Loci2path: regulatory annotation of genomic intervals based on tissue-specific expression QTLs

Tianlei Xu

29 August 2017

Abstract

Annotating a given genomic locus or a set of genomic loci is an important yet challenging task. This is especially true for the non-coding part of the genome which is enormous yet poorly understood. Since gene set enrichment analyses have demonstrated to be effective approach to annotate a set of genes, this idea can be extended to explore the enrichment of functional elements or features in a set of genomic intervals to reveal potential functional connections. In this study, we describe a novel computational strategy that takes advantage of the newly emerged, genome-wide and tissue-specific expression quantitative trait loci (eQTL) information to help annotate a set of genomic intervals in terms of transcription regulation. By checking the presence or absence of millions of eQTLs in the set of genomic intervals of interest, loci2path build a bridge connecting genomic intervals to biological pathway or pre-defined biological-meaningful gene sets. Our method enjoys two key advantages over existing methods: first, we no longer rely on proximity to link a locus to a gene which has shown to be unreliable; second, eQTL allows us to provide the regulatory annotation under the context of specific tissue types which is important.

Package version: loci2path 0.1.0

Contents

1 Prepare input dataset for query

1.1 Query regions

`loci2path` takes query regions in the format of `GenomicRanges`. Only the Genomic Locations (chromosomes, start and end position) will be used. Strand information and other metadata columns are ignored. In the demo data, 47 regions associated with Psoriasis disease were downloaded from **immunoBase.org** and used as demo query regions.

```
require(GenomicRanges)
bed.file=system.file("extdata", "query/Psoriasis.BED", package = "loci2path")
query.bed=read.table(bed.file, header=FALSE)
colnames(query.bed)=c("chr","start","end")
query.gr=makeGRangesFromDataFrame(query.bed)
```

1.2 Prepare eQTL sets.

eQTL sets are entities recording 1-to-1 links between eQTL SNPs and genes. eQTL set entity also contains the following information: tissue name for the eQTL study, IDs and genomic ranges for the eQTL SNPs, IDs for the associated genes.

eQTL set can be constructed manually by specifying the corresponding information in each slot.

eQTL set list is a list of multiple eQTL sets, usually collected from different tissues.

Below is an example to construct customized eQTL set and eQTL set list using demo data files. In the demo data folder, three eQTL sets downloaded from GTEx project are included. Due to the large size, each eQTL dataset is down sampled to 3000 records for demostration purpose.

1.2.1 construct eQTL set

```
library(loci2path)
brain.file=system.file("extdata", "eqtl/brain.gtex.txt",
                       package = "loci2path")
tab=read.table(brain.file, stringsAsFactors = FALSE, header = TRUE)
snp.gr=GRanges(seqnames=Rle(tab$snp.chr),
  ranges=IRanges(start=tab$snp.pos,
  width=1))
brain.eset=eqtlSet(tissue="brain",
  snp.id=tab$snp.id,
  snp.gr=snp.gr,
  gene=as.character(tab$gene.entrez.id))
brain.eset
## An object of class eqtlSet
##  eQTL collected from tissue: brain
##  number of eQTLs: 3000
##  number of associated genes: 815

skin.file=system.file("extdata", "eqtl/skin.gtex.txt", package = "loci2path")
tab=read.table(skin.file, stringsAsFactors = FALSE, header = TRUE)
snp.gr=GRanges(seqnames=Rle(tab$snp.chr),
  ranges=IRanges(start=tab$snp.pos,
  width=1))
skin.eset=eqtlSet(tissue="skin",
  snp.id=tab$snp.id,
  snp.gr=snp.gr,
  gene=as.character(tab$gene.entrez.id))
skin.eset
```

```
## An object of class eqtlSet
##  eQTL collected from tissue: skin
##  number of eQTLs: 3000
##  number of associated genes: 1588

blood.file=system.file("extdata", "eqtl/blood.gtex.txt",
                       package = "loci2path")
tab=read.table(blood.file, stringsAsFactors = FALSE, header = TRUE)
snp.gr=GRanges(seqnames=Rle(tab$snp.chr),
  ranges=IRanges(start=tab$snp.pos,
  width=1))
blood.eset=eqtlSet(tissue="blood",
  snp.id=tab$snp.id,
  snp.gr=snp.gr,
  gene=as.character(tab$gene.entrez.id))
blood.eset
## An object of class eqtlSet
##  eQTL collected from tissue: blood
##  number of eQTLs: 3000
##  number of associated genes: 1419
```

1.2.2 construct eQTL set list

```
eset.list=list(Brain=brain.eset, Skin=skin.eset, Blood=blood.eset)
eset.list
## $Brain
## An object of class eqtlSet
##  eQTL collected from tissue: brain
##  number of eQTLs: 3000
##  number of associated genes: 815
##
## $Skin
## An object of class eqtlSet
##  eQTL collected from tissue: skin
##  number of eQTLs: 3000
##  number of associated genes: 1588
##
## $Blood
## An object of class eqtlSet
##  eQTL collected from tissue: blood
##  number of eQTLs: 3000
##  number of associated genes: 1419
```

1.3 Prepare gene set collection

A geneset collection contains a list of gene sets, with each gene set is represented as a vector of member genes. A vector of description is also provided as the metadata slot for each gene set. The total number of gene in the geneset collection is also required to perform the enrichment test. In this tutorial the BIOCARTA pathway collection was downloaded from MSigDB.

```
biocarta.link.file=system.file("extdata", "geneSet/biocarta.txt",
                               package = "loci2path")
biocarta.set.file=system.file("extdata", "geneSet/biocarta.set.txt",
                              package = "loci2path")

biocarta.link=read.delim(biocarta.link.file, header = FALSE,
                         stringsAsFactors = FALSE)
set.geneid=read.table(biocarta.set.file, stringsAsFactors = FALSE)
set.geneid=strsplit(set.geneid[,1], split=",")
names(set.geneid)=biocarta.link[,1]

head(biocarta.link)
##                               V1
## 1         BIOCARTA_RELA_PATHWAY
## 2          BIOCARTA_NO1_PATHWAY
## 3          BIOCARTA_CSK_PATHWAY
## 4      BIOCARTA_SRCRPTP_PATHWAY
## 5          BIOCARTA_AMI_PATHWAY
## 6 BIOCARTA_GRANULOCYTES_PATHWAY
##                                                                        V2
## 1         http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_RELA_PATHWAY
## 2          http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_NO1_PATHWAY
## 3          http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_CSK_PATHWAY
## 4      http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_SRCRPTP_PATHWAY
## 5          http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_AMI_PATHWAY
## 6 http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_GRANULOCYTES_PATHWAY
head(set.geneid)
## $BIOCARTA_RELA_PATHWAY
##  [1] "8517" "1147" "2033" "5970" "7124" "3551" "7133" "8841" "7132" "7189"
## [11] "8772" "1387" "8737" "4790" "4792" "8717"
##
## $BIOCARTA_NO1_PATHWAY
##  [1] "5140"   "805"    "58"     "124827" "801"    "5577"   "3827"
##  [8] "6262"   "1128"   "7422"   "3320"   "6541"   "5139"   "5138"
## [15] "624"    "147908" "121916" "4846"   "1134"   "2321"   "3791"
## [22] "5567"   "7135"   "5568"   "2324"   "857"    "207"    "5573"
## [29] "5576"   "5575"   "808"    "5592"   "5593"
##
## $BIOCARTA_CSK_PATHWAY
##  [1] "7535" "1445" "920"  "5577" "5567" "915"  "5568" "916"  "917"  "2778"
## [11] "2792" "6957" "2782" "6955" "5573" "5576" "919"  "1387" "5575" "107"
## [21] "3932" "5788" "3123" "3122"
##
## $BIOCARTA_SRCRPTP_PATHWAY
##  [1] "1445" "6714" "994"  "995"  "5579" "2885" "993"  "5578" "891"  "5786"
## [11] "983"
##
## $BIOCARTA_AMI_PATHWAY
##  [1] "2159"  "7035"  "2147"  "1282"  "2149"  "1284"  "1285"  "1286"
##  [9] "5627"  "2266"  "5624"  "2243"  "5340"  "462"   "2244"  "5327"
## [17] "1288"  "51327" "1287"  "2155"
##
## $BIOCARTA_GRANULOCYTES_PATHWAY
```

```
##  [1] "5175" "7124" "3552" "3683" "3684" "3383" "6402" "6403" "3458" "6404"
## [11] "727"  "3689" "1440" "3576"
```

In order to build gene set, we also need to know the total number of genes in order to perform enrichment test. In this study, the total number of gene in MSigDB pathway collection is 31,847(Liberzon et al. 2015)

```
#build geneSet
biocarta=geneSet(
  gene.set=set.geneid,
  description=biocarta.link[,2],
  total.number.gene=31847)
biocarta
## An object of class geneSet
##  Number of gene sets: 217
##     6 ~ 87  genes within sets
```

2 Perform query

2.1 peroform query from one eQTL set

```
#query from one eQTL set.
res.one=query.egset(
  query.gr=query.gr,
  query.score=NULL,
  eqtl.set=skin.eset,
  gene.set=biocarta)

#enrichment result table
res.one$result.table
##                      name_pthw eQTL_pthw eQTL_total_tissue eQTL_query
## V41  BIOCARTA_CLASSIC_PATHWAY        14              3000         78
## V42     BIOCARTA_COMP_PATHWAY        14              3000         78
## V111  BIOCARTA_LECTIN_PATHWAY        14              3000         78
##      eQTL_pthw_query log_ratio pval_lr pval_fisher num_gene_set
## V41                2  1.703749      NA  0.04961954           14
## V42                2  1.703749      NA  0.04961954           19
## V111               2  1.703749      NA  0.04961954           12
##      num_gene_query num_gene_hit gene_hit log_ratio_gene pval_fisher_gene
## V41              29            1      721       4.362345       0.01267584
## V42              29            1      721       4.056964       0.01716522
## V111             29            1      721       4.516496       0.01087455

#all the genes associated with eQTLs covered by the query region
res.one$cover.gene
##  [1] "100129271" "353134"    "353144"    "353135"    "130872"
##  [6] "11127"     "64167"     "3106"      "253018"    "285834"
## [11] "5460"      "170679"    "3107"      "100130889" "100507436"
## [16] "721"       "4277"      "23586"     "10330"     "283635"
## [21] "80270"     "84148"     "29108"     "201229"    "27175"
## [26] "4669"      "11201"     "57153"     "9825"
```

2.2 peroform query from multiple eQTL sets

```
#query from one eQTL set.
res.esetlist=query.egset.list(
  query.gr=query.gr,
  query.score=NULL,
  eqtl.set.list=eset.list,
  gene.set=biocarta)
## Start query: 3 eqtl Sets...
## 1 of 3: Brain...
## 2 of 3: Skin...
## 3 of 3: Blood...
##
## done!

#enrichment result table, tissue column added
res.esetlist$result.table
##   tissue                name_pthw eQTL_pthw eQTL_total_tissue eQTL_query
## 1  Blood  BIOCARTA_LECTIN_PATHWAY        24              3000         60
## 2  Blood BIOCARTA_CLASSIC_PATHWAY        24              3000         60
## 3  Blood    BIOCARTA_COMP_PATHWAY        24              3000         60
## 4   Skin  BIOCARTA_LECTIN_PATHWAY        14              3000         78
## 5   Skin BIOCARTA_CLASSIC_PATHWAY        14              3000         78
## 6   Skin    BIOCARTA_COMP_PATHWAY        14              3000         78
## 7  Blood      BIOCARTA_DC_PATHWAY         4              3000         60
## 8  Blood BIOCARTA_CHREBP2_PATHWAY         7              3000         60
##   eQTL_pthw_query log_ratio pval_lr pval_fisher num_gene_set
## 1               2  1.427116      NA  0.08196929           12
## 2               2  1.427116      NA  0.08196929           14
## 3               2  1.427116      NA  0.08196929           19
## 4               2  1.703749      NA  0.04961954           12
## 5               2  1.703749      NA  0.04961954           14
## 6               2  1.703749      NA  0.04961954           19
## 7               1  2.525729      NA  0.07766952           22
## 8               1  1.966113      NA  0.13199866           42
##   num_gene_query num_gene_hit gene_hit log_ratio_gene pval_fisher_gene
## 1             29            2  720;721       5.209643     5.254381e-05
## 2             29            2  720;721       5.055492     7.236494e-05
## 3             29            2  720;721       4.750111     1.355988e-04
## 4             29            1      721       4.516496     1.087455e-02
## 5             29            1      721       4.362345     1.267584e-02
## 6             29            1      721       4.056964     1.716522e-02
## 7             29            1     3687       3.910360     1.984938e-02
## 8             29            1     6945       3.263733     3.756375e-02

#all the genes associated with eQTLs covered by the query region;
#names of the list are tissue names from eqtl set list
res.esetlist$cover.gene
## $Brain
##  [1] "84542"     "130872"    "64167"     "3107"      "100507436"
##  [6] "55012"     "116028"    "9810"      "84148"     "27175"
## [11] "11201"     "164592"
##
## $Skin
```

```
##  [1] "100129271" "353134"    "353144"    "353135"    "130872"
##  [6] "11127"     "64167"     "3106"      "253018"    "285834"
## [11] "5460"      "170679"    "3107"      "100130889" "100507436"
## [16] "721"       "4277"      "23586"     "10330"     "283635"
## [21] "80270"     "84148"     "29108"     "201229"    "27175"
## [26] "4669"      "11201"     "57153"     "9825"
##
## $Blood
##  [1] "130872"    "6584"      "51752"     "64167"     "100130889"
##  [6] "720"       "3106"      "5460"      "3107"      "4277"
## [11] "253018"    "100507436" "1590"      "721"       "6821"
## [16] "6231"      "280655"    "283635"    "79759"     "80270"
## [21] "3687"      "9810"      "3965"      "2548"      "2145"
## [26] "6945"      "10053"     "57153"     "147727"
```

2.3 parallel query from multiple eQTL sets

```
#query from one eQTL set.
res.paral = query.egset.list(
  query.gr = query.gr,
  query.score = NULL,
  eqtl.set.list = eset.list,
  gene.set = biocarta,
  parallel = TRUE)
## Start query: 3 eqtl Sets...
## Run in parallel mode...
##
## done!
#should return the same result as res.esetlist
```

3 explore query result

```
result=res.esetlist$result.table
```

3.1 obtain eQTL gene list

```
#all the genes associated with eQTLs covered by the query region
res.one$cover.gene
##  [1] "100129271" "353134"    "353144"    "353135"    "130872"
##  [6] "11127"     "64167"     "3106"      "253018"    "285834"
## [11] "5460"      "170679"    "3107"      "100130889" "100507436"
## [16] "721"       "4277"      "23586"     "10330"     "283635"
## [21] "80270"     "84148"     "29108"     "201229"    "27175"
## [26] "4669"      "11201"     "57153"     "9825"

#all the genes associated with eQTLs covered by the query region;
#names of the list are tissue names from eqtl set list
res.esetlist$cover.gene
```

```
## $Brain
##  [1] "84542"     "130872"    "64167"     "3107"      "100507436"
##  [6] "55012"     "116028"    "9810"      "84148"     "27175"
## [11] "11201"     "164592"
##
## $Skin
##  [1] "100129271" "353134"    "353144"    "353135"    "130872"
##  [6] "11127"     "64167"     "3106"      "253018"    "285834"
## [11] "5460"      "170679"    "3107"      "100130889" "100507436"
## [16] "721"       "4277"      "23586"     "10330"     "283635"
## [21] "80270"     "84148"     "29108"     "201229"    "27175"
## [26] "4669"      "11201"     "57153"     "9825"
##
## $Blood
##  [1] "130872"    "6584"      "51752"     "64167"     "100130889"
##  [6] "720"       "3106"      "5460"      "3107"      "4277"
## [11] "253018"    "100507436" "1590"      "721"       "6821"
## [16] "6231"      "280655"    "283635"    "79759"     "80270"
## [21] "3687"      "9810"      "3965"      "2548"      "2145"
## [26] "6945"      "10053"     "57153"     "147727"
```

3.2 obtain average tissue degree for each pathway

```
tissue.degree=res.get.tissue.degree(
  result,
  eset.list)

#check gene-tissue mapping for each gene
head(tissue.degree$gene.tissue.map)
## $`100101267`
## [1] "Brain"
##
## $`100125556`
## [1] "Brain" "Skin"  "Blood"
##
## $`100128081`
## [1] "Brain"
##
## $`100129583`
## [1] "Brain"
##
## $`100130418`
## [1] "Brain" "Skin"
##
## $`100130958`
## [1] "Brain" "Skin"  "Blood"

#check degree for each gene
head(tissue.degree$gene.tissue.degree)
## 100101267 100125556 100128081 100129583 100130418 100130958
##         1         3         1         1         2         3
```

```
#average tissue degree for the input result table
tissue.degree$mean.tissue.degree
## [1] 2 2 2 2 2 2 1 1

#add avg. tissue degree to existing result table
res.tissue=data.frame(res.esetlist$result.table,
                      t.degree=tissue.degree$mean.tissue.degree)
```

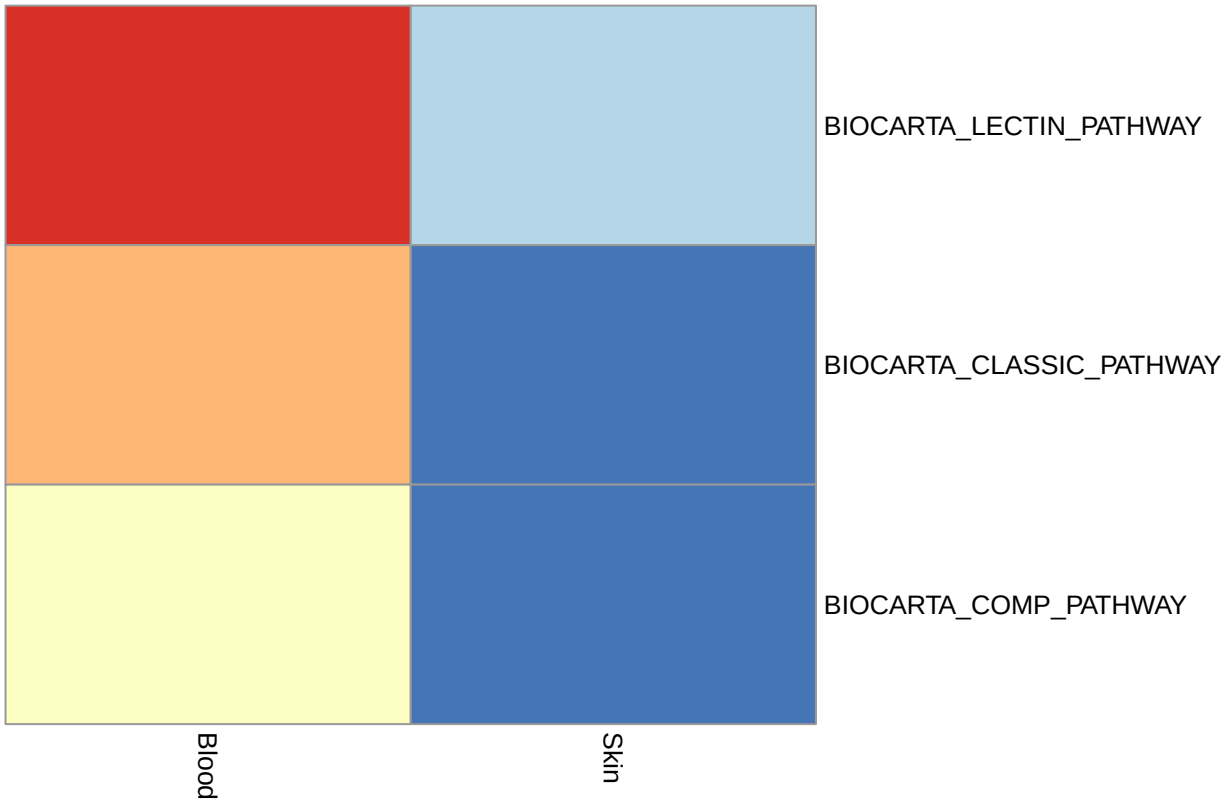
3.3 obtain tissue enrichment for query regions

```
#query tissue specificity
gr.tissue=query.tissue(query.gr, eqtl.set.list=eset.list)
gr.tissue
##       eQTL_gene_in_tissue eQTL_gene_in_query         pval         padj
## Blood                1419                 29 8.922769e-14 2.676831e-13
## Skin                 1588                 29 1.432947e-12 2.865894e-12
## Brain                 815                 12 1.653003e-05 1.653003e-05
```

3.4 extract tissue-pathway heatmap

```
#extract tissue-pathway matrix
mat=res.get.heat.mat(result, test.method = "fisher")

#plot heatmap
draw.heatmap(mat)
```



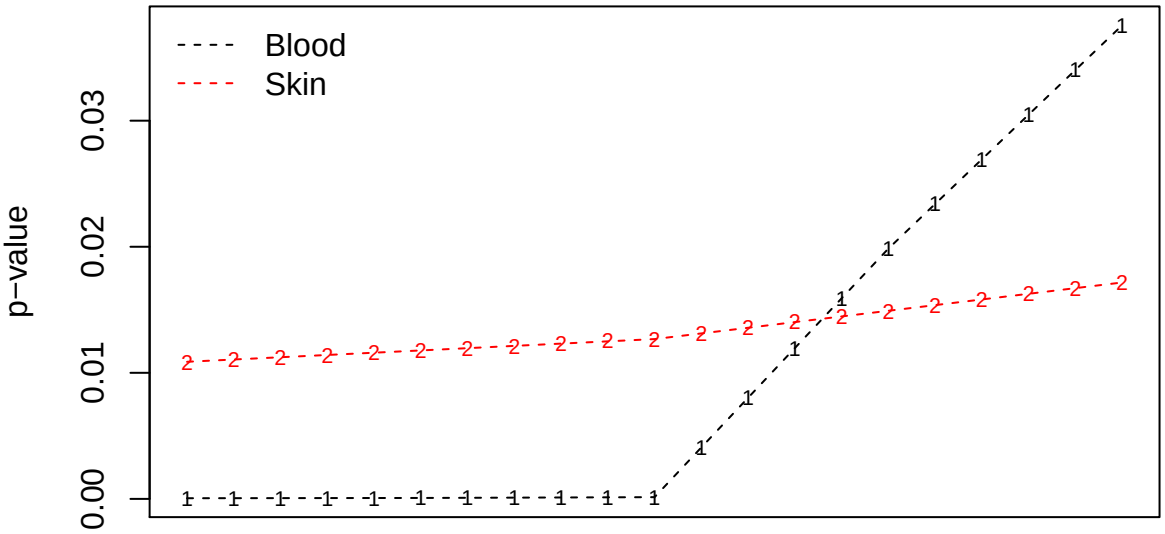

```
## $tree_row
## [1] NA
##
## $tree_col
## [1] NA
##
## $kmeans
## [1] NA
##
## $gtable
## TableGrob (5 x 6) "layout": 4 grobs
##   z     cells      name                grob
## 1 1 (1-1,3-3)      main   text[GRID.text.4]
## 2 2 (4-4,3-3)    matrix gTree[GRID.gTree.6]
## 3 3 (5-5,3-3) col_names   text[GRID.text.7]
## 4 4 (4-4,4-4) row_names   text[GRID.text.8]
```

3.5 extract word cloud from result

```
#plot word cloud
draw.wordcloud(result)
```

3.6 plot p-value distribution of result

```
#plot p-value distribution of result
draw.pval.distribution(result, test.method="fisher")
```



3.7 obtain geneset description from object

```
#obtain geneset description from object
description=get.geneset.description(biocarta, geneset.ids=result$name_pthw)
head(description)
## [1] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_LECTIN_PATHWAY"
## [2] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_CLASSIC_PATHWAY"
## [3] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_COMP_PATHWAY"
## [4] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_LECTIN_PATHWAY"
## [5] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_CLASSIC_PATHWAY"
## [6] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_COMP_PATHWAY"
```

4 Session info

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
```

```
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] loci2path_0.1.0      BiocStyle_2.4.1      pheatmap_1.0.8
##  [4] devtools_1.13.3      BiocParallel_1.10.1  GenomicRanges_1.28.4
##  [7] GenomeInfoDb_1.12.2  IRanges_2.10.2       S4Vectors_0.14.3
## [10] BiocGenerics_0.22.0  BiocCheck_1.12.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12            BiocInstaller_1.26.0
##  [3] compiler_3.4.1          RColorBrewer_1.1-2
##  [5] plyr_1.8.4              XVector_0.16.0
##  [7] iterators_1.0.8         bitops_1.0-6
##  [9] tools_3.4.1             zlibbioc_1.22.0
## [11] digest_0.6.12           evaluate_0.10.1
## [13] memoise_1.1.0           gtable_0.2.0
## [15] foreach_1.4.3           graph_1.54.0
## [17] yaml_2.1.14             GenomeInfoDbData_0.99.0
## [19] stringr_1.2.0           withr_2.0.0
## [21] httr_1.3.0              knitr_1.17
## [23] wordcloud_2.5           rprojroot_1.2
## [25] grid_3.4.1              getopt_1.20.0
## [27] data.table_1.10.4       optparse_1.4.4
## [29] Biobase_2.36.2          R6_2.2.2
## [31] XML_3.98-1.9            RBGL_1.52.0
## [33] rmarkdown_1.6           magrittr_1.5
## [35] splines_3.4.1           backports_1.1.0
## [37] scales_0.4.1            codetools_0.2-15
## [39] htmltools_0.3.6         biocViews_1.44.0
## [41] RUnit_0.4.31            colorspace_1.3-2
## [43] stringi_1.1.5           RCurl_1.95-4.8
## [45] munsell_0.4.3           slam_0.1-40
## [47] gam_1.14-4
```

References

Liberzon, Arthur, Chet Birger, Helga Thorvaldsd??ttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." *Cell Systems* 1 (6): 417–25. doi:10.1016/j.cels.2015.12.004.