BIAS IN MIRNA ENRICHMENT ANALYSIS RELATED TO GENE FUNCTIONAL ANNOTATIONS

Journal Club June 2, 2023

BIAS IN MIRNA ENRICHMENT ANALYSIS RELATED TO GENE FUNCTIONAL ANNOTATIONS

Authors

- Konstantinos Zagganas
- Thanasis Vergoulis
- Georgios K. Georgakilas
- Spiros Skiadopoulos
- Theodore Dalamagas

Institutions

- University of the Peloponnese
- Information Management Systems Institute, “Athena” Research Center
- Department of Biology, University of Patras

IS THIS PAPER INTERESTING FOR YOU?

- miRNA functional enrichment
- The standard method for functional enrichment
- bias on current standard method
- alternative statistical measure

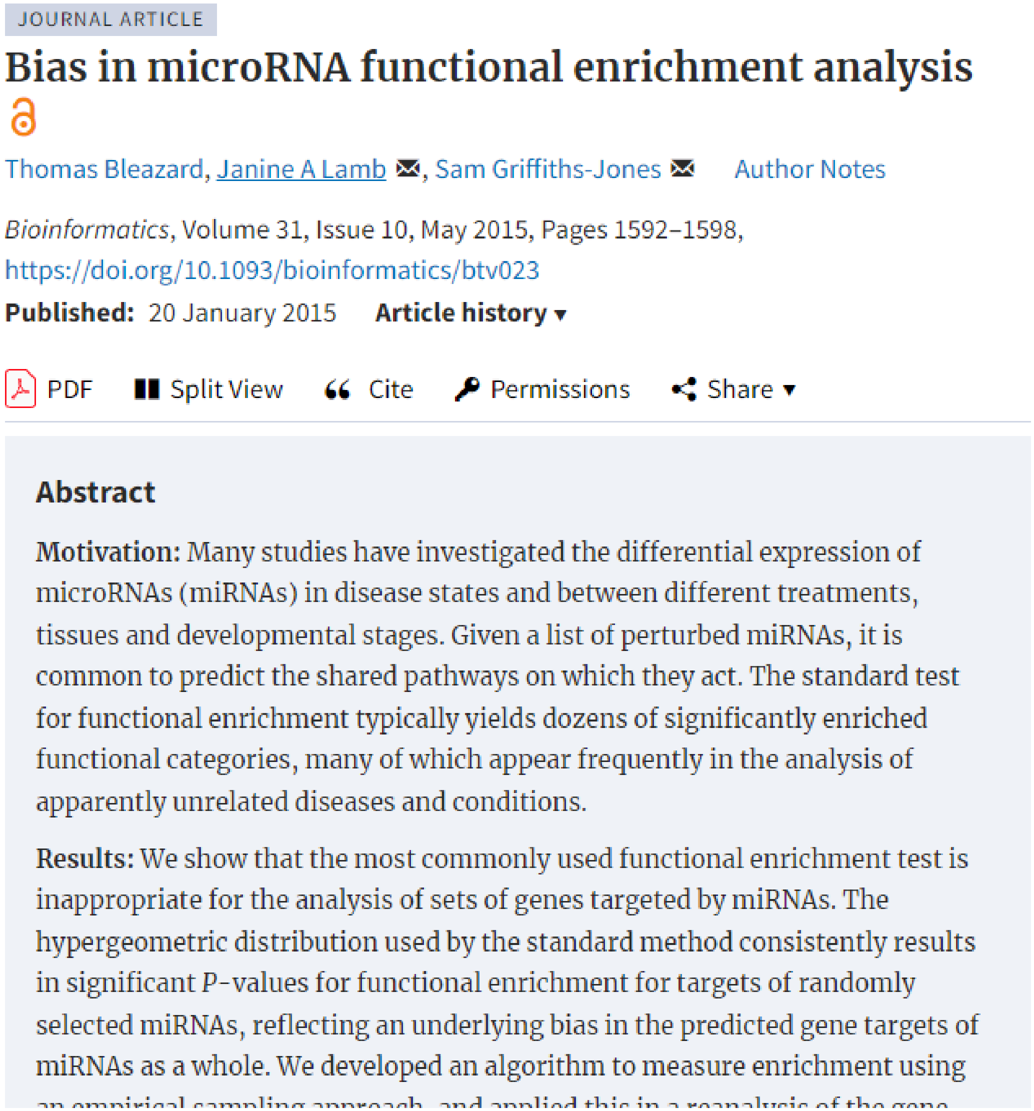
IS THIS PAPER INTERESTING FOR YOU?

- DIANA-miRPath
- miEAA
- miRPathDB
- MAGIA
- Gene Set Enrichment Analysis (GSEA)

MIRNA FUNCTIONAL ENRICHMENT ANALYSIS

1. retrieve a list of all genes targeted by the group of miRNAs
   - Union
   - Intersection
2. retrieve the list of genes that participate in the biological function
3. perform a statistical test, usually Fisher’s exact test to calculate a p-value that indicates the strength of the association between the miRNA group and the biological function

THE STANDARD METHOD OF MIRNA FUNCTIONAL ENRICHMENT ANALYSIS IS NOT SUITABLE FOR SUCH ANALYSES



Bias in microRNA functional enrichment analysis

Thomas Bleazard, Janine A Lamb ✉, Sam Griffiths-Jones ✉ Author Notes

Bioinformatics, Volume 31, Issue 10, May 2015, Pages 1592–1598, https://doi.org/10.1093/bioinformatics/btv023

Published: 20 January 2015 **Article history** ▾

Abstract

Motivation: Many studies have investigated the differential expression of microRNAs (miRNAs) in disease states and between different treatments, tissues and developmental stages. Given a list of perturbed miRNAs, it is common to predict the shared pathways on which they act. The standard test for functional enrichment typically yields dozens of significantly enriched functional categories, many of which appear frequently in the analysis of apparently unrelated diseases and conditions.

Results: We show that the most commonly used functional enrichment test is inappropriate for the analysis of sets of genes targeted by miRNAs. The hypergeometric distribution used by the standard method consistently results in significant *P*-values for functional enrichment for targets of randomly selected miRNAs, reflecting an underlying bias in the predicted gene targets of miRNAs as a whole. We developed an algorithm to measure enrichment using an empirical sampling approach, and applied this in a reanalysis of the gene

- that the standard method of miRNA functional enrichment analysis is not suitable for such analyses
- provides highly unspecific results
- The mechanics responsible for this bias are not yet fully understood
- limited amount of validated positive miRNA:target interactions
- virtually non-existent validated negative interactions

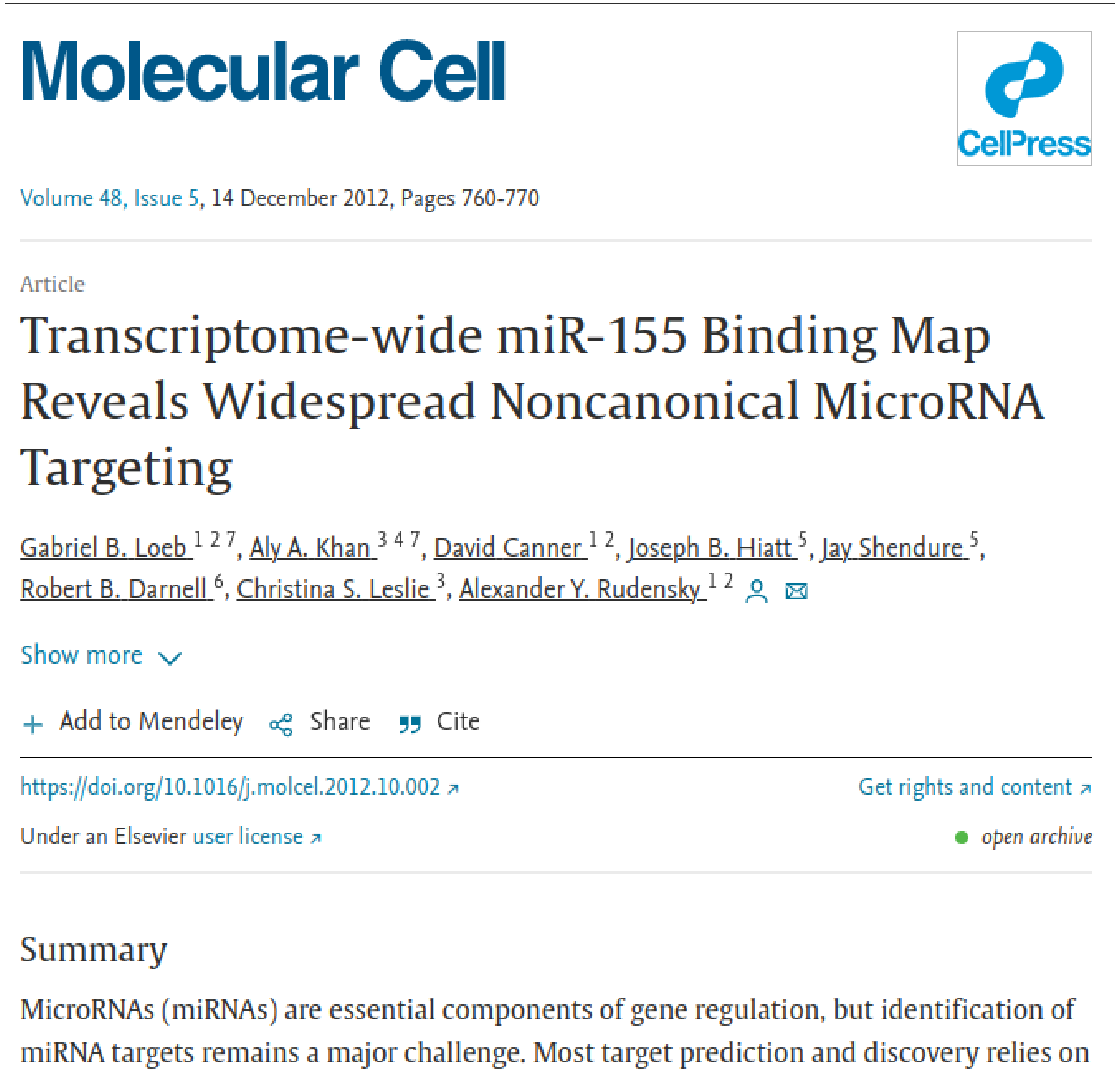
Thomas Bleazard et al

THE SEED

most target prediction algorithms have been **trained** on **seed-enriched data** sets with features extracted from the sequence surrounding the seed, even though recent evidence shows **that non-seed-based interactions are common in miRNA-mediated gene expression** regulation

“RECENT” EVIDENCE

- a transcriptome-wide identification of the endogenous targets
- miRNA-miR-155
- approximately 40% of miR-155-dependent Argonaute binding occurs at sites without perfect seed matches



Molecular Cell

CellPress

Volume 48, Issue 5, 14 December 2012, Pages 760-770

Article

Transcriptome-wide miR-155 Binding Map Reveals Widespread Noncanonical MicroRNA Targeting

Gabriel B. Loeb [1 2 7], Aly A. Khan [3 4 7], David Canner [1 2], Joseph B. Hiatt [5], Jay Shendure [5], Robert B. Darnell [6], Christina S. Leslie [3], Alexander Y. Rudensky [1 2]

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/j.molcel.2012.10.002

Get rights and content

Summary

MicroRNAs (miRNAs) are essential components of gene regulation, but identification of miRNA targets remains a major challenge. Most target prediction and discovery relies on

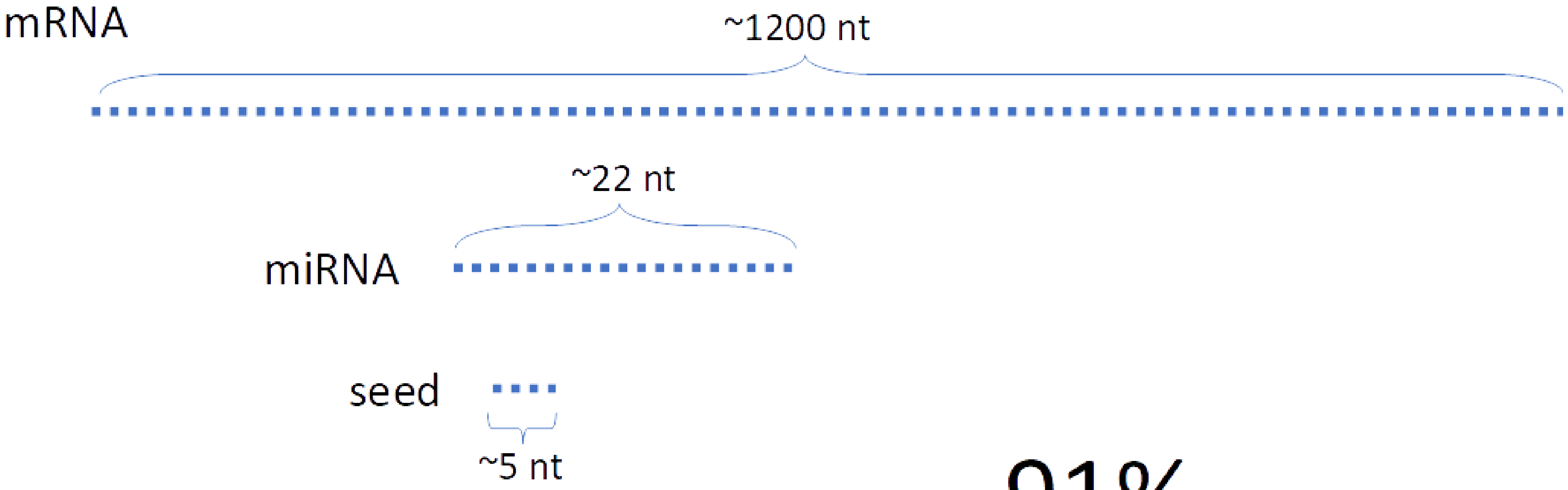
FALSE POSITIVE RATE

TargetScan: 49%

mirTarBase: 9%

MiRDB: 25%

B3GLCT predictions made by multiple algorithms in miRWalk have a high false positive rate (>96%),



mRNA

~1200 nt

~22 nt

miRNA

seed

~5 nt

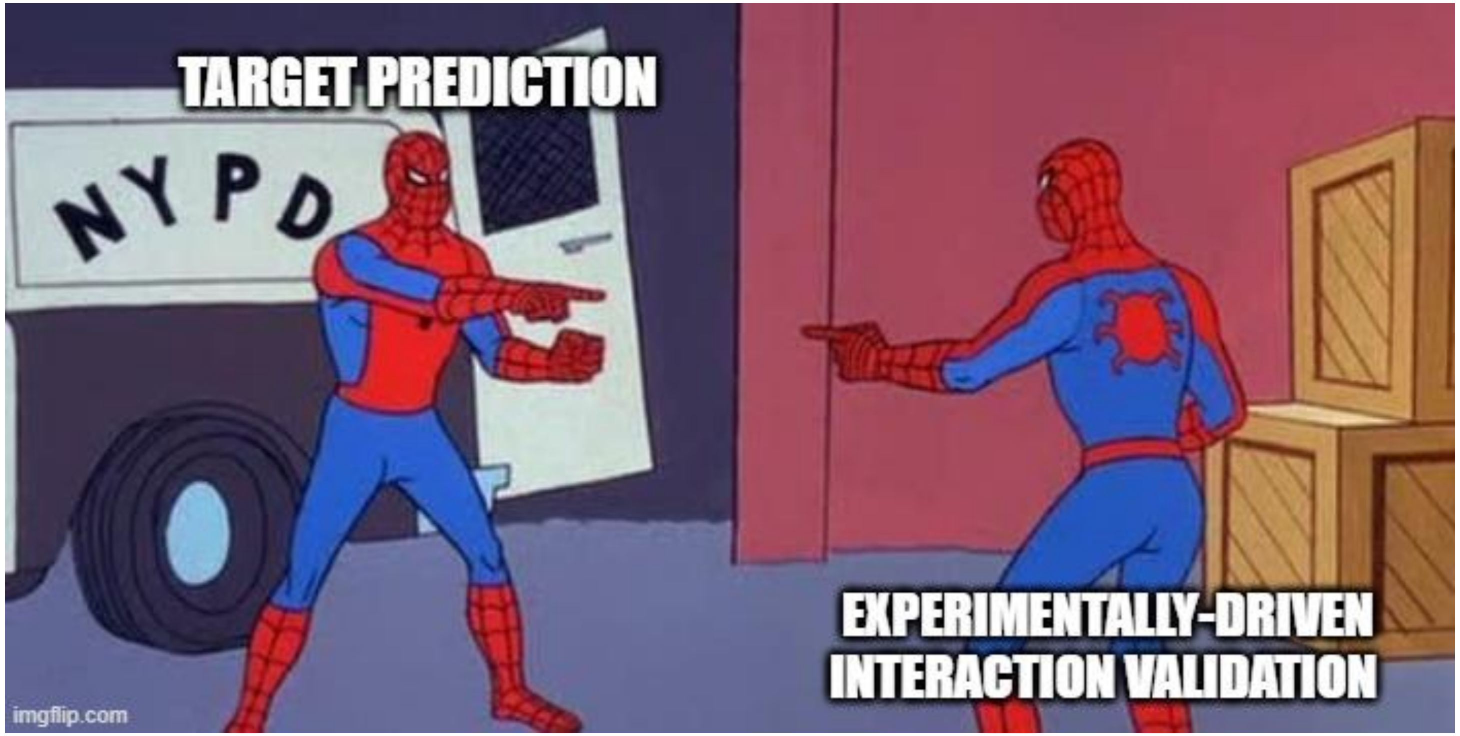
91%

Random miRNA hit random mRNA

MORE BIAS

experimentally validating miRNA binding sites is frequently driven by target prediction algorithms.

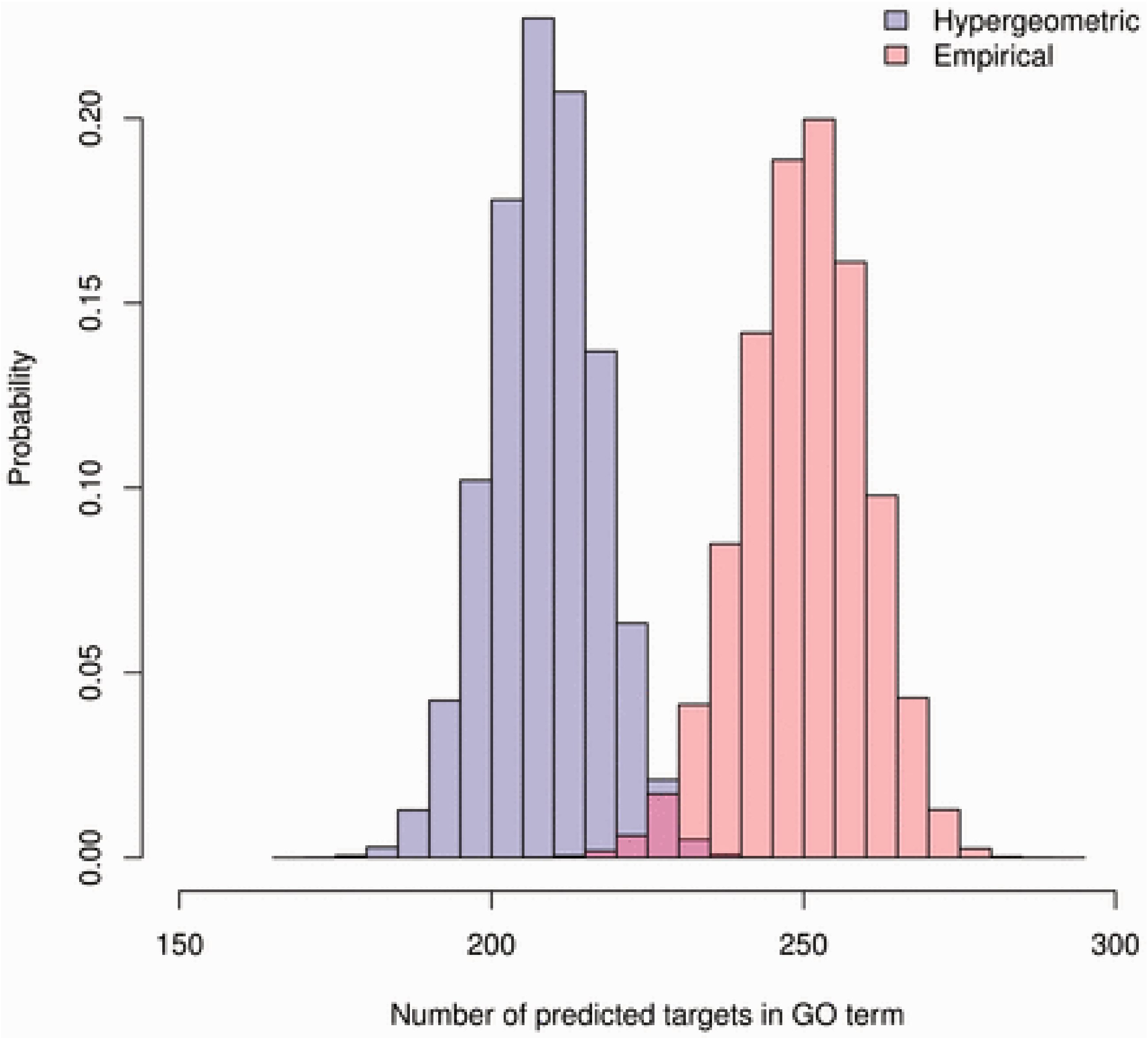
Negative results are usually not reported while the published positive interactions are inevitably enriched in seed-based binding type



invalidate the assumption made by the hypergeometric distribution:

genes are targeted by miRNAs in **a uniform fashion**

EMPIRICAL DISTRIBUTION



Thomas Bleazard et al

RANDOMIZATION TEST

1. Given a miRNA group of interest calculate a statistical measure relevant to the problem.
2. Create 1 million randomly assembled miRNA groups with the same size as the group of interest and for each of them calculate the same statistical measure.
3. The empirical p-value is then defined as the proportion of randomly assembled miRNA groups that present a better statistical behaviour compared to the behaviour of the group of interest.

GO term overlap

Thomas Bleazard et al

GO TERM OVERLAP

The proportion of genes targeted by a group of miRNAs, that are also members of a specific GO category

A: set of genes targeted by the group

B: set of genes that participate in the GO category

$$left\text{-}sided\text{-}overlap = \frac{|A \cap B|}{|A|}$$

Thomas Bleazard et al

NEW BIAS ON THE GENE-TO-BIOLOGICAL-FUNCTION ANNOTATIONS

reduced sensitivity to false negatives

predicted interactions and gene annotations

THE JACCARD COEFFICIENT

A: set of genes targeted by the group

B: set of genes that participate in the GO category

$$Jaccard\text{-}coefficient = \frac{|A \cap B|}{|A \cup B|}$$

1. randomized miRNA
   - 1 million randomly assembled miRNA groups
   - 14 miRNAs each.
2. calculated the gene members intersection
   - GO category in the data set
   - targeted by the miRNAs in the group
3. plotted the expected hypergeometric distribution for the overlaps
   - number of targeted/non-targeted genes
   - number of genes belonging/not-belonging to the same GO term

- 3106 out of a total of 15064 genes are indicated as targets in the set of interactions

- expected hypergeometric distribution following

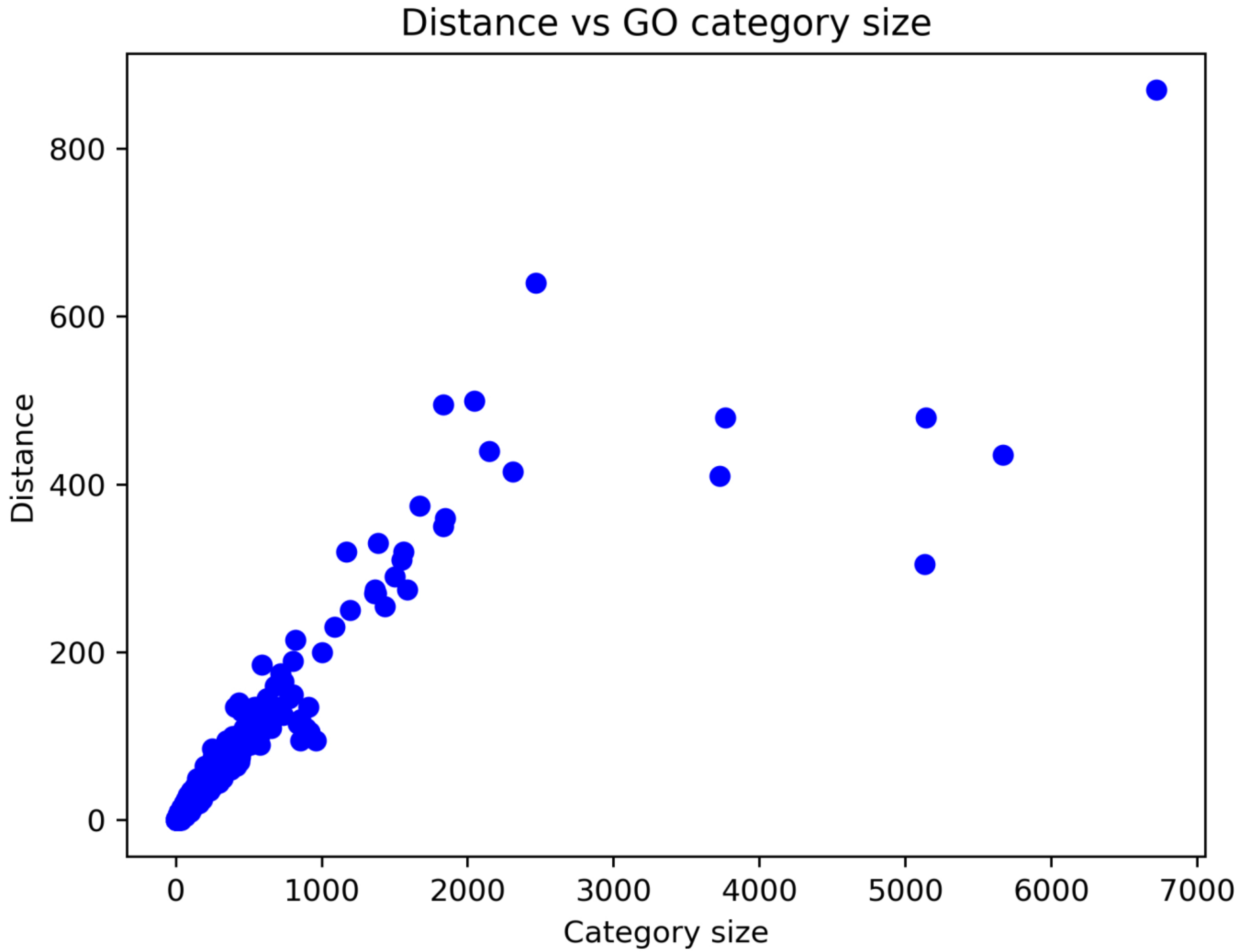
RESULTS

categories that describe more specific biological functions, tend to significantly overlap with the hypergeometric distribution.

categories that present the larger mismatch seem to be those, that contain a large number of genes

mismatch is indeed more prominent as the size of the category increase

DISTANCE BETWEEN THE DISTRIBUTIONS



THE EFFECT IS EVEN MORE PRONOUNCED FOR DISGENET

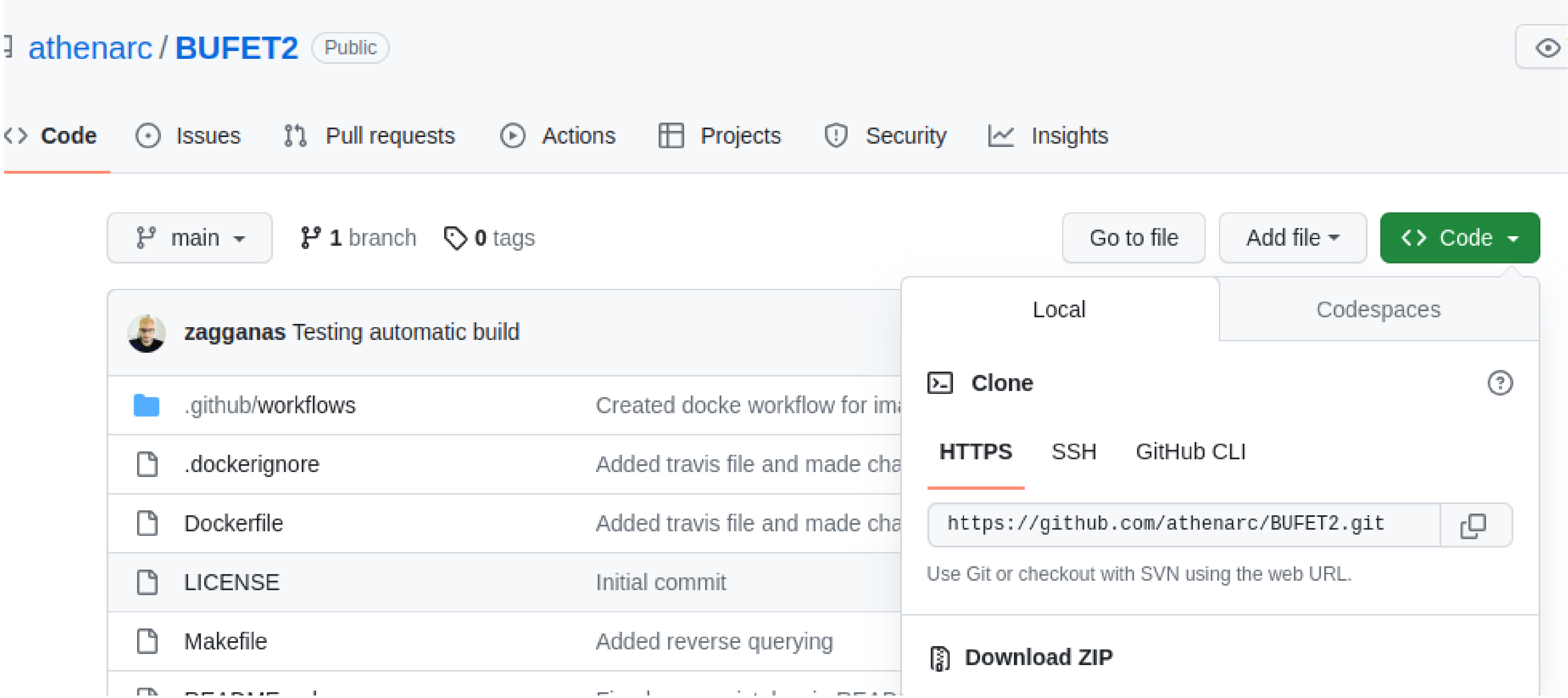
and the relationship between the disease size and the distance between the two distribution.

This can maybe be explained by the fact that **the text mining tools** used to compile the database, utilize structured vocabularies and ontologies.

Thus, the hierarchy existing between the diseases introduces the same bias as seen for GO.

KEGG has the same effect. This could maybe be attributed to complex interactions between genes in pathways that are not specified in the data set.

BUFET2



athenarc / BUFET2 Public

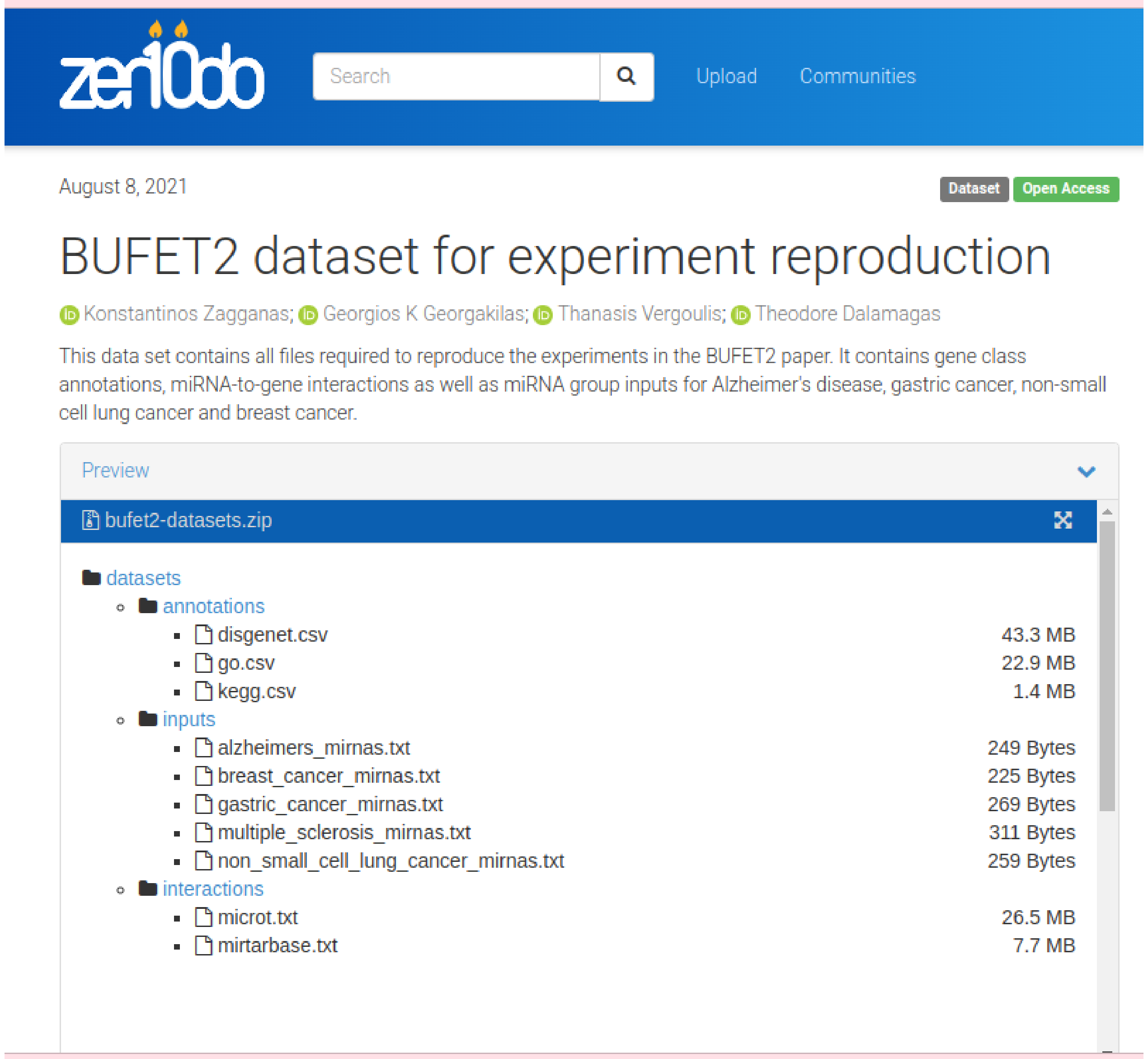
zagganas Testing automatic build

.github/workflows	Created docke workflow for ima
.dockerignore	Added travis file and made cha
Dockerfile	Added travis file and made cha
LICENSE	Initial commit
Makefile	Added reverse querying

```
https://github.com/athenarc/BUFET2.git
```

Use Git or checkout with SVN using the web URL.

EXAMPLE DATA



August 8, 2021

BUFET2 dataset for experiment reproduction

Konstantinos Zagganas; Georgios K Georgakilas; Thanasis Vergoulis; Theodore Dalamagas

This data set contains all files required to reproduce the experiments in the BUFET2 paper. It contains gene class annotations, miRNA-to-gene interactions as well as miRNA group inputs for Alzheimer's disease, gastric cancer, non-small cell lung cancer and breast cancer.

bufet2-datasets.zip

- datasets
  - annotations

disgenet.csv	43.3 MB
go.csv	22.9 MB
kegg.csv	1.4 MB

  - inputs

alzheimers_mirnas.txt	249 Bytes
breast_cancer_mirnas.txt	225 Bytes
gastric_cancer_mirnas.txt	269 Bytes
multiple_sclerosis_mirnas.txt	311 Bytes
non_small_cell_lung_cancer_mirnas.txt	259 Bytes

  - interactions

microt.txt	26.5 MB
mirtarbase.txt	7.7 MB

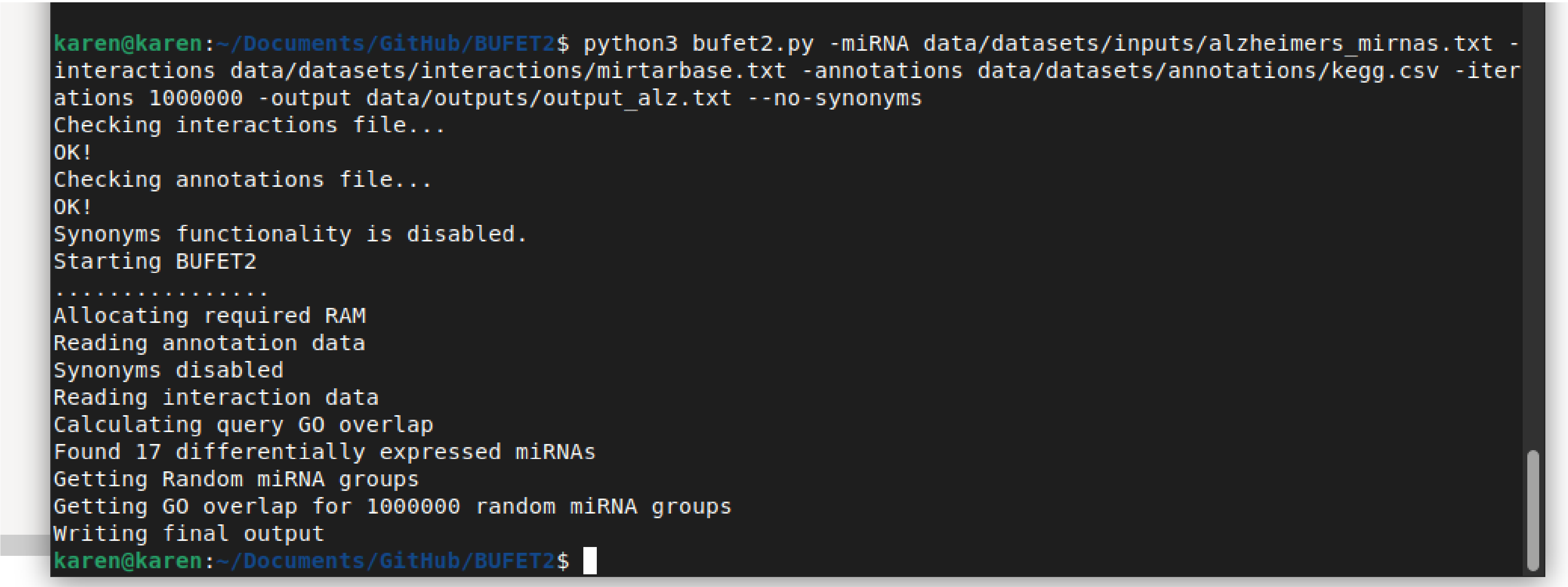
RUNNING BUFET2

```
python3 bufet2.py -miRNA microRNA_list.txt -interactions microRNA:Target_interactions.txt -annotations functional_annotations.csv -iterations 1000000 -output output_file.txt --no-synonyms
```

RUNNING BUFET2

```
python3 bufet2.py -miRNA /data/datasets/inputs/alzheimers_mirnas.txt -interactions /data/datasets/interactions/mirtarbase.txt -annotations /data/datasets/annotations/kegg.csv -iterations 1000000 -output /data/outputs/output_alz.txt --no-synonyms
```

IN CONSOLE



```
karen@karen:~/Documents/GitHub/BUFET2$ python3 bufet2.py -miRNA data/datasets/inputs/alzheimers_mirnas.txt -
interactions data/datasets/interactions/mirtarbase.txt -annotations data/datasets/annotations/kegg.csv -iter
ations 1000000 -output data/outputs/output_alz.txt --no-synonyms
Checking interactions file...
OK!
Checking annotations file...
OK!
Synonyms functionality is disabled.
Starting BUFET2
................
Allocating required RAM
Reading annotation data
Synonyms disabled
Reading interaction data
Calculating query GO overlap
Found 17 differentially expressed miRNAs
Getting Random miRNA groups
Getting GO overlap for 1000000 random miRNA groups
Writing final output
karen@karen:~/Documents/GitHub/BUFET2$ 
```

OUTPUT

#GO-term-ID	GO-term-size	Observed-Target-Left-Sided...	Mean-Random-Simulated-Left...	Left-sided-empirical-p-value	Observed-Target-Two-Sided...	Mean-Random-Simulated-Two...	Two-sided-empirical-p-value
hsa04974~Protein digestion and absorption	103	0.003261	0.004142	0.739577	0.003189	0.003946	0.715527
hsa04971~Gastric acid secretion	76	0.004014	0.005203	0.800503	0.003955	0.005029	0.782733
hsa04966~Collecting duct acid secretion	27	0.001756	0.001667	0.429571	0.001747	0.001647	0.424004