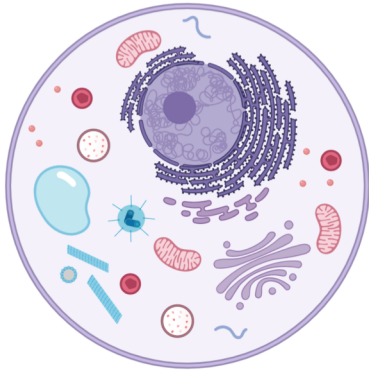
Garvan Institute
of Medical Research



Cell-type classification at single-cell resolution

Jose Alquicira Hernandez

Outline

Garvan Institute of Medical Research

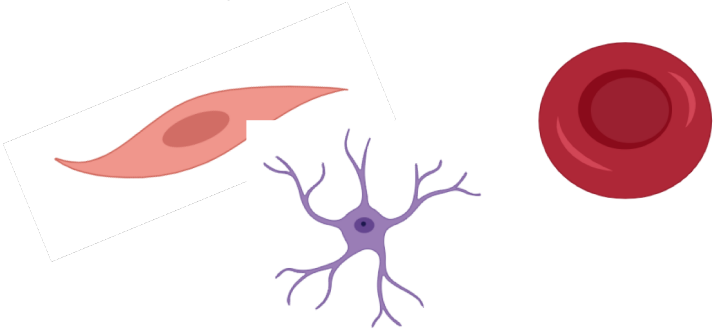
1. What is a cell type?
2. Difference between cell type and cell state
3. What is classification? What is prediction?
4. Types of classification
    a. Uni vs. multivariate
    b. Hierarchical vs. linear
    c. Cluster-based vs. supervised
5. Software
6. Recommendations
7. scPred

Garvan Institute
of Medical Research

What is a cell type?

Cell types can be defined according to different criteria

- Morphology
- Function
- Molecular composition
- End or start point of differentiation
- Determined fate
- *Evolutionary conservation*
- *Phenotype*

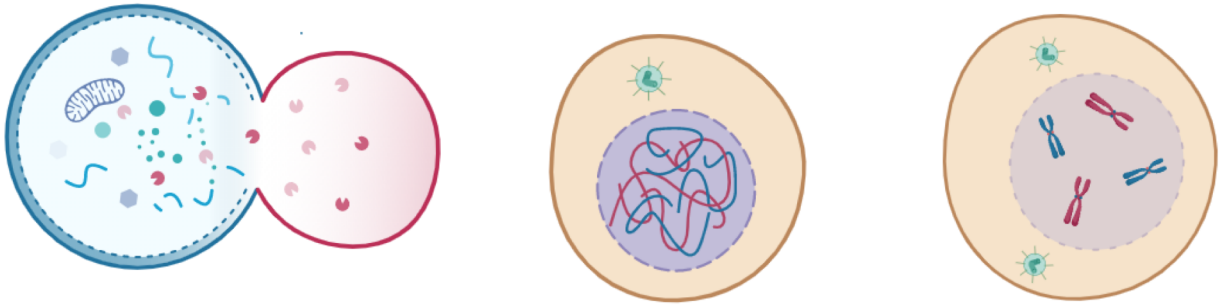


Modified from Grün 2017 - Cell types, clustering, and annotation. HCA meeting

Cell states are different to cell types

Garvan Institute of Medical Research

- Temporary instance of a cell type
- *Two cells can share a cell state and being distinct cell types*
- Cell states are more overlapped than cell types
- No determined fate



Modified from Grün 2017 - Cell types, clustering, and annotation. HCA meeting

Cell types can be detected using various methods

Garvan Institute of Medical Research

Methods

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

...

When are two cells the same?

- Depends on the question
- Context dependent
- Granularity
- Hierarchical definition using established cell ontology

Modified from Grün 2017 - Cell types, clustering, and annotation. HCA meeting

Cell type classification vs. cell type definitio

Garvan Institute of Medical Research

Cell type classification

- Process to label a cell based on a previous definition and a set of features

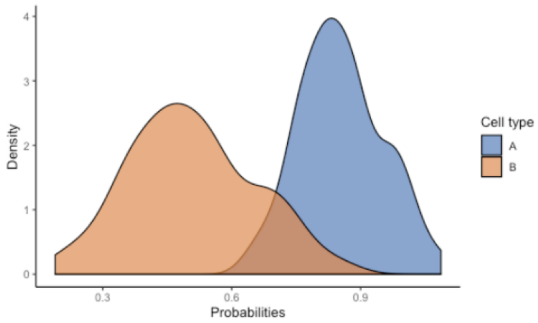
- **Cell type definition**
- Ground truth based on biological and/or statistical criteria
- Cell type discovery

Classification vs. prediction

Garvan Institute of Medical Research

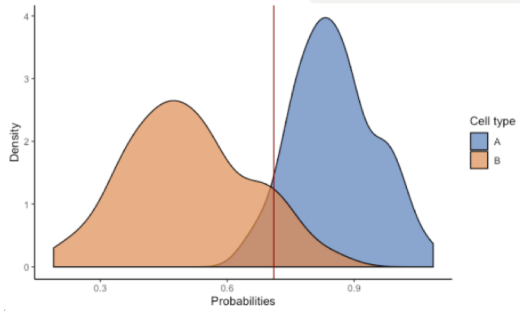
Prediction

- Modeling of tendencies: probabilities



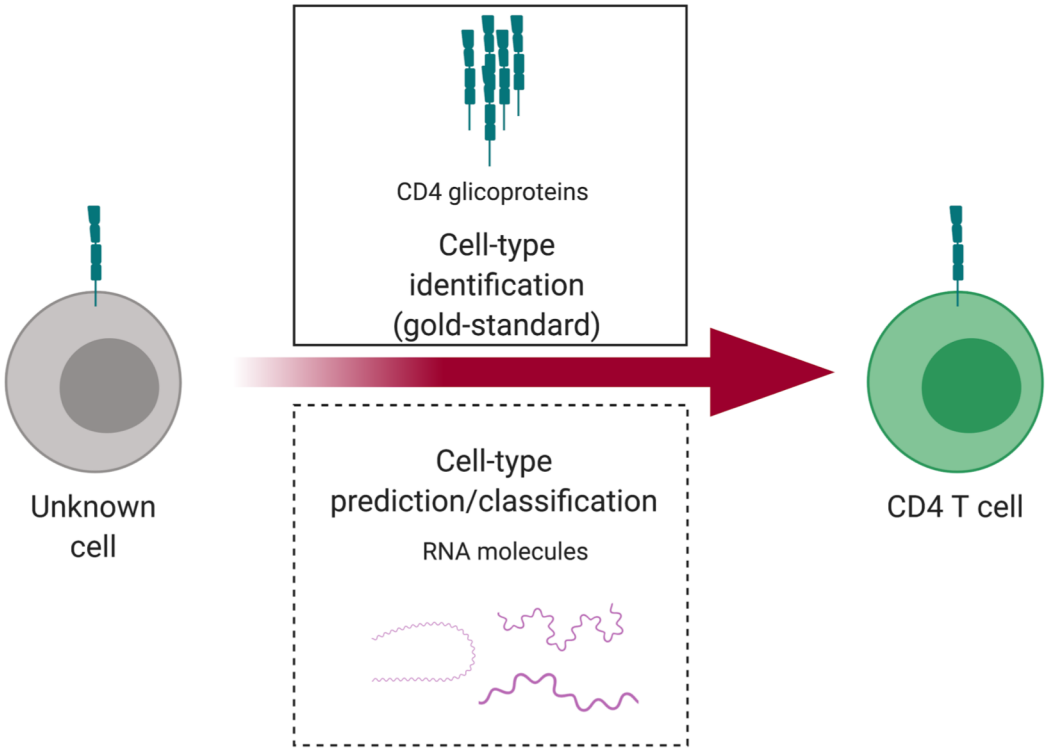
Classification

- Decision making
- Forced choice according to context
- Classification combines prediction and decision making



Frank Harrell: Classification vs. Prediction. Statistical thinking blog (2018-05-19)

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



Garvan Institute
of Medical Research

Cell-type classification approaches

Univariate vs. multivariate

Garvan Institute of Medical Research

Univariate

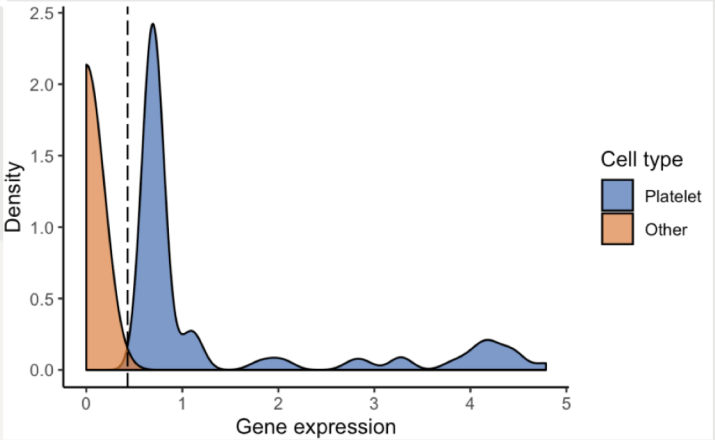
- Some cell types can be classified using a single gene marker (e.g. platelets) or protein

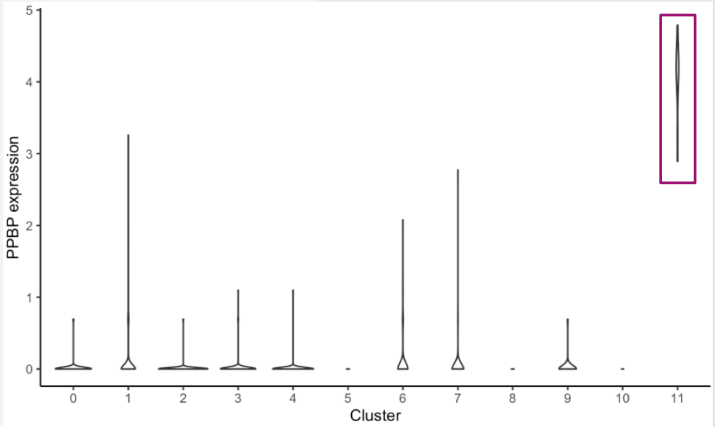
Advantages

- Easier classification
- Clear interpretation

Caveats

- Context dependent (shared expression between cells)
- Lack of correlation between canonical marker and feature
- Expression variance





Univariate vs. multivariate

Garvan Institute of Medical Research

Multivariate

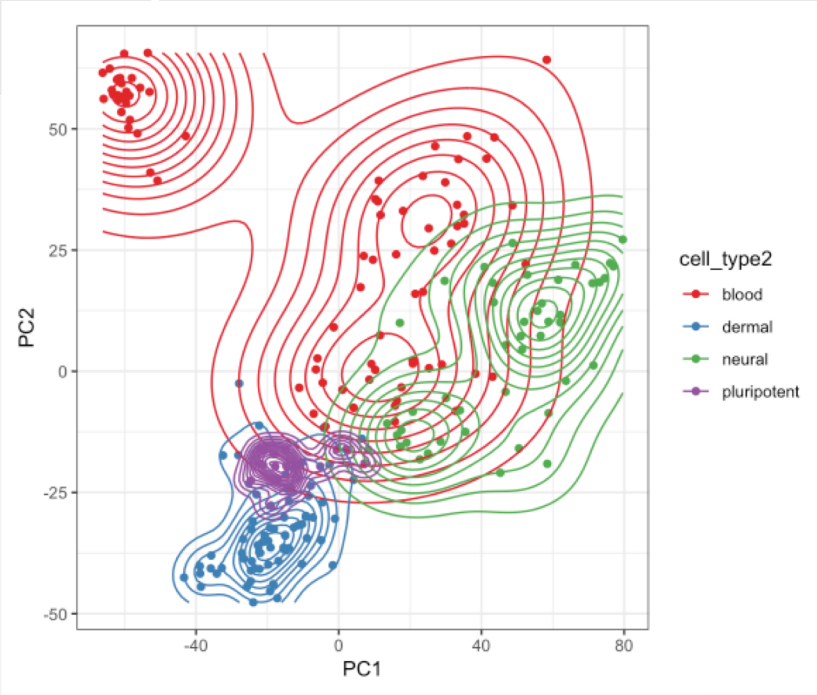
- Combination of features explain cell identity

Advantages

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

Caveats

- Feature selection (DEGs, classic markers)



Flat vs hierarchical

Garvan Institute of Medical Research

Flat

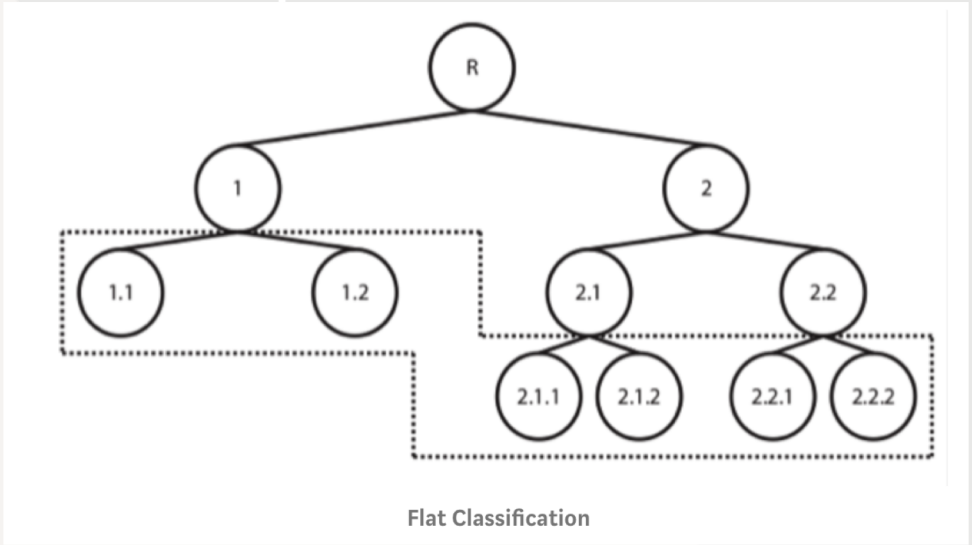
All cells are classified in a single step

Advantages

- Simple
- Fast

Caveats

- Cell heterogeneity (outlier populations)
- Cell type relatedness



From: “A Survey of Hierarchical Classification Across Different Application Domains”

Linear vs hierarchical

Garvan Institute of Medical Research

Hierarchical

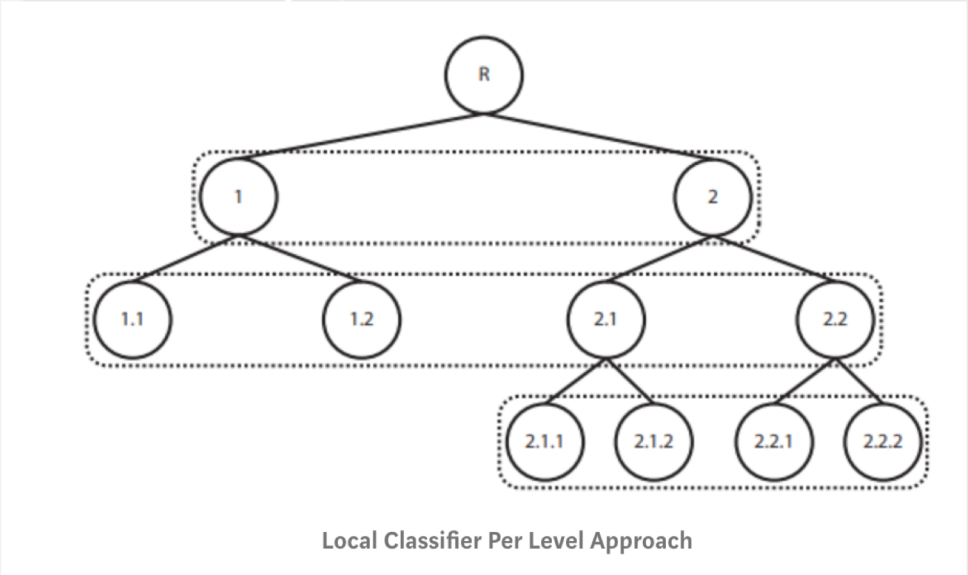
Takes into account cell organization/relatedness (e.g. hematopoietic lineage)

Advantages

- Based on biological knowledge of the population

Caveats

- Slower depending on the complexity of the hierarchy



From: “A Survey of Hierarchical Classification Across Different Application Domains”

Unsupervised classification

Garvan Institute of Medical Research

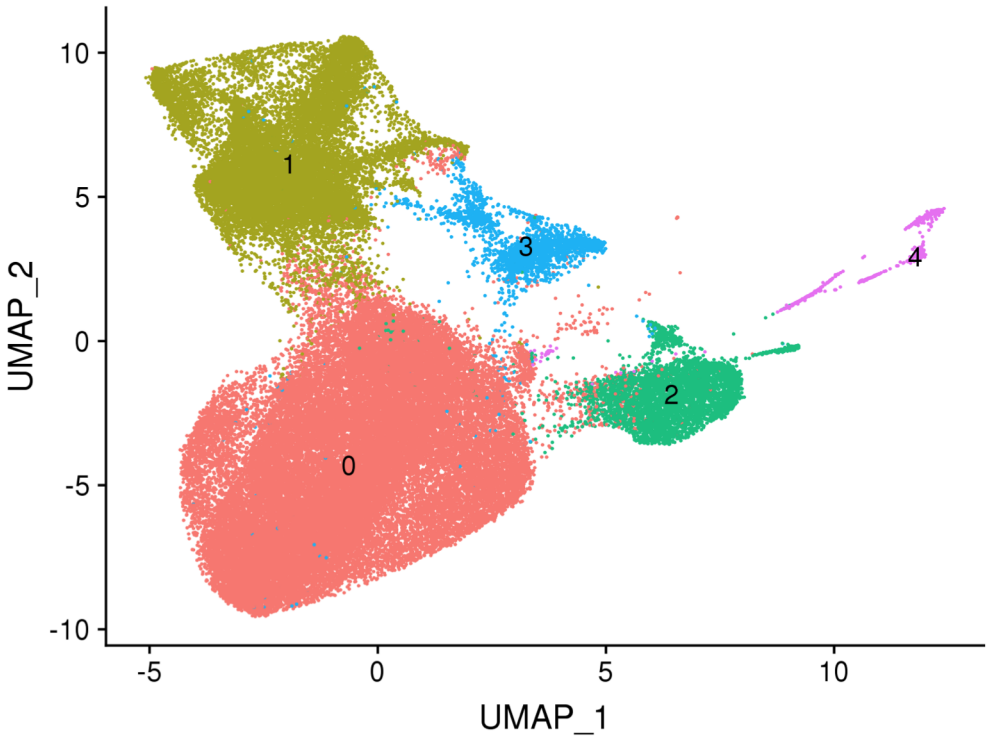
- Clustering
- Clusters are expected to correlate with real cell types or subpopulations
- All cells within a cluster are assumed to be the same cell type

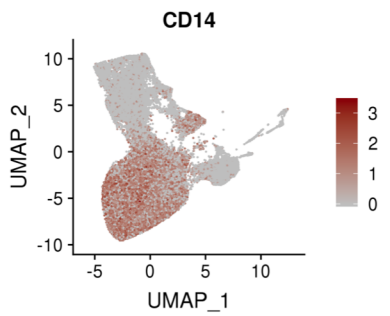
Advantages

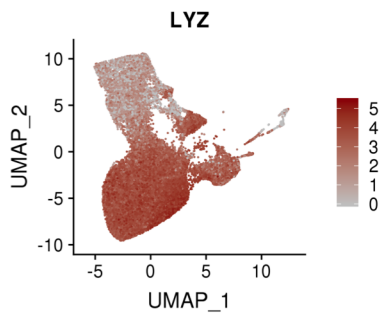
- No reference is needed
- Classic markers can be used to guide the annotation
- Cell type discovery / definition

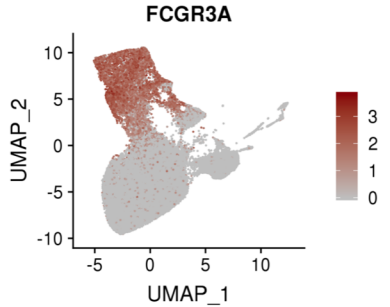
Caveats

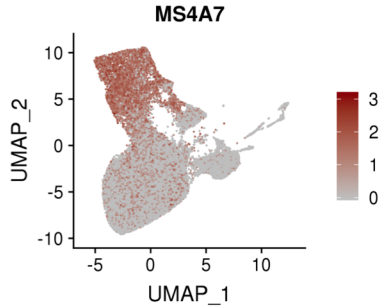
- Ambiguous results depending on resolution or hierarchy level
- Feature selection


UMAP_2
UMAP_1
10
5
0
-5
-10
-5
0
5
10
0
1
2
3
4


CD14
UMAP_2
UMAP_1
10
5
0
-5
-10
-5
0
5
10
3
2
1
0


LYZ
UMAP_2
UMAP_1
10
5
0
-5
-10
-5
0
5
10
5
4
3
2
1
0


FCGR3A
UMAP_2
UMAP_1
10
5
0
-5
-10
-5
0
5
10
3
2
1
0


MS4A7
UMAP_2
UMAP_1
10
5
0
-5
-10
-5
0
5
10
3
2
1
0

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
- Consistent classification criteria applied to different datasets

Caveats

- Lack of reference (gold standard data)
- Completeness of reference

Supervised classification

Garvan Institute of Medical Research

A comparison of automatic cell identification methods for single-cell RNA-sequencing data

Tamim Abdelaal[1,2#] (t.r.m.abdelaal-1@tudelft.nl)
Lieke Michielsen[1,2#] (l.c.m.michielsen@student.tudelft.nl)
Davy Cats[3] (d.cats@lumc.nl)
Dylan Hoogduin[3] (ddhoogduin@gmail.com)
Hailiang Mei[3] (H.Mei@lumc.nl)
Marcel J.T. Reinders[1,2] (m.j.t.reinders@tudelft.nl)
Ahmed Mahfouz[1,2*] (a.mahfouz@lumc.nl)

[1] Leiden Computational Biology Center, Leiden University Medical Center, Einthovenweg 20, 2333ZC, Leiden, The Netherlands
[2] Delft Bioinformatics Lab, Delft University of Technology, Van Mourik Broekmanweg 6, 2628XE, Delft, The Netherlands
[3] Sequencing Analysis Support Core, Department of Biomedical Data Sciences, Einthovenweg 20, 2333ZC, Leiden University Medical Center, Leiden, The Netherlands

[#] Equal contribution
* Corresponding author (a.mahfouz@lumc.nl)

Supervised classification

Garvan Institute of Medical Research

- [Cell type classification] is typically solved by **unsupervised clustering** of cells into groups based on the similarity of their gene expression profiles, followed by cell population annotation by assigning labels to each cluster

- The [cell type] annotation step is cumbersome and **time-consuming** as it involves **manual inspection** of cluster-specific marker genes

- **Manual annotations**, which are often not based on standardized ontologies of cell labels, **are not reproducible** across different experiments within and across research groups

Supervised classification

- In this study, we evaluated the performance of 20 different methods for automatic cell identification using eight scRNA-seq datasets. Several classifiers accurately performed on almost all datasets, particularly: **scPred, SVM**, **scmapcell/cluster, singleCellNet**, **scVI**, **LDA** and **ACTINN**.

scPred: Cell type prediction at single-cell resolution

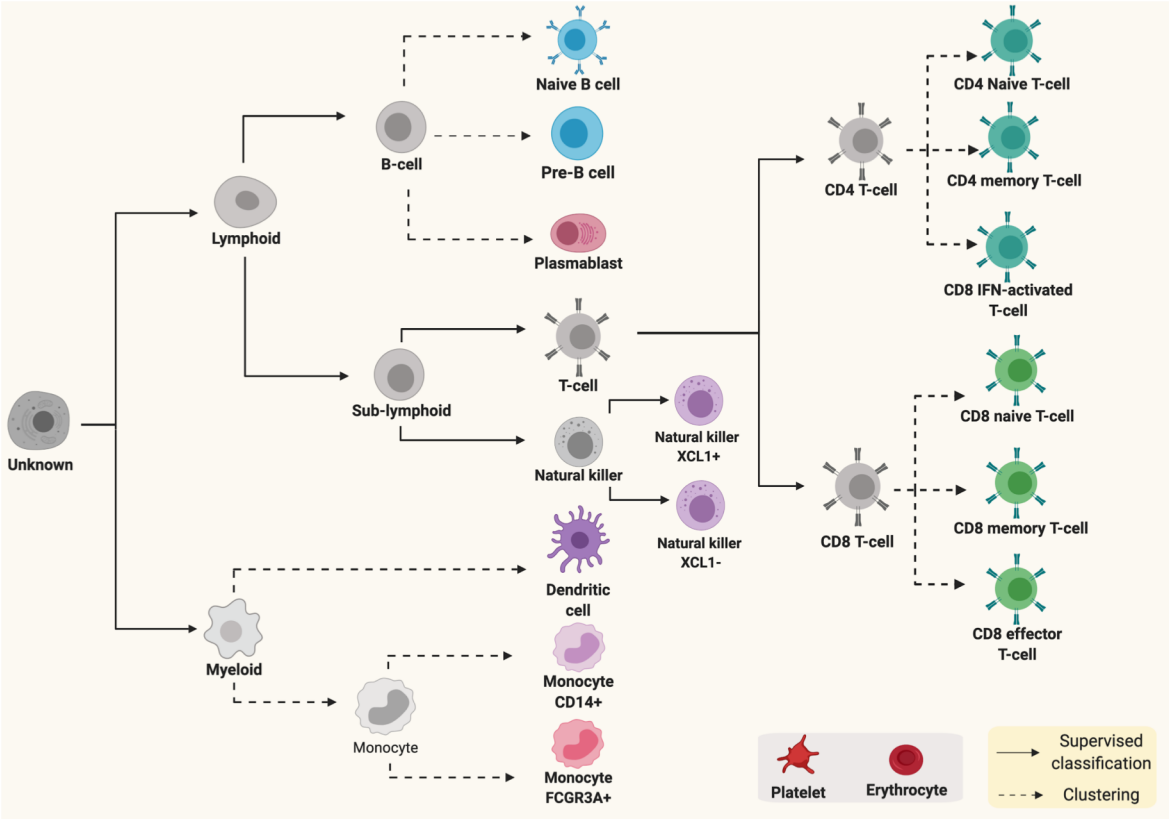
José Alquicira-Hernández[1,2], Anuja Sathe[3,4], Hanlee P Ji[3,4], Quan Nguyen[1], and Joseph E Powell[1,2,5]

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane
[2]Garvan Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Darlinghurst, Sydney
[3]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, United States
[4]Stanford Genome Technology Center, Stanford University, Palo Alto, United States
[5]Faculty of Medicine, University of New South Wales, Darlinghurst, Sydney

Under review

Hierarchical classification of PBMCs using supervised and unsupervised approaches

Garvan Institute of Medical Research



Recommendations

- **Biology, biology, and biology!**
- Know your population of study
- How many cell types do you expect?
- Expected cell proportions by cell type
- Expected cell proportions by individual
- Outlier population or cell contamination? E.g. erythrocytes, platelets
- If a cell hierarchy is known, use it to guide the classification
- Identify factors that can confound the prediction
  - Sequencing depth
  - Pool/batch
- Pay attention to feature selection

...

Recommendations

- Avoid *cluster-hacking*:

“**Exhaustive** clustering search so your data matches your hypothesis”

Acknowledgements

- Quan Nguyen
- Anuja Sathe
- Hanlee P Ji
- Joseph Powell

Single Cell and Computational Genomics

Garvan Institute of Medical Research

- Anne Senabouth
- Seyhan Yazar
- Drew Neavin
- Venessa Chin
- Angela Murphy

Institute of Molecular Bioscience

University of Queensland