Introduction to GWAS using R and GenABEL
LUPA Workshop in Statistical Methods for GWAS studies

Marcin Kierczak*

*Computational Genetics Group
Faculty of Veterinary and Animal Breeding
Swedish University of Agricultural Sciences
Uppsala, SWEDEN

04-05 May 2011, Uppsala
Introduction to R

What is R?

R is a:
- programming language
- software environment

for:
- statistical computing
- beautiful graphics
Introduction to R

Why?

- *de facto* standard among statisticians
- widely used for development and data analysis
- an implementation of the S programming language
- partially inspired by Scheme
- created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand
- source code is freely available under the GNU GPL
- mainly command line
- GUIs exist
- many tools/tests at hand
- project homepage: http://www.r-project.org
R code example

> 2 + 2
[1] 4
> p.value <- 0.05
> p.value
[1] 0.05
> -log10(p.value)
[1] 1.30103
> print("Hello!")
[1] "Hello!"
Power of R

R is modular – there is a core and you can load packages containing custom functions.

- 2984 packages available on CRAN (02.05.2011) – http://cran.r-project.org/
- 998 projects registered on R-Forge (02.05.2011) – http://r-forge.r-project.org/
- 460 packages available on BioConductor (03.05.2011) – http://www.bioconductor.org/
- from genetics to social sciences and from geology to cryptography
Installing packages

```r
> install.packages("GenABEL")
> install.packages("DatABEL",
+ repos="http://R-Forge.R-project.org")
```

Loading packages

```r
> require("GenABEL")  # Within functions
> library("GenABEL")
```
How to get help

> vignette("GenABEL") # Package level
> demo(graphics)
> help(qtscore) # Function level
> ?qtscore
> ??qtscore # Extensive search
Function for generating random genotypes

```r
> generateGenotypes <- function(num.markers = 1, missing = F) {
+   if (missing) {
+     genotypes <- sample(1:5, num.markers, replace = T)
+   } else {
+     sample(1:4, num.markers, replace = T) -> genotypes
+   }
+   genotypes[genotypes == 1] <- "A"
+   genotypes[genotypes == 2] <- "T"
+   genotypes[genotypes == 3] <- "C"
+   genotypes[genotypes == 4] <- "G"
+   genotypes[genotypes == 5] <- "X"
+   genotypes
+ }

> generateGenotypes(5)
[1] "T" "T" "G" "G" "G"
> generateGenotypes(num.markers = 5)
[1] "G" "T" "A" "G" "G"
> generateGenotypes(10, T)
[1] "C" "C" "A" "T" "A" "A" "G" "A" "G" "A"
> generateGenotypes(missing = T, num.markers = 10)
[1] "G" "A" "X" "A" "A" "C" "X" "A" "A" "C"
```
The mission of the GenABEL project is to provide a framework for collaborative, sustainable, transparent, open-source based development of statistical genomics methodology. We aim to streamline methodology discussion, development, implementation, dissemination and maintenance; through the community.

GenABEL is developed by a Team led Dr. Yurii Aulchenko, Erasmus MC, Rotterdam.
• **GenABEL** – genome-wide association analysis for quantitative, binary and time-till-event traits.

• **MetABEL** – meta-analysis of genome-wide SNP association results GWAS for quantitative, binary and time-till-event trait.

• **ProbABEL** – genome-wide association analysis of imputed data.

• **PredictABEL** – assess the performance of risk models for binary outcomes.

• **DatABEL** – file-based access to large matrices stored on HDD in binary format.

• **ParallABEL** – generalized parallelization of GWAS.

• **MixABEL** – more mixed models GWAS; experimenting with GSL, multiple input formats, iterator, parallelization through threads.
Data representation in GenABEL

- **.raw** – genotype data GenABEL internal binary format.
- **.dat** – phenotype data, e.g., as in PLINK.

Binary format = compression, e.g. for 170K SNP chip 200 individuals: data.ped – 144.4MB vs. data.raw – 32.9MB.

**.dat file format**

```
  id  sex  age  bt1  ct  ct1
"289982"  0  30.33  NA  NA  3.93
"325286"  0  36.514  1  0.49  3.61
"357273"  1  37.811  0  1.65  5.30
```
GenABEL
Importing and loading data

Importing from different data formats

- `convert.snp.text` – convert from text format.
- `convert.snp.ped` – convert from PED format.
- `convert.snp.tped` – convert from TPED format.
- `convert.snp.illumina` – convert from Illumina format.

Loading data

```r
> data <- load.gwaa.data("dataset/phenotype.dat",
+     "dataset/genotype.raw",
+     makemap = T)
```

If coords are chromosome specific, you can make them genome-wise by: `makemap = T`. 
Examine the phenotype data

> nids(data)
[1] 207
> nsnps(data)
[1] 174375
> phdata(data)[2,]
   id sex  bt  ct group response
  dog225 dog225 1 0 1.925575 3 1.569402
> phdata(data)[1:5, "sex"]
[1] 1 1 0 1 0
Get summary for trait response

```r
> summary(phdata(data)[, "response"])

   Min. 1st Qu.  Median    Mean  3rd Qu.   Max.   NA's
-0.4325  0.9996  1.4090  1.4860  1.9480  3.3860   1.0000

> hist(phdata(data)[, "response"],
+ breaks = 100,
+ col = "red")
```

Histogram of phdata(data)[, "response"]
Examine the genotype data 1st individual, markers 3-5

```r
> gtdata(data)[1,3:5]
@nids = 1
@nsnps = 3
@nbytes = 1
@idnames = dog224
@snpnames = BICF2P1383091 TIGRP2P259 BICF2P186608
@chromosome = 1 1 1
@coding = 04 01 01
@strand = 00 00 00
@map = 3212349 3249189 3265742
@male = 1
@gtps =
80 40 40
```
### Get summary for markers 2 and 3

```r
> summary(gtdata(data))[2:3,]

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Strand</th>
<th>A1</th>
<th>A2</th>
<th>NoMeasured</th>
<th>CallRate</th>
<th>Q.2</th>
<th>P.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>BICF2630707846</td>
<td>1 3082514</td>
<td>u</td>
<td>1</td>
<td>2</td>
<td>206</td>
<td>0.995169</td>
<td>0.0</td>
<td>206</td>
</tr>
<tr>
<td>BICF2P1383091</td>
<td>1 3212349</td>
<td>u A G</td>
<td>207</td>
<td>1.000000</td>
<td>0.5</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P.12</th>
<th>P.22</th>
<th>Pexact</th>
<th>Fmax</th>
<th>Plrt</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1.000000e+00</td>
<td>0</td>
<td>1.000000e+00</td>
</tr>
<tr>
<td>207</td>
<td>0</td>
<td>1.240547e-61</td>
<td>-1</td>
<td>2.282010e-64</td>
</tr>
</tbody>
</table>
```
Is the binary trait bt correlated with sex?

```r
> tab <- table(phdata(data)$bt, phdata(data)$sex)
> fisher.test(tab)

Fisher's Exact Test for Count Data

data:  tab
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5322721 1.9110534
sample estimates:
  odds ratio
1.010358
```
Is the binary trait bt related to response?

```r
> boxplot(phdata(data)$response ~ phdata(data)$bt,
+ names = c("Ctrls", "Cases"),
+ ylab = "Response")
```
Do a simple QC?

```r
> qc1 <- check.marker(data, call = 0.95,
+ perid.call = 0.95,
+ maf = 1e-08,
+ p.lev = 1e-08)
> ...
> data.clean <- data[qc1$idok, qc1$snpok]
```
Do a simple association test

```r
> an <- qtscore(bt ~ response, data,
+ trait.type="binomial", times=1)
> summary(an, top=5)

Summary for top 5 results, sorted by P1df

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Strand</th>
<th>A1</th>
<th>A2</th>
<th>N</th>
<th>effB</th>
<th>se_effB</th>
<th>chi2.1df</th>
<th>P1df</th>
<th>effAB</th>
<th>effBB</th>
<th>chi2.2df</th>
<th>P2df</th>
</tr>
</thead>
<tbody>
<tr>
<td>BICF2P506952</td>
<td>1 90475257</td>
<td>u</td>
<td>A</td>
<td>G</td>
<td>204</td>
<td>-0.002009766</td>
<td>0.0003574521</td>
<td>31.61223</td>
<td>1.882407e-08</td>
<td>0</td>
<td>NA</td>
<td>31.61223</td>
<td>1.882407e-08</td>
</tr>
<tr>
<td>BICF2G630348662</td>
<td>3 339590471</td>
<td>u</td>
<td>T</td>
<td>C</td>
<td>204</td>
<td>-0.002009766</td>
<td>0.0003574521</td>
<td>31.61223</td>
<td>1.882407e-08</td>
<td>0</td>
<td>NA</td>
<td>31.61223</td>
<td>1.882407e-08</td>
</tr>
<tr>
<td>TIGRP2P51678</td>
<td>3 339878416</td>
<td>u</td>
<td>C</td>
<td>T</td>
<td>204</td>
<td>-0.002009766</td>
<td>0.0003574521</td>
<td>31.61223</td>
<td>1.882407e-08</td>
<td>0</td>
<td>NA</td>
<td>31.61223</td>
<td>1.882407e-08</td>
</tr>
<tr>
<td>BICF2G630348969</td>
<td>3 340378977</td>
<td>u</td>
<td>T</td>
<td>G</td>
<td>204</td>
<td>-0.002009766</td>
<td>0.0003574521</td>
<td>31.61223</td>
<td>1.882407e-08</td>
<td>0</td>
<td>NA</td>
<td>31.61223</td>
<td>1.882407e-08</td>
</tr>
<tr>
<td>BICF2P628966</td>
<td>3 340641323</td>
<td>u</td>
<td>C</td>
<td>T</td>
<td>204</td>
<td>-0.002009766</td>
<td>0.0003574521</td>
<td>31.61223</td>
<td>1.882407e-08</td>
<td>0</td>
<td>NA</td>
<td>31.61223</td>
<td>1.882407e-08</td>
</tr>
</tbody>
</table>

Pc1df

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>P1df</th>
</tr>
</thead>
<tbody>
<tr>
<td>BICF2P506952</td>
<td>1.767672e-07</td>
</tr>
<tr>
<td>BICF2G630348662</td>
<td>1.767672e-07</td>
</tr>
<tr>
<td>TIGRP2P51678</td>
<td>1.767672e-07</td>
</tr>
<tr>
<td>BICF2G630348969</td>
<td>1.767672e-07</td>
</tr>
<tr>
<td>BICF2P628966</td>
<td>1.767672e-07</td>
</tr>
</tbody>
</table>
```
What is $\lambda$, show Q-Q plot...

```r
> estlambda(an[, "P1df"])
$estimate
[1] 1.159153

$se
[1] 0.0003918843
```
What about a Manhattan plot?

```r
> plot(an,
+ col = c("red", "slateblue"),
+ pch = 19,
+ cex = .5,
+ df = "1")
> bonferroni <- -log10(0.05 / nsnps(data))
> abline(h=bonferroni, col = "red")
```
GenABEL – why?
Easiness of comparing different approaches...

Load your data one time and enjoy:

- Simple association tests.
- Genomic control.
- PCA-based correction – Eigenstrat.
- Mixed models.
- Structured association.
- Any combination of the above!
Thank You! and:

- Leif Andersson
- Yurii Aulchenko
- Örjan Carlborg
- Dirk Jan de Koonig
- Kerstin Lindblad-Toh
- Xia Shen
- Katarina Tengvall
Use account:

- login: Kurs_LUPAonStatistic
- password: LupaStat2011
- DO NOT TRY TO LOGIN TO VMWARE (Windows) - you will block the whole account!!!

Website:

http://www.computationalgenetics.se/LUPA2011