METHOD ARTICLE

REVISED Enhancing gene set enrichment using networks [version 2; peer review: 2 approved, 2 approved with reservations]

Michael Prummer [1,2]

[1]NEXUS Personalized Health Technologies, ETH Zurich, Zurich, Switzerland
[2]Swiss Institute of Bioinformatics, Zurich, Switzerland

v2 **First published:** 30 Jan 2019, **8**:129 (https://doi.org/10.12688/f1000research.17824.1)
Latest published: 16 Jul 2019, **8**:129 (https://doi.org/10.12688/f1000research.17824.2)

Abstract
Differential gene expression (DGE) studies often suffer from poor interpretability of their primary results, i.e., thousands of differentially expressed genes. This has led to the introduction of gene set analysis (GSA) methods that aim at identifying interpretable global effects by grouping genes into sets of common context, such as, molecular pathways, biological function or tissue localization. In practice, GSA often results in hundreds of differentially regulated gene sets. Similar to the genes they contain, gene sets are often regulated in a correlative fashion because they share many of their genes or they describe related processes. Using these kind of neighborhood information to construct networks of gene sets allows to identify highly connected sub-networks as well as poorly connected islands or singletons. We show here how topological information and other network features can be used to filter and prioritize gene sets in routine DGE studies. Community detection in combination with automatic labeling and the network representation of gene set clusters further constitute an appealing and intuitive visualization of GSA results. The RICHNET workflow described here does not require human intervention and can thus be conveniently incorporated in automated analysis pipelines.

Keywords
differential gene expression analysis, gene set analysis, enrichment analysis, network analyis, GSEA

This article is included in the RPackage gateway.

This article is included in the Bioconductor gateway.

Open Peer Review

Reviewer Status ✔ ? ? ✔

	Invited Reviewers			
	1	**2**	**3**	**4**
REVISED **version 2** published 16 Jul 2019			? report	
version 1 published 30 Jan 2019	✔ report	? report	? report	✔ report

1 **Kimberly Glass**, Harvard Medical School, Boston, USA
Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA

2 **Monther Alhamdoosh**, CSL Limited, Parkville, Australia
The University of Melbourne, Parkville, Australia

3 **Jill P. Mesirov**, University of California, San Diego, San Diego, USA
Alexander Wenzel, University of California, San Diego, San Diego, USA

4 **Rita Casadio**, University of Bologna, Bologna, Italy

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Michael Prummer (prummer@nexus.ethz.ch)

Author roles: Prummer M: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Prummer M. **Enhancing gene set enrichment using networks [version 2; peer review: 2 approved, 2 approved with reservations]** F1000Research 2019, **8**:129 (https://doi.org/10.12688/f1000research.17824.2)

First published: 30 Jan 2019, **8**:129 (https://doi.org/10.12688/f1000research.17824.1)

REVISED **Amendments from Version 1**

This revised manuscript is based upon valuable input from the reviewers. Most importantly, installation and loading of required packages is now taken care of.

See referee reports

Introduction

Interpretation of whole-transcriptome differential expression studies is often difficult because the sheer volume of the differentially expressed genes (DEGs) can be overwhelming. It is common place in designed experiments with more than just a marginal biological effect to find several thousands of differentially expressed genes (DEGs). One way to handle the vast numbers and to identify the biological consequences of gene expression changes is to associate them with overarching processes involving a whole set of genes, such as Gene Ontology (GO) terms or Kyoto Enzyclopedia of Genes and Genomes (KEGG) pathways.

Curated genesets have been designed or discovered for a wide range of common contexts, such as, a biological process, molecular pathway, or tissue localization[1,2]. They have been introduced in the past not only to reduce complexity and to improve interpretability but also to increase statistical power by reducing the number of performed tests. As it turns out, this often results in finding hundreds of differentially regulated pathways[1].

As with co-expressed genes, many of the pathways exhibit strong mutual correlation because they contain a large proportion of shared genes which is in turn a result of the fact that many of them describe closely related aspects of an overarching biological theme. Therefore, to further increase interpretability of differential geneset regulation and to capture the global change of a biological phenotype, it would be desirable to identify possibly existing umbrella organizations among genesets.

Networks are ideal to model dependencies, interactions, and similarities among individuals[3–5], be it people, computers, genes, or genesets. The degree of connectivity between them can have an influence on information flow and defines communities or *cliques*, i.e., clusters of highly connected nodes within and infrequent connections between them.

In order to construct a geneset network, a similarity measure is required and can be defined as the fraction of common genes, also called the Jaccard index[6]. Other ways to measure similarity among genesets include, for instance, coexpression strength as implemented in WGCNA[7,8].

Community detection based on network topology is a standard problem in the analysis of social networks[9,10]. Well-established algorithms allow for computationally efficient clustering of genesets and can be used to identify highly connected sub-networks. There is no unique or optimal method available but many options exist. Popular methods to define clusters include the *edge-betweenness* criterion, the *Infomap* or the *Louvain* algorithm (`igraph`), as well as hierarchical or kmeans clustering.

Once geneset clusters are defined they can be characterized by their size and connectivity and thus prioritized and ranked. In particular, the clusters can be categorized as singletons, doublets, medium and large or dense and loose clusters.

Network analysis not only allows for detection of clusters and performance of measurements on them, networks are also straightforward and appealing visualizations of similarities among genesets. There are a couple of interactive visualization software tools available, of which Cytoscape is probably the most popular[11]. In some cases interactivity is useful but the emphasis here is to provide some of Cytoscape's features without any human intervention for easy integration into automatic analysis pipelines. For instance, automatic labeling of communities using the n most frequent terms was adopted here, similar as in Kucera *et al.*[12].

The purpose of this step-by-step workflow is to provide a fully automated and reproducible procedure for downstream analysis and visualization of differential geneset analysis results in R[13]. The focus is on supporting scientists in result interpretation by bringing order into the list of differentially regulated genesets based on biological rather than pure statistical arguments. The workflow is suitable for any kind of geneset library including new or custom sets and any kind of geneset analysis method.

[1]The terms *geneset* and *pathway* are used interchangeably throughout this document and refer to a set of genes.

Starting with differential expression analysis of a model dataset, geneset analysis is performed based on the MSigDB library. A geneset network is constructed to identify isolated genesets (singletons) and geneset pairs (doublets). Larger connected sub-networks are then split into smaller clusters of closely related genesets describing similar processes. The effect of each modification step on the network topology is visually documented in Figure 1–Figure 4. Using the most frequently occurring terms in the geneset names of a cluster, an attempt to automatically assign cluster labels is made. Finally, all labeled clusters of genesets are plotted to provide a one page overview of the results.

Preparations

The packages required for this workflow provide plotting functions (`ggplot2` and relatives), network functions (`igraph`[14], etc.), and `GGally`, text analytics functions (`wordcloud`, etc.) and gene expression analysis functions `DESeq2`[15], `limma`[16], and `org.Hs.eg.db`. The Rmarkdown version of this article further requires the Bioconductor package `BiocWorkflowTools` to be installed.

```
cranpcks = c("ggplot2", "gplots", "cowplot", "RColorBrewer", "knitr", "ggrepel",
             "reshape2", "kableExtra","igraph", "GGally", "network", "sna",
             "intergraph", "wordcloud", "tm", "SnowballC", "BiocManager")
biocpcks = c("DESeq2", "limma", "org.Hs.eg.db", "airway")
tmp = sapply(cranpcks, require, character.only=T, warn.conflicts=F, quietly=T)
toinstall = names(tmp)[!tmp]
if(length(toinstall)>0){
  install.packages(toinstall, quiet=T)
  sapply(toinstall, require, character.only=T, warn.conflicts=F, quietly=T)
}
tmp = sapply(biocpcks, require, character.only=T, warn.conflicts=F, quietly=T)
toinstall = names(tmp)[!tmp]
if(length(toinstall)>0) {
  BiocManager::install(toinstall)
  sapply(toinstall, require, character.only=T, warn.conflicts=F, quietly=T)
}
```

In addition to and often based on `igraph`, several R packages for network visualization are available and described in the form of tutorials[17,18].

Example data

We are using the popular *airway* data set[19] and perform a simple differential expression analysis.

```
data(airway)
dds = DESeqDataSetFromMatrix(countData = assay(airway),
                             colData = colData(airway),
                             design = ~ cell + dex)
dds$dex = relevel(dds$dex, "untrt")
dds = DESeq(dds, betaPrior = T)
res = results(dds, contrast = c("dex", "trt", "untrt"))
```

Mapping Ensembl IDs to ENTREZ IDs

We are using the popular `org.Hs.eg.db` package based on the UCSC annotation database and keep only genes with a unique mapping.

```
res$entrezgene = unname(mapIds(org.Hs.eg.db, keys = rownames(res),
                               column = "ENTREZID", keytype = "ENSEMBL"))
res = subset(res, subset = !is.na(res$entrezgene) & !is.na(res$stat))
res = res[-which(duplicated(res$entrezgene)), ]
```

Gene set enrichment analyis

We are using the popular KEGG, Reactome, and Biocarta pathways from the MSigDB gene set library C2. The following chunk guarantees that the gene set library list object is called `gset`.

```
url = "http://bioinf.wehi.edu.au/software/MSigDB/human_c2_v5p2.rdata"
temp.space = new.env()
bar = load(url(url), temp.space)
```

```
gset = get(bar, temp.space)
rm(temp.space)
gs.libs = sapply(names(gset), function(x) strsplit(x, "_")[[1]][1])
gset = gset[which(gs.libs %in% c("KEGG", "REACTOME", "BIOCARTA"))]
idx = ids2indices(gene.sets = gset, identifiers = res$entrezgene)
```

Alternatively, the same result can be obtained using the EGSEAdata experiment package[19b].

```
library(EGSEAdata)
library(EGSEA)
gset = EGSEA::buildMSigDBIdx(entrezIDs = res$entrezgene,
                             species = "human",
                             geneSets = "c2", min.size = 3)
idx = gset$c2@idx
gs.libs = sapply(names(idx), function(x) strsplit(x, "_")[[1]][1])
idx = idx[which(gs.libs %in% c("KEGG", "REACTOME", "BIOCARTA"))]
```

Competitive gene set enrichment analysis is performed using the function `camera()` from the `limma` package. We include uni-directional and bi-directional enrichment by using both the test statistics ("up" or "down") and its modulus ("mixed") for gene set testing. We limit the following network analysis to gene sets with a $FDR < 0.05$.

```
dat = cameraPR(res$stat, idx, sort = F)
dat$PValue.Mixed = cameraPR(abs(res$stat), idx, sort = F)$PValue
dat$FDR.Mixed = p.adjust(dat$PValue.Mixed, method = "BH")
dat$name = rownames(dat)

dat$Direction = as.character(dat$Direction)
dat$Direction[dat$FDR > 0.05] = "Mixed"
dat$Direction[dat$Direction == "Mixed" & dat$FDR.Mixed > 0.05] = "NOT"
dat$Direction = factor(dat$Direction, levels=c("NOT", "Up", "Down", "Mixed"))

idx = which(dat$Direction == "Mixed")
if(length(idx) > 0) dat$FDR[idx] = dat$FDR.Mixed[idx]
dat = dat[, -grep("\\.Mixed", names(dat))]
dat = dat[dat$Direction != "NOT", ]
dat$Direction = factor(dat$Direction, levels=c("Up", "Down", "Mixed"))
```

Starting from 1077 gene sets, 264 are found to be differentially regulated. Many of them are expected to describe similar processes and to be highly correlated.

Network construction

We construct a gene set network based on the proportion of common genes as the inverse distance measure. The nodes are gene sets which are connected by edges if the Jaccard index

$$J = \frac{\text{Number of common genes}}{\text{Number of all genes}}$$

is larger than a preset threshold, $J > 0.2$. While this threshold is somewhat arbitrary it has proven to be a reasonable one in many projects. As a guide for finding a reasonable threshold a broad distribution of disjoint cluster sizes is desired. Network analysis does not help if the cutoff is too large (no connections) or too small (all sets are connected with each other). In any case, it is strongly recommended to investigate its effect on the quality of the results.

```
# only keep gene sets present in the data
id.keep = which(names(gset) %in% dat$name)
gset = gset[id.keep]
# adjacency matrix
m.adj = sapply(gset, function(x)
  sapply(gset, function(y)
```

```
    length(intersect(unlist(x), unlist(y) ))
    )
  )
diag(m.adj) = 0
# Jaccard index matrix
NGenes = sapply(gset, length)
m.union = outer(NGenes, NGenes, "+") - m.adj
m.jacc = m.adj / m.union
```

The Jaccard matrix, or adjacency matrix, can be conveniently used to construct a network object using the function `igraph::graph_from_adjacency_matrix()`. In this example geneset, similarity is measured using all member genes irrespective of whether they were detected and present in the data. Alternatively, one could include only genes present in the data depending on whether the current data seem more relevant and trustworthy or the prior information given by the geneset definition. Graphical display is achieved here using `ggnet::ggnet2()` (Figure 1).

```
# choose node colors
palette = brewer.pal(9, "Set1")[c(1,2,9)]
names(palette) = c("Up", "Down", "Mixed")
# apply cutoff to Jaccard matrix
m.adj1 = m.adj * (m.jacc > 0.2)
# construct network object
net = graph_from_adjacency_matrix(m.adj1, "upper", diag = F, weighted = T)
# add vertex features
V(net)$size = dat$NGenes
V(net)$color = palette[dat$Direction]
V(net)$Direction = as.character(dat$Direction)
# plot
ggnet2(net, size = 2, color = "Direction", palette = palette,
       edge.size = 1, edge.color = "#99CC33")
```

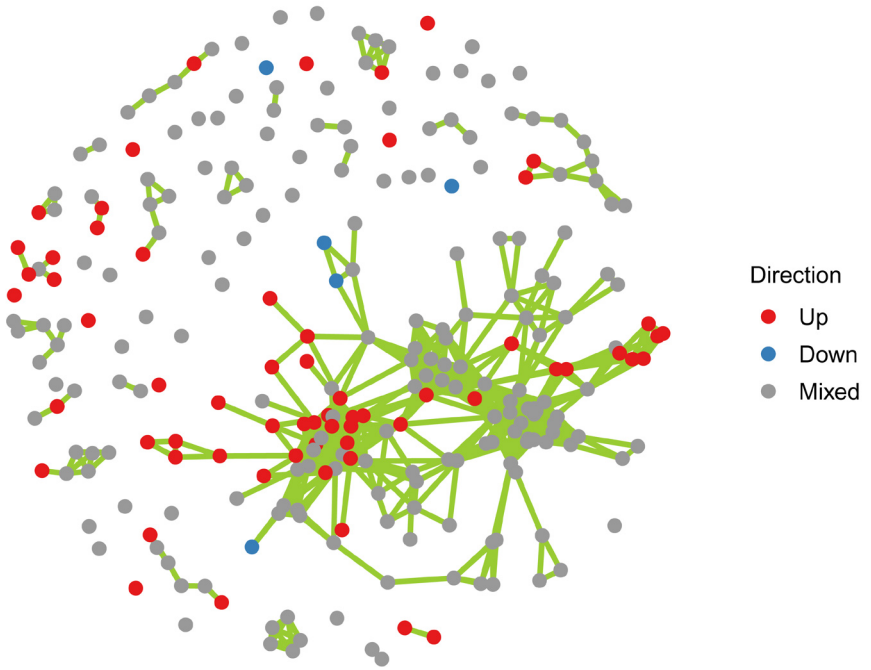

Figure 1. Graphical representation of the initial gene set network. Node colors indicate whether the member genes of a set are predominantly up or down regulated or whether there is no preferential direction (mixed).

Network modifications

In the following, components of the network for which network analysis does not improve interpretability are identified and put to aside. This includes singletons, i.e., genesets not connected to any other geneset, and doublets, also termed binary systems or dumbbells, i.e., pairs of genesets connected with each other but isolated from the rest.

Identify singletons

```
singletons = which(igraph::degree(net) == 0)
net1 = delete_vertices(net, singletons)
in.single = which(dat$name %in% V(net)$name[singletons])
tab = dat[in.single, ]
tab$FDR = signif(tab$FDR, 2)
tab$name = gsub("_", " ", tab$name)
tab = kable(tab[,c("name", "NGenes", "Direction", "FDR")],
            row.names = F, format = "latex",
            caption = "List of all singletons, i.e., genesets without
            sufficient overlap with any other geneset.")
kable_styling(tab, latex_options = "scale_down", font_size = 8)
```

In total, 49 singletons were identified and excluded from further analysis (Table 1). It is important to note that these genesets, while down-prioritized for the time being, may still be worthwhile investigating later.

```
ggnet2(net1, size = "size", max_size = 4, color = palette[V(net1)$Direction],
       size.cut = 4, edge.size = 1, edge.color = "#99CC33")
```

Figure 2 shows the remaining network clusters, with the size of the nodes representing the number of genes in the set.

Table 1. List of all singletons, i.e., genesets without sufficient overlap with any other geneset.

name	NGenes	Direction	FDR
KEGG GLYCOLYSIS GLUCONEOGENESIS	58	Mixed	0.04900
KEGG GLYCINE SERINE AND THREONINE METABOLISM	29	Mixed	0.00020
KEGG ARGININE AND PROLINE METABOLISM	52	Up	0.04100
KEGG GLUTATHIONE METABOLISM	43	Mixed	0.00210
KEGG O GLYCAN BIOSYNTHESIS	26	Mixed	0.01300
KEGG ARACHIDONIC ACID METABOLISM	48	Mixed	0.04300
KEGG NICOTINATE AND NICOTINAMIDE METABOLISM	23	Mixed	0.00290
KEGG CHEMOKINE SIGNALING PATHWAY	160	Mixed	0.02700
KEGG P53 SIGNALING PATHWAY	67	Mixed	0.01400
KEGG APOPTOSIS	78	Mixed	0.02800
KEGG TGF BETA SIGNALING PATHWAY	82	Mixed	0.00140
KEGG ADHERENS JUNCTION	72	Up	0.02900
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	99	Mixed	0.00360
KEGG PROGESTERONE MEDIATED OOCYTE MATURATION	80	Mixed	0.04900
KEGG ADIPOCYTOKINE SIGNALING PATHWAY	62	Up	0.00180
KEGG PATHOGENIC ESCHERICHIA COLI INFECTION	53	Mixed	0.04000
BIOCARTA AGR PATHWAY	33	Mixed	0.01300
BIOCARTA ATM PATHWAY	20	Mixed	0.00790
BIOCARTA BCELLSURVIVAL PATHWAY	15	Up	0.03400
BIOCARTA LAIR PATHWAY	14	Mixed	0.00910
BIOCARTA EPONFKB PATHWAY	11	Mixed	0.01300
BIOCARTA GABA PATHWAY	6	Mixed	0.04800
BIOCARTA P53HYPOXIA PATHWAY	22	Mixed	0.02600
BIOCARTA EGFR SMRTE PATHWAY	11	Mixed	0.01600

name	NGenes	Direction	FDR
BIOCARTA PPARA PATHWAY	52	Mixed	0.00091
BIOCARTA RAC1 PATHWAY	20	Mixed	0.00440
BIOCARTA NKCELLS PATHWAY	14	Mixed	0.02800
REACTOME METABOLISM OF VITAMINS AND COFACTORS	50	Mixed	0.03100
REACTOME IL 7 SIGNALING	10	Mixed	0.00062
REACTOME SULFUR AMINO ACID METABOLISM	24	Up	0.00530
REACTOME SPHINGOLIPID DE NOVO BIOSYNTHESIS	30	Mixed	0.00970
REACTOME SIGNALING BY HIPPO	20	Mixed	0.00110
REACTOME GASTRIN CREB SIGNALLING PATHWAY VIA PKC AND MAPK	171	Mixed	0.05000
REACTOME PLATELET ADHESION TO EXPOSED COLLAGEN	10	Mixed	0.00910
REACTOME VEGF LIGAND RECEPTOR INTERACTIONS	10	Mixed	0.05000
REACTOME METABOLISM OF AMINO ACIDS AND DERIVATIVES	182	Mixed	0.03100
REACTOME TRANSMISSION ACROSS CHEMICAL SYNAPSES	161	Mixed	0.01900
REACTOME INTEGRATION OF ENERGY METABOLISM	104	Mixed	0.02900
REACTOME CYTOSOLIC TRNA AMINOACYLATION	24	Down	0.03000
REACTOME OLFACTORY SIGNALING PATHWAY	65	Mixed	0.00220
REACTOME SEMA3A PLEXIN REPULSION SIGNALING BY INHIBITING INTEGRIN ADHESION	13	Mixed	0.04000
REACTOME NA CL DEPENDENT NEUROTRANSMITTER TRANSPORTERS	12	Down	0.03700
REACTOME SYNTHESIS AND INTERCONVERSION OF NUCLEOTIDE DI AND TRIPHOSPHATES	18	Up	0.04700
REACTOME ROLE OF DCC IN REGULATING APOPTOSIS	10	Mixed	0.00420
REACTOME NETRIN1 SIGNALING	35	Mixed	0.00940
REACTOME NEPHRIN INTERACTIONS	17	Up	0.03000
REACTOME RAP1 SIGNALLING	15	Mixed	0.00900
REACTOME ETHANOL OXIDATION	10	Up	0.00012
REACTOME HORMONE SENSITIVE LIPASE HSL MEDIATED TRIACYLGLYCEROL HYDROLYSIS	11	Mixed	0.01000

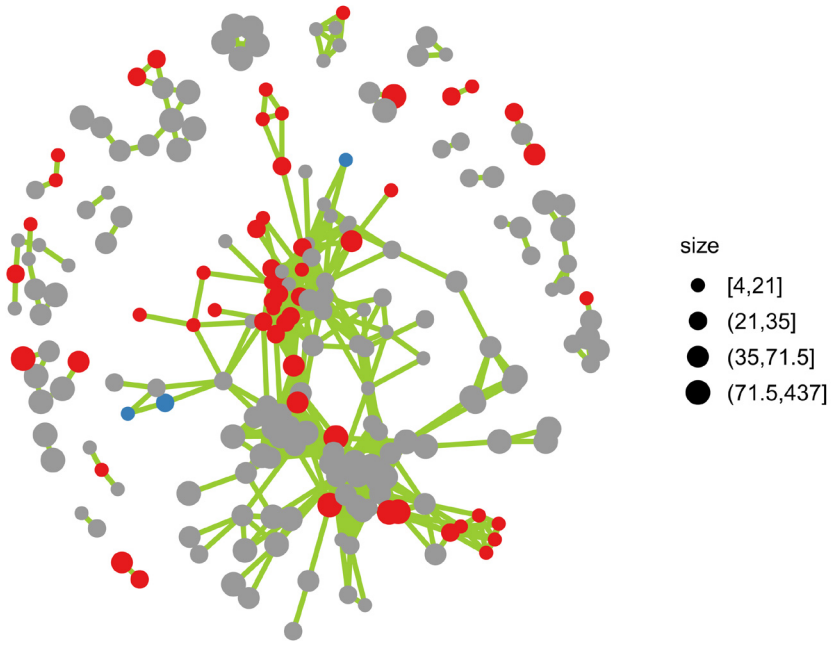

Figure 2. Gene set network with singletons removed. The color scheme is the same as above. The node size corresponds to the number of genes in a set.

Identify binary systems (2 sets)

Next we also want to separate clusters with less than 3 gene sets. To do so, we separate disjoint subnets as individual objects, count their members, and delete all vertices belonging to clusters of size smaller than 3.

```
clu1 = igraph::components(net1)
clu.lt3 = which(sizes(clu1) < 3)
v.clu.lt3 = which(clu1$membership %in% clu.lt3)
net2 = delete_vertices(net1, v.clu.lt3)
clu2 = igraph::components(net2)
in.clu.lt3 = which(dat$name %in% V(net1)$name[v.clu.lt3])
tab = dat[in.clu.lt3, ]
tab$FDR = signif(tab$FDR,2)
cludp = clu1$membership[v.clu.lt3]
cludp = data.frame(name = names(cludp), id = as.numeric(cludp))
tab = merge(tab,cludp)
tab$name = gsub("_", " ", tab$name)
tab = kable(tab[order(tab$id), c("id", "name", "NGenes", "Direction", "FDR")],
            row.names=F, format = "latex",
            caption = "List of binary clusters as indicated by the id column.")
kable_styling(tab, latex_options = "scale_down", font_size = 8)
```

In Table 2, consecutively listed gene sets with the same *id* belong to the same binary cluster. Often these are gene sets from different libraries describing the same biological process or phenotype. In total, 16 binary clusters were identified, for which network analysis would not be useful.

```
set.seed(16)
nodecol = colorRampPalette(brewer.pal(9, "Set1")[sample(9)])(max(clu2$membership))
ggnet2(net2, size = "size", max_size = 4, color = nodecol[clu2$membership],
       size.cut = 4, edge.size = 1, edge.color = "grey")
```

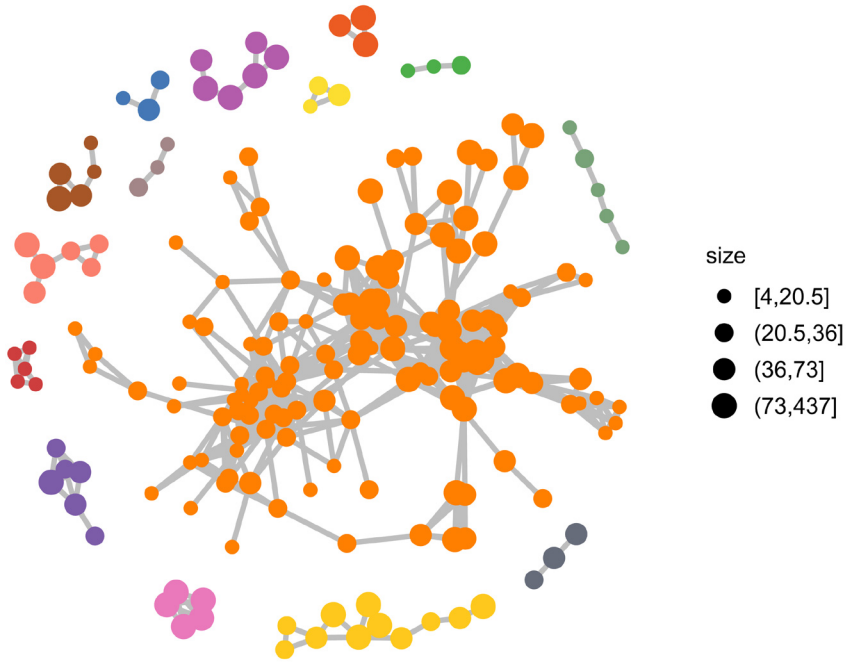

Figure 3. Gene set network with singletons and binary clusters removed. Colored according to disjoint subnetworks.

Table 2. List of binary clusters as indicated by the id column.

id	name	NGenes	Direction	FDR
3	KEGG ALANINE ASPARTATE AND GLUTAMATE METABOLISM	28	Mixed	4.9e-03
3	REACTOME AMINO ACID SYNTHESIS AND INTERCONVERSION TRANSAMINATION	16	Mixed	3.6e-04
6	KEGG INOSITOL PHOSPHATE METABOLISM	49	Mixed	2.0e-04
6	KEGG PHOSPHATIDYLINOSITOL SIGNALING SYSTEM	69	Mixed	1.5e-04
17	BIOCARTA ARF PATHWAY	16	Mixed	1.2e-02
17	BIOCARTA CTCF PATHWAY	22	Mixed	3.8e-04
18	REACTOME PLATELET ACTIVATION SIGNALING AND AGGREGATION	178	Mixed	2.9e-03
18	REACTOME RESPONSE TO ELEVATED PLATELET CYTOSOLIC CA2	72	Mixed	3.9e-02
19	REACTOME NEUROTRANSMITTER RELEASE CYCLE	28	Up	1.4e-02
19	REACTOME NOREPINEPHRINE NEUROTRANSMITTER RELEASE CYCLE	10	Up	4.5e-02
20	REACTOME AMINO ACID AND OLIGOPEPTIDE SLC TRANSPORTERS	40	Mixed	2.0e-02
20	REACTOME AMINO ACID TRANSPORT ACROSS THE PLASMA MEMBRANE	29	Mixed	8.7e-03
22	REACTOME MUSCLE CONTRACTION	42	Up	3.4e-05
22	REACTOME SMOOTH MUSCLE CONTRACTION	23	Up	0.0e+00
23	REACTOME ACTIVATION OF GENES BY ATF4	24	Mixed	7.8e-03
23	REACTOME PERK REGULATED GENE EXPRESSION	27	Mixed	2.0e-02

Without singletons and binary clusters, we are left with larger disjoint subnets (Figure 3).

Detect communities (sub-networks)

The larger disjoint clusters may consist of so-called *communities*, i.e., sub-networks of highly inter-connected nodes that stick together by only one or a few edges. We are using the popular *edge betweenness* property to identify these community-connecting edges and remove them in order to split large clusters into smaller ones.

```
net2 = delete_edge_attr(net2, "weight")
clu3 = cluster_edge_betweenness(net2)
# delete edges between communities
net3 = delete_edges(net2, which(as.vector(crossing(clu3, net2))) )
# remove clusters of size <3
small_cluster_ids = which(sizes(clu3) < 3)
small_cl_v = which(clu3$membership %in% small_cluster_ids)
net3 = delete_vertices(net3, small_cl_v)

clu3 = igraph::components(net3)
nodecol = c(brewer.pal(9, "Paired"), brewer.pal(9, "Set3") )
nodecol = colorRampPalette(nodecol)(max(clu3$membership))

ggnet2(net3, size = 0, color = nodecol[clu3$membership],
       edge.size = 1.0, edge.color = "grey") +
  geom_point(size = 2, color = "black") +
  geom_point(aes(color = color), size = 1)
```

The result of this network-based clustering is shown in Figure 4

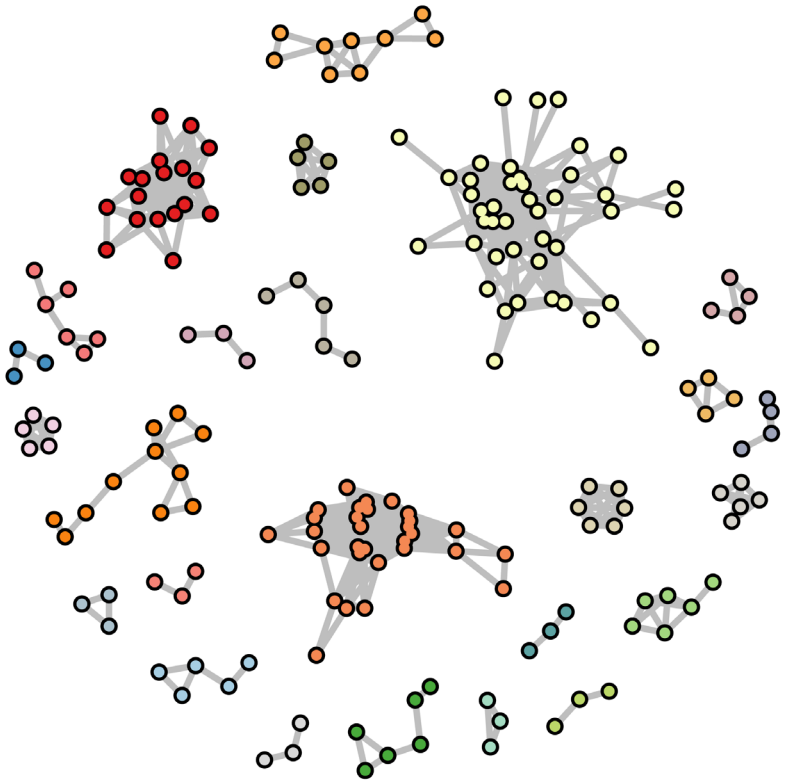

Figure 4. Disjoint clusters after community detection and splitting.

Automatic annotation of gene set clusters

In analogy to the popular interactive network visualization tool *cytoscape*[12], we attempt to generate automatic labels for gene set clusters. Gene set names are split into individual words and counted within each cluster. The four most frequent terms occurring at least twice are used as labels. The function `clust_head()` is defined for this purpose and contains an exclusion list of words not used.

```
t.rW = c("cell", "process", "regulation", "negative", "positive", "signaling",
         "response", "stimulus", "signal", "activity", "protein", "involved",
         "component", "level", "effector", "event", "projection", "organismal",
         "cellular", "modification", "pathway", "mediated", "dependent",
         "organization", "group", "target", "biocarta", "kegg", "reactome")
clust_head = function(x){
  txt = unlist(strsplit(x, "_"))
  txt = Corpus(VectorSource(txt))
  txt = tm_map(txt, PlainTextDocument)
  txt = tm_map(txt, removePunctuation)
  txt = tm_map(txt, removeNumbers)
  txt = tm_map(txt, content_transformer(tolower))
  txt = tm_map(txt, removeWords, c(t.rW, stopwords("english")))
  tdm = TermDocumentMatrix(txt)
  m = as.matrix(tdm)
  word_freqs = sort(rowSums(m), decreasing=TRUE)
  word_freqs = word_freqs[word_freqs>1]
  word_freqs = paste(names(word_freqs)[1:4], collapse=" ")
  gsub("[[:space:]]?NA[[:space:]]?", "", word_freqs)
}
```

Lattice of annotated networks

There are many possibilities to visualize geneset clusters and often a compromise between information content and crowding has to be found. Here, we are producing a lattice of network plots, one for each sub-net, with the automatic annotation as title (Figure 5). We begin by generating the cluster titles using the `clust_head()` function followed by cleaning up and ordering by cluster size.

```
clust = data.frame(cl = clu3$membership)
rownames(clust) = names(V(net3))
# generate cluster titles
cl3.lab.txt = as.character(tapply(rownames(clust), clust$cl, clust_head))
# remove NAs
cl3.lab.txt = gsub("[[:space:]]?NA[[:space:]]?", "", cl3.lab.txt)
clu3 = igraph::components(net3)
clu.order = order(clu3$csize, decreasing = T)
clu3$mem = match(clu3$membership, clu.order)
```

Then we generate a list of ggplot objects, one for each cluster or sub-net. For smaller sub-nets, the nodes are labelled with the first 4 words of their names; the first word was removed before as it is usually the name of the geneset library. For larger sub-nets, this is not feasible without overprinting. Titles are missing if none of the words from the geneset names occurred more than once. This may be indicative for a semantically mixed cluster or for sparse prior knowledge.

```
# generate a list of ggplots
g = list(max(clu3$membership))
set.seed(7042016)
for (ii in 1:max(clu3$membership)) {
  subgf = induced_subgraph(net3, which(clu3$mem == ii))
  # generate titles with one optional line break
  title = substr(toupper(cl3.lab.txt[clu.order][ii]), 1, 60)
  if (nchar(title) > 25) {
    title = sub("(^.{10,30})[[:space:]]","\\1\\\n", title)
  }
  # generate node labels using word 2-5 of the geneset name
  v.label = names(V(subgf))
  v.label = lapply(v.label, function(x) strsplit(x, "_")[[1]])
  v.label = sapply(v.label, function(x) paste(x[2:min(5, length(x))], collapse = "_"))
  # clean up geneset names
  v.label =  gsub("_PATHWAY","", v.label)
  v.label =  gsub("_SIGNALING", "", v.label)
  # introduce line breaks
  v.label =  gsub("_","\n", v.label)
  # remove node labels for large clusters
  if (length(v.label) > 5) v.label = rep(NA, length(v.label))
  g[[ii]] = ggnet2(subgf, edge.size = 1, edge.color = "#99CC33",
                   label = F, size=V(subgf)$size, max_size = 3,
                   size.cut = 4, color = palette[V(subgf)$Direction]) +
    theme(legend.position="none", plot.title = element_text(size=6),
          panel.grid = element_blank()) +
    geom_label_repel(label = v.label, size=1.2,
                     box.padding = 0.1, label.padding = 0.1) +
    ggtitle(title) }

nr.cols = min(4,max(clu3$membership))
nr.rows = ceiling(max(clu3$membership) / nr.cols)
width = sapply(g, function(x) nrow(x$data))
grid.arrange = getFromNamespace("grid.arrange", asNamespace("gridExtra"))
grid.arrange(grobs = g[seq(16)], ncol = nr.cols)
```

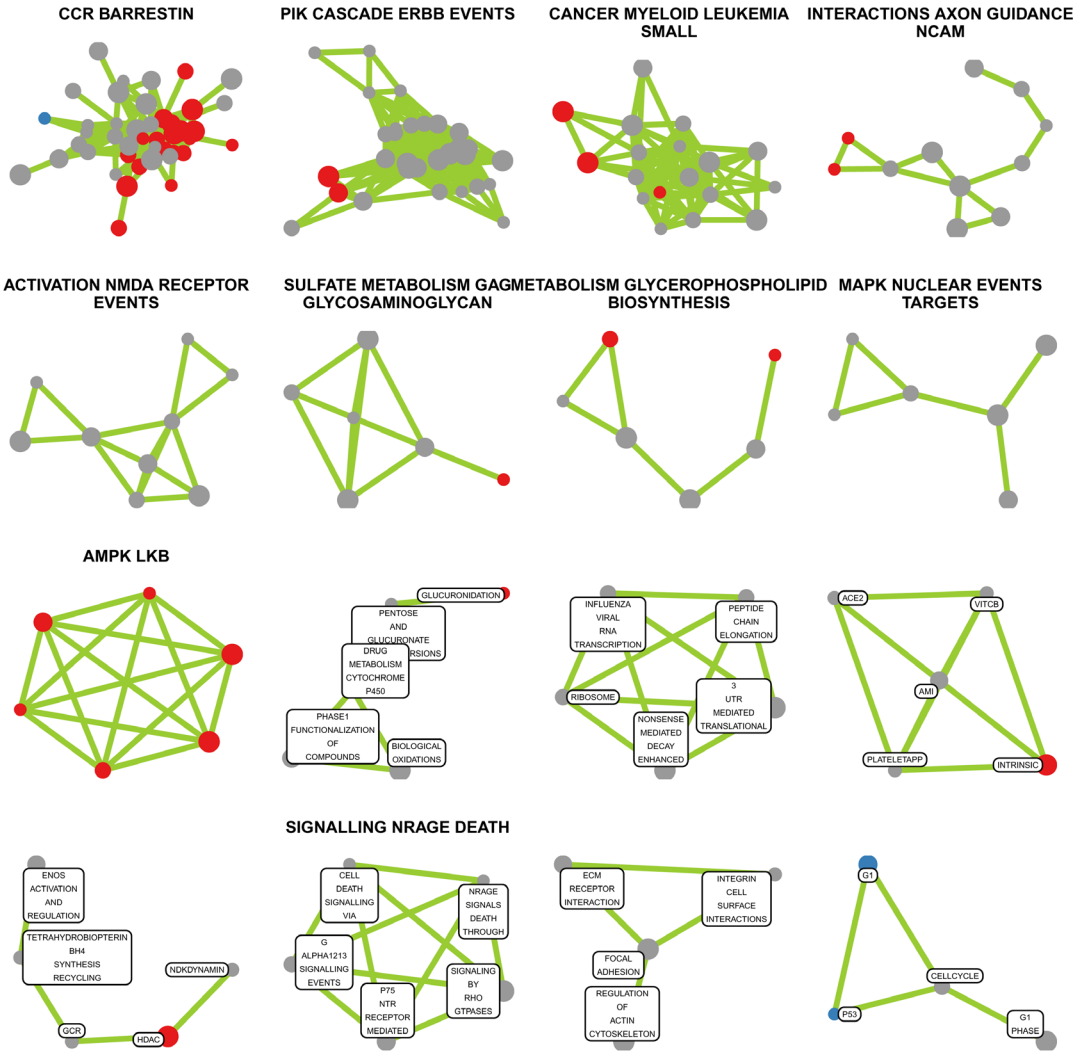

Figure 5. Geneset cluster with machine-generated titles. Only the first 16 connected subnets are shown. Geneset labels are omitted for clusters with more than 5 members.

Discussion

We have presented an automated workflow based on a small number of R packages for prioritization and visualization of gene set analysis results using networks, which we call RICHNET. We demonstrated how community detection facilitates categorization of differentially regulated gene sets into singletons and clusters of different size ranges. Automated label generation allowed to associate these clusters with biological themes or processes of which the member gene sets are part of.

The RICHNET workflow could be altered or extended quite naturally in a number of ways but the version presented here is the one we typically apply in our research service projects. One advantage over other approaches is that it does not depend on a particular geneset library or geneset analysis method. Any means of selecting genesets of interest can be used. Specific hierarchically constructed genesets, such as GO terms, would offer a straightforward way to arrive at a more global process description using higher levels in their tree structure. A second advantage is that it does not depend on the existence of a good quality gene or protein interaction network for the particular organism or disease state which is often not feasible. Only very few genesets are network-based (e.g. KEGG pathways) and would thus offer a straight-forward way to use an *a priori* network topology. Thirdly, similar as in reference 8, a geneset similarity network could be constructed in the form of a

co-enrichment network from GSVA enrichment scores[20] using weighted co-expression network analysis (WGCNA)[7]. However, this approach relies on a relatively large sample size whereas the sample size requirement of RICHNET is not more than the GSA it relies on.

As an alternative to the networks of genesets described here, networks of genes could be created in a reciprocal way. The underlying similarity metric between genes could be defined as the proportion of common genesets among all genesets they are part of. This approach would be equivalent to a STRING-DB network with "databases" as the only interaction allowed[21].

One possible future extension of the RICHNET workflow could be the introduction of a consensus similarity metric from multiple initial networks and different community detection or cluster algorithms to improve stability against noise. A second avenue forward could be the introduction of interactive graphics in 2D or 3D[17] to allow moving, pulling, rotation or zoom and display of node specific or edge specific information.

Some may argue in favor of encapsulating the RICHNET workflow in an R or Bioconductor package. However, it is our strong believe that for the sake of transparency and given the straightforward nature of the code it serves better to publish it openly. This way we encourage the users to adapt it to their specific requirements, to improve and expand on it.

Data availability

The data used in this workflow is included in the *airway* R-package[19].

Software availability

The R markdown file for this workflow is available at: https://doi.org/10.5281/zenodo.3271565[13].

Packages used

This workflow depends on various packages from version 3.7 of the Bioconductor project, running on R version 3.5.0 or higher. A complete list of the packages used for this workflow is shown below:

```
sessionInfo()

R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 17134)

Matrix products: default

locale:
[1] LC_COLLATE=English_United Kingdom.1252
[2] LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] parallel stats4 stats graphics grDevices utils datasets
[8] methods base

other attached packages:
[1] airway_1.2.0 SnowballC_0.5.1
[3] tm_0.7-5 NLP_0.2-0
[5] wordcloud_2.6 org.Hs.eg.db_3.7.0
[7] AnnotationDbi_1.44.0 limma_3.38.2
[9] DESeq2_1.22.1 SummarizedExperiment_1.12.0
[11] DelayedArray_0.8.0 BiocParallel_1.16.1
[13] matrixStats_0.54.0 Biobase_2.42.0
[15] GenomicRanges_1.34.0 GenomeInfoDb_1.18.1
```

```
[17] IRanges_2.16.0 S4Vectors_0.20.1
[19] BiocGenerics_0.28.0 igraph_1.2.2
[21] GGally_1.4.0 kableExtra_0.9.0
[23] knitr_1.20 reshape2_1.4.3
[25] ggrepel_0.8.0 cowplot_0.9.3
[27] gplots_3.0.1 ggplot2_3.1.0
[29] RColorBrewer_1.1-2

loaded via a namespace (and not attached):
[1] colorspace_1.3-2 rprojroot_1.3-2
[3] htmlTable_1.12 XVector_0.22.0
[5] base64enc_0.1-3 fs_1.2.6
[7] rstudioapi_0.8 remotes_2.0.2
[9] bit64_0.9-7 xml2_1.2.0
[11] codetools_0.2-15 splines_3.5.1
[13] geneplotter_1.60.0 pkgload_1.0.2
[15] Formula_1.2-3 annotate_1.60.0
[17] cluster_2.0.7-1 intergraph_2.0-2
[19] readr_1.2.1 compiler_3.5.1
[21] httr_1.3.1 backports_1.1.2
[23] assertthat_0.2.0 Matrix_1.2-15
[25] lazyeval_0.2.1 cli_1.0.1
[27] acepack_1.4.1 htmltools_0.3.6
[29] prettyunits_1.0.2 tools_3.5.1
[31] bindrcpp_0.2.2 coda_0.19-2
[33] gtable_0.2.0 glue_1.3.0
[35] GenomeInfoDbData_1.2.0 dplyr_0.7.8
[37] BiocWorkflowTools_1.8.0 Rcpp_1.0.0
[39] slam_0.1-43 statnet.common_4.1.4
[41] gdata_2.18.0 xfun_0.4
[43] stringr_1.3.1 network_1.13.0.1
[45] ps_1.2.1 testthat_2.0.1
[47] rvest_0.3.2 gtools_3.8.1
[49] devtools_2.0.1 XML_3.98-1.16
[51] zlibbioc_1.28.0 scales_1.0.0
[53] hms_0.4.2 yaml_2.2.0
[55] memoise_1.1.0 gridExtra_2.3
[57] rpart_4.1-13 RSQLite_2.1.1
[59] reshape_0.8.8 latticeExtra_0.6-28
[61] stringi_1.2.4 genefilter_1.64.0
[63] desc_1.2.0 checkmate_1.8.5
[65] caTools_1.17.1.1 pkgbuild_1.0.2
[67] rlang_0.3.0.1 pkgconfig_2.0.2
[69] bitops_1.0-6 evaluate_0.12
[71] lattice_0.20-38 purrr_0.2.5
[73] bindr_0.1.1 htmlwidgets_1.3
[75] bit_1.1-14 processx_3.2.0
[77] tidyselect_0.2.5 plyr_1.8.4
[79] magrittr_1.5 bookdown_0.7
[81] R6_2.3.0 Hmisc_4.1-1
[83] sna_2.4 DBI_1.0.0
[85] pillar_1.3.0 foreign_0.8-71
[87] withr_2.1.2 survival_2.43-3
[89] RCurl_1.95-4.11 nnet_7.3-12
[91] tibble_1.4.2 crayon_1.3.4
[93] KernSmooth_2.23-15 rmarkdown_1.10
[95] usethis_1.4.0 locfit_1.5-9.1
[97] grid_3.5.1 data.table_1.11.8
[99] blob_1.1.1 callr_3.0.0
[101] git2r_0.23.0 digest_0.6.18
```

```
[103] xtable_1.8-3 munsell_0.5.0
[105] viridisLite_0.3.0 sessioninfo_1.1.1
```

Author contributions

MP conceptualized the content, developed the method, performed the analysis and wrote the manuscript.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

The author would like to thank all members of NEXUS and in particular Daniel Stekhoven for fruitful discussions as well as Beate Sick (UZH) and Phil Cheng (USZ) for critically reading the manuscript.

References

1. Rouillard AD, Gundersen GW, Fernandez NF, *et al.*: **The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins.** *Database (Oxford).* 2016; **2016**: pii: baw100.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Liberzon A, Subramanian A, Pinchback R, *et al.*: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics.* 2011; **27**(12): 1739–1740.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet.* 2004; **5**(2): 101–113.
PubMed Abstract | Publisher Full Text
4. Vidal M, Cusick ME, Barabási AL: **Interactome networks and human disease.** *Cell.* 2011; **144**(6): 986–998.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Ideker T, Krogan NJ: **Differential network biology.** *Mol Syst Biol.* 2012; **8**(1): 565.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Merico D, Isserlin R, Stueker O, *et al.*: **Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.** *PLoS One.* 2010; **5**(11): e13984.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics.* 2008; **9**(1): 559.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Thorsson V, Gibbs DL, Brown SD, *et al.*: **The Immune Landscape of Cancer.** *Immunity.* 2018; **48**(4): 812–830.e14.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Girvan M, Newman ME: **Community structure in social and biological networks.** *Proc Natl Acad Sci U S A.* 2002; **99**(12): 7821–6.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Bedi P, Sharma C: **Community detection in social networks.** *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* 2016; **6**(3): 115–135.
Publisher Full Text
11. Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–504.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Kucera M, Isserlin R, Arkhangorodsky A, *et al.*: **AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations [version 1; referees: 2 approved].** *F1000Res.* 2016; **5**: 1717.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Michael P: **Enhancing gene set enrichment using networks.** *zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3271565
14. Csardi G, Nepusz T: **The igraph software package for complex network research.** *InterJournal.* Complex Sy:1695, 2006.
Reference Source
15. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 550.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Ritchie ME, Phipson B, Wu D, *et al.*: ***limma* powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.* 2015; **43**(7): e47.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Ognyanova K: **Static and dynamic network visualization with R**. 2015.
Reference Source
18. Tyner S, Briatte F, Hofmann H: **Network Visualization with ggplot2.** *R Foundation for Statistical Computing.* 2017; **9**(1): 27–59.
Publisher Full Text
19. Himes BE, Jiang X, Wagner P, *et al.*: **RNA-Seq transcriptome profiling identifies *CRISPLD2* as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells.** *PLoS One.* 2014; **9**(6): e99625.
PubMed Abstract | Publisher Full Text | Free Full Text

19b. Alhamdoosh M, Law CW, Tian L, *et al.*: **Easy and efficient ensemble gene set testing with EGSEA [version 1; peer review: 1 approved, 3 approved with reservations].** *F1000Res.* 2017; **6**: 2010.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Hänzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data.** *BMC Bioinformatics.* 2013; **14**(1): 7.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Szklarczyk D, Morris JH, Cook H, *et al.*: **The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.** *Nucleic Acids Res.* 2017; **45**(D1): D362–D368.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ? ✔

Version 2

Reviewer Report 02 October 2019

https://doi.org/10.5256/f1000research.21861.r51269

?

Jill P. Mesirov
Institute of Genomic Medicine, University of California, San Diego, San Diego, CA, USA
Alexander Wenzel
University of California, San Diego, San Diego, CA, USA

Regretfully, many of our previous comments have not been adequately addressed in the current revision and the author's response, particularly in regard to comparing the proposed method to existing work and justifying its ability to prioritize gene sets in a biologically relevant way. We also have concerns about design decisions in the algorithm and workflow, including the choice of Jaccard cut off and the use of categorical enrichment scoring, that are not adequately explained in the manuscript and responses and continuing technical problems in running the code.

1. The author acknowledges the similarity to the Enrichment Map method (Merico et al. 2010[1]) and notes that it is cited in the manuscript. However, the method is never explicitly mentioned in the text nor is its similarity to the author's method mentioned even though the similarity is mentioned in the response to review. Given this similarity we feel the text should include a comparison or a discussion of benefits or drawbacks of the respective methods. While the author responded that Merico et al.[1] is cited, the citation only refers to the use of the Jaccard method. The response to review contains the following sentence, "The present workflow implements a similar functionality in R and allows for integration in automated analysis pipelines." Why not just edit or expand on this, specifying that it refers to Enrichment Map, and insert into main text?

2. The author has not clarified whether "umbrella" is meant as a synonym for hierarchy or if it is describing some other structure. As the author does note that a collection need not have an explicitly defined hierarchy for this method to be useful, how would a potential user determine if such an "umbrella" structure was present and whether or not a given gene set collection would be a good candidate for this approach?

3. In discussing the Jaccard cut off, the author mentions the general properties preferred in an optimal network, but it is still not clear how a user would robustly evaluate their choice of Jaccard coefficient. Is it sufficient to simply assure that there are at least two distinct components in the graph or is another heuristic more appropriate? Again, some indication of quantitation would be

helpful to the user.

4. The author notes that additional validation with respect to different parameter settings and input datasets is planned for future work. We are not clear why the author cannot indicate how the results in Figure 5 recapitulate known biology in the airway dataset or if they would help the user draw new conclusions. It is unclear from the manuscript whether or not these clusters are in fact useful for further biological interpretation downstream of the proposed method. Wouldn't this be the main point of using such a method?

5. With regard to the choice of gene set collection (Hallmarks, KEGG, Reactome, etc.), we would encourage discussion as to the merits or drawbacks of different collections. We note that the Hallmark collection is the most frequently accessed gene set collection in MSigDB and was built using a method that included, in part, clustering of existing gene sets. The paper could be significantly strengthened by comparison of the results of the gene set clustering and prioritization based on enrichment scores in RICHNET to the existing Hallmark collection. We also note that collections of gene sets used to create each individual Hallmark gene set are available on MSigDB.

6. The author notes that a continuous coloring scheme for output visualizations could be developed using the negative log10 of the p-values. This is certainly possible, but the author has not addressed our concerns involving the coarse, discrete categorization of gene sets as "Up" or "Down" when many popular algorithms for calculating continuous enrichment scores exist. This design decision should be justified in a bit more detail.

7. Technical problems in running the code: The author has fixed the original bugs in the analysis. However, we encountered the following technical issues when running the new version of the R markdown notebook: a.The EGSEA package is not included in the setup code. Thus, if a user did not have it previously installed, then an error would be triggered later in the analysis. b. The R markdown notebook itself as accessed from the DOI contains notes to comment or delete code "for real submission". We are therefore uncertain if we are running the correct code for this analysis. c. We encountered the error "Error in `diag<-`(`*tmp*`,value=0): only matrix diagonals can be replace" at line 185 of the notebook and were unable to continue with the analysis. There is currently a high burden on the users of this code to identify and solve package dependencies and other errors whose origins are not immediately clear to those not familiar with the underlying packages. These concerns should be addressed to make the workflow distribution more robust and user-friendly.

References

1. Merico D, Isserlin R, Stueker O, Emili A, Bader GD: Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.*PLoS One*. 2010; **5** (11): e13984 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Version 1

Reviewer Report 10 April 2019

https://doi.org/10.5256/f1000research.19488.r45393

Rita Casadio
Biocomputing Group, CIG, Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna, Italy

The authors present an efficient and very useful method for gene set analysis. Their approach is network based and facilitates retrieval of sets of genes that are highly connected. Grouping is due to common context (molecular pathways, biological function, tissue localisation). Furthermore, this procedure facilitates gene enrichment and the identification of singletons or poorly connected islands.

The RICHNET workflow stands as a user friendly workflow that can be easily incorporated in automated analysis pipelines.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Functional annotation of protein variants; machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Jul 2019

Michael Prummer, Swiss Institute for Bioinformatics, Zurich, Switzerland

Thank you for reviewing the manuscript and for your very positive comments.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 April 2019

https://doi.org/10.5256/f1000research.19488.r45391

?

Jill P. Mesirov
Institute of Genomic Medicine, University of California, San Diego, San Diego, CA, USA
Alexander Wenzel
University of California, San Diego, San Diego, CA, USA

While the author does address an important general issue of downstream analysis of gene set enrichment results, his approach is almost identical to "*Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation*" (Merico *et al.*, 2010[1]). Furthermore, Enrichment Map provides much more information in its visualization including the size of the gene set (via the size of the node); the level of enrichment of the gene sets (via a color gradient from blue for down- to red for up-regulation); as well as the extent of overlap between gene sets (via the thickness of the edge connecting gene sets using the Jaccard index as a weight). Even considering the manuscript independent of this issue, there are a number of other concerns with the study (some minor but some more major) that require attention from the author and a major revision of the manuscript. They are listed more or less in the order that the issue appears in the manuscript.

1. The author claims that gene set analysis "often results in finding hundreds of differentially regulated pathways". What is the basis for this claim? While it is true that this may occur – it is also true that there may also be no statistically significantly differentially regulated pathways. Has he done experiments on a large number of benchmark data sets to establish this claim of "often"? Moreover, this may be a result of the particular gene set collection used for the analysis. Is it a function of the particular data set or enrichment approach he is using (CAMERA)? We note here that the author never describes the measure CAMERA uses for enrichment nor how up/down is determined. All these open issues should be addressed.

2. The use of "umbrella" organization is somewhat unclear – do you mean hierarchy? Note also the need for a hierarchy of the significantly activated or repressed pathways may result from the author's choice of gene set collection, which may not provide a significant or specific enough result. This point should be addressed.

3. The author provides little explanation for the arbitrary choice of Jaccard Index > 0.2 as the connectivity threshold. Additionally, as the edges in the networks produced by this method do not represent any other kind of continuous data, has the author considered weighting the edges by the

Jaccard index in the visualization to allow the user to see the effect of various thresholds? In fact, Enrichment Map (cited above) uses the Jaccard Index itself to determine the weight of the line connecting two overlapping gene sets. These approaches need to be compared and the choice of threshold better explained and justified.

4. Prioritizing and ranking the active pathways/gene sets by number of gene sets in a network hub and degree of connectivity seems inadequate if one is looking for biological insights. A biological measure of prioritization would be preferable which includes the levels of activation. Often a study is not done in a vacuum and so in fact some signals may be expected – this may serve as an additional measure for prioritization and even validation of the method. This issue is not addressed and should be and some validation of the results of the only test set analysed should be supplied.

5. It is important to include and describe the numerous network-based enrichment methods that have been published and specifically Enrichment Map (cited above) which is a plugin to both Cytoscape and GSEA and does exactly what the author is describing (downstream analysis of gene set enrichment results) but is never mentioned or referenced. At the very least its performance should be compared to the author's method. This is a requisite for any newly proposed method.

6. The goal of this method is to "focus is on supporting scientists in result interpretation by bringing order into the list of differentially regulated gene sets based on biological rather than pure statistical arguments." But one might ask why do this based on network measures of association and the visualization is very sparse in what it represents as noted above. Also, if this is the goal – it is incumbent on the author to show why this works better than other existing methods (see above) or even using a more sensitive/specific collection of gene sets, e.g., the Hallmark collection in MSigDB (Liberzon *et al.*, 2015[2]). This provides a collection of sets for which essentially this work has been done with additional biological curation. Comparison with other methods and with the use of other gene set collections should be included.

7. The author's example on the "airway" data set employs an analysis with KEGG, Reactome, and BioCarta. Using these 3 databases together means there will be a large amount of redundancy and overlap in the gene sets he uses – again it would be important to compare his results using these collections against the Hallmark collection. Furthermore – why just the results from one data set? The method would be better tested against multiple data sets – some where the signal is very strong and some where the signal is weaker.

8. The author uses the CAMERA method for testing gene sets. This method produces only a binary "Up" or "Down" measurement of enrichment which is used to color the nodes in the resulting network visualizations. This is a very coarse way of testing gene sets. The interpretability of the network visualizations could be improved by using a method such as GSEA, which gives a continuous enrichment score, and coloring the nodes with a gradient to compare degrees of up- or down-regulation and weighting the edges as is done in Enrichment map as noted above.

9. Finally – after application of RICHNET the author only describes and discusses the nature of his resulting networks. There is no discussion that we could see of the biological insights gained, how realistic they were, whether they recapitulated known signals in the data set, etc.

10. **Technical concerns with the code as presented:** The analysis fails with an unintuitive error ("could not coerce net to a network object") if the library "Intergraph" is not installed. While this library is listed in the sessionInfo() printout in the manuscript, this library should be included in the first library() cell of the notebook since its absence is not immediately obvious given this error.

References

1. Merico D, Isserlin R, Stueker O, Emili A, Bader GD: Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.*PLoS One*. 2010; **5** (11): e13984 PubMed Abstract | Publisher Full Text
2. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P: The Molecular Signatures Database (MSigDB) hallmark gene set collection.*Cell Syst*. 2015; **1** (6): 417-425 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
No

Is the description of the method technically sound?
No

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 08 Jul 2019

Michael Prummer, Swiss Institute for Bioinformatics, Zurich, Switzerland

Thank you for reviewing our manuscript and for your critical comments. Below are point-by-point responses to the individual comments.

The remark on the similarity of this work with prior work by others is fully justified and is discussed in the manuscript. The mentioned work by Merico *et al.* is included as reference 6.

The open nature of this workflow script allows for a straightforward implementation of any

additional graphical feature should the user of this workflow wish to do so.

By using the term "often" I was hoping to emphasize that while the statement is true for more than a small number of cases there may be cases where the statement is wrong. How often the statement is true or whether it depends on the choice of geneset or enrichment algorithm or FDR cutoff is less important as long as there exist enough cases for which the workflow presented here may be useful.

Most geneset collections consist of largely overlapping sets which describe very similar processes or properties. The workflow presented here was made for such cases. For the workflow to be useful, an explicit hierarchy, such as with GO-terms, is not required.

The network as constructed here has indeed edges weighted by the Jaccard index. Visualization by varying edge thickness has proven to be difficult to discriminate in my hands. The open nature of this workflow script should allow for an easy implementation of this feature should a user wish to do so.
To address the choice of the Jaccard index cutoff the following sentence has been added at the end of section **Network construction**: “As a guide for finding a reasonable threshold a broad distribution of disjoint cluster sizes is desired. Network analysis does not help if the cutoff is too large (no connections) or too small (all sets are connected with each other).”

One measure of quality of the network construction and community detection is the semantic purity of the clusters. This can be easily seen for the binary systems (and is discussed there) and to a lesser extent for the larger clusters. Whether biological insight can be gained depends strongly on the choice of the genesets in relation to the field of study. RICHNET provides a means of grouping and organizing results and does not generate them.

Enrichment Map is indeed very similar and a perfectly fine interactive tool (see reference 6 of the manuscript). The present workflow implements a similar functionality in R and allows for integration in automated analysis pipelines.

Many of our clients find the hallmark geneset too generic and prefer more diverse geneset collections, such as, KEGG or Reactome.

Your suggestion to include more test results with respect to different geneset libraries, Jacquard score cutoffs, community detection algorithms, and datasets is well received. We are planning to do this in the future.

Using the negative log of the enrichment p-value as enrichment score has been demonstrated in numerous cases in the literature and replacing the categorical colour scale by a continuous one is possible.

Biological interpretation of the geneset clusters produced here is beyond the scope of this work.

The package **intergraph** was included in the library installation and loading part.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 March 2019

https://doi.org/10.5256/f1000research.19488.r44250

?

Monther Alhamdoosh
[1] CSL Limited, Parkville, Victoria, Australia
[2] Bio21 Institute, The University of Melbourne, Parkville, Victoria, Australia

Summary

This workflow paper proposes an interesting approach for prioritizing gene sets in gene set analysis by utilizing network-based algorithms. The author presented a supposedly working code that is reproducible. They proposed focusing on communities of gene sets in order to identify gene sets that are more relevant to the conditions under study. Nonetheless, they show how to explore singletons and binary systems in the list of significant gene sets. Moreover, natural language processing methods were used to annotate the clusters of gene sets. I found this particular extension very useful to summarize gene sets analysis results. However, I didn't get that far in running the code. Please see my comments below for improvements.

Major comments

- Add code to install pre-requisite packages. In my case I had to run the following command to obtain missing packages:

install.packages(c("cowplot", "ggrepel", "kableExtra", "igraph", "GGally", "wordcloud", "tm", "SnowballC"))

- Add code to install the airway experiment package BiocManager::install("airway", version = "3.8")
- I managed to reproduce the analysis up to the point of generating the first network. First, it was required to install the c("network", "sna", "scales") packages, which was not explained in the text. Then, error raised while invoking ggnet2:

Error in ggnet2(net, size = 2, color = "Direction", palette = palette, : could not coerce net to a network object

- Not sure whether related to the version of R. This is my R session:

R version 3.5.0 (2018-04-23)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: RHEL

Matrix products: default
BLAS:****ps/r/3.5.0/lib64/R/lib/libRblas.so
LAPACK: ****/apps/r/3.5.0/lib64/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8 LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8 LC_NAME=C LC_ADDRESS=C
[10] LC_TELEPHONE=C LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel stats4 stats graphics grDevices utils datasets methods base

other attached packages:
[1] airway_1.2.0 SnowballC_0.6.0 tm_0.7-6
[4] NLP_0.2-0 wordcloud_2.6 org.Hs.eg.db_3.7.0
[7] AnnotationDbi_1.44.0 limma_3.38.2 DESeq2_1.22.1
[10] SummarizedExperiment_1.12.0 DelayedArray_0.8.0 BiocParallel_1.16.0
[13] matrixStats_0.54.0 Biobase_2.42.0 GenomicRanges_1.34.0
[16] GenomeInfoDb_1.18.1 IRanges_2.16.0 S4Vectors_0.20.1
[19] BiocGenerics_0.28.0 GGally_1.4.0 igraph_1.2.4
[22] kableExtra_1.0.1 knitr_1.20 reshape2_1.4.3
[25] ggrepel_0.8.0 cowplot_0.9.4 gplots_3.0.3
[28] ggplot2_3.1.0 RColorBrewer_1.1-2

loaded via a namespace (and not attached):
[1] bitops_1.0-6 bit64_0.9-8 webshot_0.5.1 httr_1.3.1
[5] rprojroot_1.3-2 tools_3.5.0 backports_1.1.2 R6_2.3.0
[9] rpart_4.1-13 KernSmooth_2.23-15 Hmisc_4.1-1 DBI_1.0.0
[13] lazyeval_0.2.1 colorspace_1.4-0 nnet_7.3-12 withr_2.1.2
[17] tidyselect_0.2.5 gridExtra_2.3 bit_1.1-14 compiler_3.5.0
[21] rvest_0.3.2 htmlTable_1.12 network_1.14-377 xml2_1.2.0
[25] slam_0.1-45 caTools_1.17.1.1 scales_1.0.0 checkmate_1.8.5
[29] readr_1.1.1 genefilter_1.64.0 stringr_1.3.1 digest_0.6.18
[33] foreign_0.8-71 rmarkdown_1.10 XVector_0.22.0 base64enc_0.1-4
[37] pkgconfig_2.0.2 htmltools_0.3.6 htmlwidgets_1.3 rlang_0.3.0.1
[41] rstudioapi_0.8 RSQLite_2.1.1 bindr_0.1.1 statnet.common_4.2.0
[45] gtools_3.8.1 acepack_1.4.1 dplyr_0.7.8 RCurl_1.96-0
[49] magrittr_1.5 GenomeInfoDbData_1.2.0 Formula_1.2-3 Matrix_1.2-15
[53] Rcpp_1.0.0 munsell_0.5.0 stringi_1.2.4 yaml_2.2.0
[57] zlibbioc_1.28.0 plyr_1.8.4 grid_3.5.0 blob_1.1.1
[61] gdata_2.18.0 crayon_1.3.4 lattice_0.20-38 splines_3.5.0
[65] annotate_1.60.0 hms_0.4.2 sna_2.4 locfit_1.5-9.1
[69] pillar_1.3.0 geneplotter_1.60.0 XML_3.99-0 glue_1.3.0
[73] evaluate_0.12 latticeExtra_0.6-28 data.table_1.11.8 gtable_0.2.0
[77] purrr_0.2.5 reshape_0.8.8 assertthat_0.2.0 xtable_1.8-3
[81] coda_0.19-2 viridisLite_0.3.0 survival_2.43-1 tibble_1.4.2
[85] memoise_1.1.0 bindrcpp_0.2.2 cluster_2.0.7-1

- There are 58 gene sets that are significant in both the uni-directional and bi-directional tests. The former was assumed to be the determinant of the direction of the gene set, which is a bit confusing here. It would be good if the author could elaborate on the reasoning behind assigning the directional status.
- It would be good to highlight the relationship between the identified clusters and annotations, and the underlying experimental conditions.

Minor comments

- EGSEAdata experiment package can be used to load huge number of gene sets including MSigDB gene sets. This is an example:

```
library(EGSEAdata)
library(EGSEA)
gset = buildMSigDBIdx(entrezIDs = res$entrezgene, species = "human", geneSets = "c2", min.size = 3)
idx = gset$c2@idx
gs.libs = sapply(names(idx), function(x) strsplit(x, "_")[[1]][1])
idx = idx[which(gs.libs %in% c("KEGG", "REACTOME", "BIOCARTA"))]
```

Please see this workflow paper for more information[1].

- The workflow is tailored towards the camera analysis. What if users want to use this workflow for other GSA methods? A more generic example is needed.
- Any possibility to wrap the network analysis code and make it easier for users to invoke as a function? Probably, an R package named RICHNET can be developed along with this wonderful workflow.

References

1. Alhamdoosh M, Law CW, Tian L, Sheridan JM, Ng M, Ritchie ME: Easy and efficient ensemble gene set testing with EGSEA.*F1000Res*. 2017; **6**: 2010 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics and AI. Developed gene set analysis method, which is widely used by the research community.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 08 Jul 2019

Michael Prummer, Swiss Institute for Bioinformatics, Zurich, Switzerland

Thank you for reviewing our manuscript and for your constructive comments. Below are point-by-point responses to the individual comments.

The missing packages were included in the initial package installation and loading part.

The source of the error in ggnet2 is fixed.

Uni- and bi-directional tests are preformed in parallel. Afterwards, priority is given to the uni-directional test as it is more stringent (either up- or down-regulated genes). The bi-directional case is biologically meaningful as well as an upregulation of and inhibitory member of a pathway has the same effect as the downregulation of an activatory member. But all these details are related to one particular choice of enrichment analysis whereas the extent of this work starts after a list of candidate genesets was generated by any appropriate method. Biological interpretation of the geneset clustering results produced here is beyond the scope of this work.

A code chunk using EGSEAdata to build the geneset library was included in the manuscript.

The following sentence is included in paragraph 2 of the discussion: “One advantage over other approaches is that it does not depend on a particular geneset library or geneset analysis method. Any means of selecting genesets of interest can be used.”

Competing Interests: No competing interests were disclosed.

Reviewer Report 07 March 2019

https://doi.org/10.5256/f1000research.19488.r44249

Kimberly Glass
[1] Department of Medicine, Harvard Medical School, Boston, MA, USA
[2] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

In "Enhancing gene set enrichment using networks" the author describes a pipeline to visualize the gene sets associated with a particular differential-expression analysis as a network. In this network connections between gene sets are based on common/shared gene annotations. The paper is clearly written and the decisions made in the pipeline are reasonable. There are only a few points on which I would like to see more discussion:

1. In the introduction the authors mention the Jaccard index as a similarity measure (alongside coexpression of genes in WGCNA). There are many similarity measures: hamming distance, cosine similarity, Fisher's exact test, as well as many other measures for continuous variables that can easily be adapted to binary variables. Which one of these is used to construct the gene-set

network could impact the structure of the network. The pros and cons of using the Jaccard index as well as other similarity measures warrants more discussion (e.g. some may better capture relationships between terms with many gene annotations, and some may better capture relationships between terms with fewer gene annotations).

2. GO terms are structured as a DAG, with genes annotated to child terms propagating to parent terms. This underlying structure will impact the structure of the similarity network between GO terms, and is worth pointing out in the manuscript.

3. In the introduction, the authors state "the clusters can be categorized as....medium and large or dense and loose clusters". The author should either include more discussion about how these can be quantified (i.e. what is a "medium" cluster) or they should illustrate the quantification of these categories in their example.

4. The equation “J=NumberofcommongenesNumberofallgenes” looks mis-formatted. It also would be better presented as: “J = intersect(set A, set B)/union(set A, set B)”.

5. The authors should also consider the structure of a general gene-set network (one not restricted to gene-sets associated with differentially-expressed genes). It is possible that the singletons/doublets that the authors remove may simply come from a sparser area of this "general" gene-set network (and the clusters a denser area), in which case the pruning step is removing relevant results (while retaining less informative ones that might be picked up by chance).

6. In naming the clusters, I would suggest normalizing the number of instances of a word against its frequency in the entire database. For example "cell" is a much more common word in KEGG/GO term names than "glycolosis". From a biological point of view, if all the terms that contain "glycolosis" are in the same cluster (even if it’s only 1-2 terms), this is likely much more interesting to highlight than if "cell" appears frequently in that cluster (but also in many other terms outside of the cluster).

7. Do the authors have any thoughts about how to interpret clusters with "missing titles" (no word appearing more than once)?

Minor comments:

- Be sure to spell out the GO and KEGG acronym for first usage.
- Some of the longer names in Figure 5 appear truncated (e.g. "cell death signalling via").

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: computational biology, systems biology, network biology, network medicine

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 08 Jul 2019

Michael Prummer, Swiss Institute for Bioinformatics, Zurich, Switzerland

Thank you for reviewing our manuscript and for your constructive comments. Below are point-by-point responses to the individual comments.

While a discussion on the influence of different distance metrics on network structure is interesting it is not in the scope of this work. Sidenote: a discussion on the influence of different community detection algorithms would be interesting as well. These aspects are critical to any attempt at clustering data but it is assumed here that the qualified reader and user of this workflow is aware of it. The issue is briefly mentioned in paragraph four of the discussion.

Avoiding the use of GO terms was an attempt to avoid having to open up the discussion about the influence of their hierarchical structure on results. I am convinced this would make the discussion unnecessarily complicated and that it is better done elsewhere.

The equation defining the Jaccard index is formatted correctly in the PDF.

In the manuscript it is emphasized that nothing is removed or deemed irrelevant and that putting aside singletons and doublets is just a means of sorting. Indeed, an unusually large proportion of singletons may indicate an unexplored area of biology. In such a situation, relying on common knowledge in the form of published genesets may not be the wisest way to go at all.

There are a number of different possibilities to obtain a representative label for a cluster and in this manuscript a relatively simple one was chosen. It may not be the most sophisticated but it is straight forward to understand.

The following text is included in section **Lattice of annotated networks** on *missing titles*: “This may be indicative for a semantically mixed cluster or for sparse prior knowledge.”

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 1

Author Response 08 Jul 2019

Michael Prummer, Swiss Institute for Bioinformatics, Zurich, Switzerland

Response to reader Shubham Choudhury:
Thank you for this suggestion which is integrated in the revised manuscript.

Competing Interests: No competing interests were disclosed.

Reader Comment 07 Feb 2019

Shubham Choudhury,

Please include the package "**intergraph**" in the list of prerequisite functions. It converts the "igraph" object into a "network" object which is given to the "ggnet2" function as an argument.

Competing Interests: No competing interests were disclosed.