On the reproducibility of functional enrichment analysis in biomedicine

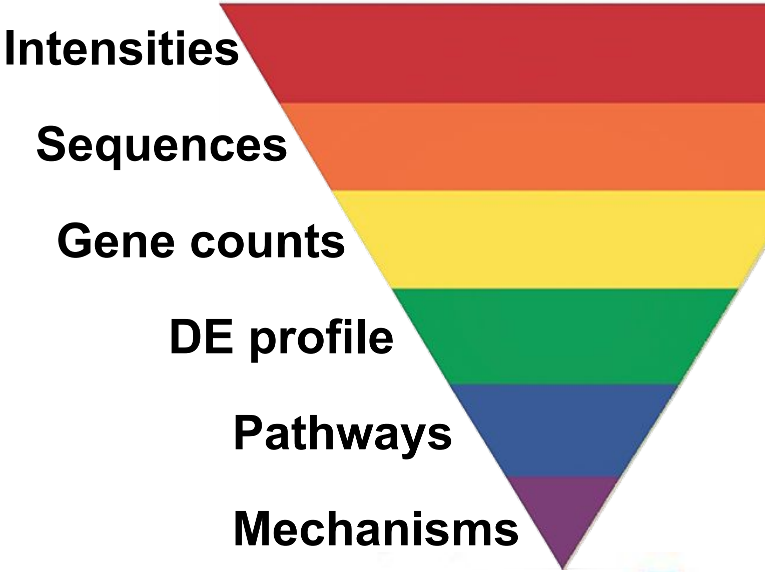
Mark Ziemann, Kaumadi Wijesooriya, Anusuiya Bora,
Sameer A Jadaan, Tanuveer Kaur, Kaushalya L Perera

2022-06-03

Outline

- Why is enrichment analysis so important?
- What are the main issues?
- How common are they?
- How to avoid them?
- What does “gold standard” analysis look like?

What is enrichment analysis and why is it so important?



Intensities

Sequences

Gene counts

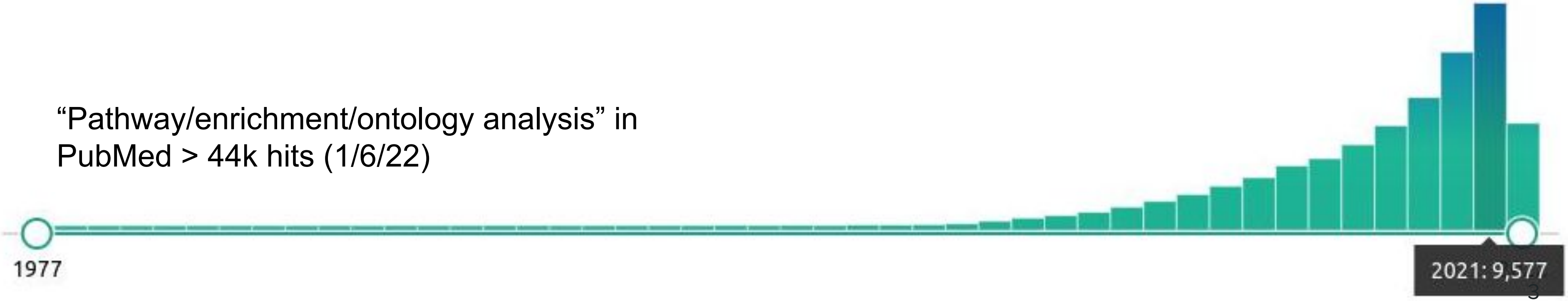
DE profile

Pathways

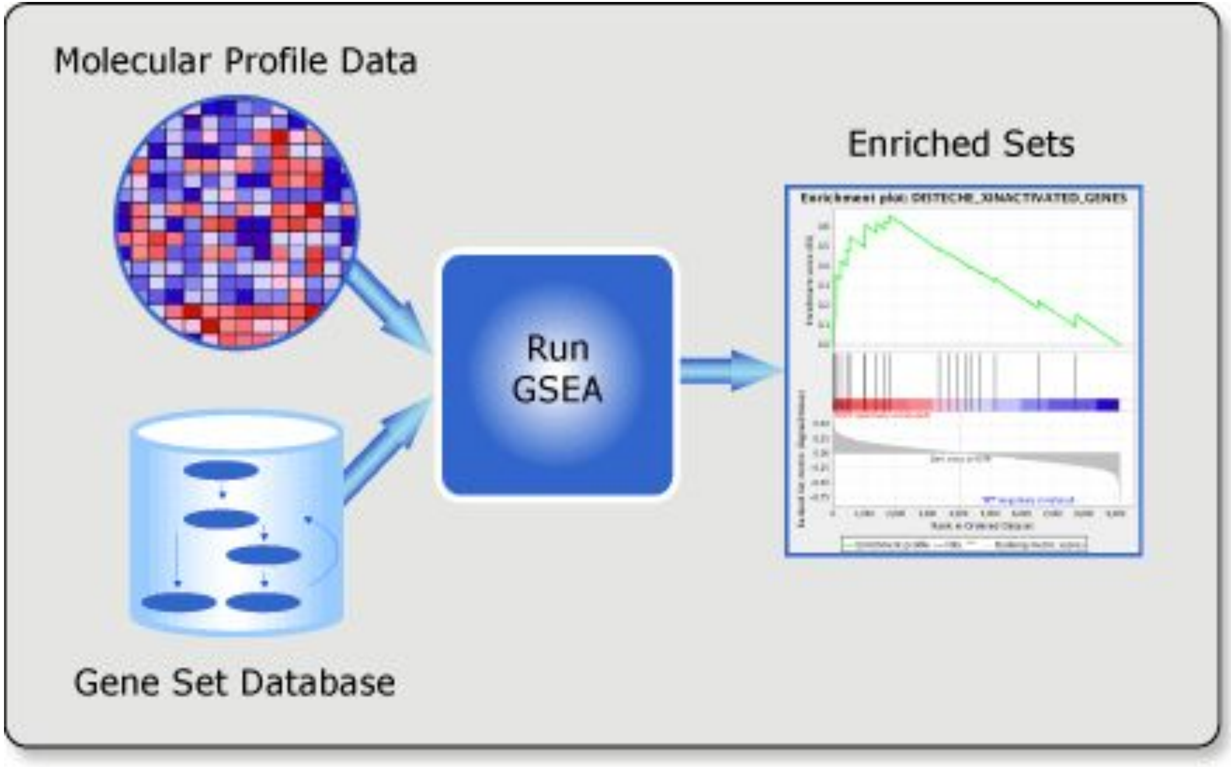
Mechanisms

- A way to summarise thousands of individual measurements into a shortlist of pathways
- May contains clues about “mechanisms”

“Pathway/enrichment/ontology analysis” in PubMed > 44k hits (1/6/22)



How does it work?



Which gene sets to use?

Literature curated

KEGG
Kyoto Encyclopedia of Genes and Genomes

GENEONTOLOGY
Unifying Biology

reactome

Data driven

MSigDB
Molecular Signatures Database

Harmonizome

Custom

ENCODE

GTEx

gnomAD
genome aggregation database

miRTarBase

ROADMAP epigenomics PROJECT

GWAS Catalog
The NHGRI-EBI Catalog of published genome-wide association studies

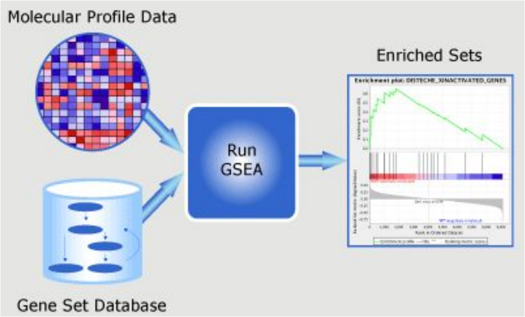
dee2.io
a repository of uniformly processed RNA-seq data

How is pathway analysis done?

Over-representation

	Non-DE	DE
Not in set	833 (87%)	121 (13%)
In set	64 (62%)	39 (38%)
Fisher Exact test p=1E-5		

Functional class scoring

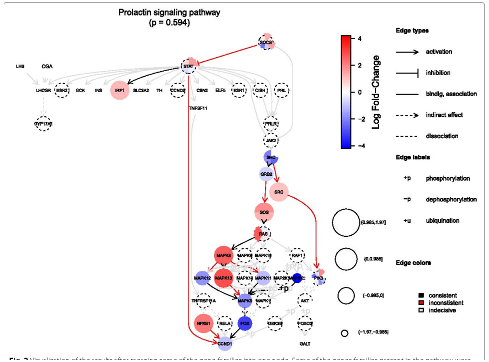

Molecular Profile Data

Run GSEA

Enriched Sets

Gene Set Database

Pathway topology


Prolactin signaling pathway
(p = 0.594)

Khatri et al, 2012, PLoS Comp Biol, https://doi.org/10.1371/journal.pcbi.1002375

ORA versus FCS

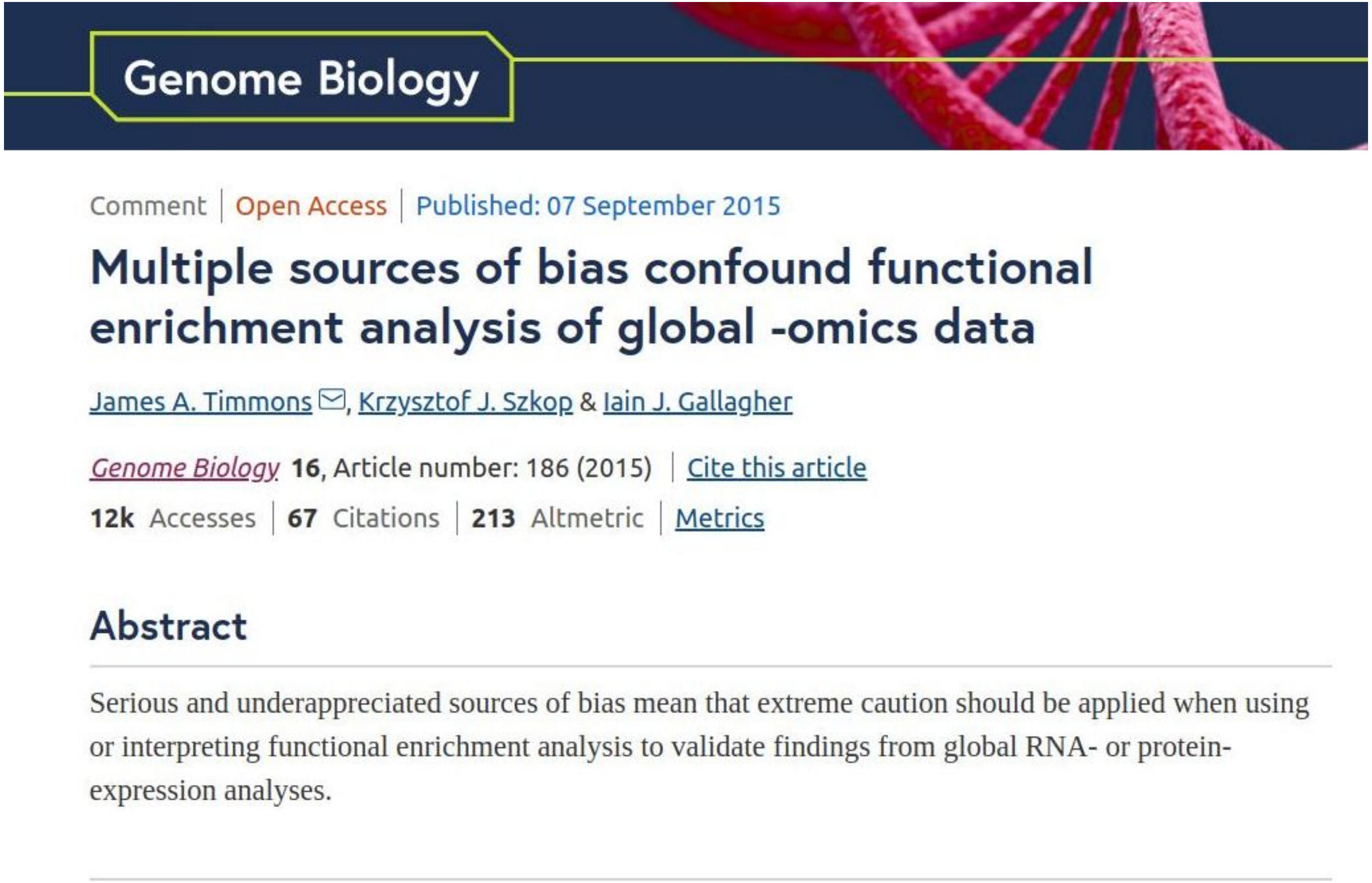
Over-representation analysis

- Treats each gene above the threshold as the same
- Treats each gene below the threshold as the same
- Selection of the threshold changes the results
- Requires careful consideration of the background list (should include all genes detected in the assay)
- As easy as submitting a list of genes to a website eg: DAVID

Functional class scoring

- Each gene has an individual weight
- Performs its own background correction
- No threshold to set
- Many ways to rank genes
- Can detect significant pathways even if no individual genes are significant
- More complicated to perform. Lack of user friendly tools. eg: GSEA

Methodological issues



Genome Biology

Comment | Open Access | Published: 07 September 2015

Multiple sources of bias confound functional enrichment analysis of global -omics data

James A. Timmons, Krzysztof J. Szkop & Iain J. Gallagher

Genome Biology **16**, Article number: 186 (2015) | Cite this article

12k Accesses | **67** Citations | **213** Altmetric | Metrics

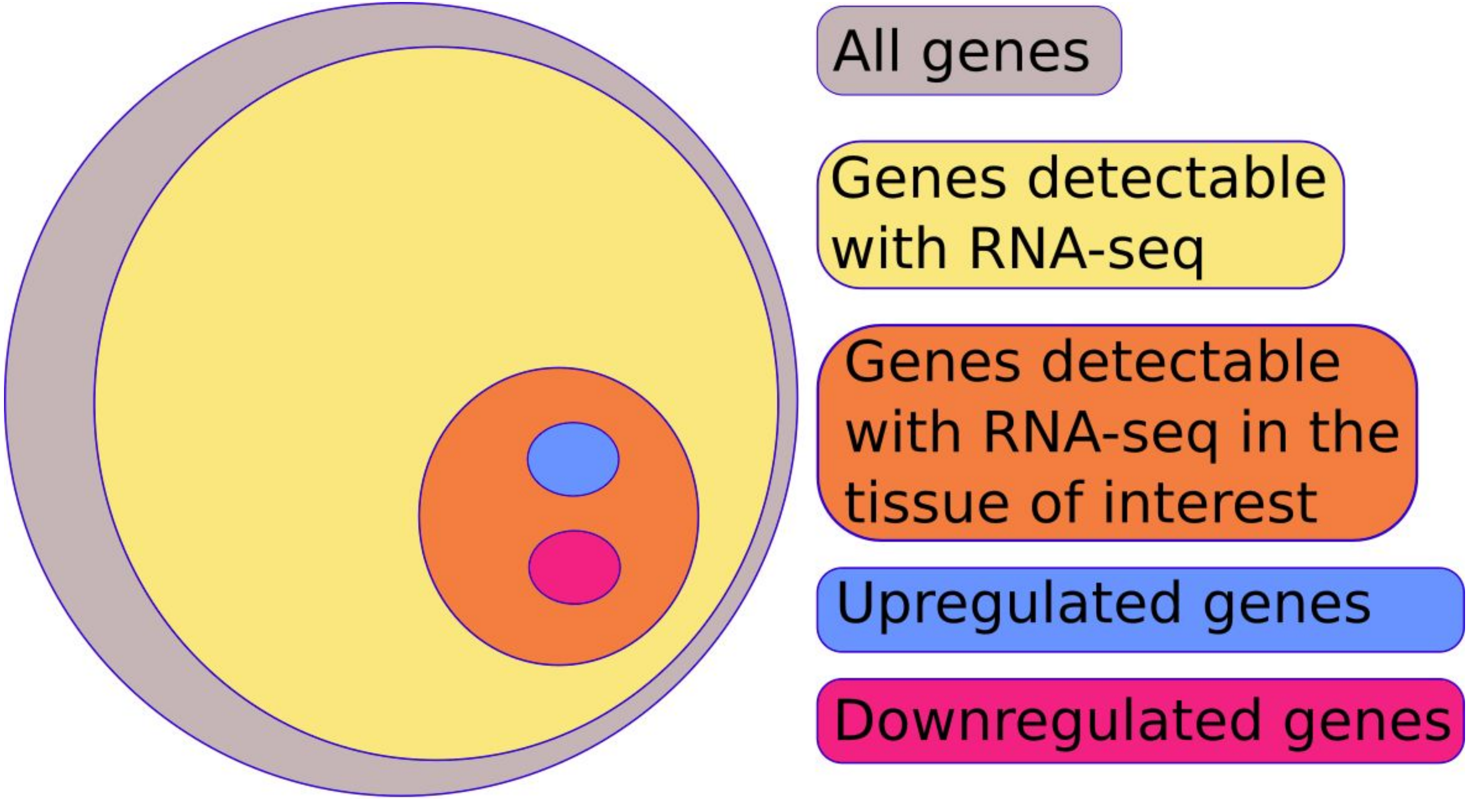
Abstract

Serious and underappreciated sources of bias mean that extreme caution should be applied when using or interpreting functional enrichment analysis to validate findings from global RNA- or protein-expression analyses.

Sources of sampling bias

- Technology/detection bias - each technology samples some genes more readily than others.
  - Affymetrix U133 GeneChip is over-represented for “Acetylation” genes compared to the whole genome
  - With RNA-seq, genes with high GC content are not well detected
  - With RNA-seq, longer genes are detected more easily
- Biological bias
  - Cells and tissues have specialised gene expression patterns, so whole genome background is inappropriate
  - When an inappropriate background is used, the results seem “truthy”

Sampling bias



All genes

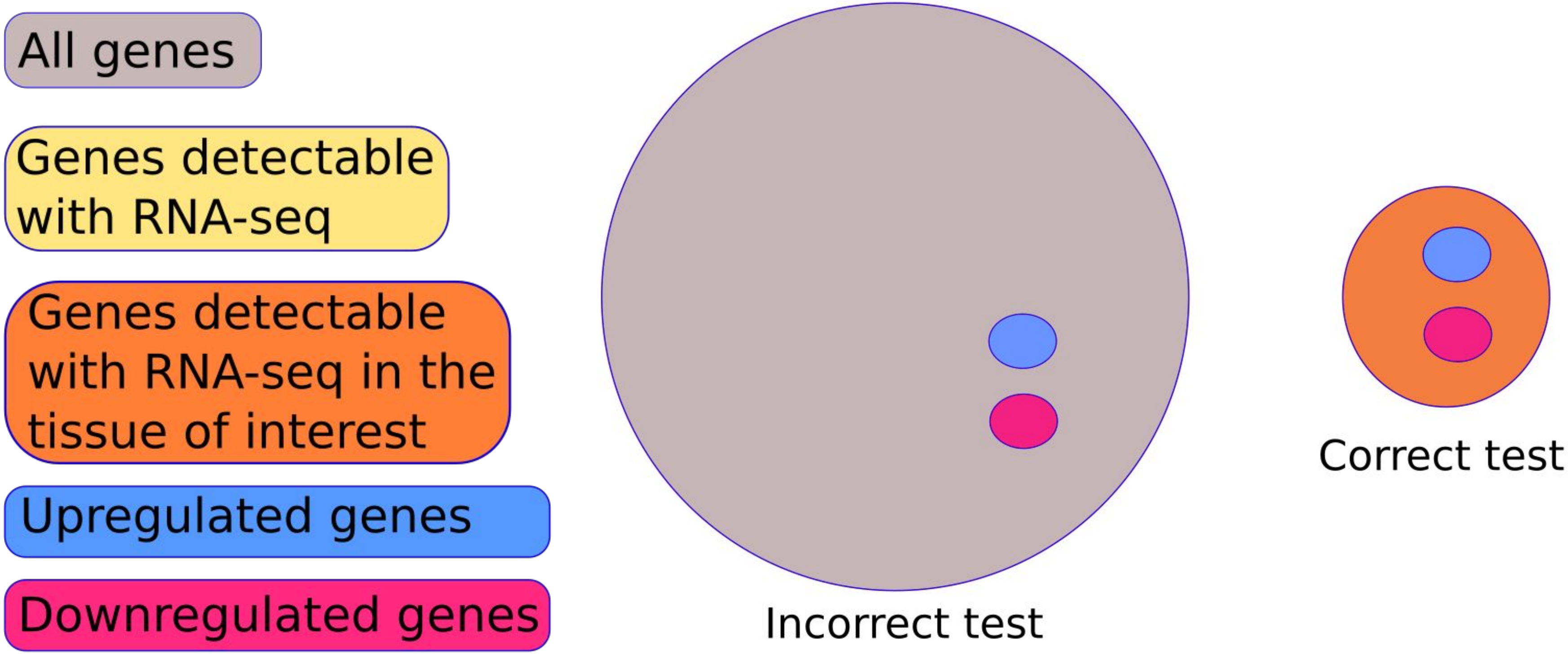
Genes detectable with RNA-seq

Genes detectable with RNA-seq in the tissue of interest

Upregulated genes

Downregulated genes

Sampling bias



All genes

Genes detectable with RNA-seq

Genes detectable with RNA-seq in the tissue of interest

Upregulated genes

Downregulated genes

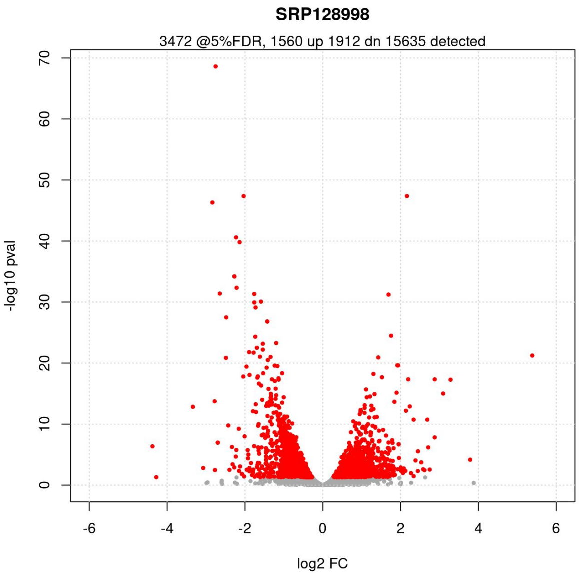
Incorrect test

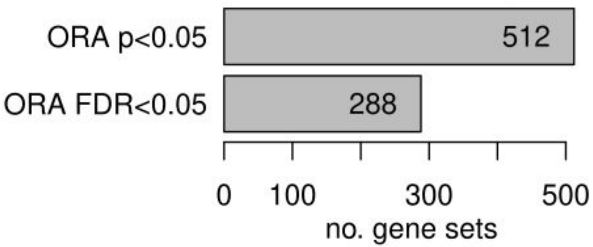
Correct test

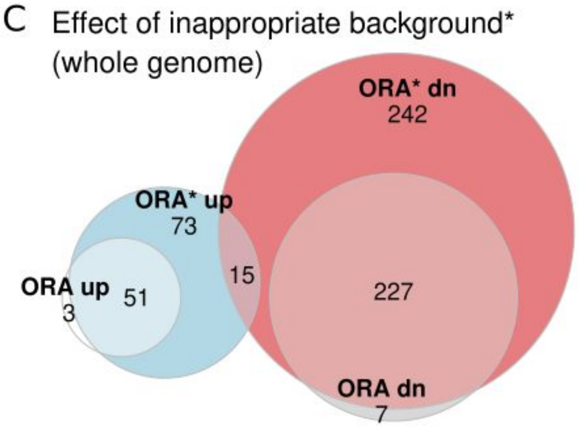
When enrichment analysis goes bad

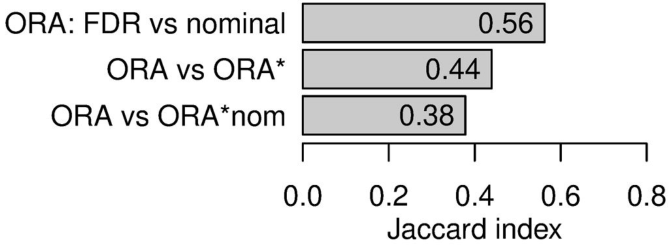
What happens when p-values are not FDR corrected for in the enrichment test?

What happens when all genes are used as the background?

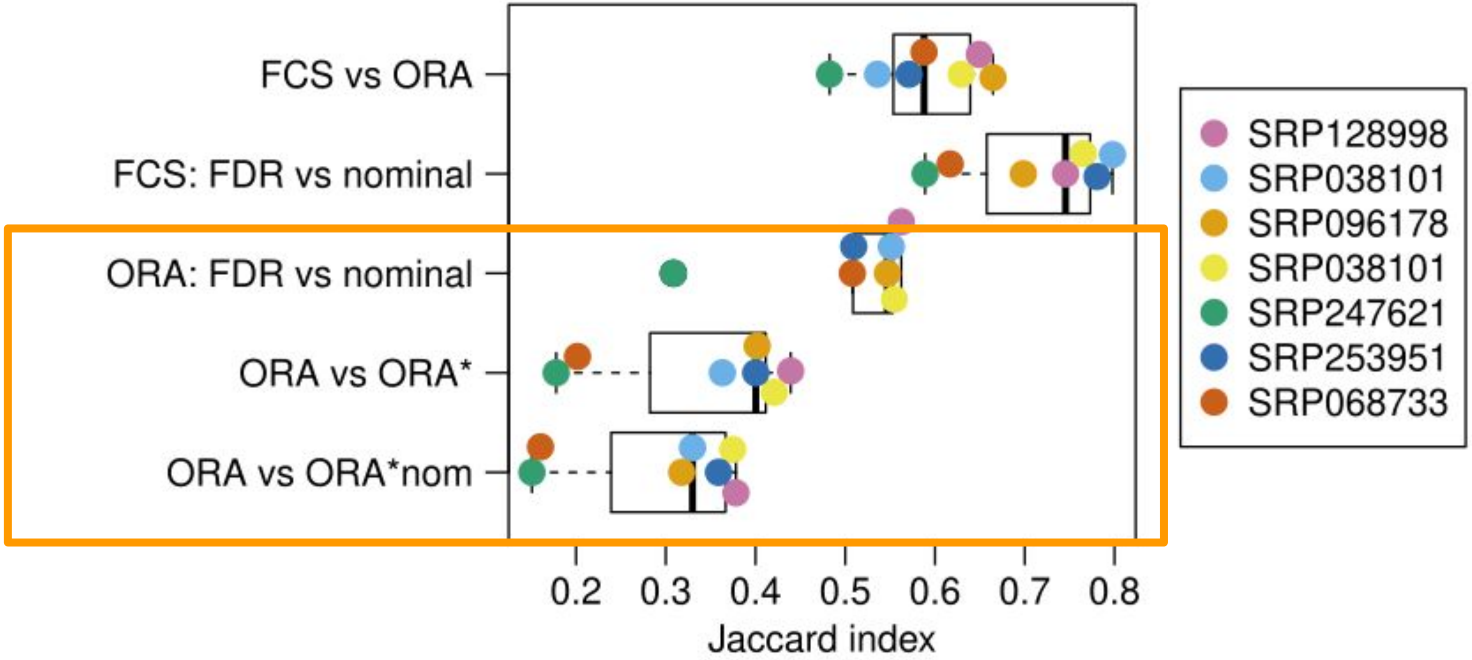








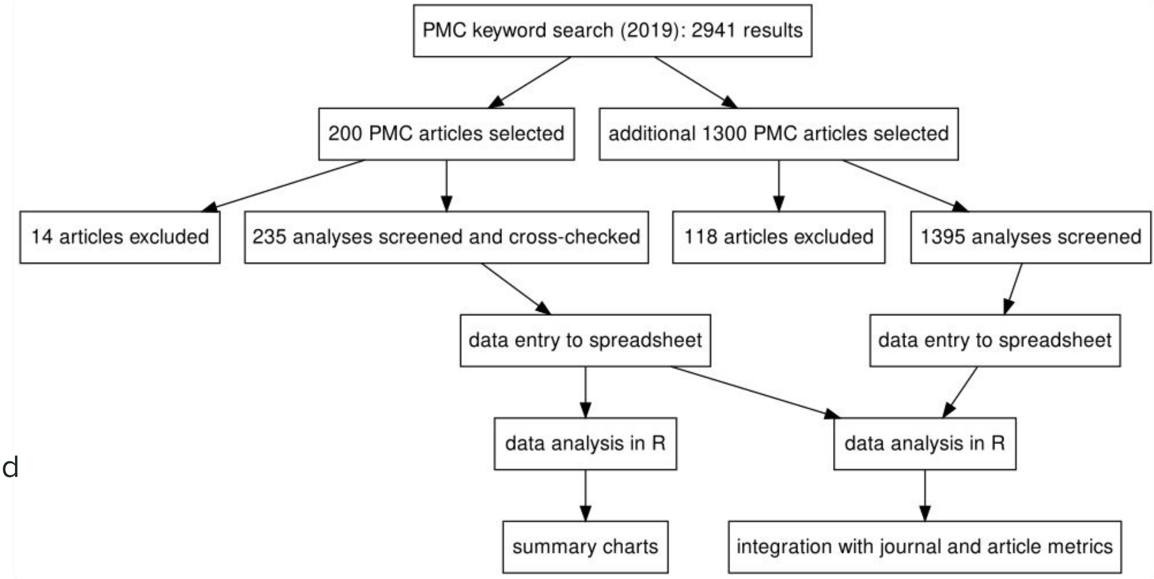
Is it a consistent pattern?



Yes.

A survey of functional enrichment practices

1. Randomly selected 1500 PMC articles from 2019 with “pathway/enrichment/ontology analysis” in abstract
2. Excluded 132 articles (new tools, reviews, conf abstracts)
3. Final set included 1363 articles, some described >1 analysis, so we have 1626 analyses in the dataset
4. We screened for methodological details:
   a. Which tool and gene set library were used (and versions)
   b. Which statistical test was used and whether FDR correction was done
   c. Whether an appropriate background was used
5. 235 analyses were double-checked



PMC keyword search (2019): 2941 results

200 PMC articles selected

additional 1300 PMC articles selected

14 articles excluded

235 analyses screened and cross-checked

118 articles excluded

1395 analyses screened

data entry to spreadsheet

data entry to spreadsheet

data analysis in R

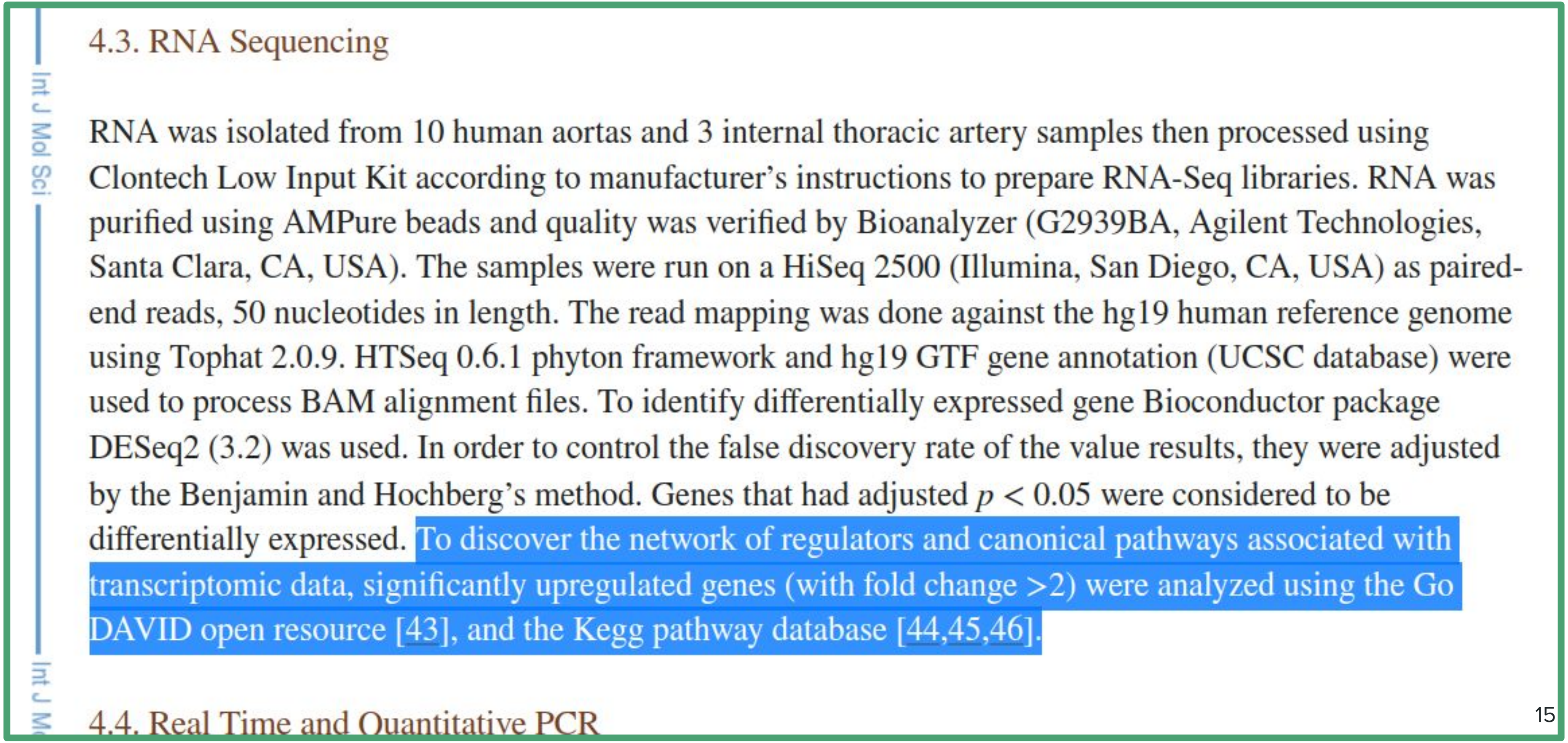
data analysis in R

summary charts

integration with journal and article metrics

Ms Kaumadi Wijesooriya
Deakin LES

Example of a methods section: PMC6425008

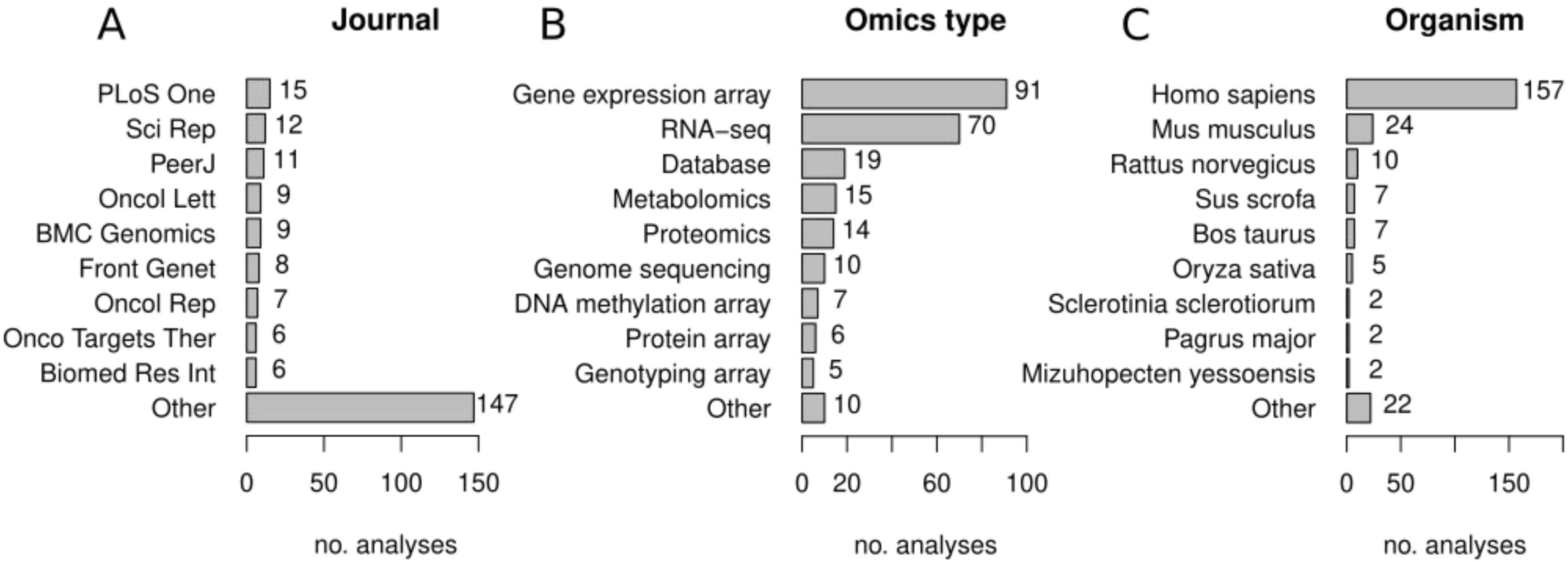


4.3. RNA Sequencing

RNA was isolated from 10 human aortas and 3 internal thoracic artery samples then processed using Clontech Low Input Kit according to manufacturer's instructions to prepare RNA-Seq libraries. RNA was purified using AMPure beads and quality was verified by Bioanalyzer (G2939BA, Agilent Technologies, Santa Clara, CA, USA). The samples were run on a HiSeq 2500 (Illumina, San Diego, CA, USA) as paired-end reads, 50 nucleotides in length. The read mapping was done against the hg19 human reference genome using Tophat 2.0.9. HTSeq 0.6.1 phyton framework and hg19 GTF gene annotation (UCSC database) were used to process BAM alignment files. To identify differentially expressed gene Bioconductor package DESeq2 (3.2) was used. In order to control the false discovery rate of the value results, they were adjusted by the Benjamin and Hochberg's method. Genes that had adjusted $p < 0.05$ were considered to be differentially expressed. To discover the network of regulators and canonical pathways associated with transcriptomic data, significantly upregulated genes (with fold change >2) were analyzed using the Go DAVID open resource [43], and the Kegg pathway database [44,45,46].

4.4. Real Time and Quantitative PCR

A survey of functional enrichment practices

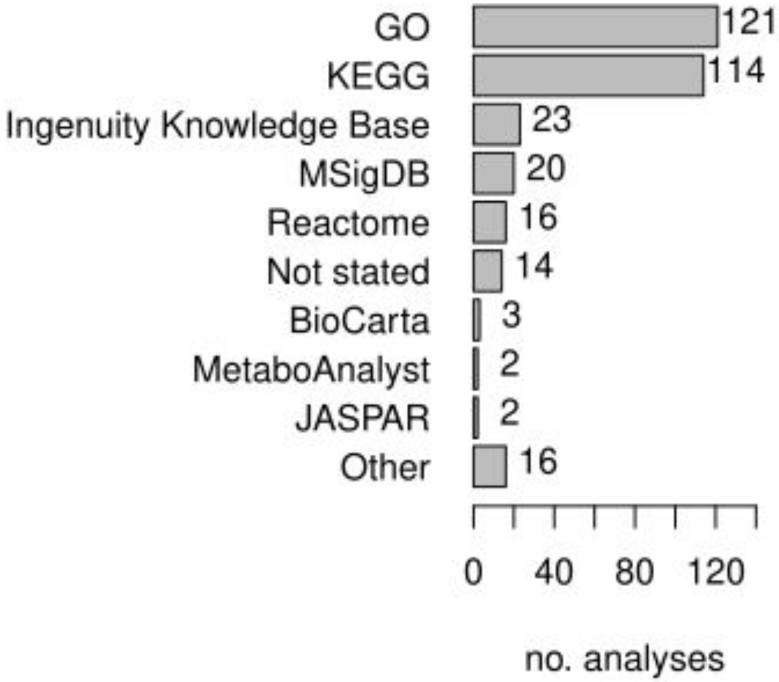


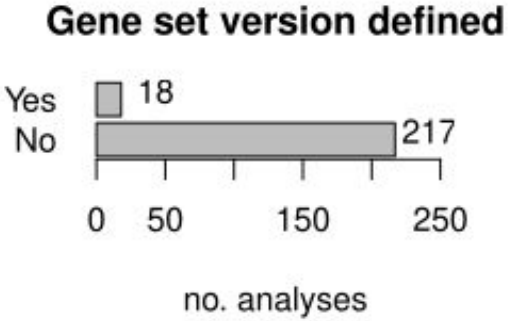
Very diverse set of journals

Gene expression analyses dominate

Mostly human focus

Gene sets used

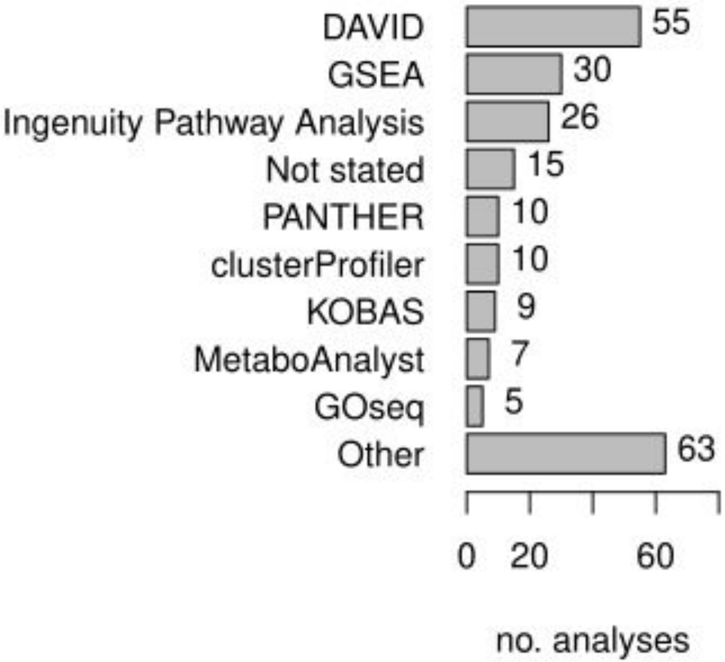




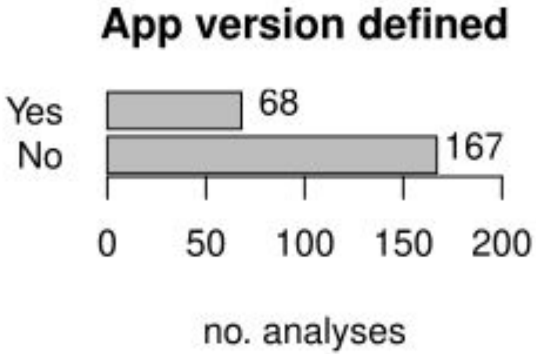
→ 92% not stated

→ GO/KEGG dominate
→ Not stated in 6% of analyses

Apps used

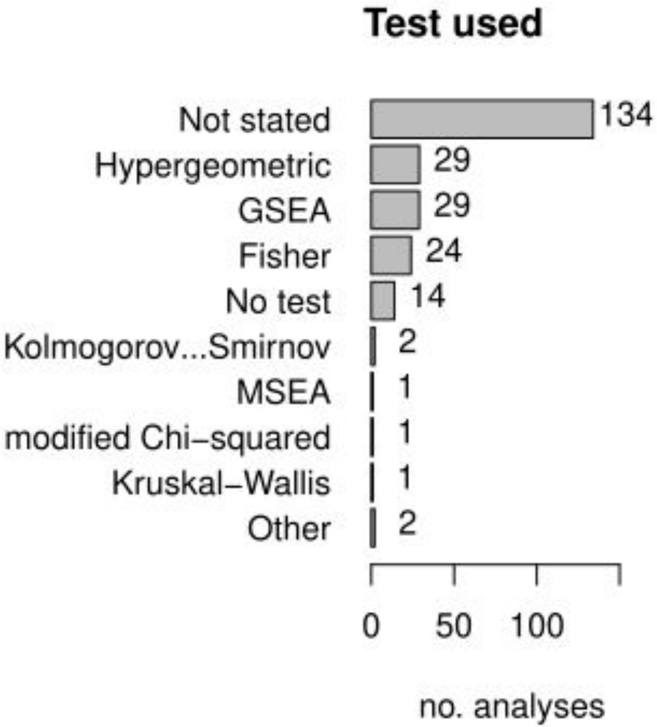


→ 71 different apps, 6% not stated

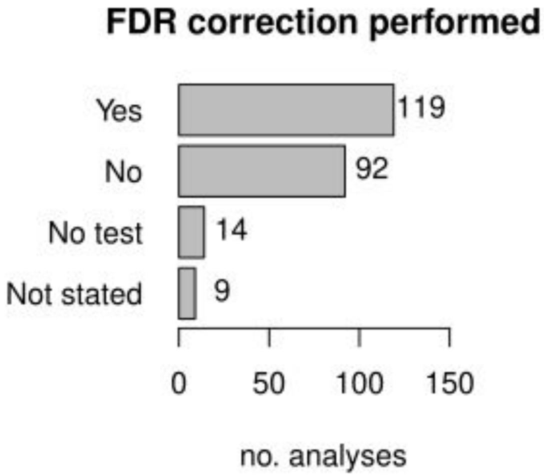


→ 71% not defined

Statistical test used



Test used

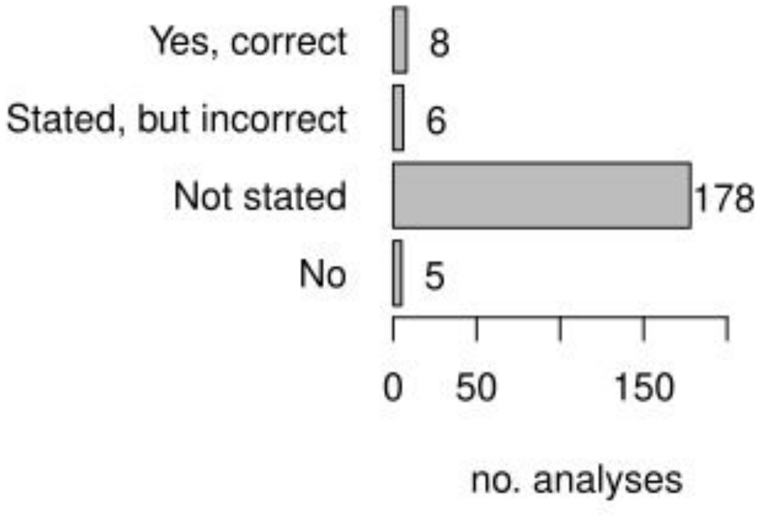


FDR correction performed

→ Only 53% did FDR correctly

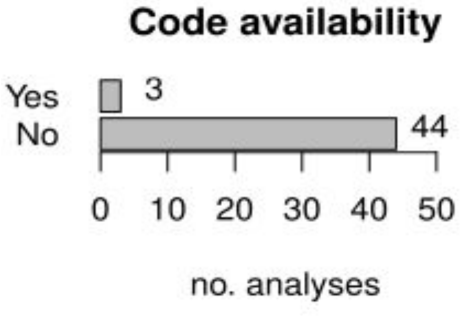
→ 29 different tests
63% not stated

Background gene lists (ORA only)



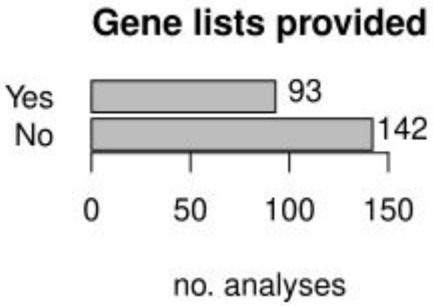
Only ~4% specified background properly

Code and data sharing

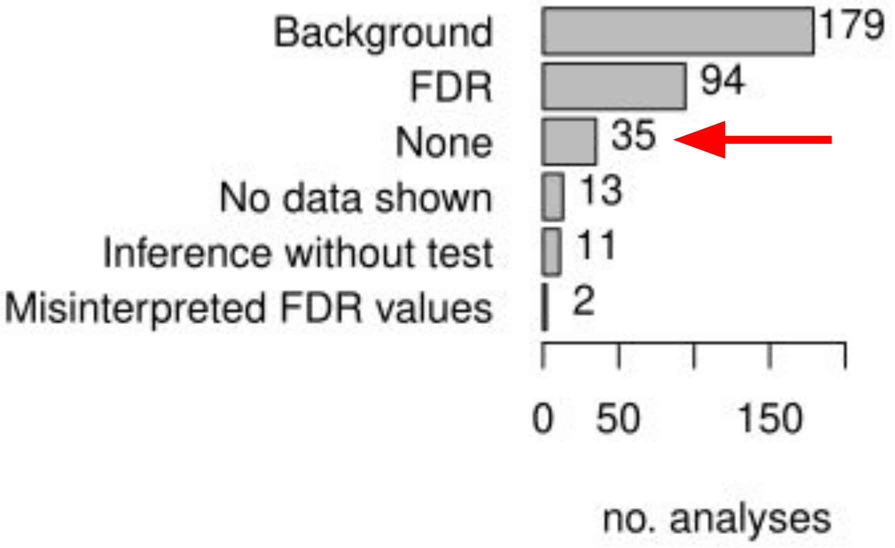


6% provided computer code

39% provided gene lists or profile data sufficient to reproduce the findings

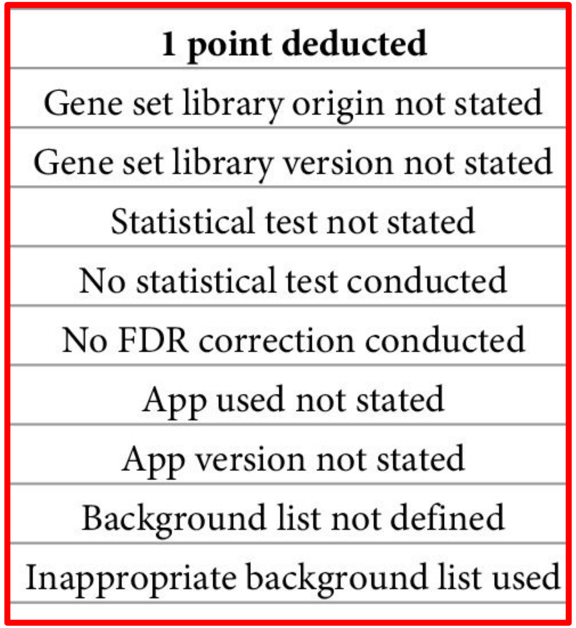


How common are major flaws?

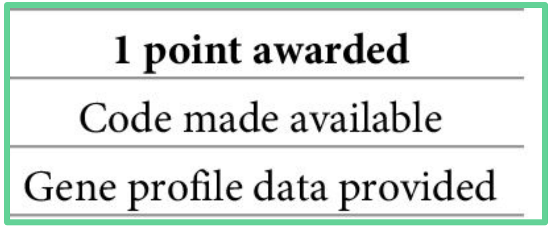


→ 15% of analyses did not have major flaws

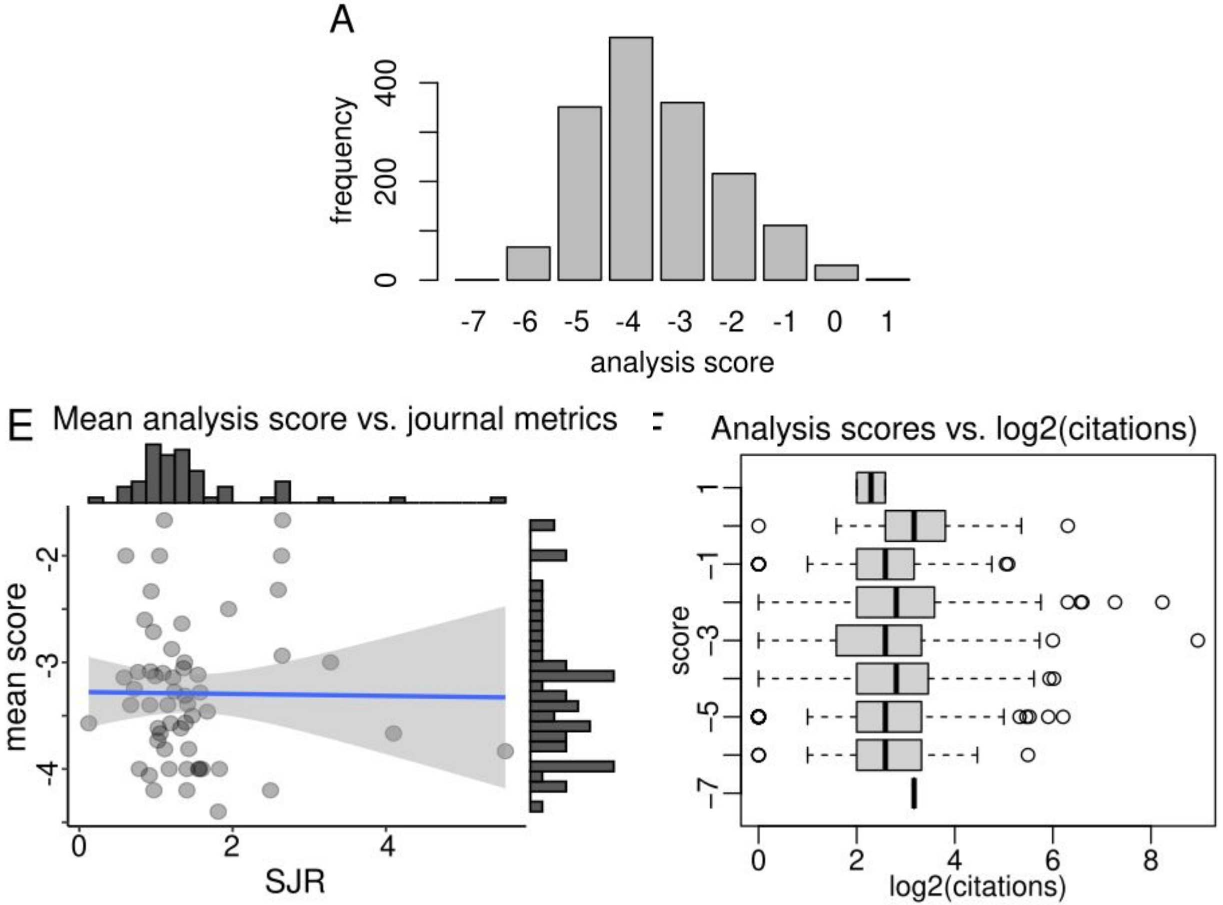
How widespread are major flaws?



1 point deducted
Gene set library origin not stated
Gene set library version not stated
Statistical test not stated
No statistical test conducted
No FDR correction conducted
App used not stated
App version not stated
Background list not defined
Inappropriate background list used



1 point awarded
Code made available
Gene profile data provided

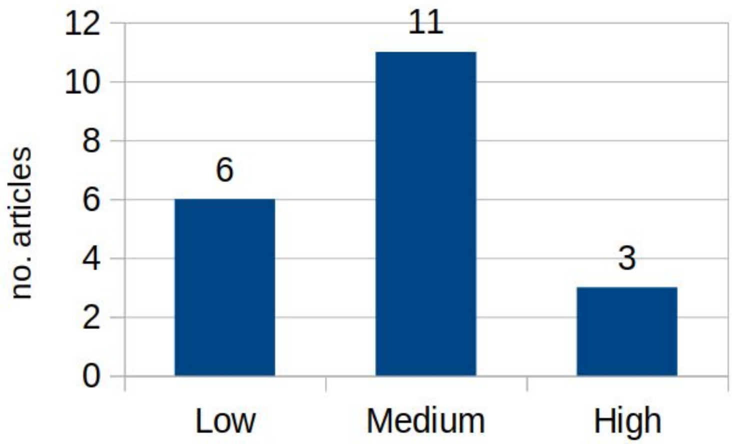


New questions arise

- Do these methodological issues invalidate the results/conclusions?
- Should up and down-regulated gene lists be examined separately or combined before ORA?
- What does best practice look like?

Pilot replication study

- 20 articles with DAVID human gene expression analysis were selected for replication <u>using the same published method</u>
- Gene lists from the supplement underwent replication using same DAVID version
- Statements from the results, discussion and conclusion were examined for consistency with replication:
    1. Low agreement
    2. Medium agreement
    3. High agreement

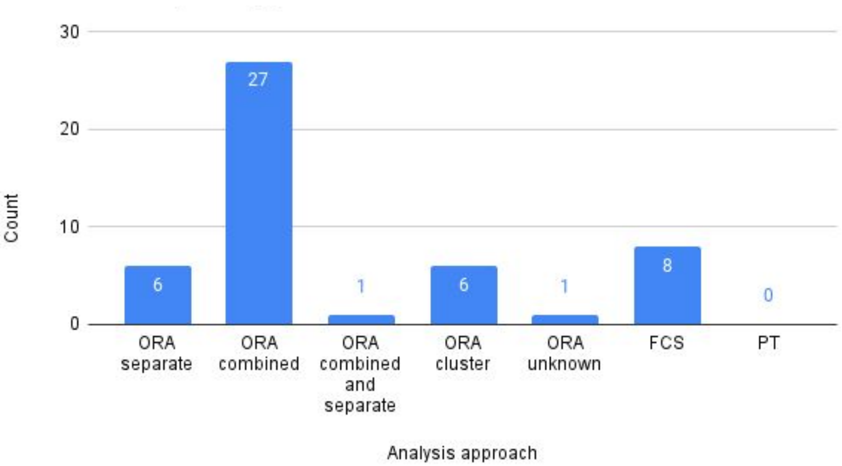


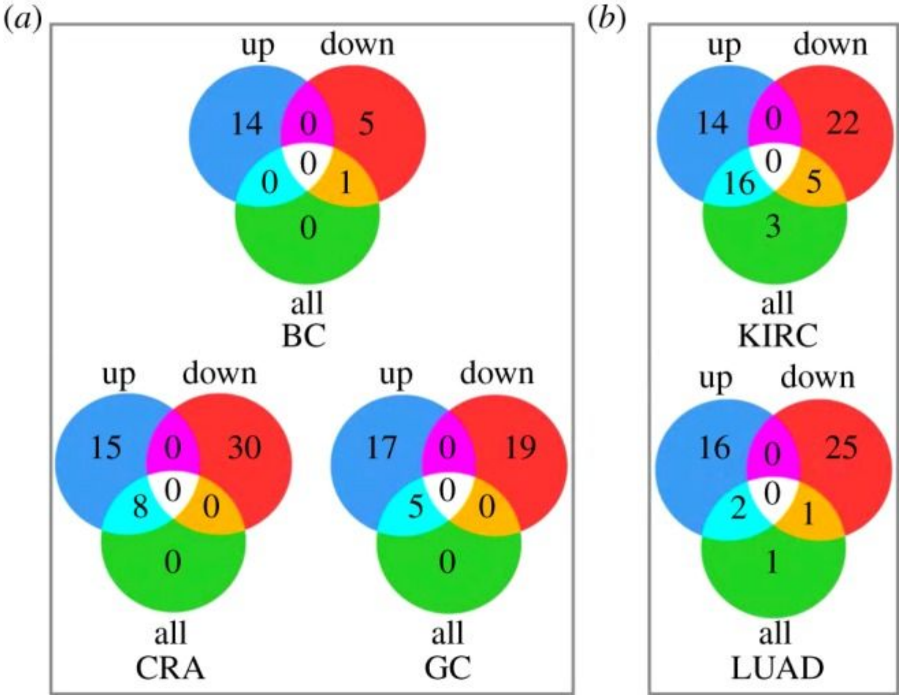
PMC6405693	PMC6368841	PMC6349697
PMC6425008	PMC6381667	PMC6444048
PMC6535219	PMC6463127	PMC6587650
PMC6539328	PMC6557785	
PMC6542760	PMC6580941	
PMC6561911	PMC6591946	
	PMC6663624	
	PMC6582306	
	PMC6333352	
	PMC6526186	
	PMC6607402	

Ms Anusuiya Bora
Vellore Institute of Technology

Should up and downregulated genes be considered separately in ORA tests?

- Hong et al (2013) found separate analysis was more sensitive (right), as most genes in pathways are positively correlated
- We found combined analysis 4x more common than separate (in a small pilot; below)



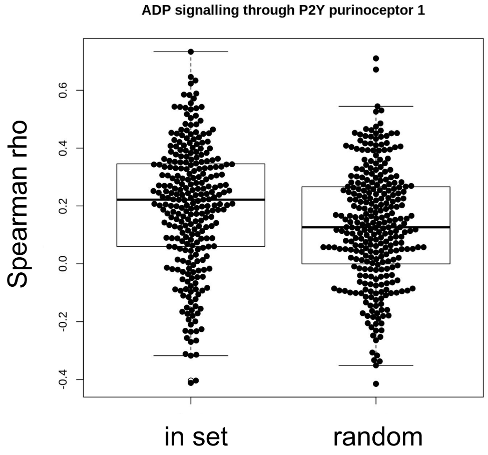


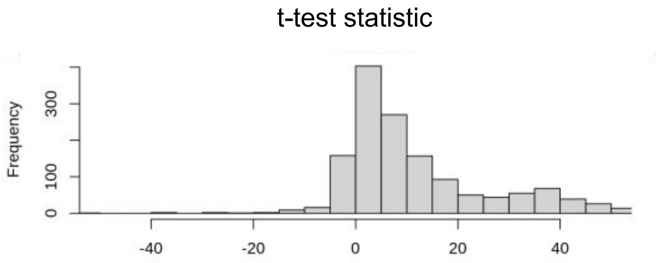
Hong et al, 2013, PMID: 24352673.

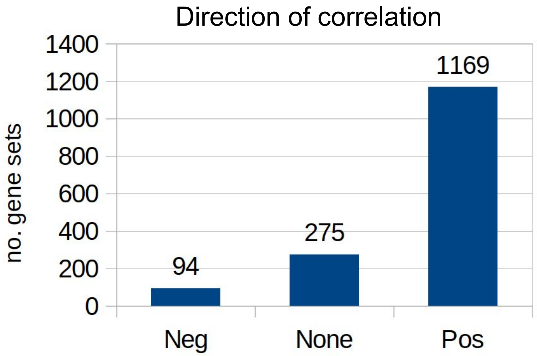
Pathway based gene sets are mostly correlated

- We examined whether genes in Reactome pathways were correlated in GTEx RNA-seq data (17383 samples)

- Generally, genes in the same set exhibited a positive correlation compared to randomly selected genes

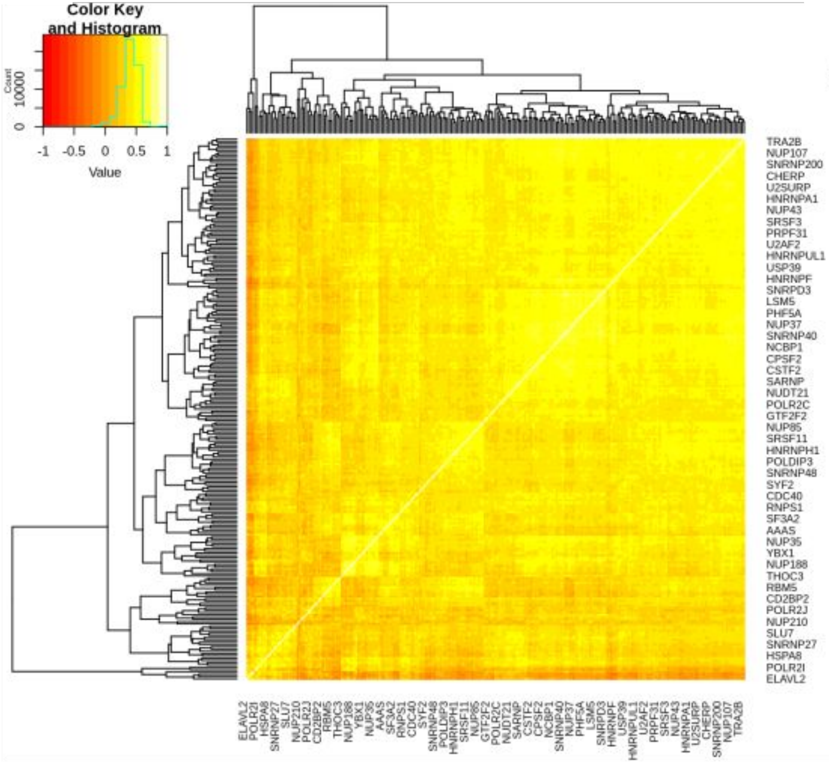




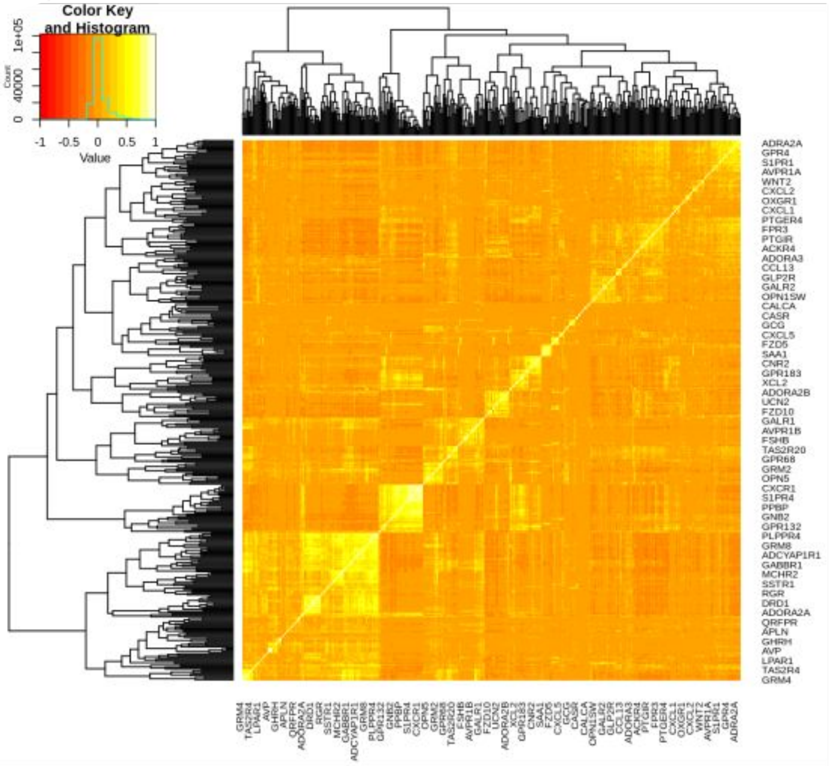


Some pathway based gene sets are not correlated

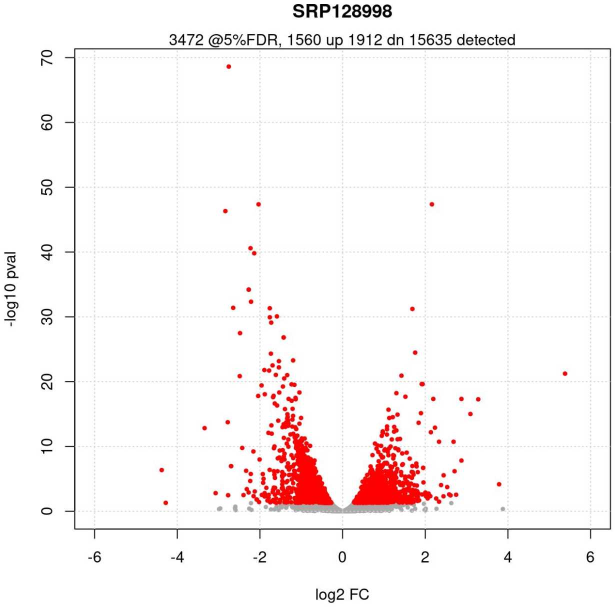
Processing of Capped Intron-Containing Pre-mRNA

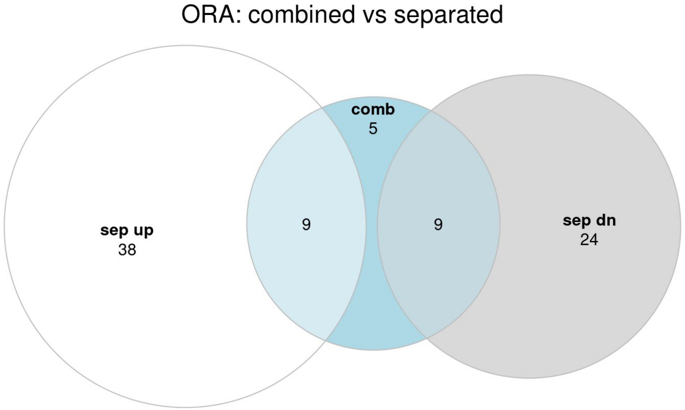


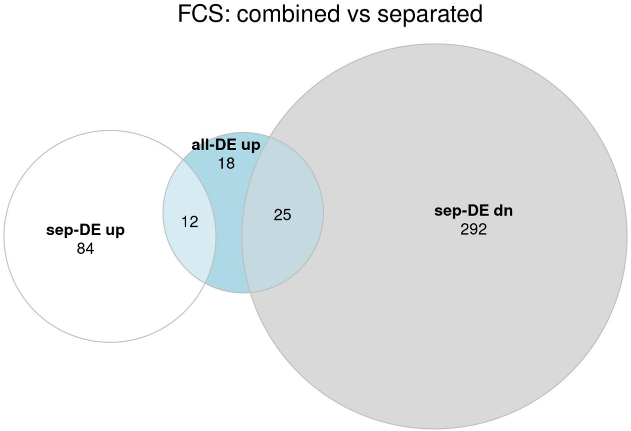
GPCR ligand binding

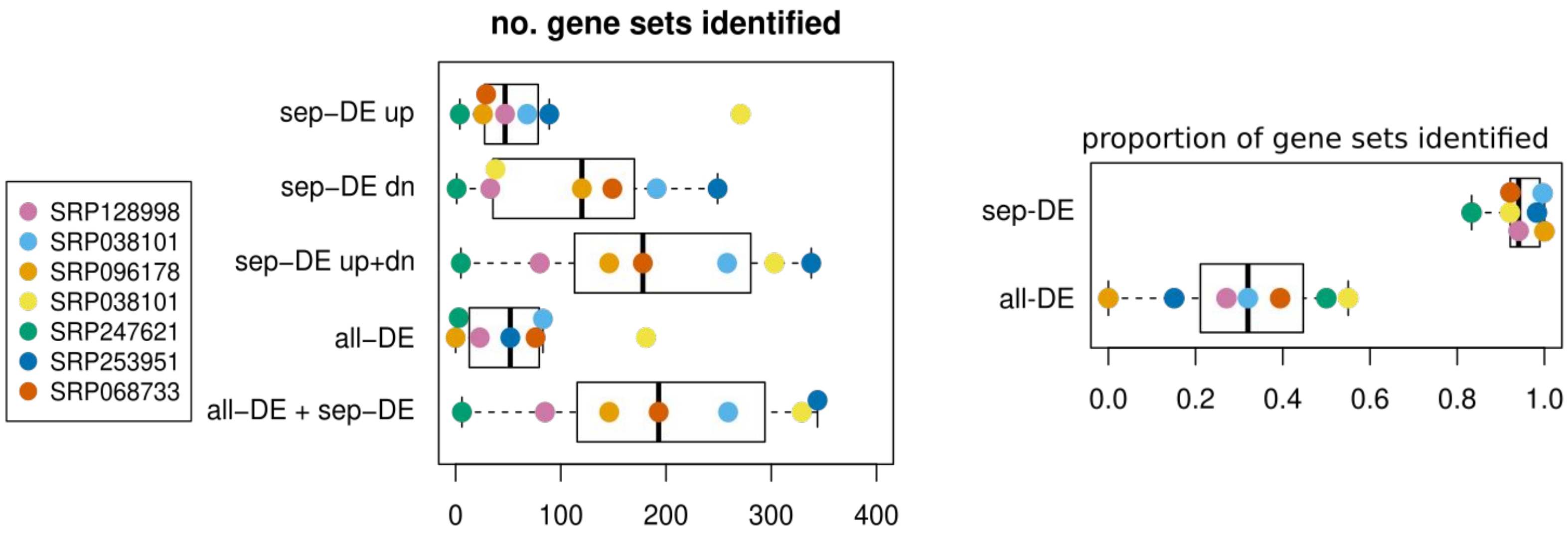


Up and down-regulated gene lists should be analysed separately







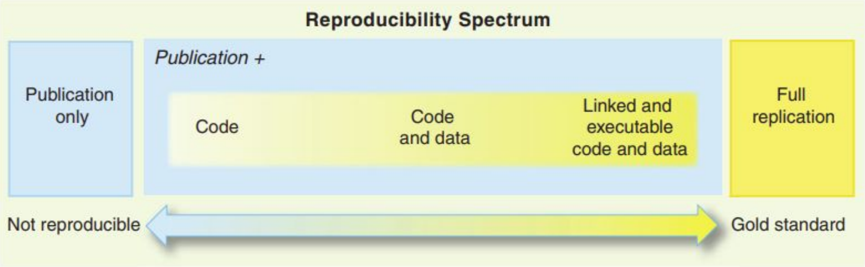

no. gene sets identified
SRP128998
SRP038101
SRP096178
SRP038101
SRP247621
SRP253951
SRP068733
sep–DE up
sep–DE dn
sep–DE up+dn
all–DE
all–DE + sep–DE
0
100
200
300
400
proportion of gene sets identified
sep-DE
all-DE
0.0
0.2
0.4
0.6
0.8
1.0

Essential minimum standards

1. Report the origin of the genesets and version
2. Report the the tool and version
3. Report the statistical test used
4. Report FDR adjusted p-values
5. For ORA, report the background used
6. Report gene selection criteria and non-default parameters
7. For ORA, perform separate analysis of up and downregulated genes

Gold standard

1. Scripted analysis rather than web app
2. Code shared at permanent repository
3. Gene profile data shared including gene lists and background
4. Code and data are linked and automatically generate tables and figures
5. Environment is recorded and managed (conda, renv, docker)



Peng 2011, PMID: 22144613.

Conclusions

- Statistical problems known since 2015, yet incredibly common in recent publications
- Most studies cannot be replicated due to lack of detail in methods
- Many common practices give suboptimal results
- Pilot study showed poor replicability
- Peer review process is failing
- A set of guidelines and reporting standards are urgently needed
- Enrichment tools need to:
  - Require a background list,
  - Report FDR values, and
  - Educate users on why both are important

Contributors

Deakin Uni School of Life and Environmental Sci.

Kaumadi Wijesooriya*

Kaushalya Perera

Tanuveer Kaur

Sameer Jadaan, Middle Technical University, Iraq

Anusuiya Bora, Vellore Institute of Technology, India

Computational resources: Nectar Research Cloud