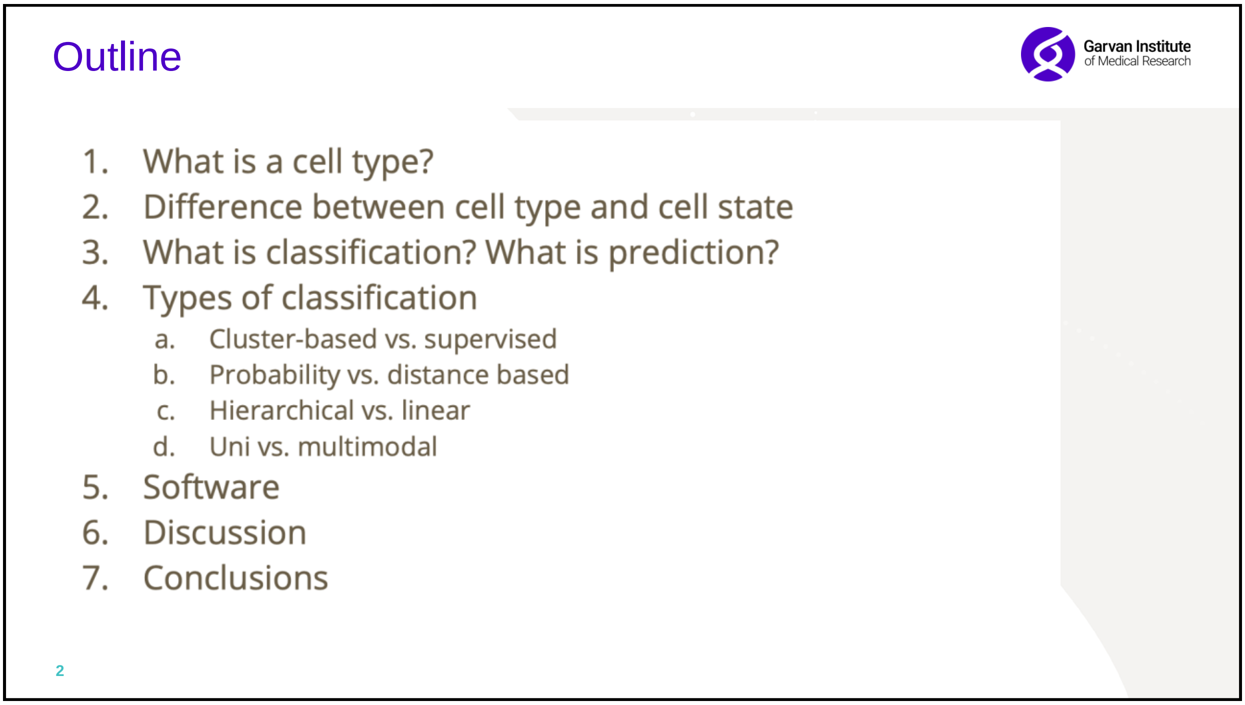
Garvan Institute of Medical Research

Single cell clustering and classification

Associate Professor Joseph Powell
Director, Garvan-Weizmann Centre for Cellular Genomics
Deputy Director, UNSW Cellular Genomics Futures Institute

SISG - 2019

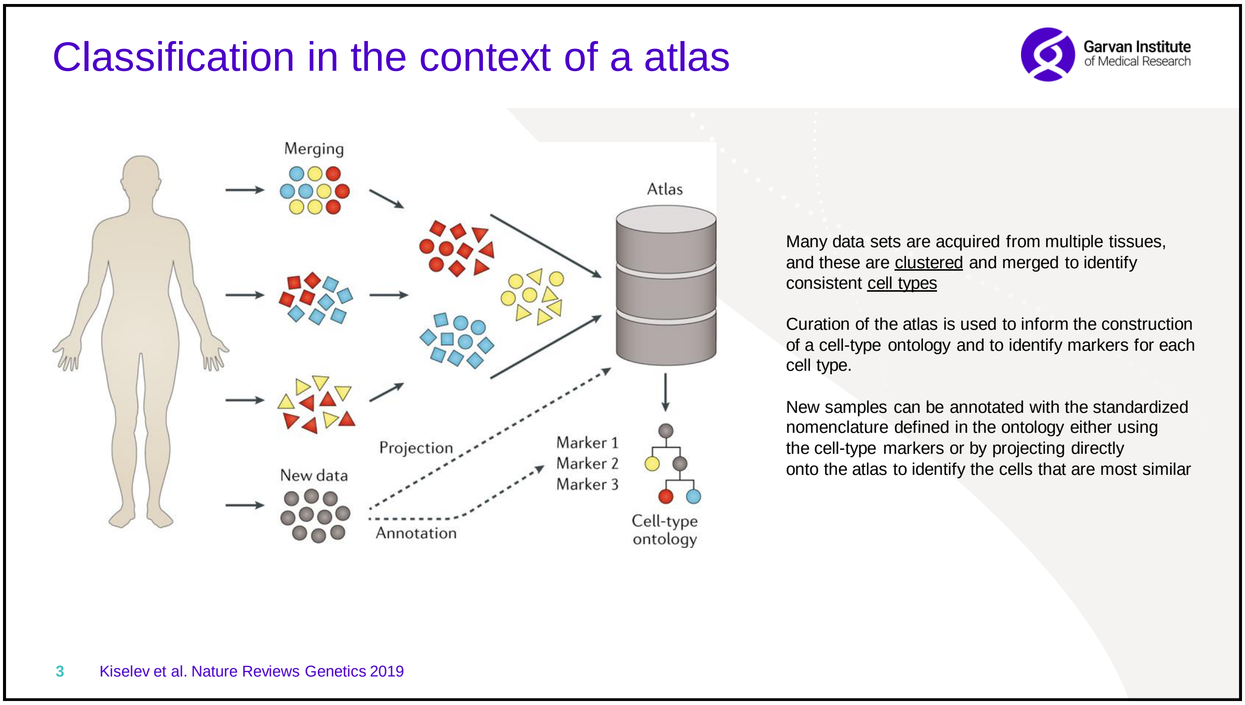


Outline

Garvan Institute of Medical Research

1. What is a cell type?
2. Difference between cell type and cell state
3. What is classification? What is prediction?
4. Types of classification
   a. Cluster-based vs. supervised
   b. Probability vs. distance based
   c. Hierarchical vs. linear
   d. Uni vs. multimodal
5. Software
6. Discussion
7. Conclusions

2



Classification in the context of a atlas

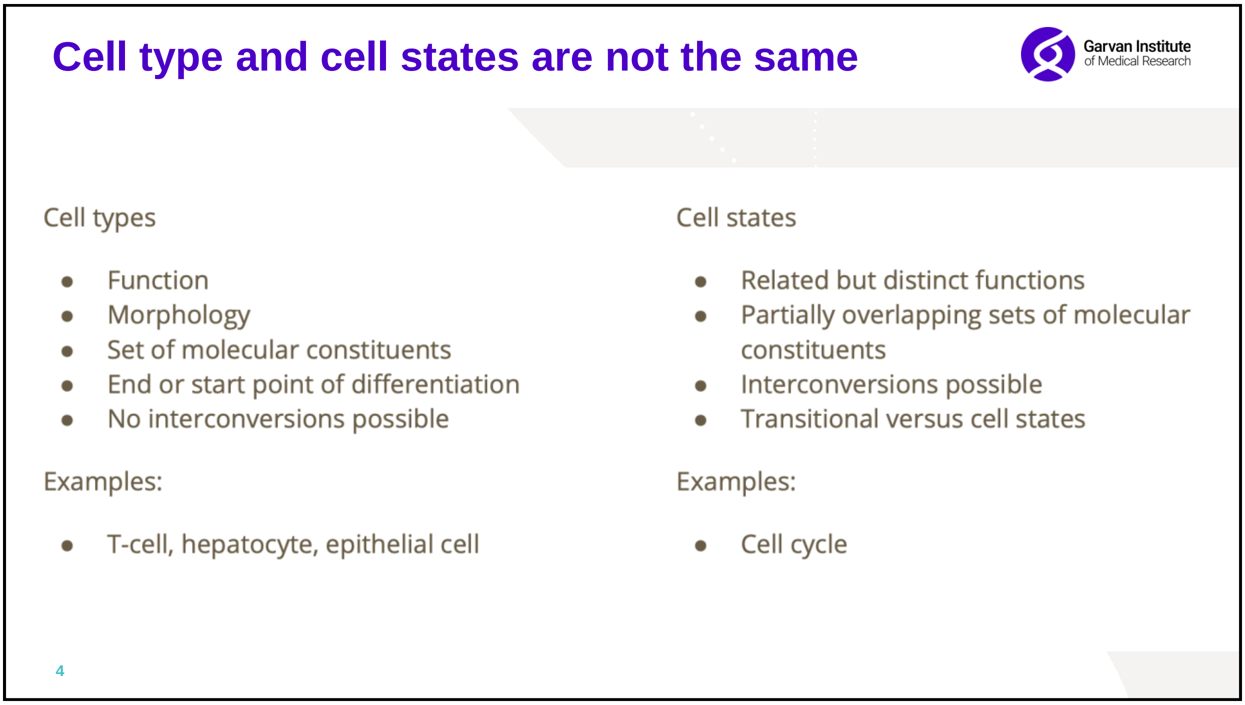
Garvan Institute of Medical Research

Many data sets are acquired from multiple tissues, and these are clustered and merged to identify consistent cell types

Curation of the atlas is used to inform the construction of a cell-type ontology and to identify markers for each cell type.

New samples can be annotated with the standardized nomenclature defined in the ontology either using the cell-type markers or by projecting directly onto the atlas to identify the cells that are most similar

Kiselev et al. Nature Reviews Genetics 2019



Cell type and cell states are not the same

Garvan Institute of Medical Research

Cell types

- Function
- Morphology
- Set of molecular constituents
- End or start point of differentiation
- No interconversions possible

Examples:

- T-cell, hepatocyte, epithelial cell

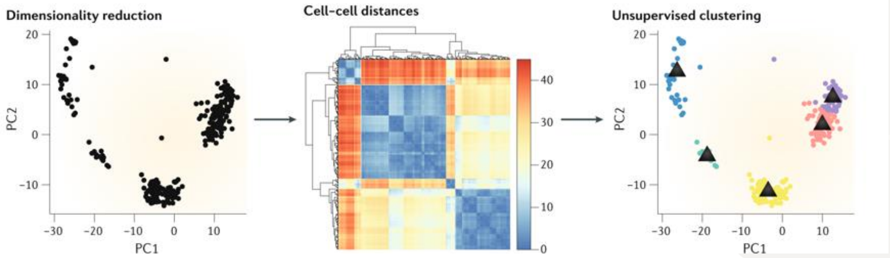
Cell states

- Related but distinct functions
- Partially overlapping sets of molecular constituents
- Interconversions possible
- Transitional versus cell states

Examples:

- Cell cycle

Clustering



Becht *et al*. Nat Biotech 2018

Duo *et al*. F1000 2018

There are lots of clustering methods, just don’t use K-means

Clustering – things to consider

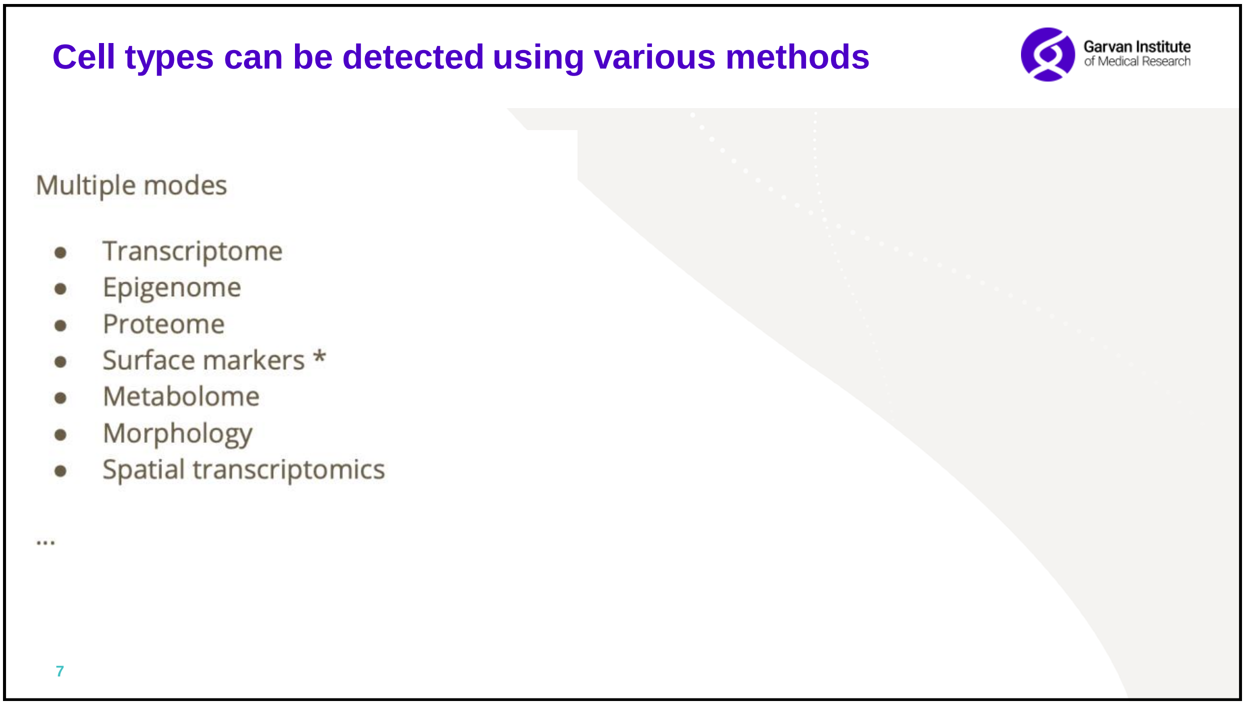
Classification

- Decision making
- Forced choice

Prediction

- Probabilistic interpretation

- Cell type prediction is based on the premise that a set of features (e.g. gene expression) are able to recapitulate the variance of the phenotype we are interested in



Cell types can be detected using various methods

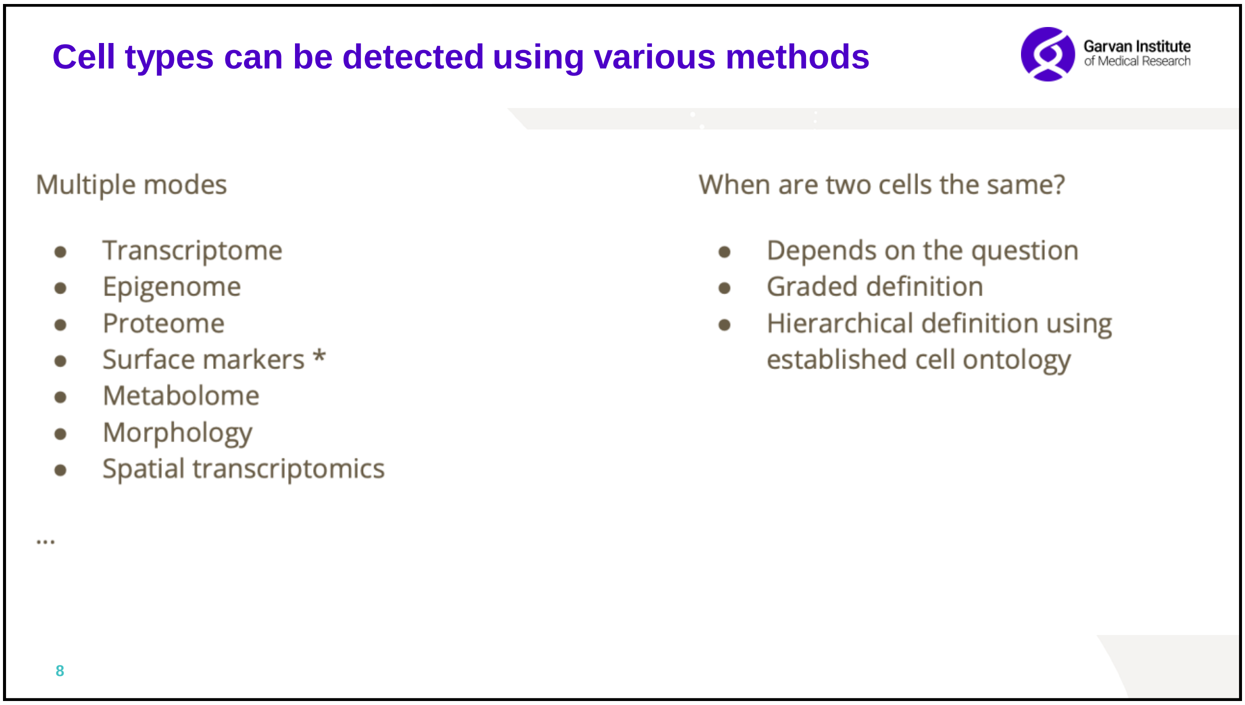
Garvan Institute of Medical Research

Multiple modes

- Transcriptome
- Epigenome
- Proteome
- Surface markers *
- Metabolome
- Morphology
- Spatial transcriptomics

…

7



Cell types can be detected using various methods

Garvan Institute of Medical Research

Multiple modes

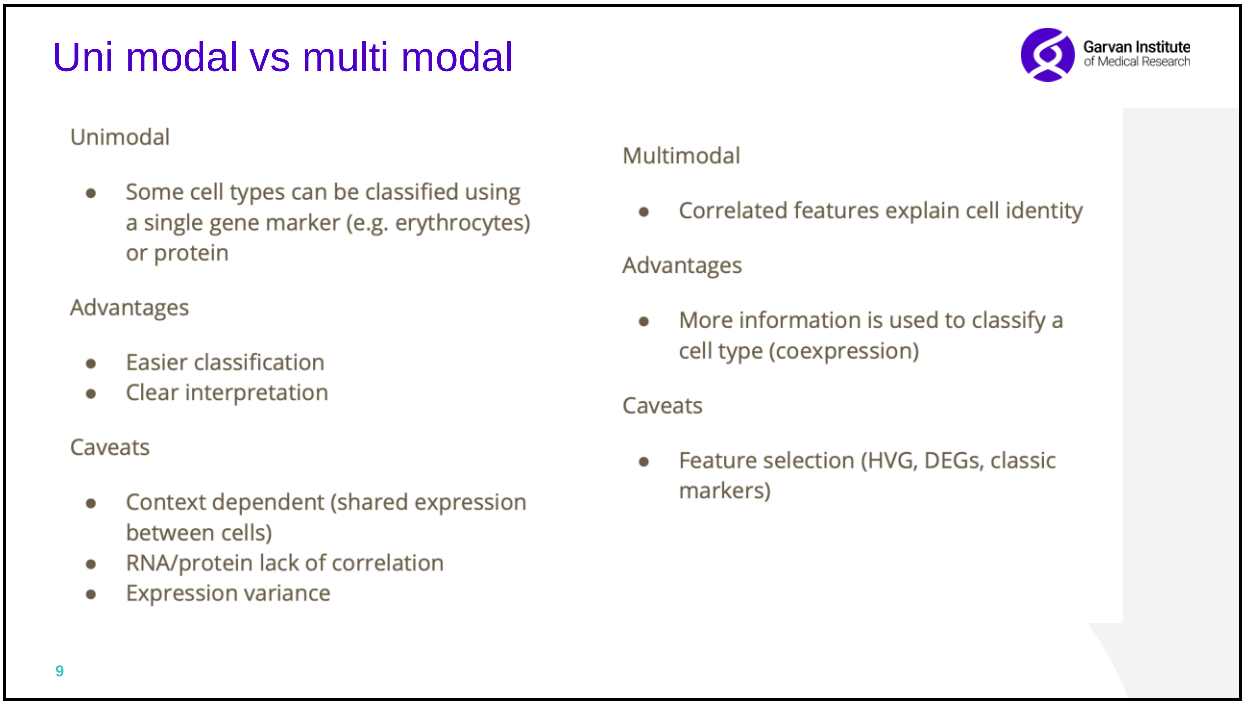
- Transcriptome
- Epigenome
- Proteome
- Surface markers *
- Metabolome
- Morphology
- Spatial transcriptomics

…

When are two cells the same?

- Depends on the question
- Graded definition
- Hierarchical definition using established cell ontology

8



Uni modal vs multi modal

Garvan Institute of Medical Research

Unimodal

- Some cell types can be classified using a single gene marker (e.g. erythrocytes) or protein

Advantages

- Easier classification
- Clear interpretation

Caveats

- Context dependent (shared expression between cells)
- RNA/protein lack of correlation
- Expression variance

Multimodal

- Correlated features explain cell identity

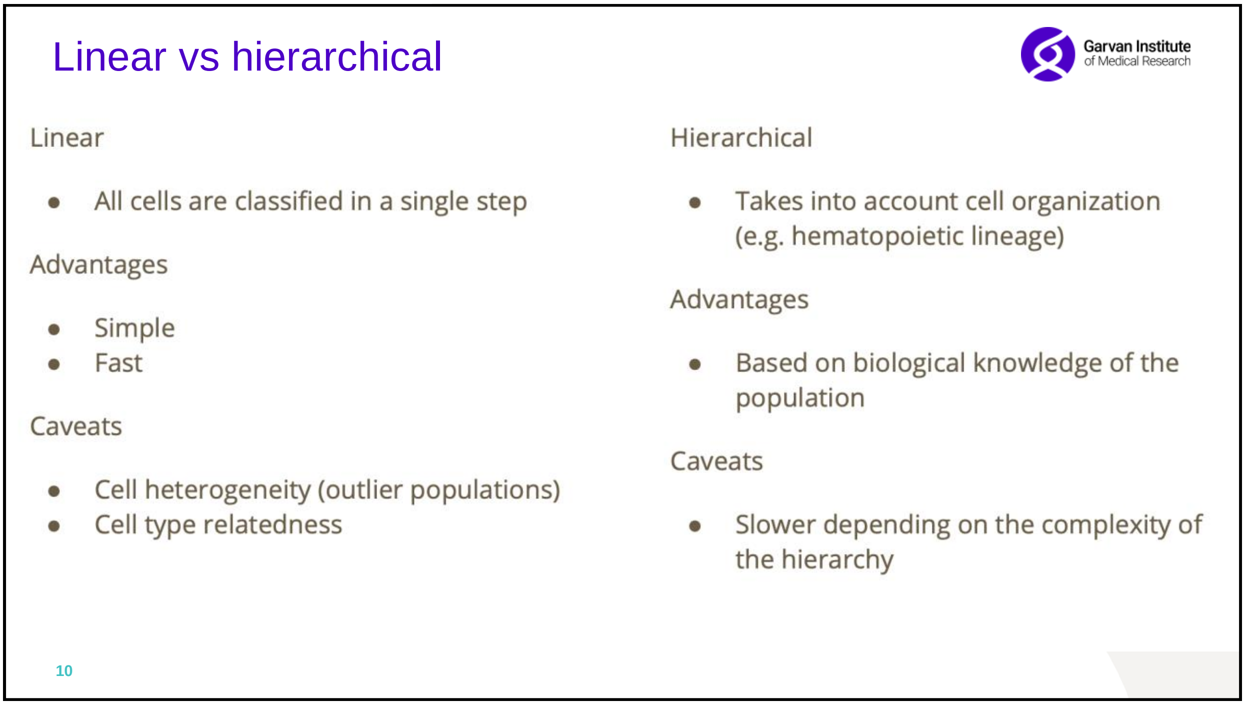
Advantages

- More information is used to classify a cell type (coexpression)

Caveats

- Feature selection (HVG, DEGs, classic markers)

9



Linear vs hierarchical

Garvan Institute of Medical Research

Linear

- All cells are classified in a single step

Advantages

- Simple
- Fast

Caveats

- Cell heterogeneity (outlier populations)
- Cell type relatedness

Hierarchical

- Takes into account cell organization (e.g. hematopoietic lineage)

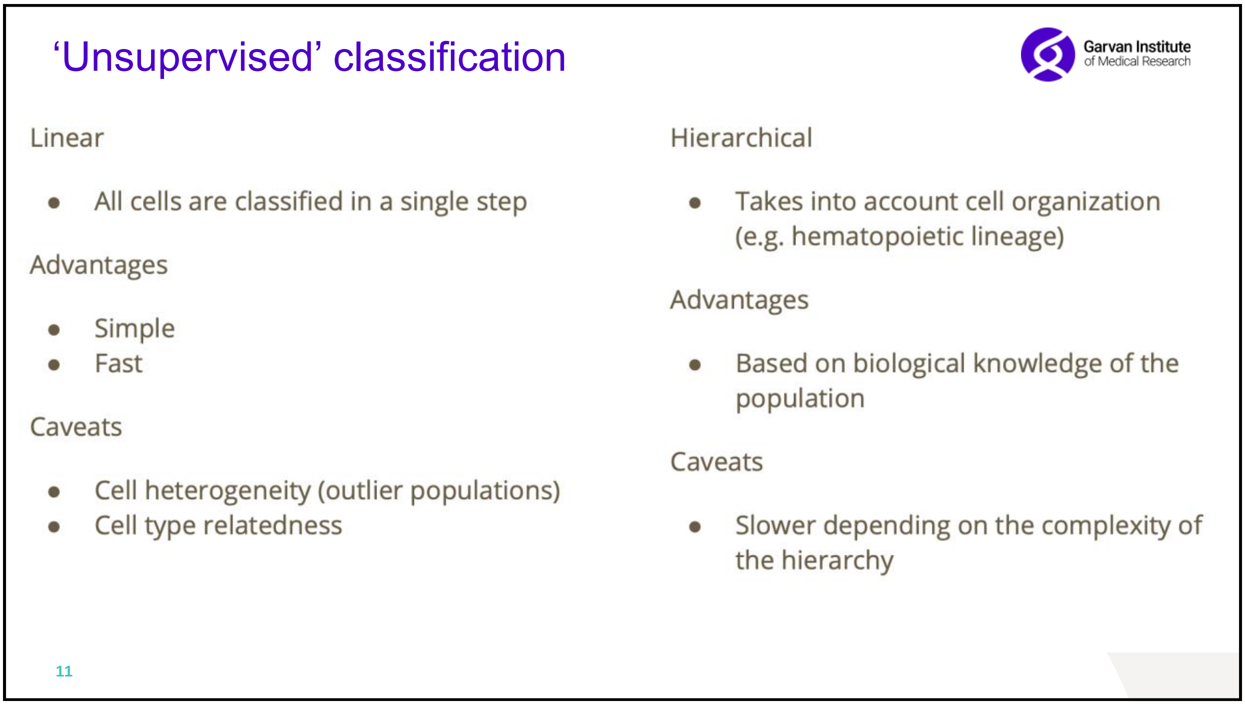
Advantages

- Based on biological knowledge of the population

Caveats

- Slower depending on the complexity of the hierarchy

10



‘Unsupervised’ classification

Linear

- All cells are classified in a single step

Advantages

- Simple
- Fast

Caveats

- Cell heterogeneity (outlier populations)
- Cell type relatedness

Hierarchical

- Takes into account cell organization (e.g. hematopoietic lineage)

Advantages

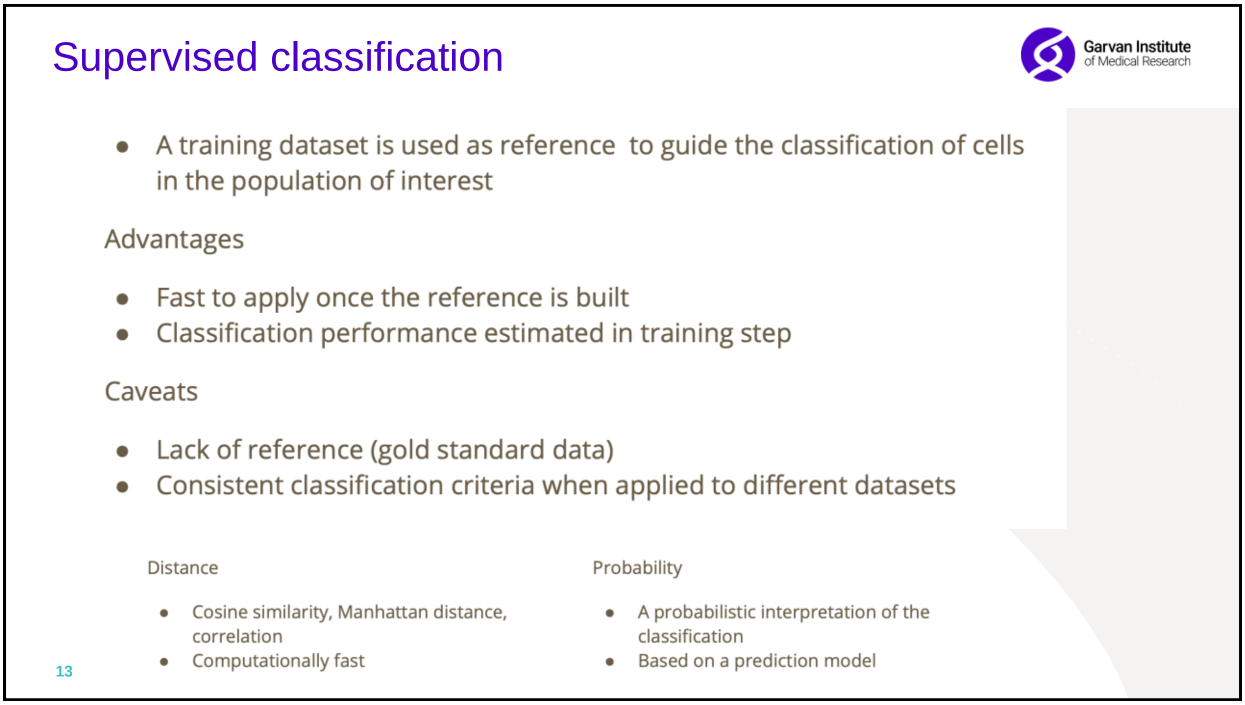
- Based on biological knowledge of the population

Caveats

- Slower depending on the complexity of the hierarchy

11

CLUSTER-HACKING
HACKERMAN
WHEN YOU ARBITRARILY CLUSTER YOUR DATA SO IT MATCHES YOUR EXPECTATIONS
imgflip.com



Supervised classification

- A training dataset is used as reference to guide the classification of cells in the population of interest

Advantages

- Fast to apply once the reference is built
- Classification performance estimated in training step

Caveats

- Lack of reference (gold standard data)
- Consistent classification criteria when applied to different datasets

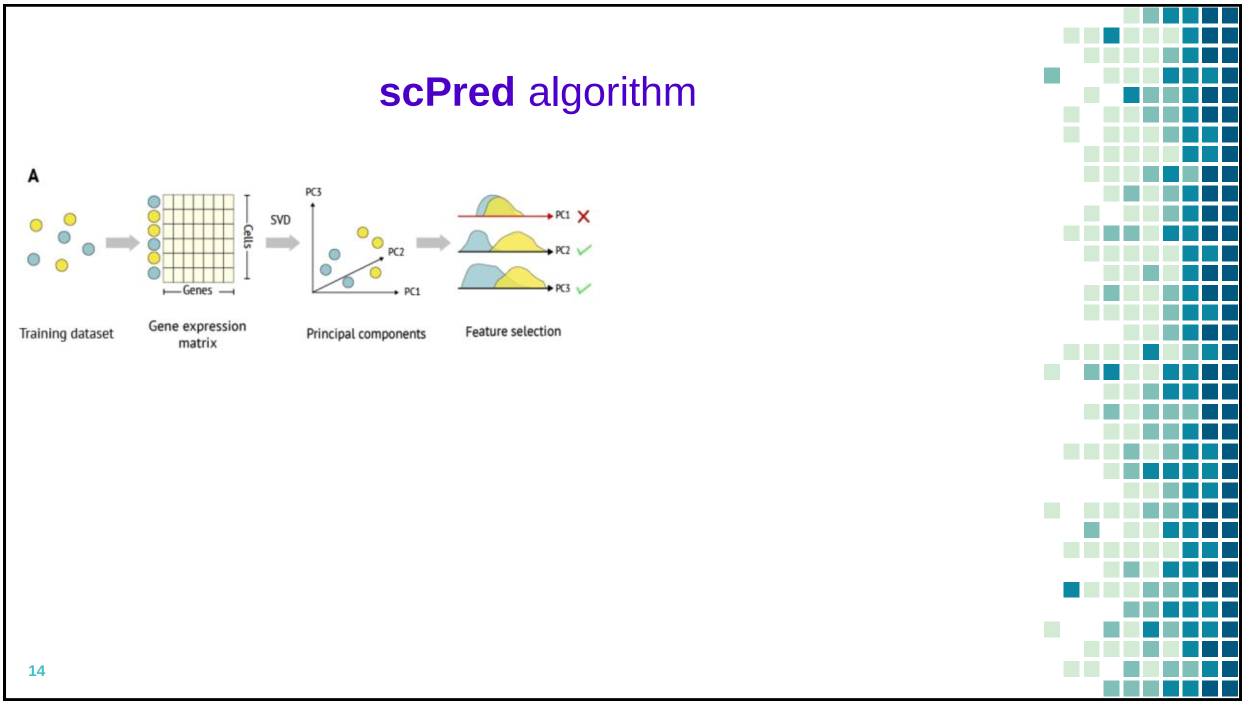
Distance

- Cosine similarity, Manhattan distance, correlation
- Computationally fast

Probability

- A probabilistic interpretation of the classification
- Based on a prediction model

13

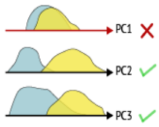


scPred algorithm

14

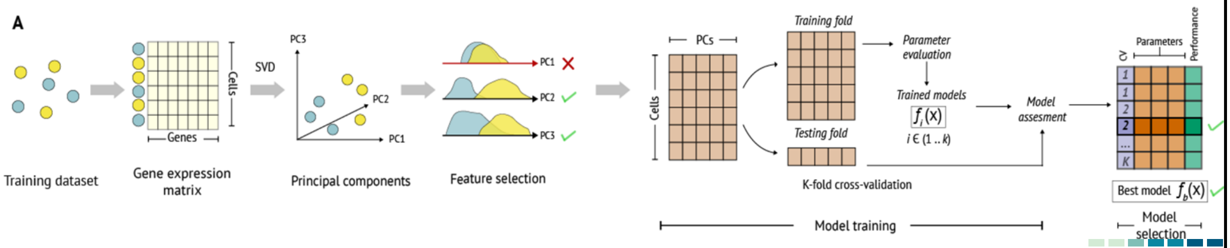
Feature selection

- Create a subspace of S (namely *R* with n rows and r columns (dimensions)), such that each dimension explains at least 0.01% of the variance of the matrix M
- Two-tailed Wilcoxon rank sum test is performed for each principal component to assess whether there is a significant difference in the distributions of principal component scores for cells in different classes



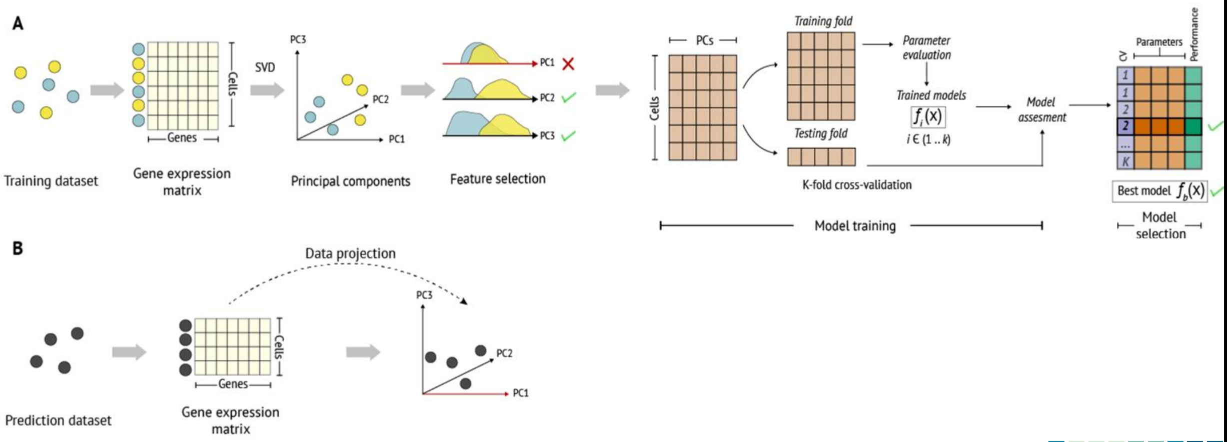
- The resulting p-values are adjusted for multiple testing using a Benjamini-Hochberg false discovery rate correction
- From *R*, we create a subspace *F* with only f columns with associated adjusted p-values less than 0.05.

scPred algorithm

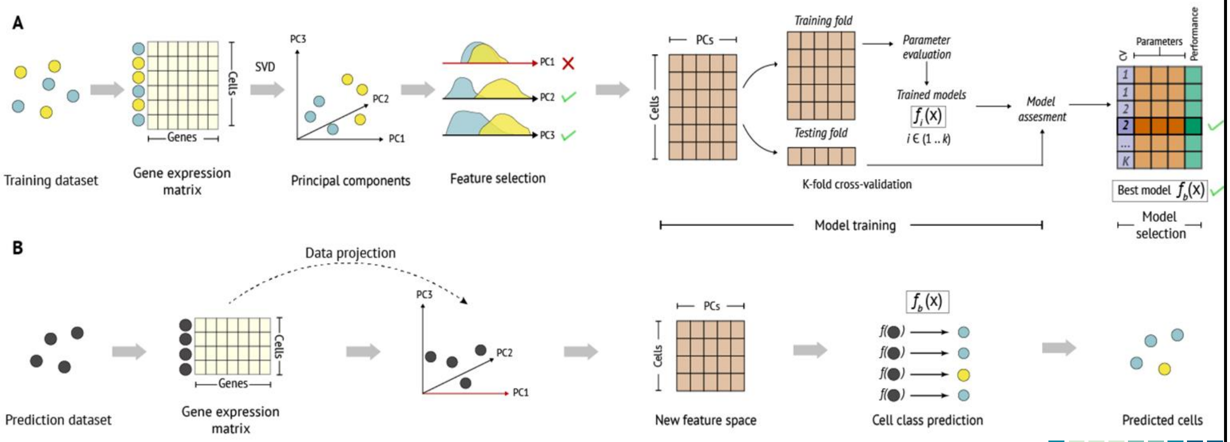


16

scPred algorithm



scPred algorithm



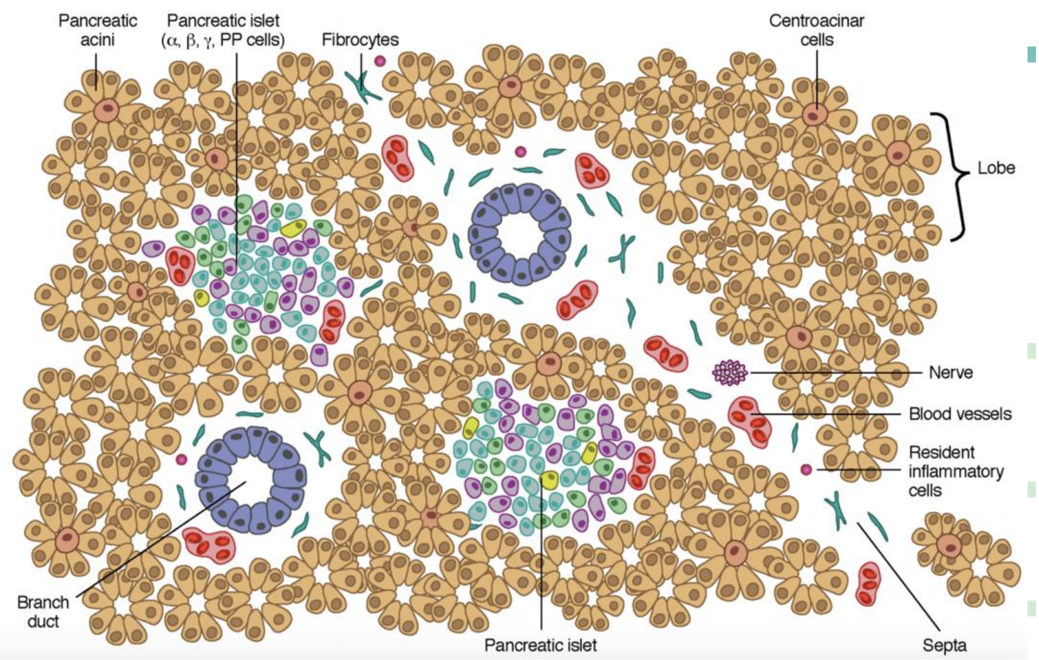
Alquicira *et al.*, Genome Biology *In Press*

1. Classification of Islets of Langerhans subtypes

Classify α (alpha), β (beta), δ (delta) and γ (gamma) cell subtypes

Training dataset

Dataset	Protocol	Number of cells
Muraro *et al.*	CEL-Seq2	2,126
Segerstolpe *et al.*	Smart-Seq2	3,514
Xin *et al.*	SMARTer	1,600



Classification of Islets of Langerhans subtypes

Classify α (alpha), β (beta), δ (delta) and γ (gamma) cell subtypes

Test dataset

Dataset	Protocol	Number of cells
Baron *et al.*	InDrop	4,964

Results

	Training			Test	
Cell type	# cells	# PCs	# Support vectors	# Cells	Accuracy
α Alpha	2584	18	362	2302	98.3
β Beta	1190	17	343	2454	96.1
δ Delta	356	14	283	596	97.1
γ Gamma	383	15	215	254	99.2
Other	0	NA	NA	2326	94.9

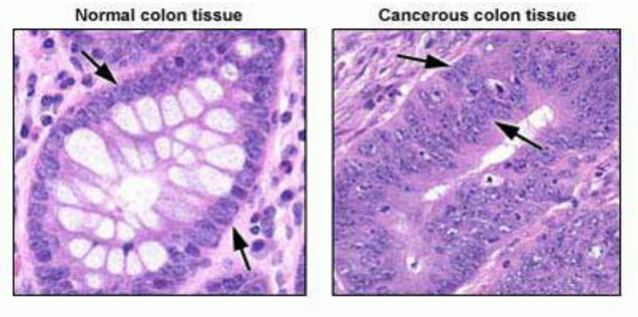
Accurate prediction of cell subtypes

Prediction of cancer cells from human colorectal cancer

Classify cancer cells vs. healthy cells

Dataset

Dataset	Protocol	Number of cells
Li *et al.*	SMARTer/C1	275



Results

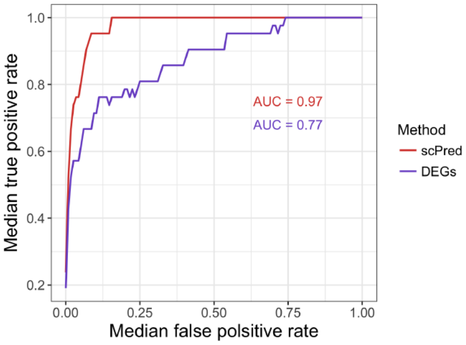


FIGURE 2.3: Prediction performance comparison of colorectal cancer stem/TA-like cells between `scPred` and differentially expressed genes (DEGs). Median values for the area under the ROC curve (AUC) are shown for method.

Software

- Implemented in R
- S4 objects
- *scPred* class
- **scPred** supports any classification model available from the *caret* package