Gene Set Analysis in Single Cell Gene Expression Data

Data Science Journal Club

Min Hu

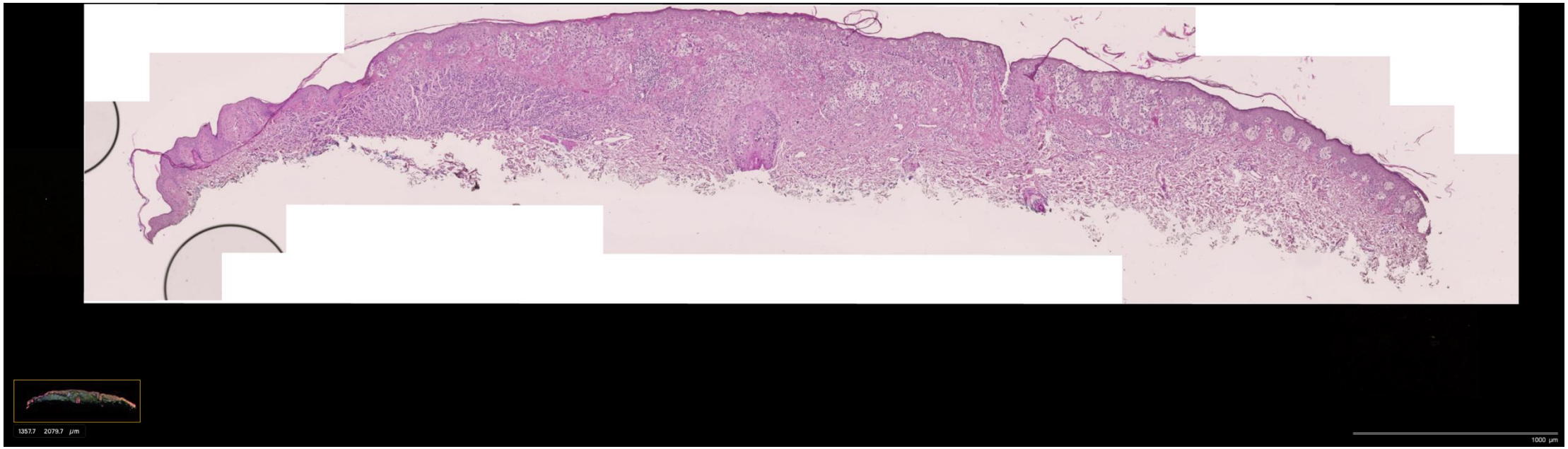
Judson-Torres and Tan labs

10/02/2025

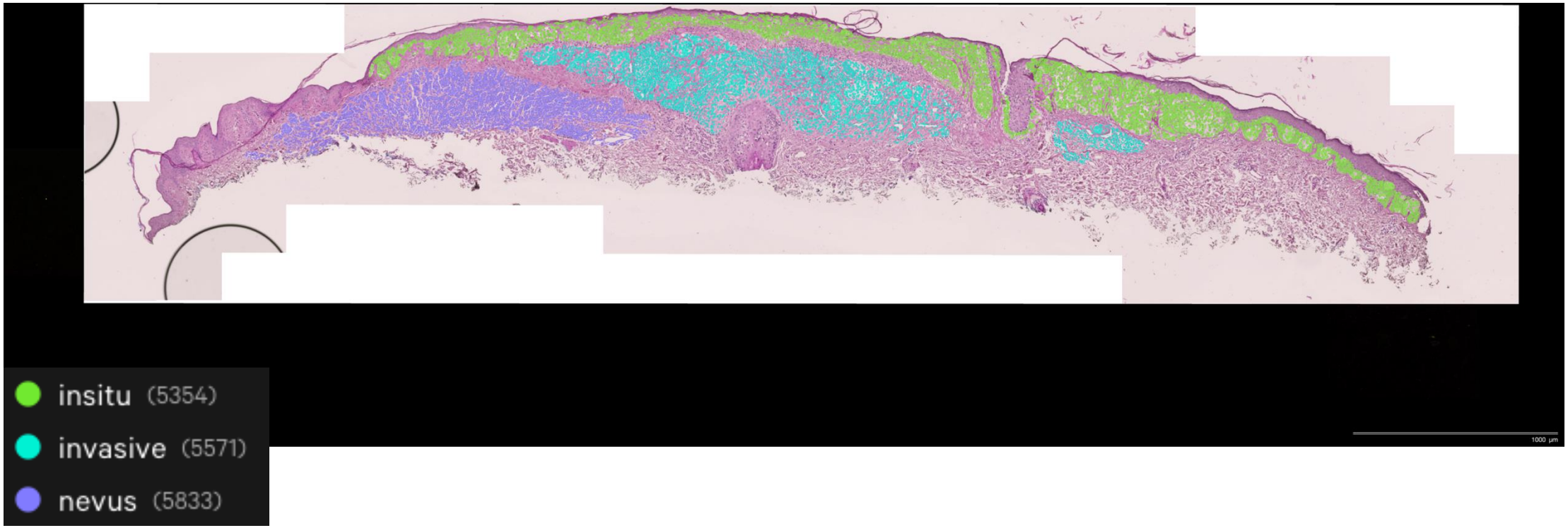
Outline

- Motivation and Dataset
- Single Cell Methods:
  - irGSEA
  - Module Score
  - AUCell
  - Ucell
- Results
- Bulk Method: GSVA

Motivation



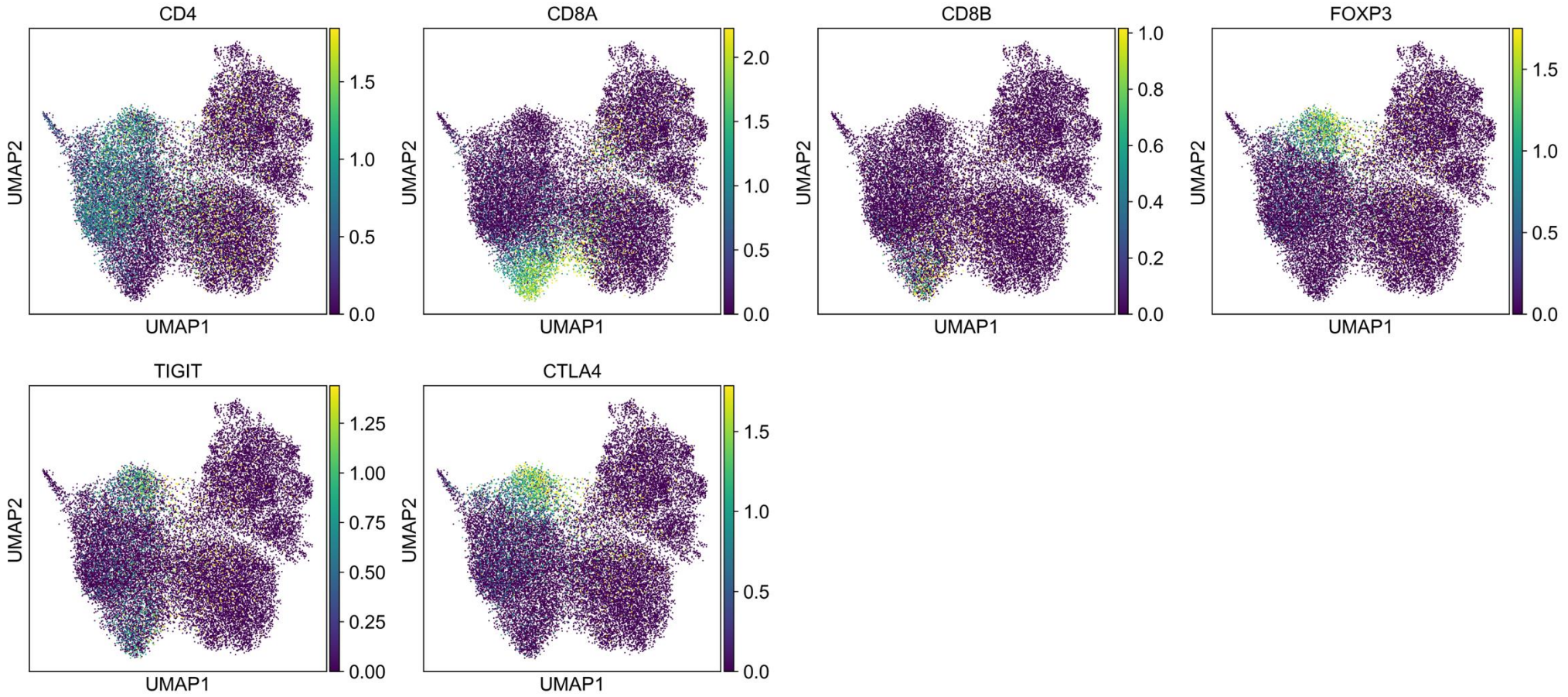
Motivation 1: Comparison between Conditions



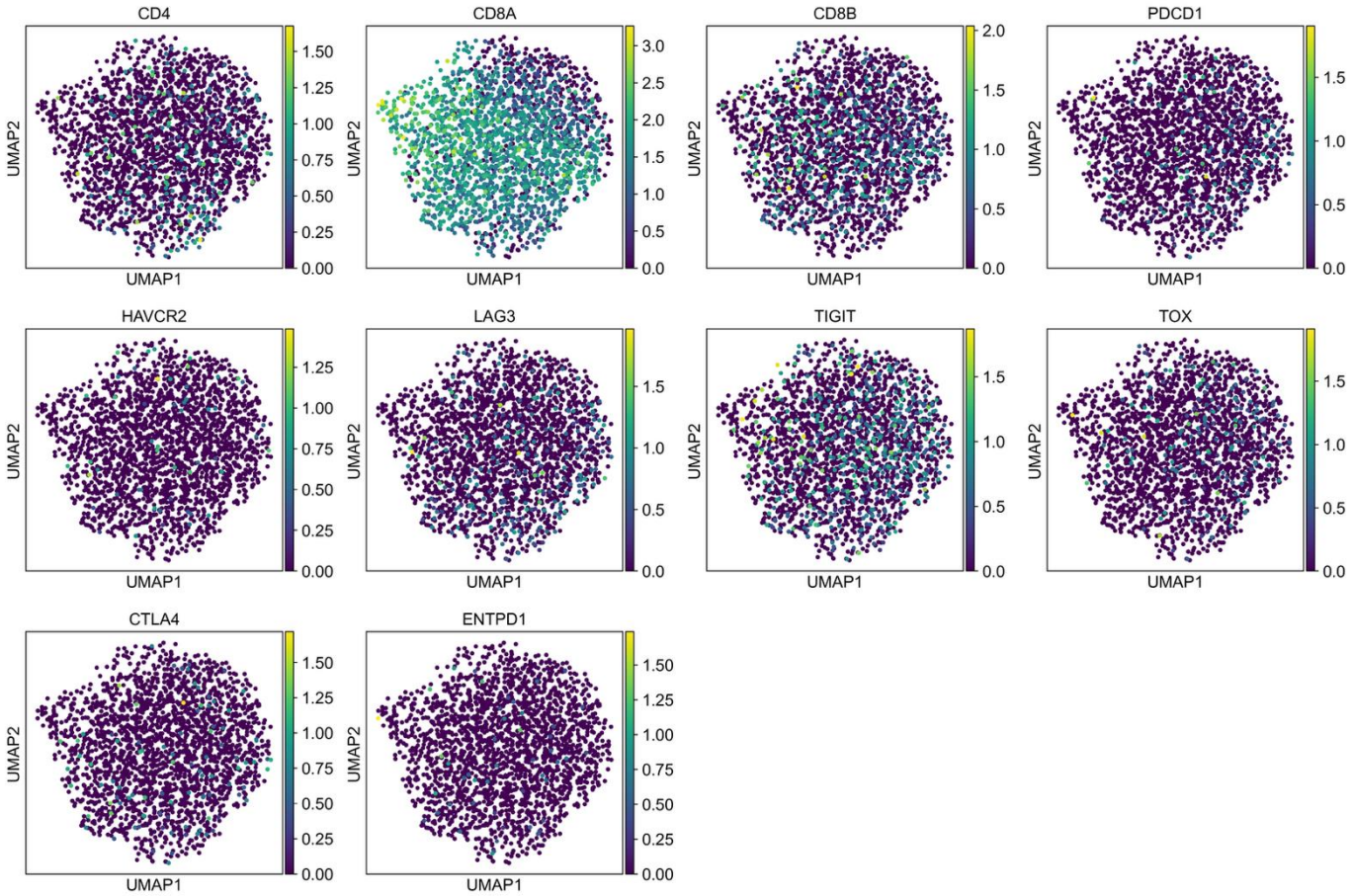
Motivation 1: Comparison between Conditions

- GSEA (https://www.gsea-msigdb.org/gsea/msigdb/human/annotate.jsp)
- EnrichR (https://maayanlab.cloud/Enrichr/)

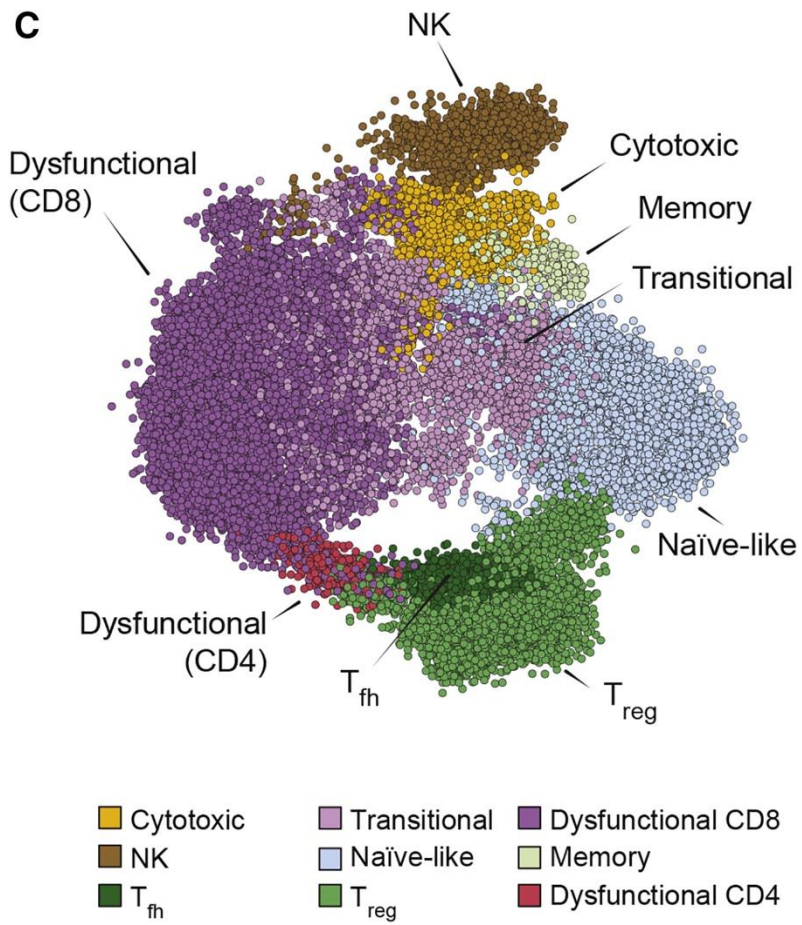
Motivation 2: Sample/Cell Annotation/Featurization



Motivation 2: Sample/Cell Annotation/Featurization

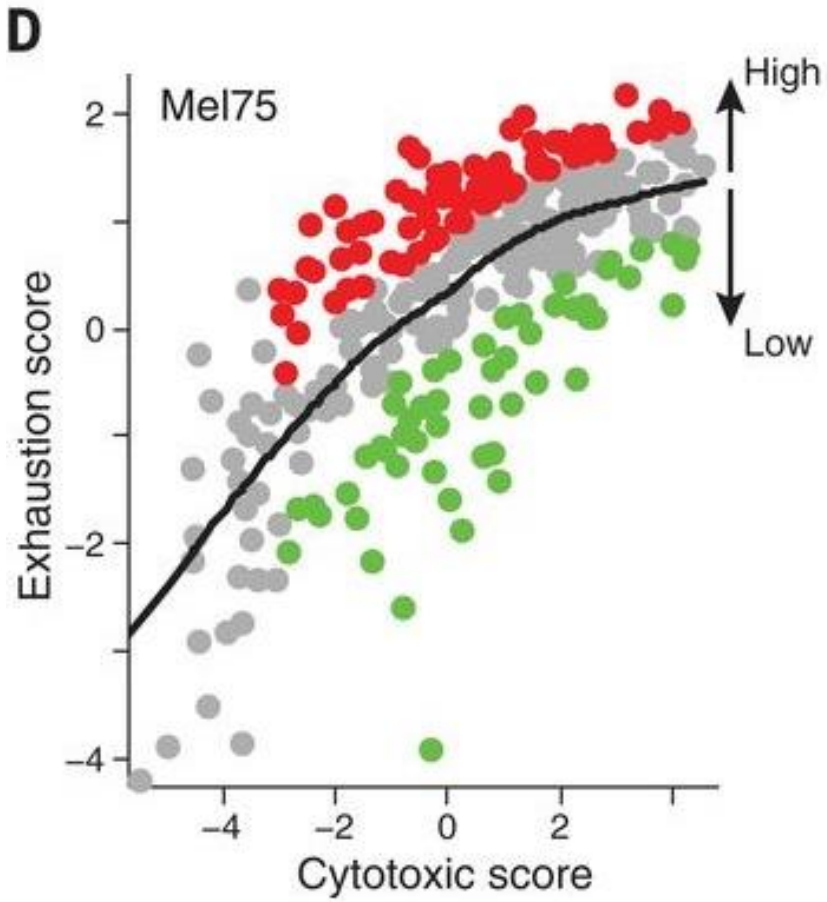


Motivation 2: Sample/Cell Annotation/Featurization



Li, et al. Cell. 2019

Motivation 2: Sample/Cell Annotation/Featurization



Tirosh et al. Science. 2016

JOURNAL ARTICLE

irGSEA: the integration of single-cell rank-based gene set enrichment analysis

Chuiqin Fan, Fuyi Chen, Yuanguo Chen, Liangping Huang, Manna Wang, Yulin Liu, Yu Wang, Huijie Guo, Nanpeng Zheng, Yanbing Liu, Hongwu Wang ✉, Lian Ma ✉

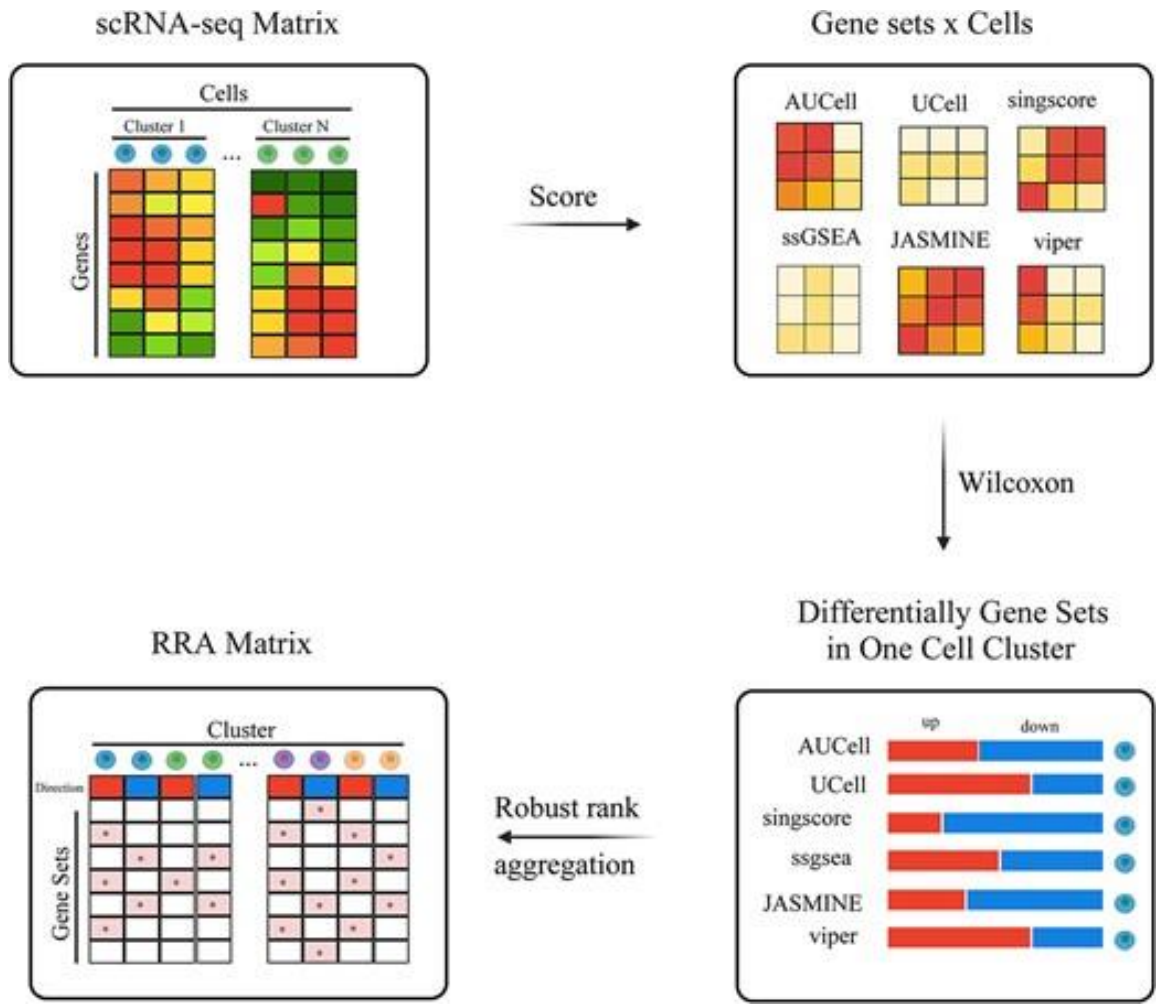
Author Notes

Briefings in Bioinformatics, Volume 25, Issue 4, July 2024, bbae243,

https://doi.org/10.1093/bib/bbae243

Published: 27 May 2024 **Article history** ▾

irGSEA Aggregates 6 Methods



irGSEA Aggregates 6 Methods

$$r_{ij} = R_{ij} / Max_{j=1}^{k} \left(R_{ij} \right)$$

i is the significant target gene set, j is the scoring method and k is the number of scoring methods used. r_{ij} is the standardized rank of the significant gene set i in the scoring method j. R_{ij} is the original rank of the significant gene set i in the scoring method j. $\mathrm{Max}_{j=1}^{k} \left(R_{ij} \right)$ is the largest original rank among k types of scoring methods.

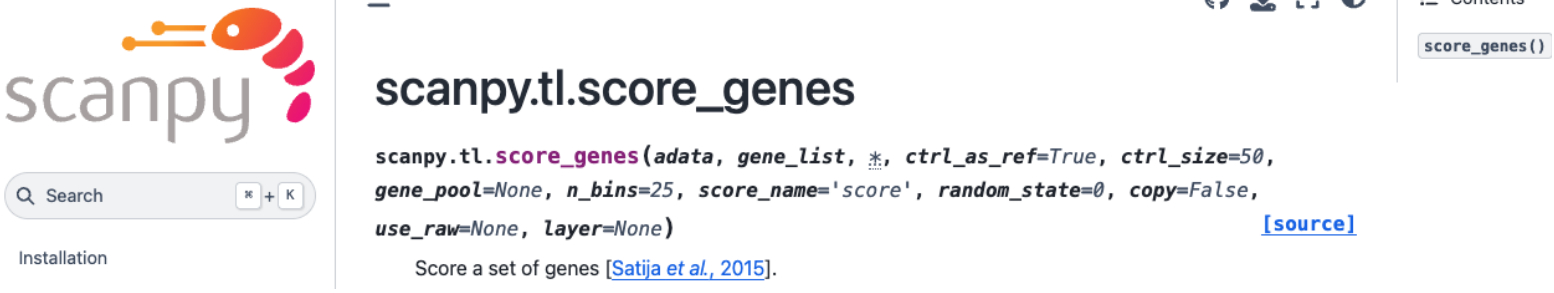
irGSEA Aggregates 6 Methods

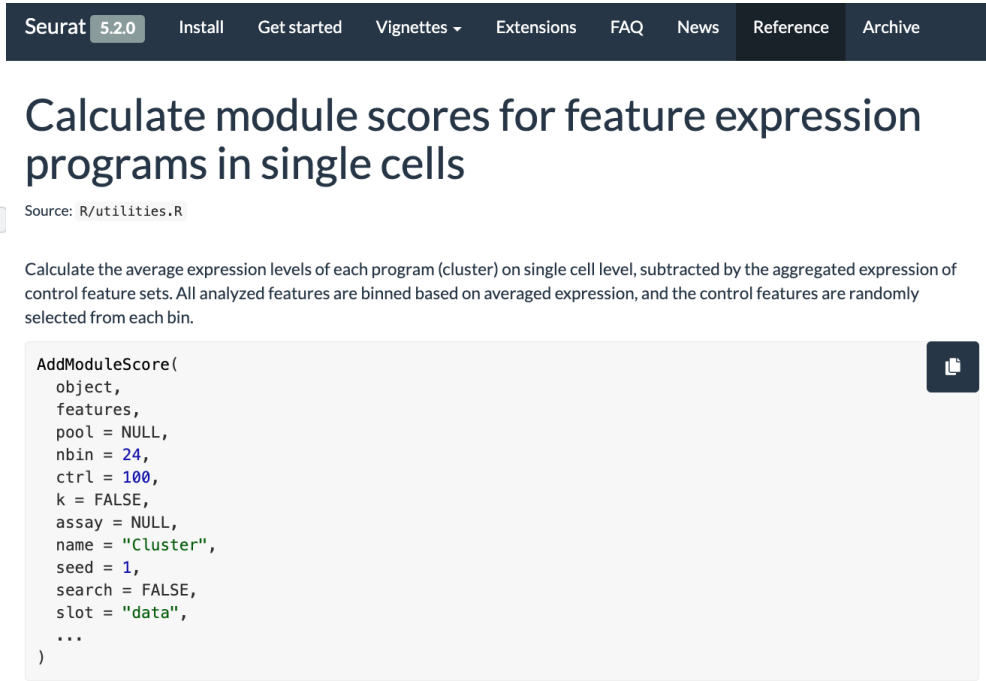
$$\beta_{x,k}\left(r_{(x)}\right) := \sum_{\ell=x}^{\mathrm{k}} \binom{\mathrm{k}}{\ell} \left(\mathrm{r}_{(\mathrm{x})}\right)^{\ell} \left(1 - \left(r_{(x)}\right)\right)^{k-\ell},$$

$$f(k, n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distribution PMF

Module Score





Module Score

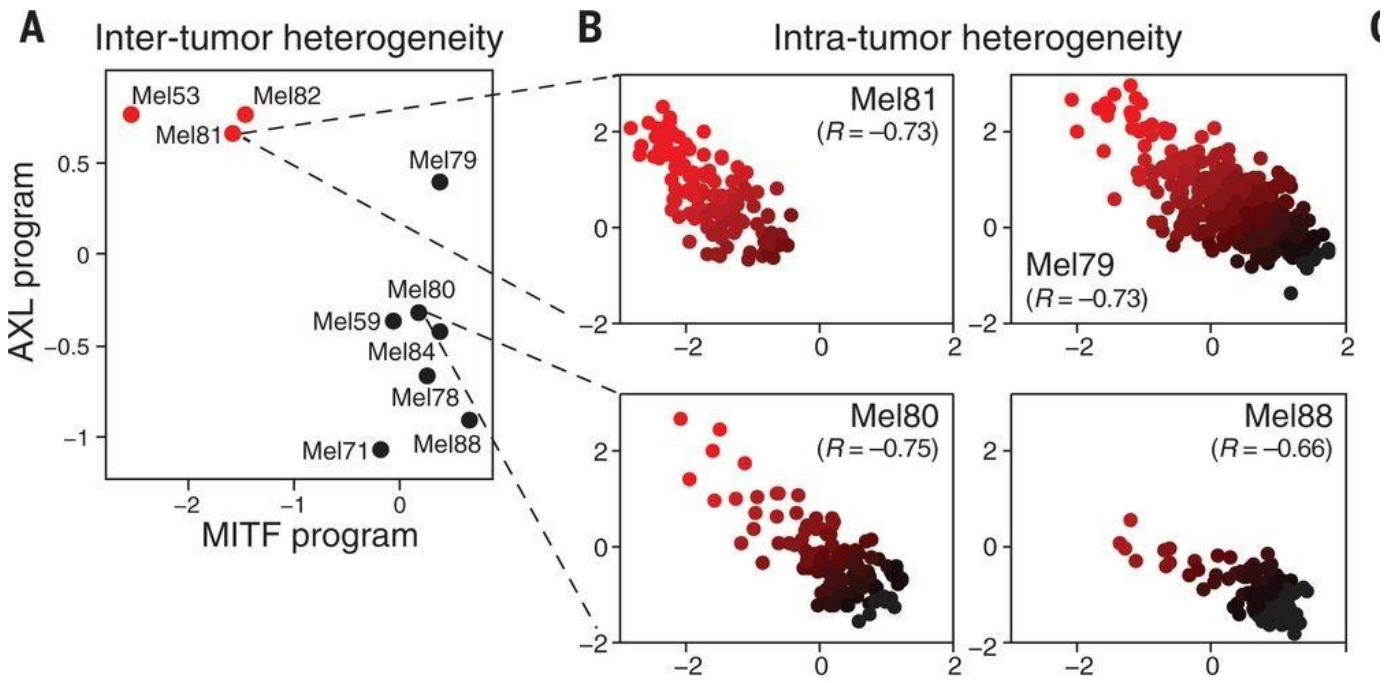
RESEARCH ARTICLE

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

ITAY TIROSH, BENJAMIN IZAR, SANJAY M. PRAKADAN, MARC H. WADSWORTH, II, DANIEL TREACY, JOHN J. TROMBETTA, ASAF ROTEM, CHRISTOPHER RODMAN, CHRISTINE LIAN, [...] , AND LEVI A. GARRAWAY +27 authors Authors Info & Affiliations

SCIENCE • 8 Apr 2016 • Vol 352, Issue 6282 • pp. 189-196 • DOI: 10.1126/science.aad0501

Module Score



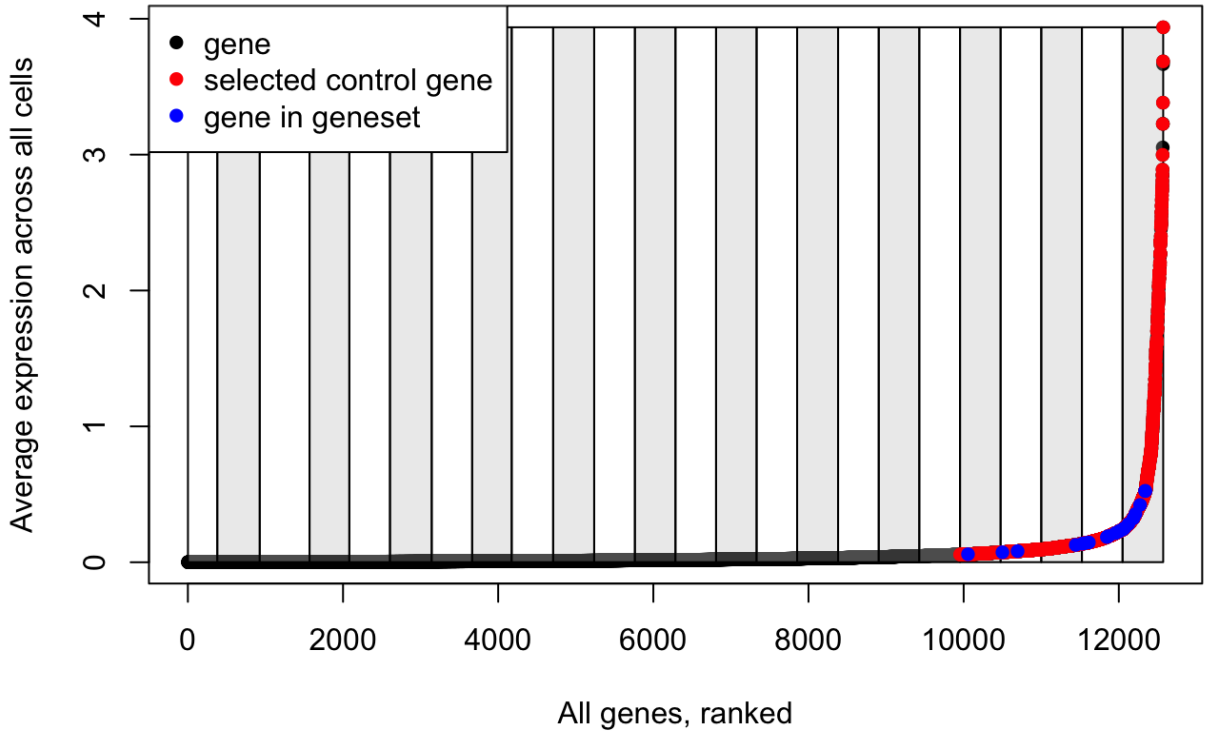
Module Score

MITF and AXL expression programs and cell scores

The top 100 MITF-correlated genes across the entire set of malignant cells were defined as the MITF program, and their average relative expression as the MITF-program cell score. The average expression of the top 100 genes that negatively correlate with the MITF program scores were defined as the AXL program and used to define AXL program cell score. To decrease the effect that the quality and complexity of each cell's data might have on its MITF/AXL scores we defined control gene-sets and their average relative expression as control scores, for both the MITF and AXL programs. These control cell scores were subtracted from the respective MITF/AXL cell scores. The control gene-sets were defined by first binning all analyzed genes into 25 bins of aggregate expression levels and then, for each gene in the MITF/AXL gene-set, randomly selecting 100 genes from the same expression bin as that gene. In this way, a control gene-sets have a comparable distribution of expression levels to that of the MITF/AXL gene-set and the control gene set is 100-fold larger, such that its average expression is analogous to averaging over 100 randomly-selected gene-sets of the same size as the MITF/AXL gene-set. To

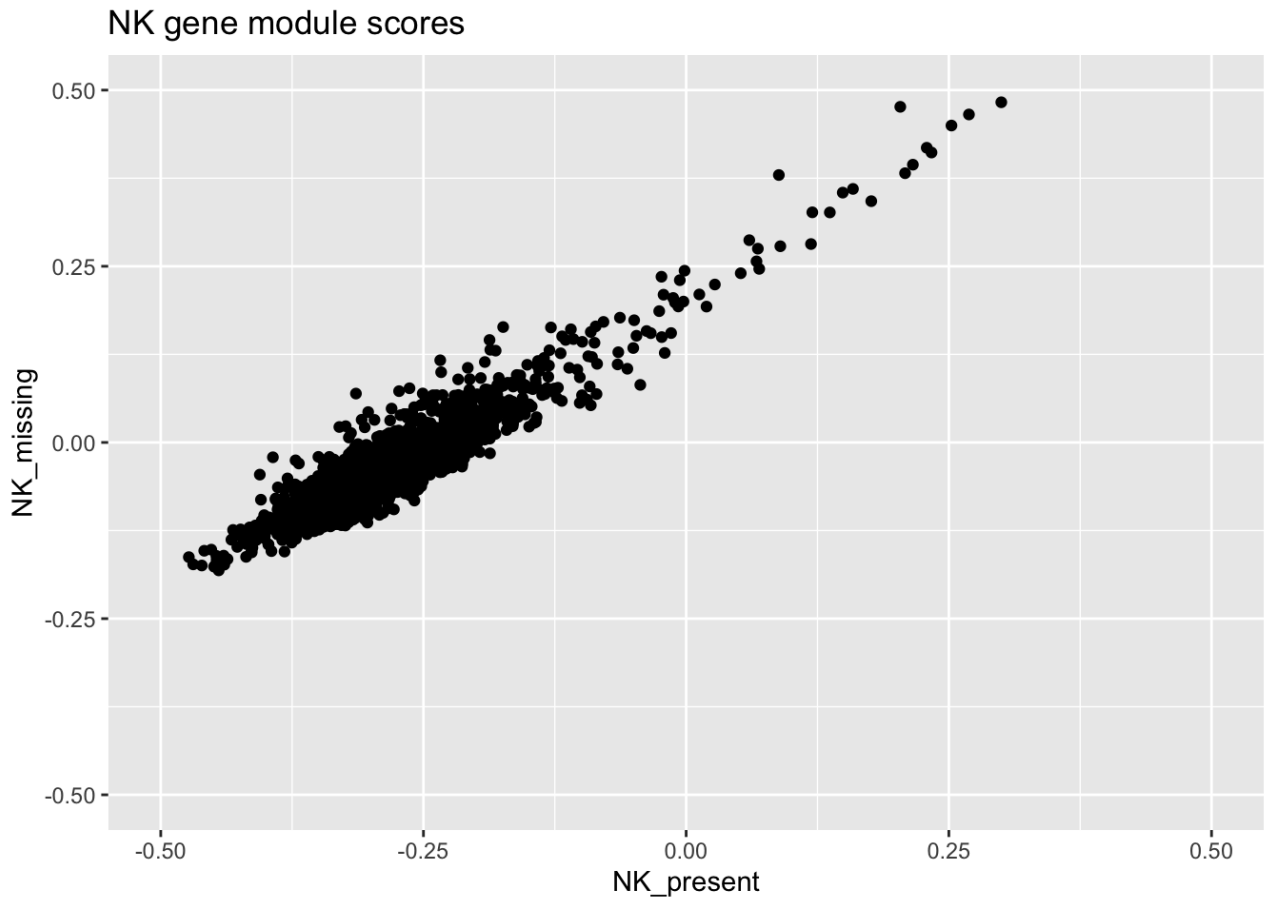
Module Score

1. Average expression of gene set
2. Binning gene expression
3. Average expression of control gene set
4. Gene set score – control gene set score



https://www.waltermuskovic.com/2021/04/15/seurat-s-addmodulescore-function/

Module Score Depends on Dataset Composition



https://www.waltermuskovic.com/2021/04/15/seurat-s-addmodulescore-function/

AUCell

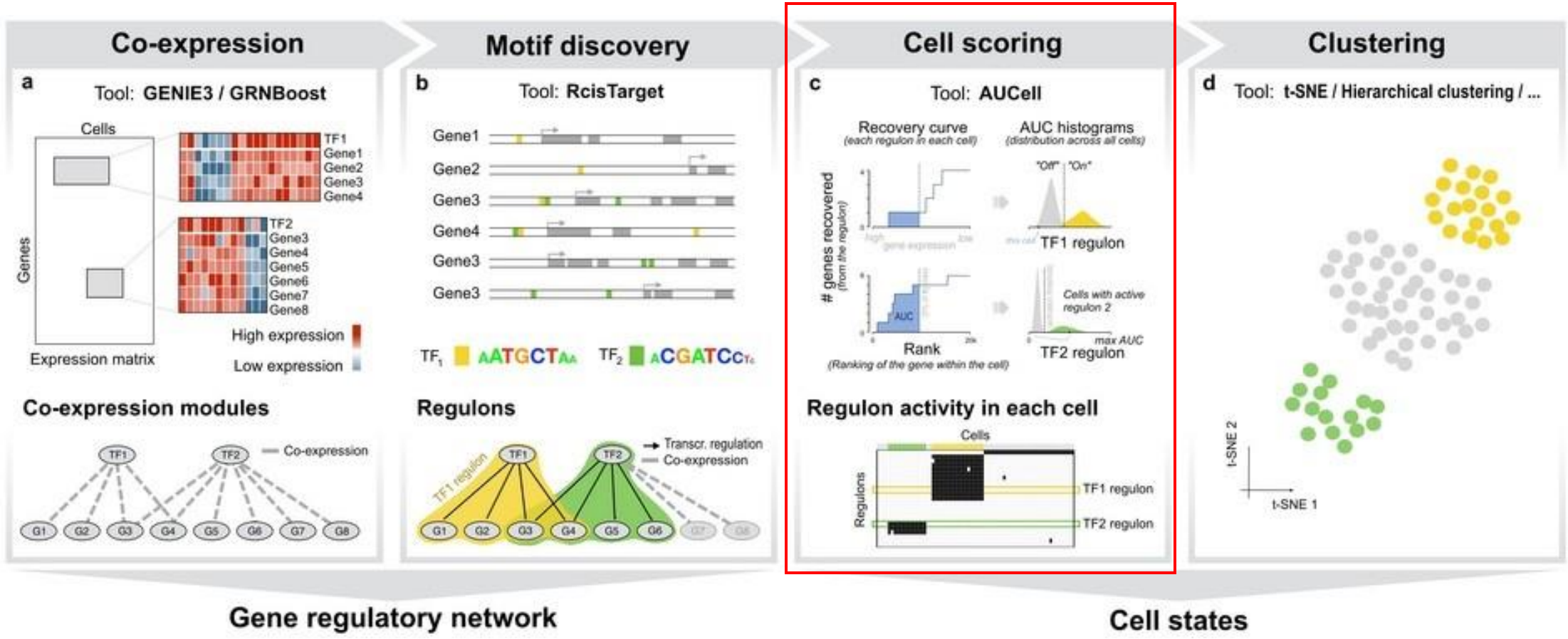
Brief Communication | Published: 09 October 2017

SCENIC: single-cell regulatory network inference and clustering

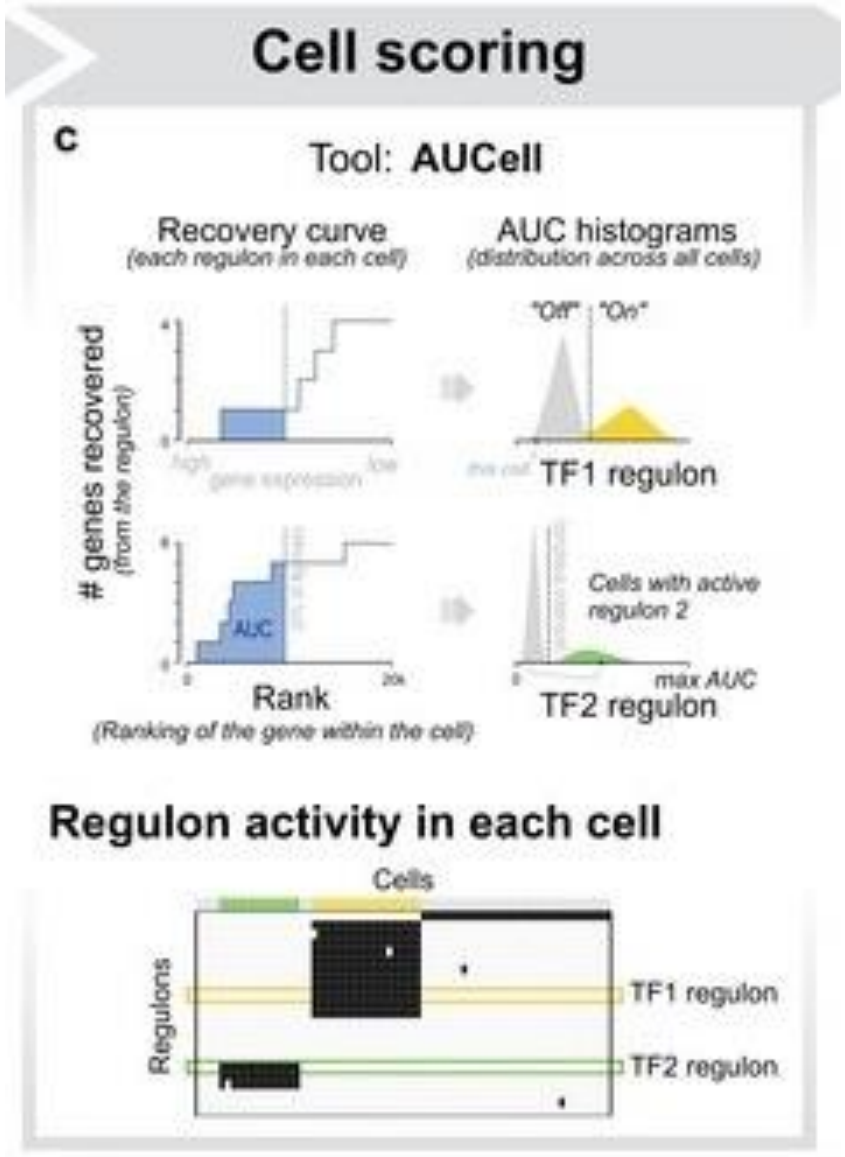
Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters & Stein Aerts

Nature Methods **14**, 1083–1086 (2017) | Cite this article

AUCell

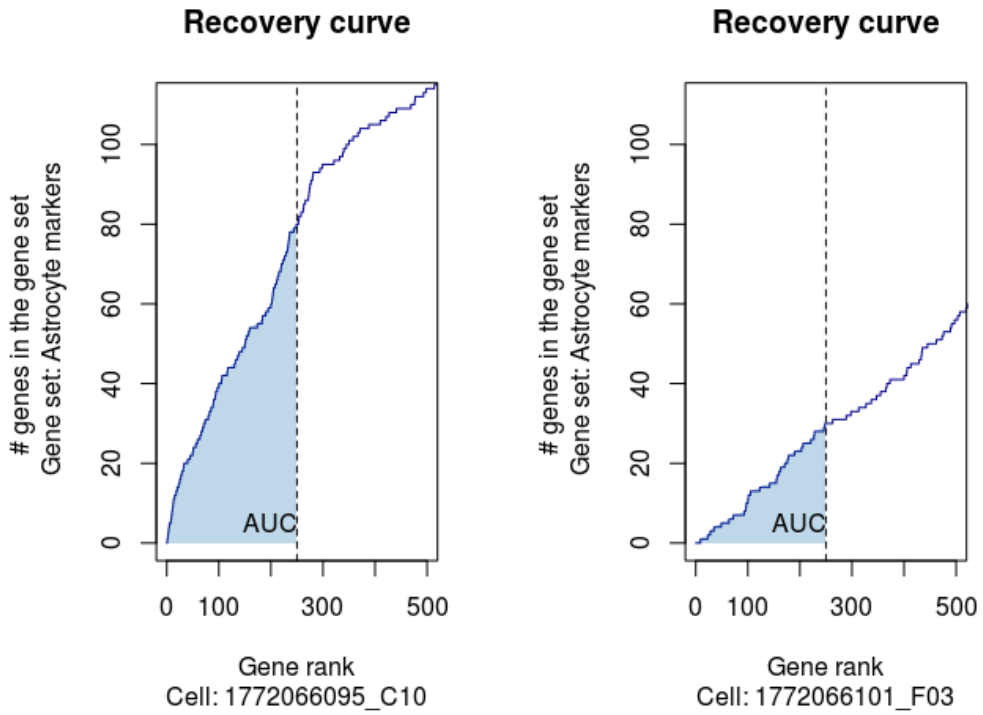

Co-expression
Motif discovery
Cell scoring
Clustering
a
Tool: GENIE3 / GRNBoost
Cells
Genes
Expression matrix
High expression
Low expression
b
Tool: RcisTarget
c
Tool: AUCell
Recovery curve
(each regulon in each cell)
AUC histograms
(distribution across all cells)
genes recovered
(from the regulon)
Rank
(Ranking of the gene within the cell)
TF1 regulon
TF2 regulon
Cells with active regulon 2
d
Tool: t-SNE / Hierarchical clustering / ...
t-SNE 2
t-SNE 1
Co-expression modules
Co-expression
Regulons
Transcr. regulation
Co-expression
Regulon activity in each cell
Cells
Regulons
Gene regulatory network
Cell states

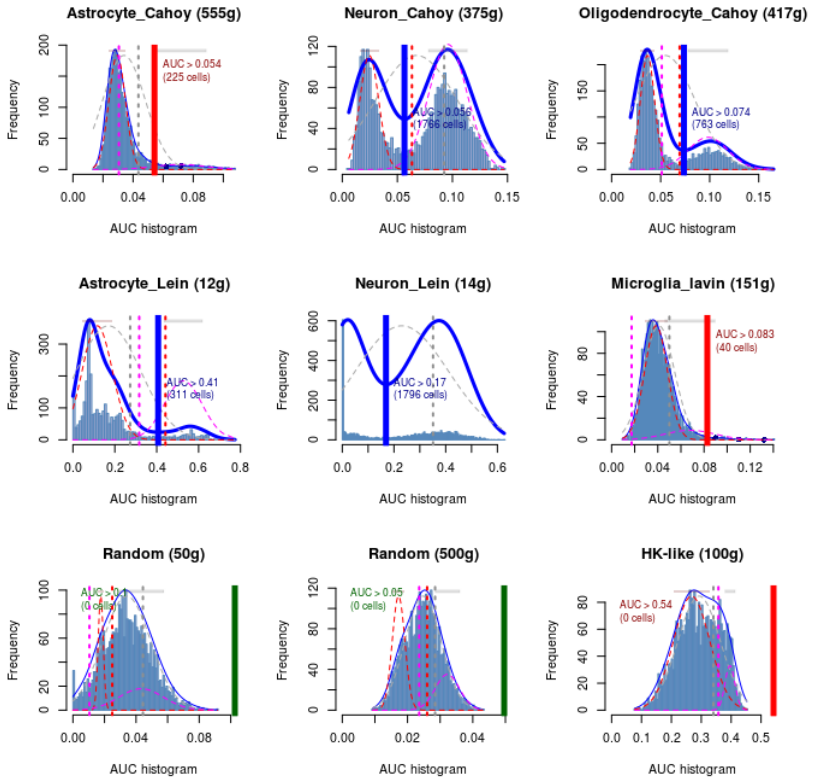
AUCell


Cell scoring
c
Tool: AUCell
Recovery curve
(each regulon in each cell)
AUC histograms
(distribution across all cells)
genes recovered
(from the regulon)
"Off"
"On"
TF1 regulon
AUC
Rank
(Ranking of the gene within the cell)
Cells with active regulon 2
max AUC
TF2 regulon
Regulon activity in each cell
Cells
Regulons
TF1 regulon
TF2 regulon

AUCell

- 1. Build gene-expression rankings for each cell
- 2. Calculate enrichment for the gene signatures (AUC)
- 3. Binarize (Optional)





https://scenic.aertslab.org/scenic_paper/tutorials/AUCell.html

UCell

Computational and Structural Biotechnology Journal

Volume 19, 2021, Pages 3796-3798

Software/Web server Article

UCell: Robust and scalable single-cell gene signature scoring

Massimo Andreatta [a b], Santiago J. Carmona [a b]

UCell

$$U'_j = 1 - \frac{U_j}{n \bullet r_{\max}}$$

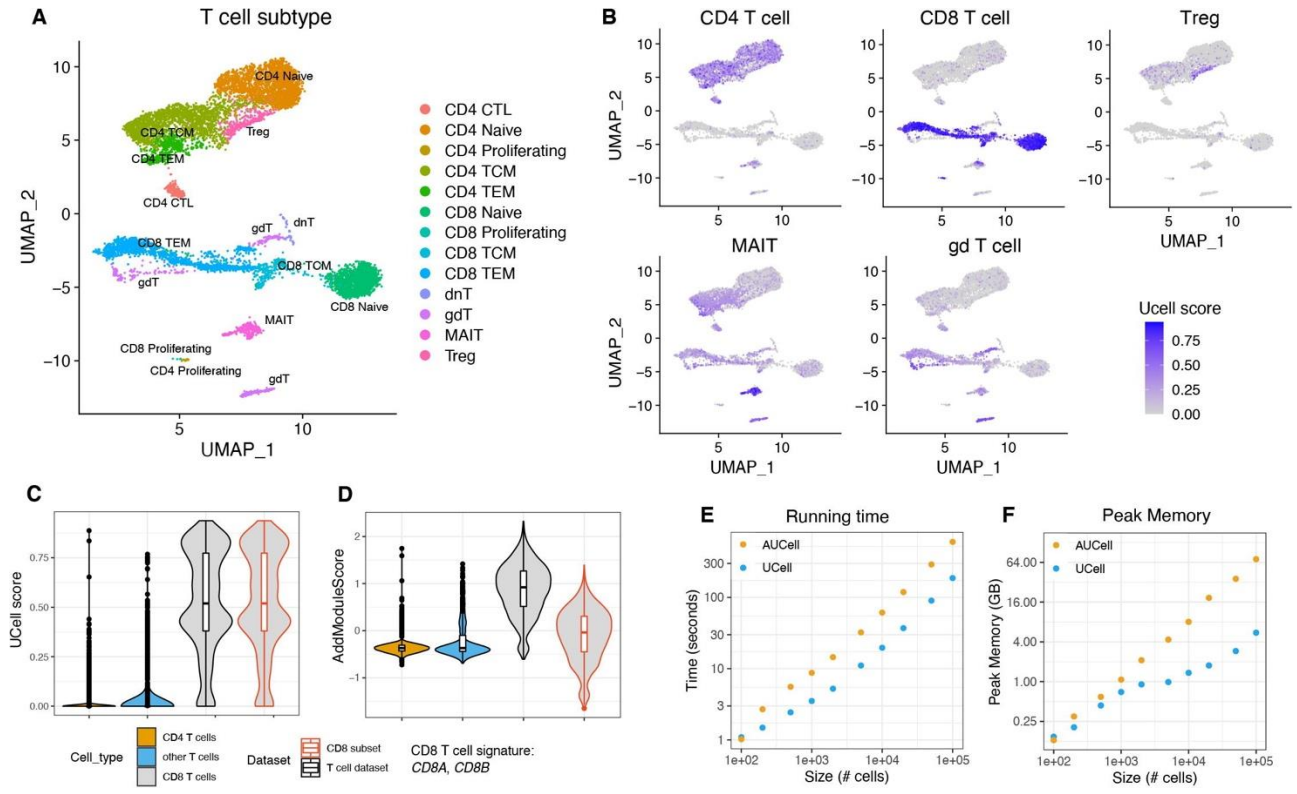
S_{max}: The largest possible sum of ranks

$$U_j = \sum_{i=1}^{n} r'_{i,j} - \frac{n(n+1)}{2}$$

S_{min} : The Smallest possible sum of ranks

U'$_j$ Ucell score for cell j
U$_j$ Mann-Whitney U statistics
n number of genes in the gene set
r'$_{i,j}$ rank of gene i in cell j

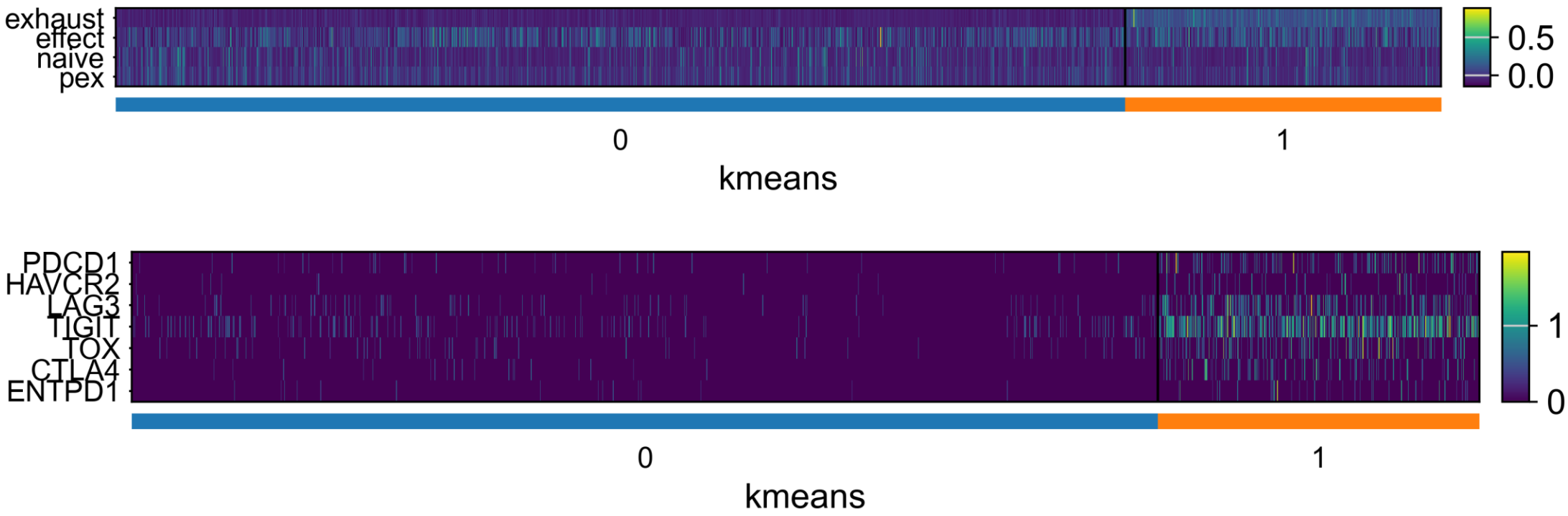
UCell Performance does not depend on dataset composition and is more efficient than AUCell



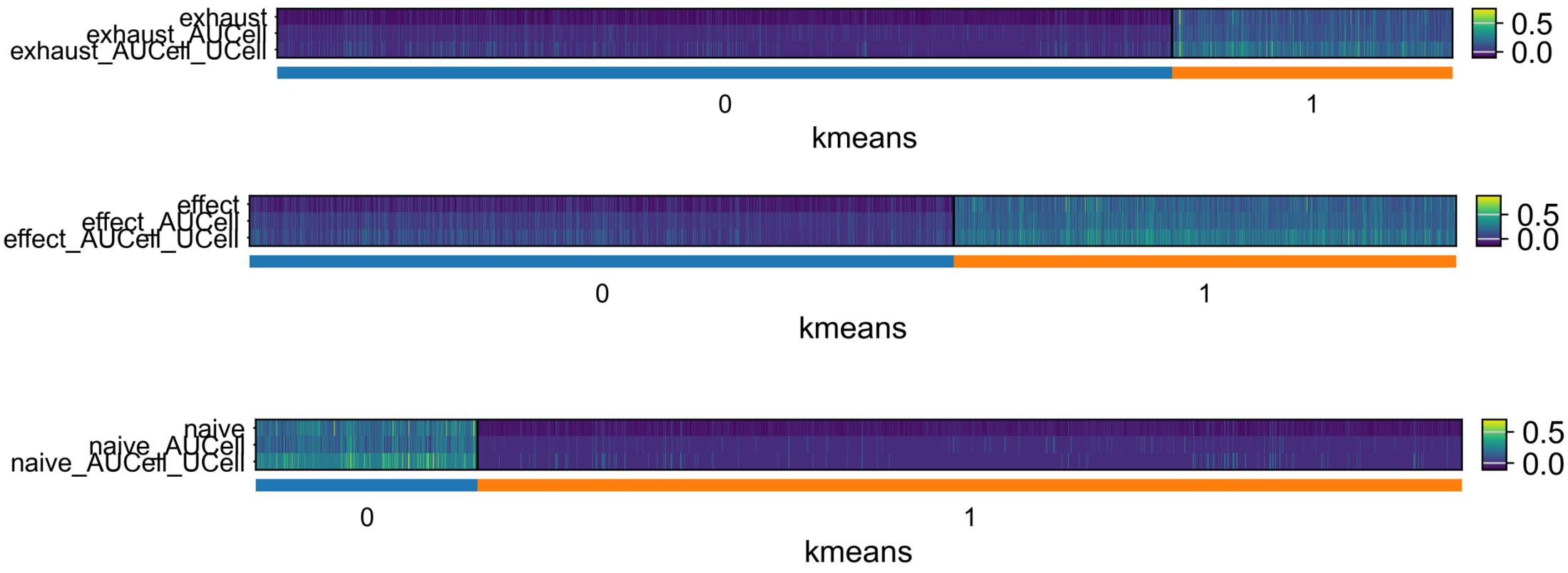
4 Gene Sets for CD8 T Cells

```
exhaust = ["PDCD1","HAVCR2","LAG3","TIGIT","TOX","CTLA4","ENTPD1"]
effect = ["GZMK","TNF","GZMA","GZMB","NKG7","PRF1","IFNG","CX3CR1"]
naive = ["TCF7","LEF1","SELL","CCR7","IL7R"]
pex = ["TCF7","LEF1","SELL","SLAMF6","BCL6","CCR7","IL7R","KLF2","KLF3"]
```

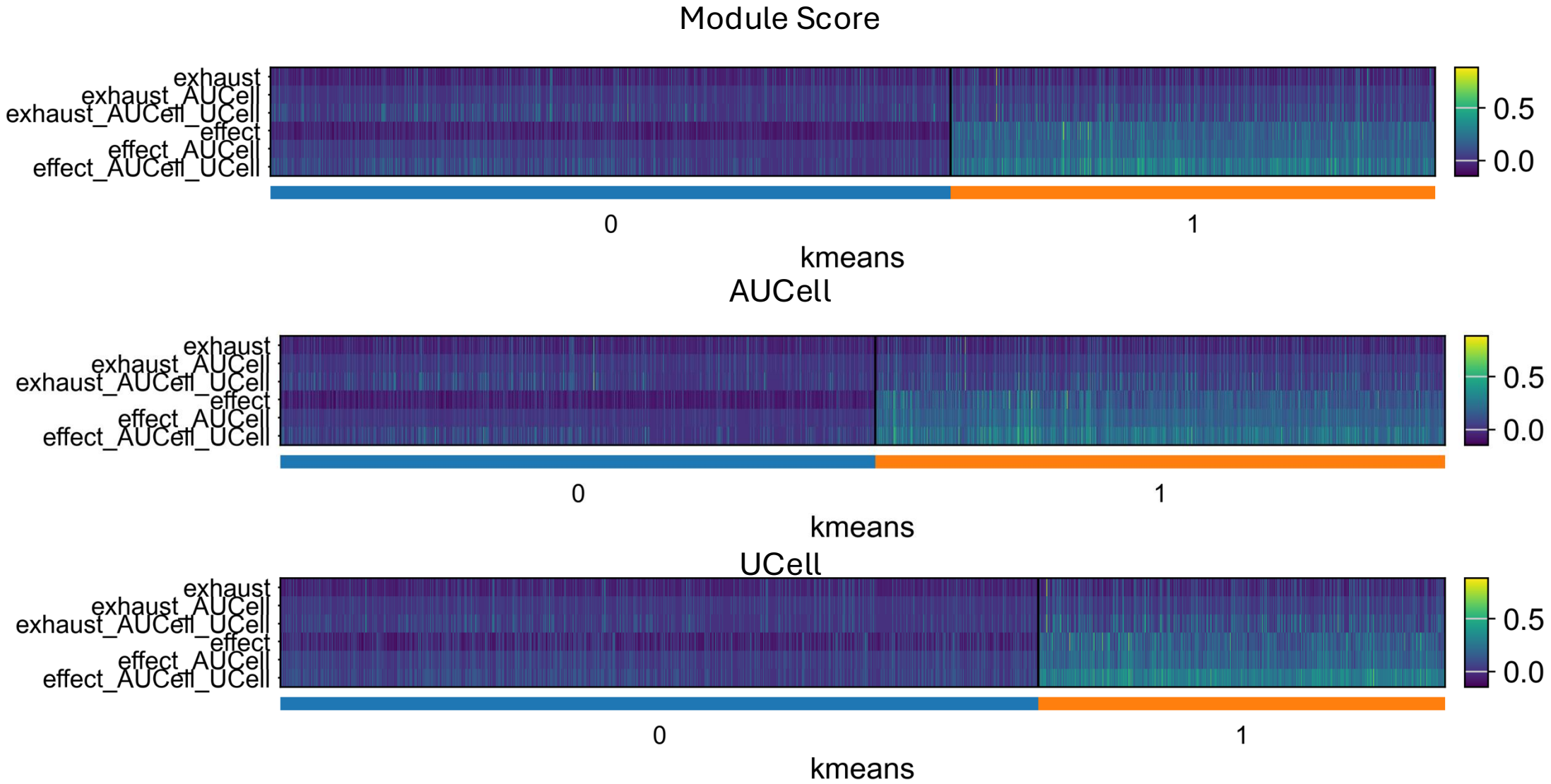
Exhaustion Module Score



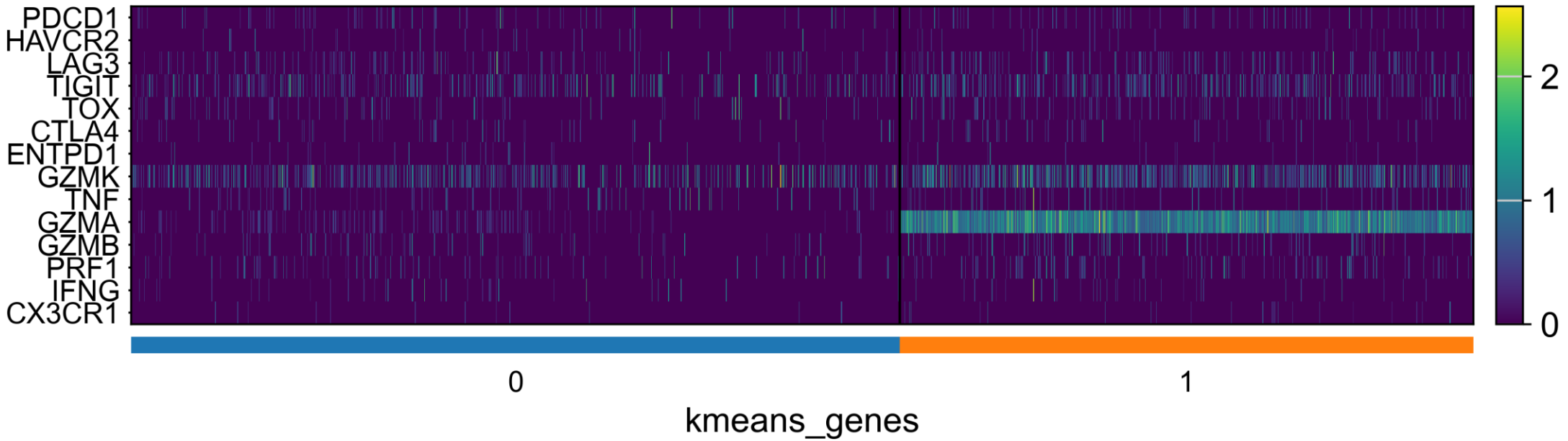
Results are consistent across methods



Exhaust VS Effect



Clustering with gene sets



single sample gene set enrichment analysis (ssGSEA)/GSVA

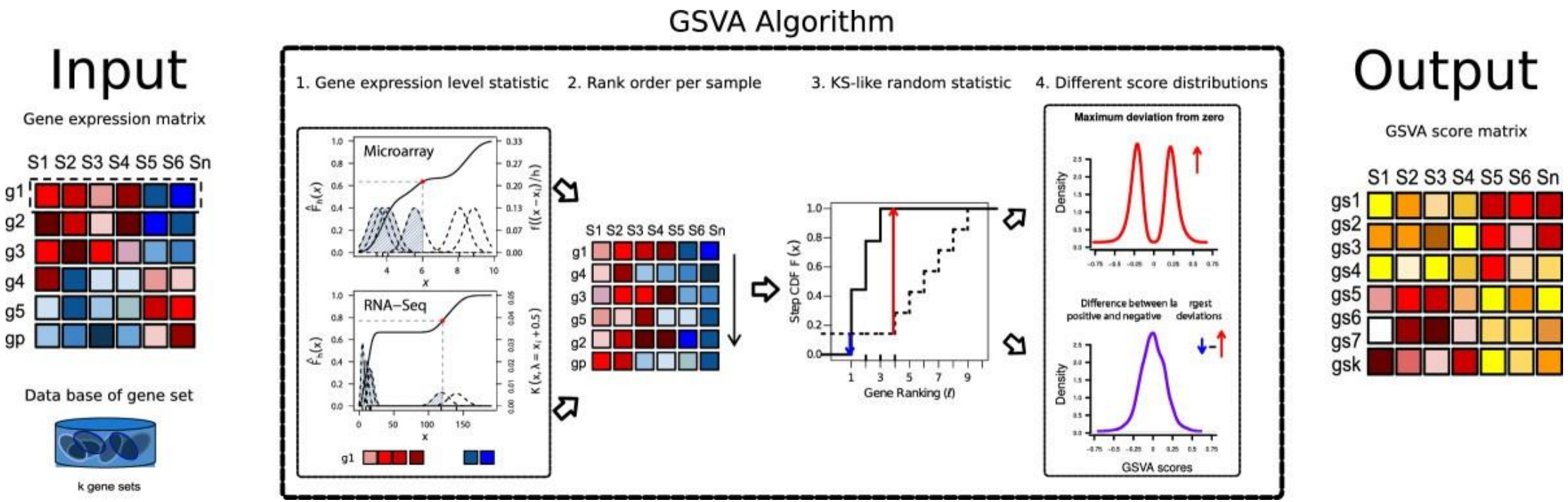
Software | Open access | Published: 16 January 2013

GSVA: gene set variation analysis for microarray and RNA-Seq data

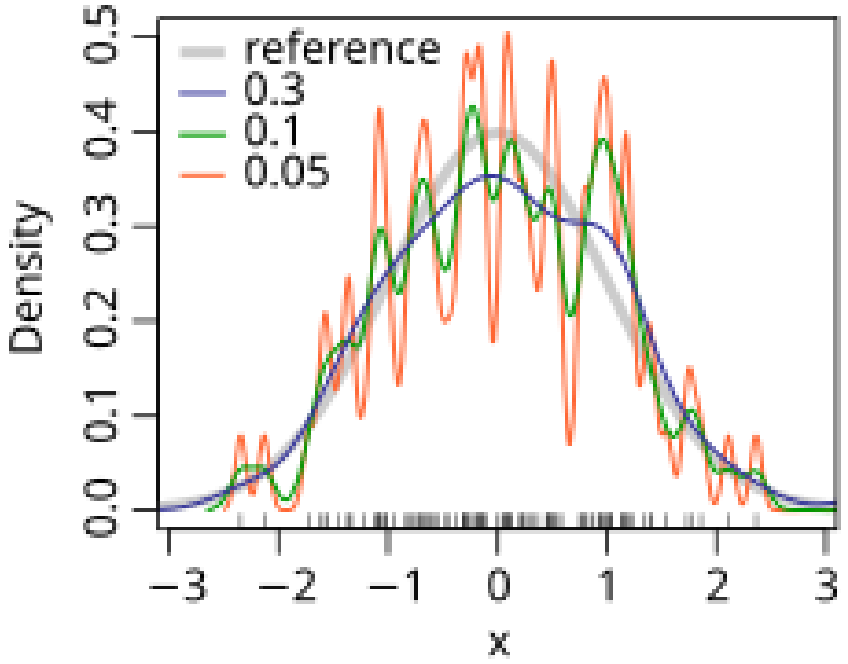
Sonja Hänzelmann, Robert Castelo & Justin Guinney

BMC Bioinformatics **14**, Article number: 7 (2013) | Cite this article

single sample gene set enrichment analysis (ssGSEA)/GSVA



Kernel Density Estimation (KDE)



https://en.wikipedia.org/wiki/Kernel_density_estimation

Step 1: gene level statistic

a matrix $X=\{x_{ij}\}_{p\times n}$
x_i the expression profile of the i-th gene
x_{ij} the specific expression value of the i-th gene in the j-th sample
γ_k the subset of row indices in X such that $\gamma_k \subset \{1,\dots p\}$ defines a set of genes forming a pathway or some other functional unit
$|\gamma_k|$ be the number of genes in γ_k

$$\widehat{F}_{h_i}\left(x_{ij}\right) = \frac{1}{n}\sum_{k=1}^{n}\int_{-\infty}^{\frac{x_{ij}-x_{ik}}{h_i}} \frac{1}{\sqrt{2\Pi}} e^{-\frac{t^2}{2}}\, dt \ ,$$

Microarray: Gaussian Kernel

$$\widehat{F}_{r}\left(x_{ij}\right) = \frac{1}{n}\sum_{k=1}^{n}\sum_{y=0}^{x_{ij}} \frac{e^{-(x_{ik}+r)}\left(x_{ik}+r\right)^{y}}{y!} \ ,$$

RNA-seq: Poisson Kernel

Step 2: Rank order per sample

a matrix $X=\{x_{ij}\}_{p\times n}$
x_i the expression profile of the i-th gene
x_{ij} the specific expression value of the i-th gene in the j-th sample
γ_k the subset of row indices in X such that $\gamma_k \subset \{1,\ldots p\}$ defines a set of genes forming a pathway or some other functional unit
$|\gamma_k|$ be the number of genes in γ_k

Convert z_{ij} to ranks $z_{(i)j}$ for each sample j and normalize further $r_{ij}=|p/2-z_{(i)j}|$

Step 3: KS like random statistic

a matrix $X=\{x_{ij}\}_{p\times n}$
x_i the expression profile of the i-th gene
x_{ij} the specific expression value of the i-th gene in the j-th sample
γ_k the subset of row indices in X such that $\gamma_k \subset \{1,\dots p\}$ defines a set of genes forming a pathway or some other functional unit
$|\gamma_k|$ be the number of genes in γ_k

$$\nu_{jk}\left(\ell\right) = \frac{\sum_{i=1}^{\ell}\left|r_{ij}\right|^{\tau} I\left(g_{(i)} \in \gamma_k\right)}{\sum_{i=1}^{p}\left|r_{ij}\right|^{\tau} I\left(g_{(i)} \in \gamma_k\right)} - \frac{\sum_{i=1}^{\ell} I\left(g_{(i)} \notin \gamma_k\right)}{p-|\gamma_k|},$$

Step 4: Scoring

Max deviation from zero

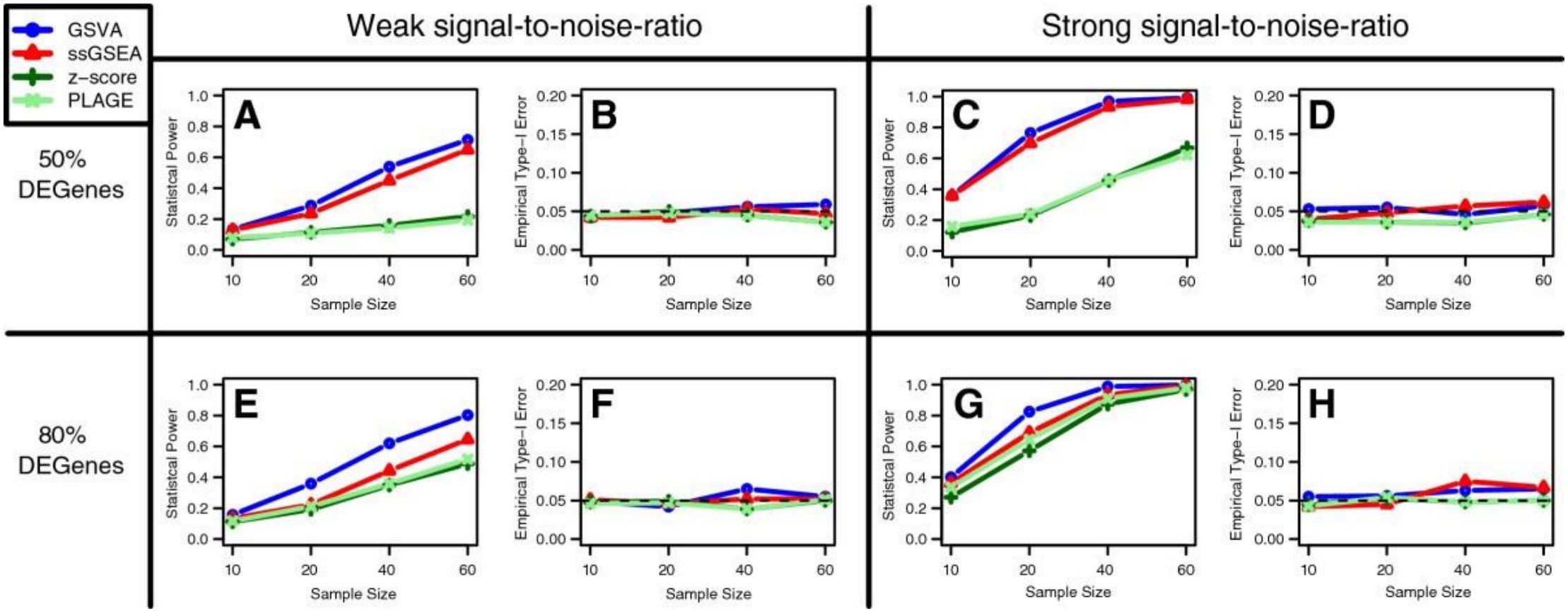
$$ES_{jk}^{\max} = \nu_{jk}\left[\arg \max_{\ell=1,\ldots,p} |\nu_{jk}(\ell)|\right].$$

Difference between positive and negative

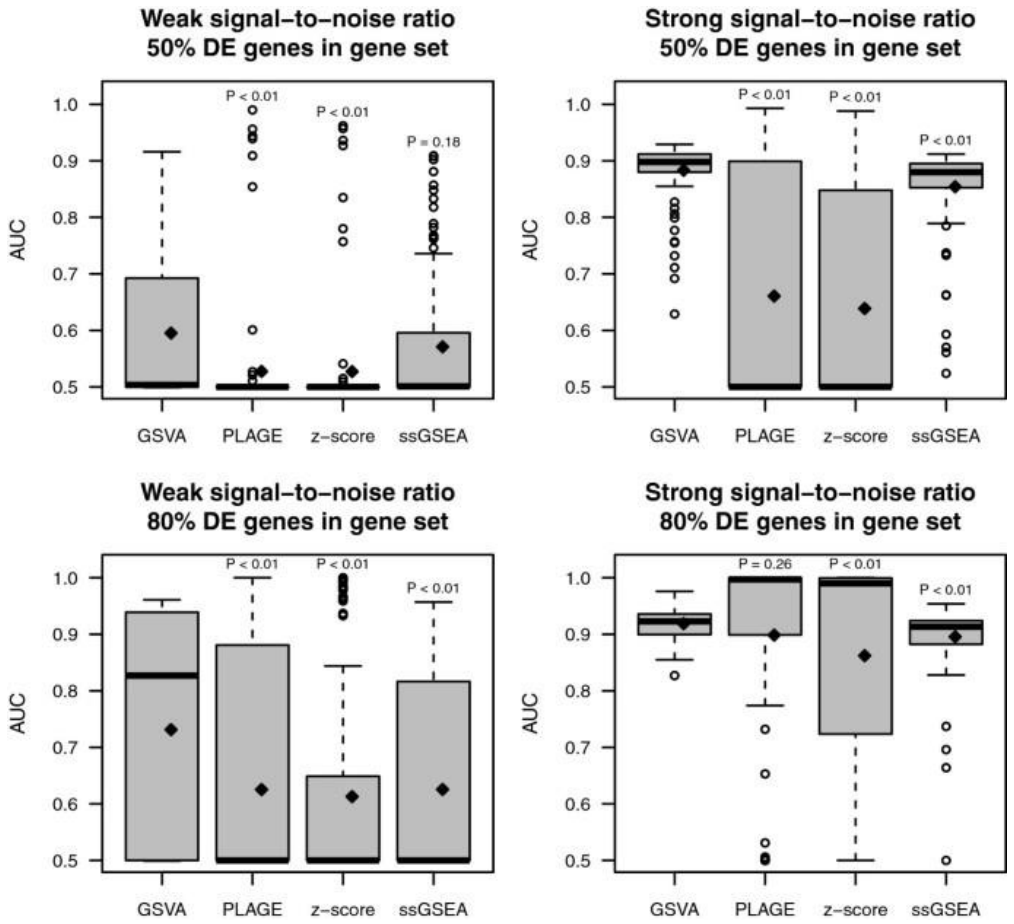
$$ES_{jk}^{\text{diff}} = \left|ES_{jk}^{+}\right| - \left|ES_{jk}^{-}\right| = \max_{\ell=1,\ldots,p}\left(0, \nu_{jk}\left(\ell\right)\right) - \min_{\ell=1,\ldots,p}\left(0, \nu_{jk}\left(\ell\right)\right)$$

ES_{jk}^{+} and ES_{jk}^{-} are the largest positive and negative random walk deviations from zero

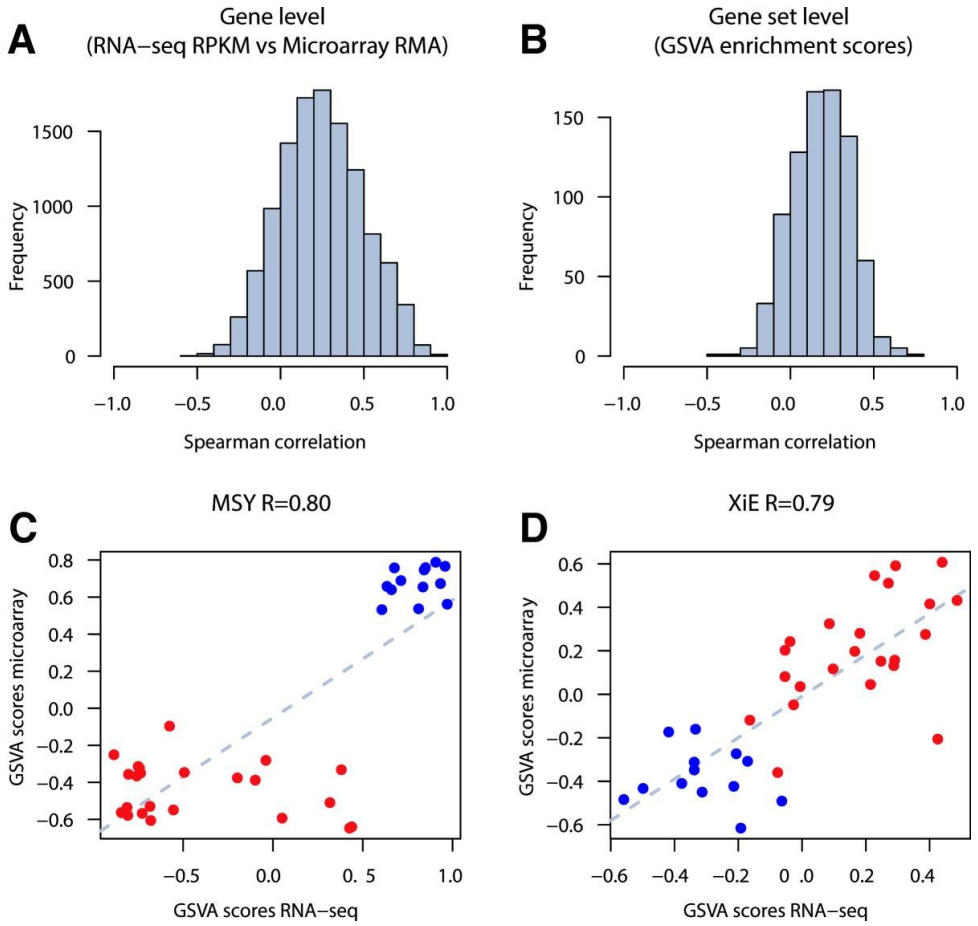
GSVA performs better than the other methods



GSVA performs better than the other methods



GSVA performs comparably in microarray and RNA-seq



Questions