Introduction to Pathway Analysis for scRNA-seq

Childhood Cancer Data Lab

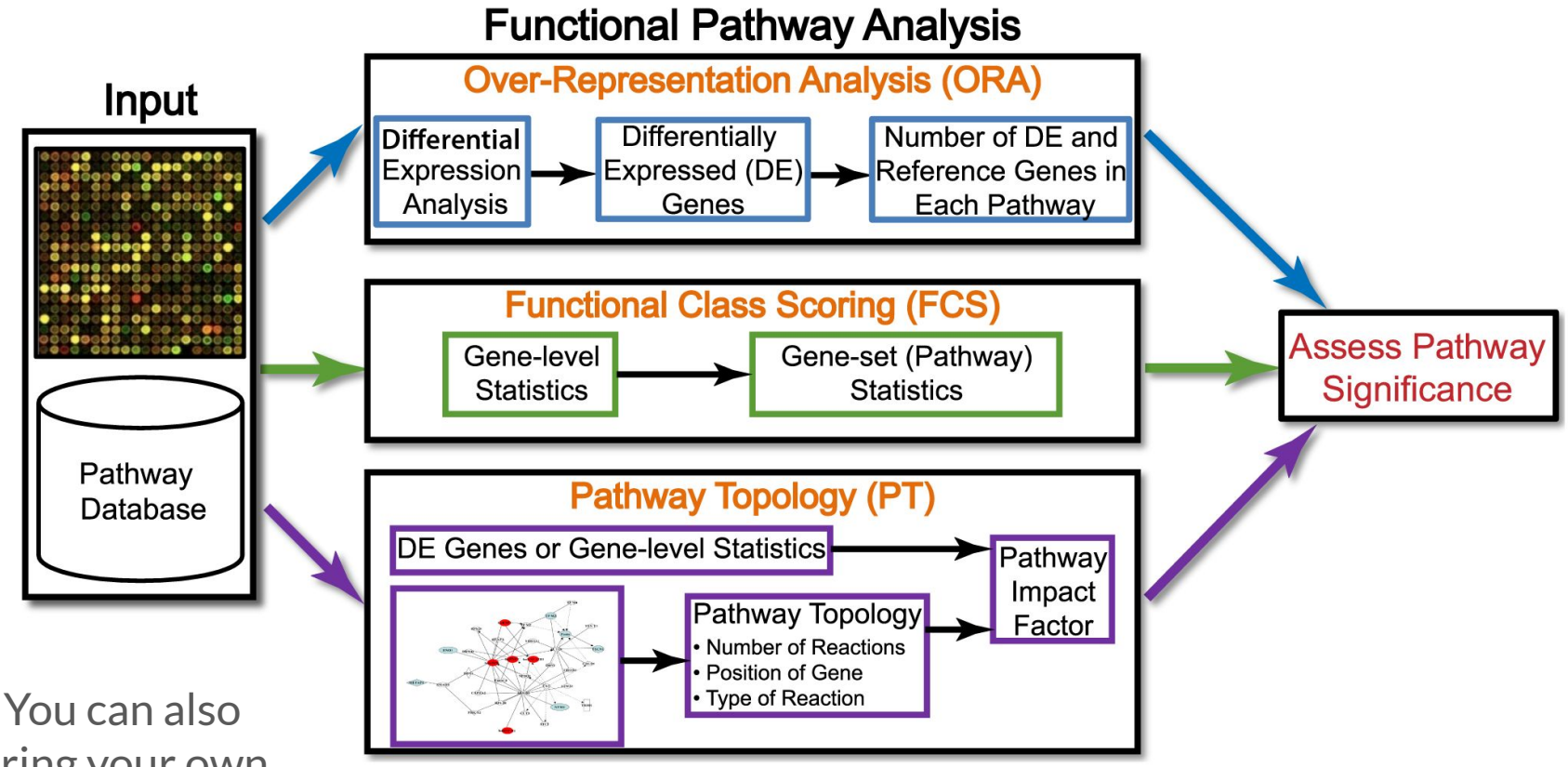
Objectives

- Explain the rationale for using pathway or gene set analysis
- Introduce principles for selecting a pathway analysis method
- Demonstrate how to run Gene Set Enrichment Analysis (GSEA) on pseudobulk DGE results
- Demonstrate how to run AUCell to estimate gene set activity in individual cells
- Briefly introduce conceptually similar (resource- and time-intensive) pathway enrichment methods that can be run on individual cells

Why pathway analysis?

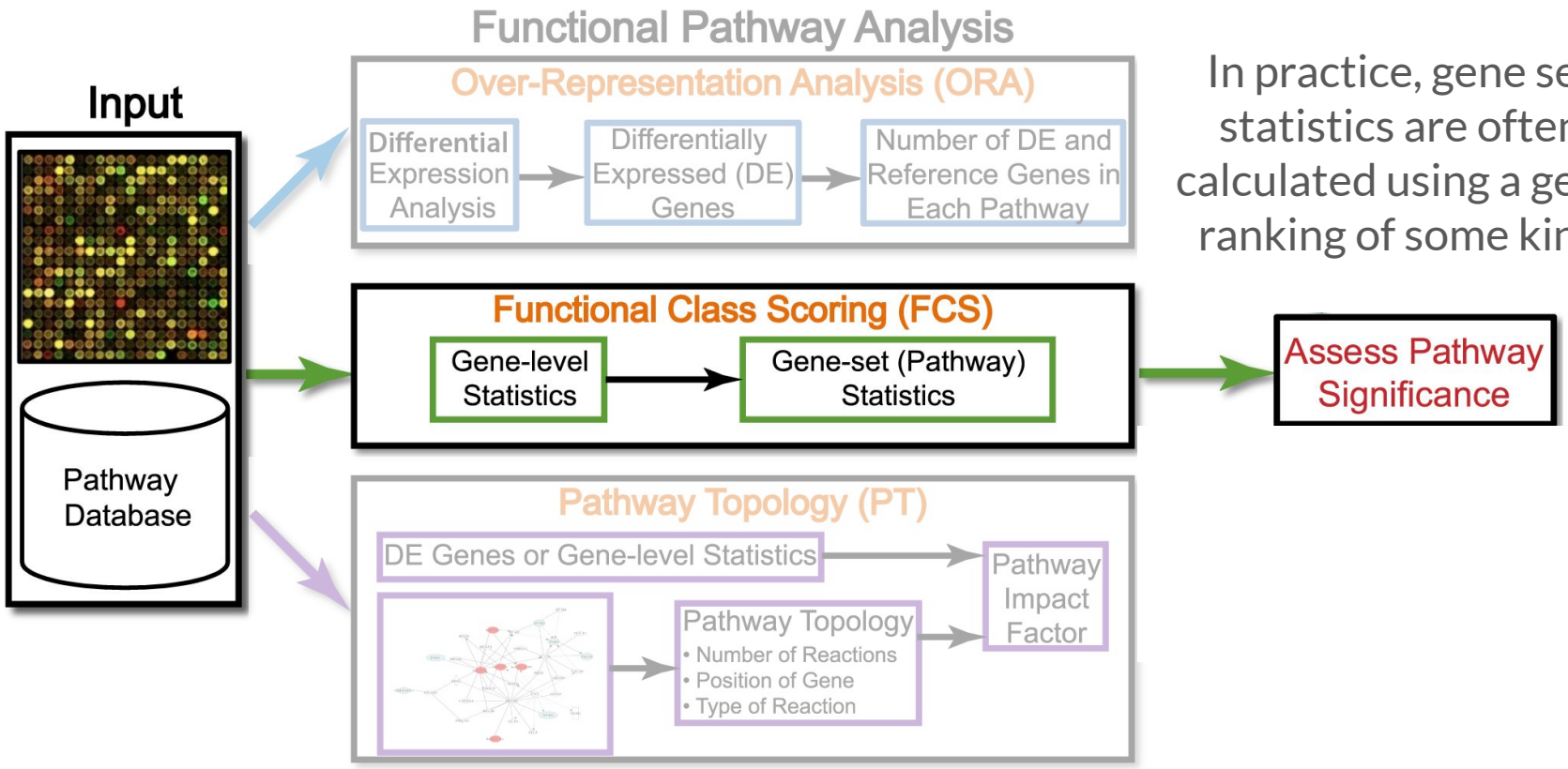
“...one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.”

- Subramanian et al. *PNAS*. 2005.



You can also bring your own gene sets!

Khatri, Sirota, and Butte. *PLoS Comp Bio*. 2012.



In practice, gene set statistics are often calculated using a gene ranking of some kind

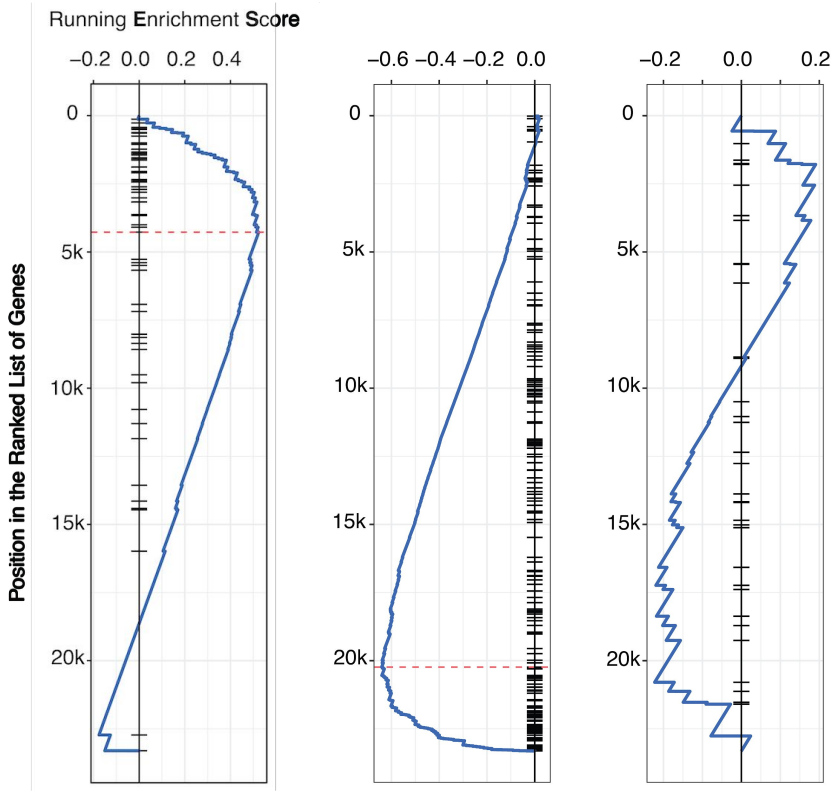
Khatri, Sirota, and Butte. *PLoS Comp Bio*. 2012.

(An incomplete list of) questions to ask when selecting a pathway analysis method

- What does it take as input (e.g., raw counts, TPM, gene-level statistics you've precomputed)?
- Does it require a group comparison (e.g., ARMS vs. ERMS)?
- How does it account for biases or characteristics present in single-cell sequencing data (e.g., library size, sparsity)?
- Are scores output from the method:
    - Comparable between different gene sets, i.e., because they have different number of genes in them?
    - Comparable between different cells or datasets?
    - Normally distributed for easier statistical analysis?
- **What does the output mean and will it address my biological question?**

Gene Set Enrichment Analysis (GSEA)

- Takes a pre-computed list of gene statistics, typically from a two-group comparison
- Genes are ranked from most positive to most negative based on the statistic and a running sum is computed
- Statistical significance is assessed through permutation testing
- Enrichment scores must be normalized to make them comparable between gene sets



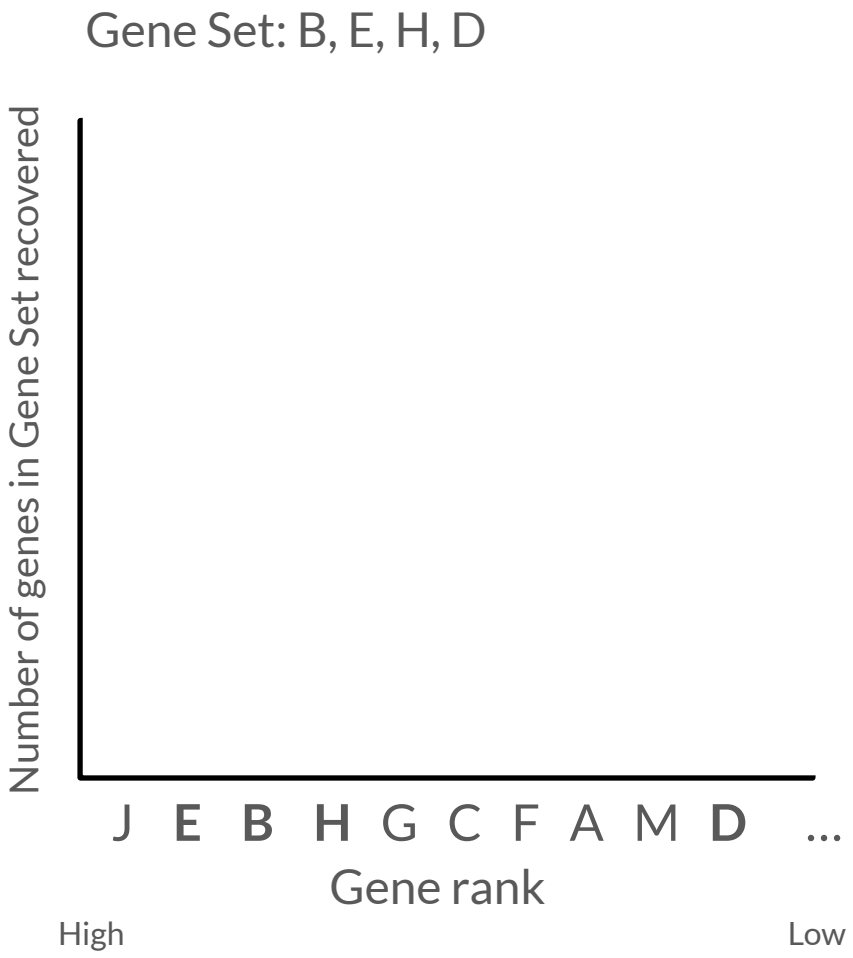
Subramanian et al. *PNAS*. 2005.

AUCell

- Genes are ranked within an individual cell from highest to lowest expression value
- The area under the recovery curve (AUC) is calculated, which represents a gene set's relative expression in an individual cell
- Ideally, the distribution of all AUC values is bimodal, allowing you to distinguish between cells where the gene set is "on" or "off"

Aibar et al. *Nature Methods*. 2017.

Diagram adapted from Malhotra. 2018.

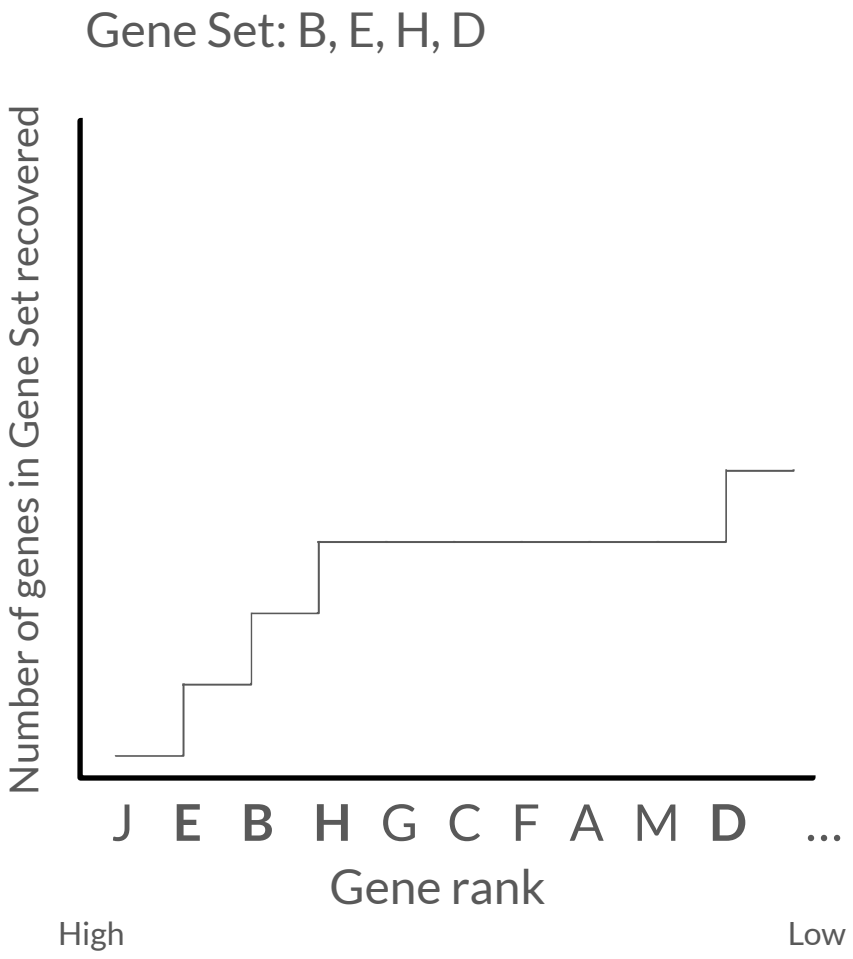


AUCell

- Genes are ranked within an individual cell from highest to lowest expression value
- The area under the recovery curve (AUC) is calculated, which represents a gene set's relative expression in an individual cell
- Ideally, the distribution of all AUC values is bimodal, allowing you to distinguish between cells where the gene set is "on" or "off"

Aibar et al. *Nature Methods*. 2017.

Diagram adapted from Malhotra. 2018.

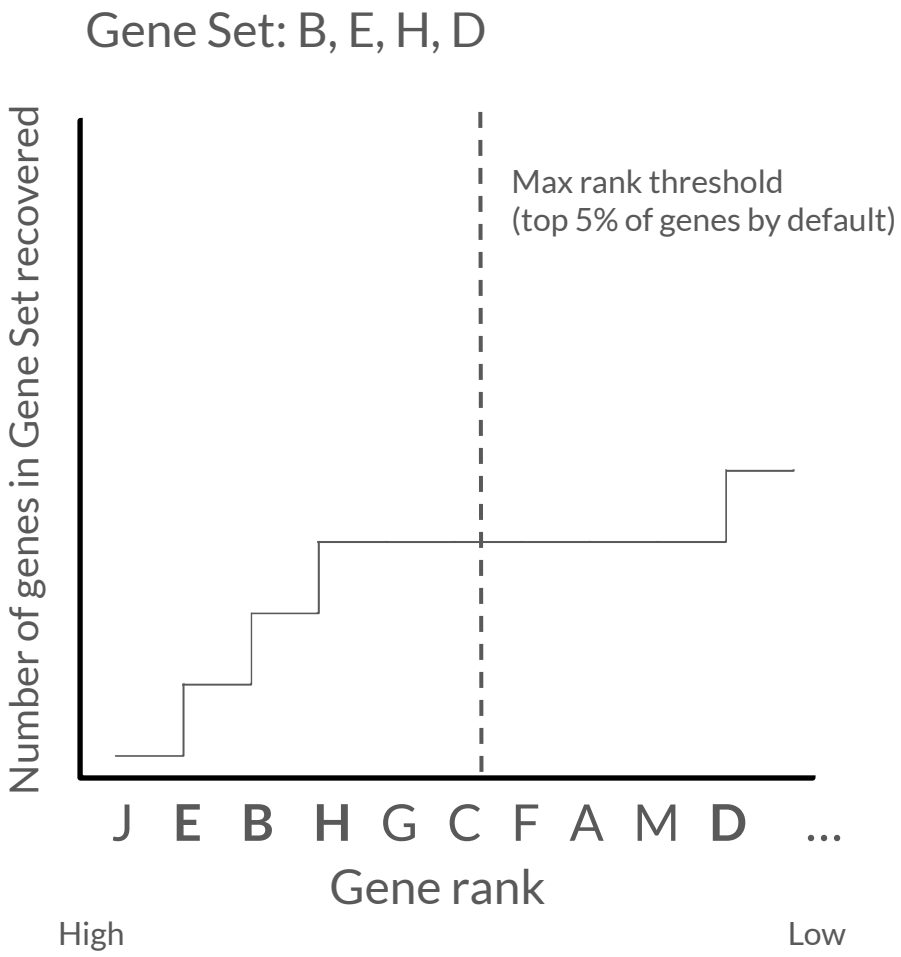


AUCell

- Genes are ranked within an individual cell from highest to lowest expression value
- The area under the recovery curve (AUC) is calculated, which represents a gene set's relative expression in an individual cell
- Ideally, the distribution of all AUC values is bimodal, allowing you to distinguish between cells where the gene set is "on" or "off"

Aibar et al. *Nature Methods*. 2017.

Diagram adapted from Malhotra. 2018.

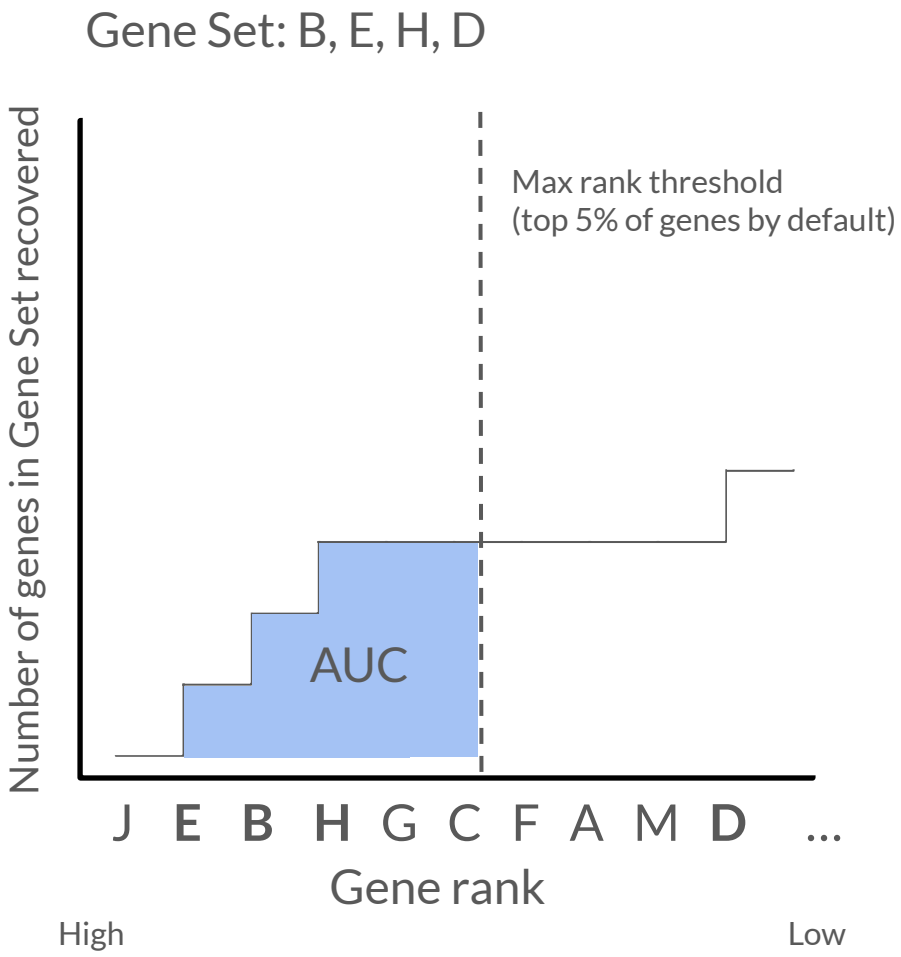


AUCell

- Genes are ranked within an individual cell from highest to lowest expression value
- The area under the recovery curve (AUC) is calculated, which represents a gene set's relative expression in an individual cell
- Ideally, the distribution of all AUC values is bimodal, allowing you to distinguish between cells where the gene set is "on" or "off"

Aibar et al. *Nature Methods*. 2017.

Diagram adapted from Malhotra. 2018.



AUCell

- Genes are ranked within an individual cell from highest to lowest expression value
- The area under the recovery curve (AUC) is calculated, which represents a gene set's relative expression in an individual cell
- Ideally, the distribution of all AUC values is bimodal, allowing you to distinguish between cells where the gene set is "on" or "off"

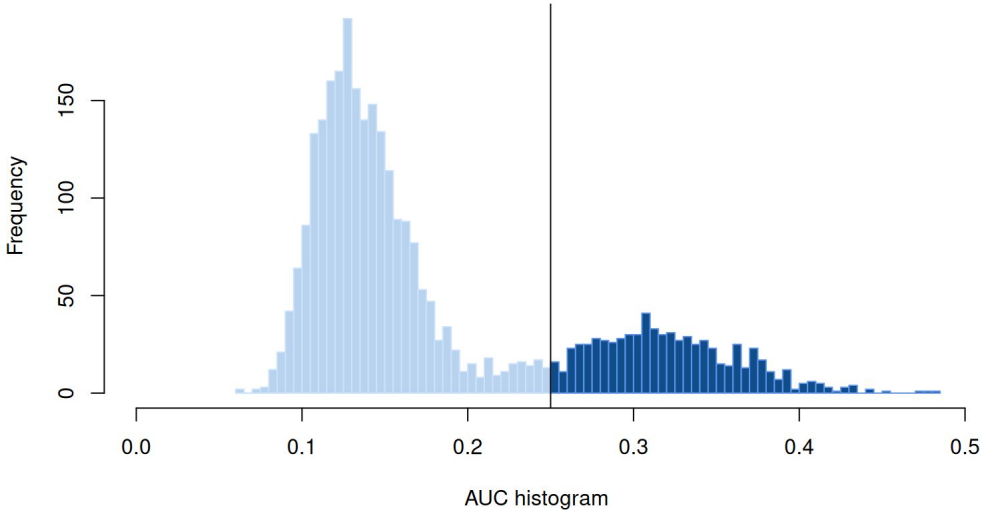
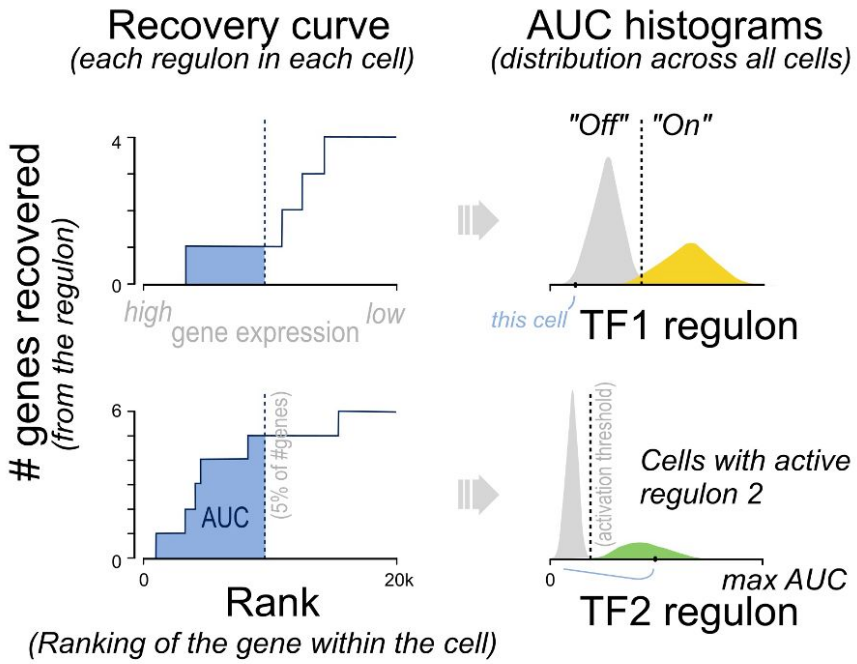


Image adapted from AUCell Vignette

Aibar et al. *Nature Methods*. 2017.

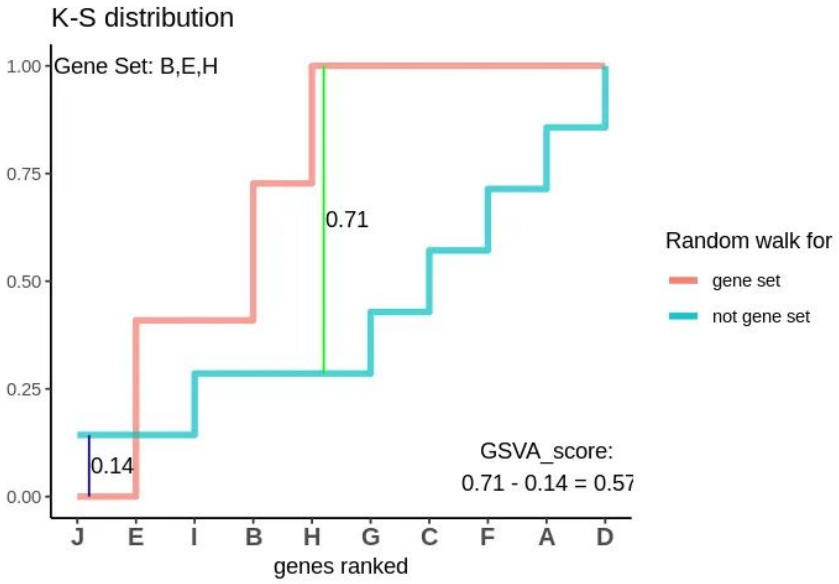
AUCell

- Genes are ranked within an individual cell from highest to lowest expression value
- The area under the recovery curve (AUC) is calculated, which represents a gene set's relative expression in an individual cell
- Ideally, the distribution of all AUC values is bimodal, allowing you to distinguish between cells where the gene set is "on" or "off"



Aibar et al. *Nature Methods*. 2017.

You can calculate scores for individual cells using single-sample GSEA (**ssGSEA**) and Gene Set Variation Analysis (**GSVA**)



Individual methods differ in:

- How the rankings are generated
- How exactly the score is calculated

Scores generally depend on gene set size

Images adapted from Malhotra. 2018.

Barbie et al. *Nature*. 2009.
Hänzelmann, Castelo, and Guinney. *BMC Bioinformatics*. 2013.

Conceptually related methods for individual cells

Single-sample GSEA (ssGSEA)

- Rank genes in an individual sample (cell) based on expression value
- Scores represent if gene set member genes are coordinately up or down (the difference in genes in the gene set vs. genes outside the gene set)
- Scores must be normalized for comparison between gene sets

Gene Set Variation Analysis (GSVA)

- Rank genes in an individual sample (cell) based on a model of a gene's expression within the population (collection of cells)
- Scores represent if gene set member genes are over- or under-expressed in a cell relative to the overall population
- Scores can be normally distributed, making downstream statistics easier, but they do depend on gene set size in practice

Barbie et al. *Nature*. 2009.
https://gsea-msigdb.github.io/ssGSEA-gpmodule/v10/

Hänzelmann, Castelo, and Guinney. *BMC Bioinformatics*. 2013.
Malhotra. 2018.

ssGSEA and GSVA, along with other methods, are implemented in the escape R package

Image Source: GitHub

But beware… zeros!

If a method ranks genes based on their expression in an individual cell, how does it deal with zeros and how will that impact the output scores?

Dropout can influence scores, particularly in methods that were originally designed for bulk data.

Noureen et al. *eLife*. 2022.

Resources

Guangchuang Yu. *clusterProfiler: universal enrichment tool for functional and comparative study*.

Harvard Chan Bioinformatics Core Training. *Intro to DGE: Functional analysis*.

Molecular Signatures Database (MSigDB)

AUCell Vignette

Borcherding and Andrews. escape Vignette.

Noureen et al. *Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data.* 2022.