Network-based gene prioritization analysis with NetGPA

Jiantao Shi

28 November 2017

Abstract

It has been demonstrated that genes who function in the same pathway tend to distribute in a coherent sub-network. NetGPA implements a network-based gene prioritization method. Briefly, a sub-network is built using a given set of seed genes that are assumed to function in the same pathway or similar pathways. To predict whether a query gene is functionally related to the seed genes, we project this gene to a global network, and test whether connection of this gene to the subnetwork is random or statistically significant.

Package

NetGPA 0.99.0

Contents

1 Intruduction 2

2 Standard workflow 2

2.1 Input data 2

2.2 Quick Start 4

3 Network database 5

4 Examples of applications 5

4.1 Pathway evaluation 5

4.2 Disease gene prioritization in Waldenstrom's macroglobulinemia . . 6

4.3 Identification of pathways that drive DNA methylation landscape . . 8

5 Citation. 9

6 Session info 10

References 11

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. Several tools have been developed to address this issue, based on diverse approaches such as pathway enrichment (Segrè 2010), text mining (Raychaudhuri 2009) or protein-protein interaction (PPI) networks (Rossin 2011). One of the most widely used tools is GRAIL (Raychaudhuri 2009), which relies on text mining of gene functions from published literature. Core functions in GRAIL have been implemented as NetGPA, which will benefit researchers who mainly use R as analytically tool.

2 Standard workflow

To demonstrate the input and output data format in NetGPA, we have included three example data sets in this package.

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

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

2.1 Input data

As input, NetGPA expect three objects: a list which contains symbols of seed genes, a vector which contains symbols of query genes and a integer matrix which contain gene networks. We will use examples to explain their structures.

2.1.1 Seed genes

As an example, genes near Crohn's disease (CD) associated SNPs (Barrett et al. 2008) are stored in a data frame `CD_GWAS`, which is part of example data `Example_NetGPA`. Since multiple genes might locate in the same SNP locus, each element in the seed list is a vector of gene symbols.

```
# genes near Crohn's disease (CD) associated SNPs
CD_GWAS = Example_NetGPA$CD_GWAS
head(CD_GWAS)
##                            GIL    Validation
## rs10010325                TET2        FAILED
## rs10045431 RNF145 UBLCP1 IL12B     VALIDATED
## rs10188217               PUS10 INDETERMINATE
```

```
## rs1040092              SLC16A7        FAILED
## rs10753415               PARP1        FAILED
## rs10758669                JAK2     VALIDATED
```

An interesting feature of this data set is that all SNPs shown in the table were validated with larger sample size. Validation status for each SNP is also included in the table as VALIDATED, INDETERMINATE or FAILED. We thus could use this information to evaluate performance of our prediction.

```
# convert to list
CD_SeedList <- strsplit(as.character(CD_GWAS$GIL), " ")
names(CD_SeedList) <- rownames(CD_GWAS)
head(CD_SeedList)
## $rs10010325
## [1] "TET2"
##
## $rs10045431
## [1] "RNF145" "UBLCP1" "IL12B"
##
## $rs10188217
## [1] "PUS10"
##
## $rs1040092
## [1] "SLC16A7"
##
## $rs10753415
## [1] "PARP1"
##
## $rs10758669
## [1] "JAK2"
```

2.1.2 Query genes

Query genes are stored in a vector of gene symbols. In a GWAS study, it's usually the union of seed genes. Of course user could provide any genes of their interest.

```
# show example query genes
CD_query = unique(unlist(CD_SeedList))
head(CD_query)
## [1] "TET2"    "RNF145"  "UBLCP1"  "IL12B"   "PUS10"   "SLC16A7"
```

2.1.3 Global networks

NetGPA use networks for gene prioritization. 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).

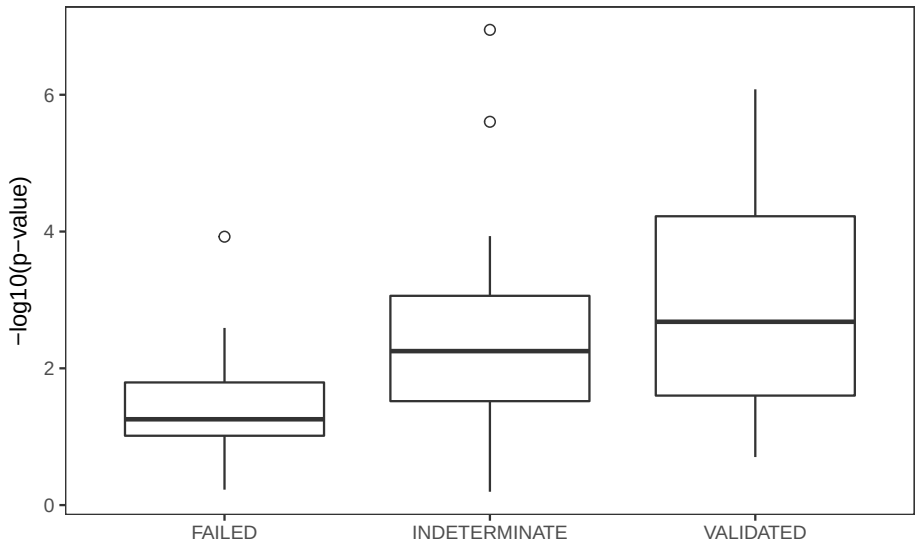
2.2 Quick Start

Now we have seed genes in `CD_SeedList`, query genes in `CD_query` and networks in `networkMatrix`.

```
# Prioritization of Crohn's disease-associated genes
rL <- NetGPA(CD_SeedList, CD_query, networkMatrix, Pfcutoff = 0.1, progressBar = FALSE)
## 73 regions loaded successfully.
## 69 regions could be found in database.
## 216 genes found in database.
queryTable <- rL$queryTable

head(queryTable[order(queryTable$queryP), ])
##               queryP     queryFDR
## IRF8    1.124462e-07 2.428838e-05
## IRF1    8.328621e-07 1.790654e-04
## IL12B   1.666499e-06 3.566308e-04
## IRGM    1.666999e-06 3.566308e-04
## IL18RAP 2.484258e-06 5.266626e-04
## IL18R1  5.407899e-06 1.141067e-03
```

NetGPA reports the prioritization p-value for each query gene, which could be used for subsequent analysis. Of note, we only included top 10% nearest genes for each gene in `networkMatrix`, so the maximum value of `Pfcutoff` is 0.1, which works well in most conditions. We now could check performance of our prediction using validation information.

```
# Prioritization of Crohn's disease-associated genes
library("ggplot2")
CD_SeedTable <- rL$seedTable
CD_SeedTable$Validation <- CD_GWAS[rownames(CD_SeedTable), "Validation"]

p <- ggplot(CD_SeedTable, aes(x = Validation, y = -log10(bestP)))
p <- p + theme_bw() + labs(x = "", y = "-log10(p-value)", title = "")
p <- p + geom_boxplot(outlier.shape = 1, outlier.size = 2)
p <- p + theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
p <- p + theme(legend.position="none")
print(p)
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



3 Network database

NetGPA could use all variety of networks, in current release, 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`).

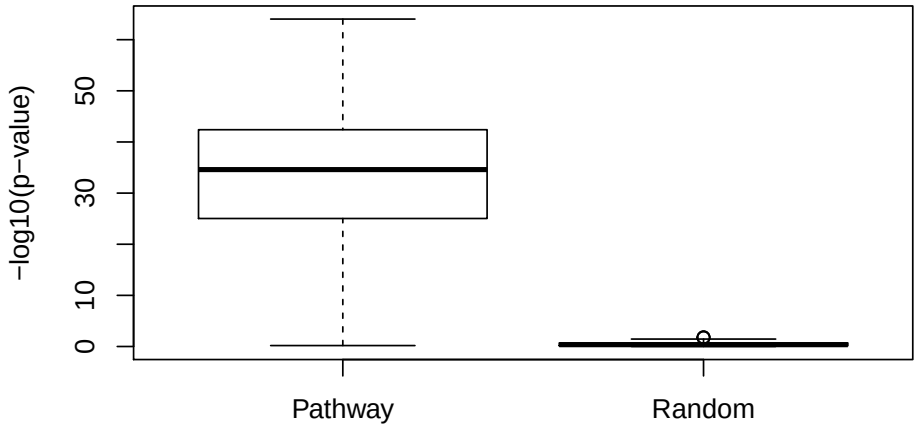
In the future, we will release more networks, including co-expression networks for Human, Mouse and Rat.

4 Examples of applications

4.1 Pathway evaluation

Pathways are manually curated, we may want to evaluate whether genes in a given pathway are functional related based on network prediction. Let's take pathway "SIGNALING_BY_FGFR" from REACTOME as example.

```
# Compare FGF pathway and a random gene set
exampleSeedFGFR <- Example_NetGPA$Cancer_GeneSet[["SIGNALING_BY_FGFR"]]

pGenes <- intersect(exampleSeedFGFR, colnames(text_2006_12_NetGPA))
rGenes <- sample(colnames(text_2006_12_NetGPA), length(pGenes))

res    <- NetGPA(as.list(pGenes), pGenes, text_2006_12_NetGPA, progressBar = FALSE)
## 108 regions loaded successfully.
## 108 regions could be found in database.
## 108 genes found in database.
pTable <- res$queryTable
pTable$group <- "Pathway"

res    <- NetGPA(as.list(rGenes), rGenes, text_2006_12_NetGPA, progressBar = FALSE)
## 108 regions loaded successfully.
## 108 regions could be found in database.
## 108 genes found in database.
rTable <- res$queryTable
rTable$group <- "Random"

mT <- rbind(pTable, rTable)

boxplot(-log10(queryP)~group, data = mT, ylab = "-log10(p-value)")
```



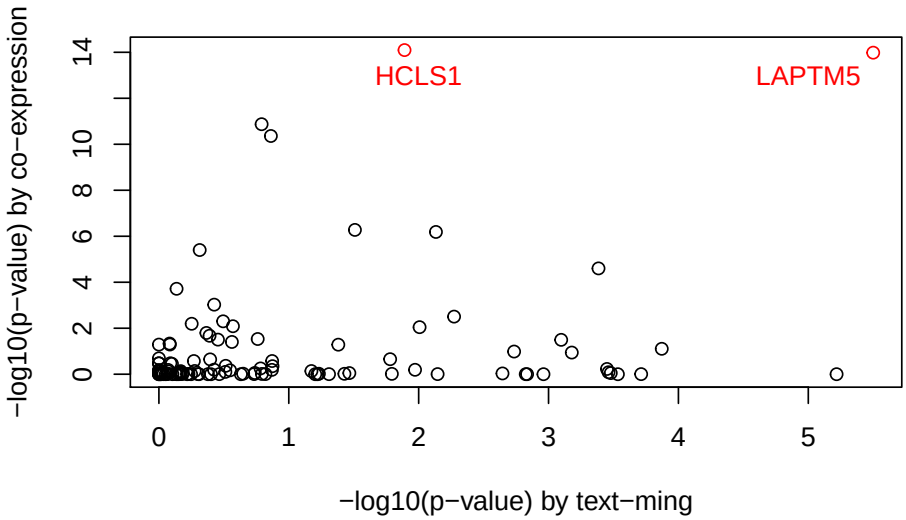
As shown above, most genes from FGFR-signaling are highly connected with other genes in the same pathway; in contract, we only observe non-significant signals when a random set of genes of the same size is tested.

4.2 Disease gene prioritization in Waldenstrom's macroglobulinemia

We have successfully used NetGPA for identification of disease genes in patients with familial Waldenstrom's macroglobulinemia (WM) (Aldo M. Roccaro and Ghobrial 2016). Briefly, we performed whole exome sequencing on germ line DNA obtained from 4 family members in which coinheritance for WM was documented in 3 of them. By using standard filtering

pipeline, 132 rare non-silent variants that are only present in affected members were identified. These variants locate in exons of 127 unique genes. It was expensive and time-consuming to validate all 127 genes. We thought it might be a good idea to prioritize these genes using gene sub-networks that were disrupted in WM. However, only few genes have been identified to be associated with WM. We thus took an alternative approach by comparing the gene expression profiles of WM B lymphocytes and normal B lymphocytes, the resulting gene expression signature is used as seed genes for prioritization.

```
data(ce_v12_08_NetGPA)

WM_Seed  <- Example_NetGPA$WM_Seed
WM_Query <- Example_NetGPA$WM_Query

# only keep genes that are present both data bases
WM_Query <- intersect(WM_Query, colnames(text_2006_12_NetGPA))
WM_Query <- intersect(WM_Query, colnames(ce_v12_08_NetGPA))

WM_text <- NetGPA(as.list(WM_Seed), WM_Query, text_2006_12_NetGPA, progressBar = FALSE)$queryTable
## 366 regions loaded successfully.
## 363 regions could be found in database.
## 108 genes found in database.
WM_coex <- NetGPA(as.list(WM_Seed), WM_Query, ce_v12_08_NetGPA,    progressBar = FALSE)$queryTable
## 366 regions loaded successfully.
## 353 regions could be found in database.
## 108 genes found in database.

# comapre the results of text-mining and coexpression
p_text  <- -log10(WM_text[WM_Query, "queryP"])
p_coex  <- -log10(WM_coex[WM_Query, "queryP"])
p_col   <- rep("black", length(WM_Query))
p_col[WM_Query %in% c("LAPTM5", "HCLS1")] <- "red"

plot(p_text, p_coex, col = p_col, type = "p",
     xlab = "-log10(p-value) by text-ming",
     ylab = "-log10(p-value) by co-expression")
text(2, 13, "HCLS1",  col = "red")
text(5, 13, "LAPTM5", col = "red")
```



In this example, It's obvious that coexpression network has better performance, since both HCLS1 and LAPTM5 have been validated using larger sample size (Aldo M. Roccaro and Ghobrial 2016). We recommend using coexpression network when seed genes are derived from gene expression signature. More details can be found in our publication.

4.3 Identification of pathways that drive DNA methylation landscape

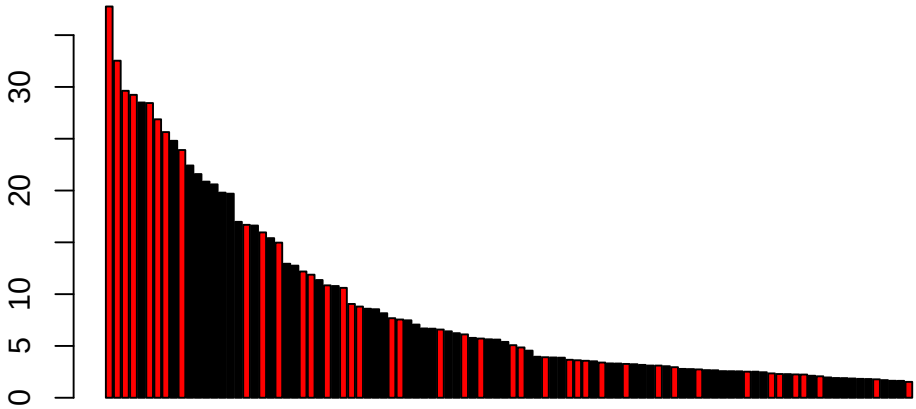
It was known for many years that DNA methylation landscape of cancer is characterized by global hypo-methylation and CpG Island (CGI) hyper-methylation, in contract to global hyper-methylation and CGI hypo-methylation in normal cells. However, genes and pathways that driver this transformation remain unclear. We hypothesized that mutated genes in cancer play a role in this transformation. We have performed pathway enrichment analysis on significantly mutated genes in 31 tumor types from TCGA and found that many pathways are recurrently mutated across majority of cancer types. Top 10 pathways are listed below.

```
CancerPathway <- Example_NetGPA$CancerPathway
cbind(CancerPathway)
##       CancerPathway
##  [1,] "SIGNALING_BY_SCF_KIT"
##  [2,] "NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE"
##  [3,] "SIGNALLING_BY_NGF"
##  [4,] "SIGNALING_BY_ERBB2"
##  [5,] "SIGNALING_BY_FGFR_IN_DISEASE"
##  [6,] "SIGNALING_BY_ERBB4"
##  [7,] "SIGNALING_BY_EGFR_IN_CANCER"
##  [8,] "DOWNSTREAM_SIGNAL_TRANSDUCTION"
##  [9,] "CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM"
## [10,] "SIGNALING_BY_FGFR"
```

It's hard to tell which pathway drive the transformation of DNA methylation landscape. Surprisingly, bifurcation of DNA methylation landscape is also observed in normal embryonic development. Specifically, compared to Epiblast, Extraembryonic Ectoderm (ExE) is globally hypo-methylated and locally hyper-methylated at CGIs. So ExE has cancer-like DNA

methylation landscape. We identified 768 ExE hyper-methylated CGIs and demonstrated that these CGIs are re-currently hyper-methylated almost all TCGA cancer types. Thus genes that locate near these CGIs represent a signature of transformation of DNA methylation. We used NetGAP to evaluate the relatedness between mutated genes in cancer and methylated genes in cancer.

```
ExE_Hyper     <- Example_NetGPA$ExE_Hyper
Cancer_gene   <- unique(unlist(Example_NetGPA$Cancer_GeneSet))

Cancer_text   <- NetGPA(as.list(ExE_Hyper), Cancer_gene, text_2006_12_NetGPA, progressBar = FALSE)
## 288 regions loaded successfully.
## 279 regions could be found in database.
## 539 genes found in database.
Cancer_qTable <- Cancer_text$queryTable
Cancer_qTable <- Cancer_qTable[order(Cancer_qTable$queryP),]

FGF_names     <- c("SIGNALING_BY_FGFR_IN_DISEASE", "SIGNALING_BY_FGFR")
FGF_pathway   <- unique(unlist(Example_NetGPA$Cancer_GeneSet[FGF_names]))
FGF_col       <- rep("black", nrow(Cancer_qTable))
FGF_col[rownames(Cancer_qTable) %in% FGF_pathway] = "red"

barplot(-log10(Cancer_qTable$queryP)[1:100], col = FGF_col[1:100])
```



Of the top 10 pathways that are mutated in cancer, FGF pathway is functionally related to cancer methylation signature with highest significance (by rank sum test). We only shows most significant 100 genes (genes in FGF signaling pathways are shown in red). We have successfully validated the role of FGF in regulation of DNA methylation in normal development (Zachary D. Smith 2017).

5 Citation

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

6 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] ggplot2_2.2.1   NetGPA_0.99.0   BiocStyle_2.6.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14     bookdown_0.5     digest_0.6.12    rprojroot_1.2
##  [5] plyr_1.8.4       grid_3.4.2       gtable_0.2.0     backports_1.1.1
##  [9] magrittr_1.5     scales_0.5.0     evaluate_0.10.1  rlang_0.1.4
## [13] stringi_1.1.6    lazyeval_0.2.1   rmarkdown_1.8    labeling_0.3
## [17] tools_3.4.2      stringr_1.2.0    munsell_0.4.3    yaml_2.1.14
## [21] compiler_3.4.2   colorspace_1.3-2 htmltools_0.3.6  knitr_1.17
## [25] tibble_1.3.4
```

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.

Barrett, Jeffrey C, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, et al. 2008. "Genome-Wide Association Defines More Than 30 Distinct Susceptibility Loci for Crohn's Disease." *Nature Genetics* 40 (8): 955–62.

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.

Rossin, Kasper AND Raychaudhuri, Elizabeth J. AND Lage. 2011. "Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology." *PLOS Genetics* 7 (1). Public Library of Science: 1–13. doi:10.1371/journal.pgen.1001273.

Segrè, Leif AND Mootha, Ayellet V. AND DIAGRAM Consortium AND MAGIC investigators AND Groop. 2010. "Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits." *PLOS Genetics* 6 (8). Public Library of Science: 1–19. doi:10.1371/journal.pgen.1001058.

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