Overview of GGtools for genetics of gene expression in Bioconductor for Bioc 2.2, 2008

May 20, 2008

Contents

1 Introduction

The *GGtools* package contains infrastructure and demonstration data for joint analysis of transcriptome and genome through combination of DNA expression microarray and high-density SNP genotyping data. For Bioconductor 2.2 we adopted a representation of genotypes due to Clayton (in package *snpMatrix*) allowing reasonably convenient storage and manipulation of 4 megaSNP phase II HapMap genotypes on all the CEPH CEU samples. This contrasts with the previous version of *GGtools* which was limited to 550 kiloSNP and 58 CEU founders.

To give an immediate taste of the capabilities, we attach the package and load some test data.

```
> library(GGtools)
> data(hmceuB36.2021)
> hmceuB36.2021

snp.matrix-based genotype set:
number of samples:  90
```

```
number of snp.matrix:  2
annotation:
 exprs: illuminaHumanv1.db
 snps: snp locs package: GGBase ; SQLite ref: hmceuAmbB36_23a_dbconn
Expression data: 47293 x 90
Phenodata: An object of class "AnnotatedDataFrame"
  sampleNames: NA06985, NA06991, ..., NA12892  (90 total)
  varLabels and varMetadata description:
    famid: hapmap family id
    persid: hapmap person id
    ...: ...
    isAdad: logical TRUE if person is a father
    (9 total)
```

Expression data are recoverable in a familiar way:

```
> exprs(hmceuB36.2021)[1:5, 1:5]
```

```
                NA06985   NA06991   NA06993   NA06994   NA07000
GI_10047089-S  5.983962  5.939529  5.912270  5.891347  5.906675
GI_10047091-S  6.544493  6.286516  6.244446  6.277397  6.330893
GI_10047093-S  9.905235 10.353804 10.380972  9.889223 10.155686
GI_10047099-S  7.993935  7.593970  8.261215  6.598430  6.728085
GI_10047103-S 11.882265 12.204753 12.249708 11.798415 12.015252
```

Genotype data have more complex representation.

```
> smList(hmceuB36.2021)
```

```
$`20`
A snp.matrix with  90 rows and  119921 columns
Row names:  NA06985 ... NA12892
Col names:  rs4814683 ... rs6090120

$`21`
A snp.matrix with  90 rows and  50165 columns
Row names:  NA06985 ... NA12892
Col names:  rs885550 ... rs10483083
```

This shows that we use a named list to hold items of the *snp.matrix* class from *snpMatrix*. We can dig in with a different accessor:

```
> rawSNP(hmceuB36.2021, 20)[1:5, 1:5]
```

```
        rs4814683 rs6076506 rs6139074 rs1418258 rs7274499
NA06985        03        03        03        03        03
NA06991        02        03        02        02        03
NA06993        01        03        01        01        03
NA06994        01        03        01        01        03
NA07000        03        03        03        03        03
```

and now we see that an unusual data representation is in use: the leading zeroes indicate that a raw byte representation is being shown.

We can coerce from this representation to allele tokens:

```
> as(smList(hmceuB36.2021)[["20"]][1:5, 1:5], "character")

     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] "B/B" "B/B" "B/B" "B/B" "B/B"
[2,] "A/B" "B/B" "A/B" "A/B" "B/B"
[3,] "A/A" "B/B" "A/A" "A/A" "B/B"
[4,] "A/A" "B/B" "A/A" "A/A" "B/B"
[5,] "B/B" "B/B" "B/B" "B/B" "B/B"
```

2 Chromosome-wide SNP screens

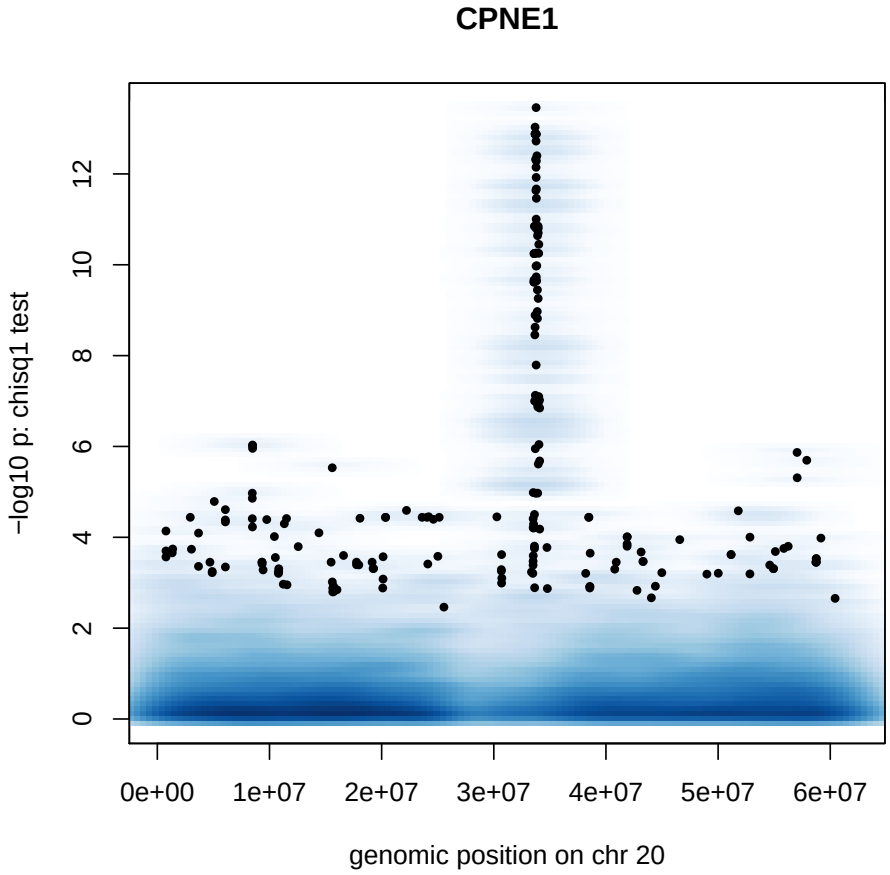
The demonstration object `hmceuB36.2021` has only chromosomes 20 and 21. An object in the *GGdata* package called `hmceuB36` has all 24 chromosomes. The *GGBase* package has the SNP locations for hmceuB36 in an external SQLite store that is exported dynamically upon attaching GGBase, on which GGtools depends.

It is fairly easy to test for eQTL for a given gene on a specified chromosome:

```
> g1 = gwSnpScreen(genesym("CPNE1"), hmceuB36.2021, chrnum(20))
> class(g1)

[1] "cwSnpScreenResult"
attr(,"package")
[1] "GGBase"

> g1

gwSnpScreenResult with 1 inference data.frames
gene used:  CPNE1 ; expression platform: illuminaHumanv1.db
```

We can visualize the results over the chromosome:

```
> plot(g1)
```



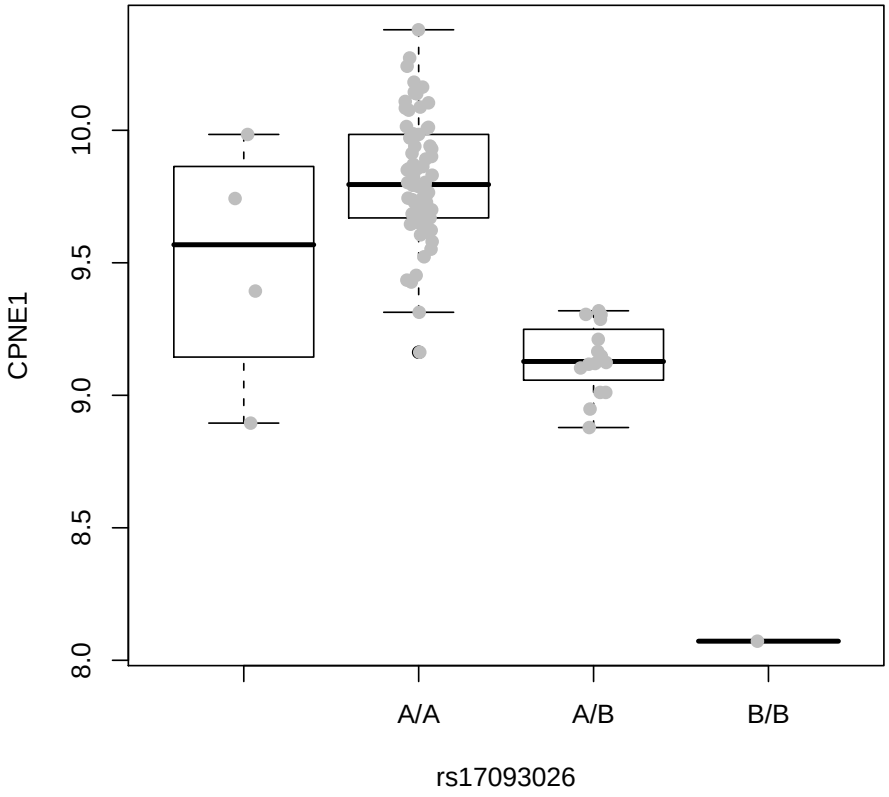
and we can get a report on the most highly associated SNPs:

```
> tt = topSnps(g1)
> tt
```

```
                  p.1df
rs17093026 3.464560e-14
rs1118233  9.322326e-14
rs7273815  1.304436e-13
rs2425078  1.330170e-13
rs1970357  1.330170e-13
rs12480408 1.330170e-13
rs6060535  1.330170e-13
rs11696527 1.330170e-13
rs6058303  1.330170e-13
rs6060578  1.330170e-13
```

To see the relationship between expression values and genotype, the `plot_EvG` method can be used.

```
> plot_EvG(genesym("CPNE1"), rsNum(rownames(tt)[1]), hmceuB36.2021)
```



3 Genome-wide SNP screens

The *GGdata* package contains a representation of all chromosomes.

```
> library(GGdata)
> if (!exists("hmceuB36")) data(hmceuB36)
```

We can use `gwSnpScreen` to find eQTL, if we have sufficient memory. On windows, we skip this step.

```
> ONWIN = FALSE
> ONWIN = (Sys.info()["sysname"] == "Windows")
> if (!ONWIN) {
+     rr = gwSnpScreen(genesym("RPS26"), hmceuB36)
+     rrt = topSnps(rr)
```

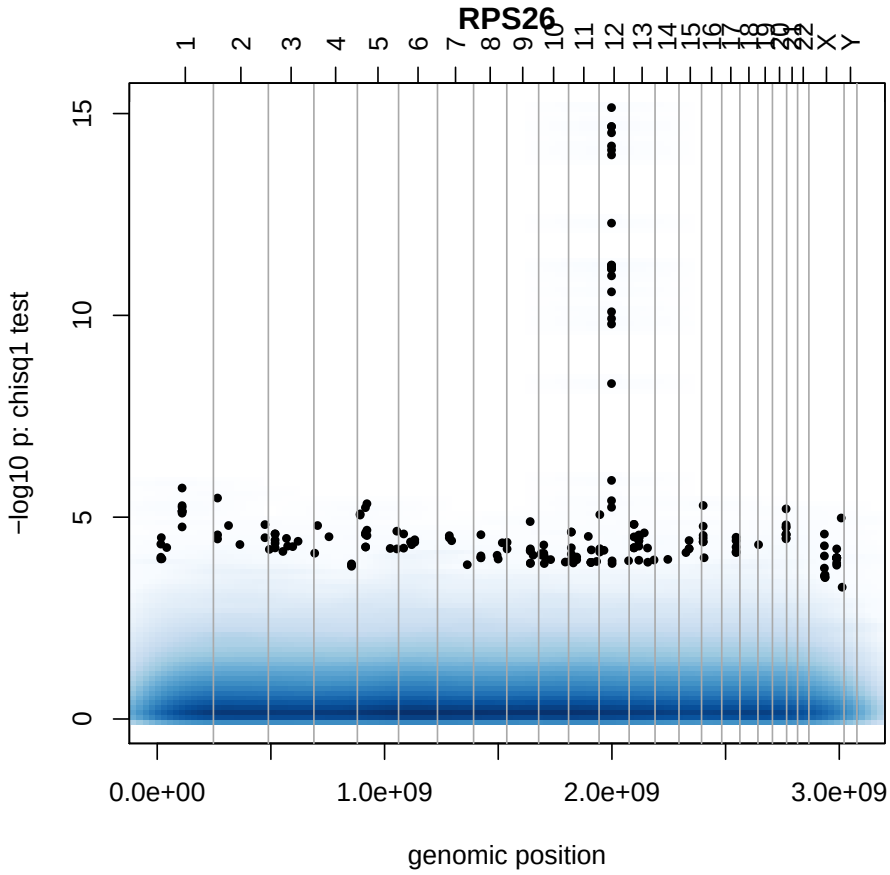
```
+     rrt[[12]]
+ }
```

```
                  p.1df
rs10876864 7.118332e-16
rs773114   2.099500e-15
rs1873914  2.099500e-15
rs1131017  2.992594e-15
rs705699   6.355083e-15
rs11171739 8.001744e-15
rs2271194  1.058659e-14
rs7312770  5.183227e-13
rs772921   5.686637e-12
rs1701704  5.686637e-12
```

Note that topSnps, when applied to a genome-wide screen output, will return a list of chromosome-specific results.

The whole-genome plot is automatic but somewhat slow; on windows, this is a placeholder. Please get the genuine vignette from the web site.

```
> if (ONWIN) {
+     rr = 1
+ }
> plot(rr)
```



4 Phenotype data available on CEPH samples

```
> names(pData(hmceuB36))
```

```
[1] "famid"     "persid"    "mothid"    "fathid"    "sampid"    "isFounder"
[7] "male"      "isAmom"    "isAdad"
```

5 Session information

```
> sessionInfo()
```

```
R version 2.7.0 (2008-04-22)
x86_64-unknown-linux-gnu

locale:
LC_CTYPE=en_US;LC_NUMERIC=C;LC_TIME=en_US;LC_COLLATE=en_US;LC_MONETARY=C;LC_MESSAGES=en
```

```
attached base packages:
 [1] grid      splines   tools     stats     graphics  grDevices utils
 [8] datasets  methods   base

other attached packages:
 [1] GGdata_0.1.1             illuminaHumanv1.db_1.0.0 GGtools_2.0.2
 [4] Biostrings_2.8.6         RColorBrewer_1.0-2       mgu74av2.db_2.2.0
 [7] geneplotter_1.18.0       hgfocus.db_2.2.0         GGBase_2.0.3
[10] snpMatrix_1.4.0          hexbin_1.14.0            lattice_0.17-8
[13] survival_2.34-1          GSEABase_1.2.1           annotate_1.18.0
[16] xtable_1.5-2             AnnotationDbi_1.2.0      RSQLite_0.6-8
[19] DBI_0.2-4                Biobase_2.0.1

loaded via a namespace (and not attached):
[1] cluster_1.11.10    graph_1.18.1       KernSmooth_2.22-22 Ruuid_1.18.0
[5] XML_1.95-2
```