THRESHOLD - *A novel gene saturation analysis GUI*

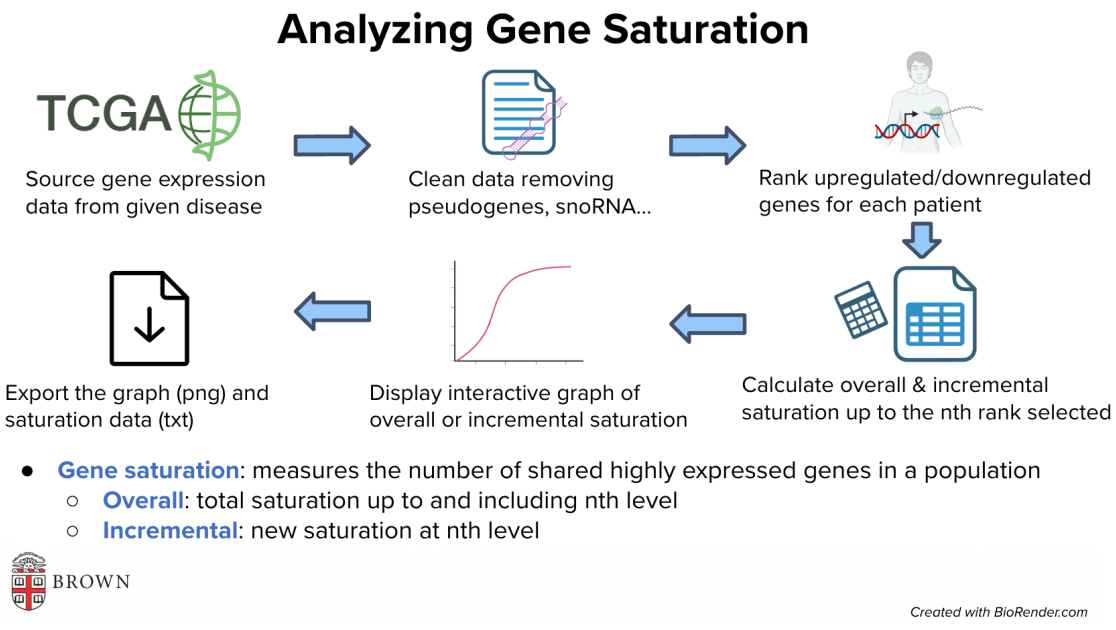
Authors

This project was developed by Finán Gammel, Jennifer Li and Dr. Alper Uzun in the Uzun Lab at Brown University.

Purpose

THRESHOLD analyzes transcriptomic data across large samples of patients to understand the cohesion of the most upregulated/downregulated genes in a given disease. This lends researchers knowledge to inform network topology analyses or to assess the relative amount of genes influencing a given disease. THRESHOLD offers several features to aid in analysis including user-inputted saturation type, restriction factors, and rank type. The tool outputs an interactive graph of saturation permitting the calculation of specific saturation thresholds and most saturated genes.

Overall Workflow



- **Gene saturation**: measures the number of shared highly expressed genes in a population
  - **Overall**: total saturation up to and including nth level
  - **Incremental**: new saturation at nth level

**Alternatively, genes may be ranked from low to high.*

Overall Saturation - The saturation of all the genes up to and including the nth level. Quantifies the count of genes up to the nth level that exceed the inputted restriction level, divided by the total count of genes up to and including the nth level.

Formula:

Overall Saturation =

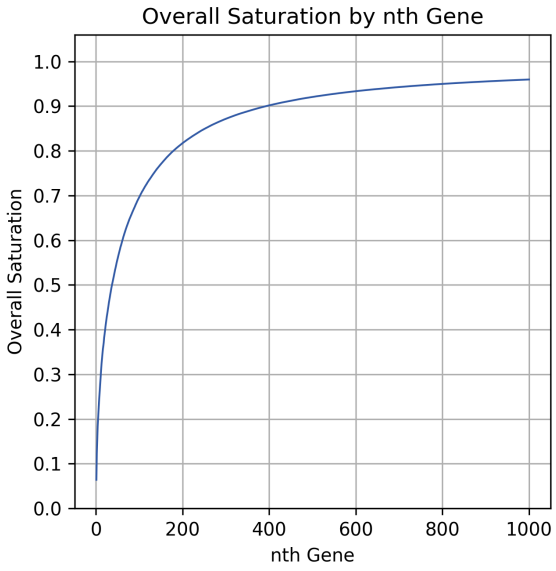
$$\frac{\sum_{e \in \{y|y \in A\}} max(|[y \in A|y = e]| - x, 0)}{|A|}$$

A = the multiset of genes up to and including the nth level

x = the restriction level

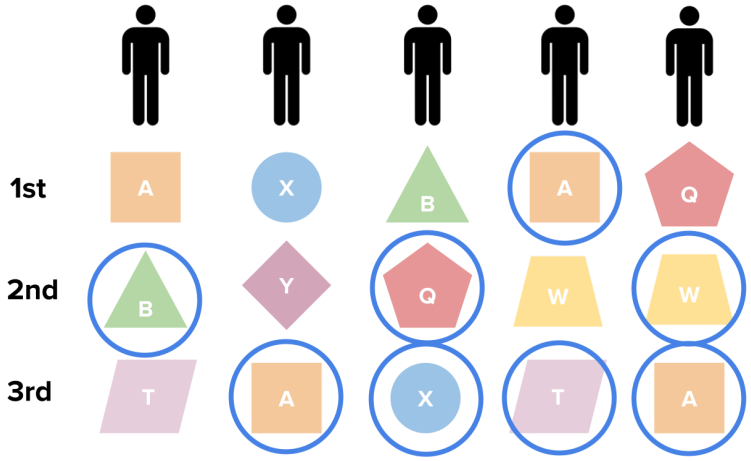
Conceptual Idea - The sum of each unique element in the multiset A's non-negative greater value between the element count in multiset A minus the restriction level divided by the number of elements in the multiset A.

Graphical Output Example:



Visual Example:

*Find the **overall saturation** up to the **3rd level** (n = 3) given a **restriction level** of **1**:*



Saturated Genes; occur > 1 time (restriction level = 1)

8 genes saturated / 15 genes up to and including nth level = ~53% overall saturation

nth level	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
1st	Gene A	Gene X	Gene B	Gene A	Gene Q
2nd	Gene B	Gene Y	Gene Q	Gene W	Gene W
3rd	Gene T	Gene A	Gene X	Gene T	Gene A

- All genes present up to and including n = {A, X, B, A, Q, B, Y, Q, W, W, T, A, X, T, A}
    - Occurrences by element: A = 4, X = 2, B = 2, Q = 2, Y = 1, W = 2, T = 2
- Saturated genes (subtract the restriction level):

  A: 4 - 1 = 3 B: 2 - 1 = 1 Y: 1 - 1 = 0 T: 2 - 1 = 1

  X: 2 - 1 = 1 Q: 2 - 1 = 1 W: 2 - 1 = 1
- *Occurrences > 1 (restriction level)*: 3 + 1 + 1 + 1 + 0 + 1 + 1 = **8**
- 8 saturated genes / 15 genes up to and including n = **~53% overall saturation**

Incremental Saturation - The saturation only at the incremental nth gene level. Quantifies the count of genes at the nth level that exceed the inputted restriction level up to and including that level, divided by the total count of genes at the nth level.

Formula:

Incremental Saturation =

$$\frac{\sum_{e \in \{y|y \in B\}} max(min(|[y \in A|y = e]| - x, |[y \in B|y = e]|), 0)}{|B|}$$

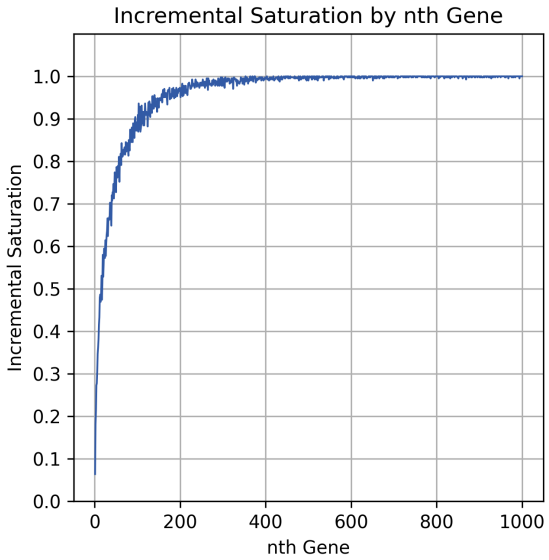
A = the multiset of genes up to and including the nth level

B = the multiset of genes at the nth level

x = restriction level

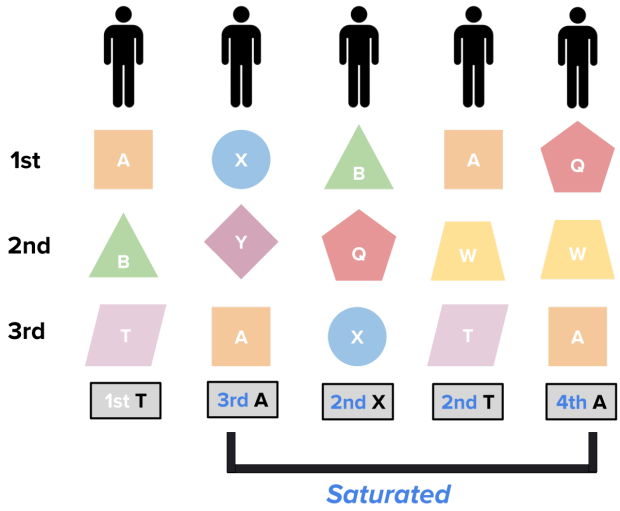
Conceptual Idea - The sum of each unique element in the multiset B's non-negative lesser value between the difference in the element count in multiset A minus the restriction level and the element count in multiset B divided by the number of elements in multiset B.

Graphical Output Example:



Visual Example:

*Find the **incremental saturation** of the **3rd level** (n = 3) given a **restriction level** of **1**:*



At n = 3 (third level) there are 4 saturated genes out of 5 genes total; 4 / 5 = 80% incremental saturation

nth level	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
1st	Gene A	Gene X	Gene B	Gene A	Gene Q
2nd	Gene B	Gene Y	Gene Q	Gene W	Gene W
3rd	Gene T	Gene A	Gene X	Gene T	Gene A

- Genes present up to and including n; A = {A, X, B, A, Q, B, Y, Q, W, W, T, A, X, T, A}
    - Occurrences by element: A = 4, X = 2, B = 2, Q = 2, Y = 1, W = 2, T = 2
- Genes present at n; B = {T, A, X, T, A}
    - Occurrences by element: T = 2, A = 2, X = 1
- Saturated genes (lesser between count in multiset A minus the restriction level and count in multiset B; must be non-negative)

  T: 2 - 1 = 1 ... 2 in B ... 1 < 2 ➔ 1 X: 2 - 1 = 1 ... 1 in B ... 1 = 1 ➔ 1

  A: 4 - 1 = 3 ... 2 in A ... 2 < 3 ➔ 2
- Occurrences *> 1 (restriction level)*: 1 + 1 + 2 = **4**
- 4 saturated genes / 5 genes at n = **80% incremental saturation**

Standardization

To compare saturation across datasets, restriction levels can be standardized.

- Increasing the restriction level by a factor of the data set size increase and produces the same saturation results and can be used for standardization.
- For example, a data set of 100 patients has a restriction level of 2 genes. To standardize a data set of 200 patients (twice the size), you would need to double the restriction level 2*2 = 4.
- Restriction levels in terms of percents are universal and do not need further standardization (potentially ±1 gene due to rounding)

Dataset Validation:

The following testing exists to verify the calculations outputted by **THRESHOLD**. From our testing, the tool has demonstrated consistent accuracy.

Saturation Validation Testing

Given the following test dataset:

Hugo_Symbol	Entrez_Gene_Id	Pat1	Pat2	Pat3	Pat4
A	-	1	5	1	5
B	-	-3	3	5	0
X	-	5	0	-1	-2
W	-	-1	-1	-2	3
Q	-	-2	-3	3	1
Y	-	3	-2	-3	-3
T	-	-4	1	0	-1

THRESHOLD passed verified calculations all of the following tests:

- *Regular rank (level = 1, n = 5)*
- *Regular rank (level = 2, n = 5)*
- *Regular rank (level = 3, n = 5)*
- *Regular rank (level = 1% n = 5)*
- *Regular rank (level = 60%, n = 5)*
- *Regular rank (level = 70%, n = 5)*
- *Reverse rank (level = 1, n = 5)*
- *Reverse rank (level = 2, n = 5)*
- *Reverse rank (level = 3, n = 5)*
- *Reverse rank (level = 1% n = 5)*
- *Reverse rank (level = 60%, n = 5)*
- *Reverse rank (level = 70%, n = 5)*

Usage:

Download

To install **THRESHOLD** simply follow the installation instructions in the GitHub:

THRESHOLD GitHub

Once you have downloaded the required folder, open the entire folder in a VScode IDE. Ensure the Python and Java extensions are added. When ready, simply navigate to the threshold.gui file and press run to begin. The GUI window should appear.

Input File Format

THRESHOLD requires the input of a file of patient transcriptomic data with gene expression data in the form of zscores or percentiles comparing expression against a control population or ranked relatively within an individual patient's expression.

The inputted .txt file must be formatted as such:

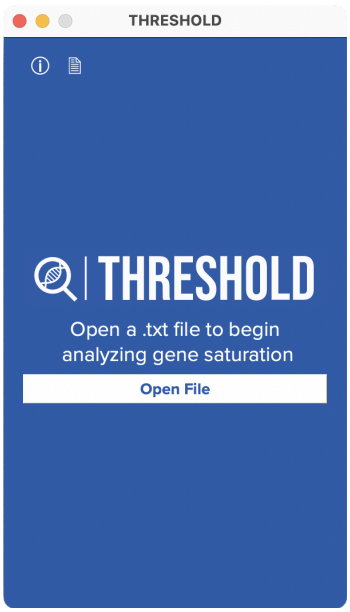
Hugo_Symbol	{Blank}	{Patient1ID}	{Patient2ID}...
Gene	...	z-score/percentile	z-score/percentile
...	...	...	...
...	...	...	...

- The Blank column is never actually used; it is a placeholder. As long as there's a

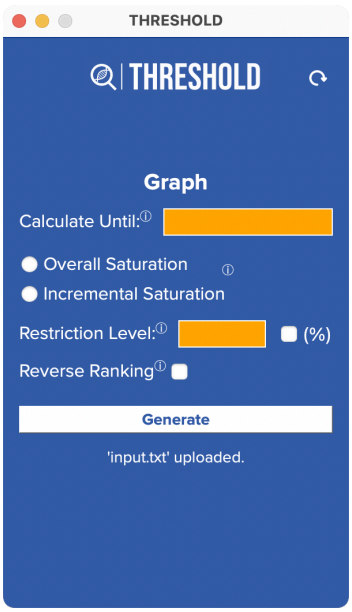
space between the Hugo_Symbol and {Patient1ID} column **THRESHOLD** will run.

- Note the first heading MUST be called Hugo_Symbol
- The columns must be separated by a tab '\t'

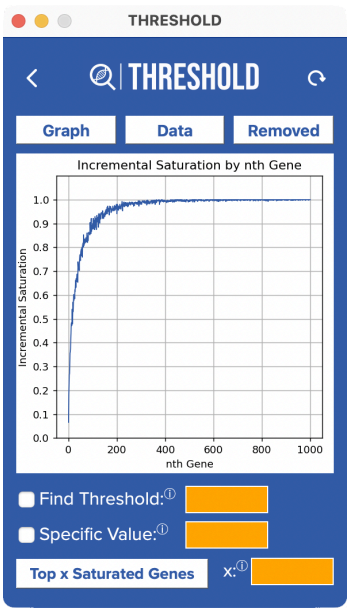
GUI Navigation:



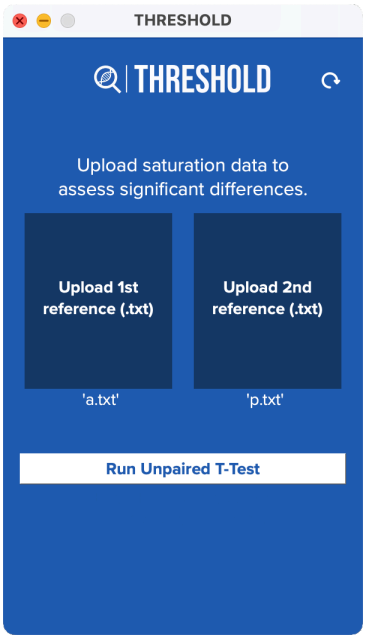
Home



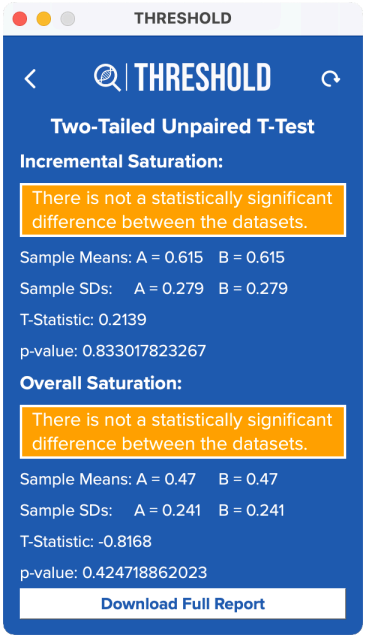
Calculate



Analyze



Run T-Test



Assess Significance

Home

- To begin the analysis, simply upload a .txt file in the appropriate file format by clicking on the Open File button. Wait about 5 seconds while the pseudogenes are removed, and you will automatically move to the next page.
- To view info and documentation of the tool, press the appropriate icons in the GUI toolbar

Calculate

- Confirm your file is uploaded with the dialog below the Generate button.
- Input the parameters according to your desired analyses. Clicking on the 'ⓘ' icons provides detailed explanations of the parameter's domains and purpose.
- Once satisfied with your inputs, press Generate. If the inputted file is formatted correctly, and the inputs accurate, an In progress... text dialog will appear. For a dataset of about 400 patients, you will need to wait about 2 minutes for the calculations to be made.
- If you would like to restart your analyses with a different file, press the restart button in the top right corner

Analyze

- A saturation curve will appear indicating saturation type by nth ranked gene.
- Hover over the graph to find specific saturation values. Alternatively, use the "Find Threshold" input to find when a certain saturation level is reached and the "Specific Value" input to evaluate the saturation of a specific nth gene. Clicking on the ⓘ icons provides detailed explanations of the parameters' domains and purpose.

- Additionally, you can find the most saturated genes with the Top x Saturated Genes button. Simply enter the number of top genes you want to find in “x:”
- To export a .png file of the Graph, a .txt file of the Data or a .txt file of removed pseudogenes, press the corresponding buttons above the graph.
- To begin a new analysis with the same file, press the back button. If you would like to restart your analysis entirely, press the restart button.

Run Unpaired T-Test

- Upload the two saturation files you would like to compare by clicking on the black fields.
- Ensure the files are in the proper, standard format (.txt) as exported by the **THRESHOLD** tool. The files should have three columns, “Nth Gene Included,” “Incremental Saturation,” and “Overall Saturation.” The file should be the same size; ie, the same number of rows.
- To run the test, simply press the Run Unpaired-T-Test button,
- To navigate back home, simply press restart in the top right corner.

Assess Significance

- This page simply outlines the result of your analyses, describing whether the differences between the saturation types in each data set yielded statistically significant differences. This result is outlined in the orange box. Additional relevant statistical measures, including the calculated p-value, are also listed.
- To download a more detailed description of the statistical analyses, press “Download Full Report” to export a (.txt) file of relevant calculated statistical measures for each of the saturation types.
- To begin a new statistical analysis, press the back button. If you would like to restart entirely and navigate back home, press the back button.