Exploring MatrixEQTL with GGtools

VJ Carey

March 5, 2012

1 Introduction

This document illustrates use of a very rudimentary interface between Bioconductor GGtools (default tests based on snpStats package of D. Clayton) and MatrixEQTL from A. Shabalin. The objective is to assess the impact of choice of test procedure on speed and content of results of a search for cis eQTL on data on human chromosomes 17-20, using illumina expression data on immortalized B cells from GENEVAR project, and Phase 2 HapMap genotypes, on cells extracted from the HapMap CEU cohort, N=90 individuals.

2 Setup

```
> library(parallel)
> options(mc.cores=12)
> suppressPackageStartupMessages(library(GGtools))
> library(MatrixEQTL)
```

3 Default GGtools run

Objective is enumeration of genes on chroms 17-20 that a) have overall variation across samples in the top 40% of the distribution of IQR (non-specific filtering), b) show evidence of association of mean expression and genotype for at least one SNP within 50000 bases of the coding extents for the gene. Expression values are pre-filtered to remove expression heterogeneity by removing the first 10 principal components from the full, column-centered $90 \times 47K$ expression matrix.

Test statistics are score statistics for the additive genetic model (linear regression with count of the number of alphabetically later nucleotides in diallelic SNP as predictor of mean expression). Plug-in estimate of FDR is obtained with 2 permutations of expression against genotype.

```
> set.seed(1234)
> u1b = unix.time(b1b <<- best.cis.eQTLs(smpack="GGdata", rhs=~1,
+  chrnames=c("17", "18", "19", "20"), geneApply=mclapply,
+  smFilter=function(x) nsFilter(MAFfilter(
+     clipPCs(x, 1:10), lower = 0.05), var.cutoff = 0.6)))

> u1b

    user   system  elapsed
1036.177  112.995  503.743

> b1b

GGtools mcwBestCis instance.  The call was:
best.cis.eQTLs(smpack = "GGdata", rhs = ~1, chrnames = c("17",
    "18", "19", "20"), geneApply = mclapply, smFilter = function(x) nsFilter(MAFfilter(
    1:10), lower = 0.05), var.cutoff = 0.6))
Best loci for 1253 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>
  GI_15451941-S       19 [18632614, 18738269]      * |     75.75   rs2283616
  GI_14149701-S       17 [ 4793630,  4898515]      * |     67.39    rs238245
  GI_33859747-S       19 [59036766, 59145763]      * |     67.29   rs8110595
  GI_31542722-S       17 [48574562, 48683211]      * |     65.83   rs8076632
                  snploc radiusUsed       fdr
                <integer>  <numeric> <numeric>
  GI_15451941-S  18682495      50000         0
  GI_14149701-S   4847443      50000         0
  GI_33859747-S  59095126      50000         0
  GI_31542722-S  48625928      50000         0
  ---
  seqlengths:
        17       18       19       20
   80951060 77948226 59145763 62957578
====
use chromsUsed(), fullreport(), etc. for additional information.

> sum(fdr(b1b)<=0.05)

[1] 325

> save(b1b, file="b1b.rda")
> save(u1b, file="u1b.rda")
```

4 GGtools run modified to use MatrixEQTL test statistics

Same objective as above, but the correlation-based statistics of MatrixEQTL. Currently hardwired to use

```
  useModel = modelLINEAR; # modelANOVA or modelLINEAR

  snps = SlicedData$new();
  snps$fileDelimiter = '\t'; # the TAB character
  snps$fileOmitCharacters = 'NA'; # denote missing values;
  snps$fileSkipRows = 1; # one row of column labels
  snps$fileSkipColumns = 1; # one column of row labels
  snps$fileSliceSize = 2000; # read file in pieces
 ...

me = Matrix_eQTL_engine(
        snps,
        gene,
        cvrt,
        output_file_name = outfile(mefob),
        pvOutputThreshold = .2,
        useModel = modelLINEAR,
        errorCovariance = numeric(),
        verbose = FALSE, #TRUE,
        pvalue.hist = 10);
```

to collect the statistics; passage of control parameters up to the main interface is pending.

To manage the voluminous output results, *ff* archives of short ints are populated with

```
me$all$eqtls[, "statistic"] * shortfac
```

where `shortfac` upscales the statistic for truncation to short integer storage. This helps reduce both RAM and disk requirements of storing comprehensive feature x feature test outputs.

```
> set.seed(1234)
> u2b = unix.time(b2b <<- best.cis.eQTLs(chrnames=c("17", "18", "19", "20"), geneAppl
+  smFilter=function(x) nsFilter(MAFfilter(clipPCs(x, 1:10), lower = 0.05), var.cutof
+  useME=TRUE))

> u2b
```

```
    user   system  elapsed
1078.842   89.597 1015.443
```

```
> b2b
```

```
GGtools mcwBestCis instance.  The call was:
best.cis.eQTLs(chrnames = c("17", "18", "19", "20"), geneApply = mclapply,
    smFilter = function(x) nsFilter(MAFfilter(clipPCs(x, 1:10),
        lower = 0.05), var.cutoff = 0.6), useME = TRUE)
Best loci for 1253 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>
  GI_15451941-S       19 [18632614, 18738269]      * |     22.42   rs2283616
  GI_14149701-S       17 [ 4793630,  4898515]      * |     16.64    rs400688
  GI_33859747-S       19 [59036766, 59145763]      * |     16.51   rs8110595
  GI_31542722-S       17 [48574562, 48683211]      * |     15.81   rs8076632
                  snploc radiusUsed       fdr
                 <integer>  <numeric> <numeric>
  GI_15451941-S  18682495      50000         0
  GI_14149701-S   4839930      50000         0
  GI_33859747-S  59095126      50000         0
  GI_31542722-S  48625928      50000         0
  ---
  seqlengths:
        17       18       19       20
   80951060 77948226 59145763 62957578
====
use chromsUsed(), fullreport(), etc. for additional information.
```

```
> save(u2b, file="u2b.rda")
> save(b2b, file="b2b.rda")
> sum(fdr(b2b)<=0.05)
```

```
[1] 325
```

```
> ok1 = which(fdr(b1b)<=0.05)
> ok2 = which(fdr(b2b)<=0.05)
> setdiff(names(fullreport(b1b))[ok1], names(fullreport(b2b))[ok2])
```

```
[1] "GI_32698729-S"
```

```
> setdiff(names(fullreport(b2b))[ok2], names(fullreport(b1b))[ok1])
```

```
[1] "GI_38327534-S"
```

5 Comment

The statistics in use differ but there is little indication of differential sensitivity for this application.

I made no attempt to parallelize any of the MatrixEQTL computations, which could easily be accomplished by splitting the gene set for example, and dispatching to cores as the default GGtools procedure does.

In summary, there should be good performance benefit for basing eQTL searches on the MatrixEQTL package tools. The use of external files as data sources is probably inevitable, but flexible feature filtering is easier if we do this in the R context. Some filtering tasks will involve multiple passes, so efficiencies of such processes should be considered as the package evolves.