Gene Set Analysis

Mario Rubio Chavarría

Why GSA?

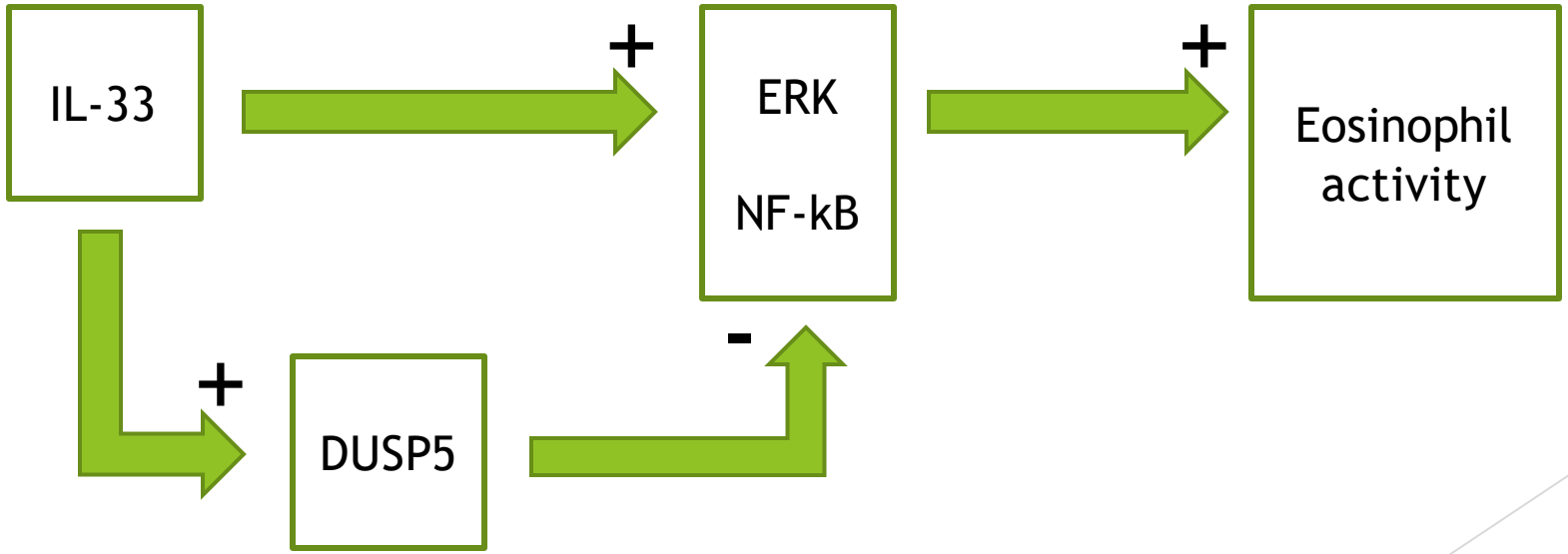
- **Gene Set Analysis (GSA)** is a family of techniques applied in genomics, proteomics, and **transcriptomics** (and poker).

More extensive analyses → More conclusions from the same data

More intensive analyses → Higher evidence

Database

- Taken from:
  Dusp 5 negatively regulates IL-33-mediated eosinophil survival and function. Derek A Holmes, Jung-Hua Yeh, Donghong Yan, Min Xu, Andrew C Chan. The EMBO Journal (2015) 34: 218-235. https://doi.org/10.15252/embj.201489456
- Mechanism studied:



Single-gene analysis
Introduction

- Essentially, a t-test between samples.
- Is the mean different between both populations?
- The traditional approach.

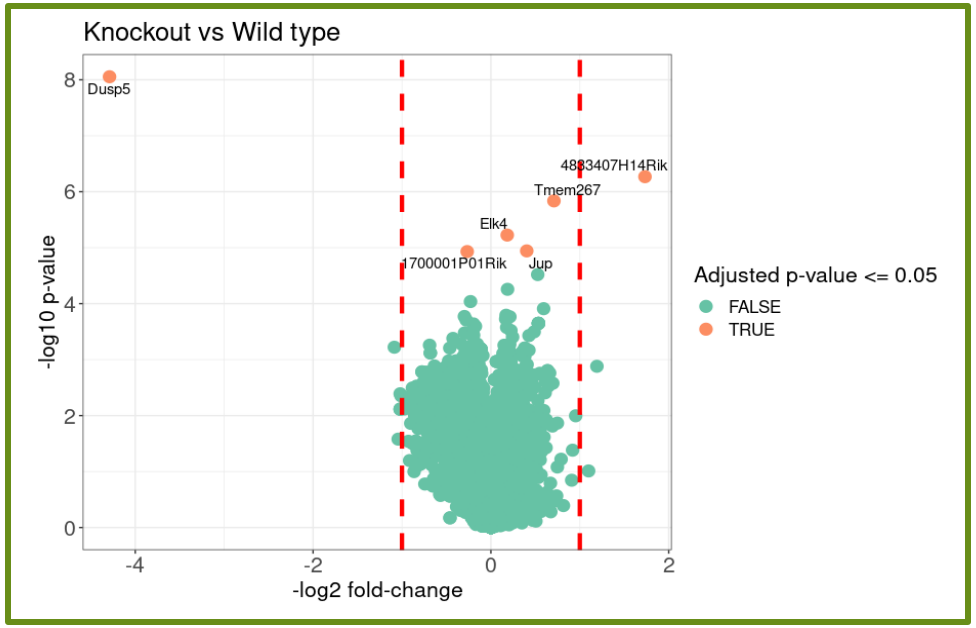
Single-gene analysis
Practice

Repository:

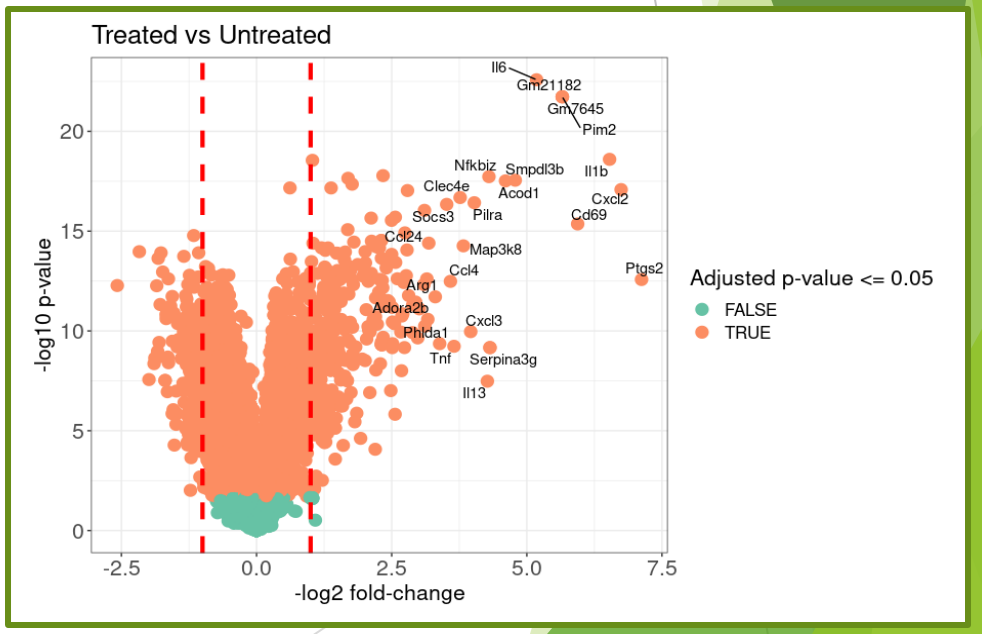
https://github.com/mrubio-chavarria/holmes_analysis

Single-gene analysis
Results

Knockout vs Wild type

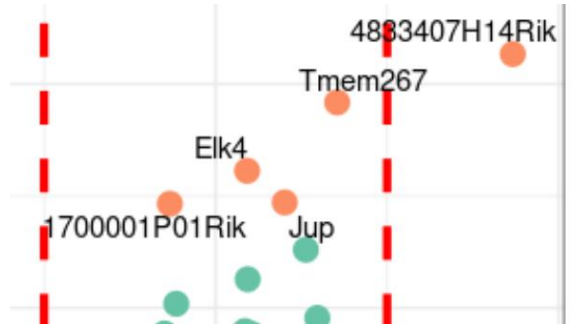


Treated vs Untreated

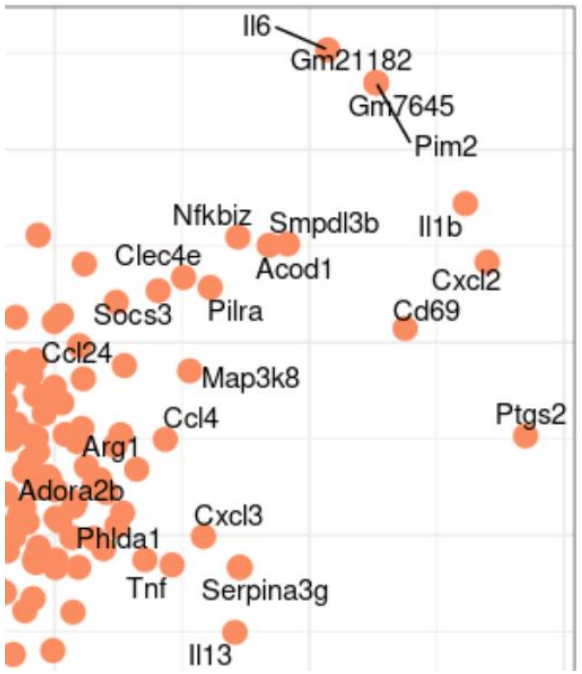


Single-gene analysis
Problems

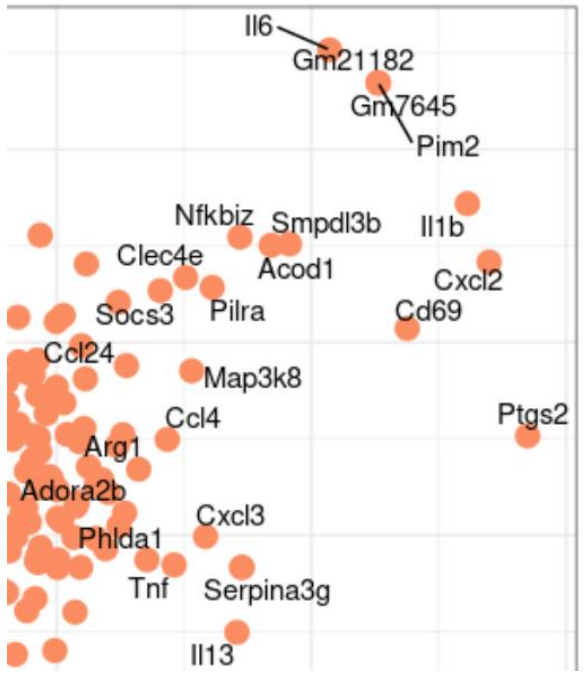
Hard interpretation



Many pathways involved



Information of just one gene



Gene Set Analysis
Introduction

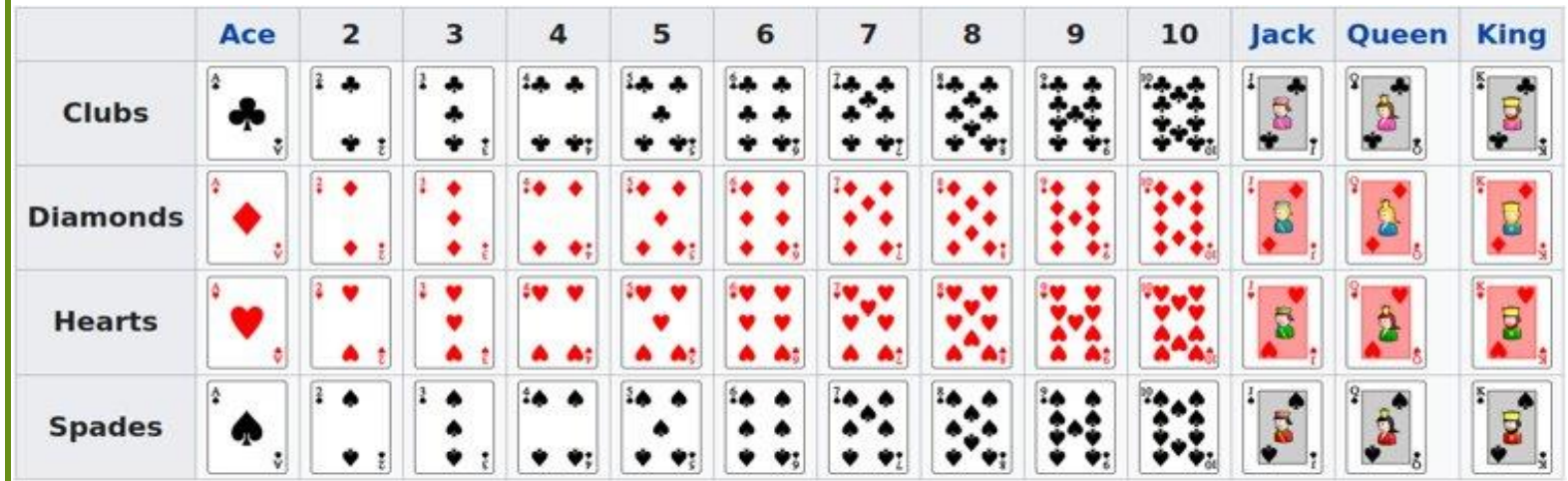
- Analysis based on the group of genes, the gene sets.
- Multiple gene sets to define specific questions.
- Robust analysis that uses the information from all the genes in the dataset.

Gene Set Analysis
Families

Over-representation analysis (ORA)	Functional class scoring methods (FCS)	Topology-based pathway analysis
• First generation. • Popular methods. • Examples: DAVID, Goana.	• Second generation. • More developed methods. • Examples: GSEA, CAMERA.	• Third generation. • Specific applications. • In development.

Over-representation analysis
Introduction

Poker deck[1]



- 13 cards per type.
- 5 cards per hand.
- Replacement: binomial distribution.
- No replacement: hypergeometric distribution

HG(N, K, n) → P[X = x]

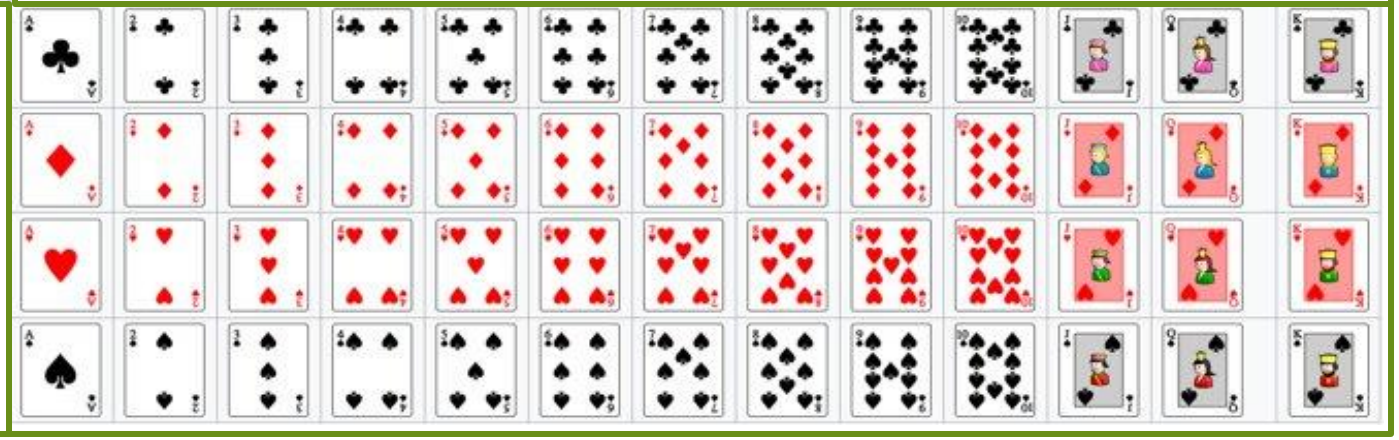
Parameters

- N: number of cards in the deck (52).
- K: number of cards of a given class (13).
- n: number of cards in your hand (5).
- x: number of cards of the given class.

1. Wikipedia

Over-representation analysis

Introduction

Genome	
	Genes
Gene sets	

- 13 cards per type.
- 5 cards per hand.
- Replacement: binomial distribution.
- No replacement: hypergeometric distribution

HG(N, K, n) → P[X = x]

Parameters
• N: number of genes in the genome (tested). • K: number of genes in the gene set. • n: number of all the DE* genes. • x: number of DE genes in the gene set.

*. Differentially expressed.

Over-representation analysis
Practice

Repository:

https://github.com/mrubio-chavarria/holmes_analysis

Over-representation analysis
Practice

Repository:

https://github.com/mrubio-chavarria/holmes_analysis

Drawbacks.

- **Unrealistic assumptions!!!!** → Genes are **not** independent
- All the genes are treated equally. It is only considered information of one gene set at a time.

Functional class scoring methods
Introduction

- There are two types: **univariate** and multivariate methods.
- All the univariate FCS consist of the following steps.
- The explanation is based on **Gene Set Enrichment Analysis** (GSEA)[2].

1	2	3	4
Compute a gene level metric	*Order the genes by the metric*	*Compute a gene set level metric*	*Compute the significance of the gene set metric*
• Metric based on gene expression. • Examples: log_2fold-change or signal-to-noise ratio.	• Ranked list with all the genes classified attending the metric.	• Computed from the ranked list. • Genes of the set should be at the top or botton of the ranked list.	• Obtained from permutation. • Phenotype permutation and gene permutation.

2. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Subramanian A, et al. Proceedings of the National Academy of Sciences Oct 2005, 102 (43) 15545-15550; DOI: 10.1073/pnas.0506580102

Functional class scoring methods
Introduction

- We focus on univariate methods (most used). They consists of the following steps:

1
Compute a gene level metric
• Metric based on gene expression. • Examples: log_2fold-change or signal-to-noise ratio.

Gene-level metric:
Log_2 fold-change

$$Log_2(I_2 - I_1)$$

I_1: average intensity of the genes in the first group (phenotype).

I_2: average intensity of the genes in the second group (phenotype).

Functional class scoring methods
Introduction

- We focus on univariate methods (most used). They consists of the following steps:

2
Order the genes by the metric
• Ranked list with all the genes classified attending the metric.

Ranked list

Gene	Log2 fold-change
Gene 1	3
Gene 2	2
Gene 3	1
…	…
Gene n-2	-1
Gene n-1	-2
Gene n	-3

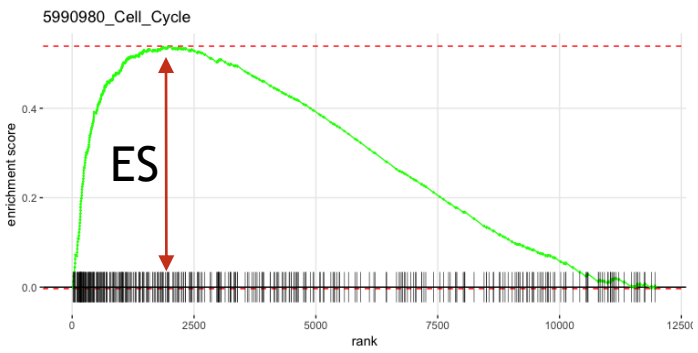
Functional class scoring methods
Introduction

- We focus on univariate methods (most used). They consists of the following steps:

3
Compute a gene set level metric
• Computed from the ranked list. • Genes of the set should be at the top or botton of the ranked list.

Gene set-level metric:
Weighted Kolmogorov-Smirnov statistic

$$S = S \pm \Delta S$$



Ranked list

Gene	Log2 fold-change
Gene 1	3
Gene 2	2
Gene 3	1
…	…
Gene n-2	-1
Gene n-1	-2
Gene n	-3

Functional class scoring methods
Introduction

- We focus on univariate methods (most used). They consists of the following steps:

4
Compute the significance of the gene set metric
• Obtained from permutation. • Phenotype permutation and gene permutation.

Gene expression matrix

Samples

Genes

$$
\begin{matrix}
a_{11} & a_{12} & \cdots & a_{1n-1} & a_{1n} \\
a_{21} & & & & a_{2n} \\
\cdots & & \cdots & & \cdots \\
a_{m-11} & & & & a_{m-1n} \\
a_{m1} & a_{m2} & \cdots & a_{mn-1} & a_{mn}
\end{matrix}
$$

Topology-based pathway analysis
Introduction

- Account for the relationships between pathways.
- For specific purposes (*ad-hoc*).
- Need further study.
- Examples[3]:
  - GSNCA.
  - TopologyGSA.
  - Clipper.

3. Gene Set Analysis: Challenges, Opportunities, and Future Research. Maleki F, et al. *Front. Genet*. June 2020 654 (11); DOI: 10.3389/fgene.2020.00654

Gene Set Enrichment Analysis
Practice

Software:

Algorithm

- GSEA:
  - Gene level: signal-to-noise ratio.
  - Permutation: gene permutation.

Gene Set Enrichment Analysis

Results

GSEA

Enrichment in phenotype: Untreated (10 samples)

- 129 / 292 gene sets are upregulated in phenotype **Untreated**
- 10 gene sets are significant at FDR < 25%
- 5 gene sets are significantly enriched at nominal pvalue < 1%
- 17 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: IL-33_4hr (10 samples)

- 163 / 292 gene sets are upregulated in phenotype **IL-33_4hr**
- 61 gene sets are significantly enriched at FDR < 25%
- 31 gene sets are significantly enriched at nominal pvalue < 1%
- 49 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Gene Set Enrichment Analysis
Results

GSEA

Enrichment in phenotype: Untreated (10 samples)

- 129 / 292 gene sets are upregulated in phenotype **Untreated**
- 10 gene sets are significant at FDR < 25%
- 5 gene sets are significantly enriched at nominal pvalue < 1%
- 17 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

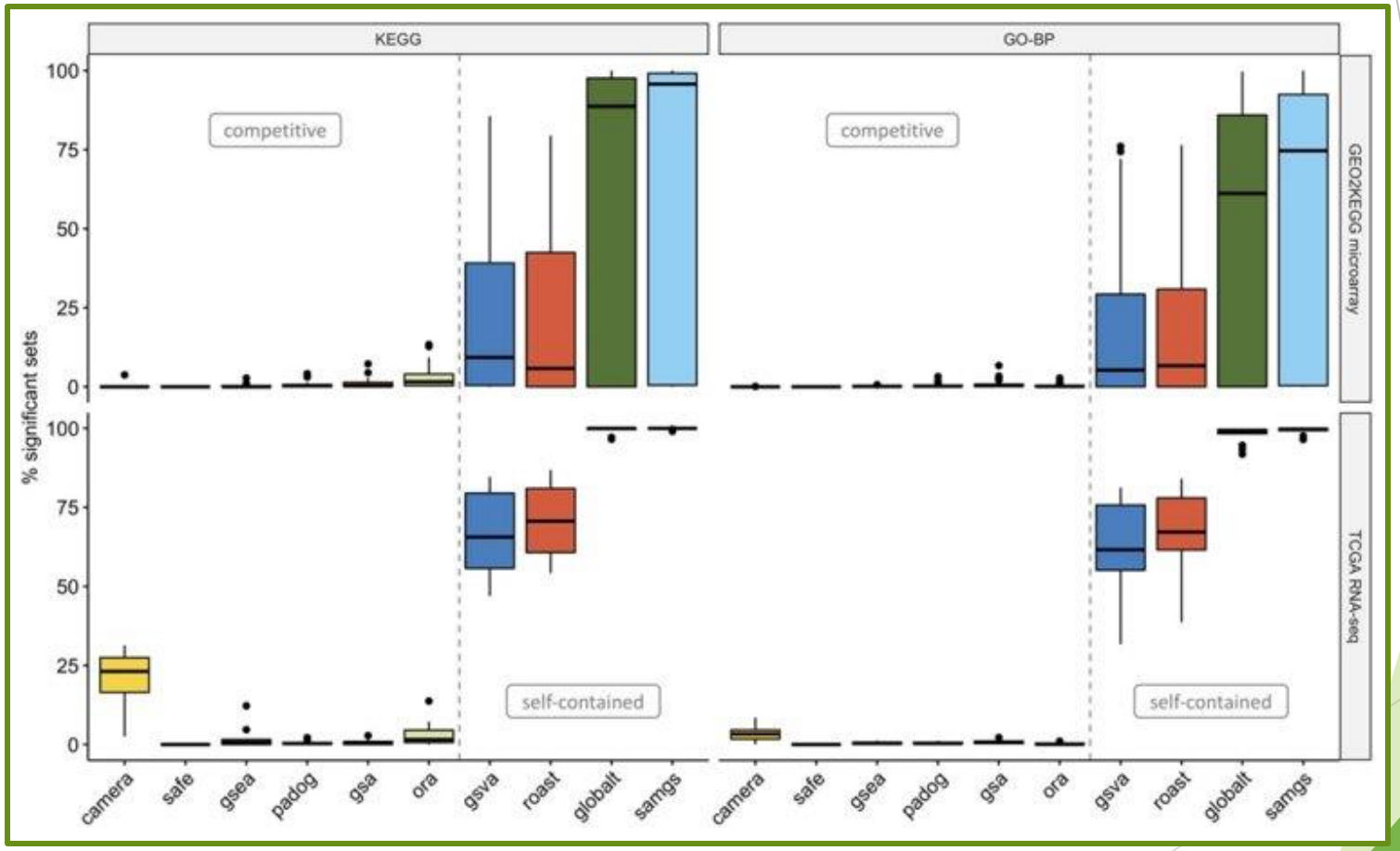
Enrichment in phenotype: IL-33_4hr (10 samples)

- 163 / 292 gene sets are upregulated in phenotype **IL-33_4hr**
- 61 gene sets are significantly enriched at FDR < 25%
- 31 gene sets are significantly enriched at nominal pvalue < 1%
- 49 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Drawbacks.

- **Unrealistic assumptions!!!!** → Gene sets are **not** independent
- Complicated.

To conclude
Benchmark



4. Geistlinger L, et al. Toward a gold standard for benchmarking gene set enrichment analysis, Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 545–556, DOI: 10.1093/bib/bbz158