Advanced topics

- Geneset testing
- Single cell RNA-seq

OPEN ACCESS Freely available online

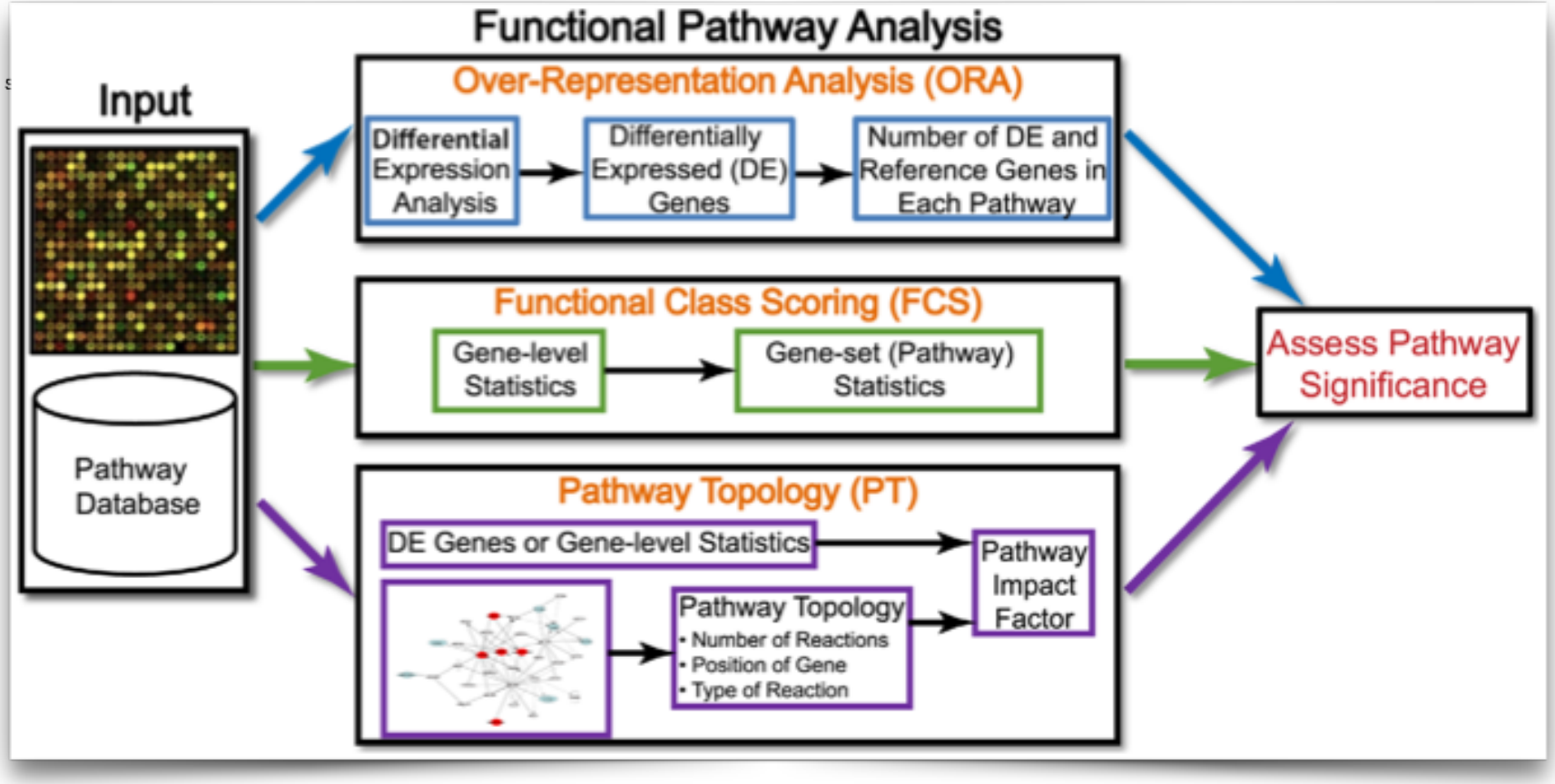
PLoS COMPUTATIONAL BIOLOGY

Review

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri[1,2*], Marina Sirota[1,2], Atul J. Butte[1,2*]

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, S
Children's Hospital, Palo Alto, California, United States of America



Casting differential expression onto biological knowledge: Functional category analysis versus gene set analysis

Motivation: DE genes might belong to a known pathway or might be the top genes from a related experiment; gene set as a whole might be altered, even if individual genes are not.

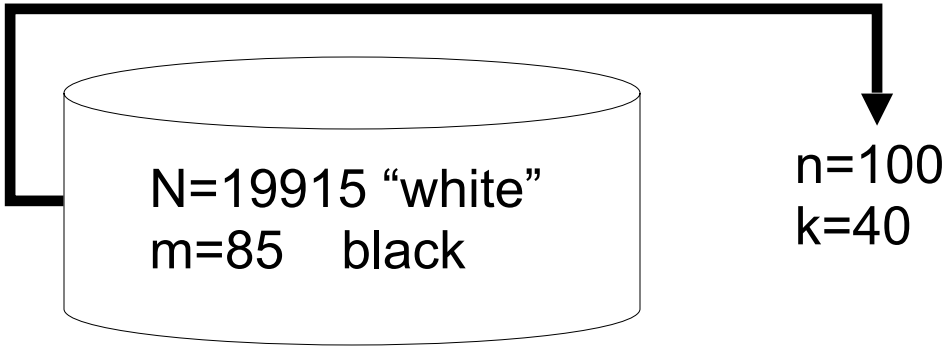
Starting point:	threshod, set of DE genes	gene-level statistics
Tool examples:	DAVID [C] **goseq** [C]	GSEA [S] **roast** [S] **CAMERA** [C]

S = self-contained
C = competitive

Functional category analysis: Overlap statistics

Question: Say you have a set of 85 genes (of a total 20000 genes) known to be associated with some function. Calculate the probability of randomly selecting 40 or more (overrepresented) of those genes in a list of 100 DE genes.

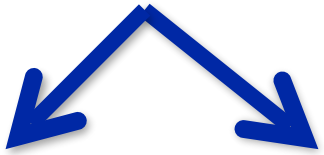
Answer: Hypergeometric (i.e. the “urn” problem).



$$P(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

e.g. FunSpec (yeast) - Robinson et al. 2002 BMC Bfx; DAVID; topGO

Gene set analysis: what is the hypothesis (test)?

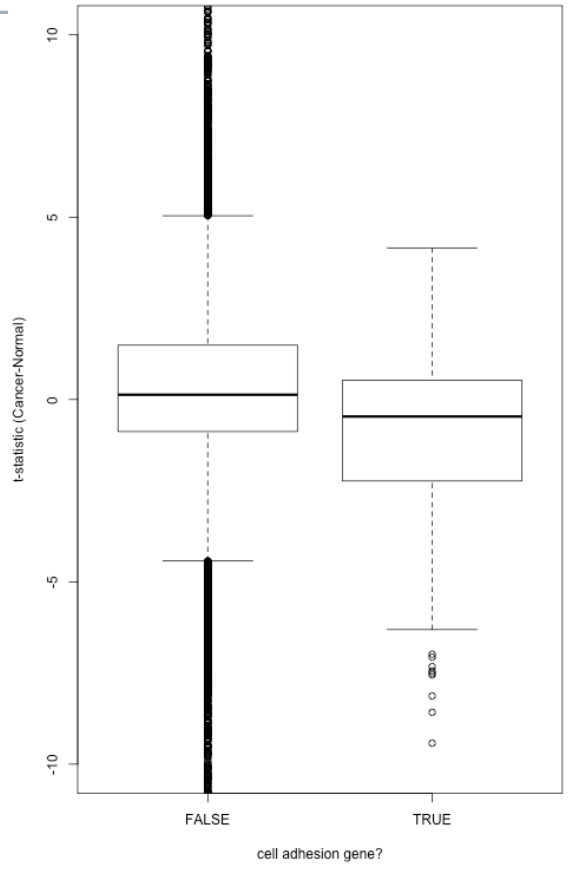


competitive	self-contained
Genes in the set tend to be more strongly DE than randomly chosen genes	*At least some genes in the set are truly DE*

Goeman & Bühlmann, *Bioinformatics* 2007

Viewing gene sets

Cell adhesion genes



Genes regulated by MYB

Gene set enrichment analysis (GSEA)

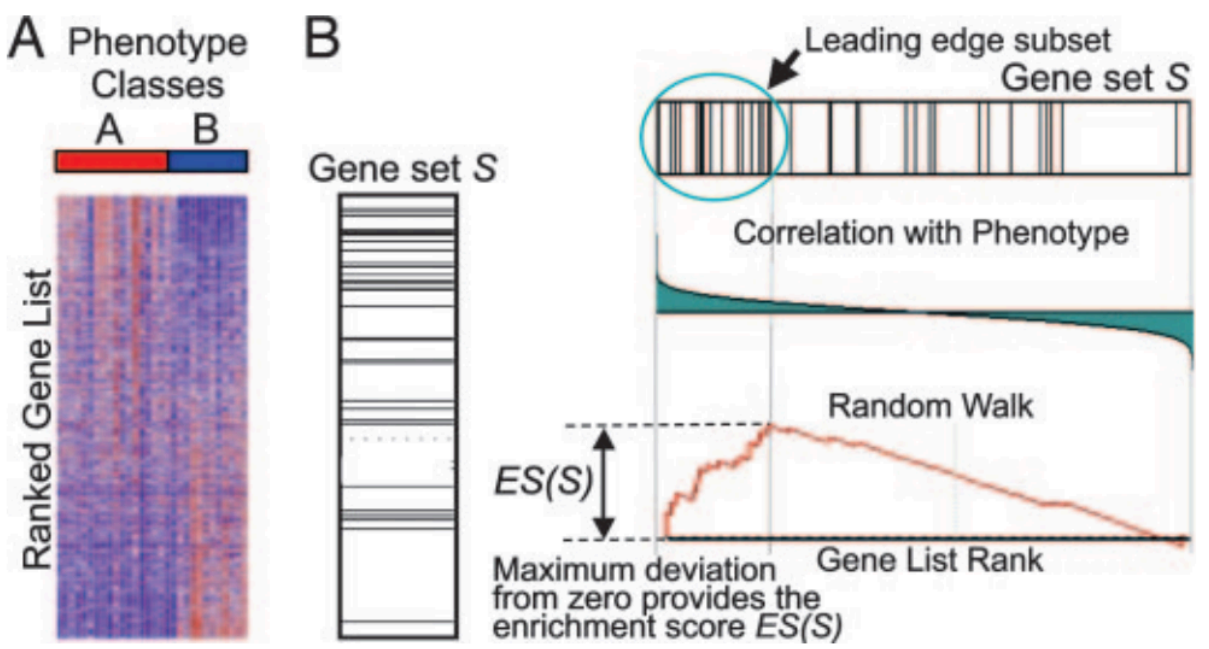


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Self-contained.

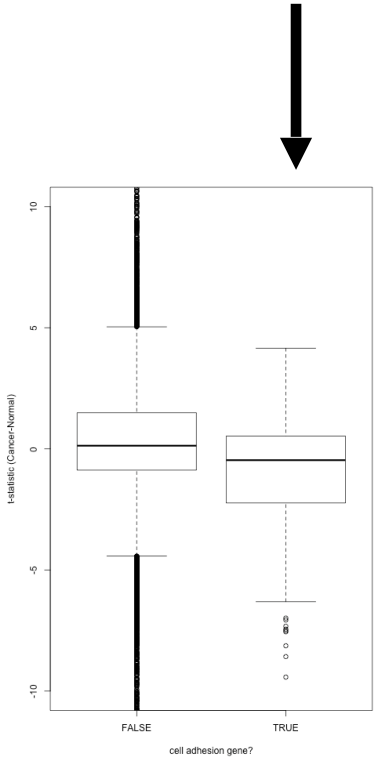
Permutation P-value: Sample permutation is done, which preserves gene correlation.

But, it has limited use in small samples (i.e. very few possible permutations).

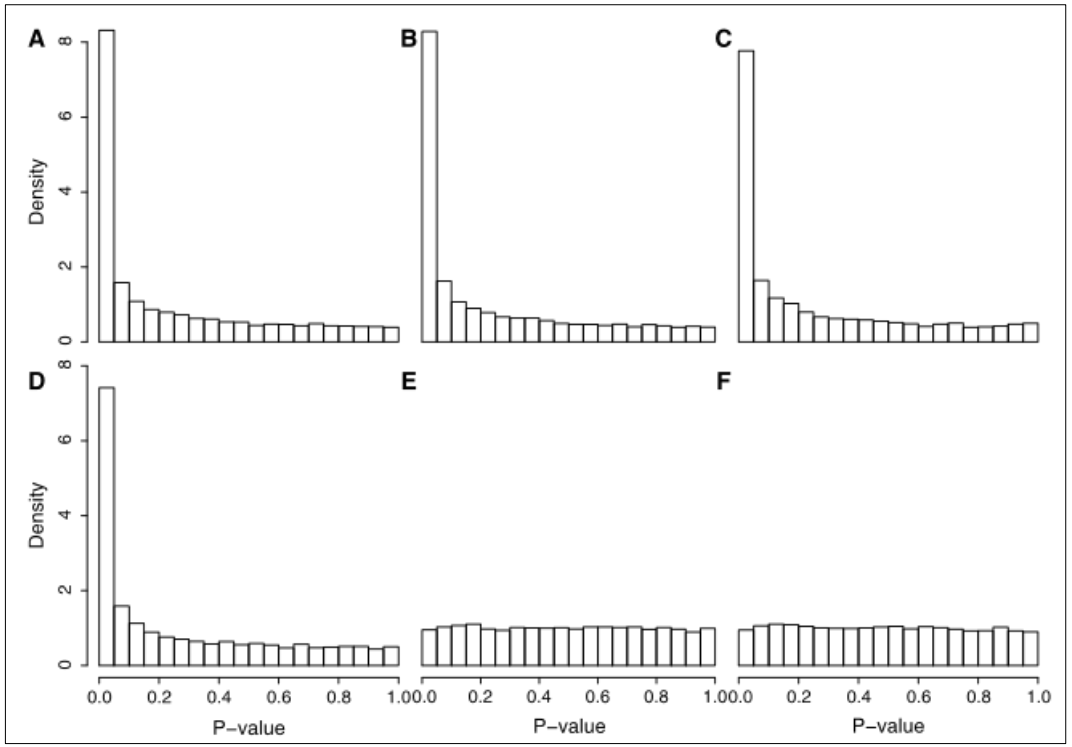
Now switches to a gene-based permutation (competitive) in small samples.

CAMERA (Correlation Adjusted MEan RAnk)

Cell adhesion genes



Main criticism of (naïve, gene-permutation) competitive tests is that the correlation structure is broken.



Distributions of p-value:

no differential expression

A geneSetTest
B geneSetTest [r]
C sigPathway
D PAGE
E CAMERA
F CAMERA [r]

Wu et al. NAR 2012

How much of this is storytelling?

A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans

Pavlos Pavlidis,*[1] Jeffrey D. Jensen,[2] Wolfgang Stephan,[3] and Alexandros Stamatakis[1]

[1]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

[2]Ecole Polytechnique Fédérale de Lausanne, School of Life Sciences, Lausanne, Switzerland

[3]Section of Evolutionary Biology, Biocenter, University of Munich, Planegg-Martinsried, Germany

***Corresponding author:** E-mail: pavlidisp@gmail.com.

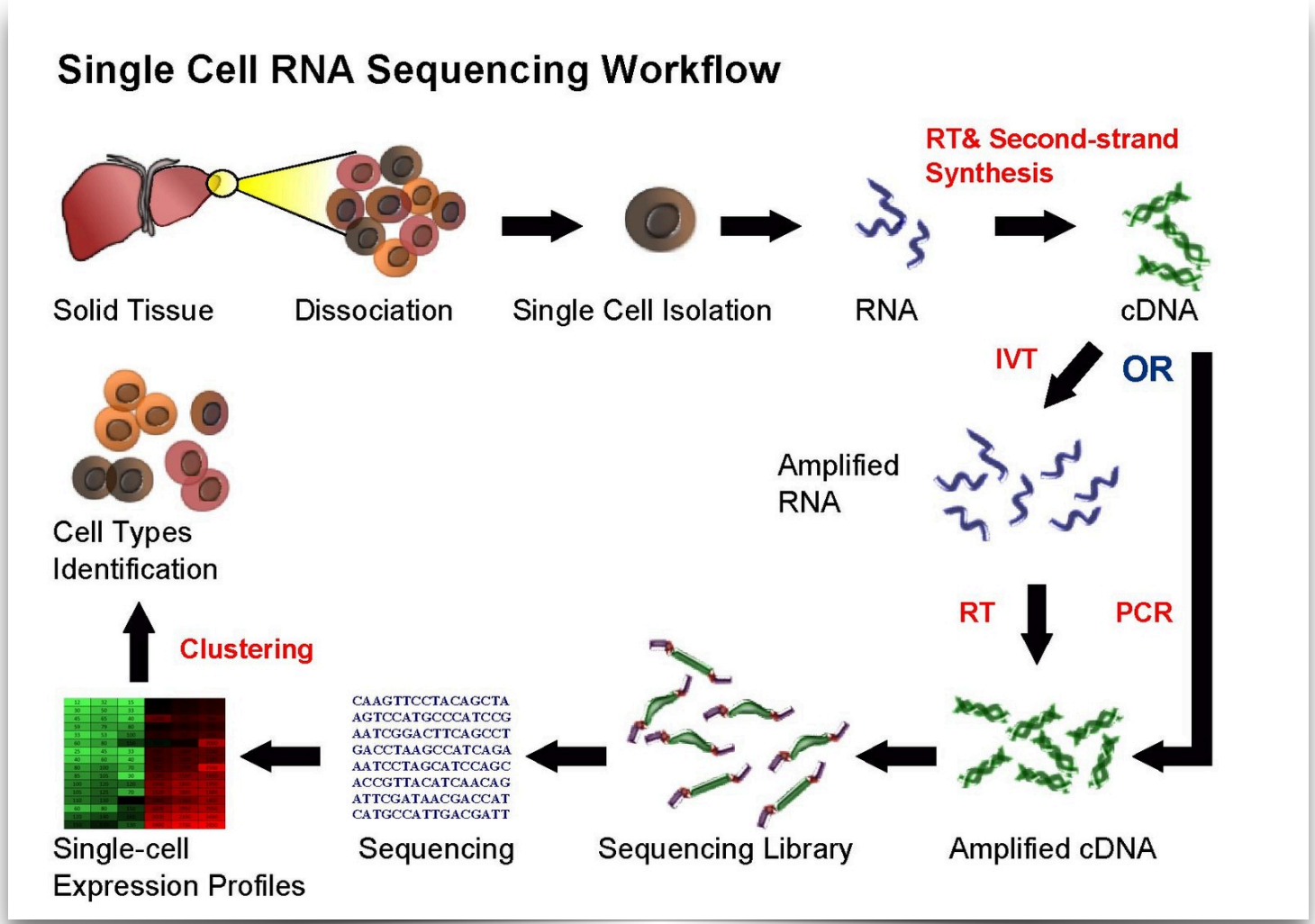
Associate editor: Arndt von Haeseler

Abstract

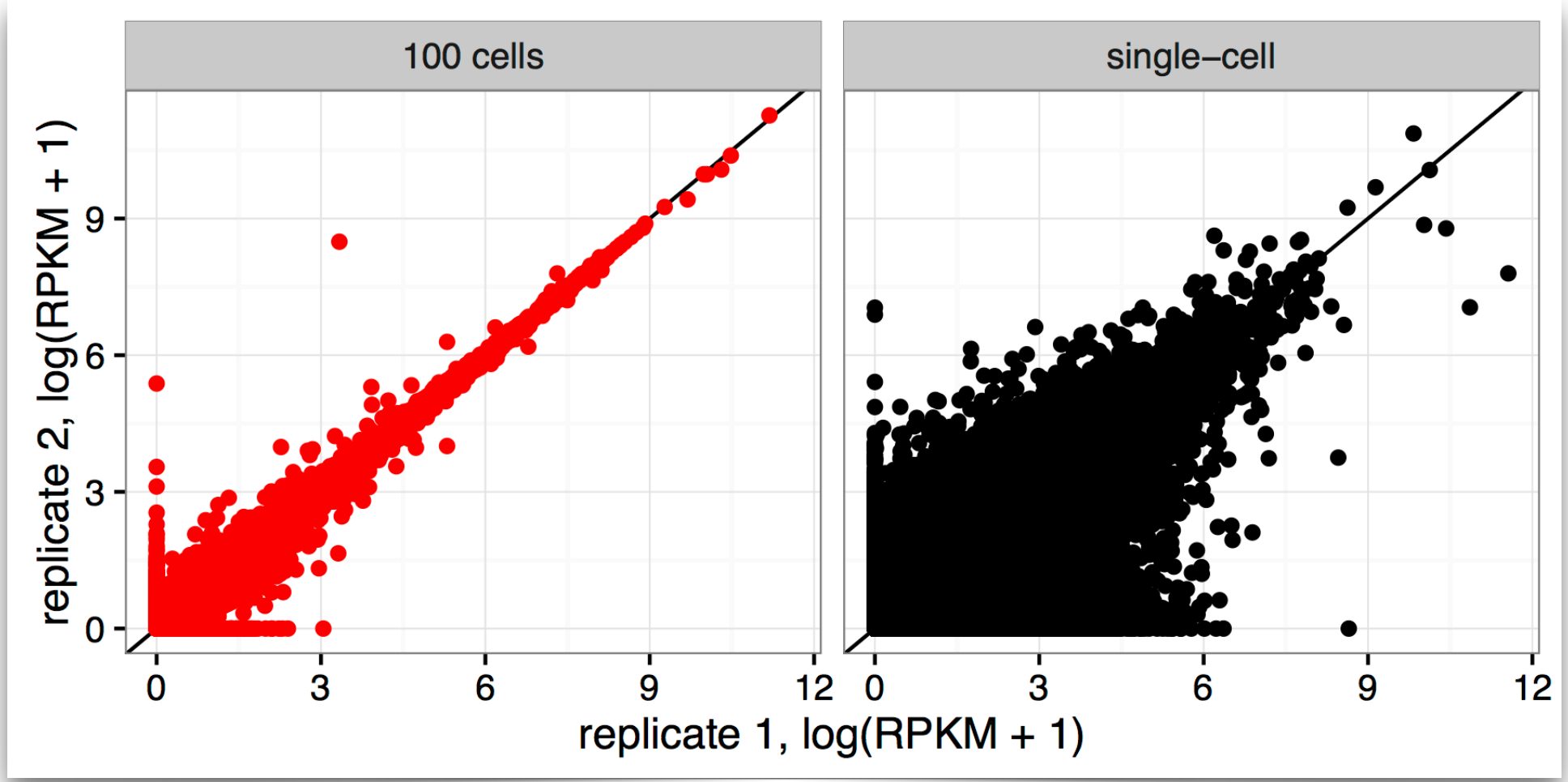
In the age of whole-genome population genetics, so-called genomic scan studies often conclude with a long list of putatively selected loci. These lists are then further scrutinized to annotate these regions by gene function, corresponding biological processes, expression levels, or gene networks. Such annotations are often used to assess and/or verify the validity of the genome scan and the statistical methods that have been used to perform the analyses. Furthermore, these results are frequently considered to validate "true-positives" if the identified regions make biological sense a posteriori. Here, we show that this approach can be potentially misleading. By simulating neutral evolutionary histories, we demonstrate that it is possible not only to obtain an extremely high false-positive rate but also to make biological sense out of the false-positives and construct a sensible biological narrative. Results are compared with a recent polymorphism data set from *Drosophila melanogaster*.

Key words: genome scanning, positive selection, gene ontology, validation, literature mining.

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



Basic properties: Variability levels



Single-cell RNA-seq: Hypothetical situations

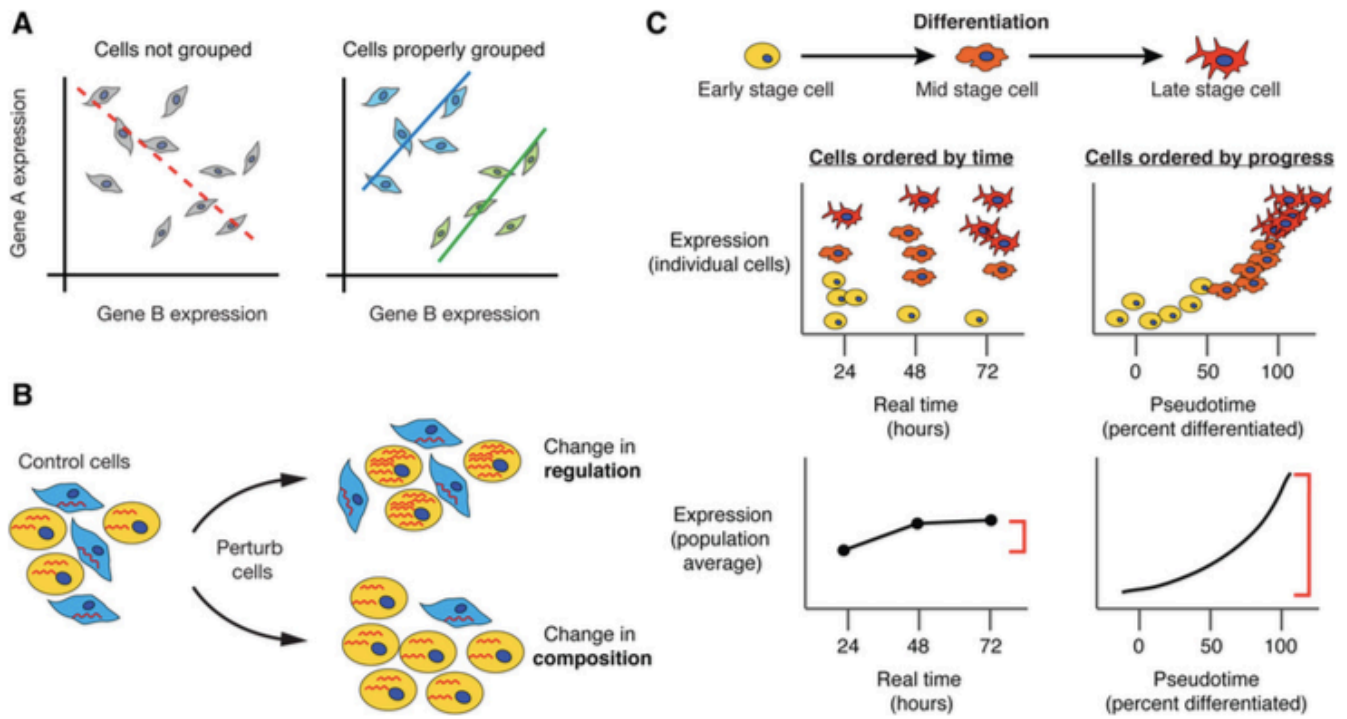
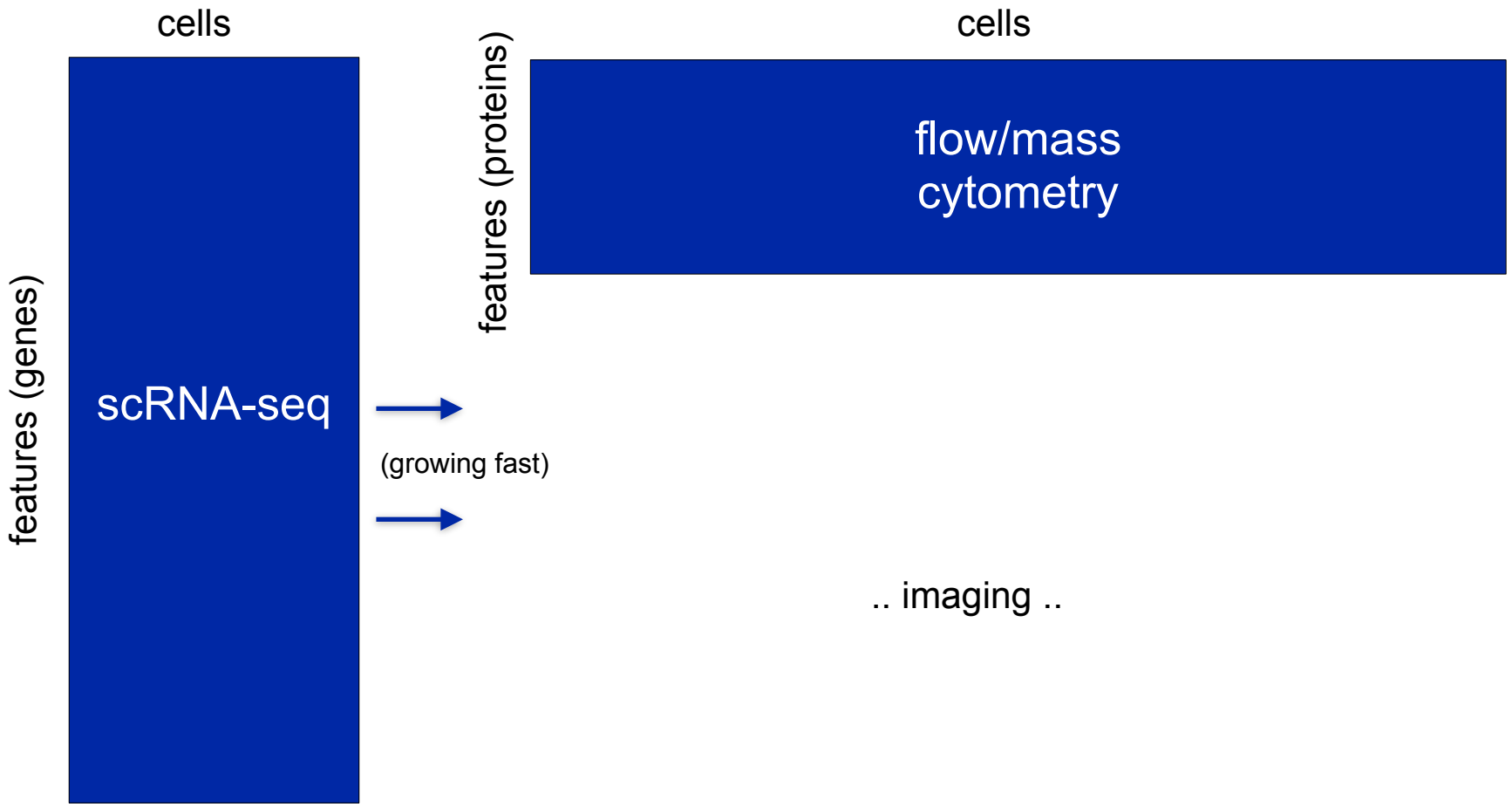


Figure 1. Single-cell measurements preserve crucial information that is lost by bulk genomics assays. (*A*) Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals. (*B*) Bulk measurements cannot distinguish changes due to gene regulation from those that arise due to shifts in the ratio of different cell types in a mixed sample. (*C*) Time series experiments are affected by averaging when cells proceed through a biological process in an unsynchronized manner. A single time point may contain cells from different stages in the process, obscuring the dynamics of relevant genes. Reordering the cells in "pseudotime" according to biological progress eliminates averaging and recovers the true signal in expression (Trapnell et al. 2014).

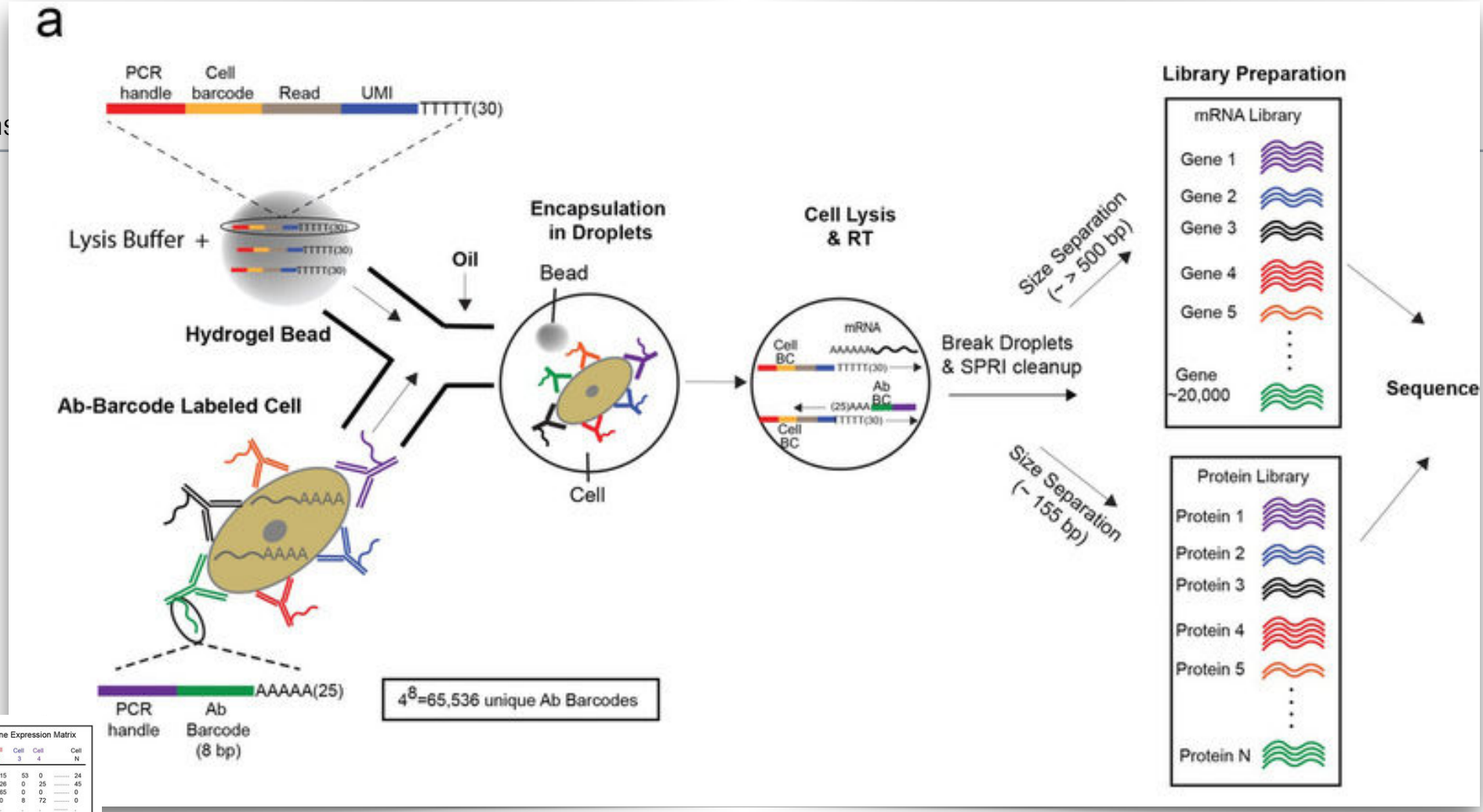
Trapnell, 2015 Genome Research

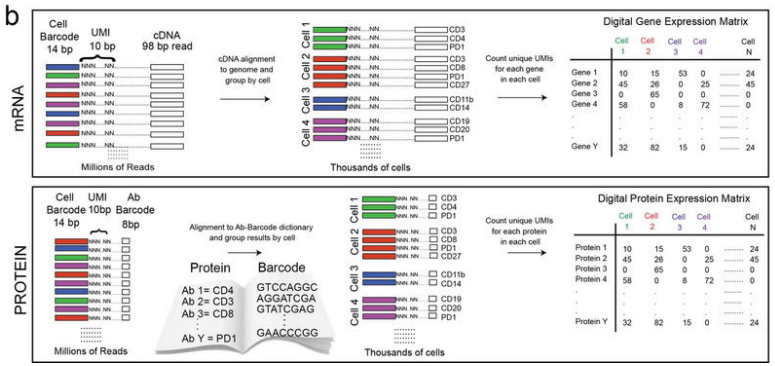
Different shapes of single cell data



Dual assays

- REAP-seq
- CITE-seq





Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]

Some terminology: Cell identity, type, state, ..

Box 1 The many facets of a cell's identity

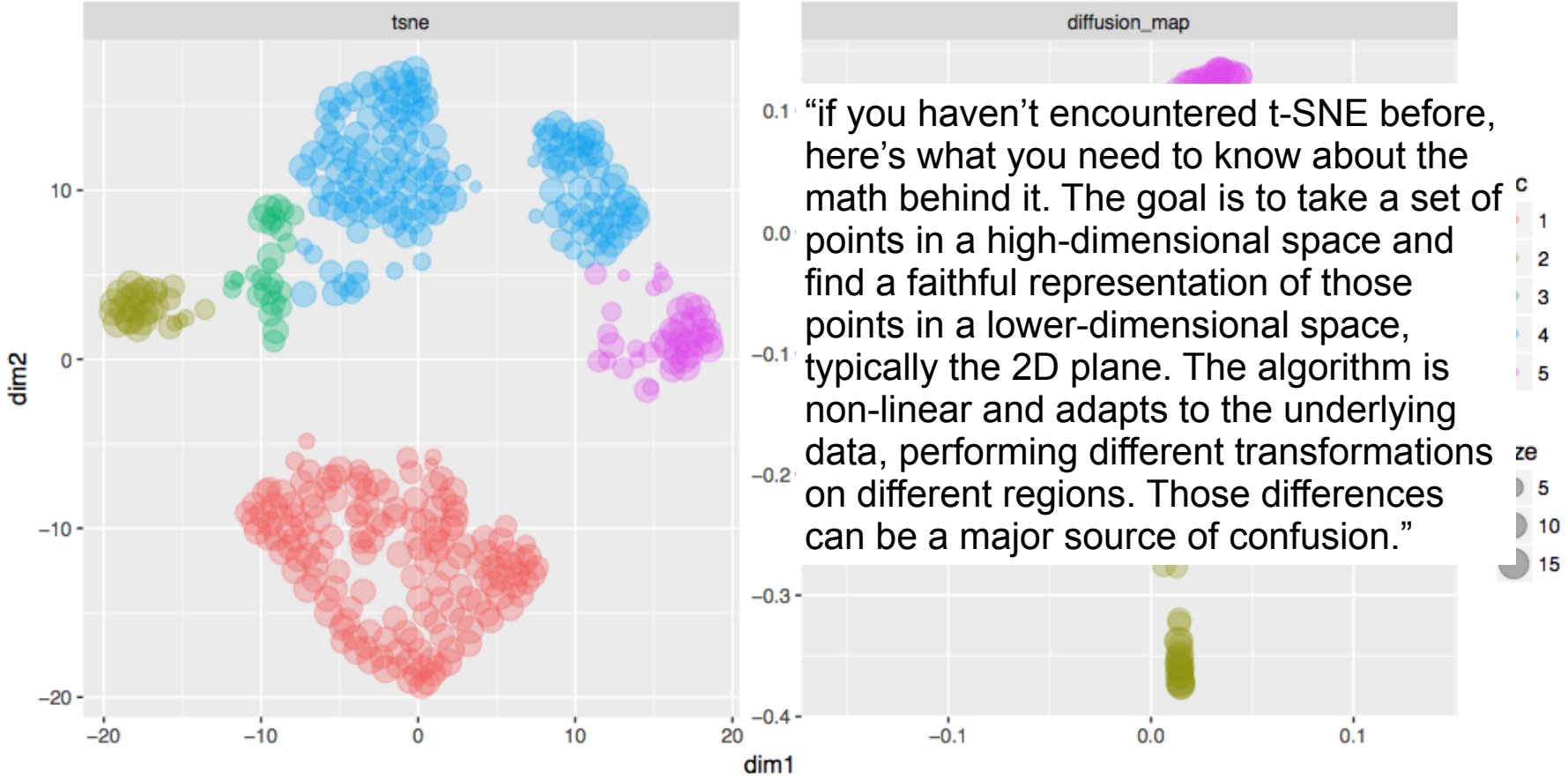
We define a cell's *identity* as the outcome of the instantaneous intersection of all factors that affect it. We refer to the more permanent aspects in a cell's identity as its *type* (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its *state*. Cell types are often organized in a hierarchical taxonomy, as types may be further divided into finer subtypes; such taxonomies are often related to a cell fate map, reflecting key steps in differentiation. Cell *states* arise transiently during time-dependent processes, either in a *temporal progression* that is unidirectional (e.g., during differentiation, or following an environmental stimulus) or in a *state vacillation* that is not necessarily unidirectional and in which the cell may return to the origin state. Vacillating processes can be *oscillatory* (e.g., cell-cycle or circadian rhythm) or can transition between states with no predefined order (e.g., due to stochastic, or environmentally controlled, molecular events). These time-dependent processes may occur transiently within a stable cell type (as in a transient environmental response), or may lead to a new,

Type: permanent

State: transient

http://distill.pub/2016/misread-tsne/

tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps

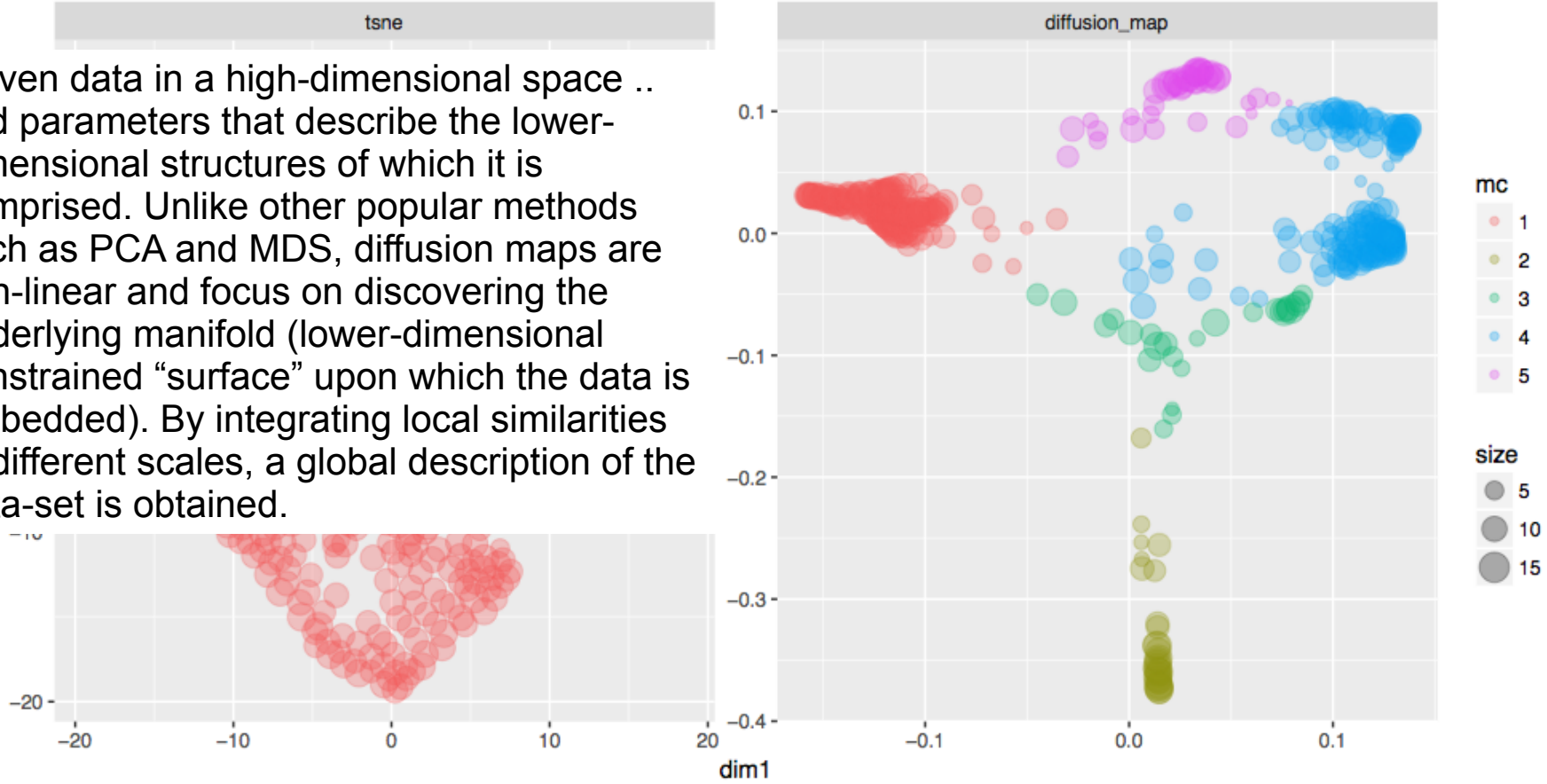


"if you haven't encountered t-SNE before, here's what you need to know about the math behind it. The goal is to take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane. The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions. Those differences can be a major source of confusion."

http://inside.mines.edu/~whereman/talks/delaPorte-Herbst-Hereman-vanderWalt-DiffusionMaps-PRASA2008.pdf

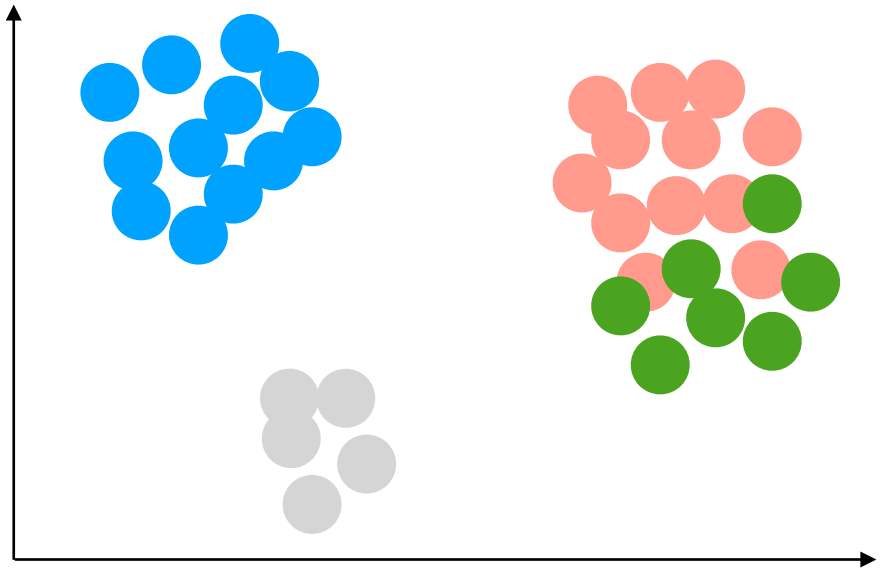
tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps

"Given data in a high-dimensional space .. find parameters that describe the lower-dimensional structures of which it is comprised. Unlike other popular methods such as PCA and MDS, diffusion maps are non-linear and focus on discovering the underlying manifold (lower-dimensional constrained "surface" upon which the data is embedded). By integrating local similarities at different scales, a global description of the data-set is obtained.

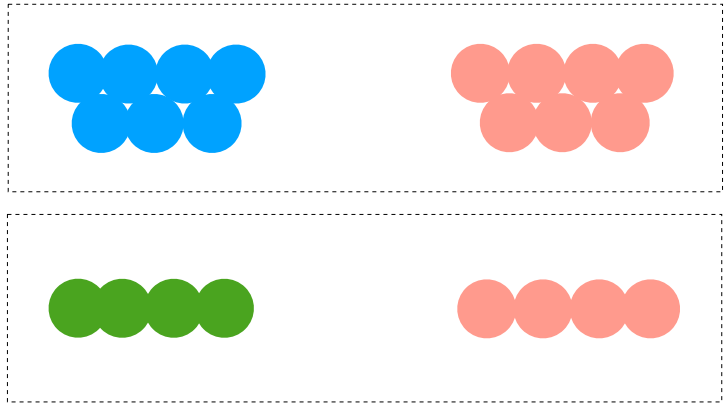


Experimental data

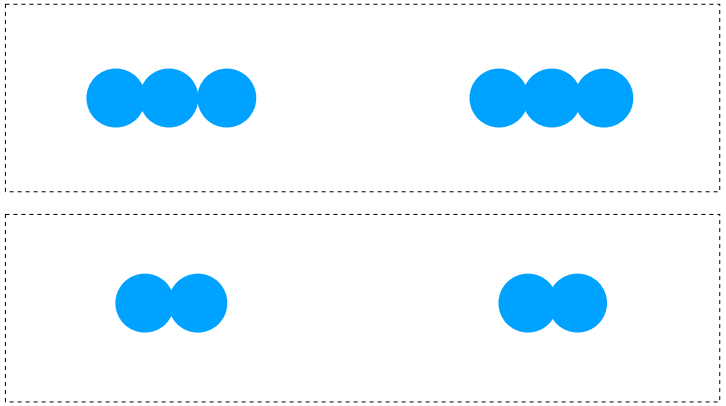
Using labels from independent truth (time point, fluoresence)



“Signal” data sets



“Mock” or null data sets



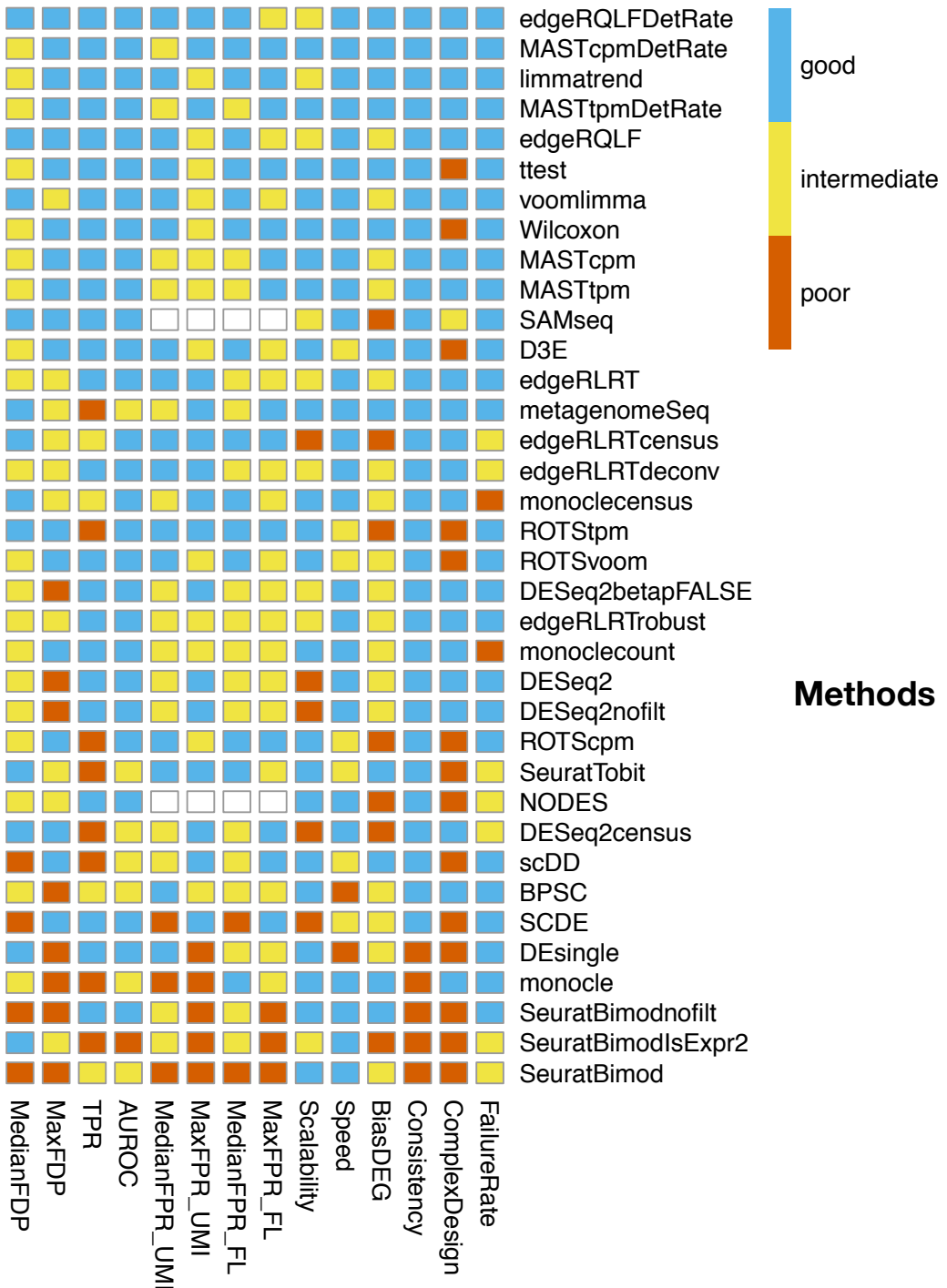
Between cell-type DE Benchmark (finding marker genes)

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson[1,2] & Mark D Robinson[1,2]

RECEIVED 6 JUNE 2017; ACCEPTED 16 JANUARY 2018; PUBLISHED ONLINE 26 FEBRUARY 2018;

"we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq."



conquer - reprocessed data analysis read public scRNA-seq datasets

- http://imlspenticton.uzh.ch:3838/conquer/
- Contains both full-length and UMI-based protocols

conquer ≡

About | scRNA-seq data sets | Changelog | Excluded samples | Tutorial

Show 10 entries Search:

	Data set	ID	organism	ncells	MultiAssayExperiment	MultiQC report	scater report	salmon archive
1	EMTAB2805 (PMID 25599176)	Buettner2015	Mus musculus	288	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-23)
2	EMTAB3929 (PMID 27062923)	Petropoulos2016	Homo sapiens	1529	Download .rds (2017-07-22)	Download .html (2017-07-22)	Download .html (2017-07-22)	Download .tar.gz (2017-07-20)
3	GSE41265 (PMID 23685454)	Shalek2013	Mus musculus	18	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-09-09)
4	GSE44183-GPL11154 (PMID 23892778)	Xue2013	Homo sapiens	29	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-24)
5	GSE44183-GPL13112 (PMID 23892778)	Xue2013	Mus musculus	17	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-24)
6	GSE44183-GPL13112-trimmed (PMID 23892778)	Xue2013	Mus musculus	17	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-28)