dsQTL: exploring DNA-variants associated with DNaseI hypersensitivity

VJ Carey

July 14, 2012

1 Introduction

Degner et al. (2012) publish information on associations between DNA variants (SNP, SNV, and indels) and DNaseI hypersensitivity measures acquired via DNase-Seq.

This package includes information from the Chicago group on normalized DNase-seq data for chromosomes 2 and 17, and genotype data from chromosome 2 only.

2 The basic data structure

```
> library(dsQTL)
> data(DSQ_17)
> exptData(DSQ_17)

SimpleList of length 2
names(2): MIAME annotation

> exptData(DSQ_17)[[1]]

Experiment data
  Experimenter name: Degner JF
  Laboratory: Department of Human Genetics, University of Chicago, Chicago, Illinois 60
  Contact information:
  Title: DNaseI sensitivity QTLs are a major determinant of human expression variation.
  URL:
  PMIDs: 22307276

  Abstract: A 252 word abstract is available. Use 'abstract' method.
```

We use summarized experiment structure for the assay data, but the imputed genotype data are kept separate, in the package, in the inst/parts folder.

The data structure on chr2, which will be used to reproduce some findings, is more mature

```
> data(DSQ_2)
> names(assays(DSQ_2))

[1] "normDHS"

> assays(DSQ_2)[[1]][1:5,1:5]

              NA18486     NA18498    NA18499    NA18501    NA18502
dhs_2_1202 -0.2684343 -0.78076674 -0.4840237  2.3894003 -1.0813642
dhs_2_1602 -1.4445813  0.92170439  0.5812017  0.8627376  0.5186581
dhs_2_2002  0.7624075 -0.12340745 -1.1821308  1.4253179  0.3125592
dhs_2_7502  0.1242963  0.60788505  0.6754706 -0.0452303  0.4876332
dhs_2_8802 -0.9554503 -0.06016578 -0.1990696  1.9383937 -1.3758668

> rowData(DSQ_2)

GRanges with 96024 ranges and 0 elementMetadata cols:
                   seqnames                 ranges strand
                      <Rle>              <IRanges>  <Rle>
       dhs_2_1202     chr2         [ 1202,  1301]      *
       dhs_2_1602     chr2         [ 1602,  1701]      *
       dhs_2_2002     chr2         [ 2002,  2101]      *
       dhs_2_7502     chr2         [ 7502,  7601]      *
       dhs_2_8802     chr2         [ 8802,  8901]      *
      dhs_2_14202     chr2         [14202, 14301]      *
      dhs_2_14302     chr2         [14302, 14401]      *
      dhs_2_34902     chr2         [34902, 35001]      *
      dhs_2_35102     chr2         [35102, 35201]      *
              ...      ...                    ...    ...
  dhs_2_242689402     chr2 [242689402, 242689501]      *
  dhs_2_242689502     chr2 [242689502, 242689601]      *
  dhs_2_242696902     chr2 [242696902, 242697001]      *
  dhs_2_242697402     chr2 [242697402, 242697501]      *
  dhs_2_242698102     chr2 [242698102, 242698201]      *
  dhs_2_242711702     chr2 [242711702, 242711801]      *
  dhs_2_242737502     chr2 [242737502, 242737601]      *
  dhs_2_242737902     chr2 [242737902, 242738001]      *
  dhs_2_242739902     chr2 [242739902, 242740001]      *
```

```
  ---
  seqlengths:
   chr2
     NA
```

To implement the GGBase protocol for on-the-fly generation of smlSet instances from getSS queries, we have an ExpressionSet instance with specific names.

```
> data(eset, package="dsQTL")
> ex

ExpressionSet (storageMode: lockedEnvironment)
assayData: 96024 features, 70 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: NA18486 NA18498 ... NA19257 (70 total)
  varLabels: naid one ... isFounder (9 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation:
```

The genotype data supplied by Degner et al are imputed to 1000 genomes haplotypes, and are reals in [0,2]. For simplicity the current image of the data uses the rounding of the fractional genotypes x with round(x,0).

The feature data refer to the retained 100bp segments that were summarized for DNaseI hypersensitivity and found to lie in the uppermost 5% of the distribution.

```
> library(Biobase)
> fData(ex)[1:5,,drop=FALSE]

data frame with 0 columns and 5 rows
```

We can get the integrated container as

```
> library(GGBase)
> ds2 = getSS("dsQTL", "roundGT_2")
```

the name indicates that we simply rounded the imputed fractional genotypes to nearest integer.

A very restricted search is:

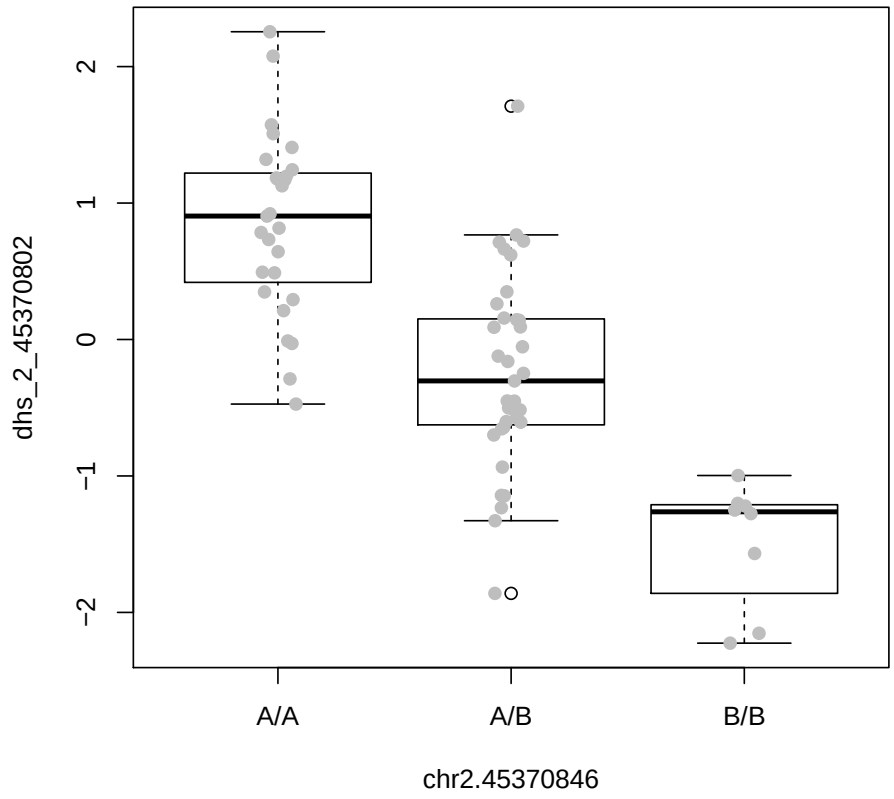
```
> # need to get rid of SNPlocs package getSNPlocs
> getSNPlocs = dsQTL::getSNPlocs  # force
> library(GGtools)
> #library(parallel)
> #options(mc.cores=12)
> n1 = best.cis.eQTLs(smpack="dsQTL", radius=2000, geneannopk="dsQTL",
+   snpannopk="dsQTL", chrnames="2", smchrpref="roundGT_",
+   smFilter = function(x) GTFfilter(x, lower=0.05)[23810:23830,],
+ #  geneApply=mclapply)
+   geneApply=lapply)

get data...build map...run smFilter...filter probes in map...tests...filter...done.
get data...build map...run smFilter...filter probes in map...tests...filter...done.
get data...build map...run smFilter...filter probes in map...tests...filter...done.

> n1

GGtools mcwBestCis instance.  The call was:
best.cis.eQTLs(smpack = "dsQTL", radius = 2000, chrnames = "2",
    smchrpref = "roundGT_", geneApply = lapply, geneannopk = "dsQTL",
    snpannopk = "dsQTL", smFilter = function(x) GTFfilter(x,
        lower = 0.05)[23810:23830, ])
Best loci for 21 are recorded.
Top  4 probe:SNP combinations:
GRanges with 4 ranges and 5 elementMetadata cols:
                 seqnames              ranges strand |     score         snpid
                    <Rle>           <IRanges>  <Rle> | <numeric>   <character>
  dhs_2_45370802        2 [45368802, 45372901]      * |     38.64 chr2.45370846
  dhs_2_45370702        2 [45368702, 45372801]      * |     29.11 chr2.45370846
  dhs_2_45369802        2 [45367802, 45371901]      * |     19.14 chr2.45370846
  dhs_2_45305002        2 [45303002, 45307101]      * |      6.43 chr2.45307016
                    snploc radiusUsed       fdr
                 <integer>  <numeric> <numeric>
  dhs_2_45370802  45370846       2000 0.0000000
  dhs_2_45370702  45370846       2000 0.0000000
  dhs_2_45369802  45370846       2000 0.0000000
  dhs_2_45305002  45307016       2000 0.1666667
  ---
  seqlengths:
    2
   NA
====
use chromsUsed(), fullreport(), etc. for additional information.
```

```
> plot_EvG(probeId("dhs_2_45370802"), rsid("chr2.45370846"), getSS("dsQTL", "roundGT_
```



3 Session information

```
> sessionInfo()

R version 2.15.1 (2012-06-22)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] stats4    splines   stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] GGtools_4.4.0       Rsamtools_1.8.5     Biostrings_2.24.1
 [4] dsQTL_0.0.18        GGBase_3.18.0       snpStats_1.6.0
 [7] Matrix_1.0-6        lattice_0.20-6      survival_2.36-14
[10] Biobase_2.16.0      GenomicRanges_1.8.7 IRanges_1.14.4
[13] BiocGenerics_0.2.0

loaded via a namespace (and not attached):
 [1] AnnotationDbi_1.18.1    BSgenome_1.24.0         DBI_0.2-5
 [4] GenomicFeatures_1.8.2   RCurl_1.91-1            RSQLite_0.11.1
 [7] VariantAnnotation_1.2.9 XML_3.9-4               annotate_1.34.1
[10] biomaRt_2.12.0          bit_1.1-8               bitops_1.0-4.1
[13] ff_2.2-7                genefilter_1.38.0       grid_2.15.1
[16] rtracklayer_1.16.2      tools_2.15.1            xtable_1.7-0
[19] zlibbioc_1.2.0
```