dsQTL: exploring DNA-variants associated with DNaseI hypersensitivity

VJ Carey

May 7, 2016

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, and genotype data from chromosome 2 only.

NOTES ADDED JANUARY 2016. The DHStop5_hg19 SummarizedExperiment instance is the primary data resource of use at this date, although most of the legacy computations described but not run throughout the later parts of this vignette can be made to work.

With the SummarizedExperiment, analysis tools in gQTLstats can be used to perform dsQTL identification, with genotype information housed in tabix-indexed VCF. This is a highly scalable approach.

Here is how we can search for dsQTL by probing into the 1000 genomes VCF in an Amazon S3 bucket for the YRI subject assayed for DNaseI hypersensitivity. We use the first 5 assay results recorded for chromosome 17. This approach assumes internet connectivity and hence is not evaluated on the build system.

```
> library(geuvPack)
> library(dsQTL)
> tf17 = gtpath(17, useS3=TRUE)
> data(DHStop5_hg19)
> dhs17 = DHStop5_hg19[ seqnames(DHStop5_hg19) == "chr17", ]
> seqlevelsStyle(dhs17) = "NCBI"
> library(gQTLstats)
> c17_1 = cisAssoc(dhs17[1:5,], tf17, cisradius=5000)
> c17_1
```

Remaining material is not executed.

2 The basic data structure

```
> library(dsQTL)
> data(DSQ_17)
> metadata(DSQ_17)
> metadata(DSQ_17)[[1]]
```

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

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
```

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]
```

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)

> n1

> plot_EvG(probeId("dhs_2_45370802"), rsid("chr2.45370846"), getSS("dsQTL", "roundGT_
+   wrapperEndo=function(x){annotation(x)="dsQTL"; x}))
```

3 Provenance

3.1 Normed DNaseI hypersensitivity scores

The dsQTL package data structures `DSQ_2` and `DSQ_17` are generated from GEO GSE31388, from which a collection of 70 compressed BED format files were acquired Aug 9 2011. These are imported using the rtracklayer package to obtain location and score information for all the recorded DNaseI hypersensitivity assay results. For example, after import for NA19257, we have

```
Browse[1]> x
[1] "NA19257"
Browse[1]> tmp
RangedData with 1465907 rows and 3 value columns across 22 spaces
           space               ranges   |        name       score   strand
        <factor>            <IRanges>   | <character>   <numeric> <factor>
1           chr1       [  402,   501]   |         NOT -0.67088720        +
2           chr1       [  502,   601]   |         NOT -1.69969288        +
3           chr1       [  602,   701]   |         NOT  0.13520754        +
...          ...                  ... ...         ...         ...      ...
1465905    chr22 [49571602, 49571701]   |         NOT  0.62742318        +
1465906    chr22 [49575102, 49575201]   |         NOT -0.09417379        +
1465907    chr22 [49581602, 49581701]   |         NOT -0.29496269        +
```

The scores for locations on chromosome 2 were collected using

```
> proc1 = function(x) {
+  library(rtracklayer)
+  tmp = import(paste(x, ".qnorm.bed.gz", sep=""))
+  stt = split(tmp, space(tmp))
+  obn = paste(x, "_dsq_chr2", sep="")
+  assign(obn, stt[["chr2"]])
```

```
+  save(list=obn, file=paste(obn, ".rda", sep=""))
+  NULL
+ }
```

The regions and scores reported are described in the GEO metadata as

> We also provide BED file format data for each individual for the top 5% of the genome in terms of total sensitivity. This data was mapped to hg18 using a custom read-mapping algorithm which we describe in detail in the associated publication. Measures of DNase sensitivity were quantile normalized within each individual to a standard normal distribution. Each individual was corrected for GC bias and the top 4 principle (sic) components were removed from the data (See manuscript).

Score data were structured as a matrix with columns corresponding to Yoruba HapMap subject, and rows corresponding to reported hypersensitivity regions.

The `RangedSummarizedExperiment` container is used to unite range and score data in the assays component, and allied metadata are available in metadata and colData components.

3.2 Genotype data

Textual representation of the allelic doses is provided at `http://eqtl.uchicago.edu/dsQTL_data/GENOTYPES/`. As of Oct 2012, these were rounded to allele counts to allow use of snpMatrix representation for chromosome 2 genotypes; propagation of dosage fractions will be undertaken in late 2012.

4 Session information

```
> sessionInfo()

R version 3.3.0 (2016-05-03)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.4 LTS

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=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

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

loaded via a namespace (and not attached):
[1] tools_3.3.0
```