Popularity and Performance of Bioinformatics Software - The Case of Gene Set Analysis

Chengshu Xie, Shaurya Jauhari, and Antonio Mora*

School of Life Sciences, Guangzhou Medical University and Guangzhou Institutes of Biomedicine and Health (Chinese Academy of Sciences), China

INTRODUCTION

Gene Set Analysis (GSA) consists on comparing a query gene set (a list or a rank of differentially expressed genes, for example) to a reference database, using a particular statistical method, in order to interpret it as a rank of significant pathways, functionally related gene sets, or ontology terms [1].

GOALS

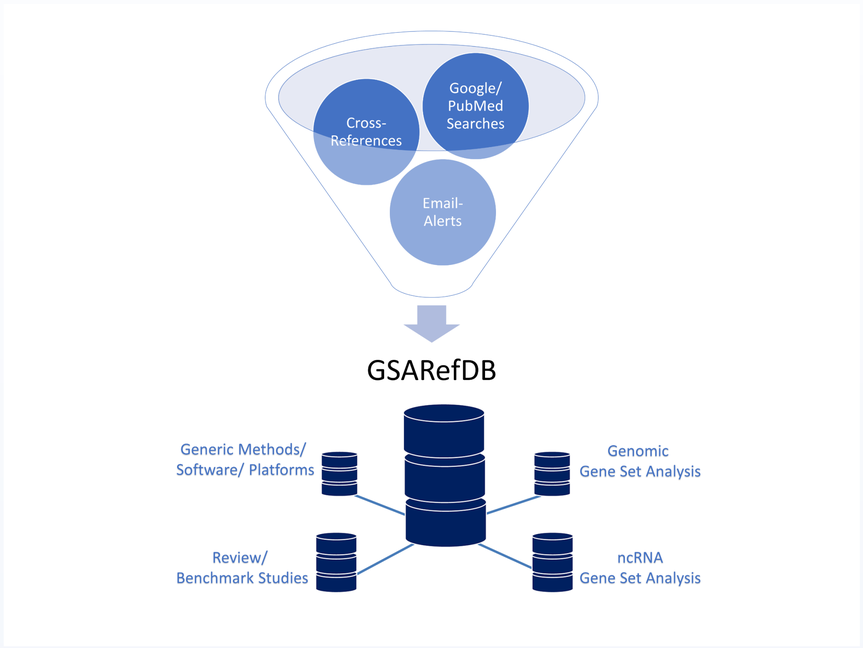
(i) Build tools to measure the popularity and performance of all available GSA methods, tools, and platforms.

(ii) Determine if the best performing methods/tools are the most popular ones.

METHODS

1. Popularity:

We have built a comprehensive and open database of GSA references (GSARefDB), including 350 papers of GSA methods, software or platforms, 91 references to papers for non-mRNA GSA tools, and 62 reviews or benchmark studies. GSARefDB stores the citation count of each paper together with other information [2].

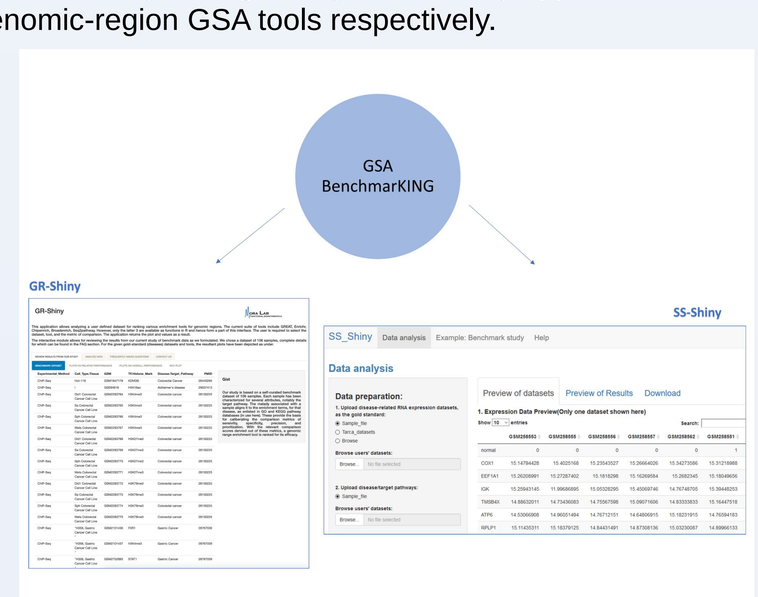


2. Performance:

We have consolidated the conclusions of all existing reviews and benchmark studies:

Authors	Year	Scope	Size	Recommended methods
Naeem *et al.* (23)	2012	ORA and FCS methods	14 methods	ANOVA, Z-SCORE, and Wilcoxon's rank sum (WRS)
Hung *et al.* (24)	2012	FCS methods	6 methods	WRS and WKS
Tarca *et al.* (25)	2013	ORA, FCS, and SS methods	16 methods	PLAGE, GLOBALTEST, and PADOG.
Bayerlova *et al.* (26)	2015	ORA, FCS, and PT methods	7 methods	CePa and PathNet.
Jaakkola *et al.* (27)	2016	ORA, FCS, and PT methods	5 methods	SPIA, CePa, NetGSA.
De Meyer *et al.* (28)	2016	ORA, FCS, and NI methods	4 methods	PADOG (specificity) and BinoX (sensitivity).
Lim *et al.* (29)	2018	SS/Pathway-activity methods	13 methods	Pathifier and SAS, followed by PLAGE and individPath.
Nguyen *et al.* (30)	2019	ORA, FCS, and PT methods	13 methods	PADOG and CePaGSA, as well as ROntoTools, CePaORA, and PathNet.
Ma *et al.* (31)	2019	FCS, PT, and NI methods	9 methods	DEGraph, followed by PathNet and NetGSA.
Zyla *et al.* (32)	2019	ORA, FCS, and SS methods	9 methods	PLAGE (sensitivity), ORA and PADOG (specificity/FPR), PADOG (prioritization), and CERNO (reproducibility).
Geistlinger *et al.* (33)	2020	ORA, FCS, and SS methods	10 methods	ROAST and GSVA (for self-contained hypothesis). ORA and PADOG (for competitive hypothesis).

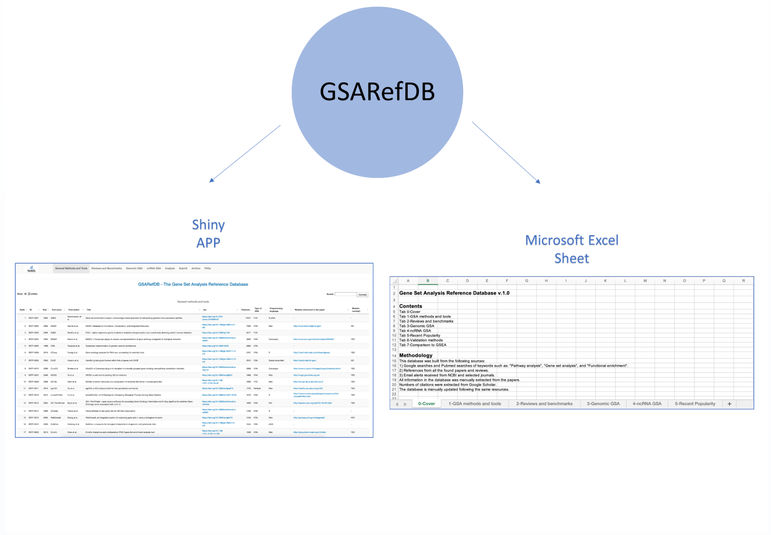
We have also built some tools for open benchmarking of GSA methods, such as jupyter notebooks and shiny apps, collected in a website called "GSA BenchmarKING" [3]. For example, 'ss-shiny' and 'gr-shiny' are R/shiny apps for benchmarking single-sample and genomic-region GSA tools respectively.



RESULTS

1. Popularity:

The most influential GSA method is "Gene Set Enrichment Analysis" (GSEA), with more than twice as many citations as its follower, the platform called DAVID. During the last year, GSEA is still the most popular method, followed by clusterProfiler, GOseq, Enrichr, DAVID, and ClueGO.



2. Performance:

Some of the most commonly recommended software include Over-Representation-Analysis methods, PADOG (among Functional Class Scoring methods), PLAGE and Pathifier (among single-sample methods), and PathNet and CePa (among topology-based methods).

3. Relationship between popularity and performance of GSA tools:

We found that performance studies are still few, small, inconsistent, and dependent on the quality of the benchmarks; however, they tend to recommend tools different to the popular and friendly ones.

CONCLUSIONS

1. Proper tool selection is essential to generate high-quality results in all scientific fields. We suggest that tool performance and tool selection studies, via the popularity-performance evaluation based on an exhaustive reference database, is a methodology that should be followed up, to keep track of the evolution of the tool selection issues in a scientific field.

2. We have also introduced examples of popularity and performance-measuring software (GSARefDB, GSA BenchmarKING) that are useful to build such studies.

3. We recommend biomedical researchers to follow such tools when facing a GSA tool selection process.

REFERENCES

1. MORA, A. (2019), Gene set analysis methods for the functional interpretation of non-mRNA data—Genomic range and ncRNA data, *Brief. Bioinf.* doi: https://doi.org/10.1093/bib/bbz090
2.GSARefDB (https://gsa-central.github.io/gsarefdb.html)
3.GSABenchmarKING (http://gsa-central.github.io/benchmarking.html)

CONTACT

https://mora-lab.github.io/