Practical Session 2

Genome-wide Association Studies Under Sample Structure

Jan 6th, 2014

STOM 2014 Workshop

Hyun Min Kang

University of Michigan, Ann Arbor

GOALS FOR THIS SESSION

- Assuming we already know how to..
  - Use PLINK to QC GWAS data
  - Use R to visualize GWAS data

- We want to learn..
  - How to detect inflation of test statistics
  - How to visualize population structure
  - How to run PCA-adjusted analysis
  - How to run mixed model association
  - How to obtain pseudo-heritability from marker-based kinship matrix.

Example Dataset

- Genotype data
  - Thinned Omni2.5 array from 1000 Genomes phase 1
  - 552 European individuals
  - Intra-continental population structure
  - Also contains related individuals (trios)

- Simulated Phenotypes (Quantitative)
  - Causal SNP have largest effect size
  - Every other SNPs have small effect sizes normally distributed, producing polygenic background
  - Random noises normally distributed

KNOWING WHERE THE FILES ARE

- To see the files for the session, type

  ```
  ls /data/stom2014/session2/
  ```

  – If you see any errors, please let me know now!

- For convenience, let's set some variables

  ```
  export S2=/data/stom2014/session2
  mkdir ~/out
  ```

RUNNING NAÏVE ASSOCIATION TESTS

- Using PLINK..

```
$S2/bin/plink --noweb --bfile $S2/data/
1000G.auto.omni.phased.EUR --pheno $S2/
data/1000G_EUR_20_1459060.phe --linear --
out ~/out/naive
```

- First, check your output file

```
less ~/out/naive.assoc.linear
```

- Because we know the causal variant, let's check the p-value at the causal variant

```
grep -w ADD ~/out/naive.assoc.linear |
grep 20:1459060
```

SANITY CHECK USING QQ PLOT

- Run the following R codes

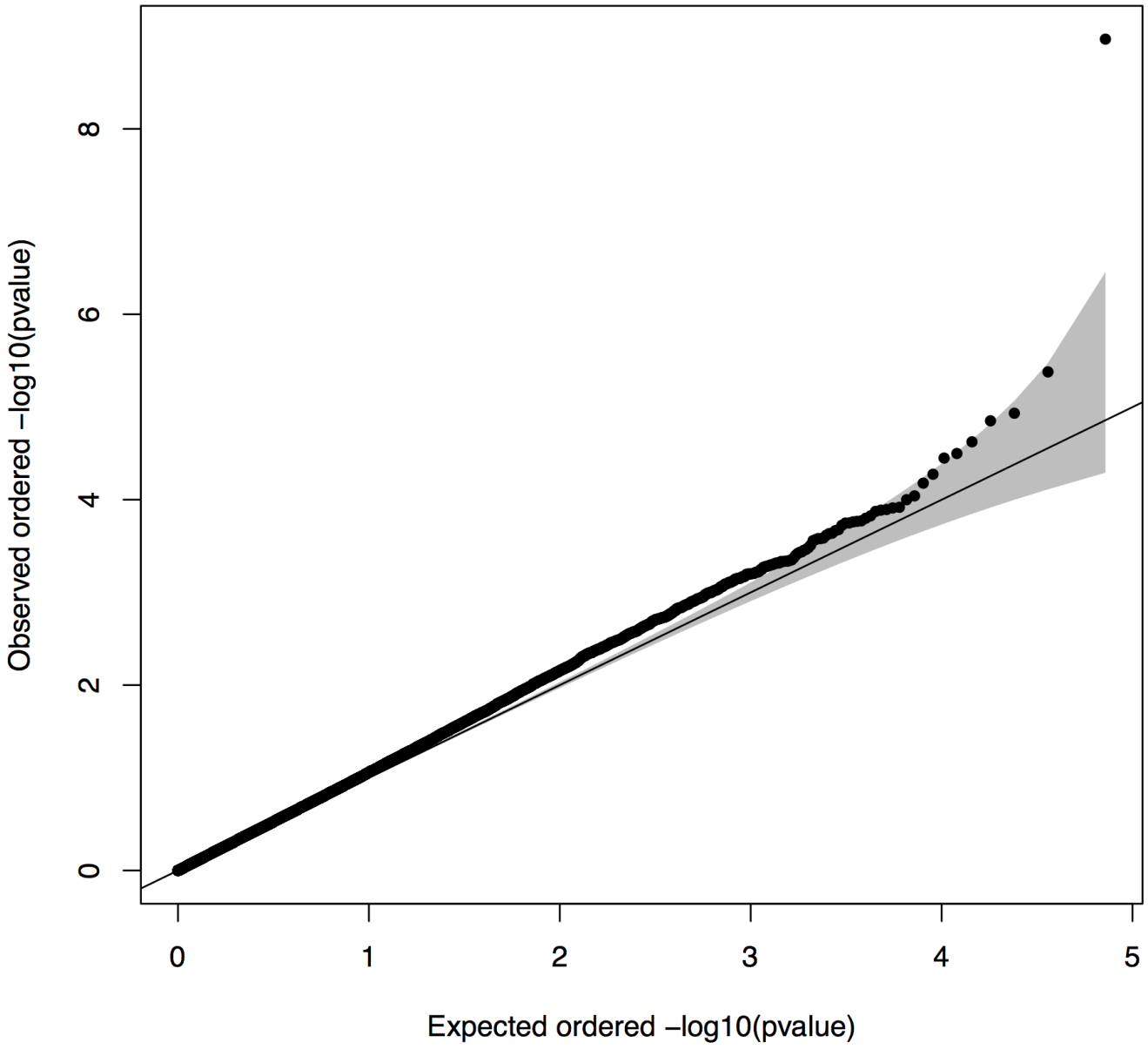
```
> source('/data/stom2014/session2/r/
qqconf.r')
> T <- read.table('~/out/
naive.assoc.linear',header=TRUE)
> pdf('~/out/naive.pdf')
> qq.conf.beta(T$P)
> dev.off()
```

- And copy the PDF file and see them

  Using WinSCP or

```
scp -P 8033 bibs.snu.ac.kr:out/
naive.pdf .
```

The QQ Plot Should Look Like..



Do you think this QQ plot is inflated or not?

How do we Know GC Inflation Factor?

$$\lambda_{GC} = \frac{\chi^2_{median}}{0.456}$$

- 0.456 is the median chi-square expected under the null
- Typical procedure to calculate GC inflation factor is
  - Get the median p-value first
  - Convert the p-value into chi-squared statistic
  - Calculate inflation factor using the equation above
- How can we achieve this?

SIMPLE WAY – USE --ADJUST OPTION

- Run the same command, with --adjust

```
$S2/bin/plink --noweb --bfile $S2/data/
1000G.auto.omni.phased.EUR --pheno $S2/
data/1000G_EUR_20_1459060.phe --linear --
adjust --out ~/out/naive
```

- Look at the messages in the screen, do you see what the GC inflation factor is?
  - Mine says 1.08875
- But… how would you know it is correct?
  - You should be able to compute yourself!

Calculating GC Inflation Factor On Your Own

- We need to know the median p-value using R

```
> T <- read.table('~/out/
naive.assoc.linear',header=TRUE)
```

- First, find the median p-value

```
> median(T$P)
> 0.4814
```

- Convert p-value into chi-square using R, and compute lambda

```
> qchisq(0.4814,1,lower.tail=FALSE)
[1] 0,4956901
> 0.4958032/0.456
[1] 1.08704
```

USING CUSTOM SCRIPT..

- See the example R script

  less $S2/r/calc.GC.lambda.r

- Feed the p-values from association results

  cut -c 96- ~/out/naive.assoc.linear | Rscript $S2/r/calc.GC.lambda.r

- Do these methods produce exactly the same inflation factor? If not, why?

PCA Analysis to Account for Population Structure

- Three ways to perform PCA analysis
  - Use Multidimensional scaling in PLINK based on IBS
    - This is a fine option, but extremely slow with large sample size
  - Use EIGENSOFT (from Alkes Price lab)
    - Provides a set of tools to adjust for PCA
  - Use EMMAX with custom scripting
    - Fast when the sample size is large
    - Let's try this option today

PCA ANALYSIS USING EMMAX

- Convert the PLINK file into EMMAX-favored format
  - (All the QCs need to be done within PLINK first)
  - Convert the binary PLINK format into transposed format

```
$S2/bin/plink --noweb --bfile $S2/data/
1000G.auto.omni.phased.EUR --recode12 --
output-missing-genotype 0 --transpose --
out ~/out/1000G.auto.omni.phased.EUR
```

- Create kinship matrix using EMMAX

```
$S2/bin/emmax-kin-intel64 -T 1 -M 0.2 -v
-d 10 ~/out/1000G.auto.omni.phased.EUR
less ~/out/
1000G.auto.omni.phased.EUR.aBN.kinf
```

Obtaining Principal Components

- Principal components can be obtained by either
  - SVD of the normalized genotype matrix, or
  - Eigendecomposition of the kinship matrix
- We can use the second approach using custom script

```
Rscript $S2/r/calc.PC.from.kinf.r ~/out/
1000G.auto.omni.phased.EUR.aBN.kinf ~/
out/1000G.auto.omni.phased.EUR.tfam ~/
out/1000G.auto.omni.phased.EUR.pc10
```

- The custom script produce a file with 10 PCs

```
less ~/out/1000G.auto.omni.phased.EUR.pc10
```

VISUALIZING SAMPLE STRUCTURE VIA PCA

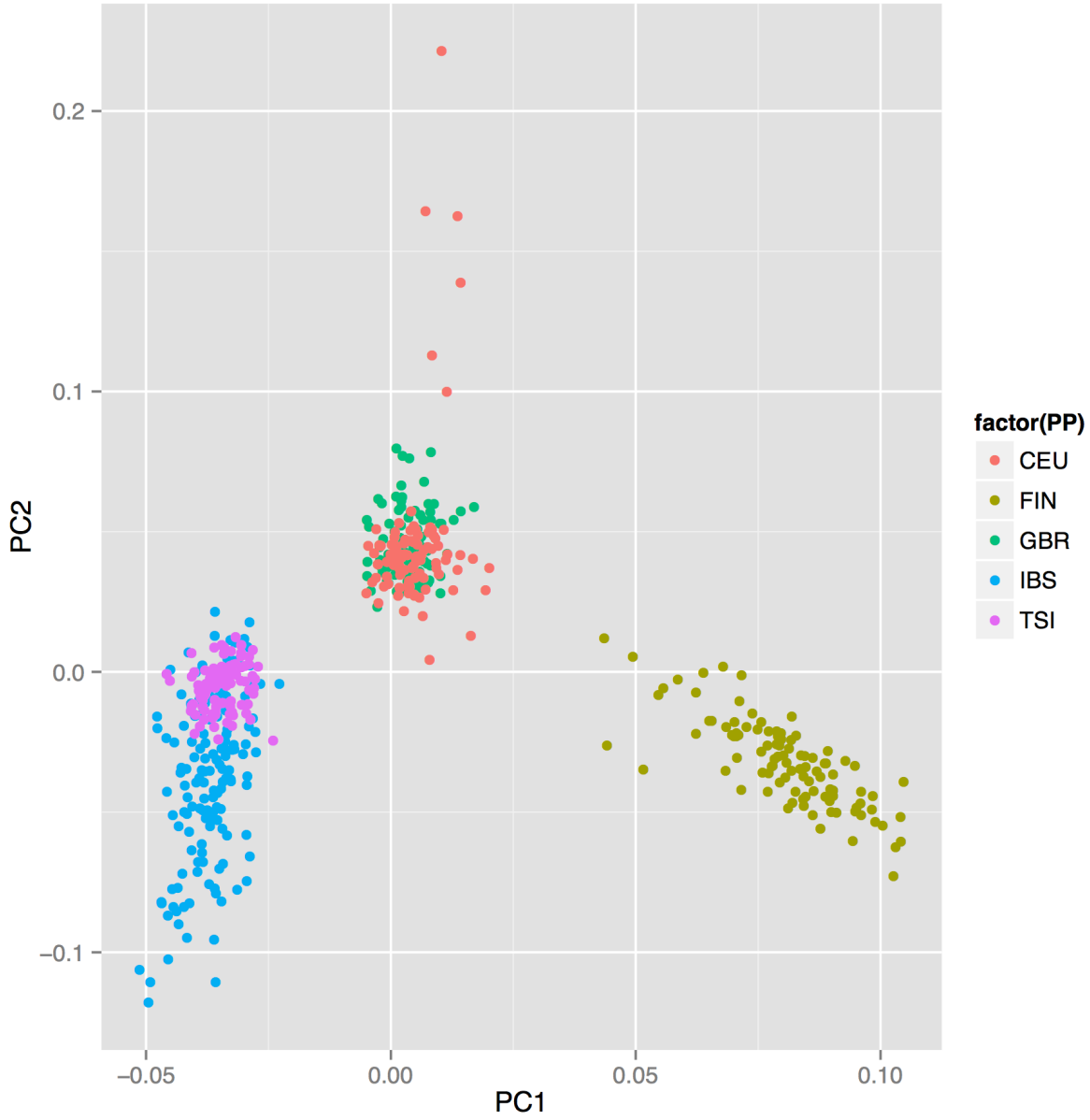
- EIGENSOFT is a widely used way to visualize sample structure

- We will use EMMAX to essentially same thing, with custom drawing script

  Rscript $S2/r/plot_pc_pop.r ~/out/ 1000G.auto.omni.phased.EUR.pc10 $S2/data/ 1000G.auto.omni.phased.EUR.pop ~/out/ 1000G.auto.omni.phased.EUR.pc10.pdf

- Copy the PDF file to your computer to see it

Output Should Look Like..



- Continental groups are well separated
- Consistent to known population history

ADJUSTING FOR POPULATION STRUCTURE

- Use PLINK with covariate option

```
$S2/bin/plink --noweb --bfile $S2/data/
1000G.auto.omni.phased.EUR --pheno $S2/
data/1000G_EUR_20_1459060.phe --covar ~/
out/1000G.auto.omni.phased.EUR.pc10 --
covar-number 1,2,3,4 --linear --adjust --
out ~/out/pca
```

- How does it affect the causal variant and inflation?

```
grep -w ADD ~/out/pca.assoc.linear | grep
20:1459060
```

```
grep -w ADD ~/out/pca.assoc.linear | cut
-c 96- | Rscript $S2/r/calc.GC.lambda.r
```

USING MIXED MODEL ASSOCIATION VIA EMMAX

- Run EMMAX

  $S2/bin/emmax-intel64 -t ~/out/1000G.auto.omni.phased.EUR -o ~/out/emmax -p $S2/data/1000G_EUR_20_1459060.phe -k ~/out/1000G.auto.omni.phased.EUR.aBN.kinf

  – Note that PCs does not have to be included as covariates

- Checking the inflation factor..

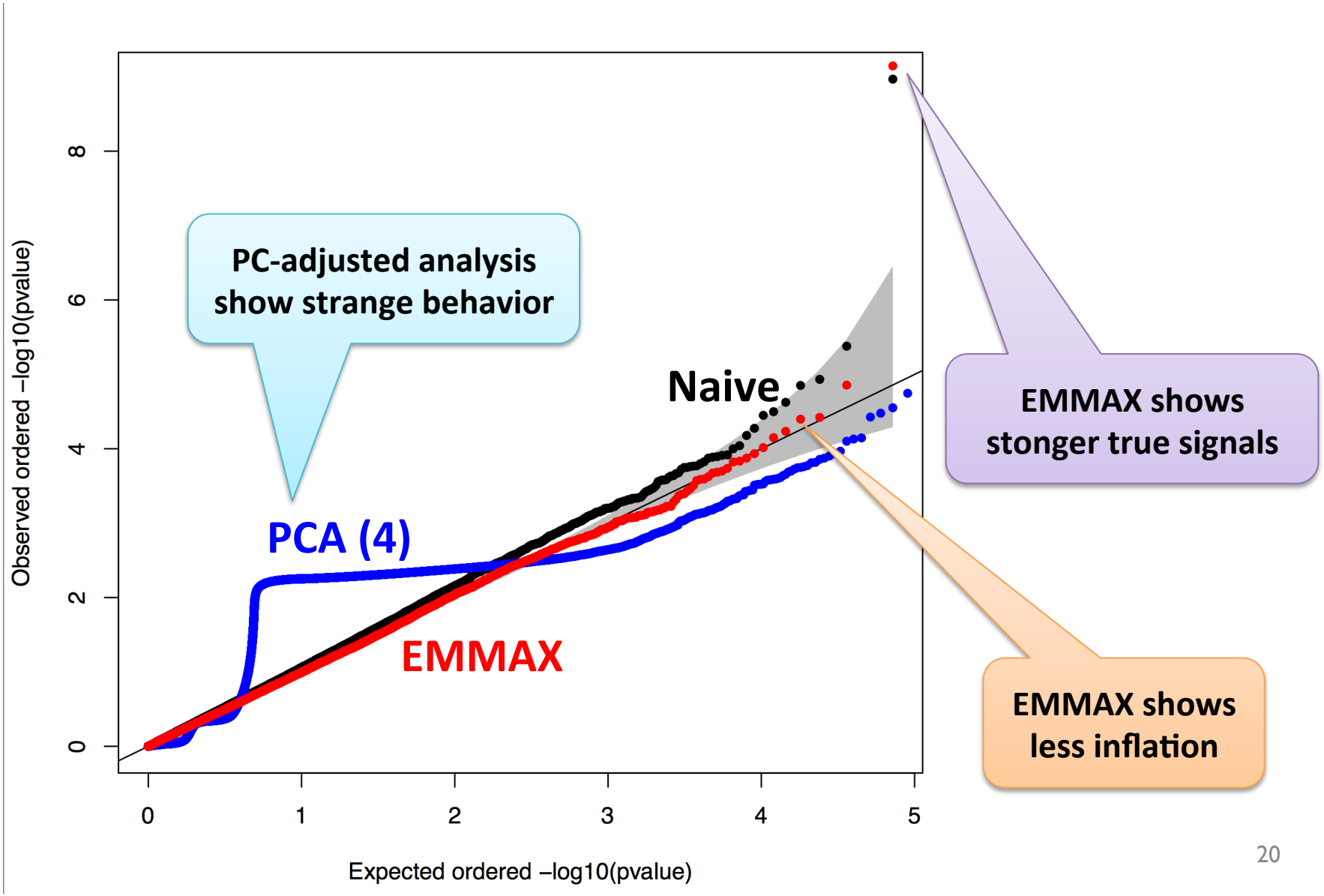
  cut -f 4 ~out/emmax.ps | Rscript $S2/r/calc.GC.lambda.r

  [1] 1.006079

DRAWING QQ PLOTS

```
> source('/data/stom2014/session2/r/qqconf.r')
> T1 <- read.table('~/out/
naive.assoc.linear',header=TRUE)
> T2 <- read.table('~/out/
pca.assoc.linear',header=TRUE)
➢ T3 <- read.table('~/out/emmax.ps')
➢ > pdf('~/out/all.pdf')
> qq.conf.beta(T1$P)
>
qq.conf.beta(T2$P,drawaxis=FALSE,ptcolor="blue")
>
qq.conf.beta(T3$V4,drawaxis=FALSE,ptcolor="red")
> dev.off()
```

Comparing QQ Plots



REVIEW

- Can you draw QQ plot from GWAS results?
- How can you check the genomic control (GC) inflation factor (λ_{GC}) after performing association test?
- Can you visualize the PCA plot given high density marker data?
- Which observations suggests that EMMAX performs better than other two methods in the example?