Network-based pathway over-representation analysis with NetORA

Jiantao Shi

30 November 2017

Abstract

We have developed a R package NetGAP for network-based gene prioritization using text-mining and coexpression networks. NetORA is an extension of this package to perform network-based pathway enrichment analysis.

Package version: NetORA 0.99.0

Contents

1 Intruduction

Genome wide association studies (GWAS) have been successfully used to identify disease-associated variants, however the causal genes in many diseases remain elusive, due to effects such as linkage disequilibrium (LD) between associated variants and long-range regulation. Direct experimental validation of the many potential causal genes is expensive and difficult, so an attractive first step is to prioritize genes with respect their biological relevance. Numerous evidences suggest genes function in pathway-level, so different disease causal genes might function in the same causal pathway. Pathways can be annotated as un-structured gene sets (Reactome), structured tree (GO) or even networks. We have implemented the algorithm of GRAIL (Raychaudhuri 2009) in R as a package NetGPA and provide both text-mining and coexpression networks. We have demonstrated coexpression network-based gene prioritization is more sensitive when gene expression signatures are provided as seeds (Aldo M. Roccaro and Ghobrial 2016). NetORA is an extension of NetGPA to perform network-based pathway enrichment analysis.

2 Standard workflow

To demonstrate the input and output data format in NetORA, we use example data sets in NetGPA.

```
library("NetGPA")
library("NetORA")
data("Example_NetGPA")

names(Example_NetGPA)
## [1] "CD_GWAS"        "WM_Seed"        "WM_Query"       "ExE_Hyper"
## [5] "CancerPathway"  "Cancer_GeneSet"
```

2.1 Input

Similar as other pathway-enrichment analysis, NetORA expect annotated pathways and signatures as input. Since it's a network-based method, a network is also required.

2.1.1 Pathways

In vignettes of NetGPA, we have used gene prioritization to identify pathways that drive DNA methylation transformation (Zachary D. Smith 2017). Here we took an more intuitive approach by directly performing network-based pathway enrichment analysis.

```
# Genes near hyper-methylated CpG Islands in Mouse ExE
ExE_Hyper     <- Example_NetGPA$ExE_Hyper
```

We treat `ExE_Hyper` as an annotated pathway.

2.1.2 Signatures

Signatures are gene sets we want to test. Here we use pathways that are frequently mutated in cancers as signatures.

```
# show example query genes
Cancer_GeneSet   <- Example_NetGPA$Cancer_GeneSet
names(Cancer_GeneSet)
##  [1] "SIGNALLING_BY_NGF"
##  [2] "SIGNALING_BY_SCF_KIT"
##  [3] "SIGNALING_BY_ERBB4"
##  [4] "SIGNALING_BY_ERBB2"
##  [5] "SIGNALING_BY_EGFR_IN_CANCER"
##  [6] "NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE"
##  [7] "SIGNALING_BY_FGFR_IN_DISEASE"
##  [8] "DOWNSTREAM_SIGNAL_TRANSDUCTION"
##  [9] "CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM"
## [10] "SIGNALING_BY_FGFR"
```

2.1.3 Background

Background represents all possible genes evaluated when signatures are defined. For microarray studies, background is all genes measured in a microarray platform; For RNA-seq studies, background is all genes in genome.

2.1.4 Global networks

NetORA us NetGPA a back-end and thus requires networks in the same format, as described below. A gene network is represented as an integer matrix, in which column names are all genes included and each column contains top nearest neighbors of the gene indicated by column name. Here we will use a global text-mining network as an example.

```
# build a example global gene-network
data(text_2006_12_NetGPA)
networkMatrix <- text_2006_12_NetGPA

dim(networkMatrix)
## [1]  1884 18835
networkMatrix[1:10, c("IL12B", "TET2")]
##       IL12B  TET2
##  [1,]  7408  2361
##  [2,]  7475 18712
##  [3,]  7453  2851
##  [4,]  7410   685
##  [5,]  7403 17880
##  [6,]  7444 18583
##  [7,]  7428 18151
##  [8,]  7327   696
##  [9,]  7411  1667
## [10,]  7459  1808
colnames(networkMatrix)[7408]
## [1] "IL12A"
```

In the example shown above, a network covers 18835 genes and the nearest neighbor of IL12B is shown as 7408, which is the 7408th element of column names (gene IL12A).

2.2 Quick Start

Now we have a pathway `ExE_Hyper`, gene signatures in `Cancer_GeneSet` and networks in `text_2006_12_NetGPA`.

```
# Network-based pathway enrichment analysis
queryTable <- NetORA_Pre(ExE_Hyper, text_2006_12_NetGPA, progressBar = FALSE)
## 288 regions loaded successfully.
## 279 regions could be found in database.
## 18835 genes found in database.
PG <- colnames(text_2006_12_NetGPA)
mergedT <- NetORA_GS(Cancer_GeneSet, queryTable, PG, FDR = 0.05)

mergedT[order(mergedT$pvalue), ]
##                                                sgID nSG  nSR nOverlap
## 10                                SIGNALING_BY_FGFR 108 1368       21
## 7                     SIGNALING_BY_FGFR_IN_DISEASE 122 1368       22
## 3                               SIGNALING_BY_ERBB4  87 1368        7
## 1                                SIGNALLING_BY_NGF 212 1368       16
## 6  NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE 134 1368        9
## 4                               SIGNALING_BY_ERBB2  97 1368        6
## 8                    DOWNSTREAM_SIGNAL_TRANSDUCTION  91 1368        4
## 2                             SIGNALING_BY_SCF_KIT  75 1368        2
## 5                      SIGNALING_BY_EGFR_IN_CANCER 105 1368        3
## 9               CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM 263 1368        3
```

```
##        ES       pvalue
## 10 2.6770 8.493137e-06
## 7  2.4830 2.030136e-05
## 3  1.1080 2.968675e-01
## 1  1.0390 3.718025e-01
## 6  0.9247 5.120482e-01
## 4  0.8516 5.632468e-01
## 8  0.6052 7.994476e-01
## 2  0.3672 9.168084e-01
## 5  0.3934 9.520597e-01
## 9  0.1571 9.999963e-01
```

You can clearly see that only FGF-related signaling pathways are statistically significant.

2.3 Pre-computed networks

Building networks is time-consuming, we thus have pre-computed networks for all canonical pathways defined in MSigDB, using FDR 0.05 as cutoff. They can be loaded as a data set in R `data(MSigDB_NetORA_GS)`. Let's revisit the example we discussed above. In our previous study (Zachary D. Smith 2017), we have identified a list of potential pathways that regulate DNA methylation of a signature genes `ExE_Hyper`. In above section, we built a network using `ExE_Hyper`, and tested enrichment potential pathways. Here, we will do it in a reverse way.

```
data(MSigDB_NetORA_GS)
PG <- colnames(text_2006_12_NetGPA)
CancerPathway <- paste0("REACTOME_", Example_NetGPA$CancerPathway)
mergedT <- NetORA_MSigDB_CP(ExE_Hyper, MSigDB_NetORA_GS[CancerPathway], PG)

mergedT[order(mergedT$pvalue), ]
##                                                    Pathway nSG  nSR
## 10                                REACTOME_SIGNALING_BY_FGFR 260 3673
## 5                       REACTOME_SIGNALING_BY_FGFR_IN_DISEASE 260 3547
## 2  REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE 260 3276
## 8                     REACTOME_DOWNSTREAM_SIGNAL_TRANSDUCTION 260 3431
## 3                                  REACTOME_SIGNALLING_BY_NGF 260 3378
## 4                                REACTOME_SIGNALING_BY_ERBB2 260 3218
## 7                         REACTOME_SIGNALING_BY_EGFR_IN_CANCER 260 3373
## 6                                  REACTOME_SIGNALING_BY_ERBB4 260 2241
## 1                               REACTOME_SIGNALING_BY_SCF_KIT 260 2449
## 9                 REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM 260 2888
##    nOverlap    ES       pvalue          FDR
## 10       97 1.913 5.401353e-12 5.401353e-11
## 5        91 1.859 1.753193e-10 1.577874e-09
## 2        66 1.459 4.120740e-04 3.296592e-03
## 8        68 1.436 5.319814e-04 3.723870e-03
## 3        65 1.394 1.577755e-03 9.466530e-03
## 4        62 1.396 1.981885e-03 9.909426e-03
## 7        62 1.332 5.987946e-03 2.395178e-02
## 6        39 1.261 5.299439e-02 1.589832e-01
## 1        38 1.124 1.901741e-01 3.803482e-01
## 9        41 1.028 3.817132e-01 3.817132e-01
```

As expected, FGF-related signaling pathways are more statistically significant than others.

3 Network databases

NetORA could use all networks that are accepted by NetGPA. In current release of NetGPA, we have provided a text-mining network(`text_2006_12_NetGPA`), a co-expression network(`ce_v12_08_NetGPA`) and a integrative network(`DEPICT_2015_01_NetGPA`).

In the future, we will release more networks, including co-expression networks for Mouse and Rat.

4 Citation

If you use NetORA in published research, please cite NetORA and also (Raychaudhuri 2009).

5 Session info

```
sessionInfo()
## R version 3.4.2 (2017-09-28)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.1
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## 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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] NetORA_0.99.0   NetGPA_0.99.0   BiocStyle_2.4.1
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.4.2  backports_1.1.1 magrittr_1.5    rprojroot_1.2
##  [5] tools_3.4.2     htmltools_0.3.6 yaml_2.1.14     Rcpp_0.12.14
##  [9] stringi_1.1.6   rmarkdown_1.6   knitr_1.17      stringr_1.2.0
## [13] digest_0.6.12   evaluate_0.10.1
```

References

Aldo M. Roccaro, Jiantao Shi, Antonio Sacco, and Irene M. Ghobrial. 2016. "Exome Sequencing Reveals Recurrent Germ Line Variants in Patients with Familial Waldenström Macroglobulinemia." *Blood* 127 (21): 2598–2606.

Raychaudhuri, Robert M. AND Rossin, Soumya AND Plenge. 2009. "Identifying Relationships Among Genomic Disease Regions: Predicting Genes at Pathogenic Snp Associations and Rare Deletions." *PLOS Genetics* 5 (6). Public Library of Science: 1–15. doi:10.1371/journal.pgen.1000534.

Zachary D. Smith, Hongcang Gu, Jiantao Shi. 2017. "Epigenetic Restriction of Extraembryonic Lineages Mirrors the Somatic Transition to Cancer." *Nature* 549 (00): 543–47.