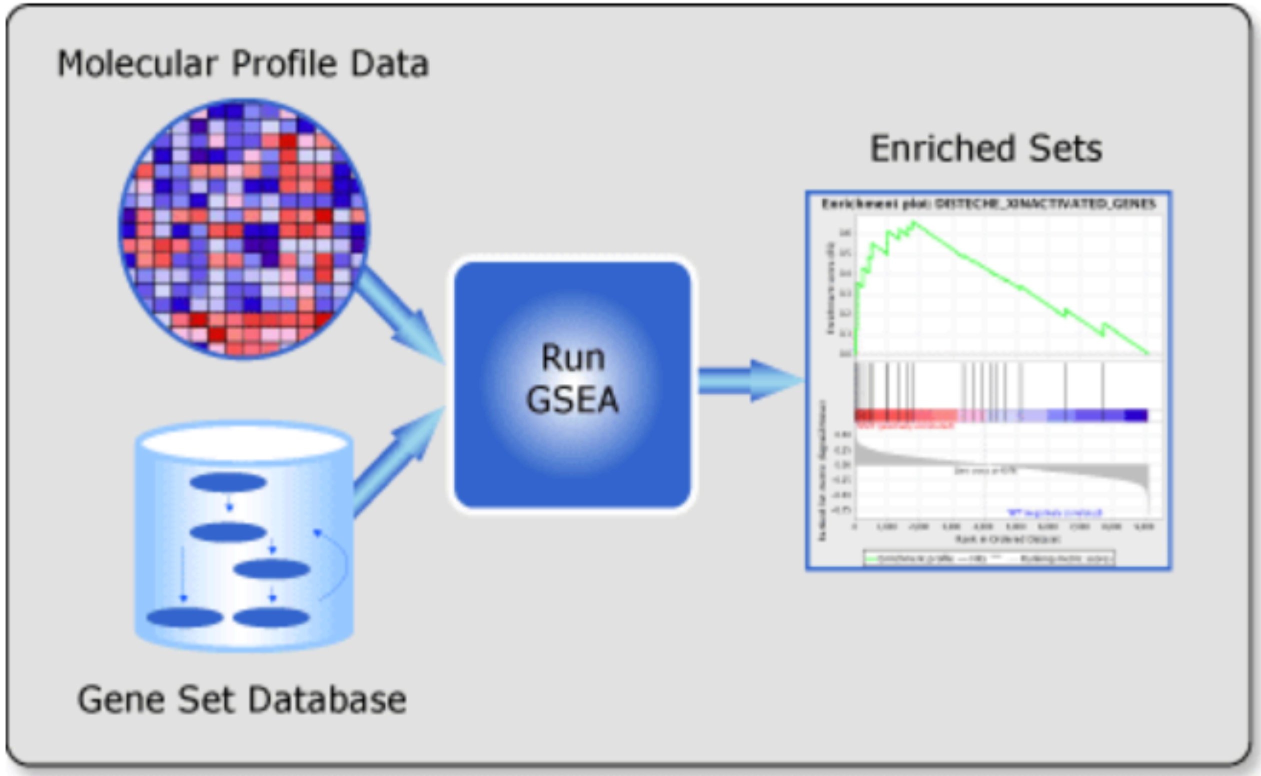
How Does Gene Set Enrichment Analysis Work?

2019-02-11

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.

GSEA
Gene Set Enrichment Analysis



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

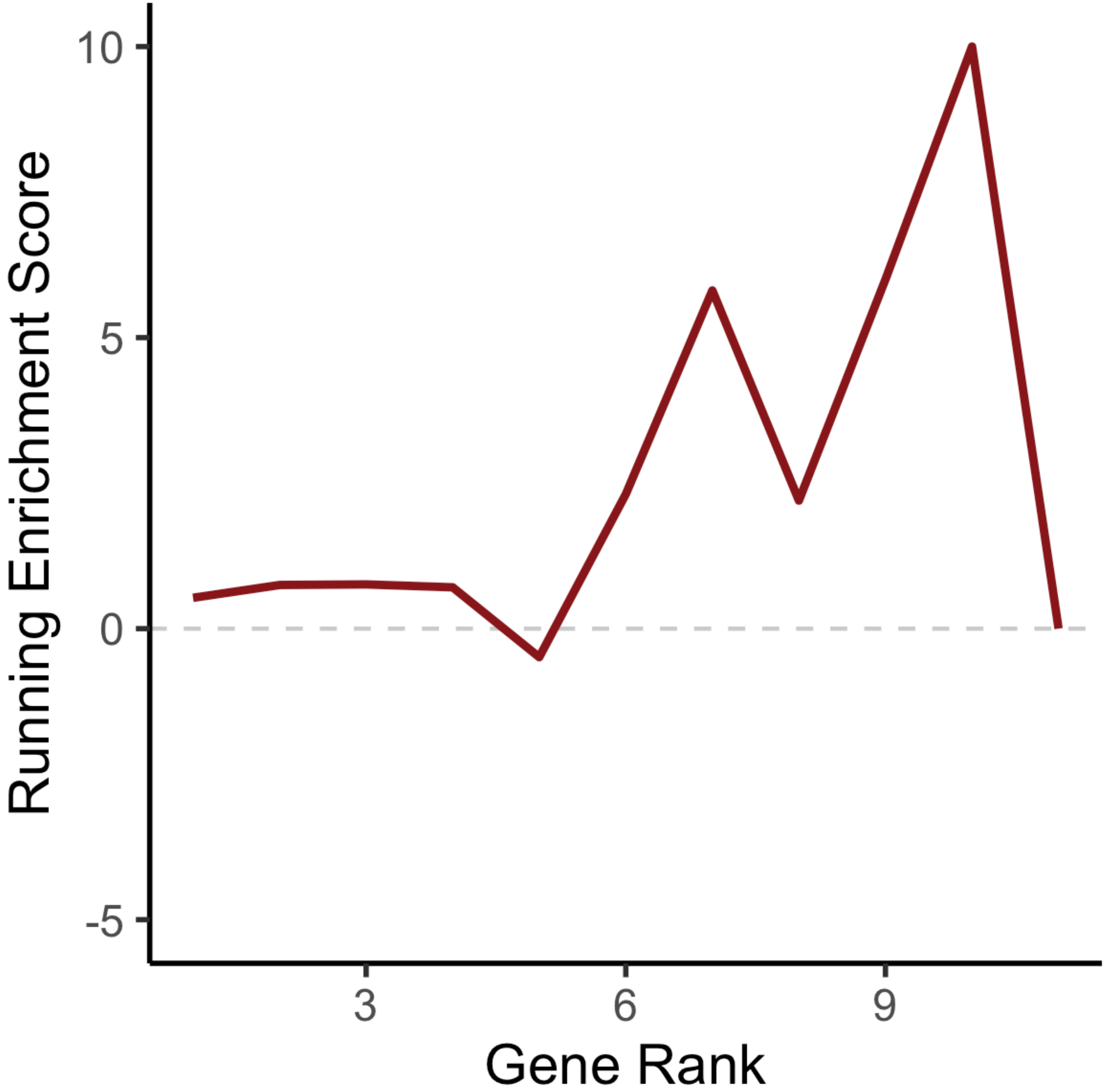
ES = 10

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?

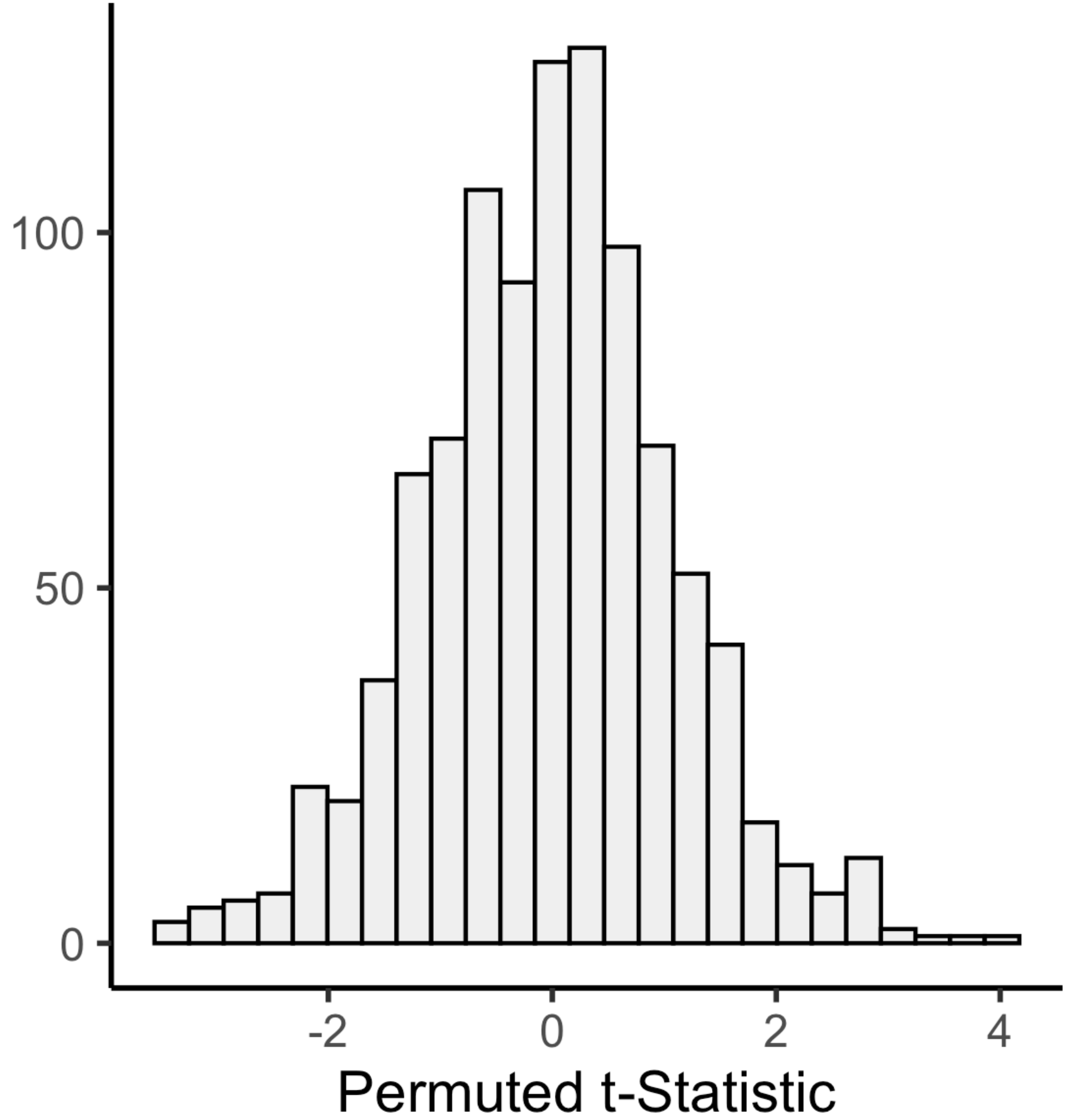
1. Permute the whether the gene is in the pathway 1,000 times

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

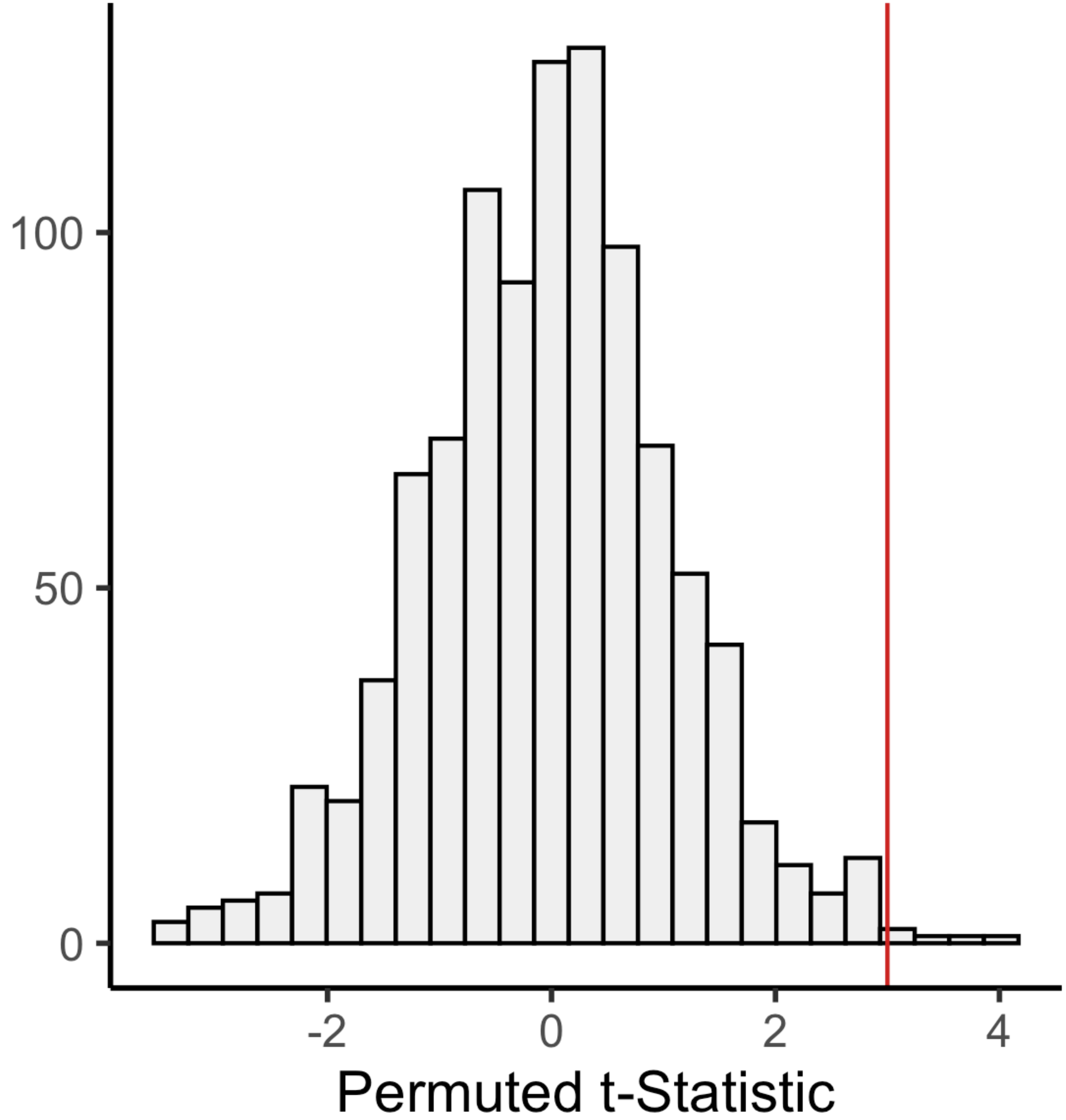
How do you get the significance of the enrichment score?

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).



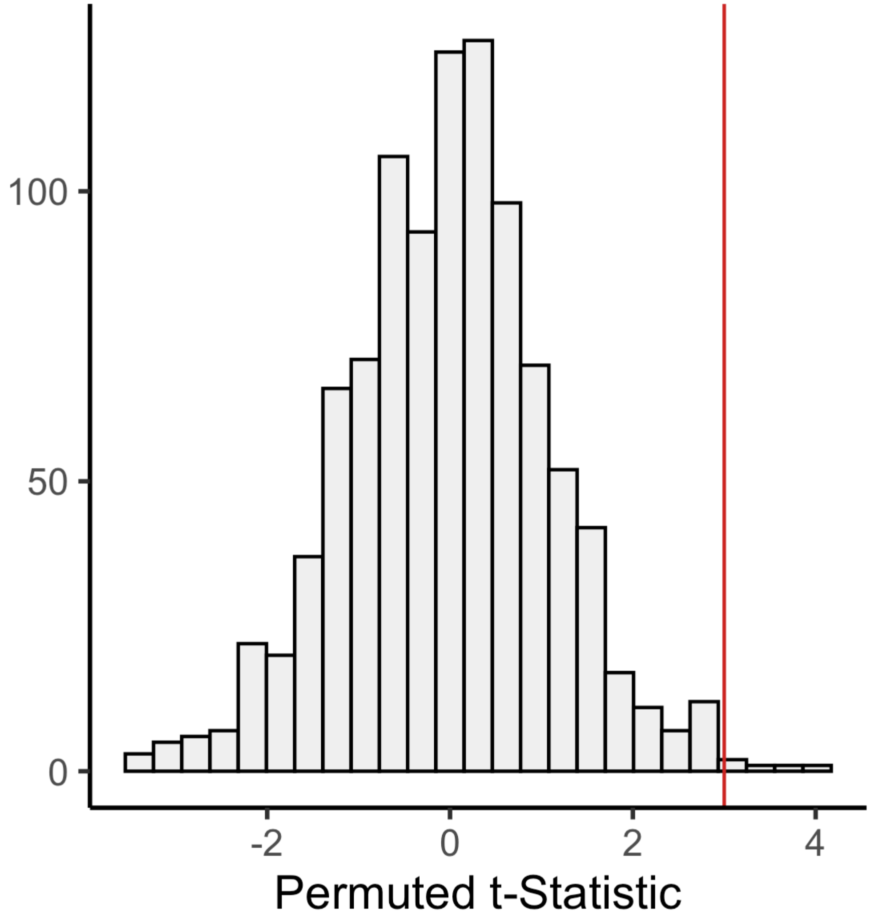
How do you get the significance of the enrichment score?

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

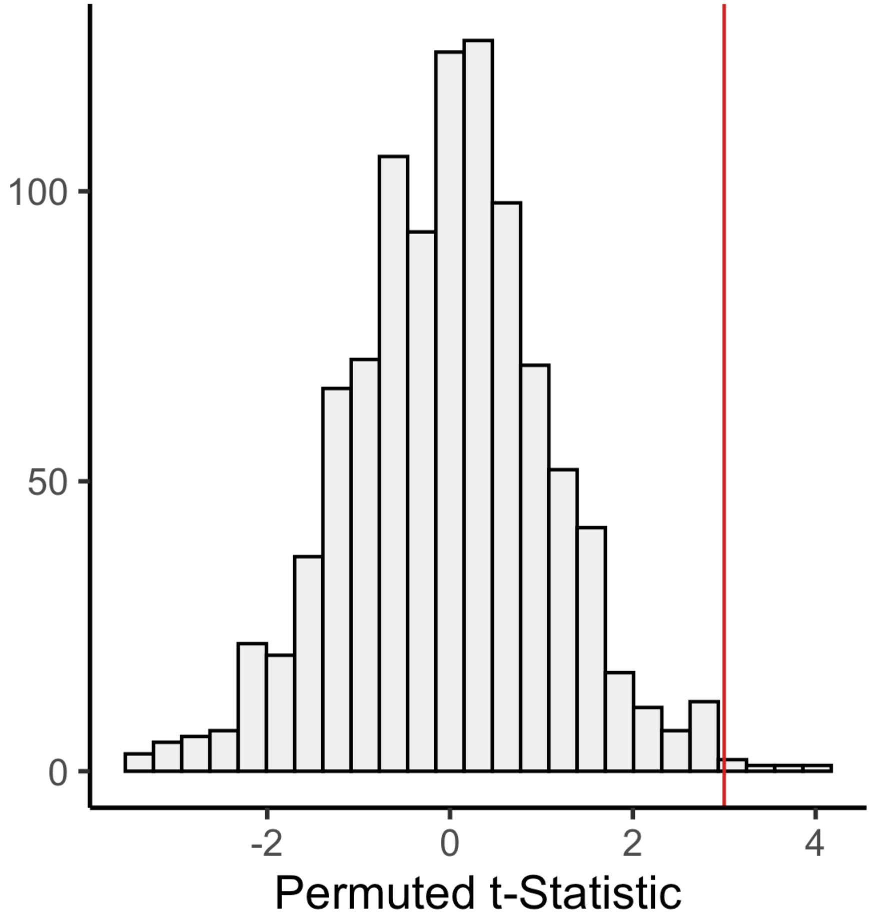


Less than ES	996
Greater than or equal to ES	4

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**