Update: Last Week of BRE

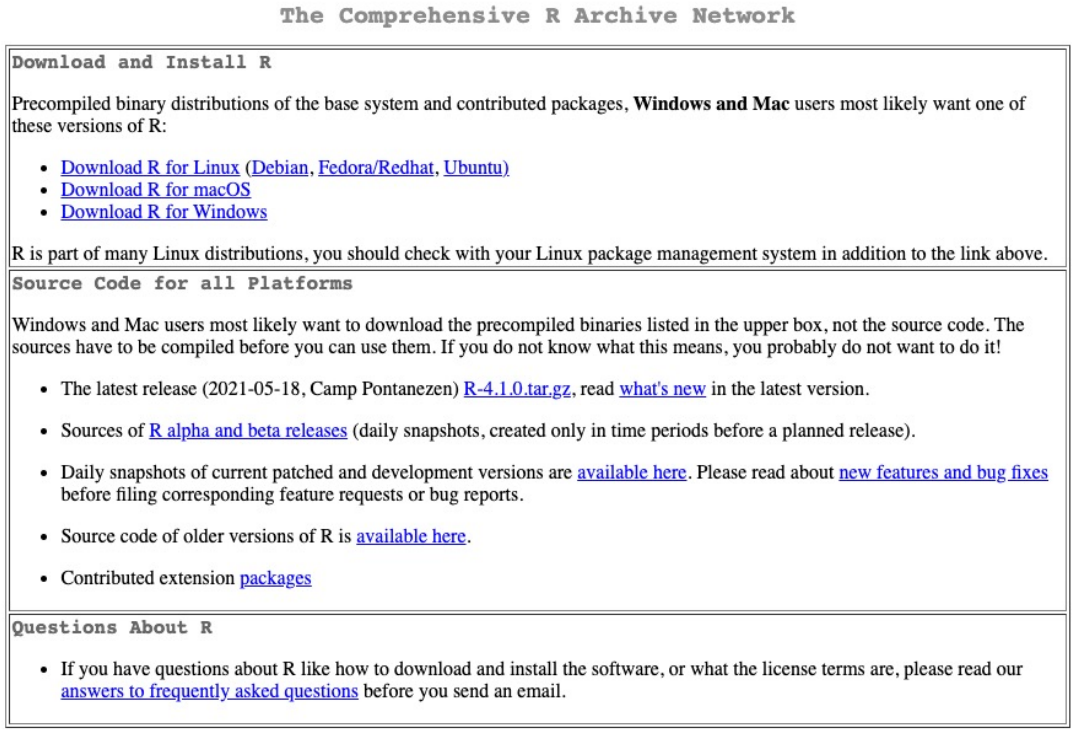
- No more homework – only have your RNA-seq assignment to complete
- No real lectures Thursday and Friday, just Journal Club in the morning
  - Thursday afternoon you will meet with me to practice your presentation for Friday
  - There will be office hours Friday morning after journal club for any last-minute questions or presentation practice
- Final presentation Friday on your RNA-seq results to which all Coriell employees will be invited

Date	Lecture
Tuesday, 7/27	Gene Set Enrichment Analysis
Wednesday, 7/28	Regular Expressions
Thursday, 7/29	Quick example RNA-seq presentation, just journal club
Friday, 7/30	No lecture, just journal club, following by morning office hours

Gene Set Enrichment Analysis

2021-07-23

CRAN and Bioconductor



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

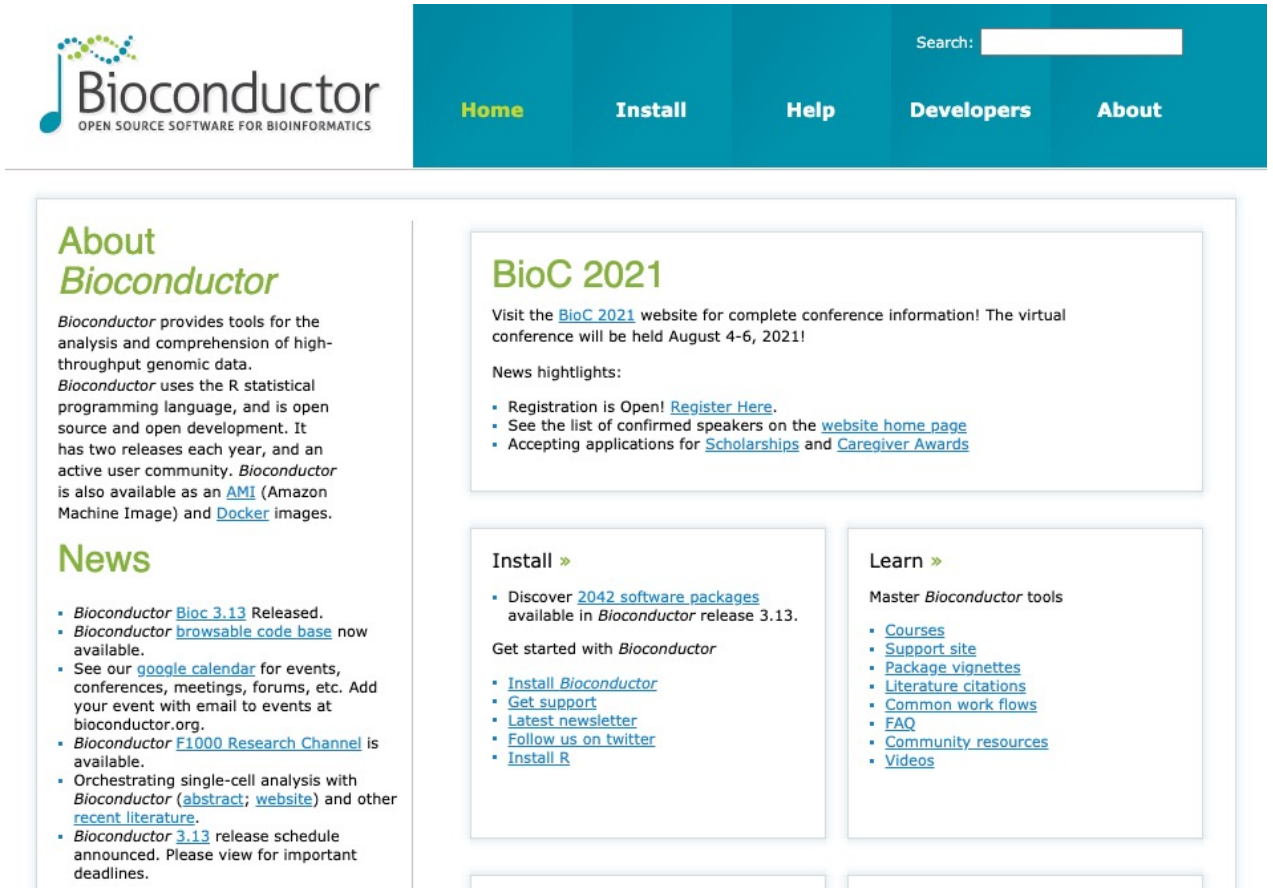
Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2021-05-18, Camp Pontanezen) R-4.1.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. *Bioconductor* uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as an AMI (Amazon Machine Image) and Docker images.

News

- *Bioconductor* Bioc 3.13 Released.
- *Bioconductor* browsable code base now available.
- See our google calendar for events, conferences, meetings, forums, etc. Add your event with email to events at bioconductor.org.
- *Bioconductor* F1000 Research Channel is available.
- Orchestrating single-cell analysis with *Bioconductor* (abstract; website) and other recent literature.
- *Bioconductor* 3.13 release schedule announced. Please view for important deadlines.

BioC 2021

Visit the BioC 2021 website for complete conference information! The virtual conference will be held August 4-6, 2021!

News hightlights:

- Registration is Open! Register Here.
- See the list of confirmed speakers on the website home page
- Accepting applications for Scholarships and Caregiver Awards

Install »

- Discover 2042 software packages available in *Bioconductor* release 3.13.

Get started with *Bioconductor*

- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

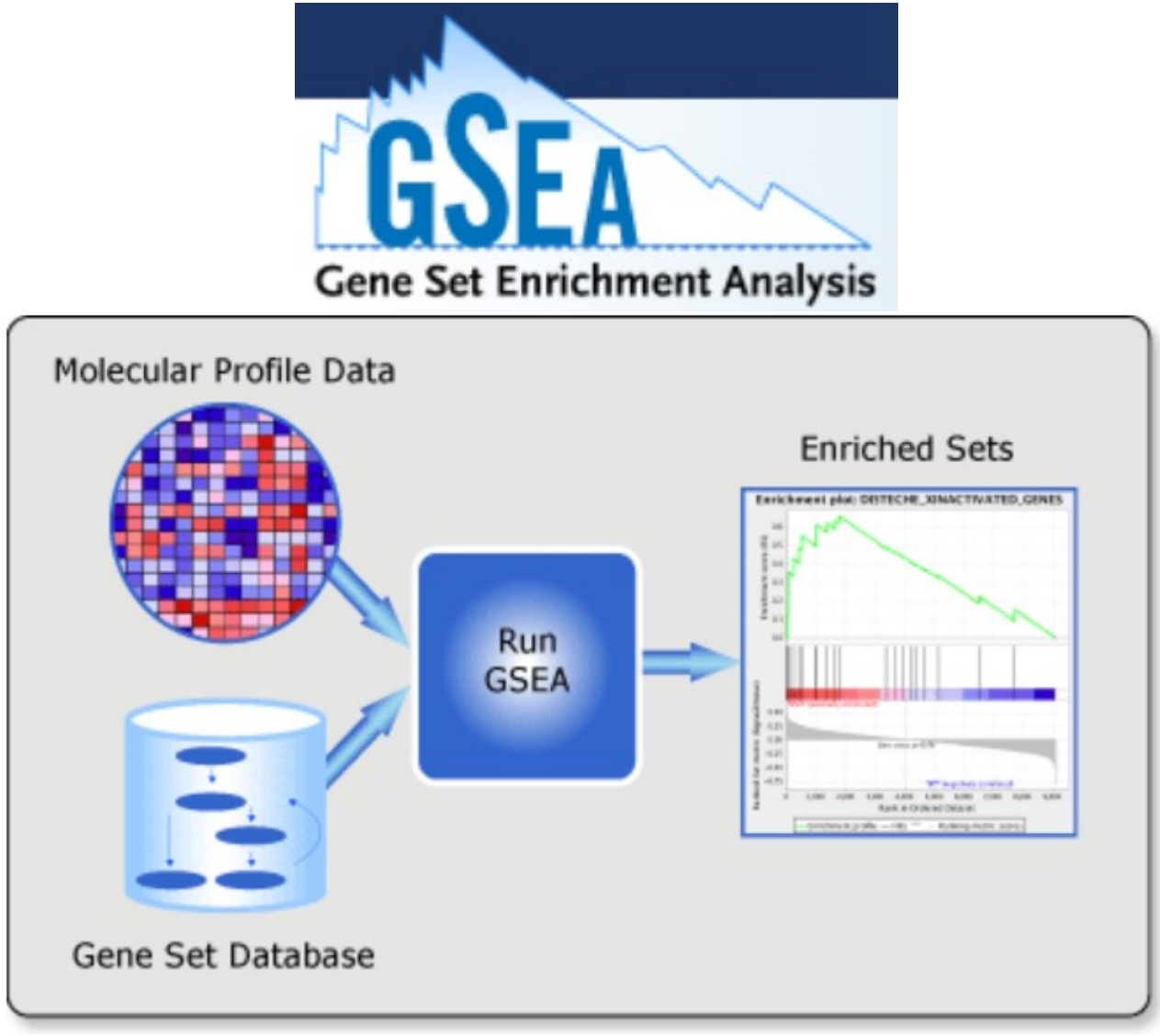
Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

What is Gene Set Enrichment Analysis?

- Problem with RNA-seq is that it's hard to derive the meaning in a list of genes.
- Gene Set Enrichment Analysis (GSEA) looks for coordinated changes in gene sets.
- Gene sets are frequently pathways, but you can use GSEA for any set of genes.



UC San Diego

BROAD INSTITUTE

How is GSEA calculated?

- For this example, we'll calculate the enrichment score for the Reactome pathway "HDMS demethylate histones"
  - Histone demethylase (HDM)
  - Contains all KDM, JDM genes

Gene	Fold Change
KDM1A	4
NCAM2	-2
ACTB	-0.01
KDM1B	3.8
SETD4	3.6
GAPDH	0.05
KDM2A	3.5
KDM2B	2.8
RAD51	-3
ERCC2	1.2

How is GSEA calculated?

- For this example, we'll calculate the enrichment score for the Reactome pathway "HDMS demethylate histones"
  - Histone demethylase (HDM)
  - Contains all KDM, JDM genes

1. Rank genes by change in expression from least to greatest significance

Gene	Rank	Fold Change
RAD51	1	-0.53
NCAM2	2	-0.22
ACTB	3	-0.01
GAPDH	4	0.05
ERCC2	5	1.20
KDM2B	6	2.80
KDM2A	7	3.50
SETD4	8	3.60
KDM1B	9	3.80
KDM1A	10	4.00

How is GSEA calculated?

- For this example, we'll calculate the enrichment score for the Reactome pathway "HDMS demethylate histones"
  - Histone demethylase (HDM)
  - Contains all KDM, JDM genes

1. Rank genes by change in expression from least to greatest significance
2. Calculate the cumulative sum of the significance over the ranked genes. Subtract the fold change if it's not in the list and add the fold change if it is in the list

Gene	Rank	Fold Change	Cumulative Sum
RAD51	1	-0.53	0.00 – (-0.53) = 0.53
NCAM2	2	-0.22	0.53 – (-0.22) = 0.75
ACTB	3	-0.01	0.75 – (-0.01) = 0.76
GAPDH	4	0.05	0.76 – 0.05 = 0.71
ERCC2	5	1.20	0.71 – 1.2 = -0.49
KDM2B	6	2.80	-0.49 + 2.80 = 2.31
KDM2A	7	3.50	2.31 + 3.50 = 5.81
SETD4	8	3.60	5.81 - 3.60 = 2.20
KDM1B	9	3.80	2.20 + 3.80 = 6.00
KDM1A	10	4.00	6.00 + 4.00 = 10.00

How is GSEA calculated?

- For this example, we'll calculate the enrichment score for the Reactome pathway "HDMS demethylate histones"
  - Histone demethylase (HDM)
  - Contains all KDM, JDM genes

1. Rank genes by change in expression from least to greatest significance
2. Calculate the cumulative sum of the significance over the ranked genes. Subtract the fold change if it's not in the list and add the fold change if it is in the list
3. Take the largest deviation from 0 as the enrichment score.

Gene	Rank	t statistic	Cumulative Sum
RAD51	1	-0.53	0.00 – (-0.53) = 0.53
NCAM2	2	-0.22	0.53 – (-0.22) = 0.75
ACTB	3	-0.01	0.75 – (-0.01) = 0.76
GAPDH	4	0.05	0.76 – 0.05 = 0.71
ERCC2	5	1.20	0.71 – 1.2 = -0.49
KDM2B	6	2.80	-0.49 + 2.80 = 2.31
KDM2A	7	3.50	2.31 + 3.50 = 5.81
SETD4	8	3.60	5.81 - 3.60 = 2.20
KDM1B	9	3.80	2.20 + 3.80 = 6.00
KDM1A	10	4.00	6.00 + 4.00 = 10.00

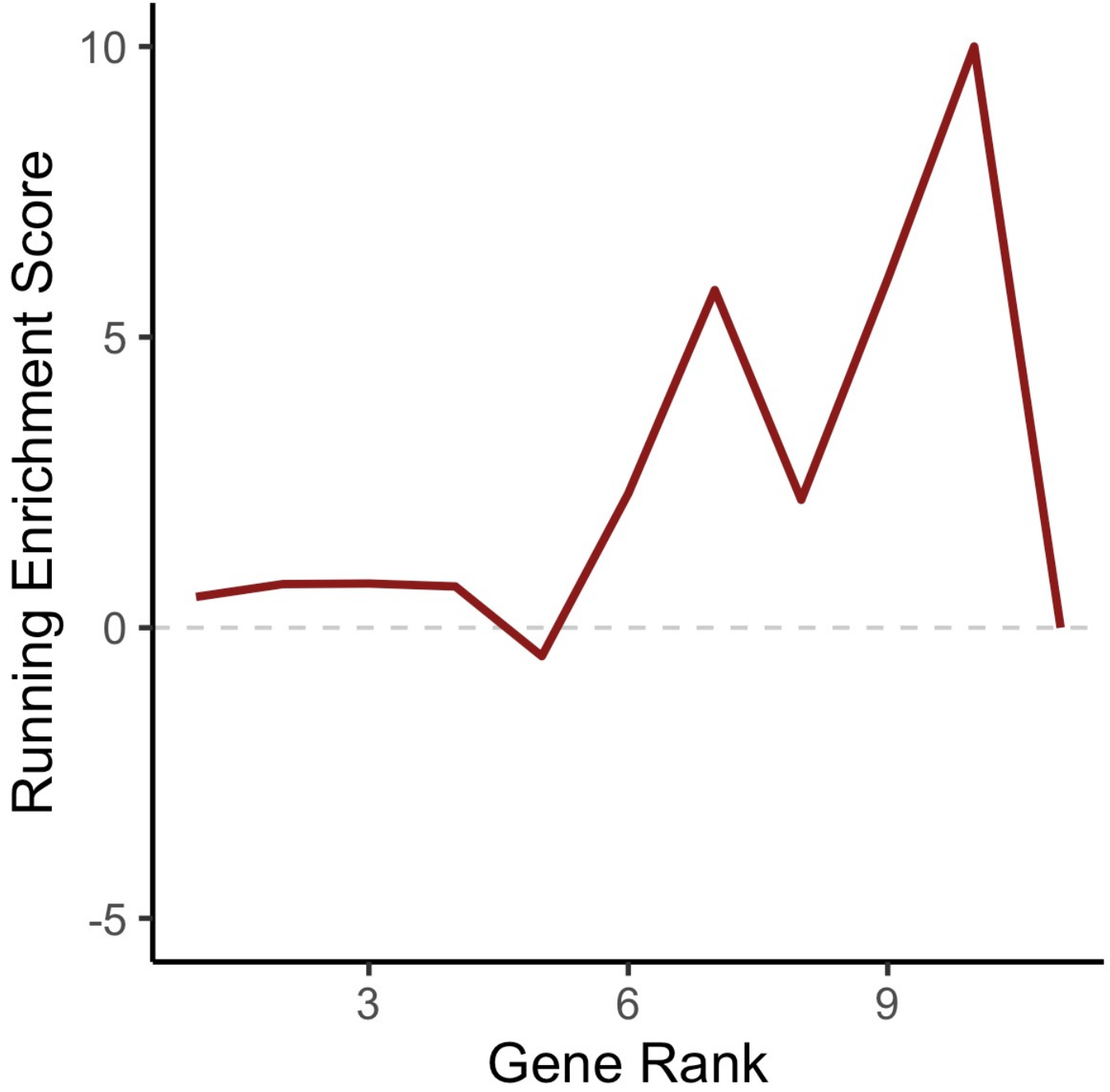
ES = 10

How is GSEA calculated?

- For this example, we'll calculate the enrichment score for the Reactome pathway "HDMS demethylate histones"
  - Histone demethylase (HDM)
  - Contains all KDM, JDM genes

1. Rank genes by change in expression from least to greatest significance
2. Calculate the cumulative sum of the significance over the ranked genes. Subtract the fold change if it's not in the list and add the fold change if it is in the list
3. Take the largest deviation from 0 as the enrichment score.

- You can visualize this with a cumulative distribution plot



How do you get the significance of the enrichment score?

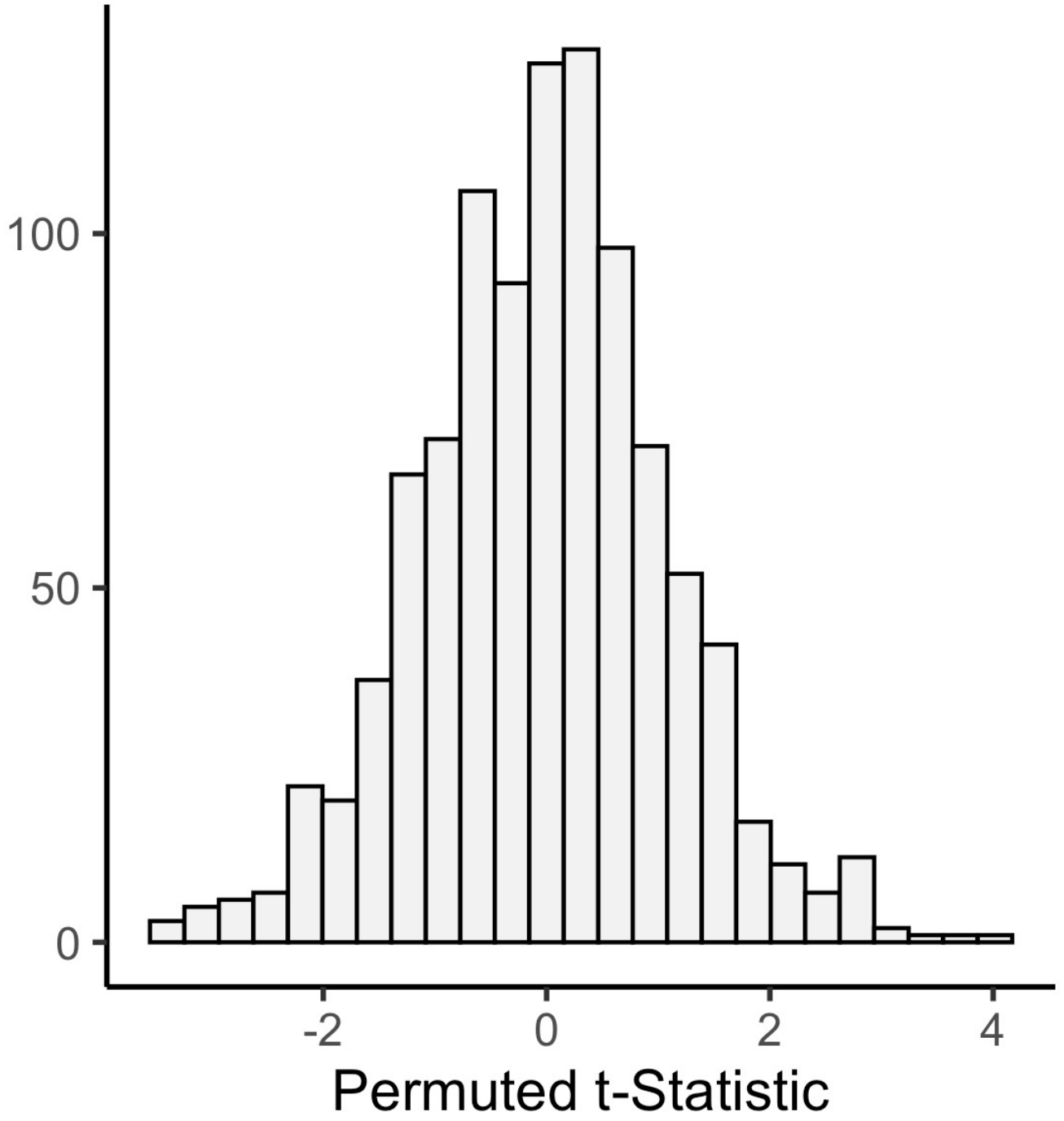
1. Permute the whether the gene is in the pathway 1,000 times

Gene	Gene	Gene
KDM1A	KDM1A	**KDM1A**
NCAM2	NCAM2	NCAM2
ACTB	**ACTB**	ACTB
KDM1B	**KDM1B**	KDM1B
SETD4	SETD4	**SETD4**
GAPDH	**GAPDH**	GAPDH
KDM2A	KDM2A	KDM2A
KDM2B	KDM2B	KDM2B
RAD51	RAD51	**RAD51**
ERCC2	**ERCC2**	**ERCC2**

X 1,000

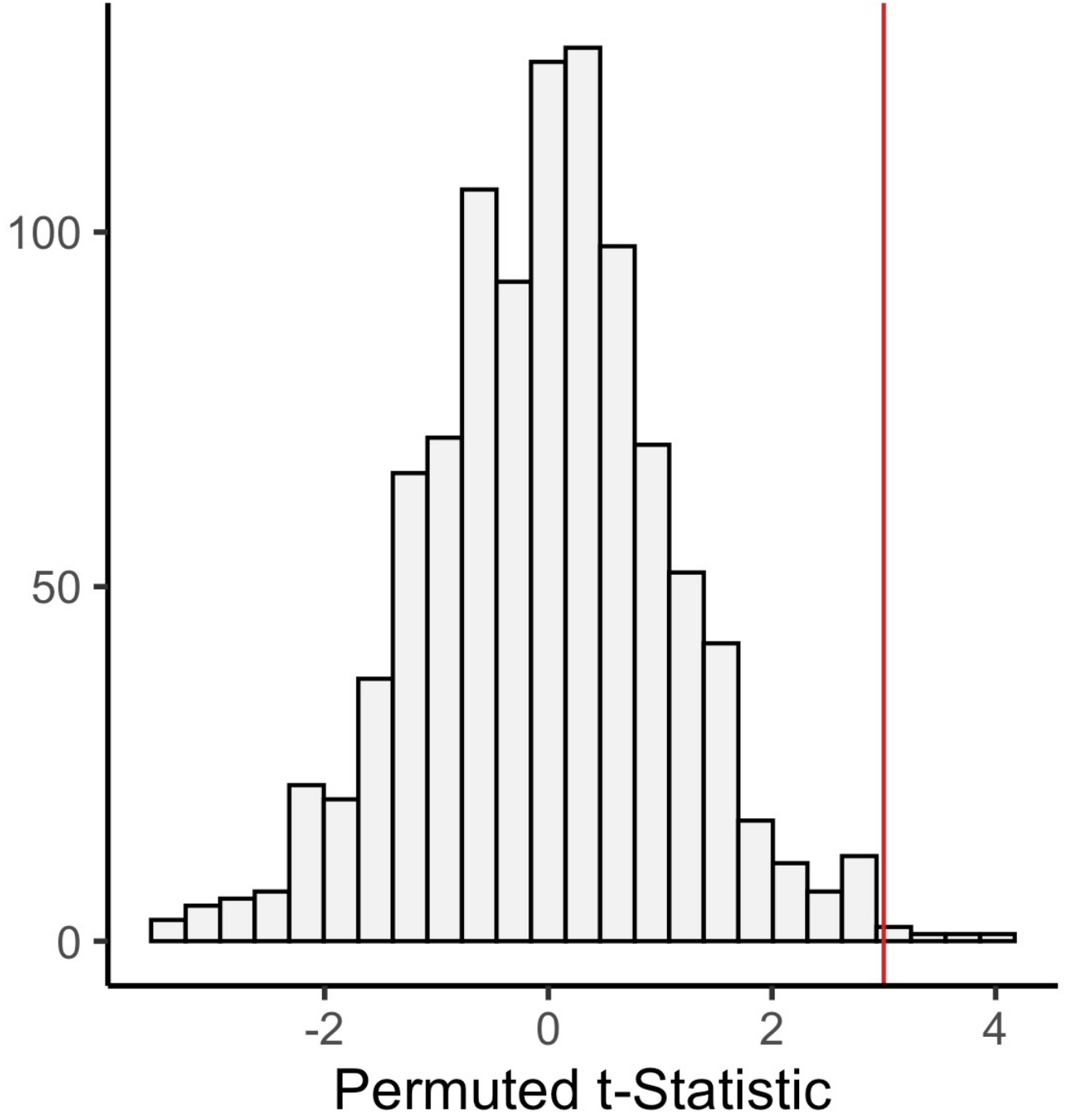
How do you get the significance of the enrichment score?

1. Permute the whether the gene is in the pathway 1,000 times
2. Calculate the significance of the enrichment score for each permutation (t-statistic).



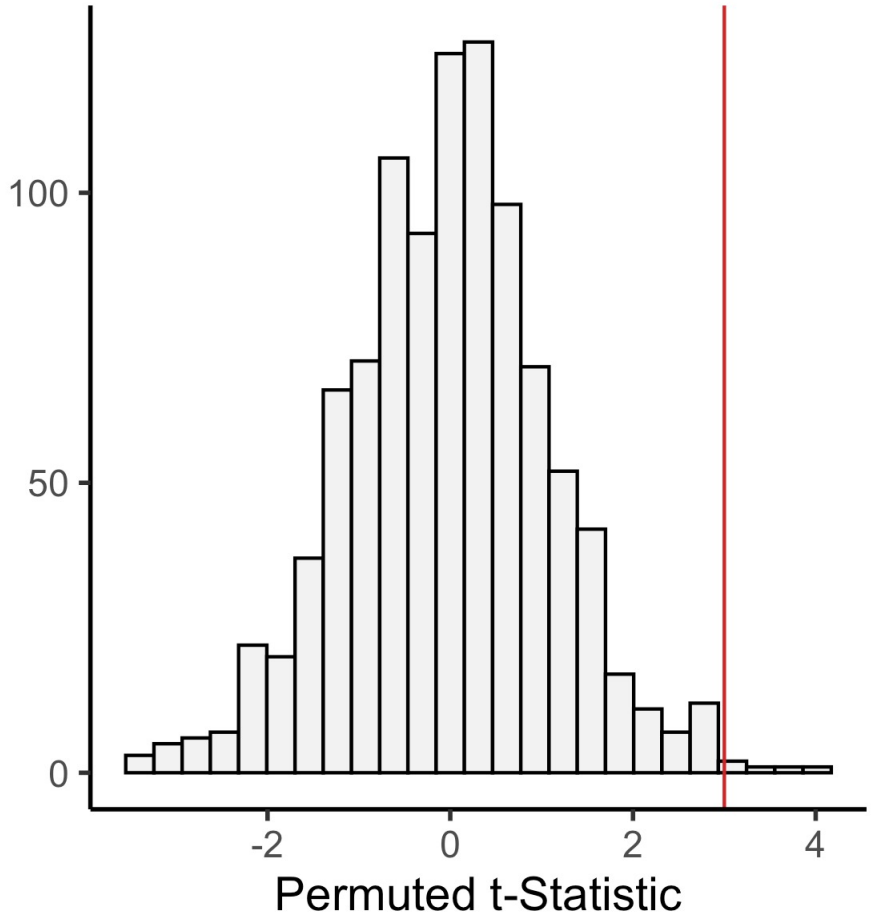
How do you get the significance of the enrichment score?

1. Permute the whether the gene is in the pathway 1,000 times
2. Calculate the significance of the enrichment score for each permutation (t-statistic).
3. Find where our score lies in the distribution



How do you get the significance of the enrichment score?

1. Permute the whether the gene is in the pathway 1,000 times
2. Calculate the significance of the enrichment score for each permutation (t-statistic).
3. Find where our score lies in the distribution
4. The significance, the empirical p-value, is the number of times the enrichment score was greater than or equal to the observed enrichment score divided by the number of permutations

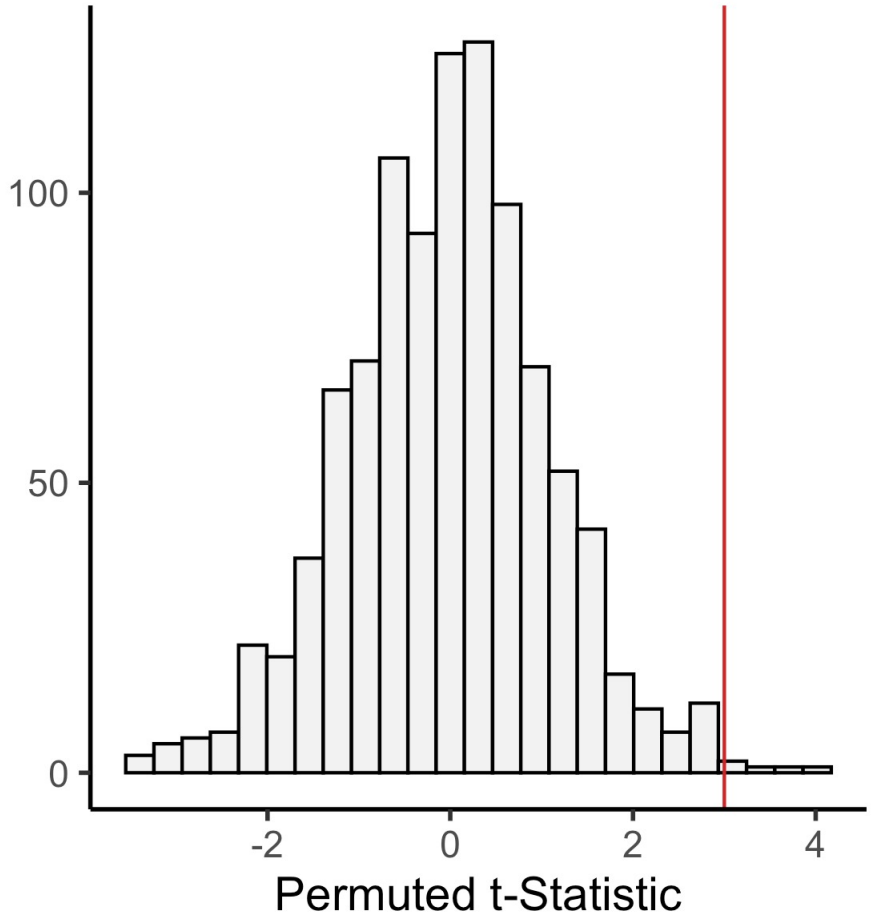


Less than ES	996
Greater than or equal to ES	4

p = 4 / 1000 = **0.0004**

How do you get the significance of the enrichment score?

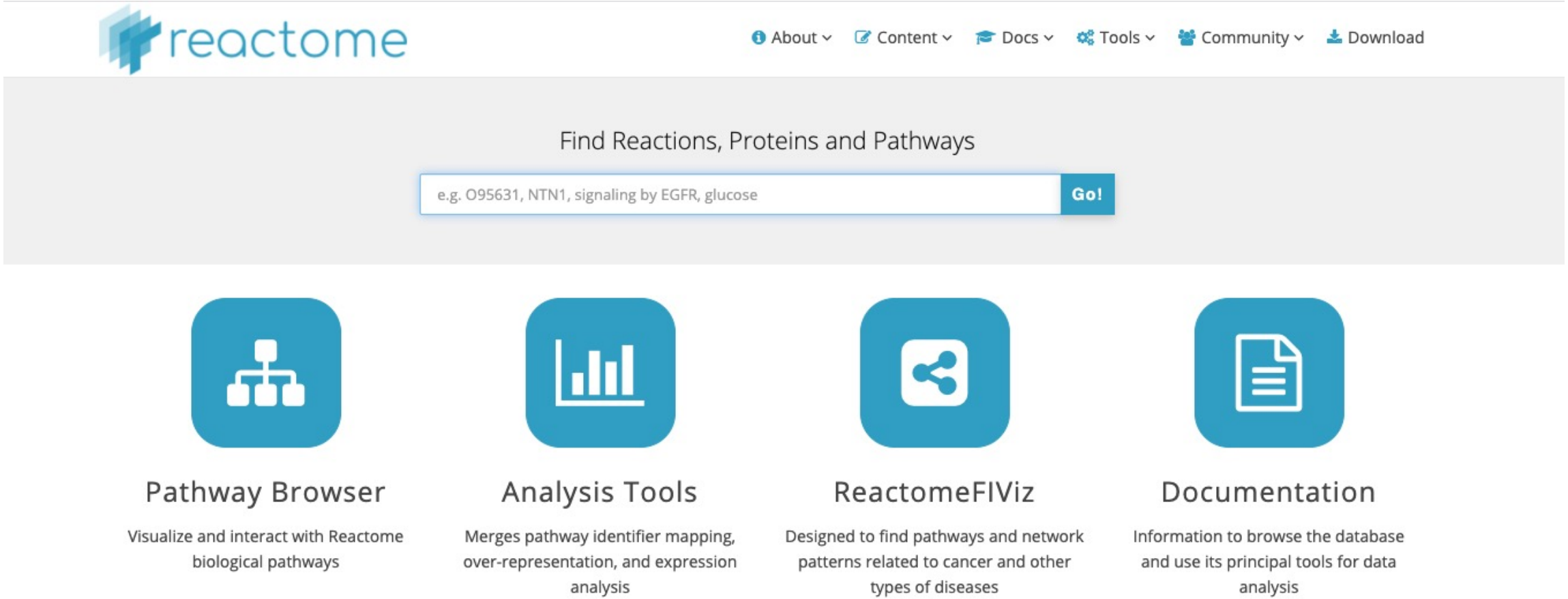
1. Permute the whether the gene is in the pathway 1,000 times
2. Calculate the significance of the enrichment score for each permutation (t-statistic).
3. Find where our score lies in the distribution
4. The significance, the empirical p-value, is the number of times the enrichment score was greater than or equal to the observed enrichment score divided by the number of permutations
5. When testing many pathways at once, the enrichment scores will be normalized by the size of the pathway and the p-values will be corrected for multiple testing.



Less than ES	996
Greater than or equal to ES	4

p = 4 / 1000 = **0.0004**

reactomedb


reactome
About
Content
Docs
Tools
Community
Download
Find Reactions, Proteins and Pathways
e.g. O95631, NTN1, signaling by EGFR, glucose
Go!
Pathway Browser
Visualize and interact with Reactome biological pathways
Analysis Tools
Merges pathway identifier mapping, over-representation, and expression analysis
ReactomeFIViz
Designed to find pathways and network patterns related to cancer and other types of diseases
Documentation
Information to browse the database and use its principal tools for data analysis

Assign RNA-seq project

- Using publicly available data downloaded from GEO, a repository for sequencing data
- Work through everything we've learned with RNA-seq
- Like exploratory data analysis project, will submit a report and give a presentation (see assignment for details)
- The presentation will be this Friday at 12PM and all Coriell employees will be invited.