From biology to statistics, and back

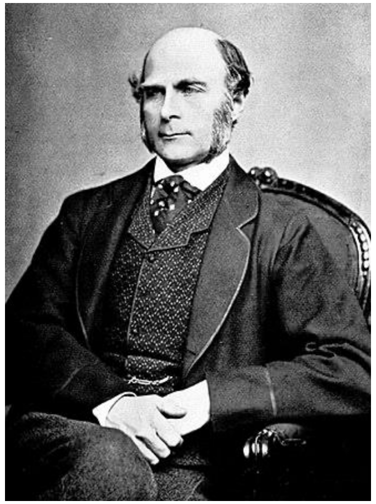
Dr Jiadong Mao

Melbourne Integrative Genomics

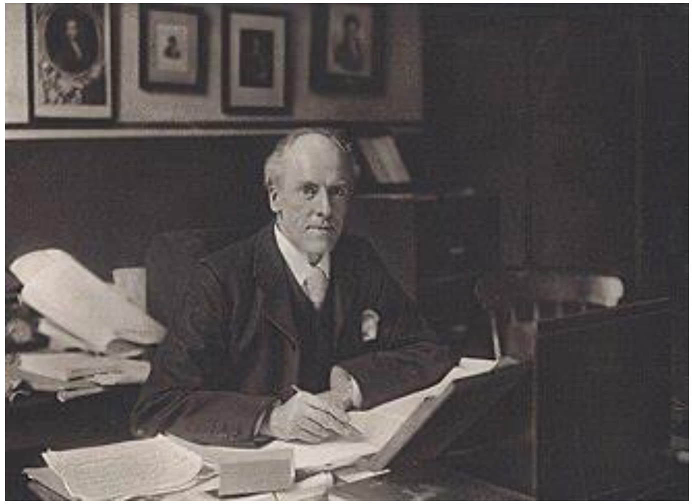
School of Mathematics and Statistics

University of Melbourne

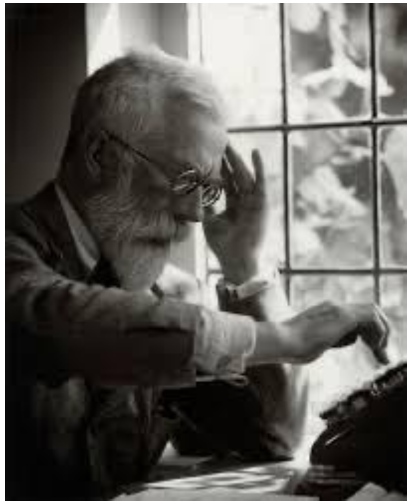
Statistics and biology



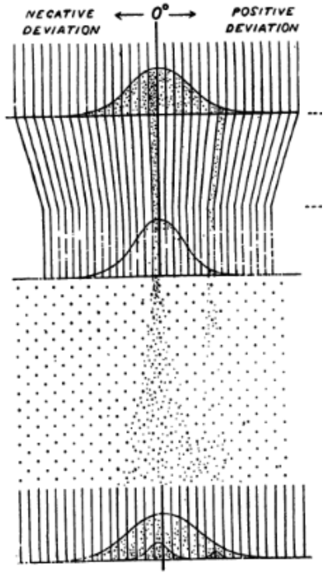
Francis Galton (1822–1911)



Karl Pearson (1857–1936)



Ronald A Fisher (1890–1962)



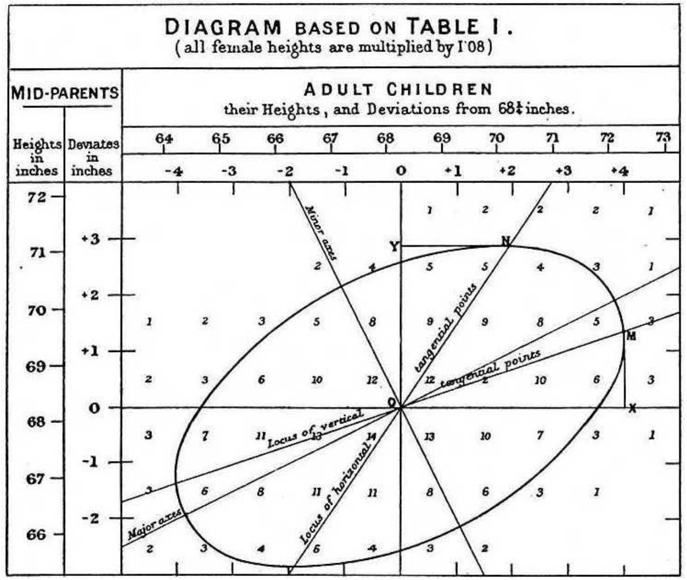
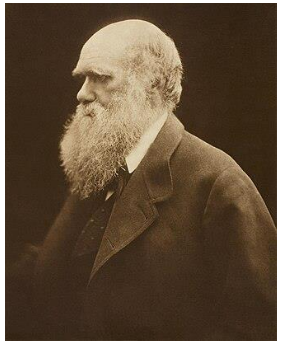


Fig credit: Wikipedia

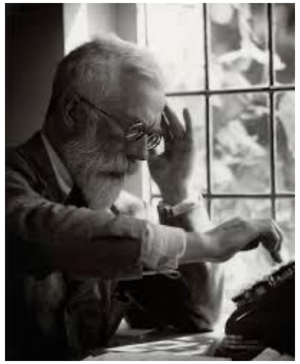
Heredity: the hidden theme of early statistics

Gregor Mendel (1822–1884)



Charles Darwin (1809–1882)

‘Modern synthesis’



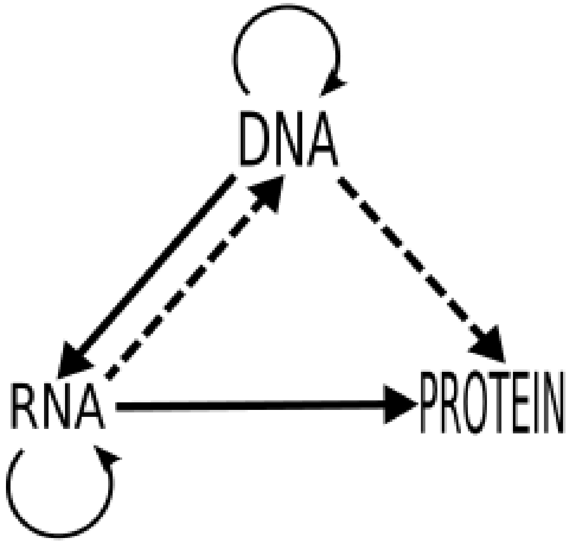
Ronald A Fisher (1890–1962)

(Fisher is) the greatest of Darwin’s successors.
-- Richard Dawkins, *The Blind Watch Maker*

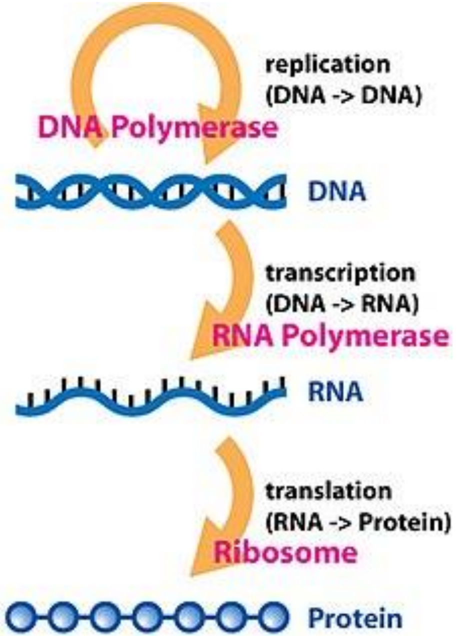
Fig credit: Wikipedia

‘Central dogma’ & omics data

- Molecular biology of the cell
- ‘Omics’ data & high-throughput sequencing
- Types of omics
  - Genomics
  - **Transcriptomics**
  - Proteomics
  - Metabolomics, epigenomics, …
- What’s so special about RNAs (transcripts/gene expression)
  - Cell activities and identities: T cell, B cell, …



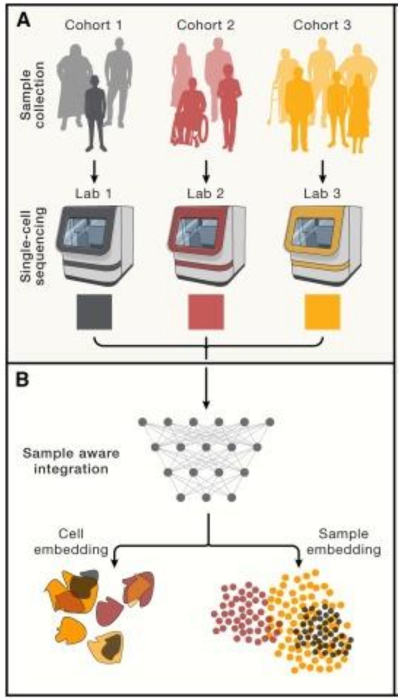
Crick’s central dogma (Wikipedia)



Watson’s central dogma (Wikipedia)

Omics and modern statistics

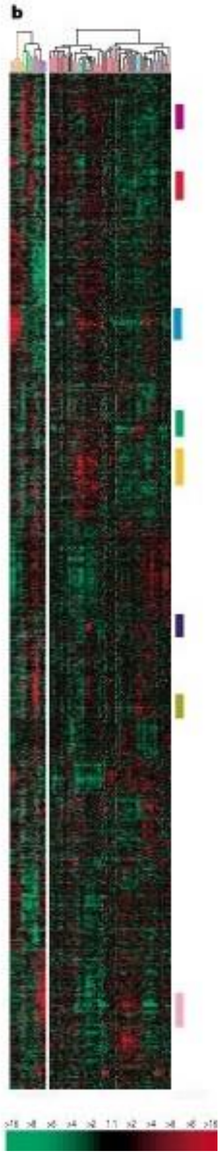
- Bulk RNA sequencing, eg microarray
- HDLSS: high dimension, low sample size
  - ~50 samples, >10,000 genes
- Variable selection, multiple testing
- Common goal: marker gene identification
  - Diagnosis, treatment, prognosis, eg cancer subtyping



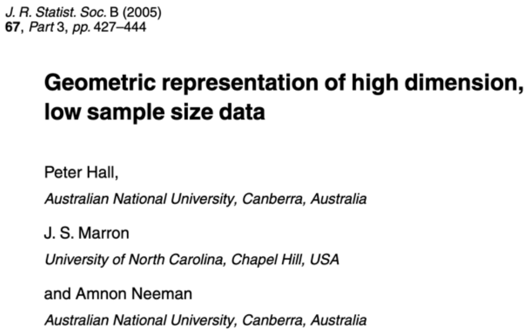
Lotfollahi et al. (2024). *Cell*.

samples

genes



Perou et al. (2000). *Nature*.



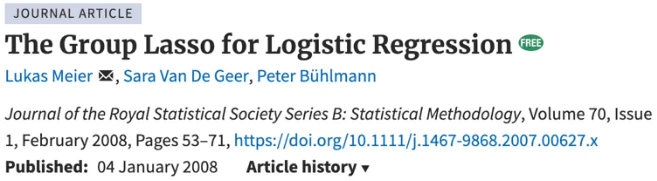
J. R. Statist. Soc. B (2005)
67, *Part* 3, *pp.* 427–444

Geometric representation of high dimension, low sample size data

Peter Hall,
Australian National University, Canberra, Australia

J. S. Marron
University of North Carolina, Chapel Hill, USA

and Amnon Neeman
Australian National University, Canberra, Australia



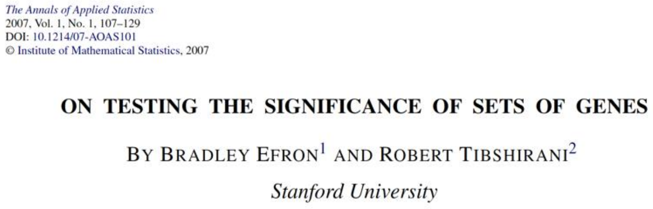
JOURNAL ARTICLE

The Group Lasso for Logistic Regression FREE

Lukas Meier ✉, Sara Van De Geer, Peter Bühlmann

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 70, Issue 1, February 2008, Pages 53–71, https://doi.org/10.1111/j.1467-9868.2007.00627.x

Published: 04 January 2008 **Article history** ▾



The Annals of Applied Statistics
2007, Vol. 1, No. 1, 107–129
DOI: 10.1214/07-AOAS101
© Institute of Mathematical Statistics, 2007

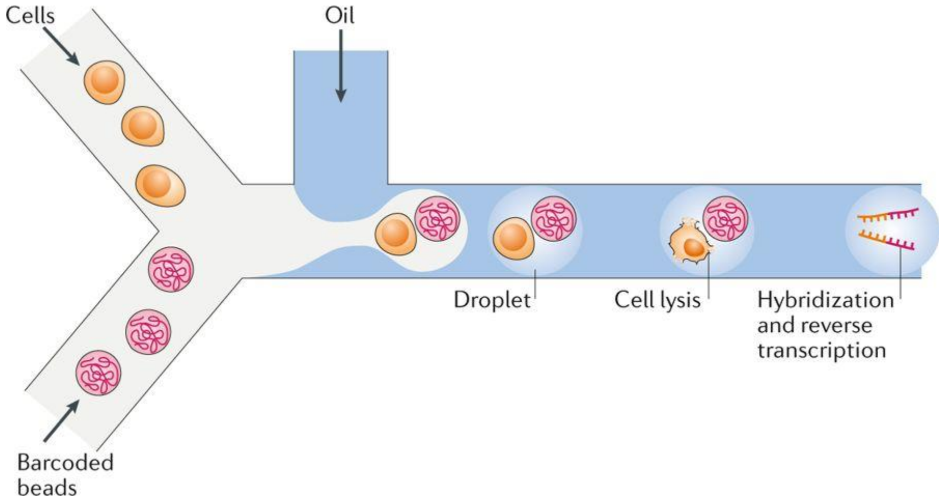
ON TESTING THE SIGNIFICANCE OF SETS OF GENES

BY BRADLEY EFRON[1] AND ROBERT TIBSHIRANI[2]

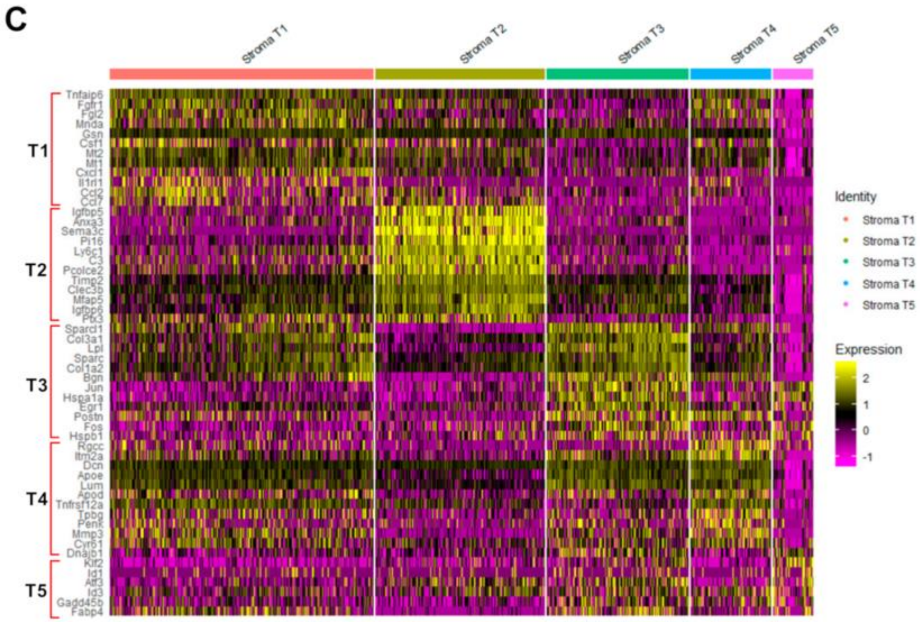
Stanford University

New challenges: big omics data

- Single-cell RNA sequencing (scRNA-seq)
  - Dissolve tissue into single cells & seq
- ‘HDHSS’: High dimension, high sample size
  - >10,000 cells per sample/donor
  - >20,000 genes
- Finding marker genes at cell (type) level



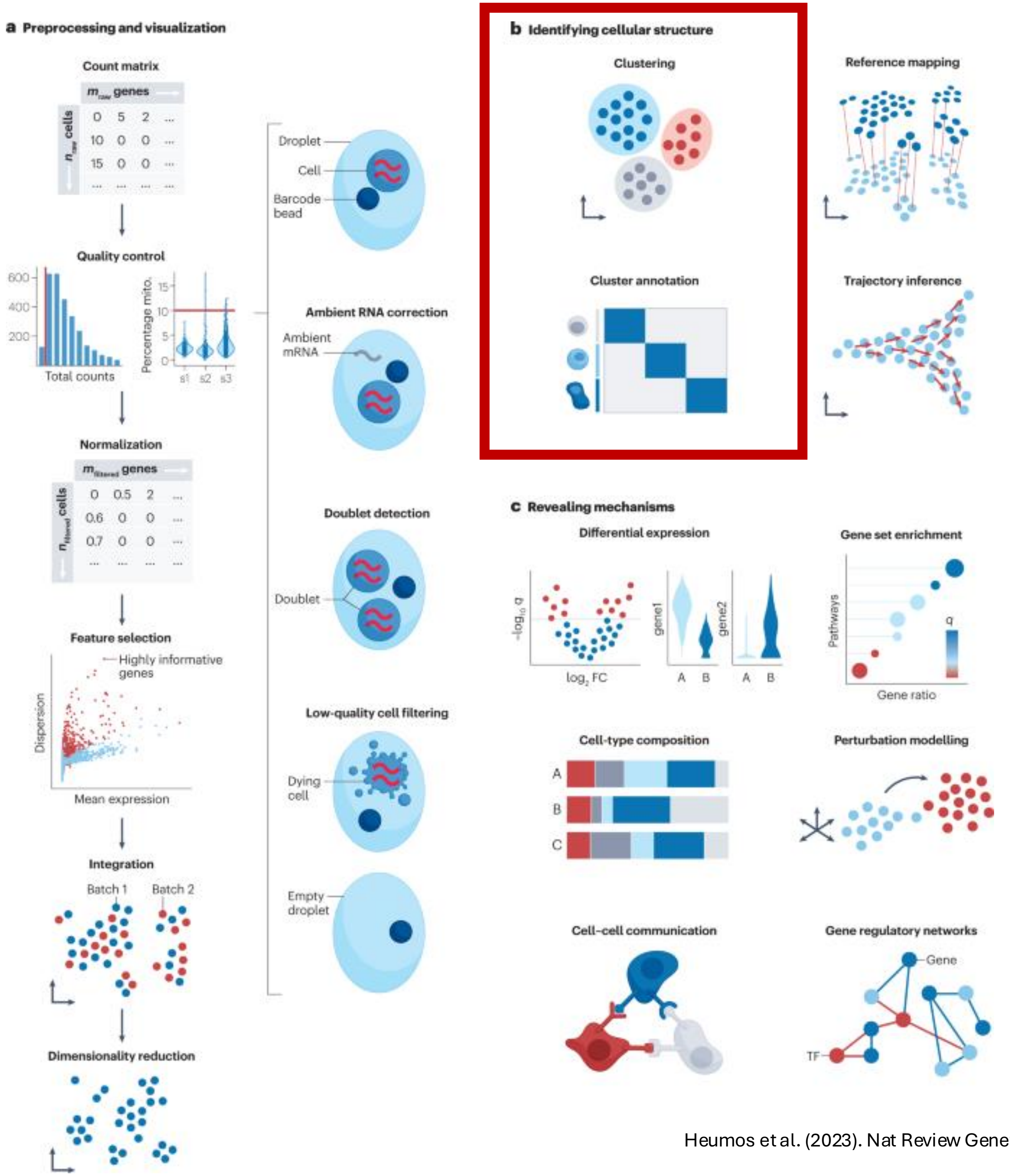
https://www.rna-seqblog.com/wp-content/uploads/2018/08/droplet.png



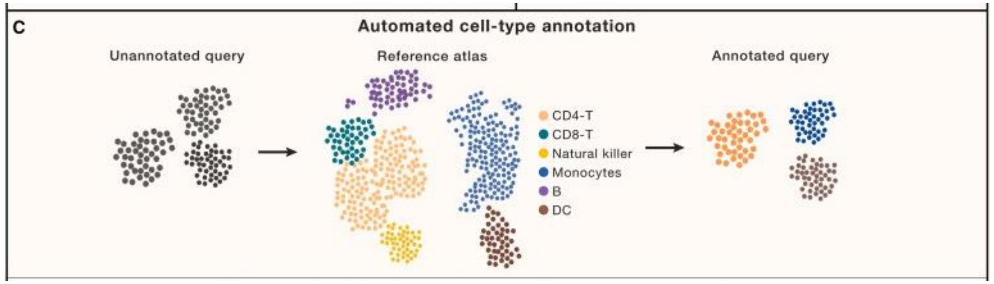
Lo et al. (2020). *Cancers*.

Cell type annotation

- Cells form (relatively) homogeneous groups
- Group cells into cell types: train cell type classifier on reference data



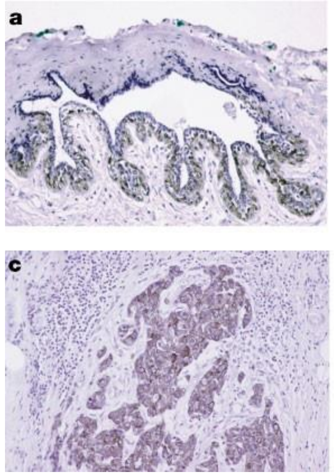
Heumos et al. (2023). Nat Review Genetics.



Lotfollahi et al. (2024). *Cell*.

Adding spatial information

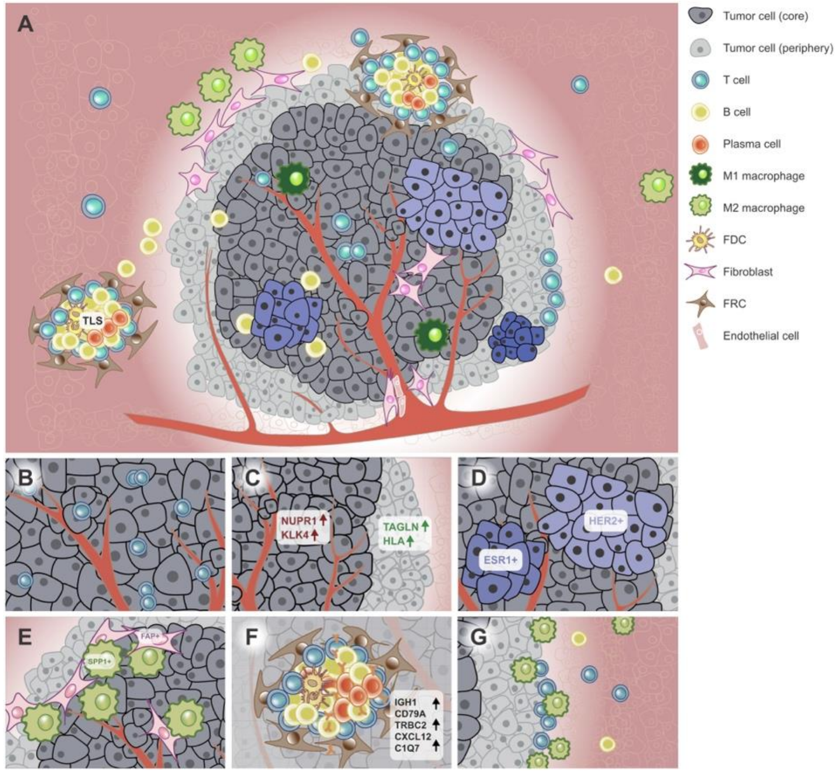
- Spatial transcriptomics (ST)
- Why ST
  - High-throughput: measuring a lot of molecules
  - scRNA-seq: which cells are doing what
  - ST: which cells are doing what, and **where they are doing it**
  - Example: tumour microenvironment
- Cell type annotation is still the key

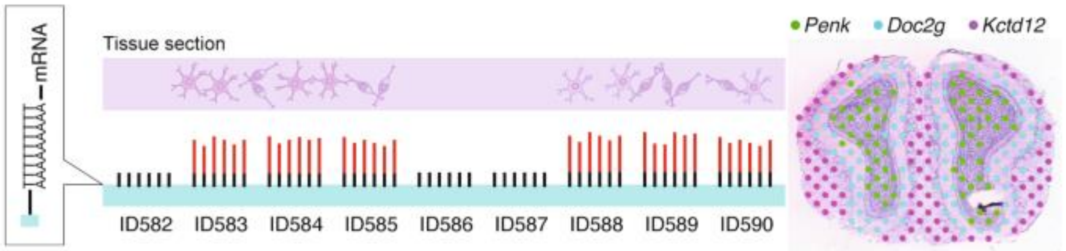


normal

tumour

Perou et al. (2000). *Nature*.

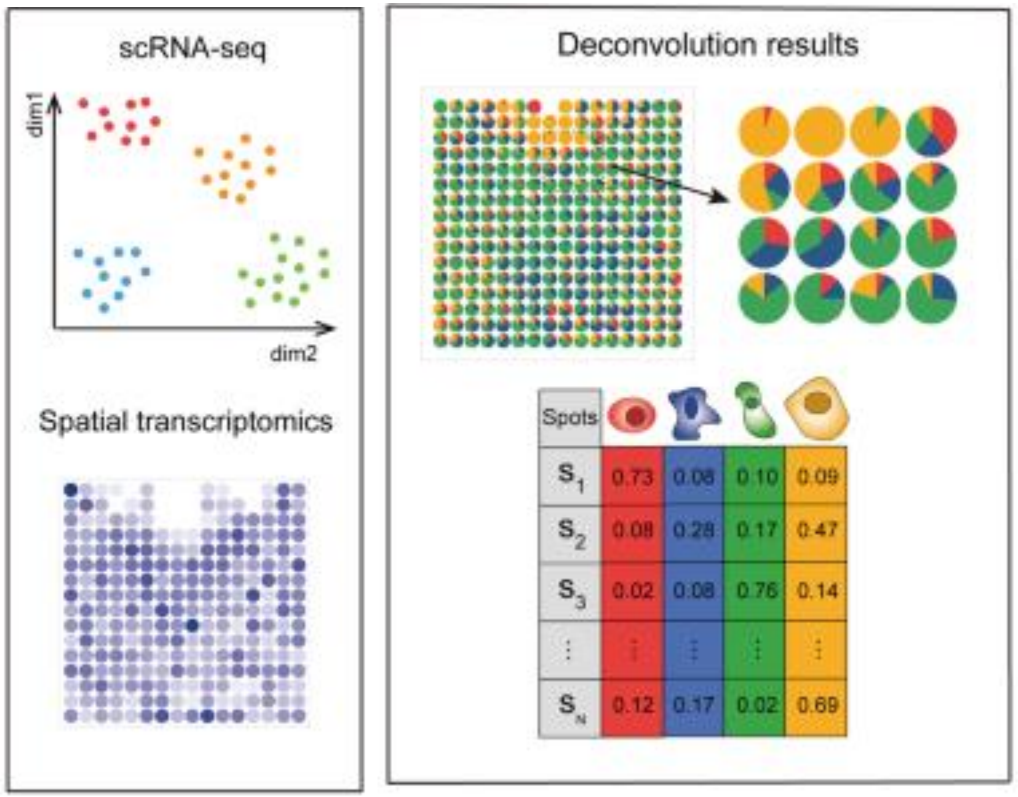




Nature Methods Method of the year 2021

Cell type deconvolution

- Main idea
  - Each ‘spot’ may contain multiple cells
  - Use scRNA-seq as ‘reference’
  - Decompose gene expression in each spot as combination of reference cell types

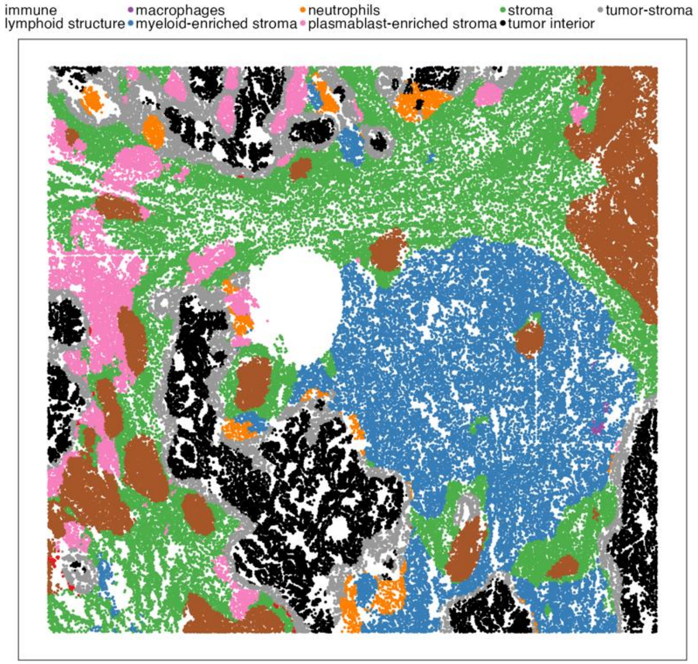


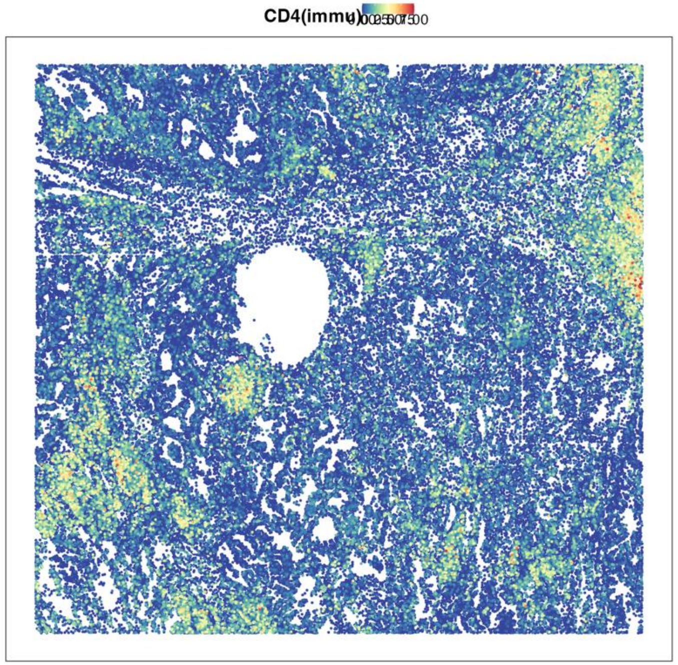
Spots				
S_1	0.73	0.08	0.10	0.09
S_2	0.08	0.28	0.17	0.47
S_3	0.02	0.08	0.76	0.14
⋮	⋮	⋮	⋮	⋮
S_N	0.12	0.17	0.02	0.69

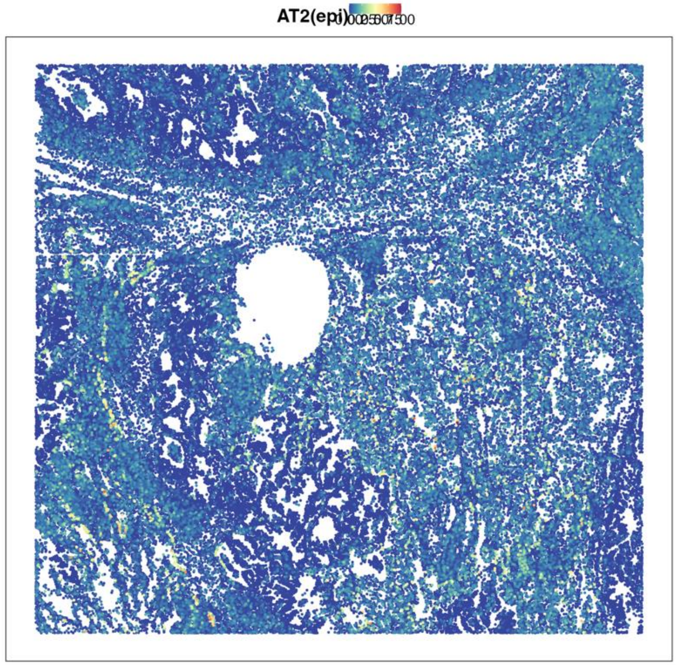
Li et al. (2023). Nat Comms

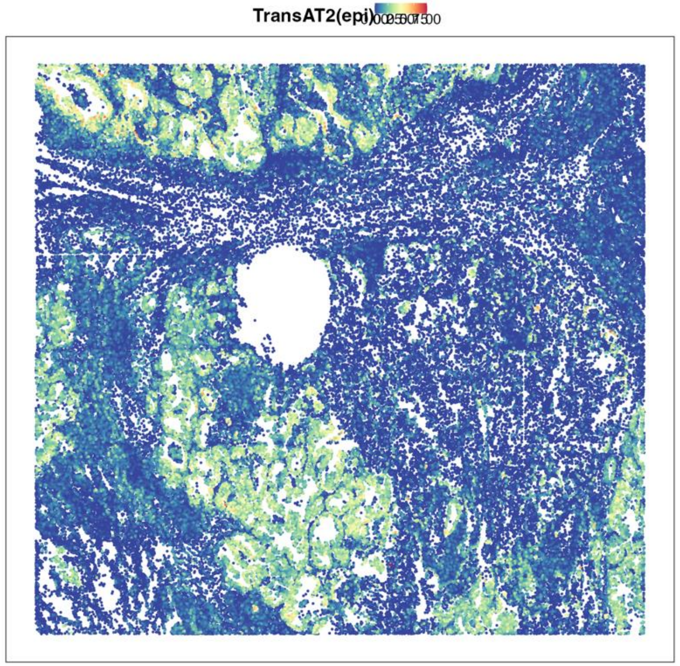
Deconvolute cancer cell states

- *Platform*: Nanostring CosMx
- *Sample:* human lung cancer (non-small cell lung cancer)
- *References*: scRNA-seq from lung fibrosis patients





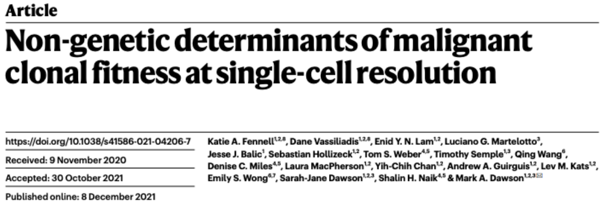




ST with cancer cell lineage tracing

- *Platform*: BGI Stereo-seq
- *Sample info*
  - mouse spleen sample with AML cells
  - SPLINTR lineage tracing, same AML **clone** same **barcode**
- *References*: scRNA-seq from mouse spleen and bone marrow

Henrietta Holze


Article

Non-genetic determinants of malignant clonal fitness at single-cell resolution

https://doi.org/10.1038/s41586-021-04206-7

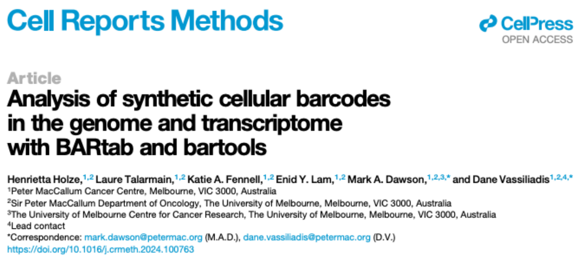
Received: 9 November 2020

Accepted: 30 October 2021

Published online: 8 December 2021

Katie A. Fennell[1,2,8], Dane Vassiliadis[1,2,8], Enid Y. N. Lam[1,2], Luciano G. Martelotto[3], Jesse J. Balic[1], Sebastian Hollizeck[1,2], Tom S. Weber[4,5], Timothy Semple[1,3], Qing Wang[6], Denise C. Miles[4,5], Laura MacPherson[1,2], Yih-Chih Chan[1,2], Andrew A. Guirguis[1,2], Lev M. Kats[1,2], Emily S. Wong[6,7], Sarah-Jane Dawson[1,2,3], Shalin H. Naik[4,5] & Mark A. Dawson[1,2,3]✉

Dane Vassiliadis


Cell Reports Methods

CellPress
OPEN ACCESS

Article

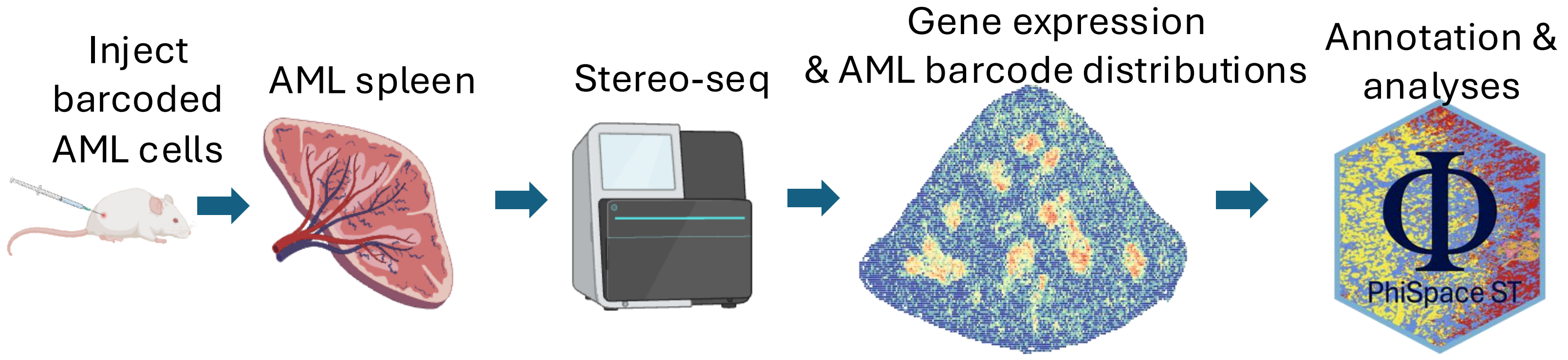
Analysis of synthetic cellular barcodes in the genome and transcriptome with BARtab and bartools

Henrietta Holze,[1,2] Laure Talarmain,[1,2] Katie A. Fennell,[1,2] Enid Y. Lam,[1,2] Mark A. Dawson,[1,2,3,*] and Dane Vassiliadis[1,2,4,*]

[1]Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia
[2]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC 3000, Australia
[3]The University of Melbourne Centre for Cancer Research, The University of Melbourne, Melbourne, VIC 3000, Australia
[4]Lead contact
*Correspondence: mark.dawson@petermac.org (M.A.D.), dane.vassiliadis@petermac.org (D.V.)
https://doi.org/10.1016/j.crmeth.2024.100763

Mark Dawson

Peter Mac
Peter MacCallum Cancer Centre
Victoria Australia

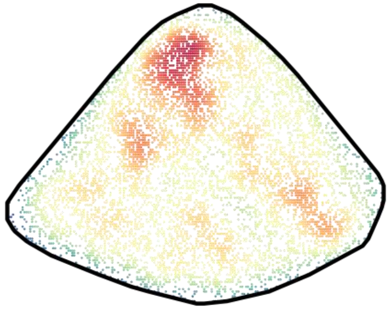

Inject barcoded AML cells → AML spleen → Stereo-seq → Gene expression & AML barcode distributions → Annotation & analyses

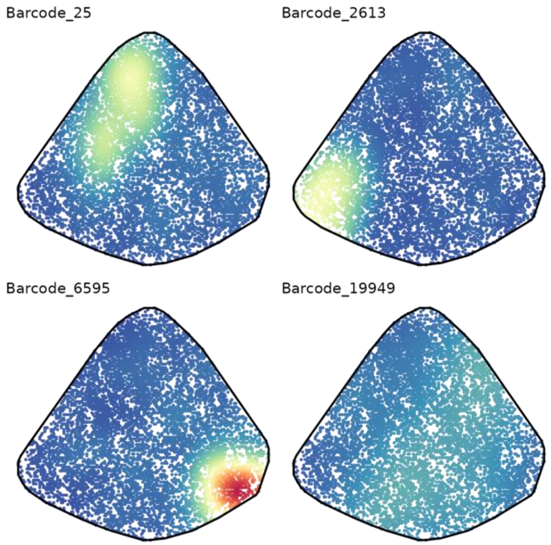
AML: acute myeloid leukemia

SPLINTR: single-cell profiling and lineage tracing

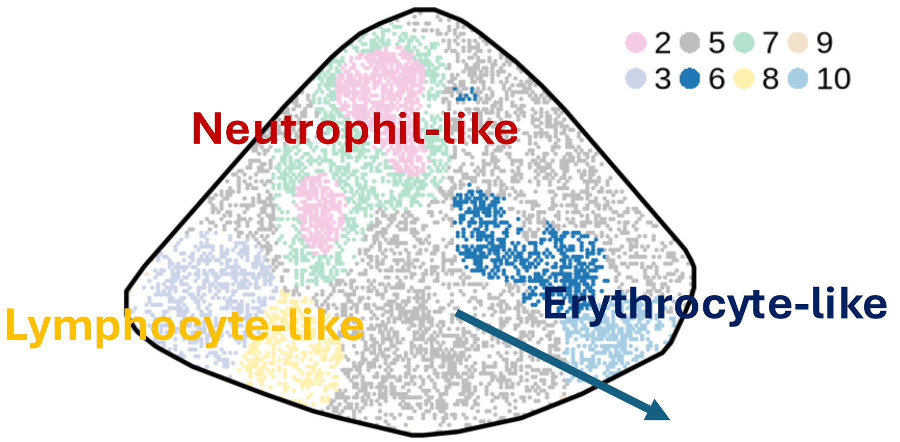
Cell states of ‘meta-clones’

Density of all AML barcodes





AML Meta-clones



Uncoloured: nonbarcoded background

Meta-clone-specific enriched cell types

	Meta-clone 2	Meta-clone 3	Meta-clone 6	Meta-clone 7	Meta-clone 8	Meta-clone 10
1	Granulo (BM)	Imm NK (BM)	RBC (CITE)	HPC (BM)	HPC (BM)	Naive B (BM)
2	Neutro (Spleen)	Trans B (Spleen)	RBC (Spleen)	IMM 1 (Neutro)	T (BM)	ProEryThBla (BM)
3	HPC (BM)	T (BM)	ErythBla (BM)	Neutro (Spleen)	Pre-B cycl (Spleen)	Pre-B cycl (Spleen)
4	T1 (Neutro)	MAT 3 (Neutro)	ProEryThBla (BM)	Pre-B cycl (Spleen)	Imm NK (BM)	ICOS+ Tregs (CITE)
5	Imm NK (BM)	CD8 T (Spleen)	IMM 2 (Neutro)	Pre-B (BM)	Pre-B (BM)	CD4 T (CITE)

Reflections

- Fast-evolving biotech, reliable stats method needed
  - 3D spatial, spatio-temporal, …
- Collaborative culture
  - Wet: Biologists, bioinformaticians;
  - Bridge: computational biologists;
  - Dry: statisticians, mathematicians, computer scientists
- (Effective) visualisation
  - Most commonly used: ggplot2 & plotly
  - What you want to show ≠ what viewers see