Practicals Bioinformatics 2011-2012

Olivier Stern olivier.stern@ulg.ac.be

Tom Cattaert tom.cattaert@ulg.ac.be

13 December 2011: GenABEL and FAM-MDR

The class gwaa.data

- Install package GenABEL

```
> install.packages('GenABEL')
> library(GenABEL)
```

- Load example data of class gwaa.data

```
> data(srdta)
```

- This class provides an efficient way of storing GWA data in R
- You can get specific data from the internal representation using specific functions

The class gwaa.data

- Get phenotype data as a data frame with first column subject ID, second column subject gender, and further columns phenotypes

```
> phdata(srdta)[1:5, ]
  id sex  age   qt1    qt2  qt3 bt
1 p1   1 43.4 -0.58   4.46 1.43  0
2 p2   1 48.2  0.80   6.32 3.90  1
3 p3   0 37.9 -0.52   3.26 5.05  1
4 p4   1 53.8 -1.55 888.00 3.76  1
5 p5   1 47.5  0.25   5.70 2.89  1
```

- Specific rows and columns can be accessed in the usual way using subsetting and the $ operator

The class gwaa.data

- Get genotype data

```
> gtdata(srdta[1:5, 1:5])
@nids = 5
@nsnps = 5
@nbytes = 2
@idnames = p1 p2 p3 p4 p5
@snpnames = rs10 rs18 rs29 rs65 rs73
@chromosome = 1 1 1 1 1
@coding =  08 0b 0c 03 04
@strand =  01 01 02 01 01
@map = 2500 3500 5750 13500 14250
@male = 1 1 0 1 1
@gtps =
40 40 40 80 40
40 40 00 00 40
```

The class gwaa.data

- Get number of individuals

```
> nids(srdta)
[1] 2500
```

- Get number of SNPs

```
> nsnps(srdta)
[1] 833
```

- Get subject ID

```
> idnames(srdta)[1:5]
[1] "p1" "p2" "p3" "p4" "p5"
```

The class gwaa.data

- Get subject gender

```
> male(srdta)[1:5]
p1 p2 p3 p4 p5
 1  1  0  1  1
```

- Evaluate numbers of females and males

```
> table(male(srdta))
   0    1
1225 1275
```

- Get SNP name

```
> snpnames(srdta)[1:5]
[1] "rs10" "rs18" "rs29" "rs65" "rs73"
```

The class gwaa.data

- Get SNP chromosome

```
> chromosome(srdta)[1:5]
[1] "1" "1" "1" "1" "1"
```

- Get SNP map position

```
> map(srdta)[1:5]
 rs10  rs18  rs29  rs65  rs73
 2500  3500  5750 13500 14250
```

- Obtain SNPs between 1,100,000 and 1,105,000 b.p

```
> snpnames(srdta)[map(srdta)>1100000 & map(srdta)<1105000]
[1] "rs4180" "rs4186" "rs4187"
```

The class gwaa.data

- Get SNP coding (where effect of second versus first is considered)

```
> coding(srdta)[1:5]
rs10 rs18 rs29 rs65 rs73
"TG" "GA" "GT" "AT" "AG"
```

- Get SNP strand information

```
> strand(srdta)[1:5]
rs10 rs18 rs29 rs65 rs73
 "+"  "+"  "-"  "+"  "+"
```

Adding and deleting phenotypes

- When modifying the phenotype data frame you should not modify it directly, but use special functions for this
- One can add phenotypes using add.phdata()

```
> age2 <- phdata(srdta)$age^2
> srdta <- add.phdata(srdta, newph = age2, name = "age_squared")
> phdata(srdta)[1:5, ]
   id sex  age   qt1    qt2  qt3 bt age_squared
p1 p1   1 43.4 -0.58   4.46 1.43  0     1883.56
p2 p2   1 48.2  0.80   6.32 3.90  1     2323.24
p3 p3   0 37.9 -0.52   3.26 5.05  1     1436.41
p4 p4   1 53.8 -1.55 888.00 3.76  1     2894.44
p5 p5   1 47.5  0.25   5.70 2.89  1     2256.25
```

Adding and deleting phenotypes

- Multiple phenotypes can be added if supplying a data frame containing subject ID besides new phenotypes

```
> newvalues <- matrix(rnorm(3 * 5), 3, 5)
> newdata <- data.frame(id = c("p1", "p2", "p4"), ph1 = 1, ph2 = 1, ph3 = 1, ph4 = 1, ph5 = 1)
> newdata[, c(2:6)] <- newvalues
> srdta <- add.phdata(srdta, newdata)
> phdata(srdta)[1:5, ]
   id sex  age   qt1    qt2  qt3 bt age_squared        ph1        ph2        ph3       ph4        ph5
p1 p1   1 43.4 -0.58   4.46 1.43  0     1883.56 -0.27962804  1.0457324 -0.6573852 0.9601471 -0.3230608
p2 p2   1 48.2  0.80   6.32 3.90  1     2323.24  0.02432089  1.5948616 -0.3376420 0.7067874 -1.0845408
p3 p3   0 37.9 -0.52   3.26 5.05  1     1436.41          NA         NA         NA        NA         NA
p4 p4   1 53.8 -1.55 888.00 3.76  1     2894.44 -1.50233785 -0.8274808 -0.6995663 0.9530038 -0.6804120
p5 p5   1 47.5  0.25   5.70 2.89  1     2256.25          NA         NA         NA        NA         NA
```

Adding and deleting phenotypes

- We can delete the newly created phenotypes using del.phdata()

```
> srdta <- del.phdata(srdta, c("age_squared", "ph1", "ph2", "ph3", "ph4", "ph5"))
> phdata(srdta)[1:5, ]
   id sex  age   qt1    qt2  qt3 bt
p1 p1   1 43.4 -0.58   4.46 1.43  0
p2 p2   1 48.2  0.80   6.32 3.90  1
p3 p3   0 37.9 -0.52   3.26 5.05  1
p4 p4   1 53.8 -1.55 888.00 3.76  1
p5 p5   1 47.5  0.25   5.70 2.89  1
```

Sub-setting and coercing gwaa.data

- Objects of class gwaa.data can be subsetted

```
> ssubs <- srdta[1:5, 1:3]
```

- The result is a smaller object of class gwaa.data
- The object ssubs contains phenotype data for the first 5 individuals

```
> phdata(ssubs)
  id sex  age   qt1    qt2  qt3 bt
p1 p1   1 43.4 -0.58   4.46 1.43  0
p2 p2   1 48.2  0.80   6.32 3.90  1
p3 p3   0 37.9 -0.52   3.26 5.05  1
p4 p4   1 53.8 -1.55 888.00 3.76  1
p5 p5   1 47.5  0.25   5.70 2.89  1
```

Sub-setting and coercing gwaa.data

- … and genotype data for the first 5 individuals and first 3 SNPs

```
> gtdata(ssubs)

@nids = 5

@nsnps = 3

@nbytes = 2

@idnames = p1 p2 p3 p4 p5

@snpnames = rs10 rs18 rs29

@chromosome = 1 1 1

@coding =  08 0b 0c

@strand =  01 01 02

@map = 2500 3500 5750

@male = 1 1 0 1 1

@gtps =

40 40 40

40 40 00
```

Sub-setting and coercing gwaa.data

- To get a human-readable format, the genotype object should be coerced to a regular R data type
- The genotype object can be coerced to character

```
> as.character(gtdata(ssubs))
   rs10  rs18  rs29
p1 "T/T" "G/G" "G/G"
p2 "T/T" "G/G" NA
p3 "T/T" "G/G" NA
p4 "T/T" "G/G" NA
p5 "T/T" "G/A" "G/G"
```

Sub-setting and coercing gwaa.data

- The genotype object can also be coerced to numeric

```
> as.numeric(gtdata(ssubs))
   rs10 rs18 rs29
p1    0    0    0
p2    0    0   NA
p3    0    0   NA
p4    0    0   NA
p5    0    1    0
```

- The coding 0, 1, 2 corresponds to homozygotes for the first allele, heterozygotes and homozygotes for the second allele

Sub-setting and coercing gwaa.data

- This is done with respect to the coding given by

```
> coding(ssubs)
rs10 rs18 rs29
"TG" "GA" "GT"
```

- Hence e.g. for rs18 G/G is converted to 0, G/A is converted to 1 and A/A is converted to 2
- We can also select on ID and SNP names

```
> ssubs2 <- srdta[c("p141", "p147", "p2000"), c("rs10", "rs29")]
> gtdata(ssubs2)
@nids = 3
...
```

Exploring genetic data

- A summary can be obtained

```
> summary(ssubs)
    Chromosome Position Strand A1 A2 NoMeasured CallRate Q.2 P.11 P.12 P.22
Pexact      Fmax      Plrt
rs10        1    2500     +  T  G        5      1.0 0.0    5    0    0      1  0.0000000 1.0000000
rs18        1    3500     +  G  A        5      1.0 0.1    4    1    0      1 -0.1111111 0.7386227
rs29        1    5750     -  G  T        2      0.4 0.0    2    0    0      1  0.0000000 1.0000000
```

- Information is given on chromosome, position, strand, reference and coding alleles, callrate, coding allele frequency, genotype frequencies, exact HWE p-value

Exploring genetic data

- Restrict to individuals 65 and older

```
> vec <- (phdata(srdta)$age >= 65)
> table(vec)
vec
FALSE  TRUE
 2450    50
```

- A summary for the same 3 SNPs for individuals 65 and older

```
> summary(gtdata(srdta[vec, 1:3]))
```

	Chromosome	Position	Strand	A1	A2	NoMeasured	CallRate	Q.2	P.11	P.12	P.22	Pexact	Fmax	Plrt
rs10	1	2500	+	T	G	48	0.96	0.1354167	36	11	1	1.0000000	0.02131603	0.8843626
rs18	1	3500	+	G	A	47	0.94	0.2765957	25	18	4	0.7245853	0.04298643	0.7697067
rs29	1	5750	-	G	T	45	0.90	0.1555556	32	12	1	1.0000000	-0.01503759	0.9188943

Exploring genetic data

- Test for HWE in controls (bt=0) for 5 first SNPs

```
> summary(gtdata(srdta[phdata(srdta)$bt == 0, 1:5]))

    Chromosome Position Strand A1 A2 NoMeasured  CallRate       Q.2 P.11 P.12 P.22       Pexact
Fmax        Plrt

rs10         1     2500      + T  G       1177 0.9453815 0.12744265  897  260   20 7.933317e-01
0.006751055 8.178295e-01

rs18         1     3500      + G  A       1185 0.9518072 0.27426160  623  474   88 9.418133e-01 -
0.004812165 8.683219e-01

rs29         1     5750      - G  T       1188 0.9542169 0.13215488  897  268   23 5.288436e-01
0.016525913 5.737373e-01

rs65         1    13500      + A  T       1183 0.9502008 0.71344041   98  482  603 8.871139e-01
0.003540522 9.031273e-01

rs73         1    14250      + A  G       1188 0.9542169 0.01641414 1154   29    5 6.941219e-06
0.244001185 5.537568e-06
```

Exploring genetic data

- One can also obtain per individual summary data

```
> perid.summary(srdta[1:5, ])
```

	NoMeasured	NoPoly	Hom	E(Hom)	Var	F	CallPP	Het
p1	790	636	0.7987342	0.6389451	0.3960760	0.44256165	0.9483794	0.2012658
p2	792	640	0.7474747	0.6393834	0.4430435	0.29974032	0.9507803	0.2525253
p3	783	635	0.6206897	0.6371009	0.3834696	-0.04522259	0.9399760	0.3793103
p4	789	635	0.6070976	0.6383291	0.4703640	-0.08635342	0.9471789	0.3929024
p5	790	637	0.6658228	0.6387869	0.5566545	0.07484741	0.9483794	0.3341772

- The output contains the call rate and heterozygosity
- Outliers who have increased average heterozygosity may be suggestive of contaminated DNA samples.

GWA analysis using family-based data

- Load data

```
> load("erfsmall.RData")
```

- The object erfs contains GWA data

```
> class(erfs)
[1] "gwaa.data"
attr(,"package")
[1] "GenABEL"
```

- The object pkins contains the expected kinship matrix derived from pedigree data

```
> class(pkins)
[1] "matrix"
```

GWA analysis using family-based data

- Number of individuals and of markers

```
> nids(erfs)
[1] 150
> nsnps(erfs)
[1] 5827
```

- Distribution of markers over chromosomes

```
> table(chromosome(erfs))
  1  10  11  12  13  14  15  16  17  18  19   2  20  21  22  23   3   4   5   6   7   8   9   X
484 251 224 285 188 210 206 189 159 170 149 481 130 100 132  26 403 300 320 397 286 253 206 278
```

- Chromosome 23 refers to autosomal region of X chromosome

GWA analysis using family-based data

- Generate summary statistics of markers

```
> descriptives.marker(gtdata(erfs))
$`Minor allele frequency distribution`
     X<=0.01 0.01<X<=0.05 0.05<X<=0.1 0.1<X<=0.2    X>0.2
No    17.000       26.000      75.000    437.000 5272.000
Prop   0.003        0.004       0.013      0.075    0.905
$`Cumulative distr. of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No          2   14.000  83.000 319.000  5827
Prop        0    0.002   0.014   0.055     1
```

- Most SNPs have high MAFs, there is little deviation from HWE

GWA analysis using family-based data

- Callrates are also outputted

$`Distribution of porportion of successful genotypes (per person)`

	X<=0.9	0.9<X<=0.95	0.95<X<=0.98	0.98<X<=0.99	X>0.99
No	0	1.000	2.000	12.00	135.0
Prop	0	0.007	0.013	0.08	0.9

$`Distribution of porportion of successful genotypes (per SNP)`

	X<=0.9	0.9<X<=0.95	0.95<X<=0.98	0.98<X<=0.99	X>0.99
No	77.000	33.000	214.000	208.000	5295.000
Prop	0.013	0.006	0.037	0.036	0.909

- Callrates are generally high, both per person and per SNP

GWA analysis using family-based data

- There is also information on the distribution of heterozygosity

$`Mean heterozygosity for a SNP`

[1] 0.4402752

$`Standard deviation of the mean heterozygosity for a SNP`

[1] 0.08287253

$`Mean heterozygosity for a person`

[1] 0.4354001

$`Standard deviation of mean heterozygosity for a person`

[1] 0.01305448

- The low standard deviation on per person heterozygosity also indicates high quality of genotypic data

Kinship coefficients

- Look at the pedigree kinship matrix pkins

```
> pkins[1:5, 1:5]
            id1         id2         id3         id4         id5
id1 5.00000e-01 8.56146e-05 1.01984e-04 2.33397e-04 8.56146e-05
id2 8.56146e-05 5.00000e-01 3.96513e-03 2.56896e-05 2.51269e-01
id3 1.01984e-04 3.96513e-03 5.00000e-01 1.21593e-05 3.96513e-03
id4 2.33397e-04 2.56896e-05 1.21593e-05 5.00000e-01 2.56896e-05
id5 8.56146e-05 2.51269e-01 3.96513e-03 2.56896e-05 5.00000e-01
```

- Symmetric matrix
- Values between 0 (unrelateds) and 1/2 (diagonal + identical twins)
- Sib pairs or parent offspring pairs have kinship 1/4, e.g. id2 and id5 with some inbreeding making the kinship > 1/4

Kinship coefficients

- Summarize distribution of pedigree kinship coefficients

```
> summary(pkins[lower.tri(pkins)])
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.0000000 0.0004999 0.0028500 0.0062810 0.0053780 0.2633000
```

- Average kinship is quite low
- This can also be visualized in a histogram

```
> hist(pkins[lower.tri(pkins)])
```

- Almost all probability mass is on low values of kinship

Kinship coefficients

- Kinship can also be estimated from genomic data

```
> gkins <- ibs(erfs[, autosomal(erfs)], weight = "freq")
> gkins[1:5, 1:5]
            id1          id2          id3          id4          id5
id1  0.513083834  5.439000e+03  5.446000e+03 5441.00000000 5440.0000000
id2 -0.012569446  4.931959e-01  5.524000e+03 5524.00000000 5521.0000000
id3  0.001516184 -8.551323e-03  5.044065e-01 5529.00000000 5528.0000000
id4  0.010459422 -1.471218e-02 -3.467894e-03    0.50739823 5523.0000000
id5 -0.007957596  2.561484e-01 -8.925127e-03   -0.02205665    0.4943105
```

- Nr of informative pairs used for estimation given above diagonal
- Genomic kinship gives unbiased estimate but can be lower than 0

Kinship coefficients

- Summary of genomic kinship coefficients

```
> summary(gkins[lower.tri(gkins)])
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.038980 -0.011930 -0.005826 -0.003362  0.000481  0.268000
```

- High correlation between pedigree and genomic kinship

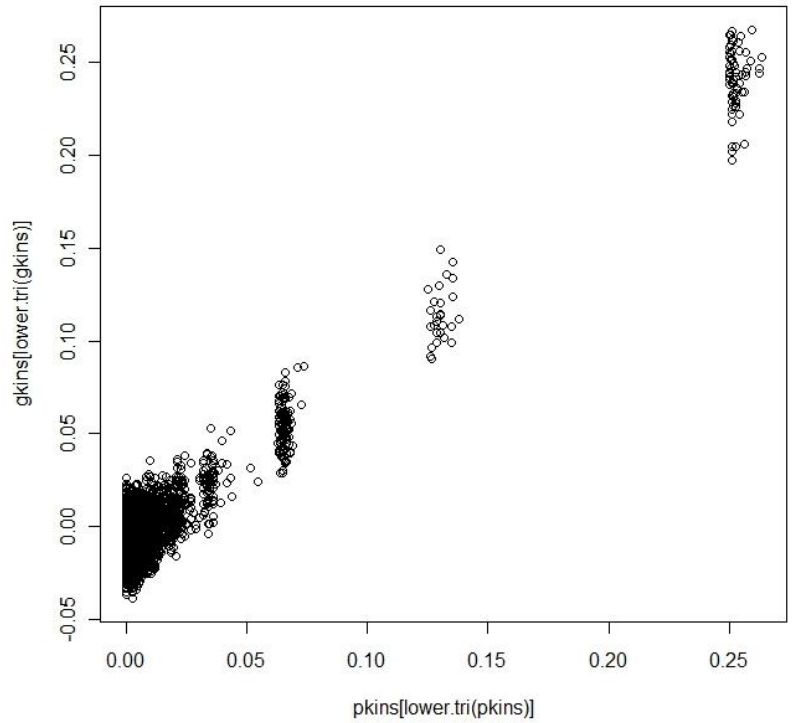
```
> cor(pkins[lower.tri(pkins)], gkins[lower.tri(gkins)])
[1] 0.91615
```

- Despite the high correlation, these coefficients are not identical
- Generally, pedigrees are more prone to errors than genotypic, hence genomic kinship is prepared if enough SNPs are available for its estimation (i.e. GWA, not candidate gene studies)

Kinship coefficients

- Plot pedigree versus genomic kinship

```
> plot(pkins[lower.tri(pkins)], gkins[lower.tri(gkins)])
```



Polygenic model and GRAMMAS analysis

- Estimate the polygenic model for the trait qtbas

```
> h2 <- polygenic(qtbas, kin = pkins, data = erfs)
> h2$h2an
$minimum
[1] 173.9357
$estimate
[1] 0.1509599 0.5833869 1.2264689
$gradient
[1] -2.842171e-08  5.684342e-08  1.407633e-07
$code
[1] 1
$iterations
[1] 10
```

Polygenic model and GRAMMAS analysis

- Estimates are given for
  - intercept
  - heritability (fraction of polygenic out of total variance): 58%
  - total variance (sum of polygenic and environmental): 1.23
- Run GRAMMAS analysis

```
> grs <- qtscore(h2$pgres, data = erfs, clam = FALSE)
```

- Obtain ‘genomic control’ inflation factor

```
> lambda(grs)$est
[1] 0.9122118
```

- Note that it is actually a deflation factor, necessary to get non-conservative test statistics (allowed by option))

Polygenic model and GRAMMAS analysis

- Report top findings

```
> descriptives.scan(grs, sort = "Pc1df")

Summary for top 10 results, sorted by Pc1df

          Chromosome   Position Strand A1 A2   N       effB    se_effB  chi2.1df        P1df
effAB      effBB  chi2.2df        P2df        Pc1df

rs1075456         15 3036078968      u  1  2 150 -0.1935121 0.05157344 14.078742
0.0001753141 -0.06334842 -0.4094758 17.155466 0.0001882512 8.545384e-05

rs1781670         13 2694164735      u  1  2 150  0.1793708 0.05328373 11.332208
0.0007617427  0.19536920  0.3572530 11.377006 0.0033846560 4.241281e-04

rs1264007          X 4256199578      u  1  2 150 -0.1554757 0.04733800 10.787113
0.0010220916 -0.17590133 -0.3050234 10.805101 0.0045050754 5.843354e-04

rs1054889          2  324154337      u  1  2 150  0.1866985 0.05727787 10.624488
0.0011159984  0.08202318  0.4069583 12.925494 0.0015605031 6.430777e-04...
```

Polygenic model and GRAMMAS analysis

- Various 2df (genotypic) and 1df (allelic) effect sizes and p-values are reported
- Effect sizes are underestimated with respect to full measured genotype model
- Most relevant to us is the p-value Pc1df which is 1df and corrected for deflation of test statistics
- GRAMMAS is very fast compared to the full MG model
- It is possible to perform permutations in order to correct for multiple testing

```
> grs.e <- qtscore(h2$pgres, data = erfs, times = 200, clam = FALSE, quiet = TRUE)
```

Polygenic model and GRAMMAS analysis

- Report top findings: no SNP meets genome-wide significance

```
> descriptives.scan(grs.e, sort = "Pc1df")

Summary for top 10 results, sorted by Pc1df

          Chromosome   Position Strand A1 A2   N       effB    se_effB  chi2.1df  P1df Pc1df
effAB       effBB  chi2.2df P2df

rs1075456         15 3036078968      u  1  2 150 -0.19351205 0.05157344 14.078742
0.595 0.360 -0.06334842 -0.40947582 17.155466 0.49

rs1781670         13 2694164735      u  1  2 150  0.17937084 0.05328373 11.332208
0.955 0.855  0.19536920  0.35725296 11.377006 1.00

rs1264007          X 4256199578      u  1  2 150 -0.15547569 0.04733800 10.787113
0.995 0.920 -0.17590133 -0.30502341 10.805101 1.00

rs1054889          2  324154337      u  1  2 150  0.18669851 0.05727787 10.624488 0.995
0.935  0.08202318  0.40695828 12.925494 1.00 ...
```

FAM-MDR showcase

- See zip file FAMMDRwithCPP.zip
- No exam material

Exercises

- Which are the reference/coding alleles of SNP rs114 in dataset srdta? Obtain their frequencies in the total sample and in males and females separately.
- Repeat the GRAMMAS analysis for the trait qt, also available in the erfs data. Interpret your findings.