Analysis of single-cell RNA-seq data (V)

Hao Wu
Department of Biostatistics
and Bioinformatics
Rollins School of Public Health
Emory University

Ziyi Li
Department of Biostatistics
The University of Texas MD
Anderson Cancer Center

ENAR 2021 short course
March 2021

Course outline

- 8-9:15: Intro and data preprocessing.
- 9:15-9:45: Lab: preprocessing and visualization.
- 10-11:15: Normalization, batch effect, imputation, DE, simulator.
- 11:15-12: Lab: Normalization, batch effect, imputation, DE, simulator

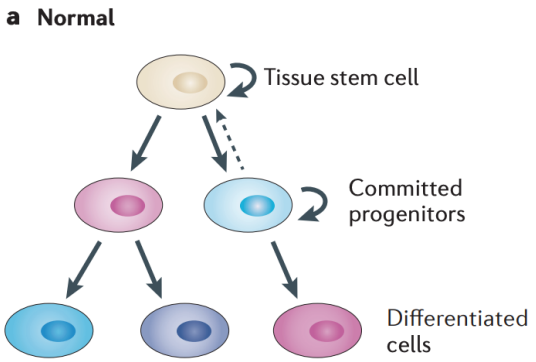
- 12-1: Lunch break

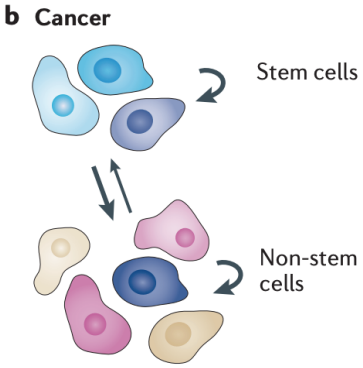
- 1-2: Clustering and pseudotime construction
- 2-2:30: Lab: Clustering and pseudotime construction
- 2:45–3:30: Supervised cell typing & related single cell data sources
- 3:30-4: Lab: supervised cell typing.
- **4:15-5: scRNA-seq in cancer**

Outline for this session

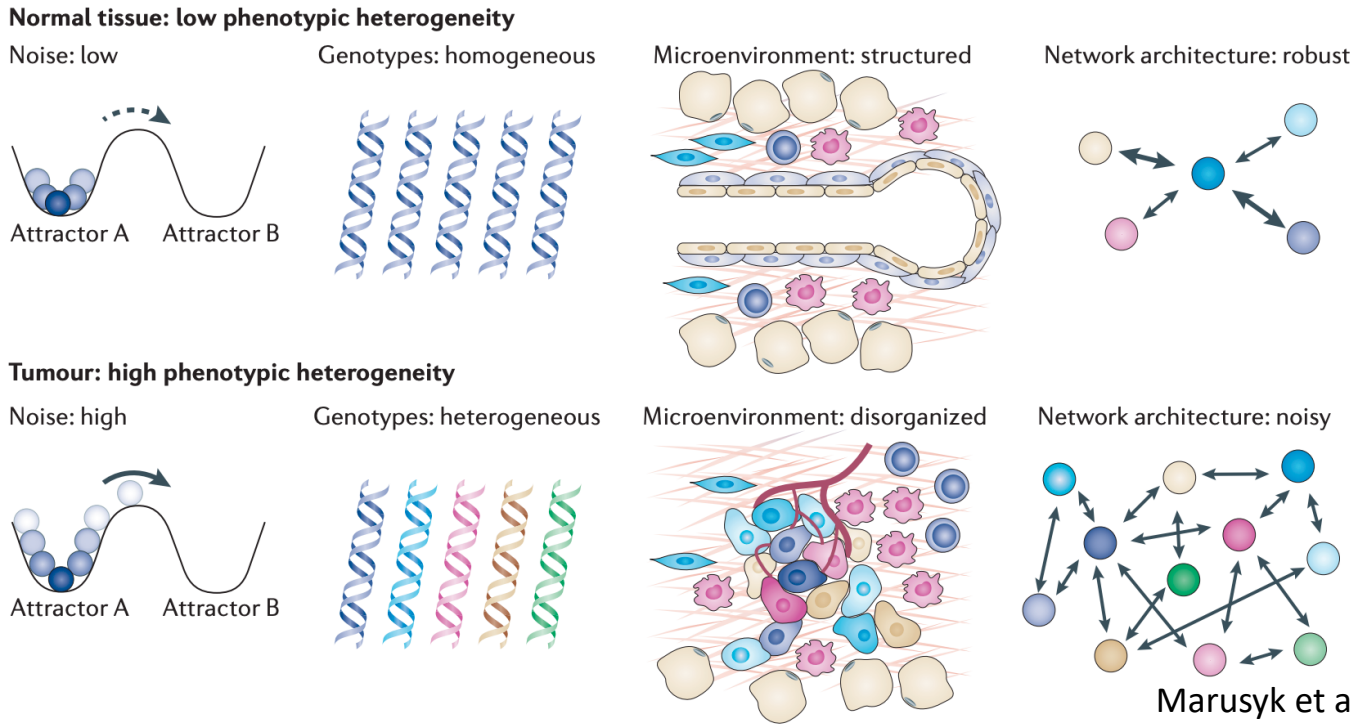
- **Background**
  - Uniqueness of tumor tissue
  - Opportunities and challenges
- **Relevant computational methods**
  - Unified analysis across condition and multiple samples
  - Distinguishing neoplastic from nonneoplastic cells
  - Inferring communication with tumor microenvironment
  - Delineating tumoral and microenvironment evolution
  - Other tumor-specific topics
- **Future opportunities**

Uniqueness of tumor tissue



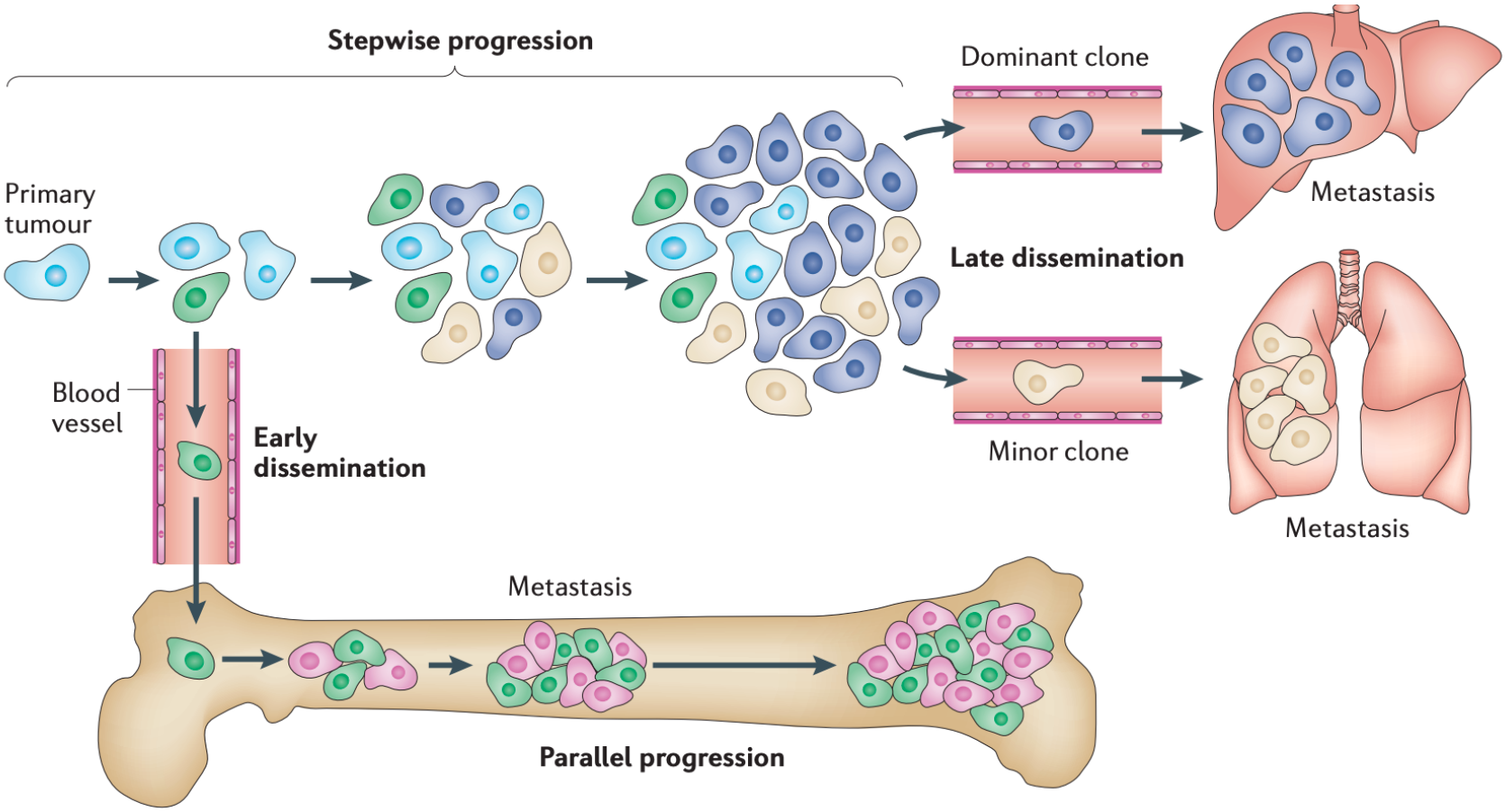


- Differentiation hierarchies is changed in cancer cells
- Different factors shape cellular phenotypes



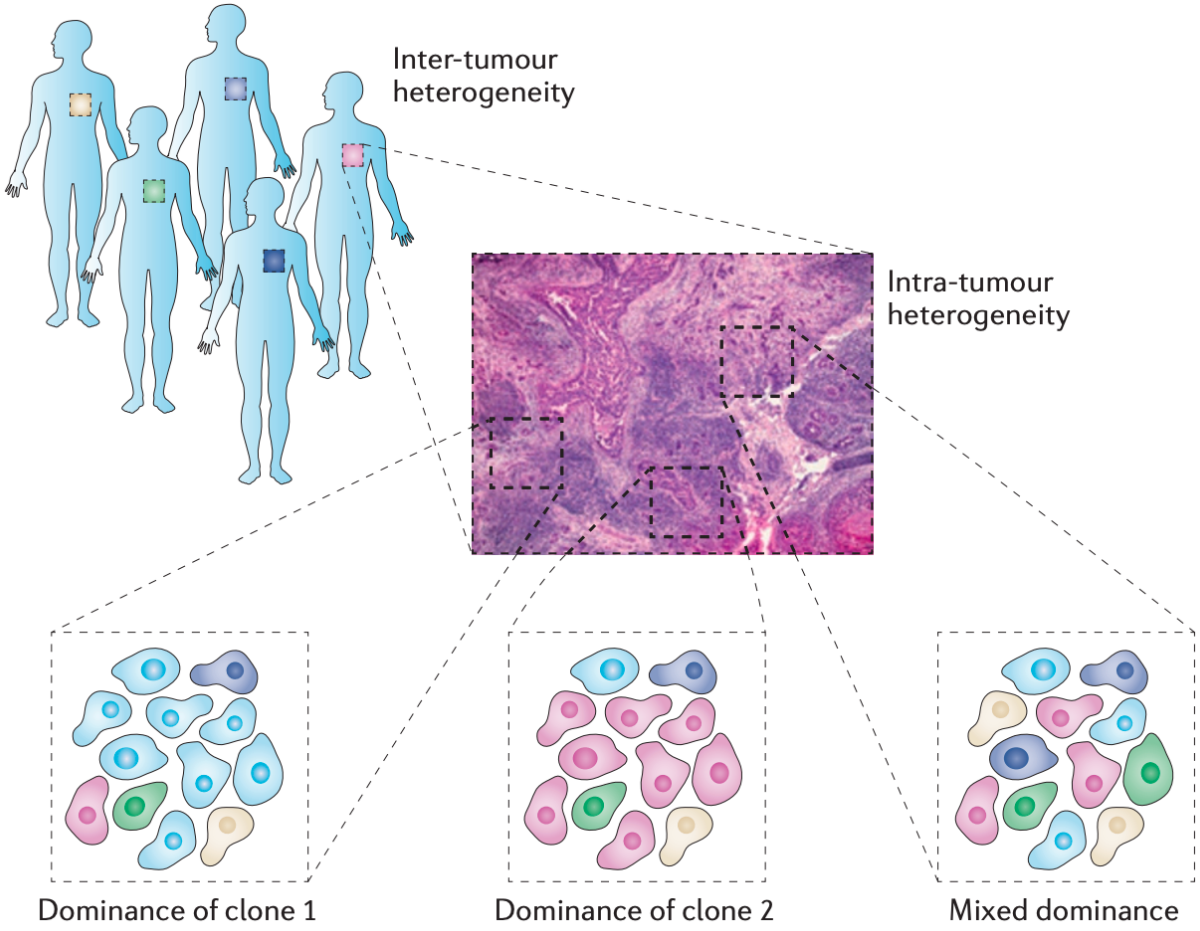
Marusyk et al. 2012, Nat. Rev. Cancer

Uniqueness of tumor tissue



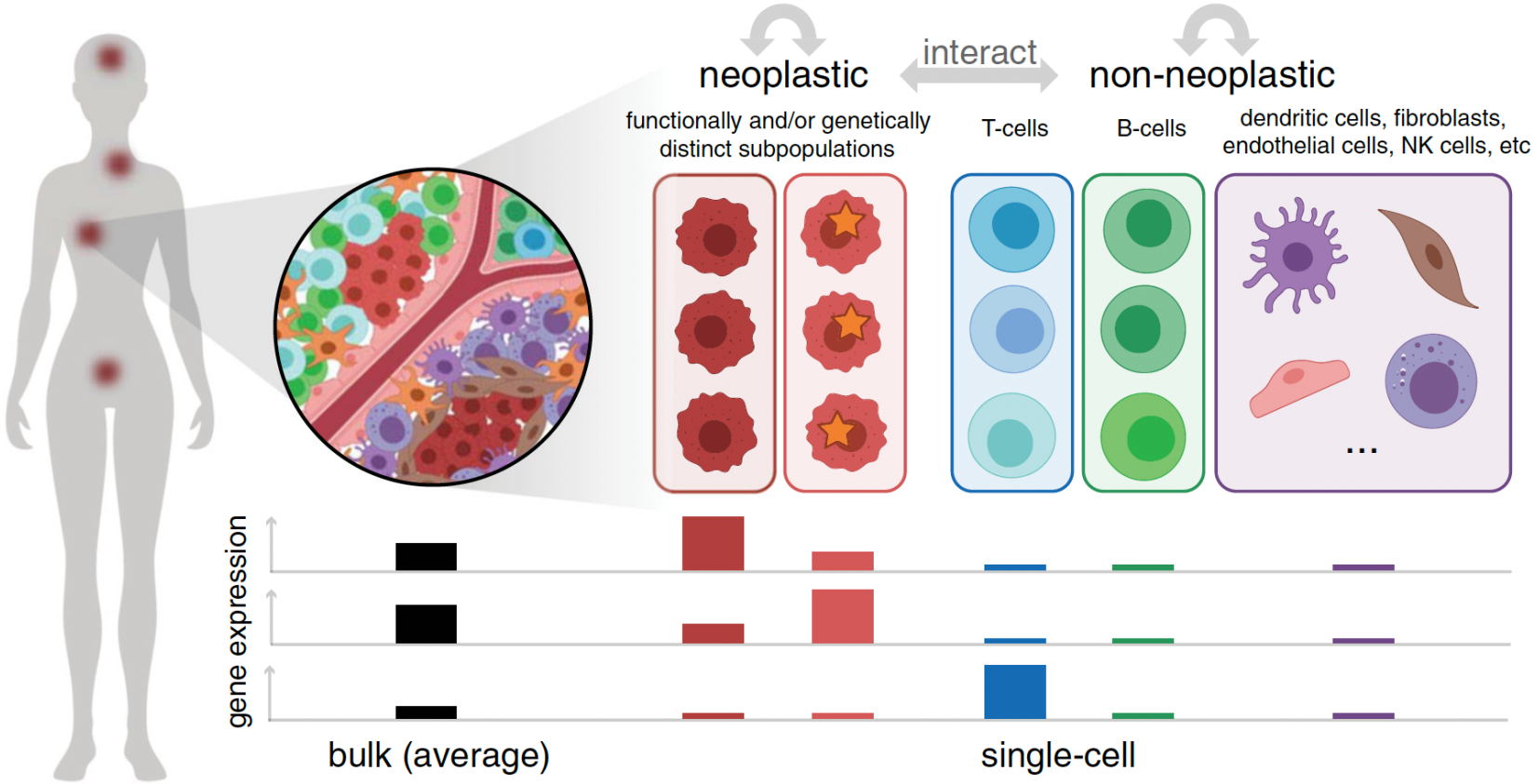
Marusyk et al. 2012, Nat. Rev. Cancer

Uniqueness of tumor tissue



Marusyk et al. 2012, Nat. Rev. Cancer

Single cell RNA-seq provides unbiased characterization of cell profiles in tumor environment



Fan et al. 2020. Experimental & Molecular Medicine

Unified analysis across many patients and disease states

- Goal: identifying common cell types and states shared across patients and disease states from multiple scRNA-seq datasets.
- Batch effect is a big concern here.
- Batch correction tools: MultiCCA, MNN, combat, etc.
- Newly emerged tools: LIGER, Harmony, scVI, SAUCIE

Challenges in clustering: neoplastic cells

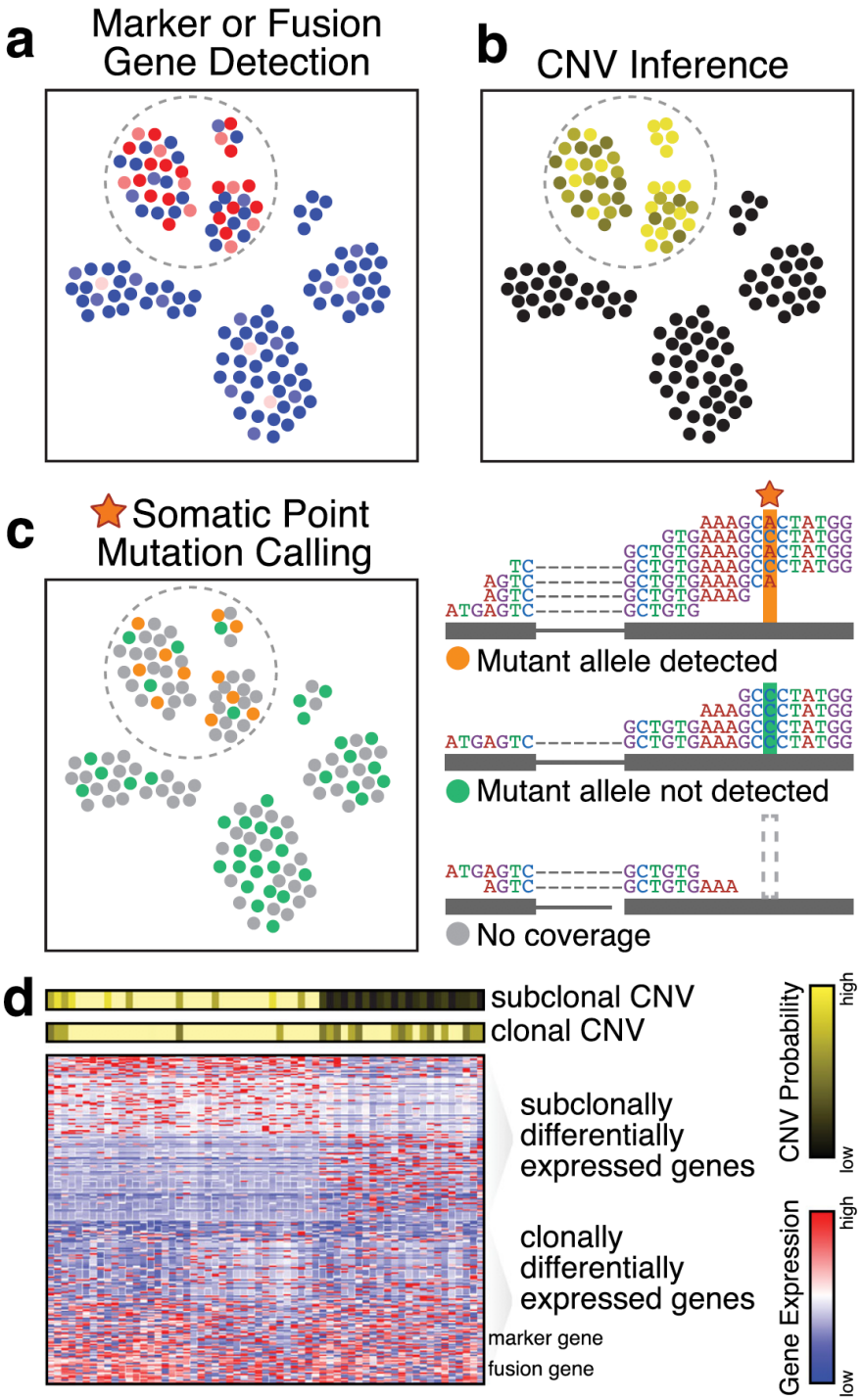
- Neoplastic cells aggregate by patient due to the inter-patient heterogeneity for neoplastic versus non-neoplastic cells
- Neoplastic cells need to be considered separately from nonneoplastic cells
- Clustering per patient is also recommended to avoid over-correction
- Perform generic batch correction with caution

Distinguishing neoplastic from nonneoplastic cells

- Neoplastic cells generally exhibit extensive alterations in a variety of biochemical pathways and oncogenic programs emblematic of cancer
- Certain cancers - distinct marker genes or combinations of marker genes
  - E.g. multiple myeloma cells are marked by CD38+/CD138+ antigen expression

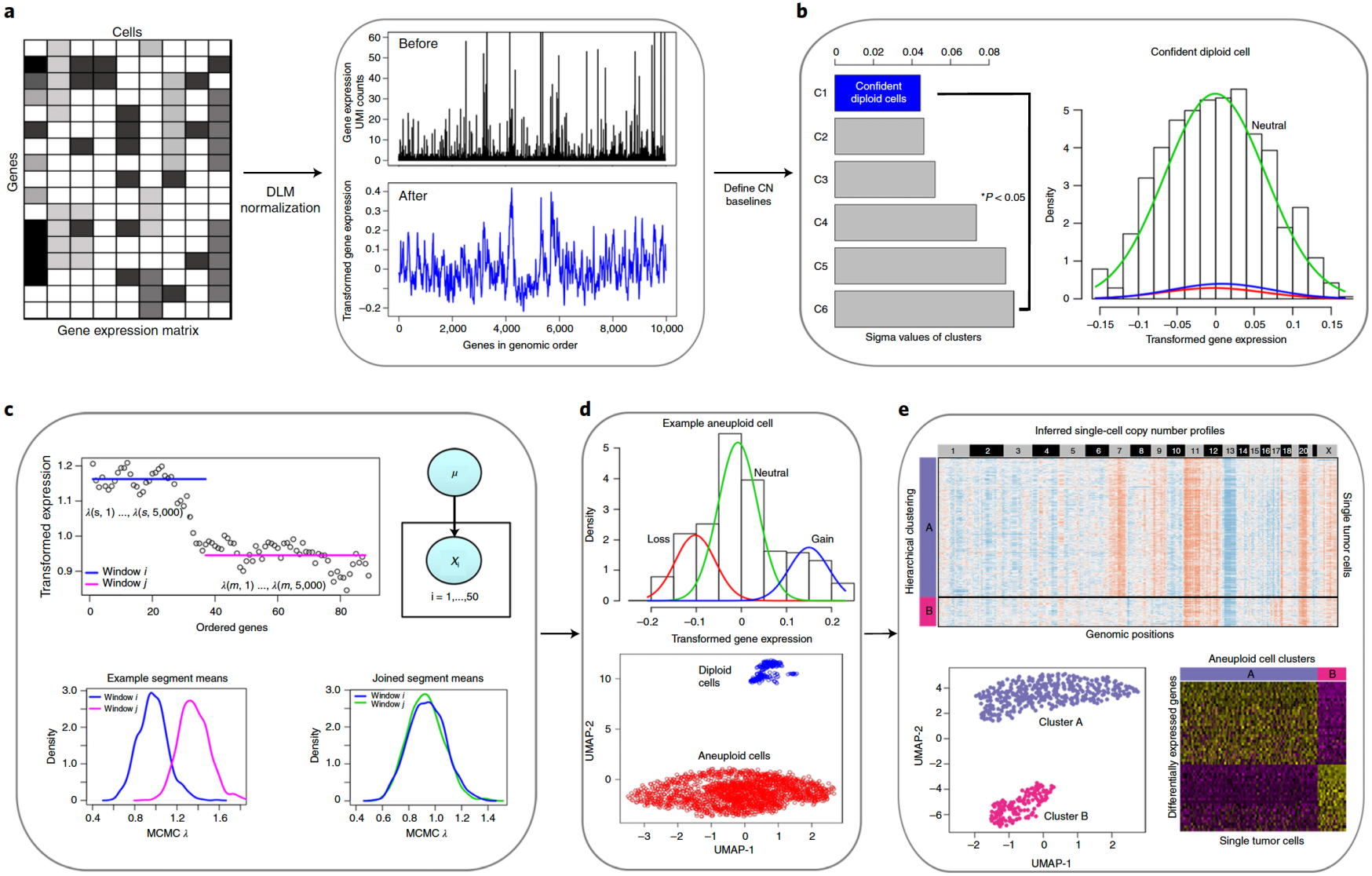
Distinguishing neoplastic from nonneoplastic cells (continue)

- Other cancers - marker gene or pathway are not enough
  - Neoplastic cells can also express genes and pathways typically associated with canonical nonneoplastic cells in ways that we might not expect.
  - CNV inference-based detection: InferCNV, CopyKAT
  - Point mutation-based detection: HoneyBADGER
  - some cancers are not well defined by either large-scale CNVs or somatic point mutations (chronic myeloid leukemia – BCR-ABLfusion)



Fan et al. 2020.
Experimental & Molecular Medicine

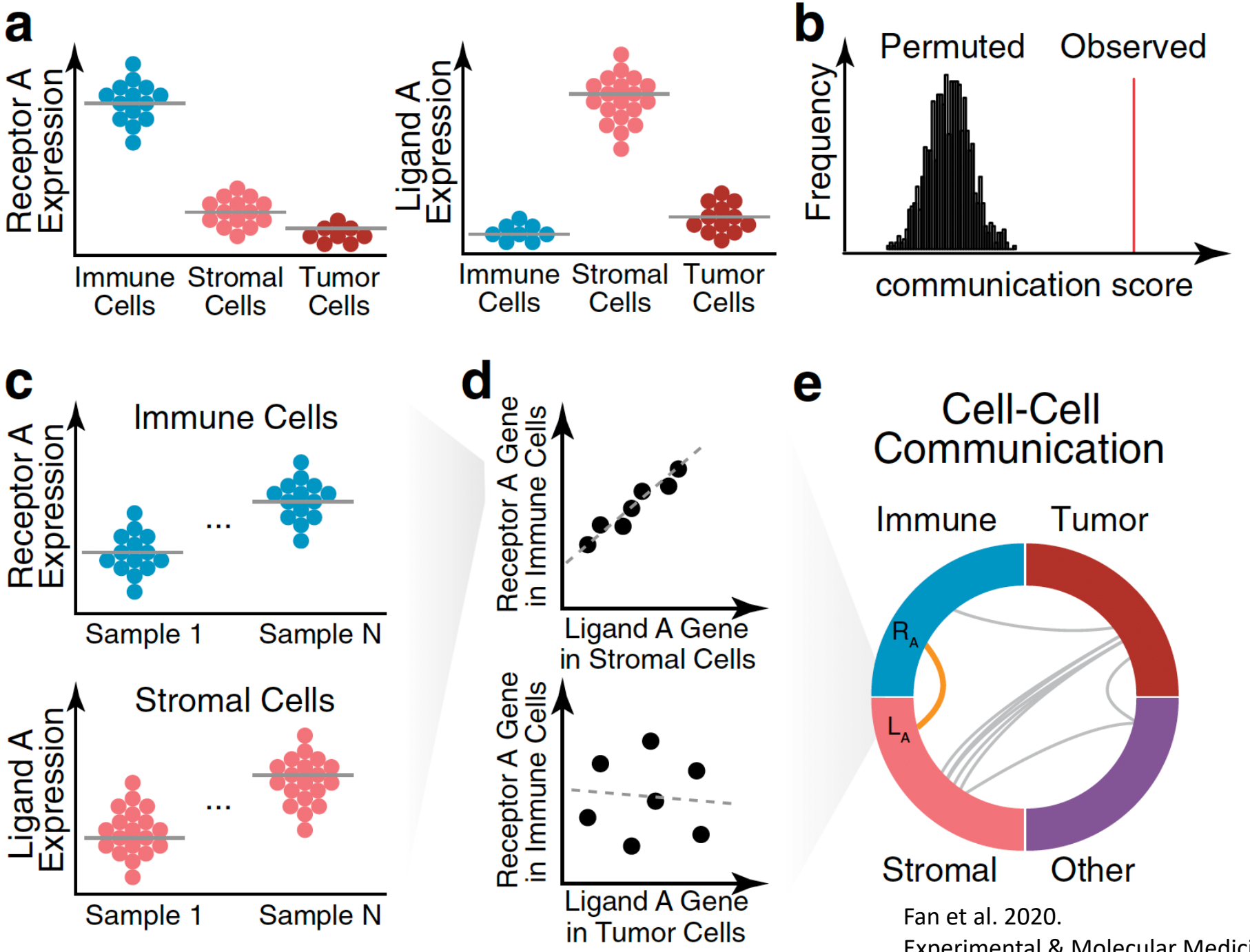
CopyKAT



Gao et al. 2021. Nature Biotech

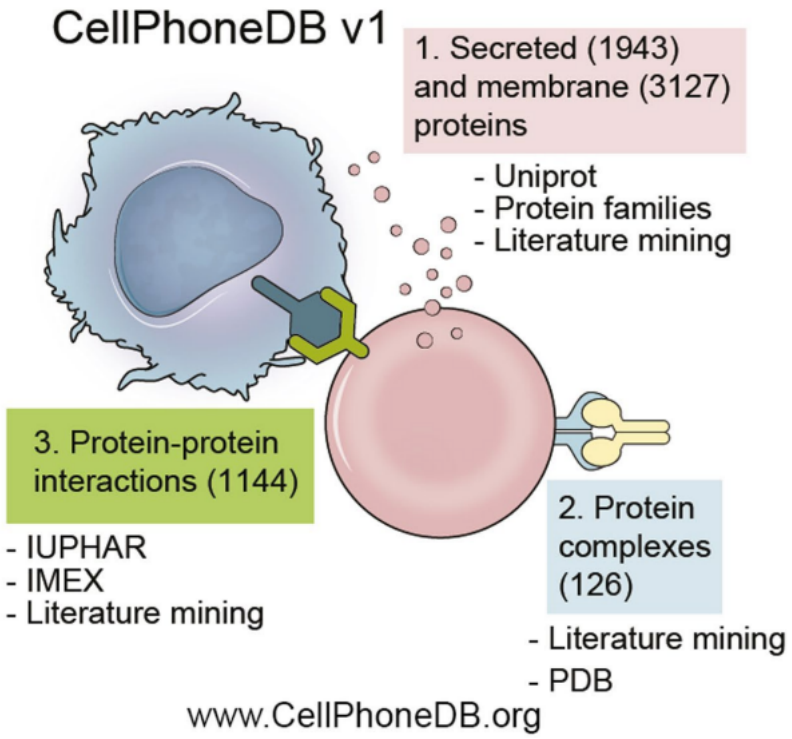
Inferring communication within the tumor microenvironment

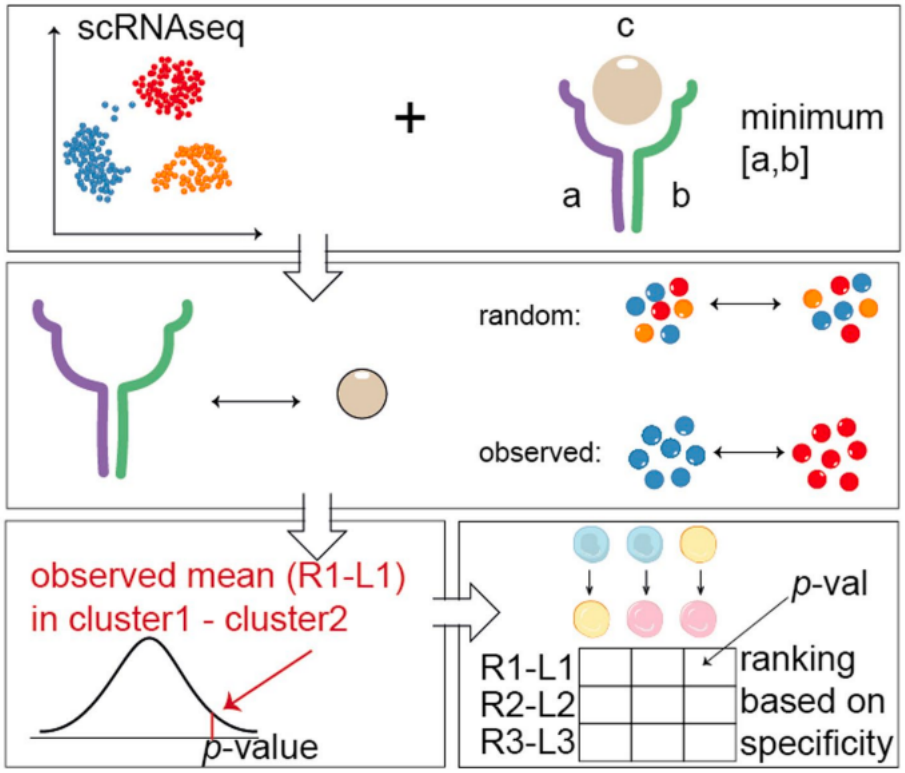
- Infer putative communication of cell types: comparison of the expression levels of a receptor gene in one cell type and a corresponding ligand gene in another cell type
- Methods: CellPhoneDB (2018 Nature, 2020 Nature protocol), CellTalker (2020 Immunity), NicheNet (2020 Nature Methods), iTalk (bioRxiv, 2019)



Fan et al. 2020.
Experimental & Molecular Medicine

CellPhoneDB





Deconvolution of bulk RNA-seq using single cell RNA-seq data

- Infer proportions of different immune and stromal cell type
- Assumption: bulk sample is a mixture of multiple transcriptionally distinguishable cell types
- Methods: MuSiC (Wang et al. 2019, Nat Comm), CibersortX (Newman et al. 2019, Nature)
- Other: TIMER, Cibersort, MCP-counter, xCell
- Marker gene selection is important - cancer cells may aberrantly express genes associated with canonical immune or nonneoplastic cell types

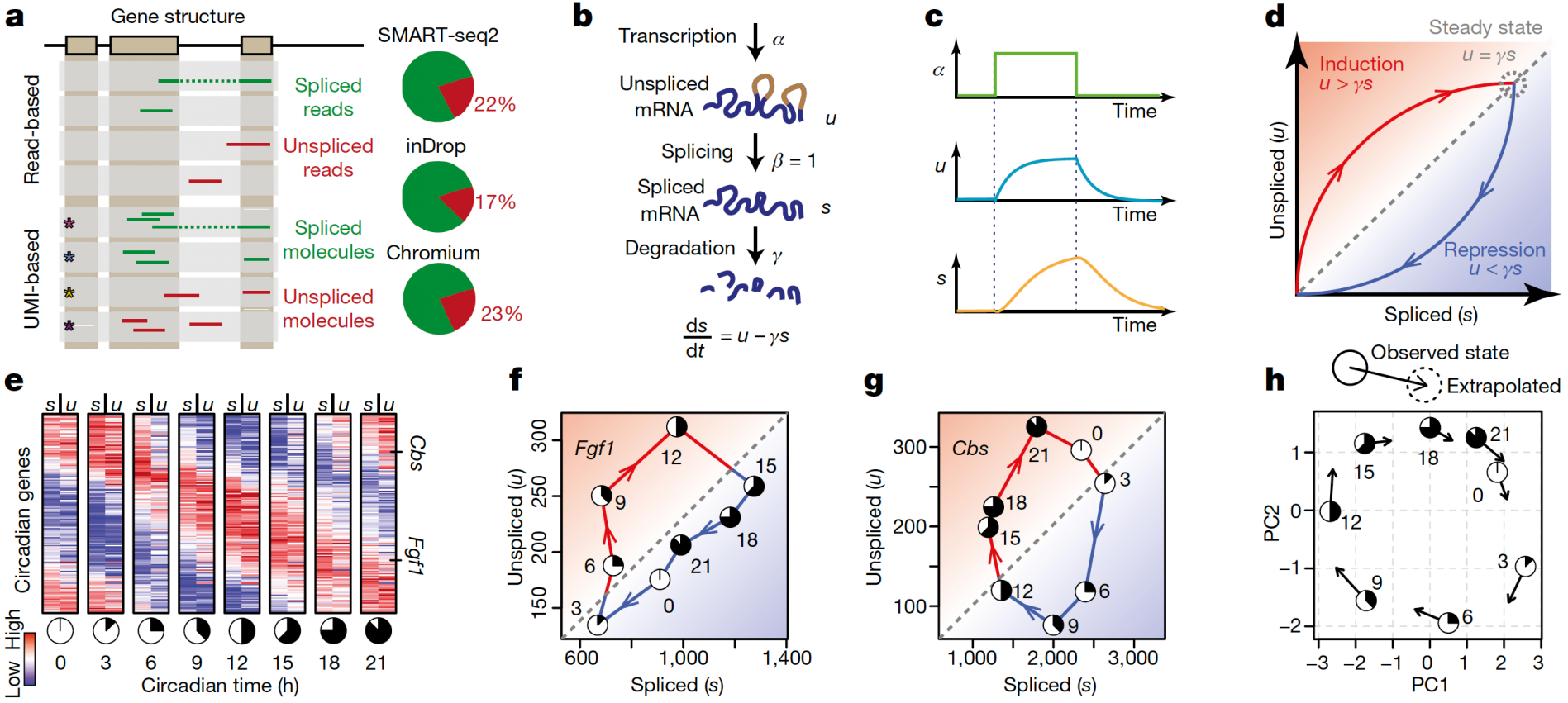
Delineating tumoral and microenvironmental evolution

- Pseudotime construction methods could be used for trajectory construction
- Special attention is needed for determining start and end point in trajectory construction
- RNA velocity analysis

RNA velocity analysis

- Has been applied on some cancer studies, but not cancer specific. It is originally designed for capturing developmental trajectory.
- Balance between unspliced and spliced mRNAs is predictive of cellular state progression
- Increase in the transcription rate: a rapid increase in unspliced mRNA -> increase in spliced mRNA -> new steady state
- Drop in the rate of transcription: a rapid drop in unspliced mRNA -> reduction in spliced mRNA -> stead state
- Such spliced/unspliced states can be identified using protocols of SMART-seq2, inDrop, STRT/C1, and 10x genomics

RNA velocity analysis



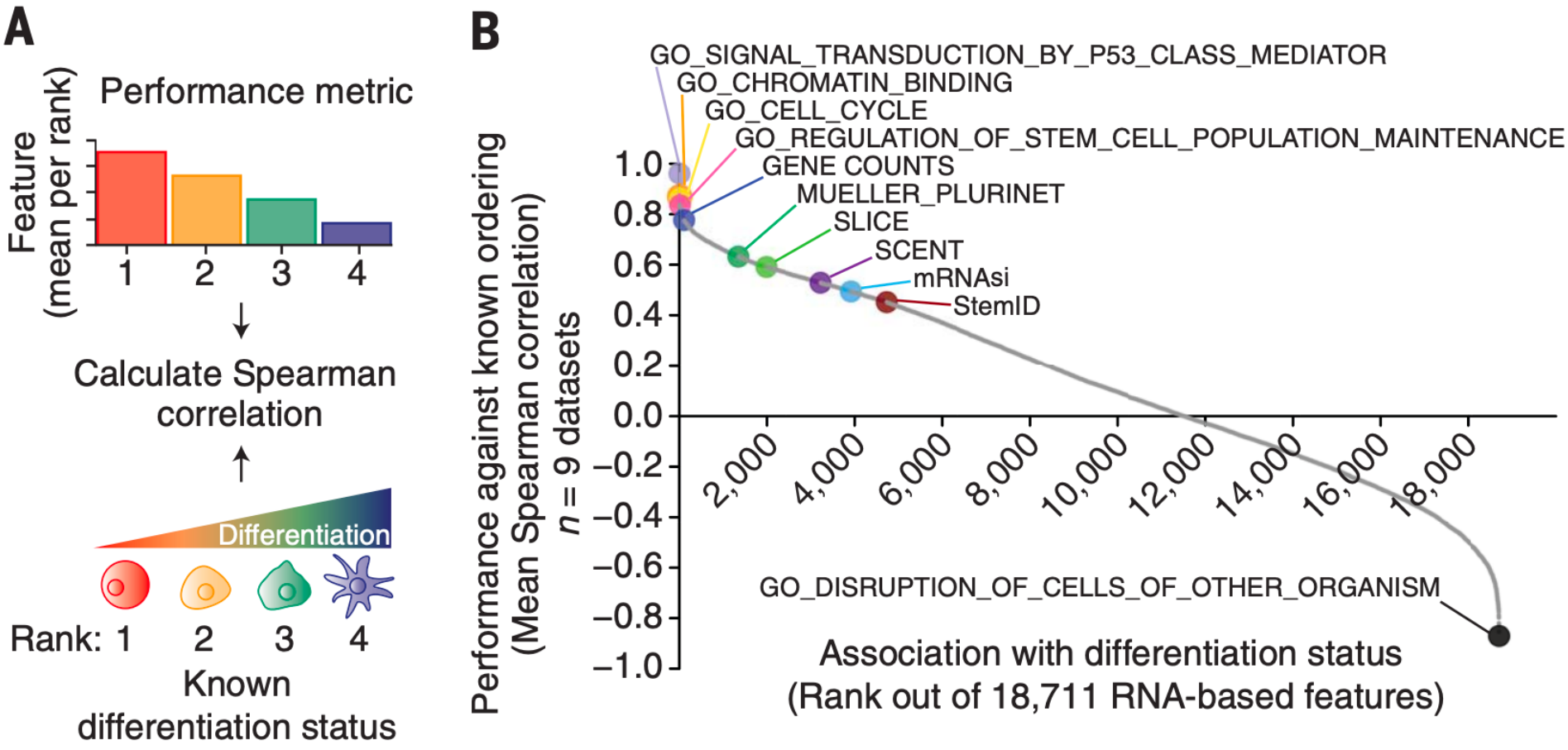
CytoTRACE

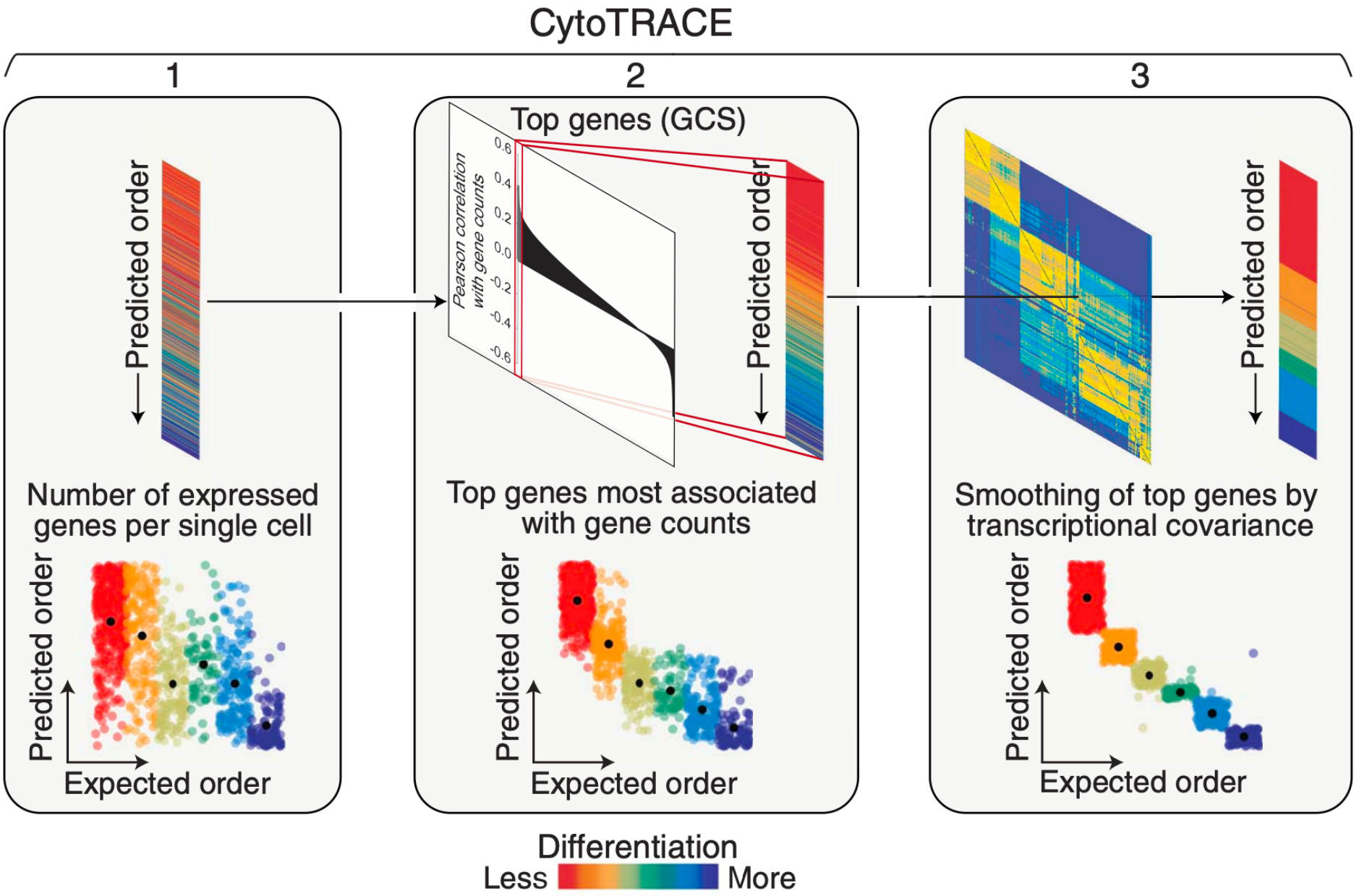
RESEARCH ARTICLE

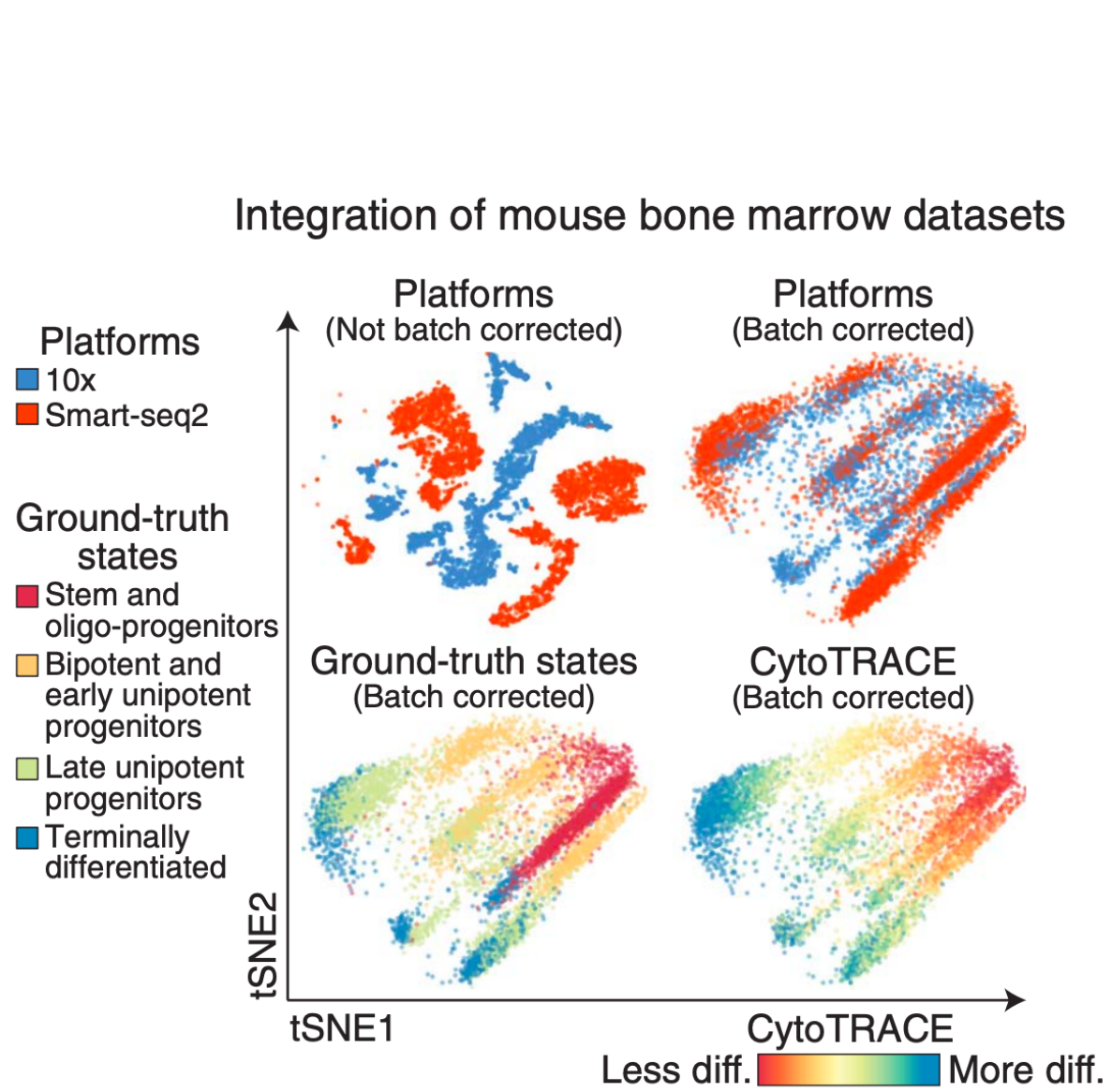
RESEARCH METHODS

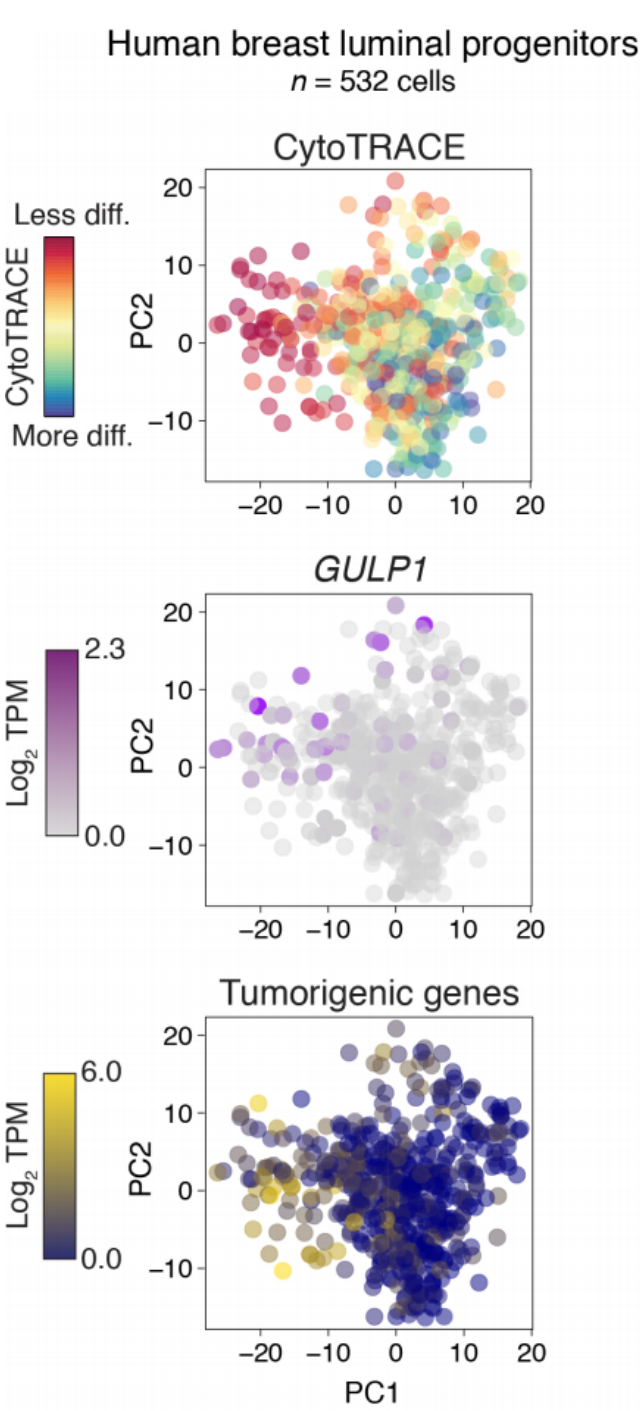
Single-cell transcriptional diversity is a hallmark of developmental potential

Gunsagar S. Gulati[1]*, Shaheen S. Sikandar[1]*, Daniel J. Wesche[1], Anoop Manjunath[1], Anjan Bharadwaj[1], Mark J. Berger[2]†, Francisco Ilagan[1], Angera H. Kuo[1], Robert W. Hsieh[1], Shang Cai[3], Maider Zabala[1]‡, Ferenc A. Scheeren[4], Neethan A. Lobo[1]‡, Dalong Qian[1], Feiqiao B. Yu[5], Frederick M. Dirbas[6], Michael F. Clarke[1,7], Aaron M. Newman[1,8]§


A
Performance metric
Feature
(mean per rank)
1
2
3
4
Calculate Spearman
correlation
Differentiation
Rank: 1
2
3
4
Known
differentiation status
B
GO_SIGNAL_TRANSDUCTION_BY_P53_CLASS_MEDIATOR
GO_CHROMATIN_BINDING
GO_CELL_CYCLE
GO_REGULATION_OF_STEM_CELL_POPULATION_MAINTENANCE
GENE COUNTS
MUELLER_PLURINET
SLICE
SCENT
mRNAsi
StemID
Performance against known ordering
(Mean Spearman correlation)
n = 9 datasets
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
GO_DISRUPTION_OF_CELLS_OF_OTHER_ORGANISM
Association with differentiation status
(Rank out of 18,711 RNA-based features)


CytoTRACE
1
2
3
Predicted order
Top genes (GCS)
Pearson correlation with gene counts
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
Predicted order
Predicted order
Number of expressed genes per single cell
Top genes most associated with gene counts
Smoothing of top genes by transcriptional covariance
Predicted order
Expected order
Predicted order
Expected order
Predicted order
Expected order
Differentiation
Less
More


Integration of mouse bone marrow datasets
Platforms
10x
Smart-seq2
Ground-truth states
Stem and oligo-progenitors
Bipotent and early unipotent progenitors
Late unipotent progenitors
Terminally differentiated
Platforms
(Not batch corrected)
Platforms
(Batch corrected)
Ground-truth states
(Batch corrected)
CytoTRACE
(Batch corrected)
tSNE2
tSNE1
CytoTRACE
Less diff.
More diff.


Human breast luminal progenitors
n = 532 cells
CytoTRACE
Less diff.
More diff.
CytoTRACE
PC2
GULP1
Log2 TPM
2.3
0.0
Tumorigenic genes
6.0
0.0
PC1
20
10
0
−10
−20

Futures

- Single cell multi-omics: epigenetic heterogeneity and its interplay with transcriptional heterogeneity at the single-cell level
- Spatial transcriptomics
- International consortium: Human Cell Atlas, Human Developmental Cell Atlas, Pediatric Cell Atlas, HuBMAP, Human Tumor Atlas Network, LifeTime EU Flagship