**BICF Nanocourse: Genome Analysis**

**Workshop for: Exome-/genome-sequencing in population based studies**

Julia Kozlitina

[Julia.Kozlitina@UTSouthwestern.edu](mailto:Julia.Kozlitina@UTSouthwestern.edu)

May 2, 2019

**Today we are going to:**

* Convert VCF file into format suitable for association analysis
* Perform QC of population sequencing data
* Perform association analysis
  + Single-variant test for common variants
  + Gene-based test for rare variants
* Summarize and assess the quality of the results

**This tutorial will use the following software:**

PLINK (<https://www.cog-genomics.org/plink2>)

Command-line genetic analysis toolset

Haploview (<https://www.broadinstitute.org/haploview/haploview>)

Graphical tool for viewing PLINK results and SNP analysis

Locuszoom ([locuszoom.org/](http://locuszoom.org/))

Graphical tool for visualizing regional association results

EPACTS (<https://genome.sph.umich.edu/wiki/EPACTS>)

versatile software pipeline to perform various statistical tests for identifying genome-wide association from sequence data

#############################################################################

1. **Getting started**

#############################################################################

Log into BioHPC and log into the compute node

* Set up a WebGUI (<https://portal.biohpc.swmed.edu/terminal/webgui>) session on BioHPC
* Launch via "connect with VNC client", open using TurboVNC (<https://sourceforge.net/projects/turbovnc/>).
* You can also launch the session by "connect via web" but copying and pasting may not work under this mode.
* Open a terminal window - you should be in your home directory.

/home2/trainXX

Now prepare the environment and data

1. Copy session4 material into your directory and work from there

cp -r /archive/nanocourse/genome\_analysis/shared/session4 .

cd session4

1. Load the necessary modules

module load R/3.4.1-gccmkl

module load locuszoom/1.4

module load epacts/3.3.2

export PATH=~/session4/bin/:$PATH

Check that PLINK is working by typing:

plink

This will provide a description of PLINK, basic syntax example and a list of some commands.

#############################################################################

**2. Datasets**

#############################################################################

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 the whole-genome sequencing data, a subset of ~171,000 bi-allelic SNPs, mostly in exonic regions, was extracted. The genotypes along with a simulated disease status, quantitative phenotype and some covariates are contained in the following files.

1kg\_Exome.vcf.gz genotype data for 1164 individuals

1kg\_data.covar additional covariates to be used in analysis

1kg\_data.BIN.pheno case-control status

1kg\_data.QT.pheno quantitative phenotype

1kg\_sample\_info.txt sample information

###############################################################################

**3. Explore the data and convert VCF to PLINK BED/BIM/FAM Format**

###############################################################################

* 1. Go to the data directory by typing at the command prompt:

cd ~/session4/data

* 1. Check the VCF file

gunzip -c 1kg\_Exome.vcf.gz | head -265 | cut –f 1-10

* 1. Convert the VCF file to PLINK format for QC and analysis:

plink --vcf 1kg\_Exome.vcf.gz --keep-allele-order --double-id --make-bed --out 1kg\_data

The --keep-allele-order option keeps the REF and ALT alleles as defined in the VCF file. PLINK by default forces the more common allele to be REF allele (A2), and the less common allele to be ALT allele (A1), regardless of which is REF and ALT in the VCF. To learn more, type:

plink --help --keep-allele-order

The --make-bed command above will produce the following output files:

1kg\_data.bed genotype data in binary format

1kg\_data.bim chromosomal map file for SNPs included in .bed file

1kg\_data.fam family pedigree information

1kg\_data.log log file containing all the commands and options

1. 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

…

1. BIM file contains information on the SNPs included in the .bed file. 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 rs201186828 0 865545 A G

1 rs148711625 0 865584 A G

1 rs146327803 0 865625 A G

…

To see how many variants are in the genotype file:

wc -l 1kg\_data.bim

1. FAM file contains the pedigree information, the same as the first 6 columns of a standard PED file. It has 6 columns: family ID, individual ID, paternal ID, maternal ID, sex (1 = Male, 2 = Female, 0 = unknown), and phenotype (1=unaffected control, 2=affected case, 0 or -9 = missing).

head 1kg\_data.fam

HG00096 HG00096 0 0 0 -9

HG00097 HG00097 0 0 0 -9

HG00099 HG00099 0 0 0 -9

HG00100 HG00100 0 0 0 -9

HG00101 HG00101 0 0 0 -9

…

Notice that sex variable is set to unknown for all individuals (since this information was not provided in the VCF). We can update this information using the following command:

plink --bfile 1kg\_data --keep-allele-order --update-sex 1kg\_sample\_info.txt 3 --make-bed --out 1kg\_data\_temp

head 1kg\_data\_temp.fam

HG00096 HG00096 0 0 1 -9

HG00097 HG00097 0 0 2 -9

HG00099 HG00099 0 0 2 -9

HG00100 HG00100 0 0 2 -9

HG00101 HG00101 0 0 1 -9

1. **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.BIN.pheno

To view the file, type:

head 1kg\_data.BIN.pheno

head 1kg\_data.QT.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

..

1. **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

head 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

#############################################################################

**4. 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

PLINK 1.07: <http://pngu.mgh.harvard.edu/purcell/plink/>

PLINK 2.0: <https://www.cog-genomics.org/plink2>

#############################################################################

**5. Data QC**

#############################################################################

**Note:** In this exercise, we assume that samples have already undergone standard quality control steps: all gender discordant samples, duplicates, discordant duplicate pairs, have been excluded. These steps can be implemented using the commands below. For more details, see Anderson et al., 2010 [PMID: 21085122].

plink --bfile 1kg\_data --check-sex --out 1kg\_data

plink --bfile 1kg\_data --het --out 1kg\_data

5.1 Exclude SNPs with a missing genotype call rate of >10% (--geno 0.1) and individuals with a missing genotype call rate of >10% across SNPs (--mind 0.1).

plink --bfile 1kg\_data\_temp --keep-allele-order --geno 0.1 --mind 0.1 --make-bed --out 1kg\_data\_temp2

5.2 Exclude SNPs with minor allele count (mac) <5 (since single-variant tests have low power to detect the effects of extremely rare variants).

plink --bfile 1kg\_data\_temp2 --keep-allele-order --mac 5 --make-bed --out 1kg\_data\_temp3

5.3 Compute HWE p-values (we have to do it separately for AFR and EUR):

plink --bfile 1kg\_data\_temp3 --keep-allele-order --filter 1kg\_sample\_pop.txt AFR --hardy midp gz --out 1kg\_AFR

This will produce a file 1kg\_AFR.hwe.gz. Extract all SNPs with P(HWE) < 1e-6:

gunzip -c 1kg\_AFR.hwe.gz | awk '{if($9 <= 1e-6) print $0}' > SNPs\_fail\_HWE\_AFR.txt

head SNPs\_fail\_HWE\_AFR.txt

awk '{print $2}' SNPs\_fail\_HWE\_AFR.txt > SNPs\_fail\_AFR.txt

Follow the same steps for EUR population:

plink --bfile 1kg\_data\_temp3 --keep-allele-order --filter 1kg\_sample\_pop.txt EUR --hardy midp gz --out 1kg\_EUR

gunzip -c 1kg\_EUR.hwe.gz | awk '{if($9 <= 1e-6) print $0}' > SNPs\_fail\_HWE\_EUR.txt

awk '{print $2}' SNPs\_fail\_HWE\_EUR.txt > SNPs\_fail\_EUR.txt

Combine the two lists:

cat SNPs\_fail\_AFR.txt SNPs\_fail\_EUR.txt > SNPs\_fail\_HWE.txt

Filter out the failed SNPs:

plink --bfile 1kg\_data\_temp3 --keep-allele-order --exclude SNPs\_fail\_HWE.txt --make-bed --out 1kg\_data.pass

Remove all the temp files:

rm \*temp\*

5.4. (Optional) Summarize the allele frequencies:

plink --bfile 1kg\_data --keep-allele-order --filter 1kg\_sample\_pop.txt AFR --freq gz --out 1kg\_AFR

plink --bfile 1kg\_data --keep-allele-order --filter 1kg\_sample\_pop.txt EUR --freq gz --out 1kg\_EUR

#############################################################################

**6. PCA**

#############################################################################

We will use pre-computed PCs included in the 1kg\_data.covar file for analysis. To calculate principal components, you could follow the steps below. For more details, see Anderson et al. (2010).

1. Extract high-quality independent variants (i.e., not in linkage disequilibrium) from the QC’d genotype file:

#plink --bfile 1kg\_data.pass --geno 0.05 --maf 0.05 --indep-pairwise 50 5 0.2 --out 1kg\_data.pass

This will produce a list of variants to include and to exclude:

1kg\_data.prune.in

1kg\_data.prune.out

1. Extract the pruned-in variants from your genotype file and, optionally, from a reference file (e.g. 1000 genomes):

#plink --bfile 1kg\_data.pass --extract 1kg\_pca\_snps.txt --make-bed --out 1kg\_pca\_temp

#plink --bfile 1000G\_data\_full --extract 1kg\_data.prune.in --make-bed --out 1000G\_pca

#plink --bfile 1000G\_pca --bmerge 1kg\_pca\_temp --make-bed --out 1kg\_ merged

1. Compute the PC’s:

#plink --bfile 1kg\_merged --pca --out 1kg\_data\_pca

This will produce two files containing PCA summary and the first 20 PC scores for each individual in the sample:

1kg\_data.eigenval ## Summary

1kg\_data.eigenvec ## PC’s

#############################################################################

**7. Association analysis for binary trait (case/control status)**

#############################################################################

Create a directory to write plink output:

cd ~/session4

mkdir plink\_out

Update the phenotype value:

plink --bfile ~/session4/data/1kg\_data.pass --pheno ~/session4/data/1kg\_data.BIN.pheno --make-bed --out ~/session4/data/1kg\_data.pass

1. **Basic association test (allelic).** To perform a basic χ2 test, which compares frequencies of alleles in cases versus controls, type:

cd ~/session4/plink\_out

plink --bfile ~/session4/data/1kg\_data.pass --assoc --out data

This will create an output file ‘data.assoc’. 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 χ2 test statistic [CHISQ], the asymptotic *P-*value [P] and the estimated OR for association between the minor allele and disease [OR].

head data.assoc

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 rs146327803 865625 A 0.004078 0.001815 G 0.9918 0.3193 2.252

1 rs41285790 865628 A 0.001631 0.002722 G 0.3223 0.5702 0.5986

1 rs9988179 865694 T 0.01631 0.01089 C 1.259 0.2618 1.506

1 rs116730894 865700 T 0.003263 0.001815 C 0.4732 0.4915 1.8

1 rs149677938 874456 A 0.002447 0.001815 G 0.1082 0.7422 1.349

**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.

1. **Association between genotype frequencies and disease status.** When there are no covariates to consider, carry out a simple χ2 test of association which compares genotype frequencies in cases versus controls, by using the --model option:

plink --bfile ~/session4/data/1kg\_data.pass --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

This creates 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 χ2 test statistic [CHISQ], the degrees of freedom for the test [DF] and the asymptotic *P* value [*P*].

head data.model

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

1 rs148711625 A G DOM 25/588 27/524 NA NA NA

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 χ2 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 ~/session4/data/1kg\_data.pass --model --cell 3 --out data

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

plink --bfile ~/session4/data/1kg\_data.pass --model fisher --out fisher

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

head 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

1 rs148711625 A G DOM 25/588 27/524 0.5703

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

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

plink --bfile ~/session4/data/1kg\_data.pass --logistic --hide-covar --covar ~/session4/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).

head 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 rs146327803 865625 A ADD 1164 1.809 0.7036 0.4817

1 rs41285790 865628 A ADD 1164 0.8182 -0.2181 0.8273

1 rs9988179 865694 T ADD 1164 1.333 0.7688 0.442

1 rs116730894 865700 T ADD 1164 1.631 0.5621 0.5741

1 rs149677938 874456 A ADD 1164 1.124 0.1277 0.8984

To output top association results:

awk '{if($9 <= 1e-4) print $0}' data.assoc.logistic > data.assoc.logistic.top.txt

head data.assoc.logistic.top.txt

2 rs17188434 157096776 C ADD 1164 3.251 4.002 6.29e-05

7 rs2108225 107453103 G ADD 1164 0.6469 -4.51 6.475e-06

#############################################################################

**8. Data visualization and interpretation**

#############################################################################

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

1. Start R software (type R at the prompt).
2. To create a q-q plot based on the results of chi-square tests (performed in 7.2 above), copy and paste the following commands at the prompt:

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));

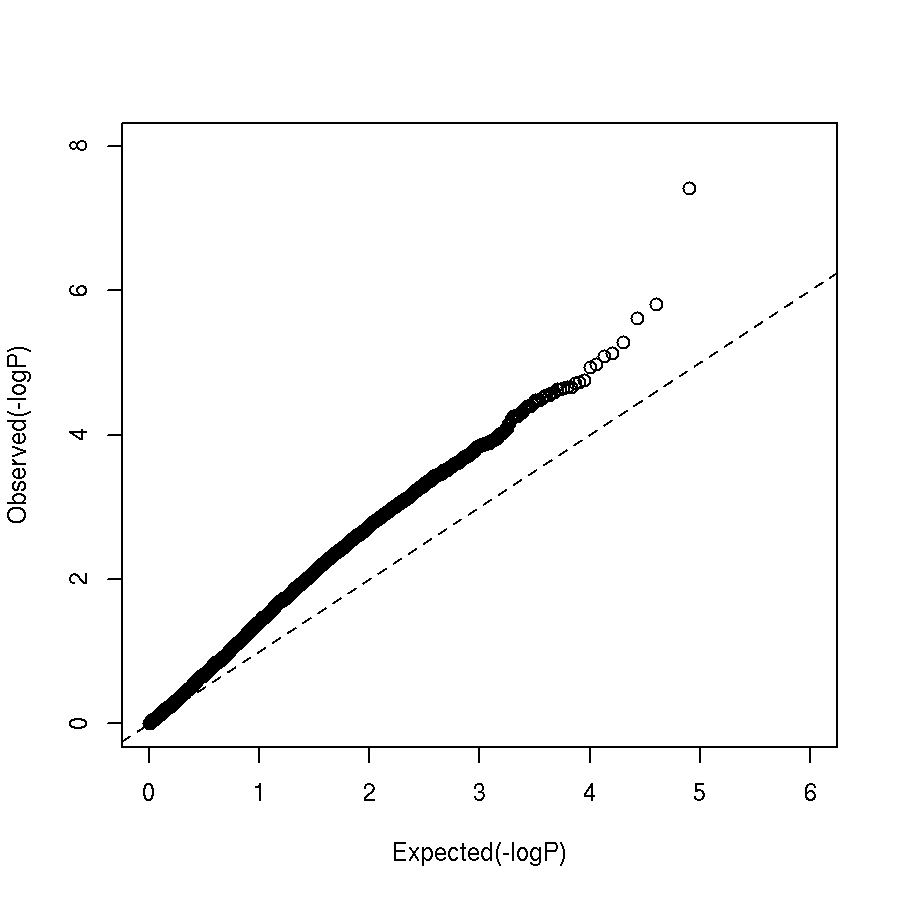
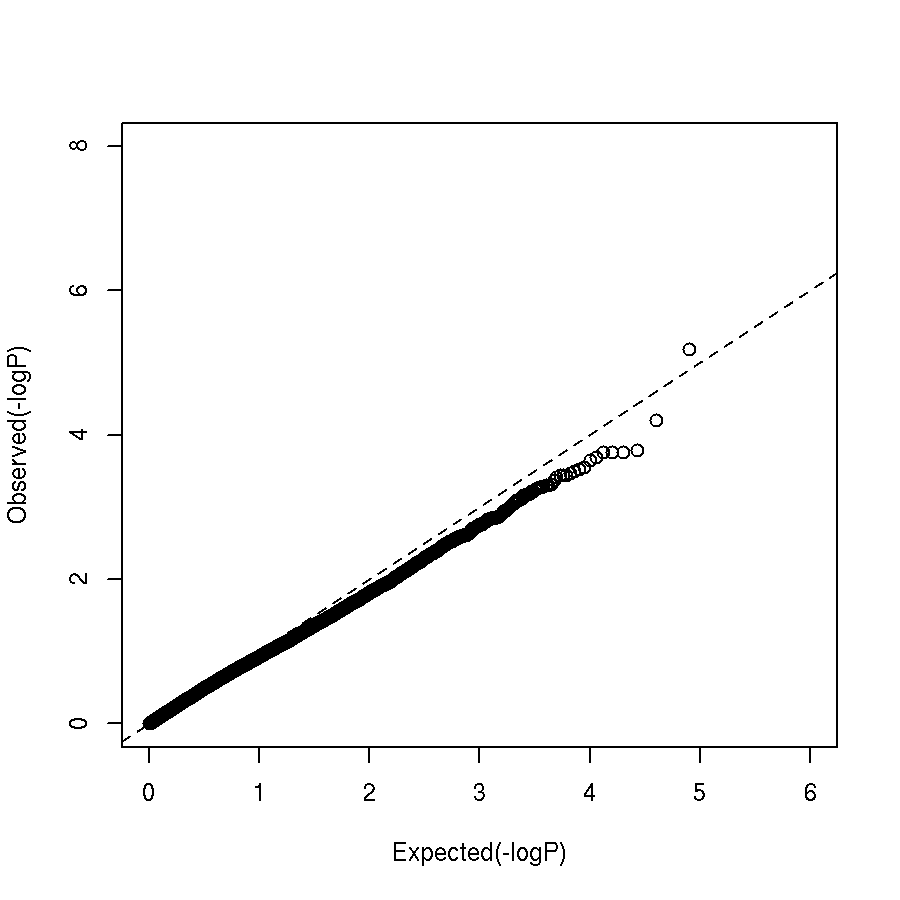
abline(a=0, b=1, col=1, lwd=1.5, lty=2);

dev.off()

q()

Open the file "pvalue.logistic.qq.plot.pdf". 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.**

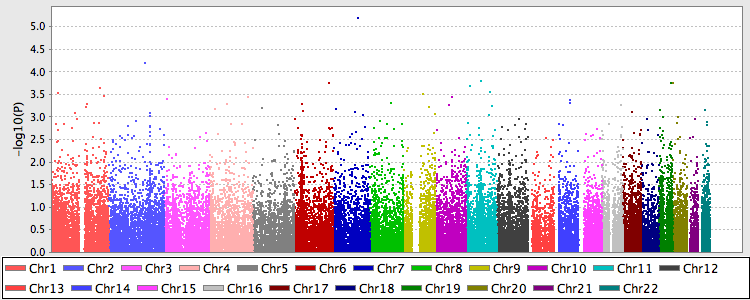
1. To obtain the inflation factor, include the --adjust option in any of the PLINK commands described in Step 4. For example, the inflation factor based on logistic regression assuming a multiplicative model is obtained by typing

plink --bfile ~/session4/data/1kg\_data.pass --logistic --hide-covar --covar ~/session4/data/1kg\_data.covar --adjust --out data

1. Open the PLINK log file ‘data.log’, which records the inflation factor. The inflation factor for our GWA study is 1.0077, indicating that no population stratification is detected in our GWA data (values <1.1 are considered “acceptable”)
2. 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.**

1. Start Haploview (java -jar ~/session4/bin/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 7. 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.pass.bim’. Leave other options as they are (ignore pairwise comparison of markers > 500 kb apart and exclude individuals with > 50% missing genotypes). Click ‘OK’.
2. 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.
3. 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.

(iv) To create a Manhattan plot in R, start R software and copy the following commands:

source('~/session4/bin/ Manhattan.plot.R')

data <-read.table("data.assoc.logistic", header=TRUE);

cl <-c("red", "navyblue", "darkgreen", "gold", "deepskyblue4", "magenta4", "slategray")

png('Manhattan\_plot.png', width = 8.5, height = 3.5, units = "in", res=300)

par(mar=c(4.1,4.1,1.6,1.1), cex.lab=1.4, cex.axis=1.3, mgp=c(2.75, .95, 0), las=1, font=2)

m.plot(data$P, data$CHR, data$BP, cex=0.75, pch=16, cex.axis=1.3, cex.lab=1.5, col=cl, mgp=c(2.75, .95, 0), pt.cex=0.9, main="", ylab=expression(paste(-log[10], ' P-value')));

abline(h=-log10(0.05/sum(!is.na(data$P))), lty=2, col='gray37', lwd=0.75)

dev.off()

#############################################################################

**9. Quantitative traits**

#############################################################################

1. **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 ~/session4/data/1kg\_data.pass --pheno ~/session4/data/1kg\_data.QT.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 rs146327803 865625 1164 -2.423 7.75 8.412e-05 -0.3127 0.7546

1 rs41285790 865628 1164 -7.605 9.159 0.000593 -0.8303 0.4065

1 rs9988179 865694 1164 -2.285 3.664 0.0003345 -0.6236 0.533

1. 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 ~/session4/data/1kg\_data.pass --linear --pheno ~/session4/data/1kg\_data.QT.pheno --pheno-name Pheno --hide-covar --covar ~/session4/data/1kg\_data.covar --out data

View 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 rs146327803 865625 A ADD 1164 -1.361 -0.1759 0.8604

1 rs41285790 865628 A ADD 1164 -9.112 -0.995 0.3199

1 rs9988179 865694 T ADD 1164 -0.5524 -0.1503 0.8805

1. 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?

1. 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('~/session4/data/1kg\_data.QT.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()

q()

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

plink --bfile ~/session4/data/1kg\_data.pass --linear --pheno ~/session4/data/1kg\_data.QT.pheno --pheno-name lPheno --hide-covar --covar ~/session4/data/1kg\_data.covar --out data2

R

data <- read.table("data2.assoc.linear", header=TRUE);

obs <- -log10(sort(data[data$TEST == "ADD", ]$P));

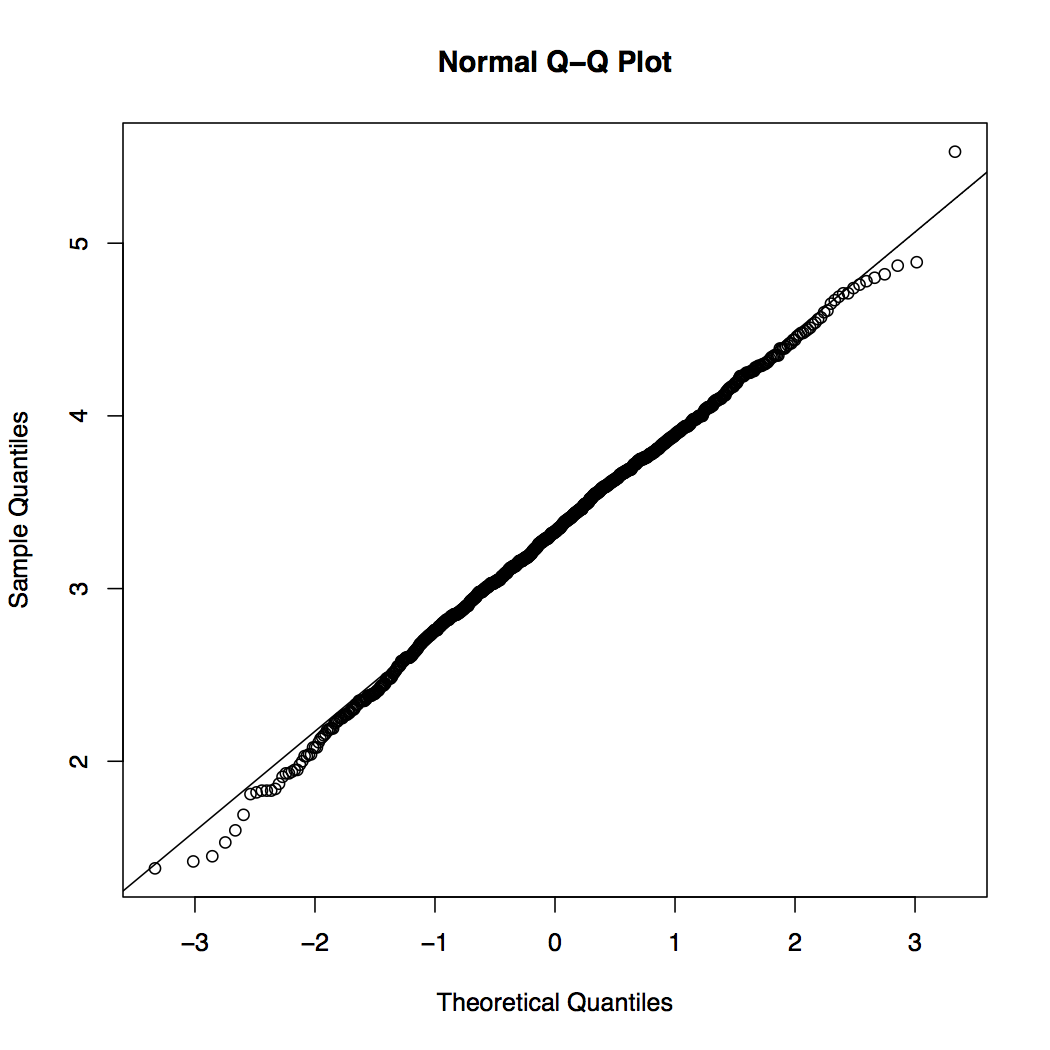
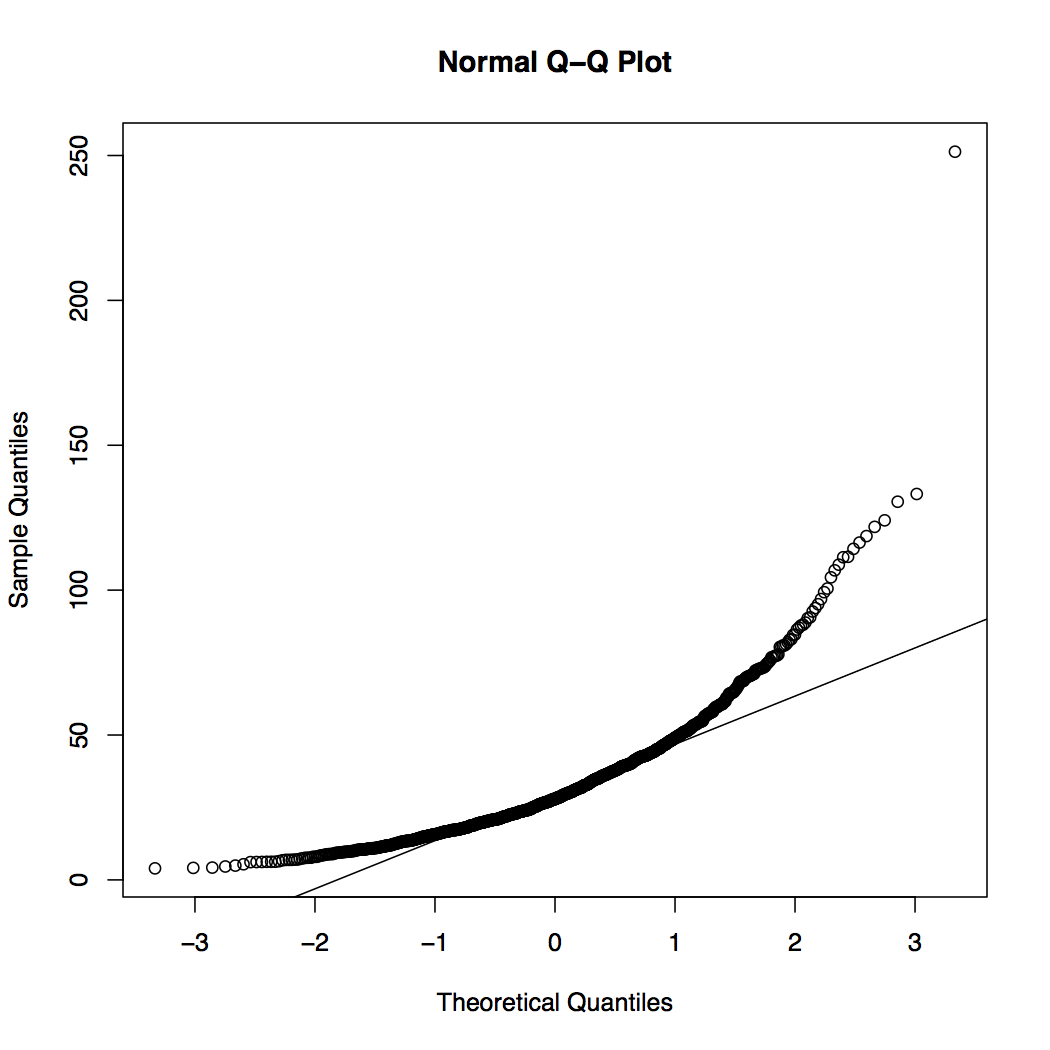
exp <- -log10(c(1:length(obs))/(length(obs) + 1));

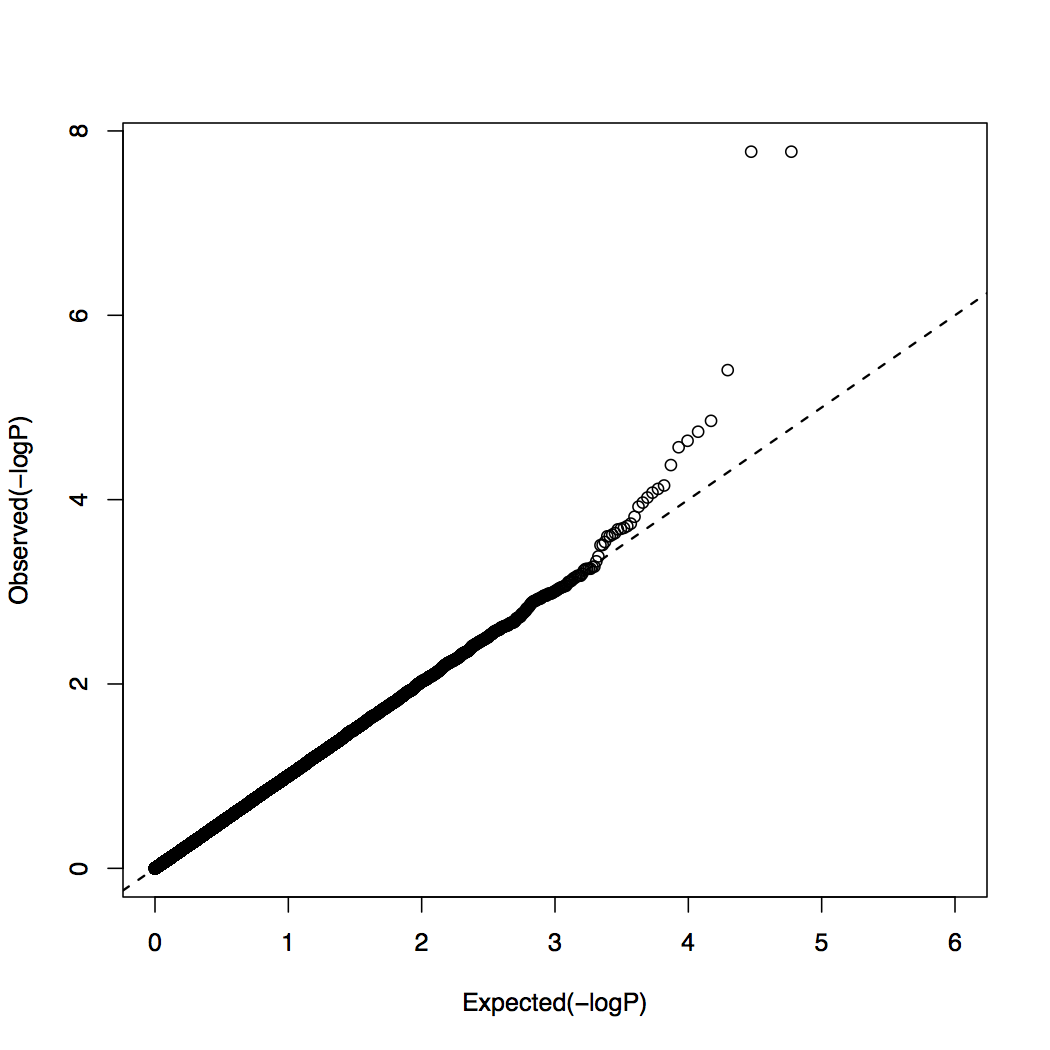
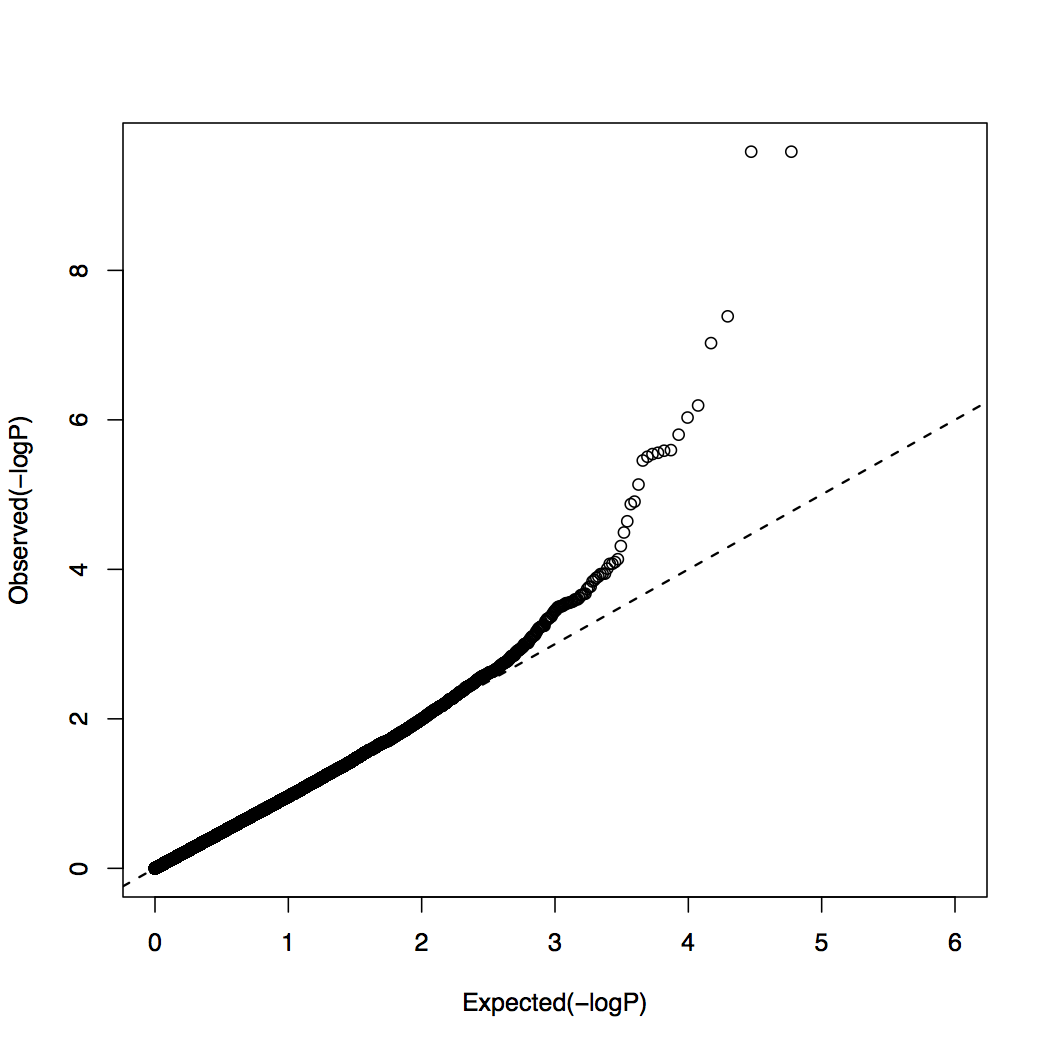
pdf("pvalue.linear.logpheno.qq.plot.pdf");

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).

1. 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

64190 16 rs1421085 53800954 C ADD 1164 0.1754 5.683 1.678e-08

64191 16 rs1558902 53803574 A ADD 1164 0.1754 5.683 1.678e-08

64198 16 rs9941349 53825488 T ADD 1164 0.1295 4.637 3.931e-06

71675 19 rs141060900 7691062 G ADD 1164 -1.0920 -4.601 4.678e-06

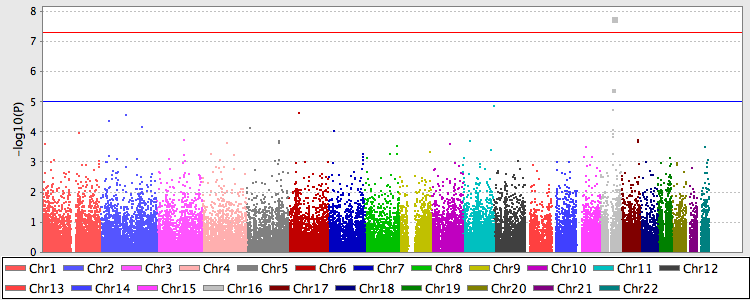
51756 11 rs6590705 133334522 A ADD 1164 -0.1640 -4.363 1.396e-05

64193 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.



#############################################################################

**10. Gene-based tests**

#############################################################################

To perform gene based tests for rare variants, we will use [EPACTS](https://genome.sph.umich.edu/wiki/EPACTS). We will use a built-in example in EPACTS package, using genotype data from chr 20 on a subset of 1000 genomes project participants.

cd ~/session4

run\_epacts –shell

./myrun\_epacts.sh

exit

The script above will perform single-variant association analysis for a binary phenotype DISEASE and then a burden test for variants with MAF<0.05. The annotated results can be viewed in the directory epacts\_out

**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.