yriMulti – HapMap YRI population, multiassay interfaces

Vincent J. Carey, stvjc at channing.harvard.edu

February 2015

Contents


```
## Now getting the GODb Object directly
## Now getting the OrgDb Object directly
## Now getting the TxDb Object directly
## Warning: replacing previous import 'ggplot2::Position' by
## 'BiocGenerics::Position' when loading 'erma'
```

1 Introduction

The EBV-transformed B-cells from Yoruban donors are assayed for genotype and various genomic features in a number of prominent studies. This package helps work with relevant datasets and data structures as use cases for biocMultiAssay package development. A particular concern is accommodation of distributed genotype data, in this case, based on the 1000 genomes VCF files in an S3 bucket.

2 Basic data resources

2.1 Expression data

We will use the RNA-seq expression data in the *geuvPack* package.

```
library(geuvPack)
data(geuFPKM)
```

```
geuFPKM
## class: RangedSummarizedExperiment
## dim: 23722 462
## metadata(3): MIAME constrHist colDataSource
## assays(1): exprs
## rownames(23722): ENSG00000152931.6 ENSG00000183696.9 ...
##   ENSG00000257337.1 ENSG00000177494.5
## rowData names(18): source type ... tag ccdsid
## colnames(462): HG00096 HG00097 ... NA20826 NA20828
## colData names(35): Source.Name Comment.ENA_SAMPLE. ...
##   Factor.Value.laboratory. popcode
```

2.2 Methylation data

We have added 450k data from Banovich, Lan, McVicker, van de Geijn, Degner, Blischak, Roux, Pritchard, and Gilad (2014) paper to the yriMulti package.

```
library(yriMulti)
data(banovichSE)
```

```
banovichSE
## class: RangedSummarizedExperiment
## dim: 329469 64
## metadata(0):
## assays(1): betas
## rownames(329469): cg00000029 cg00000165 ... ch.9.98989607R
##   ch.9.991104F
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(64): NA18498 NA18499 ... NA18489 NA18909
## colData names(35): title geo_accession ... data_row_count naid
```

2.3 DnaseI hypersensitivity data

```
library(dsQTL)
data(DHStop5_hg19)
```

```
DHStop5_hg19
## class: RangedSummarizedExperiment
## dim: 1465442 70
## metadata(2): MIAME annotation
## assays(1): scores
## rownames(1465442): dhs_chr1_10402 dhs_chr1_10502 ...
##   dhs_chr22_51228236 dhs_chr22_51234736
## rowData names(0):
## colnames(70): NA18486 NA18498 ... NA19239 NA19257
## colData names(9): naid one ... male isFounder
```

2.4 Genotype data

We take advantage of a function (`gtpath`) that generates paths to S3-resident VCF from the 1000 genomes project.

```
litvcf = readVcf(gtpath(20),
      param=ScanVcfParam(which=GRanges("20",
           IRanges(3.7e7,3.701e7))), genome="hg19")
```

```
litvcf
## class: CollapsedVCF
## dim: 347 2504
## rowRanges(vcf):
##   GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
## info(vcf):
##   DataFrame with 27 columns: CIEND, CIPOS, CS, END, IMPRECISE, MC, MEIN...
## info(header(vcf)):
##                 Number Type    Description
##    CIEND        2      Integer Confidence interval around END for impr...
##    CIPOS        2      Integer Confidence interval around POS for impr...
##    CS           1      String  Source call set.
##    END          1      Integer End coordinate of this variant
##    IMPRECISE    0      Flag    Imprecise structural variation
##    MC           .      String  Merged calls.
##    MEINFO       4      String  Mobile element info of the form NAME,ST...
##    MEND         1      Integer Mitochondrial end coordinate of inserte...
##    MLEN         1      Integer Estimated length of mitochondrial insert
##    MSTART       1      Integer Mitochondrial start coordinate of inser...
##    SVLEN        .      Integer SV length. It is only calculated for st...
##    SVTYPE       1      String  Type of structural variant
##    TSD          1      String  Precise Target Site Duplication for bas...
##    AC           A      Integer Total number of alternate alleles in ca...
##    AF           A      Float   Estimated allele frequency in the range...
##    NS           1      Integer Number of samples with data
##    AN           1      Integer Total number of alleles in called genot...
##    EAS_AF       A      Float   Allele frequency in the EAS populations...
##    EUR_AF       A      Float   Allele frequency in the EUR populations...
##    AFR_AF       A      Float   Allele frequency in the AFR populations...
##    AMR_AF       A      Float   Allele frequency in the AMR populations...
##    SAS_AF       A      Float   Allele frequency in the SAS populations...
##    DP           1      Integer Total read depth; only low coverage dat...
##    AA           1      String  Ancestral Allele. Format: AA|REF|ALT|In...
##    VT           .      String  indicates what type of variant the line...
##    EX_TARGET    0      Flag    indicates whether a variant is within t...
##    MULTI_ALLELIC 0     Flag    indicates whether a site is multi-allelic
## geno(vcf):
##   SimpleList of length 1: GT
## geno(header(vcf)):
##       Number Type   Description
##    GT 1      String Genotype
length(colnames(litvcf))
## [1] 2504
length(intersect(colnames(litvcf), colnames(banovichSE)))
## [1] 52
length(intersect(colnames(litvcf), colnames(geuFPKM)))
## [1] 445
length(intersect(colnames(litvcf), colnames(DHStop5_hg19)))
## [1] 59
```

3 Some computations focused on methylation-expression association

The yriMulti package is currently a scratch-pad for some integrative infrastructure thoughts.

3.1 Gene-centric selection

With mexGR, a GRanges instance is formed with methylation scores for CpG near a gene. The assay data are placed in the mcols, with one range devoted to the expression measures.

```
m1 = mexGR(banovichSE, geuFPKM, symbol="ORMDL3")
m1
## mexGR instance with 44 metadata columns:
## NA18498, NA18499, ..., NA18909, type
## and 7 ranges
mcols(m1)[1:4,1:4]
## DataFrame with 4 rows and 4 columns
##      NA18498    NA18499     NA18502    NA18517
##    <numeric>  <numeric>   <numeric>  <numeric>
## 1 -0.5920581 -0.1775026 -0.00553931 -1.5660101
## 2 -0.4633549  0.2876612  0.55386303 -1.0968065
## 3 -0.2245340 -1.1792216  1.09071357  0.4539331
## 4 -0.1583621 -0.8991156  1.02087023  0.5399715
table(mcols(m1)$type)
##
## expr meth
##    1    6
```

3.2 DNA-methylation association with expression

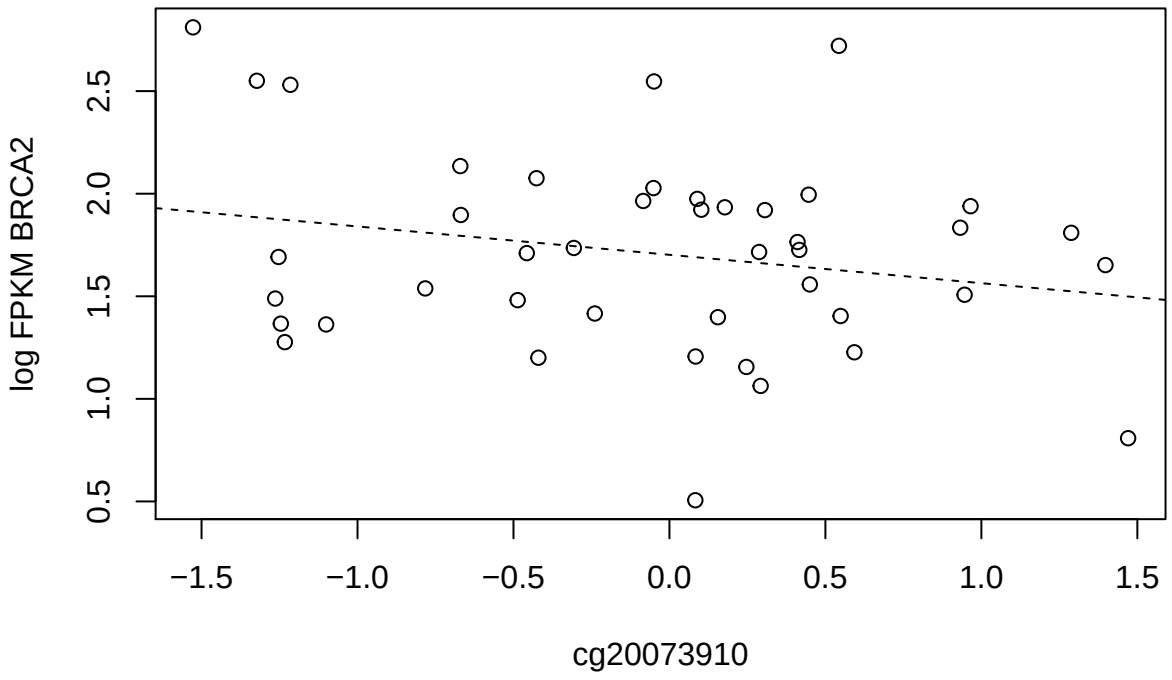
bindelms computes the regressions of the selected gene's expression values on the methylation scores. We have options to transform the expression value (parameter ytx is a function) and can indicate the radius around the gene coding region to search for CpG (parameter gradius).

We'll examine a region around gene BRCA2 for a CpG whose methylation score is negatively associated with BRCA2 expression.

```
b1 = bindelms(geuFPKM, banovichSE, symbol="BRCA2", ytx=log,
   gradius=20000)
b1
## class: RangedSummarizedExperiment
## dim: 29 43
## metadata(6): theCall symbol ... pwd txexpr
## assays(1): betas
## rownames(29): cg00031759 cg00214044 ... cg26458617 cg26941801
## rowData names(15): addressA addressB ... t p
## colnames(43): NA18498 NA18499 ... NA18489 NA18909
## colData names(35): title geo_accession ... data_row_count naid
mcols(b1)[1:3,]
## DataFrame with 3 rows and 15 columns
##      addressA    addressB channel platform percentGC
##   <character> <character>   <Rle>    <Rle> <numeric>
## 1    11602350                Both    HM450      0.50
## 2    33707391                Both    HM450      0.62
```

```
## 3    12675375                Both    HM450      0.54
##                 sourceSeq probeType  probeStart    probeEnd probeTarget
##            <DNAStringSet>     <Rle> <character> <character>   <numeric>
## 1 CGGGTATTTC...GCATCCCAAC        cg    32889486    32889535    32889534
## 2 CGGGCACCAG...ACCCATATTT        cg    32885906    32885955    32885906
## 3 GCCCACCTGA...TTCATTCCCG        cg    32984237    32984286    32984237
##        lms        slope         se           t         p
##     <list>    <numeric>  <numeric>   <numeric> <numeric>
## 1 ########  0.007409067 0.10065080  0.07361161 0.9416774
## 2 ########  0.039618908 0.10276459  0.38553072 0.7018374
## 3 ######## -0.024122378 0.08494654 -0.28397129 0.7778616
summary(mcols(b1)$t)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.4740 -0.3809  0.3426  0.2091  0.5894  1.6640
mintind = which.min(mcols(b1)$t)
mincpg = names(b1)[mintind]
mincpg
## [1] "cg20073910"
```

```
plotEvM(b1)
```



4 Using the biocMultiAssay infrastructure

4.1 Construction of the MultiAssayExperiment

We will use a unified object design to reproduce the BRCA2 display just obtained.

We need a list of relevant objects and a phenodata component.

```
library(biocMultiAssay)
myobs = list(geuvRNAseq=geuFPKM, yri450k=banovichSE, yriDHS=DHStop5_hg19)
cold = colData(geuFPKM)
suppressWarnings({
mm = MultiAssayExperiment(myobs, as.data.frame(cold))
})
mm
```

4.2 Restriction by range

We compute the BRCA2 'gene range'.

```
library(erma)
brr = range(genemodel("BRCA2"))
brr
```

Subset the multiassay structure to features in the vicinity of this range.

```
.subsetByRanges = function(ma, r) {
  el2 = lapply(elist(ma)@listData, function(x)subsetByOverlaps(x,r))
  names(el2) = names(elist(ma)) # needed?
  ma@elist = new("elist", listData=el2)
  ma
}
newmm = .subsetByRanges(mm, brr+20000)
newmm
```

We now have all the relevant features and samples. In fact we have more genes than we really wanted. But we will proceed with this selection.

4.3 All pairwise regressions

We will introduce a formula idiom to specify a collection of models of interest. `allLM_pw` is defined in *biocMultiAssay*.

```
pp = allLM_pw(geuvRNAseq~yri450k, newmm, ytx=log)
names(pp)
summary(pp[[1]][[1]])
which.min(unlist(pp[[2]])) # not BRCA2 but FRY
```

The formula idiom can be used to isolate assays and features.

```
pwplot(geuvRNAseq~yri450k, ENSG00000139618.9~cg20073910, newmm, ytx=log)
```

4.4 Integrating dense genotypes in a VcfStack instance

We set up a reference to a collection of VCF. We'll use 1000 genomes VCF for chr21, chr22, and chrY.

```
library(gQTLstats)
pa = paths1kg(paste0("chr", c(21:22,"Y")))
```

```
library(Homo.sapiens)  # necessary?
stopifnot(requireNamespace("GenomeInfoDb"))
ob = VcfStack(pa, seqinfo(Homo.sapiens))
ob
```

Now we set up a region of interest and bind it to the stack. This yields an instance of `RangedVcfStack`, for which we have `samples`, `features`, and `assay` methods defined in gQTLstats.

```
myr = GRanges("22", IRanges(20e6,20.01e6))
ob = bindRanges(ob, myr)
hasInternetConnectivity = function()
 !is.null(nsl("www.r-project.org"))
if (hasInternetConnectivity()) lka = assay(ob)
myobs = list(geuvRNAseq=geuFPKM, yri450k=banovichSE, yriDHS=DHStop5_hg19,
    yriGeno=ob)
cold = colData(geuFPKM)
suppressWarnings({
mm = MultiAssayExperiment(myobs, as.data.frame(cold))
})
mm
```

References

[1] N. E. Banovich, X. Lan, G. McVicker, et al. "Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels". In: *PLoS Genetics* 10.9 (2014), p. e1004663. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004663. .