   1  rs148711625    A    G     GENO     0/25/588    0/27/524         NA   NA         NA
   1  rs148711625    A    G    TREND      25/1201     27/1075     0.4593    1      0.498
   1  rs148711625    A    G  ALLELIC      25/1201     27/1075     0.4488    1     0.5029
```

Note: Genotypic, dominant and recessive tests will not be conducted if any one of the cells in the table of case control by genotype counts contains less than five observations. This is because the χ^2 approximation may not be reliable when cell counts are small. To change the behavior, use the ‘--cell’ option. For example, to lower the threshold to 3, one would type

```
plink --bfile ~/course_files/data/1kg_data --model --cell 3  --out data
```

3. Another option for small counts is to use Fisher’s exact test. Type

```
plink --bfile ~/course_files/data/1kg_data --model fisher --out fisher
```

This will create an output file ‘fisher.model’.

```
CHR          SNP   A1   A2     TEST          AFF       UNAFF          P
   1   rs75333668    T    C     GENO     3/69/541    2/58/491      0.904
   1   rs75333668    T    C    TREND      75/1151     62/1040     0.6176
   1   rs75333668    T    C  ALLELIC      75/1151     62/1040     0.6595
   1   rs75333668    T    C      DOM       72/541      60/491     0.7113
   1   rs75333668    T    C      REC        3/610       2/549          1
   1  rs148711625    A    G     GENO     0/25/588    0/27/524     0.5703
   1  rs148711625    A    G    TREND      25/1201     27/1075      0.498
   1  rs148711625    A    G  ALLELIC      25/1201     27/1075     0.5748
```

Warning: still reports Cochran-Armitage test results under allelic test (Chi-square, 1df)

4. When there are covariates (usually sex, age, principal components of ancestry), perform association tests using logistic regression:

```
plink --bfile ~/course_files/data/1kg_data --logistic --hide-covar --covar
~/course_files/data/1kg_data.covar --out data
```

By default, this command performs a test of association assuming a multiplicative model. To specify a genotypic, dominant or recessive model in place of a multiplicative model, include the model option --genotypic, --dominant or --recessive, respectively. To include sex as a covariate, include the option --sex (in our case, sex is included in the covariate file, so will be automatically used).

Open the output file ‘data.assoc.logistic’.

```
CHR          SNP         BP   A1       TEST    NMISS         OR         STAT            P
   1   rs75333668     762320    T        ADD     1164     0.9238      -0.4272       0.6692
   1  rs148711625     865584    A        ADD     1164     0.6893       -1.284       0.1992
   1    rs9988179     865694    T        ADD     1164      1.333       0.7688        0.442
   1  rs116362966     879381    T        ADD     1164     0.8677      -0.4269       0.6695
   1  rs113383096     879481    C        ADD     1164     0.9491      -0.2387       0.8113
   1   rs35471880     881918    A        ADD     1164       1.08       0.2933       0.7693
```

###

V. Data visualization and interpretation

###

(a) **Quantile-quantile plots.** To create a quantile-quantile plot of p-values, follow these steps.

i. Start R software.
ii. To create a q-q plot based on the results of chi-square tests (performed in IV.2 above), type the following at the command line:

```
data <- read.table("data.model", header=TRUE);
obs <- -log10(sort(data[data$TEST == "TREND", ]$P));
exp <- -log10(c(1:length(obs))/(length(obs) + 1));
pdf("pvalue.chisq.qq.plot.pdf");
plot(exp, obs, ylab="Observed(-logP)", xlab="Expected(-logP)", ylim=c(0, 8),
xlim=c(0,6));
abline(a=0, b=1, col=1, lwd=1.5, lty=2);
dev.off()
```

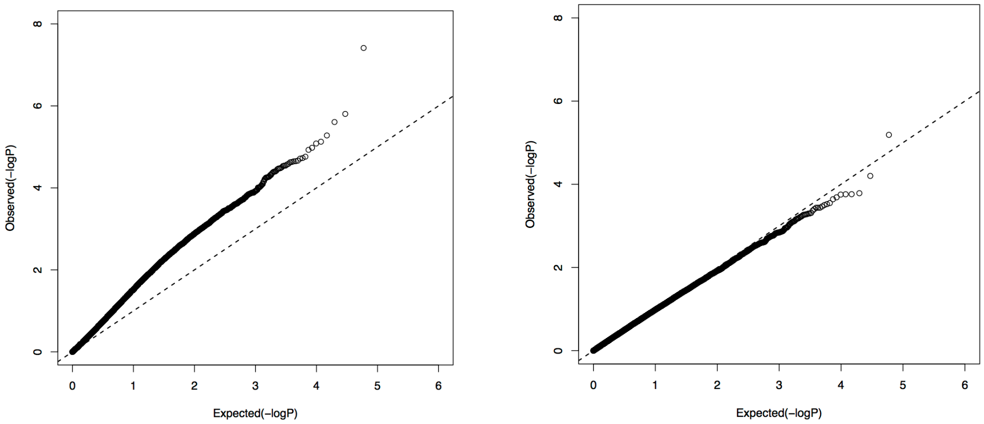
Open the file `"pvalue.chisq.qq.plot.pdf"`. What do you think about this plot?

iii. Now generate a similar plot based on the results of logistic regression analysis.

```
data <- read.table("data.assoc.logistic", header=TRUE);
obs <- -log10(sort(data[data$TEST == "ADD", ]$P));
exp <- -log10(c(1:length(obs))/(length(obs) + 1));
pdf("pvalue.logistic.qq.plot.pdf");
plot(exp, obs, ylab="Observed(-logP)", xlab="Expected(-logP)", ylim=c(0, 8),
xlim=c(0,6));
abline(a=0, b=1, col=1, lwd=1.5, lty=2);
dev.off()
q()
```

Open the file. What do you think about this plot?

Figure 1: Q-Q plots based on association analysis of a binary trait.



(b) Calculate the genomic control inflation factor λ for GWa studies.

(i) To obtain the inflation factor, include the --adjust option in any of the PLINK commands described in Step 4. For example, the inflation factor based on logistic regression assuming a multiplicative model is obtained by typing

```
plink --bfile ~/course_files/data/1kg_data --logistic --hide-covar --
covar ~/course_files/data/1kg_data.covar --adjust --out data
```

(ii) Open the PLINK log file 'data.log', which records the inflation factor. The inflation factor for our GWA study is 1.0078, indicating that no population stratification is detected in our GWA data (values <1.1 are considered "acceptable")

(iii) GC adjustment is based on the median p-value, and does not capture other features of the distribution (e.g., tail behavior), so can over- or under-correct. Use a diagnostic (to detect if there is evidence of population stratification) rather than to correct p-values.

(c) Manhattan plots.

(i) Start Haploview (java -jar Haploview.jar). In the 'Welcome to Haploview' window, select the 'PLINK Format' tab. Click the 'browse' button and select the SNP association output file created in Step IV. We select association results from the file 'data.assoc.logistic'. Select the corresponding MAP file, which will be the '.bim' file for the binary file format.

We select our GWA study file ‘1kg_data.bim’. Leave other options as they are (ignore pairwise comparison of markers > 500 kb apart and exclude individuals with > 50% missing genotypes). Click ‘OK’.

(ii) Select the association results relevant to the test of interest by selecting ‘TEST’ in the dropdown tab to the right of ‘Filter:’, ‘ = ’ in the dropdown menu to the right of that and the PLINK keyword corresponding to the test of interest in the window to the right of that. We select PLINK keyword ‘ADD’ to visualize results for allelic tests of association in our GWA study. Click the gray ‘Filter’ button. Click the gray ‘Plot’ button. Leave all options as they are so that ‘Chromosomes’ is selected as the ‘X-Axis’. Choose ‘P’ from the drop-down menu for the ‘Y-Axis’ and ‘-log10’ from the corresponding dropdown menu for ‘Scale:’. Click ‘OK’ to display the Manhattan plot.

(iii) To save the plot as a scalable vector graphics file, click the button ‘Export to scalable vector graphics:’ and then click the ‘Browse’ button (immediately to the right) to select the appropriate title and directory. **Or, after the plot is generated, right click with your mouse and choose “Save as…” from the menu, to save the graph as a PNG file.**

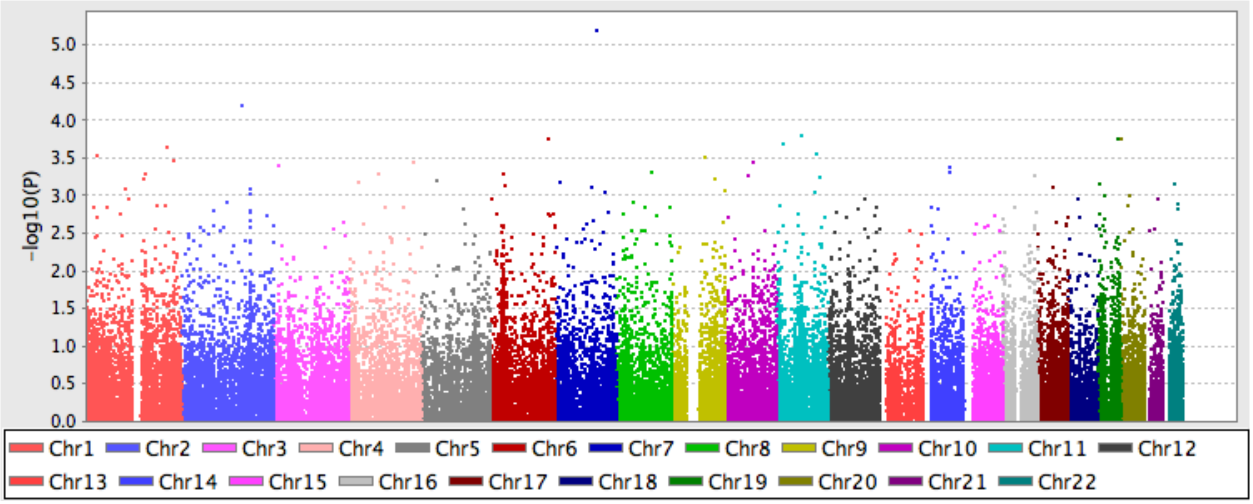


Figure 2: Manhattan plot.

##

VI. Quantitative traits

##

(a) **Basic quantitative trait association.** To load a quantitative phenotype, use the option --pheno. To obtain a basic association test between genotype and a quantitative trait, type:

```
   plink --bfile ~/course_files/data/1kg_data --pheno
~/course_files/data/1kg_data.pheno --assoc --out data
```

This will generate the file ‘data.qassoc’, with the following columns:

```
CHR           SNP         BP    NMISS       BETA         SE         R2        T            P
   1   rs75333668     762320     1164     -3.015      1.785    0.00245   -1.689      0.09141
   1  rs148711625     865584     1164     -2.871      2.899  0.0008431  -0.9902       0.3223
   1    rs9988179     865694     1164     -2.285      3.664  0.0003345  -0.6236        0.533
   1  rs116362966     879381     1164    -0.9656       3.33  7.238e-05    -0.29       0.7719
   1  rs113383096     879481     1164     0.7891       2.14   0.000117   0.3688       0.7124
   1   rs35471880     881918     1164      7.032      2.573   0.006384    2.732     0.006383
```

(b) As with a binary trait, we typically want to include covariates (such as age, gender and ancestry). To do that, use linear regression (--linear) to test the association.

```
   plink --bfile ~/course_files/data/1kg_data --linear --pheno
~/course_files/data/1kg_data.pheno --pheno-name Pheno --hide-covar --covar
~/course_files/data/1kg_data.covar --out data
```

Open the file “data.assoc.linear”.

```
CHR           SNP         BP  A1       TEST    NMISS        BETA         STAT            P
   1   rs75333668     762320    T        ADD     1164      -1.031       -0.551       0.5818
   1  rs148711625     865584    A        ADD     1164      -1.042      -0.3552       0.7225
   1    rs9988179     865694    T        ADD     1164     -0.5524      -0.1503       0.8805
   1  rs116362966     879381    T        ADD     1164      0.8358       0.2487       0.8036
   1  rs113383096     879481    C        ADD     1164       2.935        1.332       0.1832
   1   rs35471880     881918    A        ADD     1164       5.106        1.936      0.05315
```

(c) Generate a q-q plot of the results in R. Start R software.

```
data <- read.table("data.assoc.linear", header=TRUE);
obs <- -log10(sort(data[data$TEST == "ADD", ]$P));
exp <- -log10(c(1:length(obs))/(length(obs) + 1));
pdf("pvalue.linear.qq.plot.pdf");
plot(exp, obs, ylab="Observed(-logP)", xlab="Expected(-logP)", ylim=c(0, max(obs)),
 xlim=c(0,6));
abline(a=0, b=1, col=1, lwd=1.5, lty=2);
dev.off()
```

What do you think?

(d) What are the assumptions of linear regression analysis? What was the distribution of the quantitative trait? Generate a normal q-q plot.

```
pheno <-read.table('~/course_files/data/1kg_data.pheno', h=T); dim(pheno)
pheno[1:2,]

pdf("Normal.qq.plot.pheno.pdf");
qqnorm(pheno$Pheno); qqline(pheno$Pheno)
dev.off()

pdf("Normal.qq.plot.logpheno.pdf");
qqnorm(pheno$lPheno); qqline(pheno$lPheno)
dev.off()
```

To quit R, type:

```
q()
```

(e) Now re-run the association analysis using a log-transformed phenotype. Create a new q-q plot and compare the results.

```
plink --bfile ~/course_files/data/1kg_data --linear --pheno
~/course_files/data/1kg_data.pheno --pheno-name lPheno --hide-covar --covar
~/course_files/data/1kg_data.covar --out data2
```

```
data <- read.table("data2.assoc.linear", header=TRUE);
obs <- -log10(sort(data[data$TEST == "ADD", ]$P));
exp <- -log10(c(1:length(obs))/(length(obs) + 1));
pdf("pvalue.linear.logpheno.qq.plot.pdf");
plot(exp, obs, ylab="Observed(-logP)", xlab="Expected(-logP)", ylim=c(0, max(obs)),
 xlim=c(0,6));
abline(a=0, b=1, col=1, lwd=1.5, lty=2);
dev.off()
```

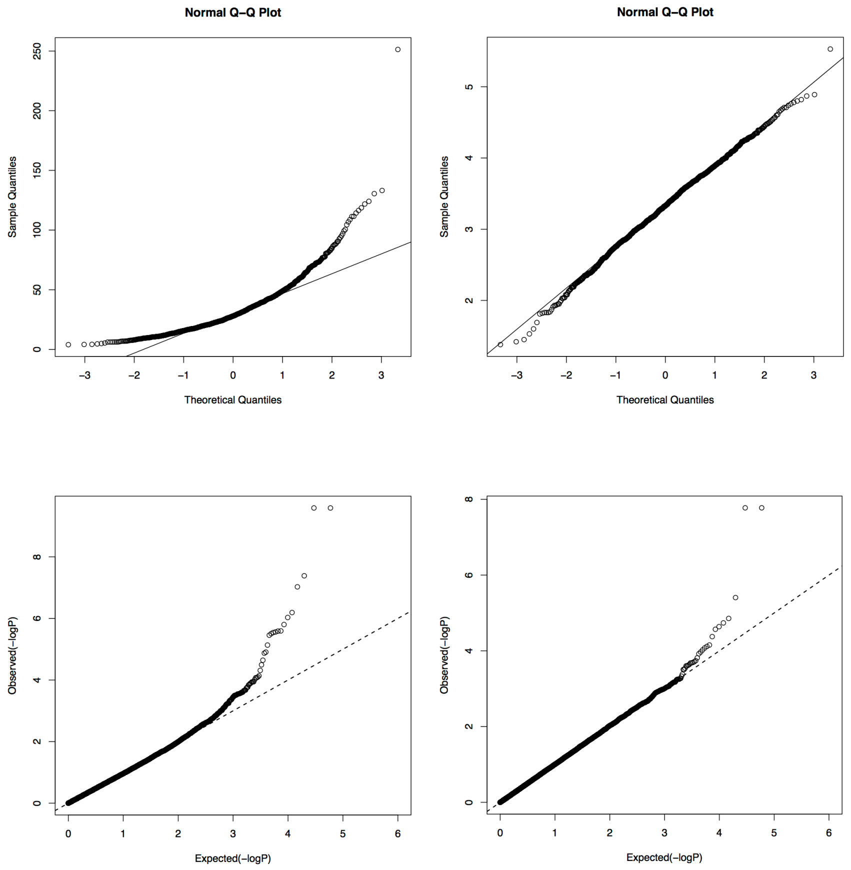


Figure 3: Top panels: Normal q-q plot of raw phenotype data (top left) and log-transformed values (top right). Lower panels: q-q plots of p-values based on the association analysis of raw phenotype data (lower left) and log-transformed values (lower right).

(f) Generate a Manhattan plot and create a plot of regional association results for the top hit.

```
data[order(data$P), ][1:10,]
```

```
CHR         SNP        BP A1  TEST NMISS    BETA   STAT         P
 16   rs1421085  53800954  C   ADD  1164  0.1754  5.683 1.678e-08
 16   rs1558902  53803574  A   ADD  1164  0.1754  5.683 1.678e-08
 16   rs9941349  53825488  T   ADD  1164  0.1295  4.637 3.931e-06
 11   rs6590705 133334522  A   ADD  1164 -0.1640 -4.363 1.396e-05
 16 rs17817449  53813367  G   ADD  1164  0.1043  4.302 1.832e-05
```

Go to Locuszoom website: http://locuszoom.sph.umich.edu/locuszoom/

- Click on “Plot Using your data”
- Choose file: “data2.assoc.linear”
- P-Value Column Name: P
- Marker Column Name: SNP
- Column Delimiter: WhiteSpace
- SNP Reference Name: rs1421085
- Choose Genome Build/LD Population (we will leave EUR)
- Click on “Plot the data” at the bottom of the page.

Figure 4: Manhattan plot of association results for a quantitative trait.

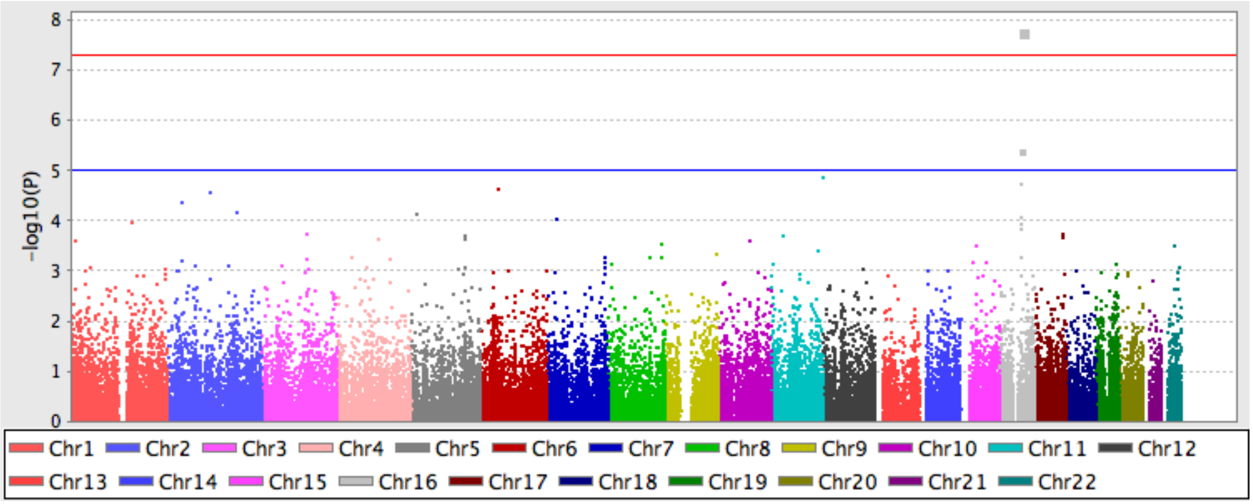
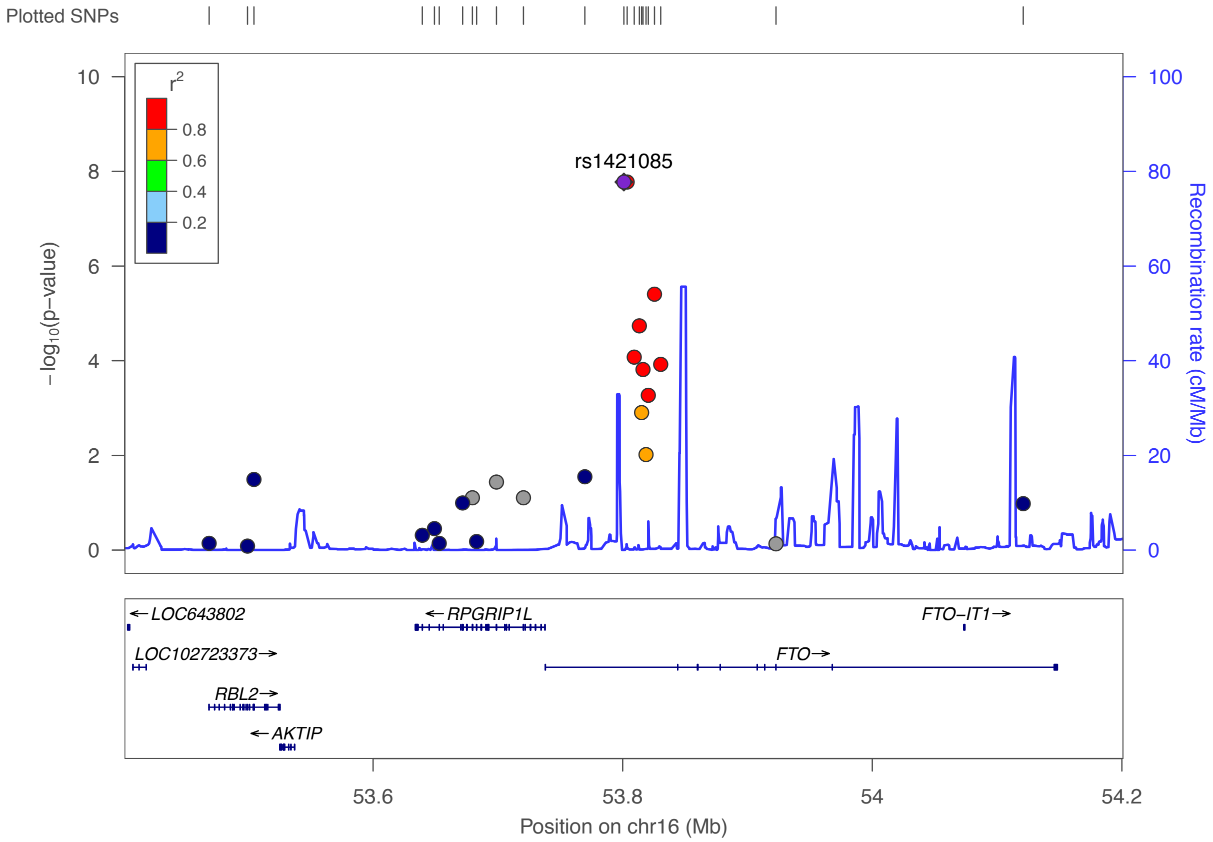


Figure 5: Plot of regional association results around the top SNP.



References:

Anderson et al. (2010) Data quality control in genetic case-control association studies. Nature Protocols, 5(9), 1564.

Clarke et al. (2011). Basic statistical analysis in genetic case-control studies. Nature Protocols, 6(2), 121.