Practicals Bioinformatics 2011-2012

Olivier Stern olivier.stern@ulg.ac.be

Tom Cattaert tom.cattaert@ulg.ac.be

6 December 2011: Multiple testing and interactions

Multiple testing: Bonferroni

- Recall medium-scale analysis of SNPs data

```
> library(SNPassoc)

> data(SNPs)

> myData<-setupSNP(data=SNPs,colSNPs=6:40,sep="")

> myData.o<-setupSNP(SNPs, colSNPs=6:40, sort=TRUE,info=SNPs.info.pos, sep="")

> ans<-WGassociation(protein~1,data=myData.o)

> ans

          comments    codominant dominant recessive overdominant log-additive

snp10004  Monomorphic -          -        -         -            -

…

snp10002  -           0.78525    0.93292  0.48600   0.87267      0.76807

…
```

Multiple testing: Bonferroni

- Bonferroni correction for number of tests performed

```
> Bonferroni.sig(ans, model="log-add", alpha=0.05,include.all.SNPs=FALSE)

number of tests:  21

alpha: 0.05

corrected alpha: 0.002380952

          comments log-additive

snp10001  -        0.001143723

snp100024 -        0.002231790
```

- The corrected alpha equals alpha divided by number of tests

```
> 0.05/21

[1] 0.002380952
```

Multiple testing: false discovery rate

- Recall medium-scale analysis of HapMap data

```
> data(HapMap)

> myDat.HapMap<-setupSNP(HapMap, colSNPs=3:9307, sort =
TRUE,info=HapMap.SNPs.pos, sep="")

> resHapMap<-WGassociation(group, data=myDat.HapMap, model="log-add")

> summary(resHapMap)

     SNPs (n) Genot error (%) Monomorphic (%) Significant* (n)  (%)

chr1    796          3.8           18.6              163 20.5

chr2    789          4.2           13.9              161 20.4

chr3    648          5.2           13.0              132 20.4

chr4    622          6.3           17.7              104 16.7

…
```

Multiple testing: false discovery rate

- Get p-values and remove monomorphic SNPs

```
> pval<-additive(resHapMap)

> pval<-pval[!is.na(pval)]
```

- Calculate q-values

```
> install.packages('qvalue')

> library(qvalue)

> qobj<-qvalue(pval)

> qobj$qvalues[1:4]

[1] 1.128563e-01 2.309632e-07 2.930540e-10 2.777937e-01
```

- Obtaining the false discovery rate (FDR) for e.g. p-value 0.001

```
> max(qobj$qvalues[qobj$pvalues <= 0.001])

[1] 0.0006046515
```

Multiple testing: multtest package

- Install multtest package

```
> source("http://www.bioconductor.org/biocLite.R")

> biocLite("Biobase")

> install.packages('multtest')

> library(multtest)
```

- Apply several multiple testing strategies

```
> procs<-c("Bonferroni","Holm","Hochberg","SidakSS","SidakSD","BH","BY")

> res2<-mt.rawp2adjp(pval,procs)

> res2$adjp[1:10,]

             rawp   Bonferroni         Holm     Hochberg SidakSS SidakSD           BH           BY
 [1,] 1.740932e-32 1.274362e-28 1.274362e-28 1.274362e-28       0       0 1.274362e-28 1.207541e-27

 [2,] 3.914510e-32 2.865421e-28 2.865030e-28 2.865030e-28       0       0 1.432711e-28 1.357586e-27

…
```

Multiple testing: multtest package

- Obtain number of rejected hypotheses at various significance levels

```
> mt.reject(res2$adjp,seq(0,0.1,0.001))$r
      rawp Bonferroni Holm Hochberg SidakSS SidakSD   BH   BY
0        0          0    0        0     220     220    0    0
0.001 3342       1518 1537     1537    1518    1537 3099 2453
0.002 3549       1591 1650     1650    1591    1650 3322 2642
0.003 3731       1671 1705     1705    1671    1705 3487 2779
0.004 3785       1710 1782     1782    1711    1782 3540 2829
0.005 3845       1751 1811     1811    1751    1812 3611 2875
0.006 3893       1800 1831     1831    1800    1831 3735 2926
0.007 4009       1817 1855     1855    1817    1855 3764 3035
0.008 4045       1831 1873     1873    1832    1874 3801 3051
…
```

Multiple testing: permutation tests

- Permute cases and controls 1000 times

```
> resHapMap.perm<-scanWGassociation(group, data=myDat.HapMap,model="log-add", nperm=1000)

> summary(resHapMap.perm)

     SNPs (n) Genot error (%) Monomorphic (%) Significant* (n)  (%)
chr1      796             0            18.6              143 18.0
chr2      789             0            13.9              143 18.1
chr3      648             0            13.0              115 17.7
chr4      622             0            17.7               92 14.8
chr5      587             0            14.7              104 17.7
chr6      556             0            16.9               86 15.5
…
```

Multiple testing: permutation tests

- Perform the actual permutation test calculations

```
> res.perm<- permTest(resHapMap.perm)

> print(res.perm)

Permutation test analysis (95% confidence level)

Number of SNPs analyzed:  9305

Number of valid SNPs (e.g., non-Monomorphic and passing calling rate):  7320

P value after Bonferroni correction:  6.83e-06

P values based on permutation procedure:

P value from empirical distribution of minimum p values:  2.883e-05

P value assuming a Beta distribution for minimum p values:  2.445e-05
```

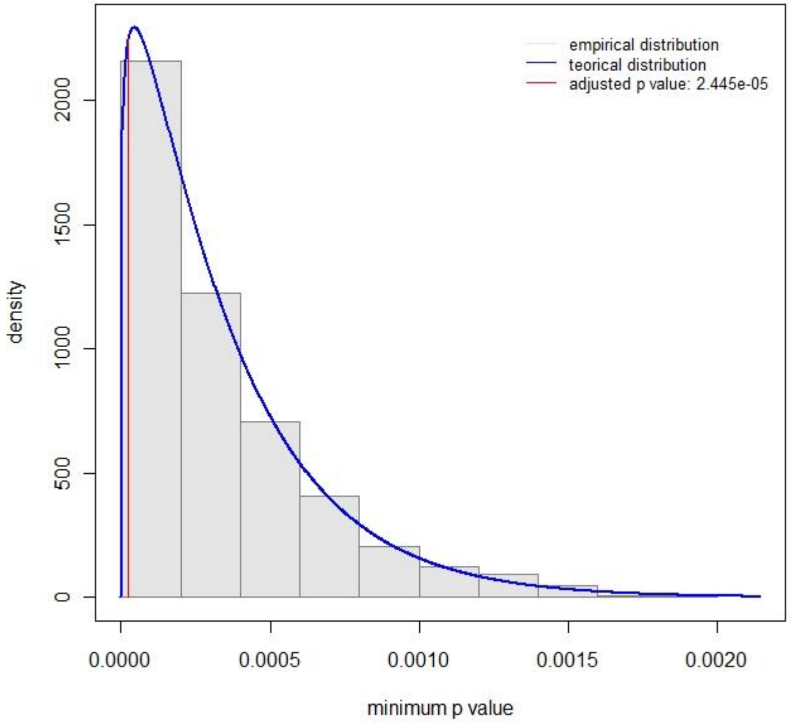
- Get the p-values in the permuted datasets

```
> perms <- attr(resHapMap.perm, "pvalPerm")

> dim(perms)

[1] 9305 1000
```

Multiple testing: permutation tests

- Plot permutation test results

```
> plot(res.perm)
```



Multiple testing: permutation tests

- Rank truncated product [Dudbridge et al. 2006] is also implemented

```
> res.perm.rtp<- permTest(resHapMap.perm,method="rtp",K=20)

> print(res.perm.rtp)

Permutation test analysis (95% confidence level)

Number of SNPs analyzed:  9305

Number of valid SNPs (e.g., non-Monomorphic and passing calling rate):  7320

P value after Bonferroni correction:  6.83e-06

Rank truncated product of the K=20 most significant p-values:

Product of K p-values (-log scale):  947.2055

Significance:  <0.001
```

Interaction analysis: GxE

- Analyze SNP interacting with gender

```
> ans<-association(log(protein)~snp10001*sex+blood.pre,data=myData,model="codominant")

> print(ans,dig=2)

     SNP: snp10001  adjusted by: blood.pre

 Interaction

      Male       dif lower upper    Female         dif lower upper

T/T 40    11 0.08  0.00     NA     NA 52    10.6 0.079 -0.026 -0.29  0.24

C/T 27    11 0.10 -0.13 -0.45  0.19 26    10.2 0.184 -0.472 -0.79 -0.15

C/C  8    10 0.35 -0.64 -1.13 -0.14  4     9.8 0.286 -0.887 -1.56 -0.22

p interaction: 0.36051
```

Interaction analysis: GxE

- Analyze SNP interacting with gender: more output

```
sex within snp10001
T/T
       n me    se    dif lower upper
Male   40 11 0.080  0.000    NA    NA
Female 52 11 0.079 -0.026 -0.29  0.24
C/T
       n me   se   dif lower  upper
Male   27 11 0.10  0.00    NA     NA
Female 26 10 0.18 -0.34 -0.69 0.0086
C/C
      n   me   se   dif lower upper
Male   8 10.0 0.35  0.00    NA    NA
Female 4  9.8 0.29 -0.25  -1.0  0.53
p trend: 0.26575
```

```
snp10001 within sex
Male
     n me   se   dif lower upper
T/T 40 11 0.08  0.00    NA    NA
C/T 27 11 0.10 -0.13 -0.45  0.19
C/C  8 10 0.35 -0.64 -1.13 -0.14
Female
     n   me    se   dif lower upper
T/T 52 10.6 0.079  0.00    NA    NA
C/T 26 10.2 0.184 -0.45 -0.75 -0.14
C/C  4  9.8 0.286 -0.86 -1.52 -0.20
p trend: 0.36051
```

Interaction analysis: GxG

- Analyze two interacting SNPs

```
> ans<-association(log(protein)~snp10001*factor(recessive(snp100019))+blood.pre,data=myData,
model="codominant")

> print(ans,dig=2)

      SNP: snp10001  adjusted by: blood.pre

 Interaction

       G/G-C/G         dif lower  upper    C/C         dif lower upper

T/T 60      11 0.063  0.00    NA     NA 32  11 0.11 -0.038 -0.32  0.24

C/T 53      10 0.106 -0.30 -0.54 -0.053  0   0 0.00     NA    NA    NA

C/C 12      10 0.244 -0.72 -1.13 -0.313  0   0 0.00     NA    NA    NA

p interaction: NA
```

Interaction analysis: GxG

- Analyze two interacting SNPs: more output

```
factor(recessive(snp100019)) within
snp10001
T/T
         n me    se    dif lower upper
G/G-C/G 60 11 0.063  0.000    NA    NA
C/C     32 11 0.112 -0.038 -0.32  0.24
C/T
         n me   se dif lower upper
G/G-C/G 53 10 0.11   0    NA    NA
C/C      0  0 0.00  NA    NA    NA
C/C
         n me   se dif lower upper
G/G-C/G 12 10 0.24   0    NA    NA
C/C      0  0 0.00  NA    NA    NA
p trend: NA
```

```
snp10001 within
factor(recessive(snp100019))
G/G-C/G
     n me    se   dif lower  upper
T/T 60 11 0.063  0.00    NA     NA
C/T 53 10 0.106 -0.30 -0.54 -0.053
C/C 12 10 0.244 -0.72 -1.13 -0.313
C/C
     n me   se dif lower upper
T/T 32 11 0.11   0    NA    NA
C/T  0  0 0.00  NA    NA    NA
C/C  0  0 0.00  NA    NA    NA
p trend: NA
```

Interaction analysis: GxG

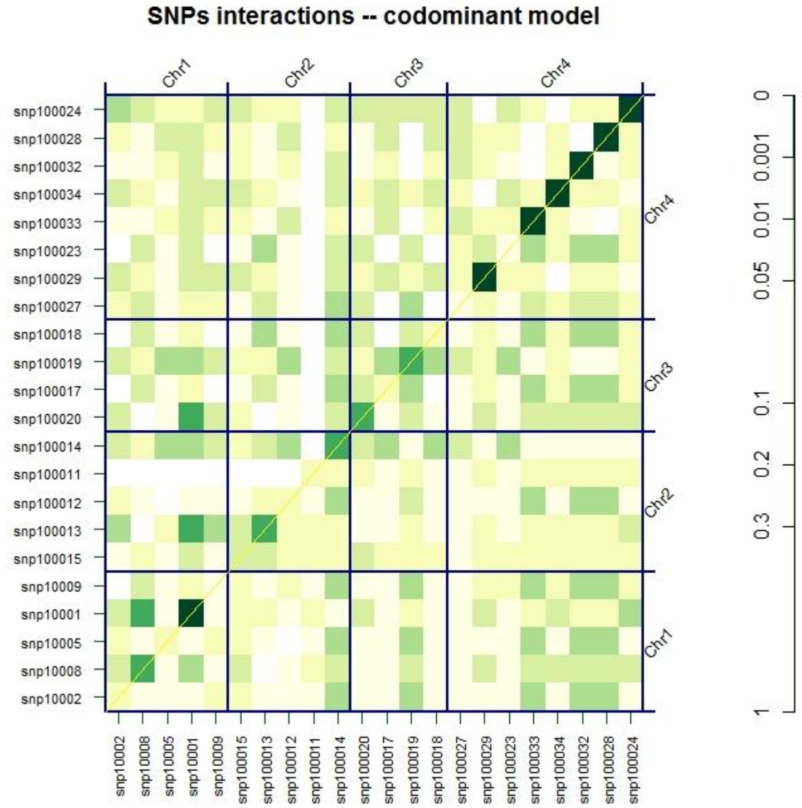
- Study gene-gene interaction

```
> ansCod<-interactionPval(log(protein)~sex, data=myData.o,model="codominant")
> ansCod[1:7,1:7]
         snp10004 snp10007 snp100010  snp10002 snp10003   snp10008  snp10005
snp10004       NA       NA        NA        NA       NA         NA        NA
snp10007       NA       NA        NA        NA       NA         NA        NA
snp100010      NA       NA        NA        NA       NA         NA        NA
snp10002       NA       NA        NA 0.4670088       NA 0.06423172 0.4187811
snp10003       NA       NA        NA        NA       NA         NA        NA
snp10008       NA       NA        NA 0.6488757       NA 0.00577702 0.6412163
snp10005       NA       NA        NA 0.6984826       NA 0.72232141 0.3777925
```

Interaction analysis: GxG

- Plot results of interaction analysis

```
> plot(ansCod)
```



Interactions and CART

- Trees allow discovery of a specific form of conditional association
- Trees do not specifically search for statistical interaction
- Consider the situation of a trait **y** with a very strong independent effect of covariate $\mathbf{x}_1$ such that the first split of the tree is on $\mathbf{x}_1$
- Suppose there is a second predictor $\mathbf{x}_2$ and that there is statistical interaction between $\mathbf{x}_1$ and $\mathbf{x}_2$, i.e. there is a difference γ in effect of $\mathbf{x}_2$ for both levels $\mathbf{x}_1 = 0$ and $\mathbf{x}_1 = 1$
- Suppose that there is also a variable $\mathbf{x}_3$ with an independent effect on **y**, regardless of the level of $\mathbf{x}_1$
- Formally we are looking at the linear model

  $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \gamma \mathbf{x}_1 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \varepsilon$

Interactions and CART

- After the initial split on $\mathbf{x}_1$, the model becomes
    - for $\mathbf{x}_1 = 0$: $\mathbf{y} = \beta_0 + \beta_2\, \mathbf{x}_2 + \beta_3\, \mathbf{x}_3 + \varepsilon$
    - for $\mathbf{x}_1 = 1$: $\mathbf{y} = (\beta_0 + \beta_1) + (\beta_2 + \gamma)\, \mathbf{x}_2 + \beta_3\, \mathbf{x}_3 + \varepsilon$
- The next split within the daughter nodes depends on relative magnitude of the regression coefficients
- E.g. if β_3 is large compared to β_2 and $\beta_2 + \gamma$, it is likely that the next split will be on the variable $\mathbf{x}_3$ in both daughter nodes, although only $\mathbf{x}_1$ and $\mathbf{x}_2$ interact statistically
- Hence, for trees conditional association is more relevant than statistical interaction

Exercises

- Within chromosome 6 of the HapMap data perform an association analysis of the group variable using the dominant model. Correct for multiple testing using different approaches that control the family-wise error rate at 5% (e.g. Bonferroni, permutations), or that control the false discovery rate at 5% (e.g. Benjamini-Hochberg, qvalue approach)
- Investigate gene-environment interaction of snp100025 and sex in determining case-control status in the SNPs dataset, adjusted for protein level
- Visualize gene-gene interactions within chromosome 4 of the SNPs data with respect to the case-control status