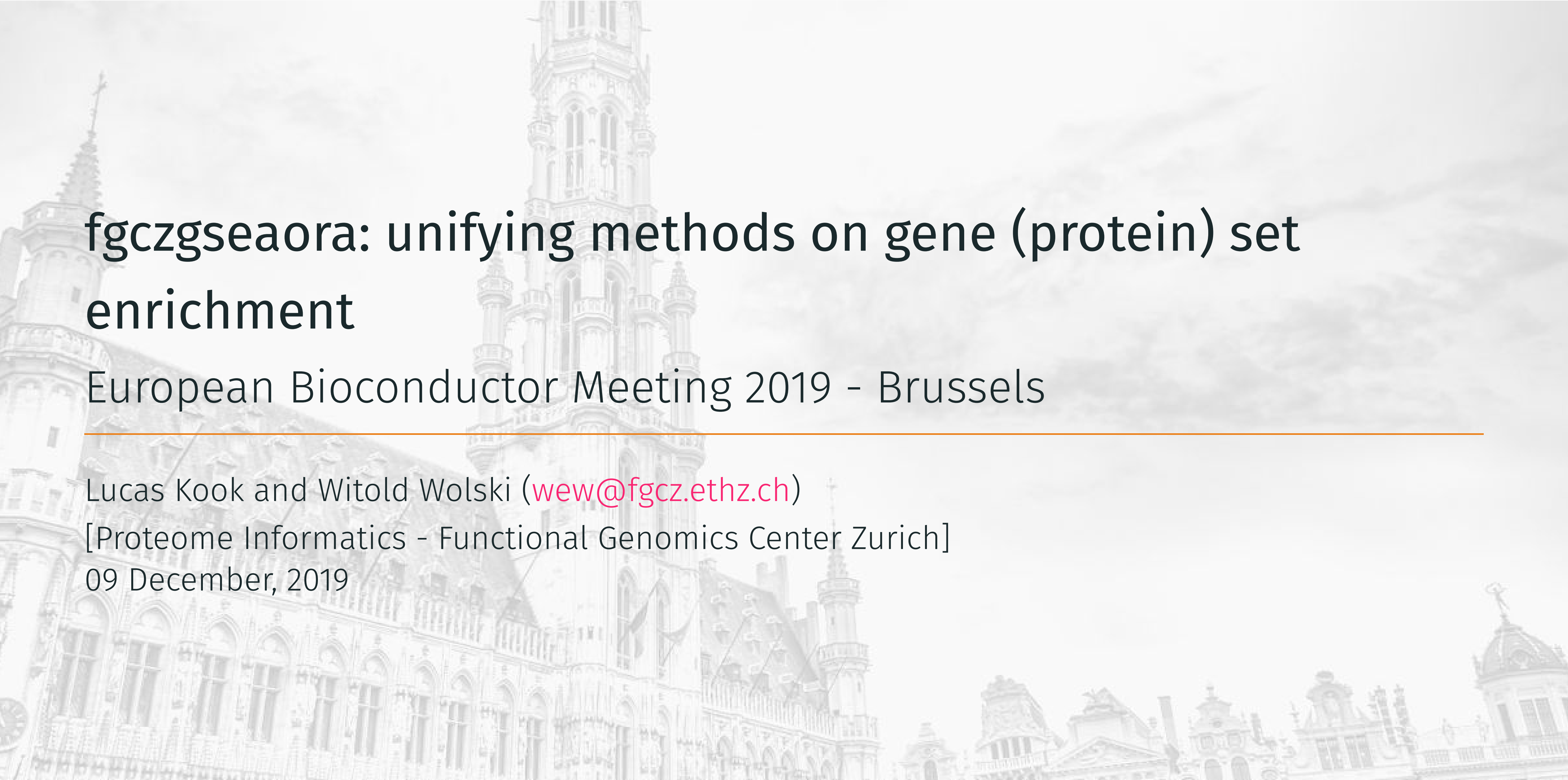


fgczgseaora: unifying methods on gene (protein) set enrichment

European Bioconductor Meeting 2019 - Brussels

Lucas Kook and Witold Wolski (wew@fgcz.ethz.ch)
[Proteome Informatics - Functional Genomics Center Zurich]

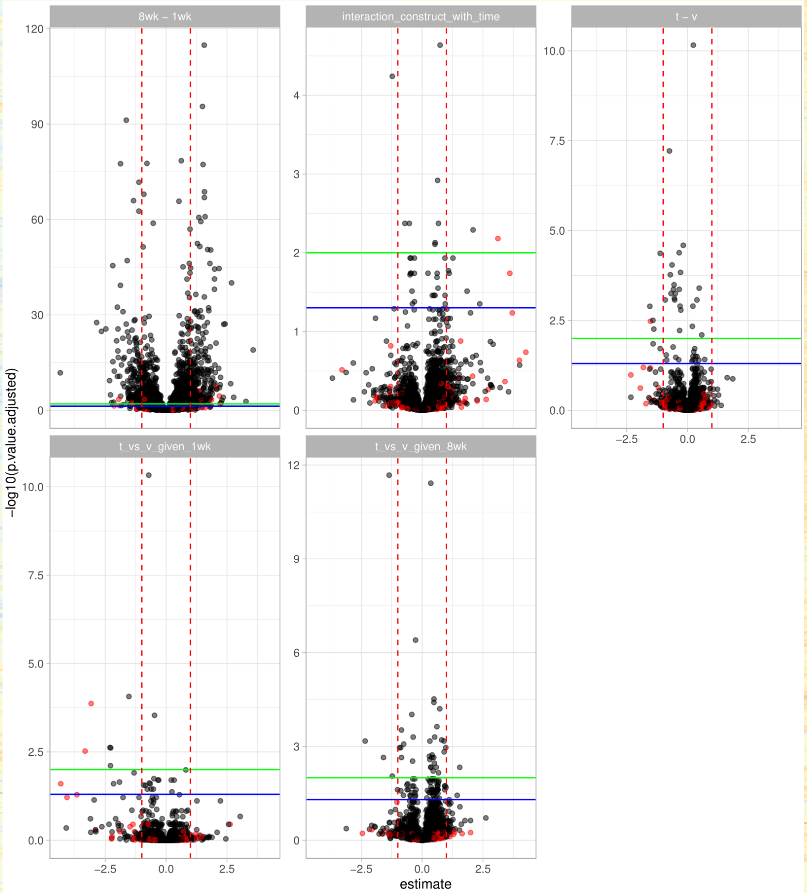
09 December, 2019

Overview

- Pathway analysis for proteomics quantification experiments
- fgczgseaora
- Outlook

Protein quantification experiments

- determine protein foldchanges
  for various contrasts (comparisons of treatments)
- up to thousands of proteins
- only *abundant* proteins quantifed (detection bias)



Pathway analysis

- Over-Represenation Analysis (ORA)
- Gene Set Enrichment Analysis (GSEA)

Pathway analysis uses a priori gene sets that have been grouped together by their involvement in the same biological pathway, or by proximal location on a chromosome. Examples of gene set database are Gene Ontology (GO), KEGG, Reactome and many more.

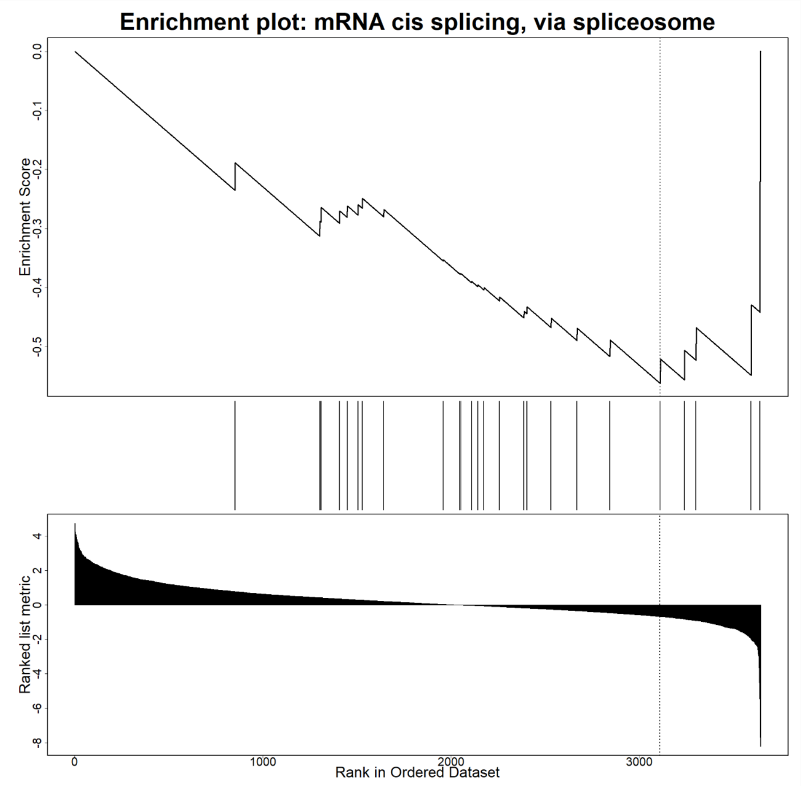
Over-Representation Analysis (ORA)

- Dychotomize list of proteins
  (e.g. using a *threshold* into overexpressed - Yes/No).
- Test if a geneset is *over-represented*
  in on of the sublists
  (e.g. Fischers Exact Test).
- how to choose the threshold?

```
## Pathway GO:0003091

##                Differentially expressed
## GO Term             Yes         No
##   Contained              12          3
##   Not Contained           7         24

## p-value: 0.00034
```

Gene Set Enrichment Analysis (GSEA)

- Ranked list (no threshold required)
- locate genes of genesets in ranked list
- compute enrichment score

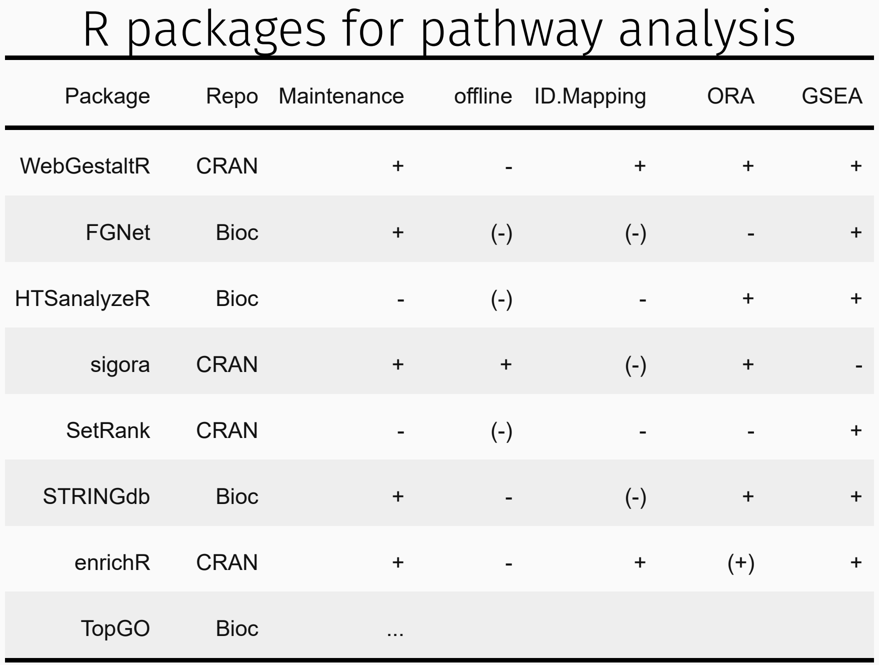


Gene Sets can be highly correlated, because they contain the same proteins. Multiplicity adjustment assumes indpendence (FDR).

fgczgseaora

- Easily generate reports to be delivered to biologists.
- For ORA We can only use tools which allow to specify detection background.
- Map identifiers - support for *sp* identifiers
- Ideally run packages locally
- Provide a similar **R** and command line interface to run ORA GSEA.

Many R packages are available



R packages for pathway analysis

Package	Repo	Maintenance	offline	ID.Mapping	ORA	GSEA
WebGestaltR	CRAN	+	-	+	+	+
FGNet	Bioc	+	(-)	(-)	-	+
HTSanalyzeR	Bioc	-	(-)	-	+	+
sigora	CRAN	+	+	(-)	+	-
SetRank	CRAN	-	(-)	-	-	+
STRINGdb	Bioc	+	-	(-)	+	+
enrichR	CRAN	+	-	+	(+)	+
TopGO	Bioc	...				

- We did integrate:
  - `WebgestaltR` (online only)
  - `sigORA` (offline)

WebgesaltR - Various gene set databases, id mapping, allows for downloading html results. sigORA - uses gene pair signatures. Searches background and pathways for protein pairs unique to a given pathway. By this it decreases the correlation among gene sets.

Common R interface

```r
runWebGestaltGSEA(
  data = dd,
  fpath = "",
  ID_col = "UniprotID",
  score_col =  "estimate",
  organism =  "hsapiens",
  target = "geneontology_Biological_Process",
  nperm = 500,
  outdir = file.path(odir, "WebGestaltGSEA")
)
```

```r
runWebGestaltORA(
  data = dd,
  fpath = "",
  ID_col = "UniprotID",
  score_col =  "estimate",
  organism =  "hsapiens",
  threshold = 1,
  greater = TRUE,
  target = "geneontology_Biological_Process",
  nperm = 500,
  outdir = file.path(odir, "WebGestaltORA")
)
runSIGORA(
  data = dd,
  score_col = "estimate",
  threshold = 1,
  greater = TRUE,
  target = "GO",
  outdir = file.path(odir, "sigORA")
)
```

Command line interface

```
Rscript lfq_multigroup_gsea.R ./foldchange_estimates.xlsx -o hsapiens
Rscript lfq_multigroup_ora.R ./foldchange_estimates.xlsx -t uniprotswissprot
```

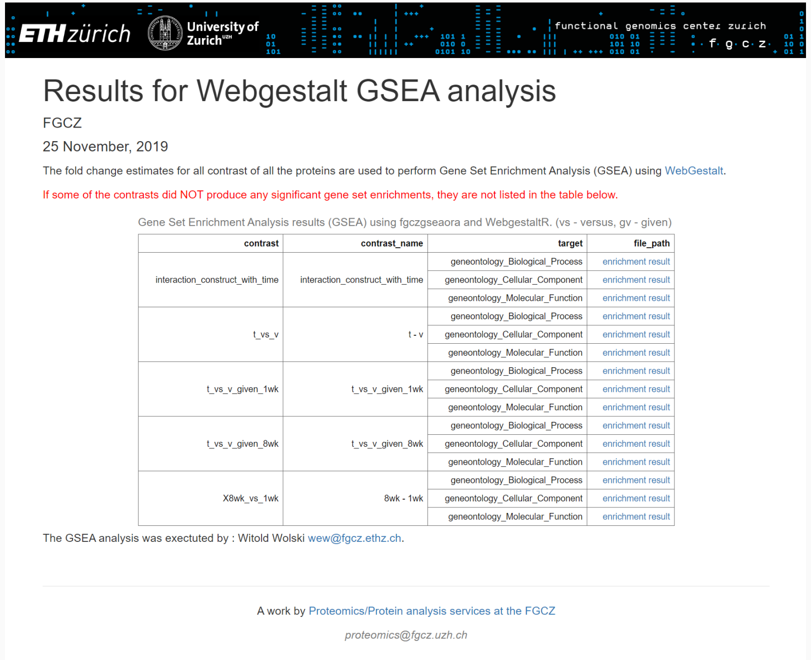
The enrichment methods in this package (ORA, GSEA sigORA) come with a `docopt` based command line tool to facilitate analysing batches of files.

Command line interface

```
"WebGestaltR GSEA for multigroup reports

Usage:
  lfq_multigroup_gsea.R <grp2file> [--organism⩽organism>] [--outdir⩽outdir>] [--

Options:
  -o --organism⩽organism> organism [default: hsapiens]
  -r --outdir⩽outdir> output directory [default: results_gsea]
  -t --idtype⩽idtype> type of id used for mapping [default: uniprotswissprot]
  -i --ID_col⩽ID_col> Column containing the UniprotIDs [default: UniprotID]
  -n --nperm⩽nperm> number of permutations to calculate enrichment scores [defau
  -e --score_col⩽score_col> column containing fold changes [default: pseudo_esti
  -c --contrast⩽contrast> column containing fold changes [default: contrast]

Arguments:
  grp2file  input file
" → doc
library(docopt)
```

HTML outputs - Multiple Contrasts and Targets

- creates folder structure with HTML files visualizing the ORA and GSEA results:
    - For all contrasts
      e.g. t - v, 8wk - 1wk etc.
    - and all selected target
      e.g. GO Bioprocess, GO Molecular Function
- These files are linked from an `index.html`
- can easily be stored and delivered as part of analysis.



Results for Webgestalt GSEA analysis

FGCZ

25 November, 2019

The fold change estimates for all contrast of all the proteins are used to perform Gene Set Enrichment Analysis (GSEA) using WebGestalt.

If some of the contrasts did NOT produce any significant gene set enrichments, they are not listed in the table below.

Gene Set Enrichment Analysis results (GSEA) using fgczgseaora and WebgestaltR. (vs - versus, gv - given)

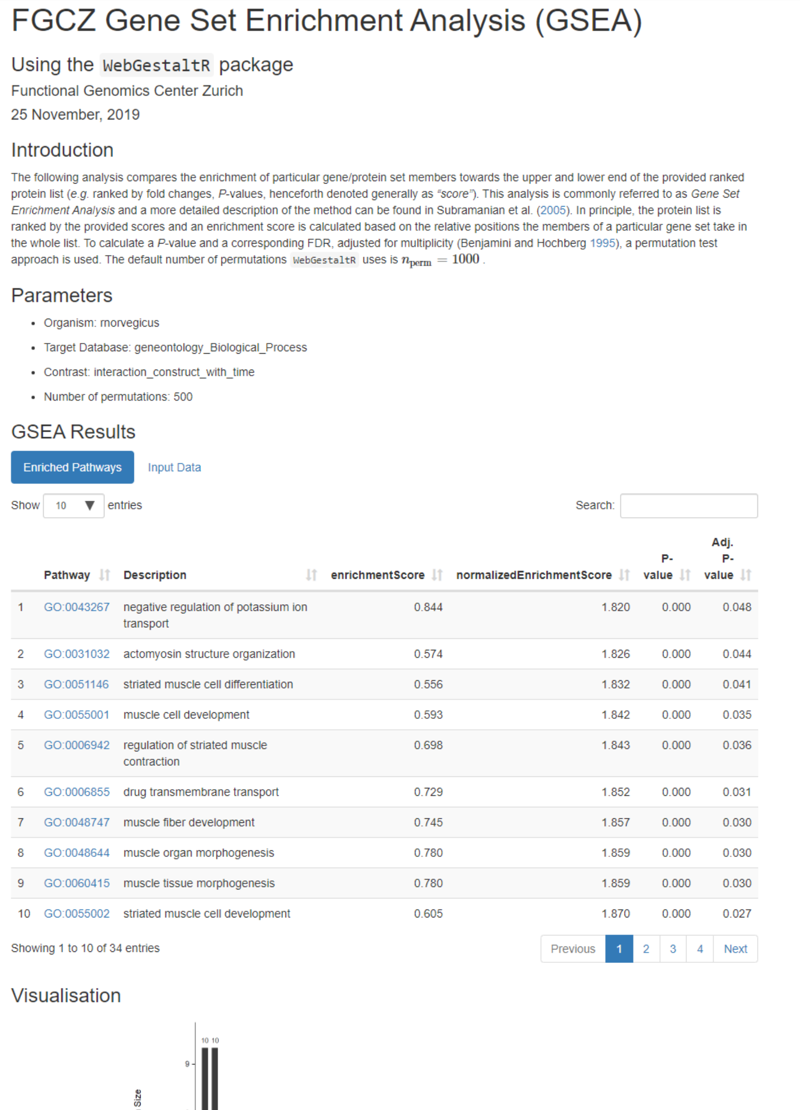
contrast	contrast_name	target	file_path
interaction_construct_with_time	interaction_construct_with_time	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result
		geneontology_Molecular_Function	enrichment result
t_vs_v	t - v	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result
		geneontology_Molecular_Function	enrichment result
t_vs_v_given_1wk	t_vs_v_given_1wk	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result
		geneontology_Molecular_Function	enrichment result
t_vs_v_given_8wk	t_vs_v_given_8wk	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result
		geneontology_Molecular_Function	enrichment result
X8wk_vs_1wk	8wk - 1wk	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result
		geneontology_Molecular_Function	enrichment result

The GSEA analysis was exectuted by : Witold Wolski wew@fgcz.ethz.ch.

A work by Proteomics/Protein analysis services at the FGCZ

proteomics@fgcz.uzh.ch

HTML output - HTML report with method description



FGCZ Gene Set Enrichment Analysis (GSEA)

Using the `WebGestaltR` package

Functional Genomics Center Zurich

25 November, 2019

Introduction

The following analysis compares the enrichment of particular gene/protein set members towards the upper and lower end of the provided ranked protein list (*e.g.* ranked by fold changes, *P*-values, henceforth denoted generally as *"score"*). This analysis is commonly referred to as *Gene Set Enrichment Analysis* and a more detailed description of the method can be found in Subramanian et al. (2005). In principle, the protein list is ranked by the provided scores and an enrichment score is calculated based on the relative positions the members of a particular gene set take in the whole list. To calculate a *P*-value and a corresponding FDR, adjusted for multiplicity (Benjamini and Hochberg 1995), a permutation test approach is used. The default number of permutations `WebGestaltR` uses is $n_{perm} = 1000$.

Parameters

- Organism: rnorvegicus
- Target Database: geneontology_Biological_Process
- Contrast: interaction_construct_with_time
- Number of permutations: 500

GSEA Results

Enriched Pathways | Input Data

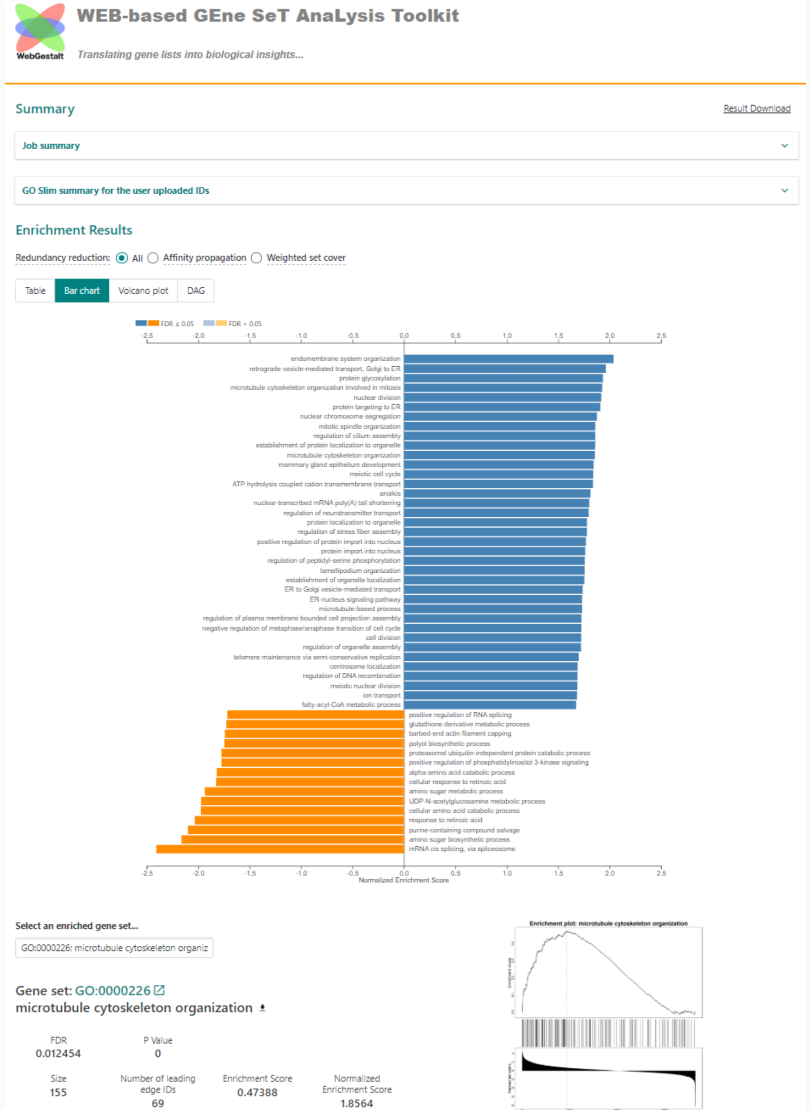
Show 10 ▼ entries | Search:

	Pathway	Description	enrichmentScore	normalizedEnrichmentScore	P-value	Adj. P-value
1	GO:0043267	negative regulation of potassium ion transport	0.844	1.820	0.000	0.048
2	GO:0031032	actomyosin structure organization	0.574	1.826	0.000	0.044
3	GO:0051146	striated muscle cell differentiation	0.556	1.832	0.000	0.041
4	GO:0055001	muscle cell development	0.593	1.842	0.000	0.035
5	GO:0006942	regulation of striated muscle contraction	0.698	1.843	0.000	0.036
6	GO:0006855	drug transmembrane transport	0.729	1.852	0.000	0.031
7	GO:0048747	muscle fiber development	0.745	1.857	0.000	0.030
8	GO:0048644	muscle organ morphogenesis	0.780	1.859	0.000	0.030
9	GO:0060415	muscle tissue morphogenesis	0.780	1.859	0.000	0.030
10	GO:0055002	striated muscle cell development	0.605	1.870	0.000	0.027

Showing 1 to 10 of 34 entries

Previous 1 2 3 4 Next

Visualisation



WEB-based GEne SeT AnaLysis Toolkit

Translating gene lists into biological insights...

Summary

Result Download

Job summary

GO Slim summary for the user uploaded IDs

Enrichment Results

Redundancy reduction: ● All ○ Affinity propagation ○ Weighted set cover

Table
Bar chart
Volcano plot
DAG
FDR ≤ 0.05
FDR > 0.05
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
endomembrane system organization
retrograde vesicle-mediated transport, Golgi to ER
protein glycosylation
microtubule cytoskeleton organization involved in mitosis
nuclear division
protein targeting to ER
nuclear chromosome segregation
mitotic spindle organization
regulation of cilium assembly
establishment of protein localization to organelle
microtubule cytoskeleton organization
mammary gland epithelium development
meiotic cell cycle
ATP hydrolysis coupled cation transmembrane transport
anoikis
nuclear-transcribed mRNA poly(A) tail shortening
regulation of neurotransmitter transport
protein localization to organelle
regulation of stress fiber assembly
positive regulation of protein import into nucleus
protein import into nucleus
regulation of peptidyl-serine phosphorylation
lamellipodium organization
establishment of organelle localization
ER to Golgi vesicle-mediated transport
ER-nucleus signaling pathway
microtubule-based process
regulation of plasma membrane bounded cell projection assembly
negative regulation of metaphase/anaphase transition of cell cycle
cell division
regulation of organelle assembly
telomere maintenance via semi-conservative replication
centrosome localization
regulation of DNA recombination
meiotic nuclear division
ion transport
fatty-acyl-CoA metabolic process
positive regulation of RNA splicing
glutathione derivative metabolic process
barbed-end actin filament capping
polyol biosynthetic process
proteasomal ubiquitin-independent protein catabolic process
positive regulation of phosphatidylinositol 3-kinase signaling
alpha-amino acid catabolic process
cellular response to retinoic acid
amino sugar metabolic process
UDP-N-acetylglucosamine metabolic process
cellular amino acid catabolic process
response to retinoic acid
purine-containing compound salvage
amino sugar biosynthetic process
mRNA cis splicing, via spliceosome
Normalized Enrichment Score

Select an enriched gene set...

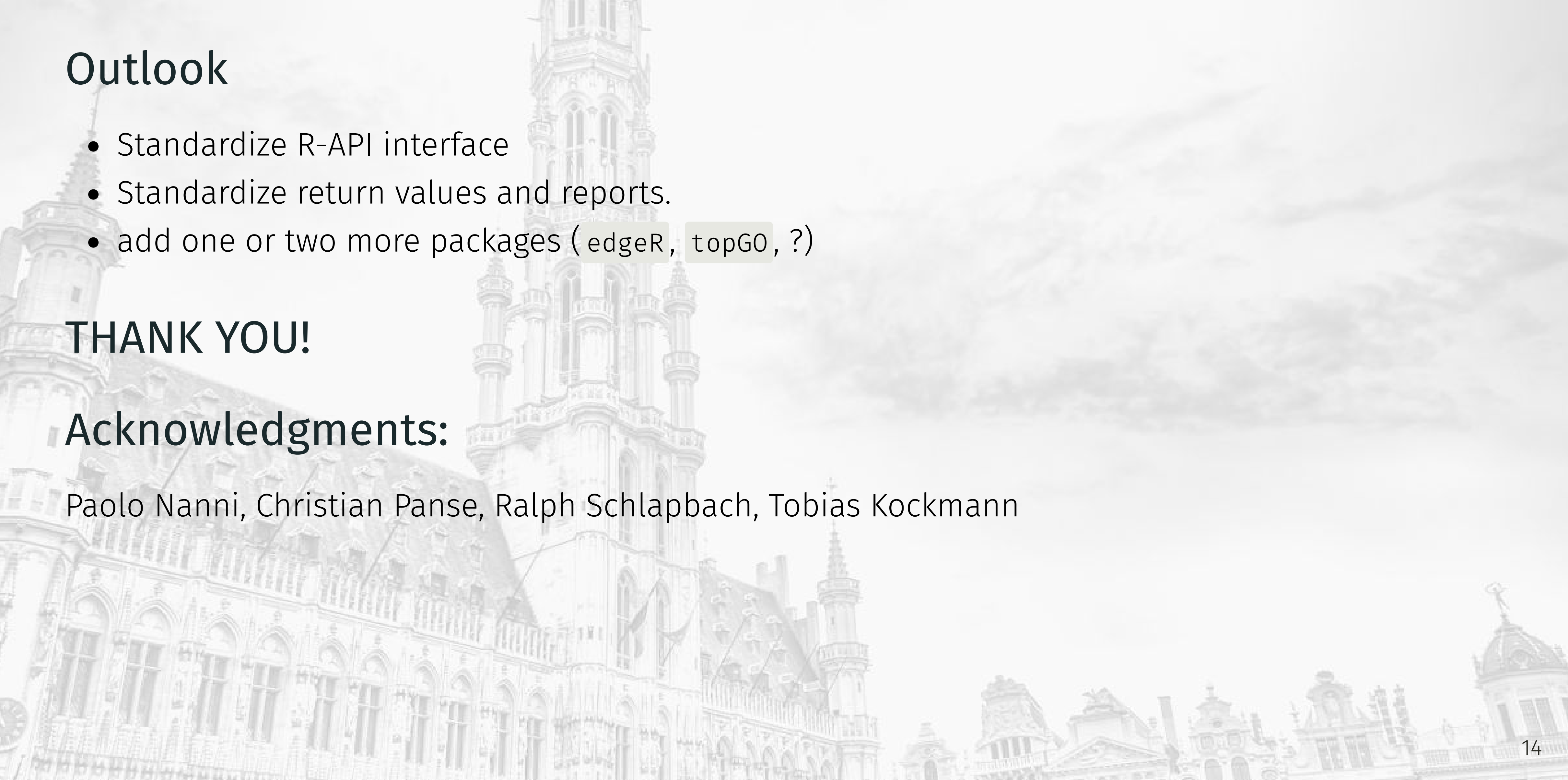
GO:0000226: microtubule cytoskeleton organiz

Gene set: GO:0000226

microtubule cytoskeleton organization

FDR	P Value		
0.012454	0		
Size	**Number of leading edge IDs**	**Enrichment Score**	**Normalized Enrichment Score**
155	69	0.47388	1.8564

Outlook



Outlook

- Standardize R-API interface
- Standardize return values and reports.
- add one or two more packages (`edgeR`, `topGO`, ?)

THANK YOU!

Acknowledgments:

Paolo Nanni, Christian Panse, Ralph Schlapbach, Tobias Kockmann