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

Tianlei Xu
Zhaohui Qin

19 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.

```
bed.file=system.file("extdata", "query/Psoriasis.BED", package = "loci2path")
query.bed=read.table(bed.file, header=F)
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

```
brain.file=system.file("extdata", "eqtl/brain.gtex.txt", package = "loci2path")
tab=read.table(brain.file, stringsAsFactors = F, header = T)
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 = F, header = T)
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 = F, header = T)
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 = F, stringsAsFactors = F)
set.geneid=read.table(biocarta.set.file, stringsAsFactors = F)
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

3 explore query result

3.1 extract tissue-pathway heatmap

3.2 extract word cloud from result

3.3 obtain eQTL gene list

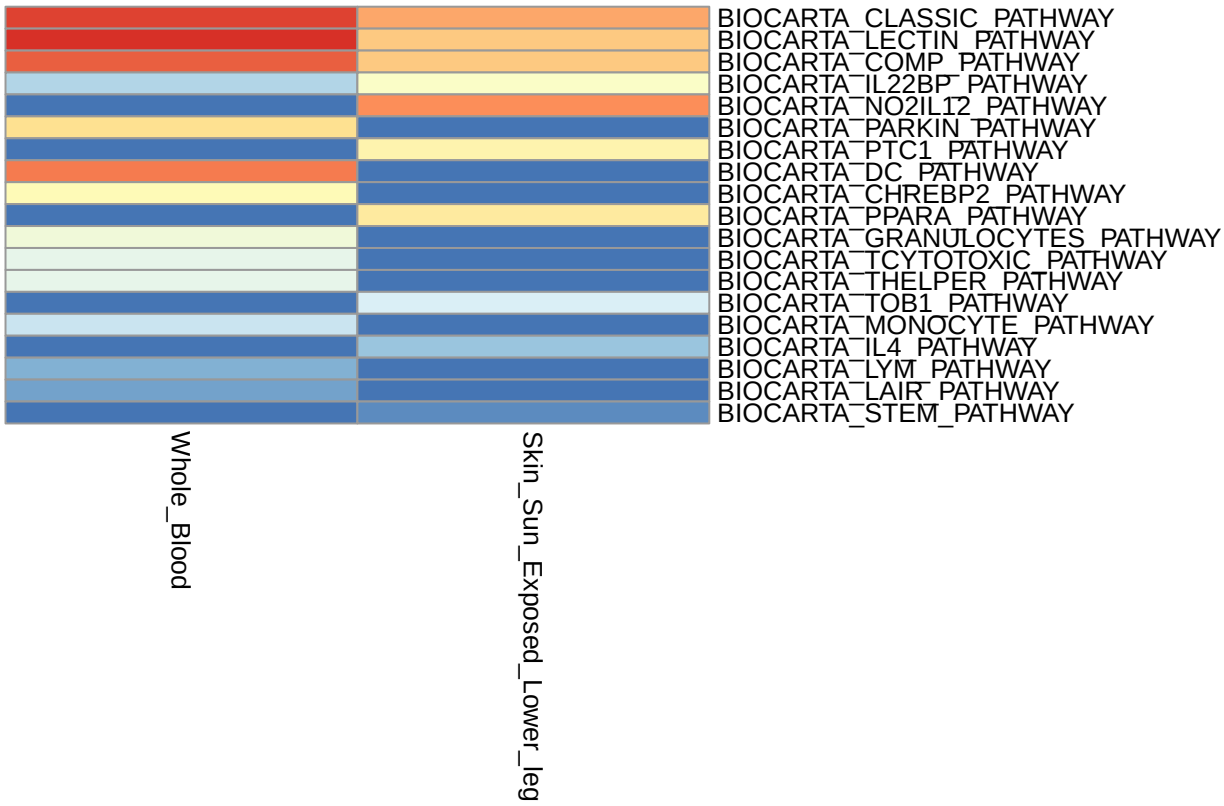
3.4 obtain average tissue degree for each pathway

3.5 obtain tissue enrichment for query regions

```
## GRanges object with 47 ranges and 0 metadata columns:
##        seqnames                ranges strand
##           <Rle>             <IRanges>  <Rle>
##    [1]     chr1 [  8200690,   8306031]      *
##    [2]     chr1 [152536784, 152785813]      *
##    [3]     chr1 [ 24461438,  24527816]      *
##    [4]     chr1 [ 67594559,  67767993]      *
##    [5]     chr1 [ 25224957,  25308276]      *
##    ...      ...                   ...    ...
##   [43]    chr18   [52210075, 52409477]      *
##   [44]    chr19   [10634264, 11164781]      *
##   [45]    chr19   [10390709, 10628548]      *
##   [46]    chr20   [48408615, 48662582]      *
##   [47]    chr22   [21809185, 22003928]      *
##   -------
##   seqinfo: 16 sequences from an unspecified genome; no seqlengths
## $Brain_Cortex
## An object of class eqtlSet
##  eQTL collected from tissue: Brain_Cortex
##  number of eQTLs: 131424
##  number of associated genes: 1796
##
## $Skin_Sun_Exposed_Lower_leg
## An object of class eqtlSet
##  eQTL collected from tissue: Skin_Sun_Exposed_Lower_leg
##  number of eQTLs: 712745
##  number of associated genes: 6171
##
## $Whole_Blood
## An object of class eqtlSet
##  eQTL collected from tissue: Whole_Blood
##  number of eQTLs: 594632
##  number of associated genes: 5073
## An object of class geneSet
```

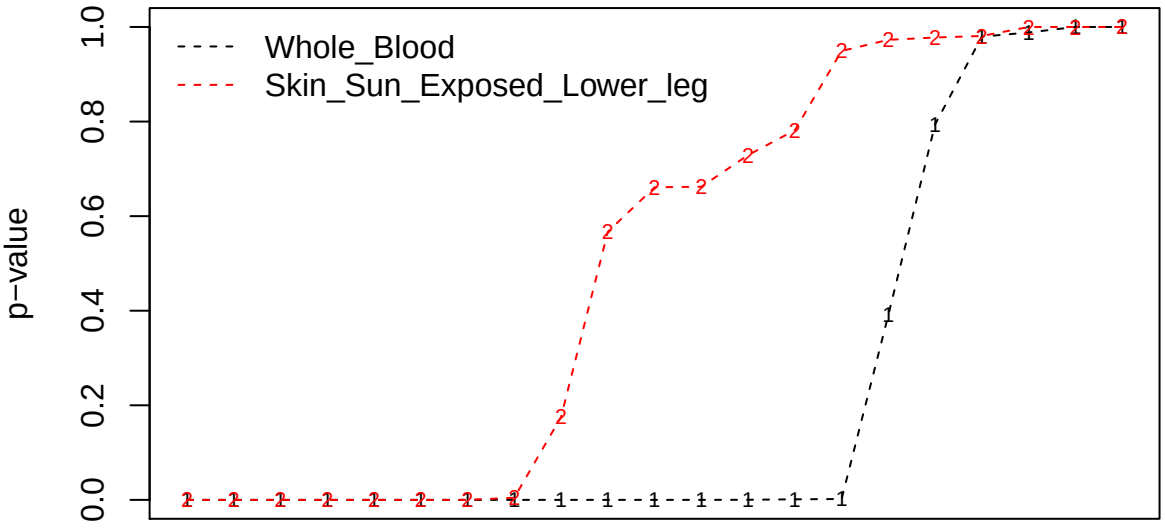
```
##  Number of gene sets: 217
##     6 ~ 87  genes within sets
## Start query: 3 eqtl Sets...
## 1 of 3: Brain_Cortex...
## 2 of 3: Skin_Sun_Exposed_Lower_leg...
## 3 of 3: Whole_Blood...
##
## done!
##                       tissue                     name_pthw eQTL_pthw
## 1                Whole_Blood     BIOCARTA_MONOCYTE_PATHWAY       157
## 2                Whole_Blood       BIOCARTA_LECTIN_PATHWAY      3227
## 3                Whole_Blood BIOCARTA_GRANULOCYTES_PATHWAY        96
## 4                Whole_Blood      BIOCARTA_CLASSIC_PATHWAY      3233
## 5 Skin_Sun_Exposed_Lower_leg       BIOCARTA_LECTIN_PATHWAY      2859
## 6                Whole_Blood         BIOCARTA_COMP_PATHWAY      3325
##   eQTL_total_tissue eQTL_query eQTL_pthw_query log_ratio pval_lr
## 1            594632      11943              14  1.490609      NA
## 2            594632      11943             538  2.116348      NA
## 3            594632      11943              14  1.982507      NA
## 4            594632      11943             538  2.114490      NA
## 5            712745      17062             262  1.342387      NA
## 6            594632      11943             538  2.086431      NA
##     pval_fisher num_gene_set num_gene_query num_gene_hit  gene_hit
## 1  4.108902e-06           11             77            2 3684;3383
## 2 8.096423e-312           12             77            2   721;720
## 3  8.812780e-09           14             77            2 3684;3383
## 4 2.149518e-311           14             77            2   721;720
## 5  2.372477e-74           12            108            2   721;720
## 6 5.197609e-305           19             77            2   721;720
##   log_ratio_gene pval_fisher_gene
## 1       4.320145     0.0003129003
## 2       4.233134     0.0003748918
## 3       4.078983     0.0005152773
## 4       4.078983     0.0005152773
## 5       3.894808     0.0007355142
## 6       3.773601     0.0009607091
## Start query: 3 eqtl Sets...
## Run in parallel mode...
##
## done!
##                       name_pthw eQTL_pthw eQTL_total_tissue eQTL_query
## 1     BIOCARTA_MONOCYTE_PATHWAY       157            594632      11943
## 2       BIOCARTA_LECTIN_PATHWAY      3227            594632      11943
## 3 BIOCARTA_GRANULOCYTES_PATHWAY        96            594632      11943
## 4      BIOCARTA_CLASSIC_PATHWAY      3233            594632      11943
## 5       BIOCARTA_LECTIN_PATHWAY      2859            712745      17062
## 6         BIOCARTA_COMP_PATHWAY      3325            594632      11943
##   eQTL_pthw_query log_ratio pval_lr   pval_fisher num_gene_set
## 1              14  1.490609      NA  4.108902e-06           11
## 2             538  2.116348      NA 8.096423e-312           12
## 3              14  1.982507      NA  8.812780e-09           14
## 4             538  2.114490      NA 2.149518e-311           14
## 5             262  1.342387      NA  2.372477e-74           12
## 6             538  2.086431      NA 5.197609e-305           19
```

```
##   num_gene_query num_gene_hit  gene_hit log_ratio_gene pval_fisher_gene
## 1             77            2 3684;3383       4.320145     0.0003129003
## 2             77            2   721;720       4.233134     0.0003748918
## 3             77            2 3684;3383       4.078983     0.0005152773
## 4             77            2   721;720       4.078983     0.0005152773
## 5            108            2   721;720       3.894808     0.0007355142
## 6             77            2   721;720       3.773601     0.0009607091
```




```
## Warning in wordcloud(words = names(pthw), freq = pthw, min.freq =
## min.freq.gset, : BIOCARTA_TCYTOTOXIC_PATHWAY could not be fit on page. It
## will not be plotted.
## Warning in wordcloud(words = names(pthw), freq = pthw, min.freq =
## min.freq.gset, : BIOCARTA_TH1TH2_PATHWAY could not be fit on page. It will
## not be plotted.
## Warning in wordcloud(words = names(pthw), freq = pthw, min.freq =
## min.freq.gset, : BIOCARTA_THELPER_PATHWAY could not be fit on page. It will
## not be plotted.
## Warning in wordcloud(words = names(pthw), freq = pthw, min.freq =
## min.freq.gset, : BIOCARTA_TNFR2_PATHWAY could not be fit on page. It will
## not be plotted.
## Warning in wordcloud(words = names(pthw), freq = pthw, min.freq =
## min.freq.gset, : BIOCARTA_TOB1_PATHWAY could not be fit on page. It will
## not be plotted.
```


BIOCARTA_PTC1_PATHWAY
BIOCARTA_PPARA_PATHWAY
BIOCARTA_NKT_PATHWAY
BIOCARTA_LYM_PATHWAY
BIOCARTA_INFLAM_PATHWAY
BIOCARTA_IL4_PATHWAY
BIOCARTA_G2_PATHWAY
BIOCARTA_CTL_PATHWAY
BIOCARTA_CD40_PATHWAY
BIOCARTA_BLYMPHOCYTE_PATHWAY
BIOCARTA_ARENRF2_PATHWAY
BIOCARTA_NO2IL12_PATHWAY
BIOCARTA_IL22BP_PATHWAY
BIOCARTA_DC_PATHWAY
BIOCARTA_COMP_PATHWAY
BIOCARTA_CHREBP2_PATHWAY
BIOCARTA_CLASSIC_PATHWAY
BIOCARTA_IL12_PATHWAY
BIOCARTA_IL6_PATHWAY
BIOCARTA_LECTIN_PATHWAY
BIOCARTA_41BB_PATHWAY
BIOCARTA_ASBCELL_PATHWAY
BIOCARTA_CELLCYCLE_PATHWAY
BIOCARTA_CYTOKINE_PATHWAY
BIOCARTA_GATA3_PATHWAY
BIOCARTA_GRANULOCYTES_PATHWAY
BIOCARTA_IL5_PATHWAY
BIOCARTA_LAIR_PATHWAY
BIOCARTA_MONOCYTE_PATHWAY
BIOCARTA_PARKIN_PATHWAY
BIOCARTA_STEM_PATHWAY
Whole_Blood
Skin_Sun_Exposed_Lower_leg


Whole_Blood
Skin_Sun_Exposed_Lower_leg
p−value
0.0
0.2
0.4
0.6
0.8
1.0

```
## [1] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_MONOCYTE_PATHWAY"
## [2] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_LECTIN_PATHWAY"
## [3] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_GRANULOCYTES_PATHWAY"
## [4] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_CLASSIC_PATHWAY"
## [5] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_LECTIN_PATHWAY"
## [6] "http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_COMP_PATHWAY"
```

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.