IFA:
Integrative Functional Analysis User's Guide

Juan C. Rodriguez[1,2] Elmer A Fernández[1,3]
jcrodriguez@bdmg.com.ar efernandez@bdmg.com.ar
[1] UA AREA CS. AGR. ING. BIO. Y S, Universidad Católica de Córdoba, CONICET, Córdoba, Argentina
[2] Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Córdoba, Argentina
[3] Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba, Argentina

First edition September 19, 2016
Last revised September 21, 2016

This free open-source software implements academic research by the authors and co-workers. If you use it, please support the project by citing the journal article listed in Section 1.

1 Citing IFA

IFA implements a body of methodological research by the authors and co-workers. As citations are the main means by which the authors receive professional credit for their work. Please cite the IFA software package as:

Rodriguez, J. C., González, G. A., Fresno, C., Llera, A. S., Fernández, E. A. (2016). Improving Information Retrieval in Functional Analysis. *Computers in biology and medicine*, in press.

2 R requirement

R (`http://www.r-project.org`) is a language and environment for statistical computing and graphics. We assume R (version 3.1.0 or higher) has been installed in your local machine. The latest version can be installed following instructions below for different platforms (Windows, Mac, and Linux).

- Quick link for Windows: Download R for Windows `https://cran.r-project.org/bin/windows/base/`.
- Quick link for Mac: Download R for Mac OS X 10.6 (Snow Leopard or higher) `https://cran.r-project.org/bin/macosx/`.
- Quick link for Linux: Download R for Linux `https://cran.r-project.org/bin/linux/`.

3 Downloading IFA

IFA source code can be downloaded as a zip file from `https://github.com/jcrodriguez1989/IFA/archive/master.zip`. Unzip this downloaded file, it will be named as "IFA-master.zip" or "master.zip". After this step you will have a folder named "IFA-master". Hereafter, the full path of IFA-master folder will be referred as "IFA_FOLDER", for example "C://Downloads/IFA-master/". Make sure that in your "IFA_FOLDER" you have a folder named "Code".

If you have Linux you can follow these steps to download IFA (lines which start with a "$" sign are meant to be copied and pasted into your terminal, don't copy the starting "$"):

1. Open a Linux terminal.
2. Download IFA:

   ```
   $ wget https://github.com/jcrodriguez1989/IFA/archive/master.zip
   ```

3. Unzip file:

   ```
   $ unzip master.zip
   ```

4. Enter IFA-master folder:

   ```
   $ cd IFA-master
   ```

   Now you are into your "IFA_FOLDER", to check its full path type:

   ```
   $ pwd
   ```

   ```
   /home/jcrodriguez/Downloads/IFA-master
   ```

   This means my "IFA_FOLDER" is */home/jcrodriguez/Downloads/IFA-master*.
   Lets make sure it has a "Code" folder in it.

   ```
   $ ls
   ```

   ```
   Code
   ```

4 R dependencies

In order to get IFA working, the following R libraries must be installed:

- dnet
- limma
- mGSZ
- org.Hs.eg.db

To install them, open R and follow these steps (lines which start with a ">" sign are meant to be copied and pasted into your R terminal, don't copy the starting ">"):

```
> source("http://bioconductor.org/biocLite.R");
> biocLite("limma");
> biocLite("org.Hs.eg.db");
> install.packages("mGSZ");
> install.packages("dnet");
```

5 IFA description

IFA *Integrative Functional Analysis*

Description

Runs an Integrative Functional Analysis of the desired gene sets by means of mGSZ and dEnricher.

Usage

IFA(exprMatrix, classes, genesets=NULL, SEAcutoff=0.01, GSEAcutoff=0.01, br=NULL, treatLfc=0, adjMethod="fdr", pAdjCutOff=0.01, ...);

Arguments

exprMatrix	Gene expression matrix. Genes as rows, samples as columns. Rownames must be EntrezGene IDs.
classes	String vector of classes representing each column from exprMatrix. Must contain exactly two different classes, classes length must be the same as exprMatrix's number of columns.
genesets	List of gene sets, each one a vector of strings (genes). If dEnricher parameters are correctly set in "..." then dEnricher gene sets will be used. If not, then the last Gene Ontology is loaded from org.Hs.eg.db.
SEAcutoff	Gene set enrichment cutoff value for SEA.
GSEAcutoff	Gene set enrichment cutoff value for GSEA.
br	Choosen background reference to use for SEA analysis. It can be a vector of Entrez IDs; or the "BRI", "BRIII" strings in order to automatically load them.
treatLfc	Treat log fold change cutoff for treat function when determining differentially expressed genes.
adjMethod	P-value adjust method passed to p.adjust function to be applied to genes p-value calculation in order to define differentially expressed ones.
pAdjCutOff	Cutoff value to determine differentially expressed genes.
...	Adittional parameters passed to dEnricher function.

Value

mGSZ returns a data.frame object with following columns:

gene.sets	Gene set ID.
SEA_pval	P-value returned by SEA.
SEA_Enriched	Logical indicating wether gene set was enriched or not by SEA.
GSEA_set.size	Gene set genes count.
GSEA_gene.set.scores	Score returned by GSEA.
GSEA_pvalue	P-value returned by GSEA.
GSEA_Enriched	Logical indicating wether gene set was enriched or not by GSEA.
Name	If gene.sets are GO IDs then it returns its name.

6 Sample IFA Session

This is a quick overview of what an IFA analysis might look like. We will run the analysis for the TCGA's breast cancer dataset, the contrast will be Basal vs. Luminal A, and it will be tested over 500 Gene Ontology's gene sets.
Lines which start with a ">" sign are meant to be copied and pasted into your terminal, don't copy the starting ">" or "+".
Note: Don't forget to put your full "IFA_FOLDER" path where it says "IFA_FOLDER", i.e., if your "IFA_FOLDER" is "C://Downloads/IFA-master/" and in the code it says "IFA_FOLDER/Code" then you should put "C://Downloads/IFA-master/Code".

```
> setwd("IFA_FOLDER/Code");
> source("IFA.R");
> load("PaperCode/tcgaInput.RData");
> set.seed(8818);
> names(tcga);
```

```
[1] "M"      "labels"
```

The tcga object is a list which contains the expression matrix (tcga$M) and the vector of classes (tcga$labels).

```
> head(tcga$labels);
```

```
[1] "Basal" "Basal" "Basal" "Basal" "Basal" "Basal"
```

```
> table(tcga$labels);
```

```
Basal  LumA
   86   198
```

We have 86 Basal and 198 Luminal A subjects.

```
> dim(tcga$M);
```

```
[1] 17117   284
```

We have 17,117 genes and 284 subjects in total.
Lets get the Gene Ontology's gene sets, and keep the first 500.

```
> GO <- loadGO()[1:500];
```

This line will start the IFA analysis, where we input the expression matrix, the classes, the 500 gene sets and a treat log fold change of 1.

```
> ifaResults <- IFA(tcga$M, tcga$labels, GO, treatLfc=1);
```

```
[1] "Using own gene sets"
[1] "Starting SEA analysis"
[1] "DE genes 909 of a total of 17117 ( 5.31 %)"
[1] "Using BRI: 7121 genes."
[1] "18 enriched terms"
[1] "Starting GSEA analysis"
[1] "20  enriched terms"
```

Lets see the results for the first 20 gene sets.

```
> head(ifaResults[,c("gene.sets", "GSEA_Enriched", "SEA_Enriched")], n=20);
```

```
    gene.sets GSEA_Enriched SEA_Enriched
1  GO:0000002         FALSE        FALSE
2  GO:0000003         FALSE        FALSE
3  GO:0000012         FALSE        FALSE
4  GO:0000018         FALSE        FALSE
5  GO:0000019            NA        FALSE
6  GO:0000022            NA         TRUE
7  GO:0000023            NA        FALSE
8  GO:0000027            NA        FALSE
9  GO:0000028         FALSE        FALSE
10 GO:0000038          TRUE        FALSE
11 GO:0000041         FALSE        FALSE
12 GO:0000042         FALSE        FALSE
13 GO:0000045         FALSE        FALSE
14 GO:0000046            NA        FALSE
15 GO:0000050         FALSE        FALSE
16 GO:0000052         FALSE        FALSE
17 GO:0000053            NA        FALSE
18 GO:0000054         FALSE        FALSE
19 GO:0000055            NA        FALSE
20 GO:0000056            NA        FALSE
```

Lets see how many gene sets were enriched by both SEA and GSEA, and show them.

```
> ifaResults$Enriched <- rowSums(ifaResults[, c("GSEA_Enriched", "SEA_Enriched")],
+ na.rm=T);
> in_both <- ifaResults[ ifaResults$Enriched == 2, ];
> nrow(in_both);
```

```
[1] 10
```

```
> in_both[, "Name"];
```

```
 [1] "mitotic sister chromatid segregation"
 [2] "DNA replication checkpoint"
 [3] "G1/S transition of mitotic cell cycle"
 [4] "regulation of transcription involved in G1/S transition of mitotic cell cycle"
 [5] "microtubule cytoskeleton organization"
 [6] "mitotic cell cycle"
 [7] "nuclear division"
 [8] "sister chromatid segregation"
 [9] "columnar/cuboidal epithelial cell differentiation"
[10] "epithelial cell maturation"
```

Lets see how many gene sets were enriched by any method, and show them.

```
> in_any <- ifaResults[ ifaResults$Enriched > 0, ];
> nrow(in_any);

[1] 28

> in_any[, "Name"];

 [1] "mitotic spindle elongation"
 [2] "very long-chain fatty acid metabolic process"
 [3] "mitotic sister chromatid segregation"
 [4] "cell cycle checkpoint"
 [5] "DNA replication checkpoint"
 [6] "regulation of cyclin-dependent protein serine/threonine kinase activity"
 [7] "G1/S transition of mitotic cell cycle"
 [8] "regulation of transcription involved in G1/S transition of mitotic cell cycle"
 [9] "G2/M transition of mitotic cell cycle"
[10] "establishment of mitotic spindle orientation"
[11] "meiotic spindle organization"
[12] "microtubule cytoskeleton organization"
[13] "spliceosomal tri-snRNP complex assembly"
[14] "mitotic cell cycle"
[15] "nuclear division"
[16] "DNA strand renaturation"
[17] "sister chromatid segregation"
[18] "antral ovarian follicle growth"
[19] "urogenital system development"
[20] "histamine metabolic process"
[21] "neural crest cell migration"
[22] "serotonin secretion"
[23] "morphogenesis of an epithelium"
[24] "positive regulation of neuroblast proliferation"
[25] "epithelial cell development"
[26] "columnar/cuboidal epithelial cell differentiation"
[27] "glandular epithelial cell differentiation"
[28] "epithelial cell maturation"
```

7 Session Info

```
> sessionInfo();

R version 3.1.2 (2014-10-31)
Platform: i686-pc-linux-gnu (32-bit)

locale:
```

```
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=es_AR.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=es_AR.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=es_AR.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_AR.UTF-8 LC_IDENTIFICATION=C

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

other attached packages:
 [1] GO.db_3.0.0          mGSZ_1.0             ismev_1.41
 [4] mgcv_1.8-12          nlme_3.1-126         MASS_7.3-45
 [7] limma_3.22.7         GSA_1.03             org.Hs.eg.db_3.0.0
[10] RSQLite_1.0.0        DBI_0.3.1            AnnotationDbi_1.28.2
[13] GenomeInfoDb_1.2.5   IRanges_2.0.1        S4Vectors_0.4.0
[16] Biobase_2.26.0       BiocGenerics_0.12.1  dnet_1.0.8
[19] supraHex_1.4.0       hexbin_1.27.1        igraph_1.0.1

loaded via a namespace (and not attached):
[1] ape_3.5          graph_1.44.1     grid_3.1.2       lattice_0.20-33
[5] magrittr_1.5     Matrix_1.2-4     Rgraphviz_2.10.0 tools_3.1.2
```