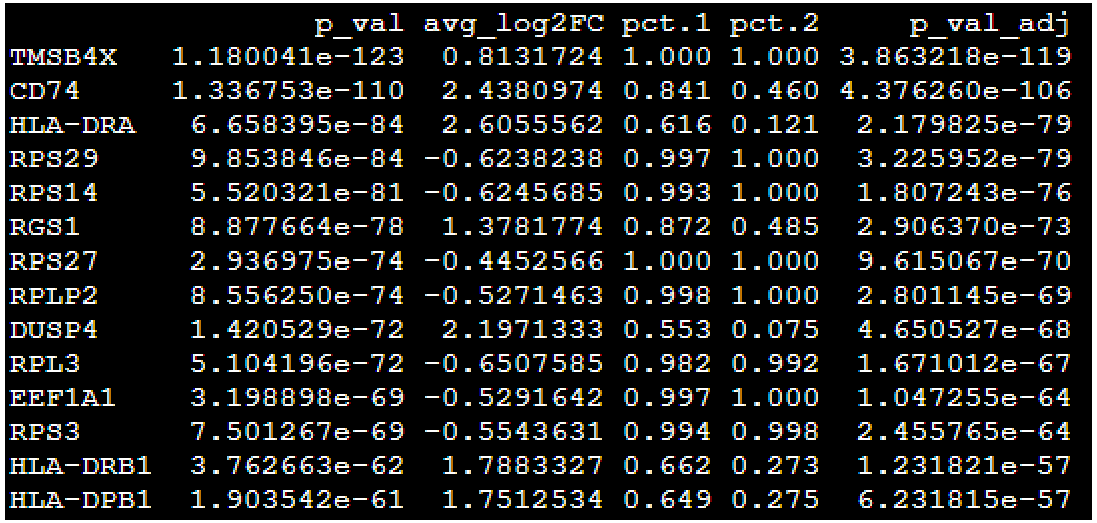
Single-cell RNA-sequencing

• Differentially expressed gene analysis between two groups

- Wilcoxon rank sum test between two groups for each cell type (or cluster)

→ Nonparametric approach (does not require a specific distribution of data)

→ Adjusted p-value, average_Log2FC + expression cell ratio (pct.1, pct.2)



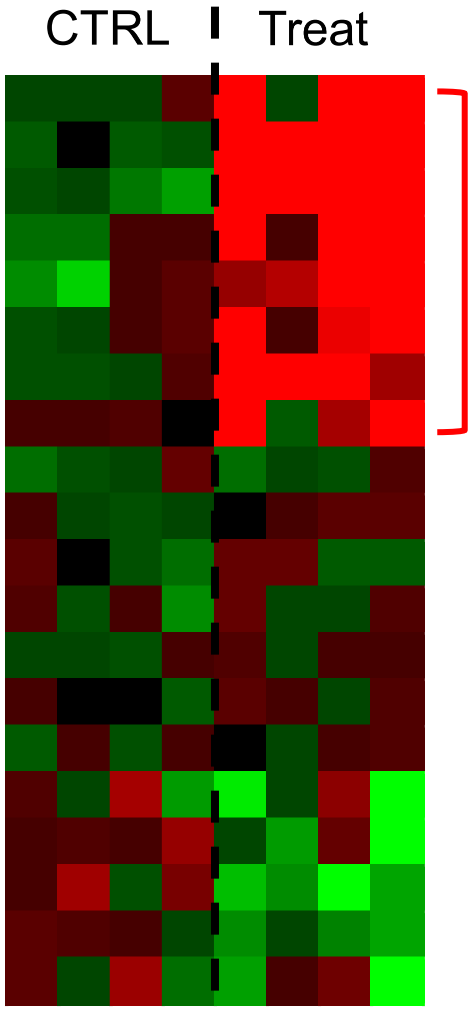
```
                 p_val avg_log2FC pct.1 pct.2     p_val_adj
TMSB4X   1.180041e-123  0.8131724 1.000 1.000 3.863218e-119
CD74     1.336753e-110  2.4380974 0.841 0.460 4.376260e-106
HLA-DRA   6.658395e-84  2.6055562 0.616 0.121  2.179825e-79
RPS29     9.853846e-84 -0.6238238 0.997 1.000  3.225952e-79
RPS14     5.520321e-81 -0.6245685 0.993 1.000  1.807243e-76
RGS1      8.877664e-78  1.3781774 0.872 0.485  2.906370e-73
RPS27     2.936975e-74 -0.4452566 1.000 1.000  9.615067e-70
RPLP2     8.556250e-74 -0.5271463 0.998 1.000  2.801145e-69
DUSP4     1.420529e-72  2.1971333 0.553 0.075  4.650527e-68
RPL3      5.104196e-72 -0.6507585 0.982 0.992  1.671012e-67
EEF1A1    3.198898e-69 -0.5291642 0.997 1.000  1.047255e-64
RPS3      7.501267e-69 -0.5543631 0.994 0.998  2.455765e-64
HLA-DRB1  3.762663e-62  1.7883327 0.662 0.273  1.231821e-57
HLA-DPB1  1.903542e-61  1.7512534 0.649 0.275  6.231815e-57
```

GZMA: Tumor_CD8 T cell > Normal_CD8 T cell
→ Cytotoxic

HLA-DRA, HLA-DRB1, HLA-DPB1: Tumor_CD8 T cell > Normal_CD8 T cell
→ activated

- Geneset analysis

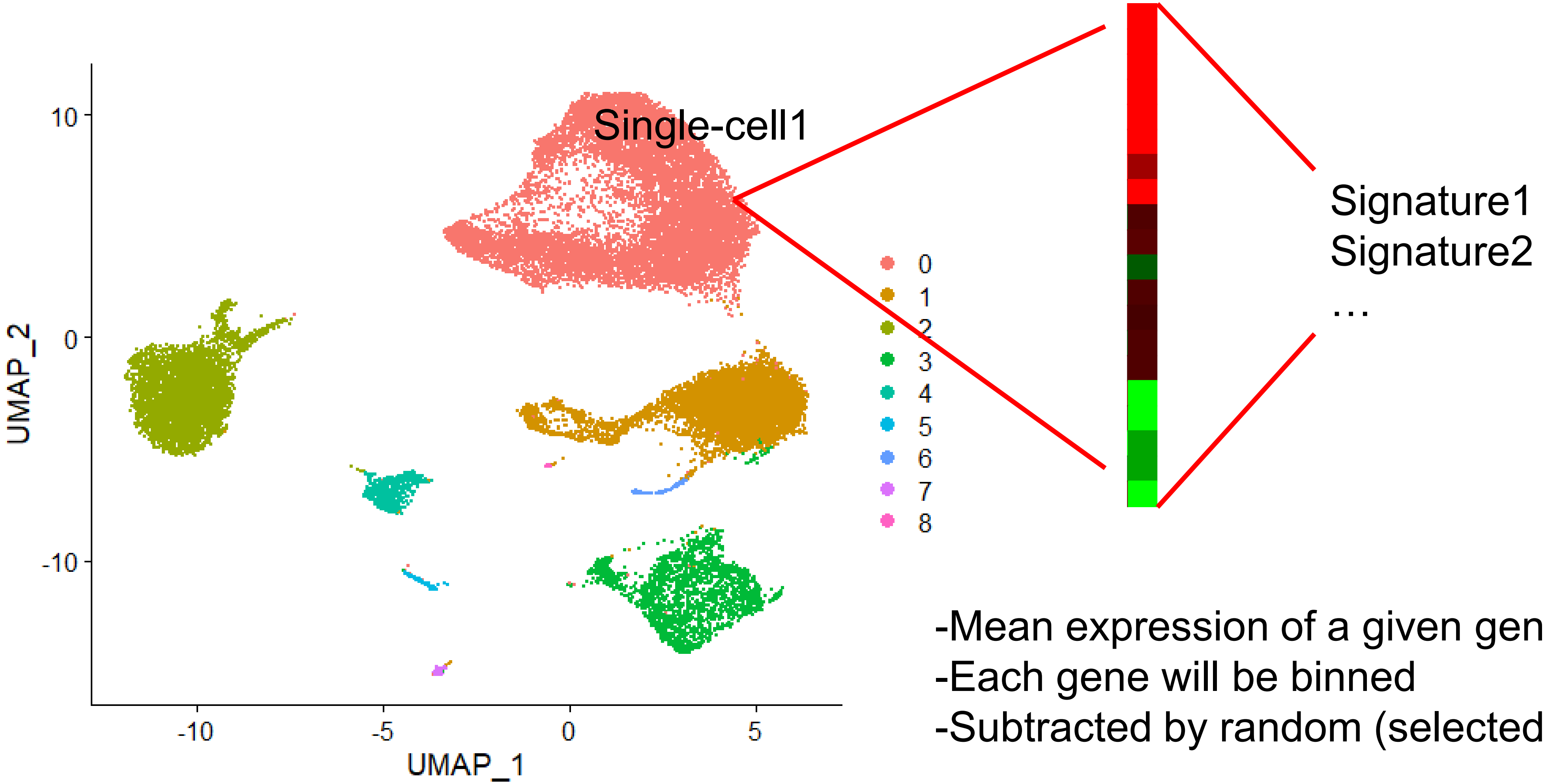
Celltype1 **Celltype2 …**



→ Fisher’s exact test or GSEA

→ Or measuring “signature score” for each cell
→ Kind of supervised dimension reduction (?)

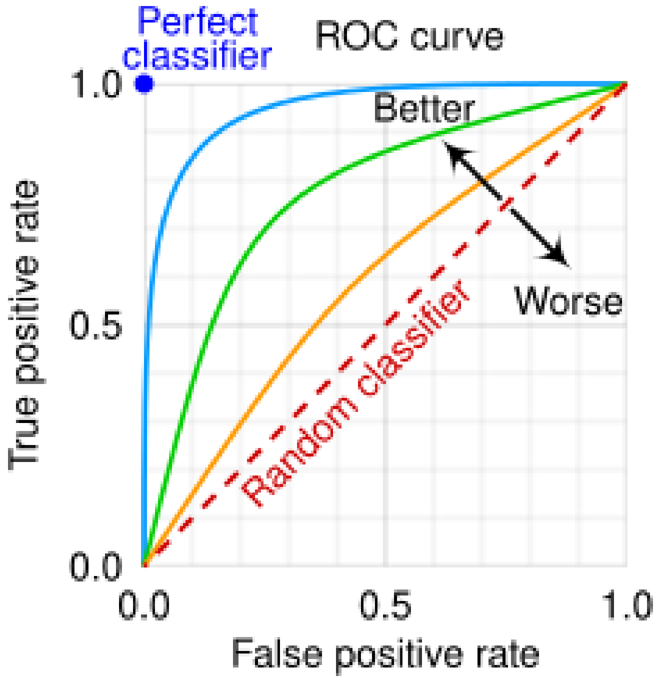
- AddModuleScore



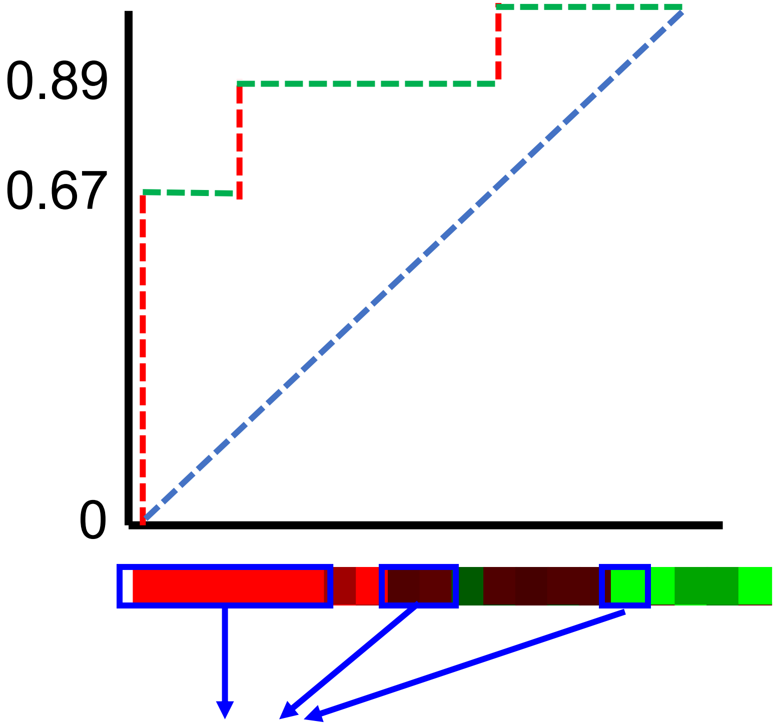
-Mean expression of a given geneset
-Each gene will be binned
-Subtracted by random (selected by each bin) noise

- AUCell

-Order genes by expression for each cell
-Measure AUC for a given geneset



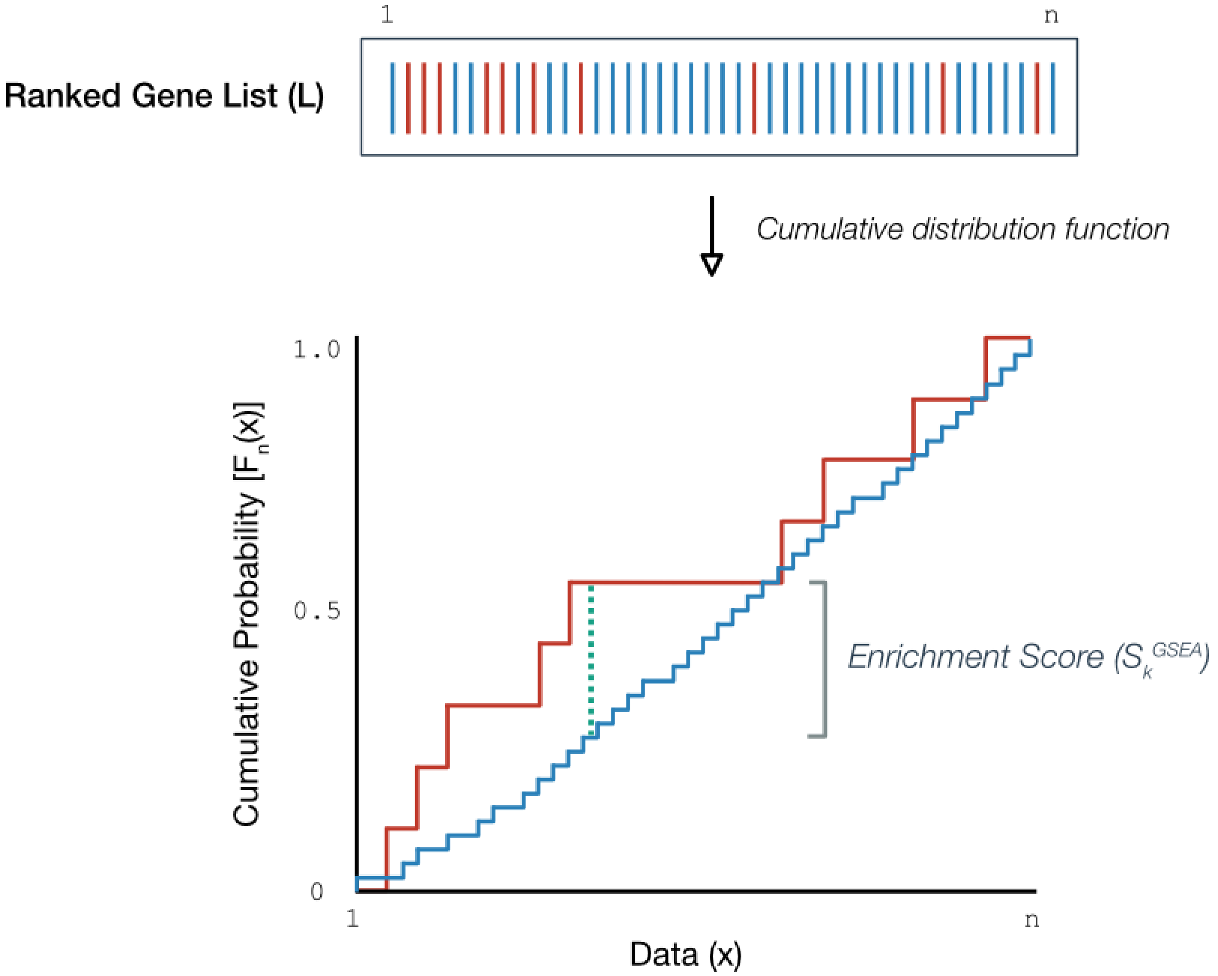
-Receiver operating characteristic (ROC) curve: performance measurement
-AUC: area under curve
→ Quantification



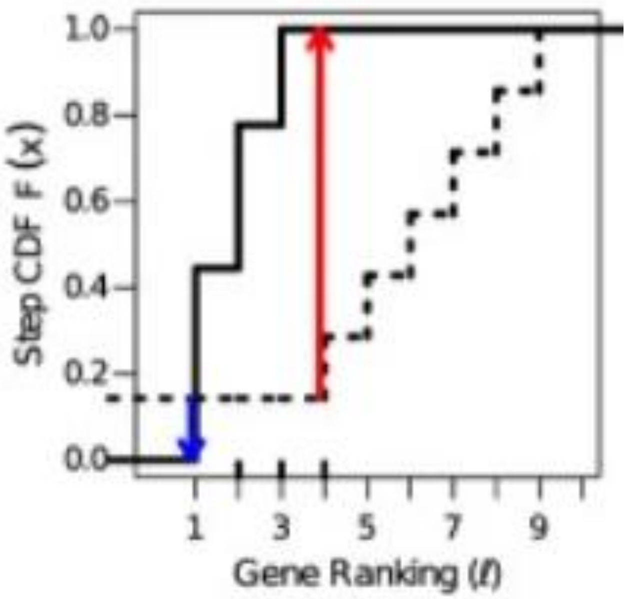
Given genes (6,2,1: 9)

ssGSEA & GSVA

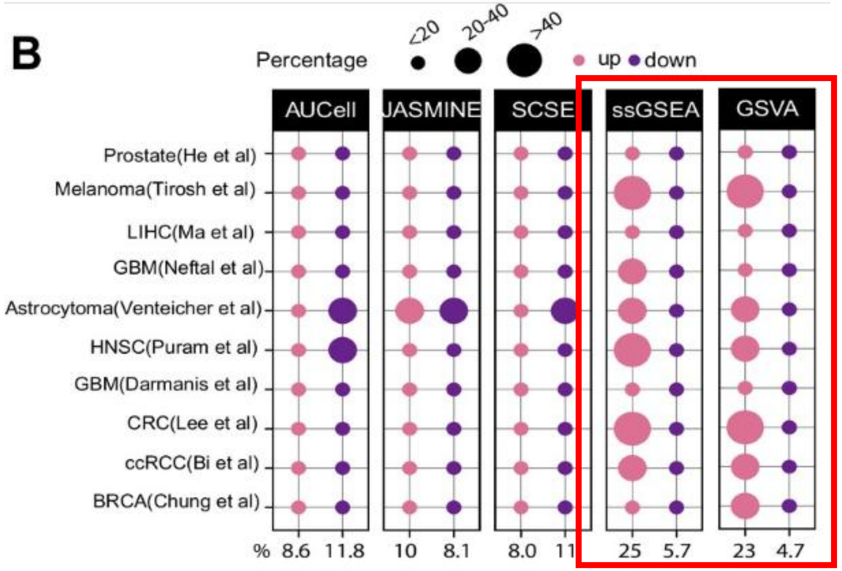
*ssGSEA
-Order genes by expression for each cell
-Make a ECDF (empirical cumulative distribution function) for a given geneset and remaining genes, respectively
-Integration of difference between two ECDFs

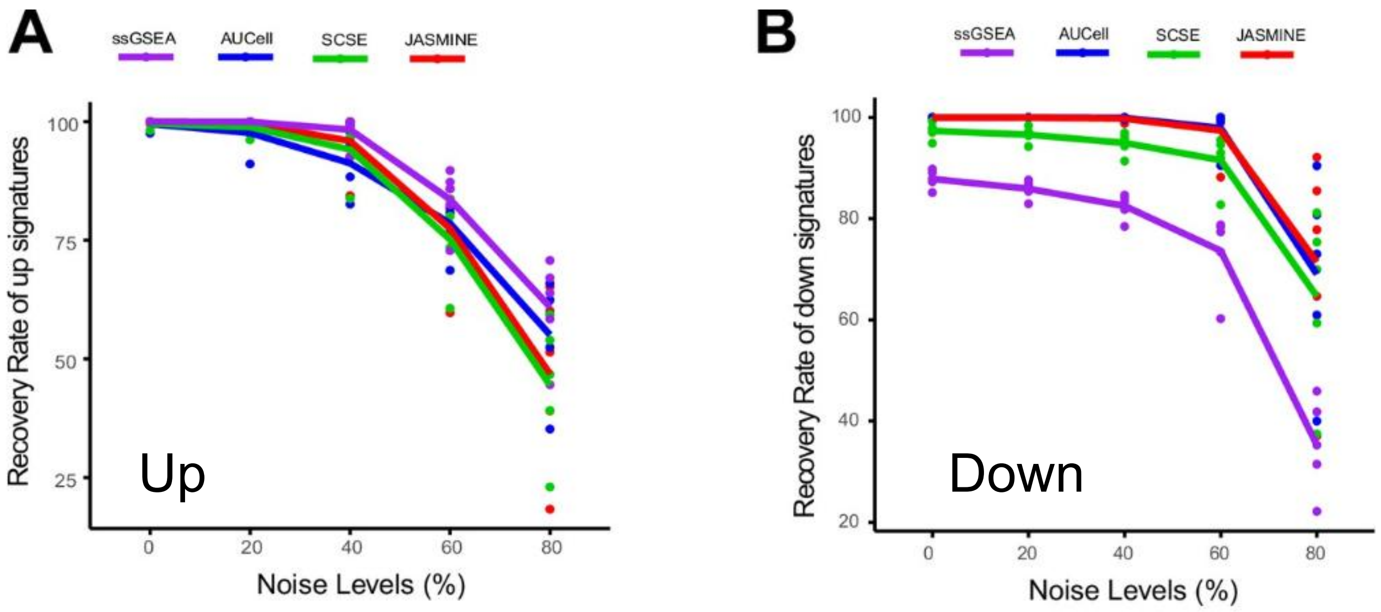


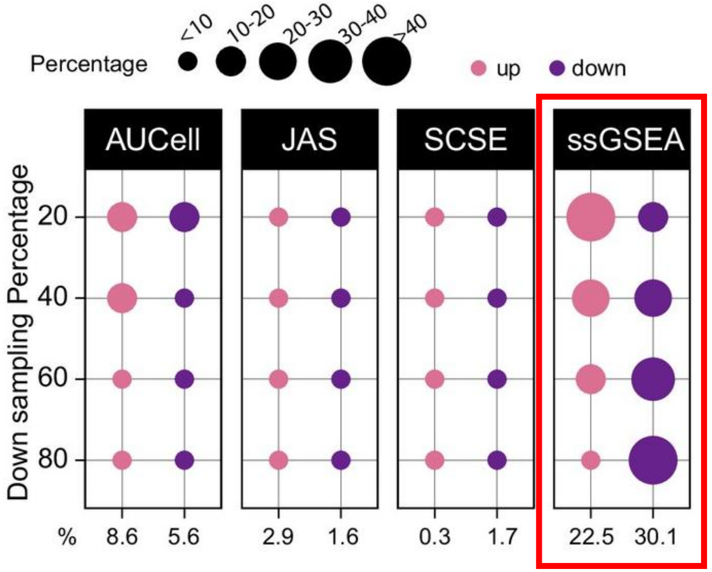
*GSVA
-Order genes by expression for each cell
-Make a ECDF
-KS-test (Kolmogorov-Smirnov test) Statistics by maximum difference



- Is not always good







-Biased to cancer (upregulation >> down regulation)
-Robustness (against noise): Up > Down
-Robustness (against down-sampling): ssGSEA (worst)

Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data

- cNMF

(consensus non-negative matrix factorization)

*Unsupervised approach
NMF: Decomposition method
Make W,H be close to X
Gradient descending

$$X = WH$$

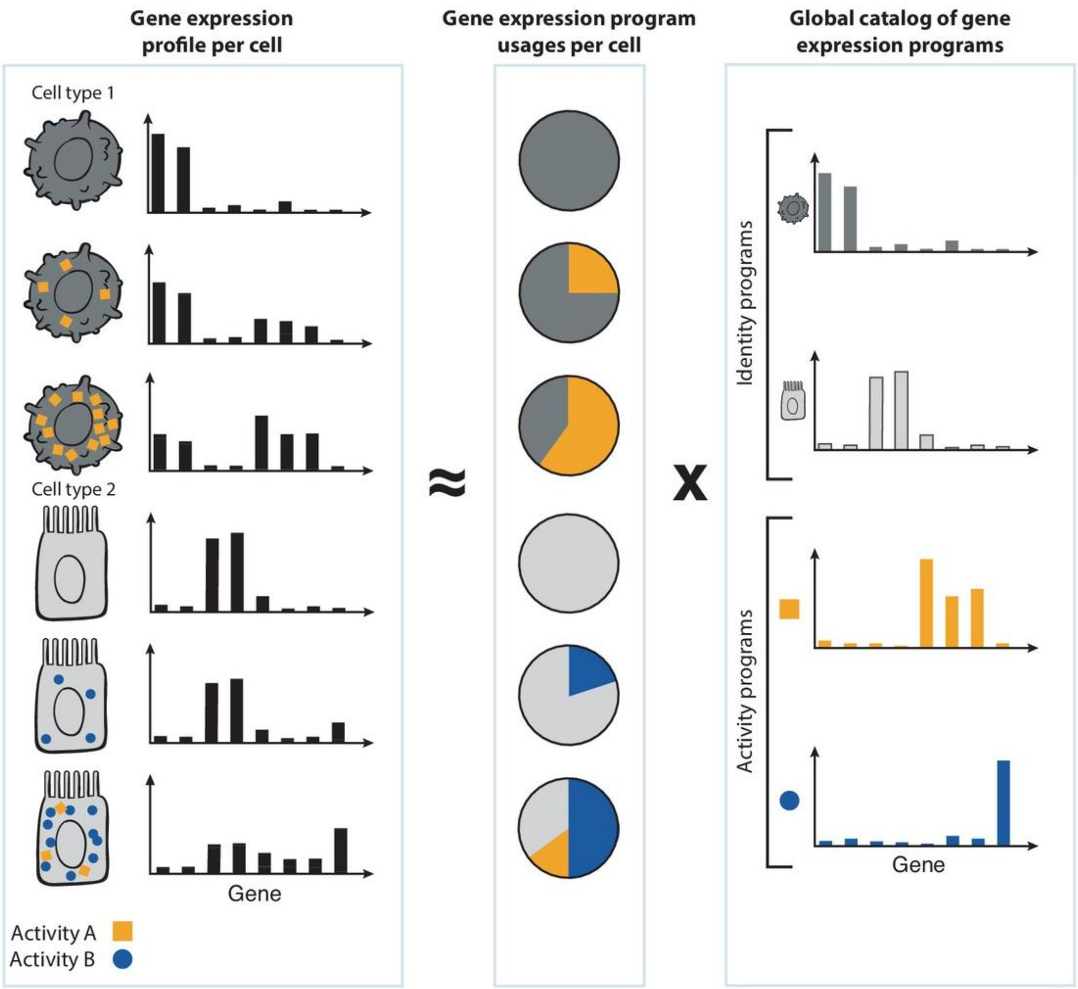
$$H := H - \eta_H \circ \boxed{\nabla_H \|X - WH\|_F^2}$$

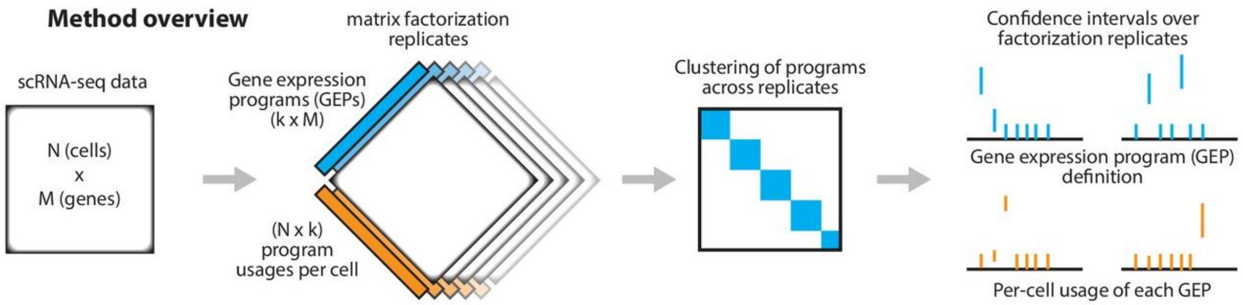
$$W := W - \eta_W \circ \boxed{\nabla_W \|X - WH\|_F^2}$$

$$\therefore H := H \circ \frac{W^T X}{W^T W H}$$

식 $(36) \Rightarrow W := W + \frac{W}{WHH^T} \circ (XH^T - WHH^T)$

$$= W + W \circ \frac{XH^T}{WHH^T} - W \circ \frac{WHH^T}{WHH^T} = W \circ \frac{XH^T}{WHH^T}$$


Gene expression profile per cell
Gene expression program usages per cell
Global catalog of gene expression programs
Cell type 1
Cell type 2
Gene
Activity A
Activity B
≈
X
Identity programs
Activity programs
Gene


Method overview
scRNA-seq data
N (cells)
x
M (genes)
matrix factorization replicates
Gene expression programs (GEPs) (k x M)
(N x k) program usages per cell
Clustering of programs across replicates
Confidence intervals over factorization replicates
Gene expression program (GEP) definition
Per-cell usage of each GEP

- cNMF

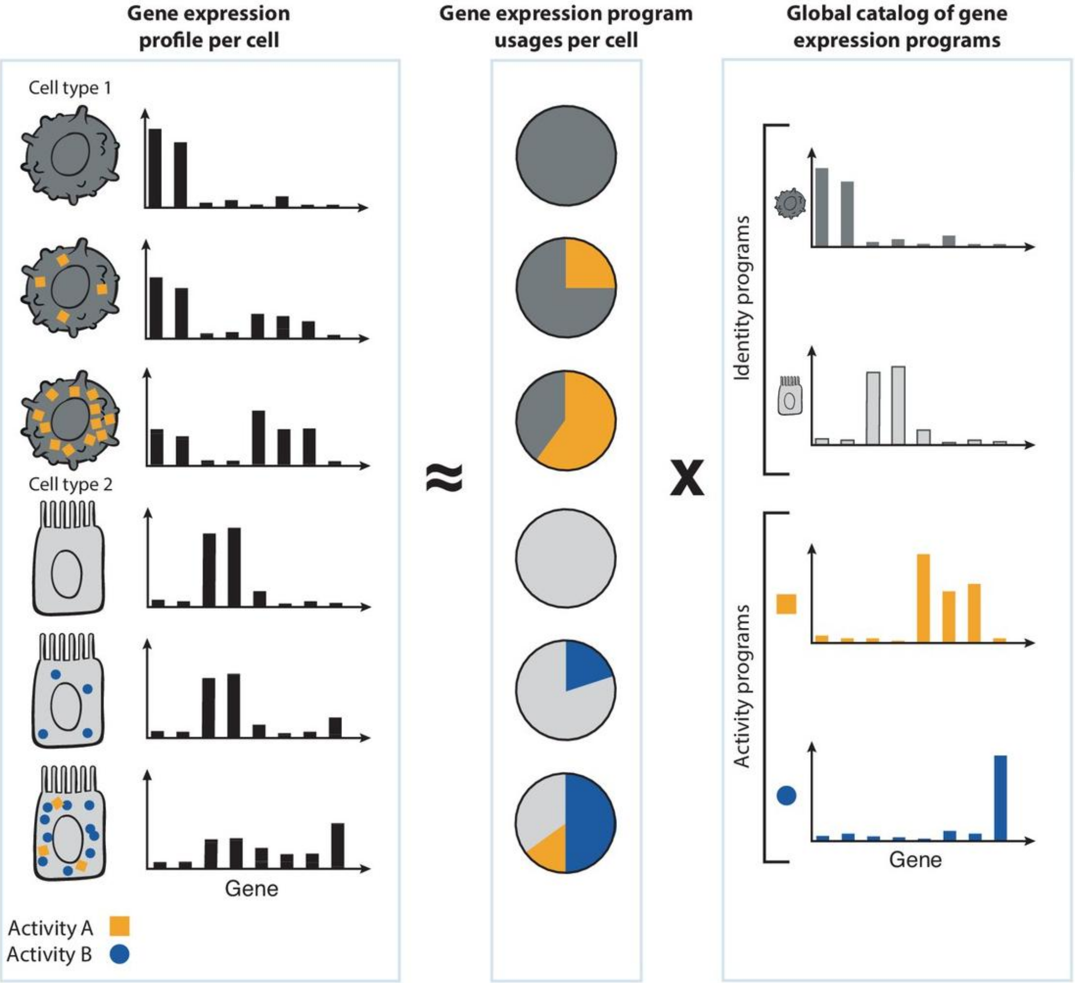
X = WH

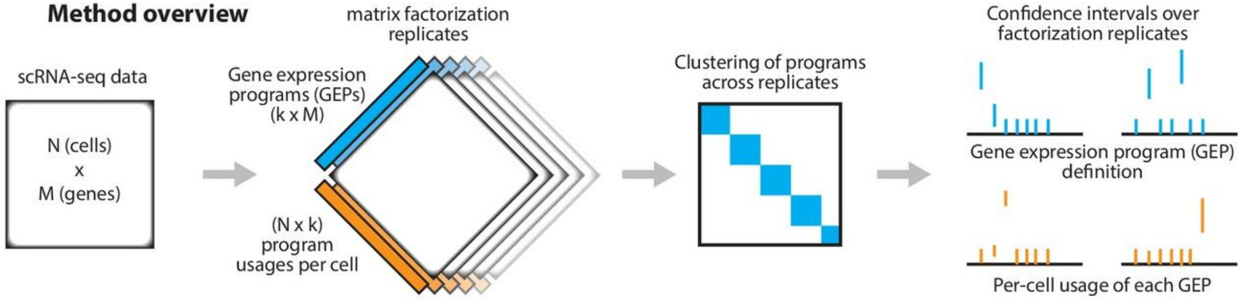
-X: gene expression (N x M)
N: cell, M: gene

-W: program usage (activity): N x k
k: number of program

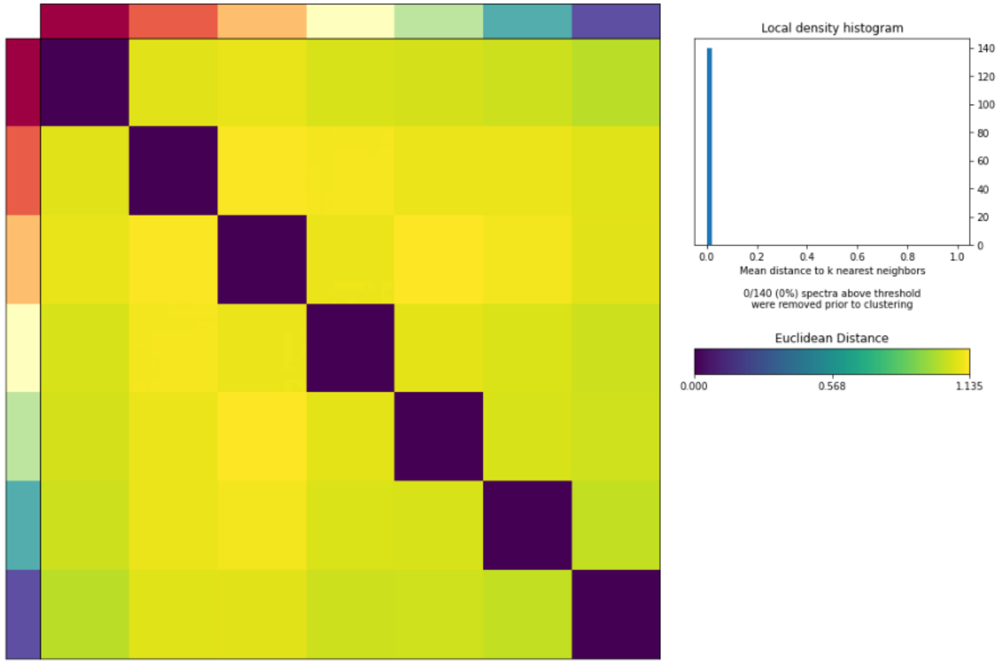
-H: gene expression program: weight of each gene
k x M

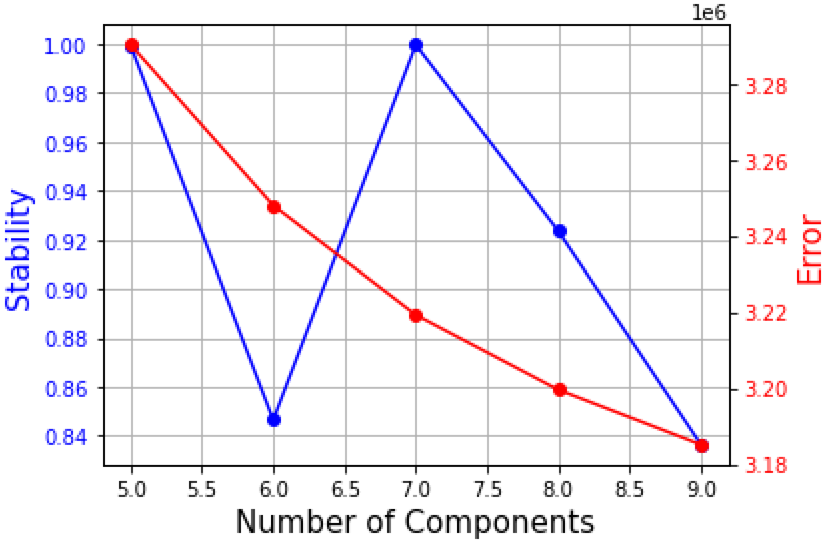
Consensus → robustness
Take median value of each gene


Gene expression profile per cell
Gene expression program usages per cell
Global catalog of gene expression programs
Cell type 1
Cell type 2
Gene
Activity A
Activity B
≈
x
Identity programs
Activity programs
Gene


Method overview
scRNA-seq data
N (cells)
x
M (genes)
matrix factorization replicates
Gene expression programs (GEPs) (k x M)
(N x k) program usages per cell
Clustering of programs across replicates
Confidence intervals over factorization replicates
Gene expression program (GEP) definition
Per-cell usage of each GEP

- cNMF





-K selection → stability: high, error: low

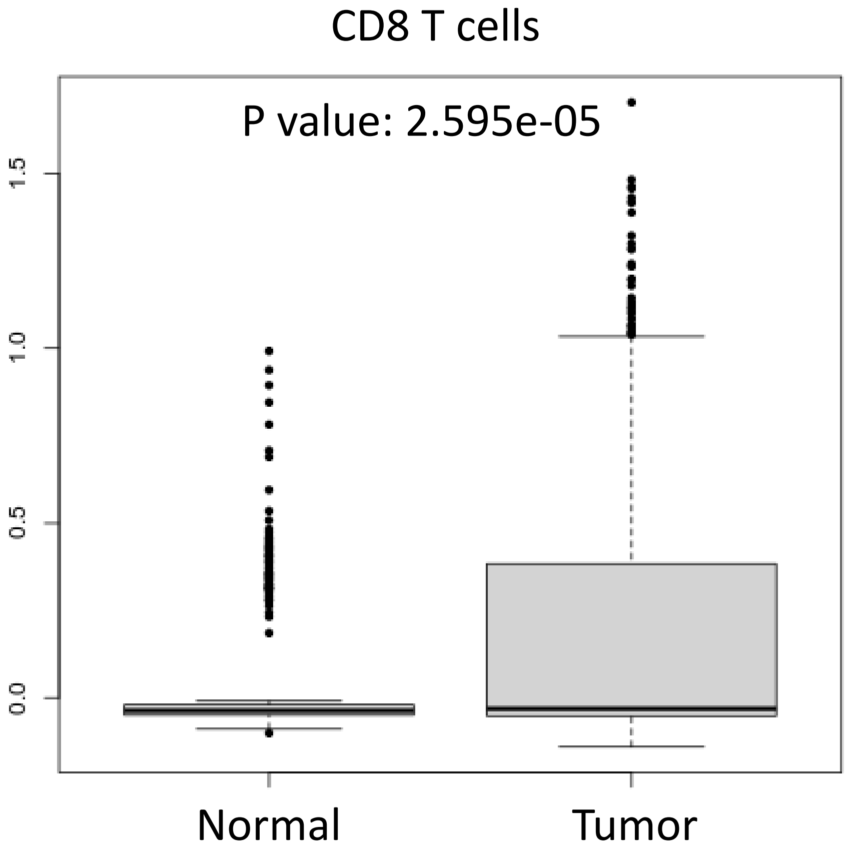
-Define density_threshold by KNN distance distribution

-Batch correction for input count matrix (harmony) moe_correct_ridge (same algorithm in harmony)

- # Signature analysis

Tumor infiltrating T cell → might be exhausted

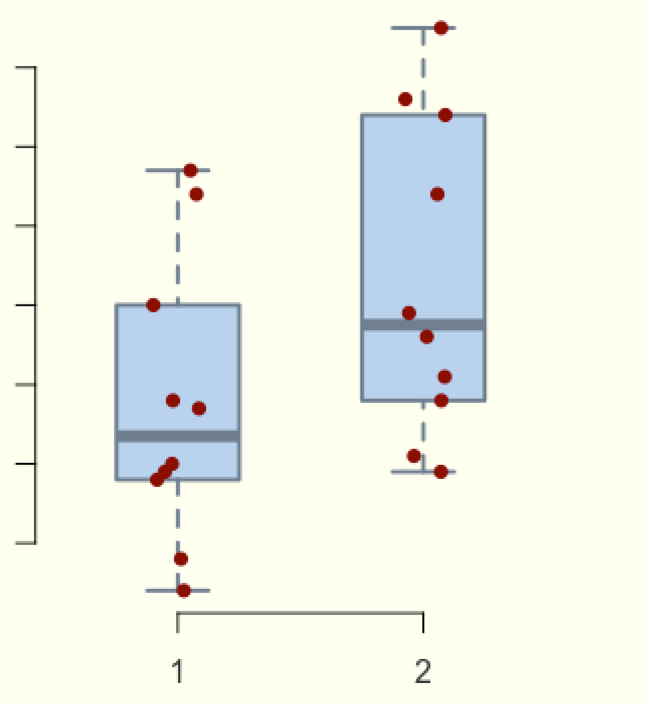
Exhaustion signature: PDCD1, CTLA4, HAVCR2, LAG3, TOX



→ T cells in the tumor-microenvironment are exhausted

- Cell abundance

*T-test, Wilcoxon



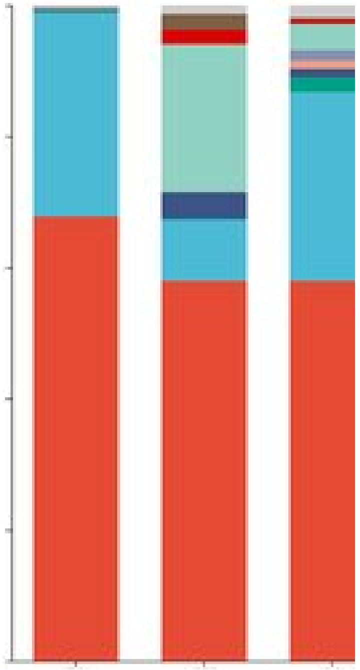
Always! Relative abundance
Why? The cell counts for each sample is always different

Sample size is usually very small for scRNA-seq
→ Poor power analysis (less significant)

*Fisher’s exact test

-Comparing by group-level

-Very sensitive; high false-positive

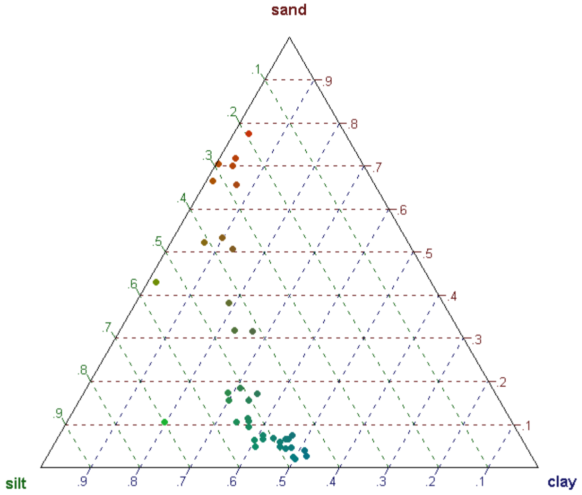


*Dirichlet Regression

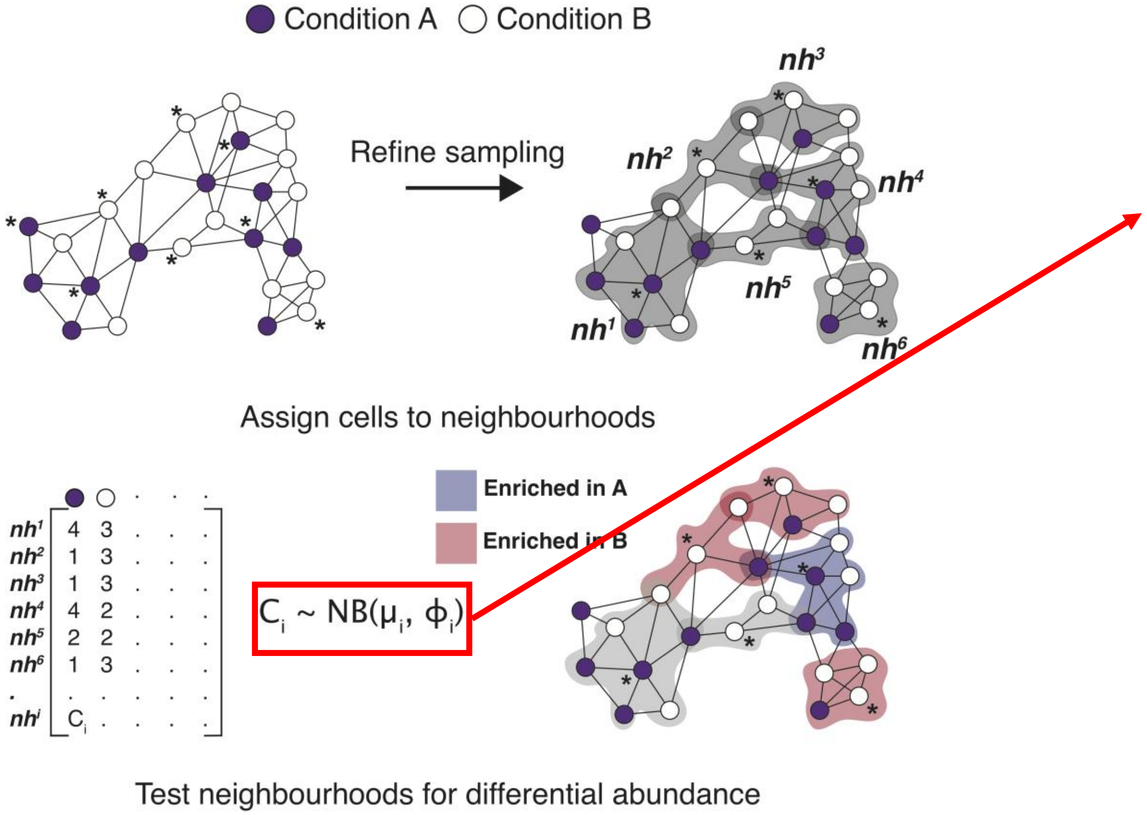
-one celltype ↑→ one celtype ↓

-**prior** reference

celltype selection



- Cell abundance (MILO)

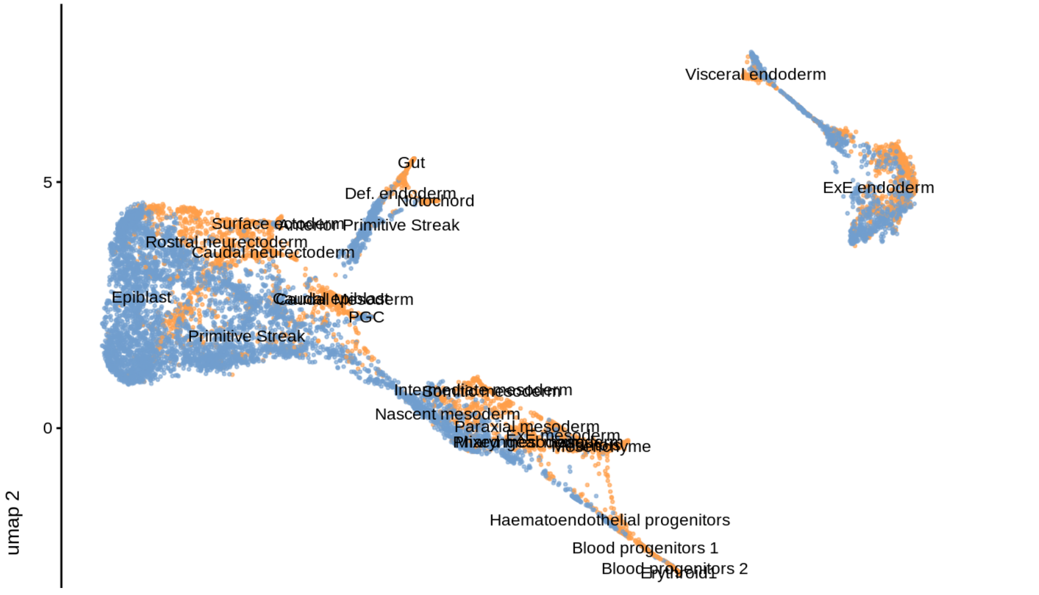

● Condition A ○ Condition B
Refine sampling
nh1
nh2
nh3
nh4
nh5
nh6
Assign cells to neighbourhoods
Enriched in A
Enriched in B
Test neighbourhoods for differential abundance

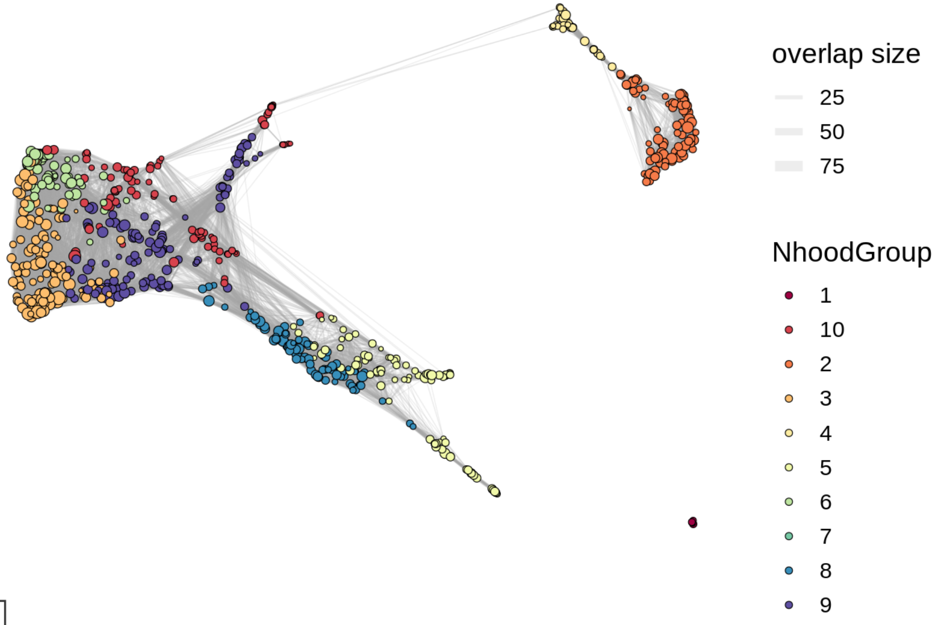
	●	○	.	.	.
nh^1	4	3	.	.	.
nh^2	1	3	.	.	.
nh^3	1	3	.	.	.
nh^4	4	2	.	.	.
nh^5	2	2	.	.	.
nh^6	1	3	.	.	.
.	.	.	.	.	.
nh^i	C_i	.	.	.	.

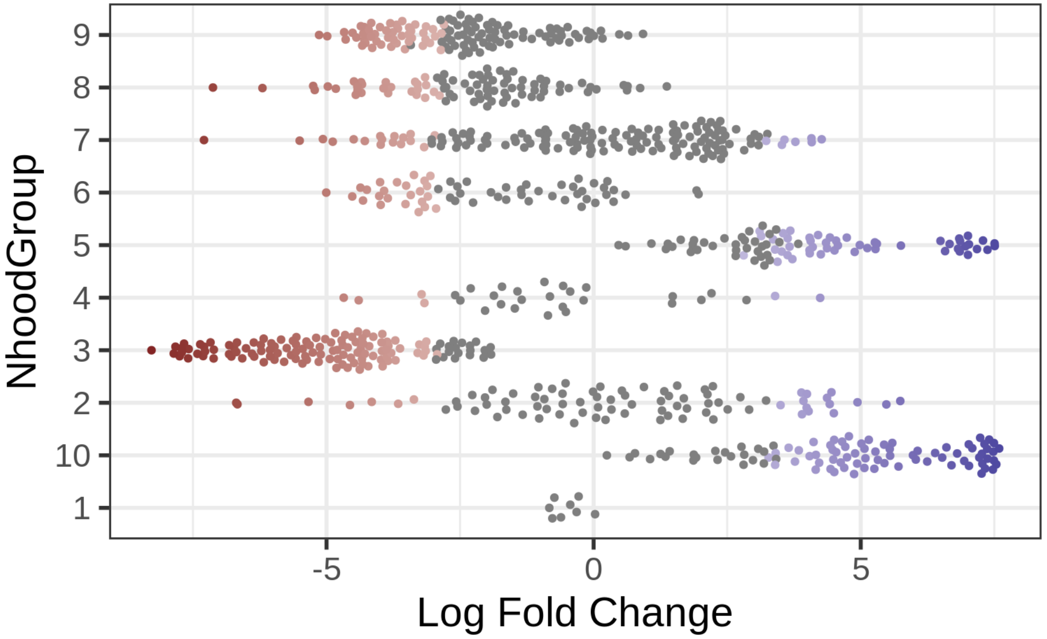
$C_i \sim NB(\mu_i, \phi_i)$

-KNN graph of cells
-Sampling to increase statistical power
-Perform enrichment test for each sampling
Which **condition** has more in the neighborhood

- Cell abundance (MILO)

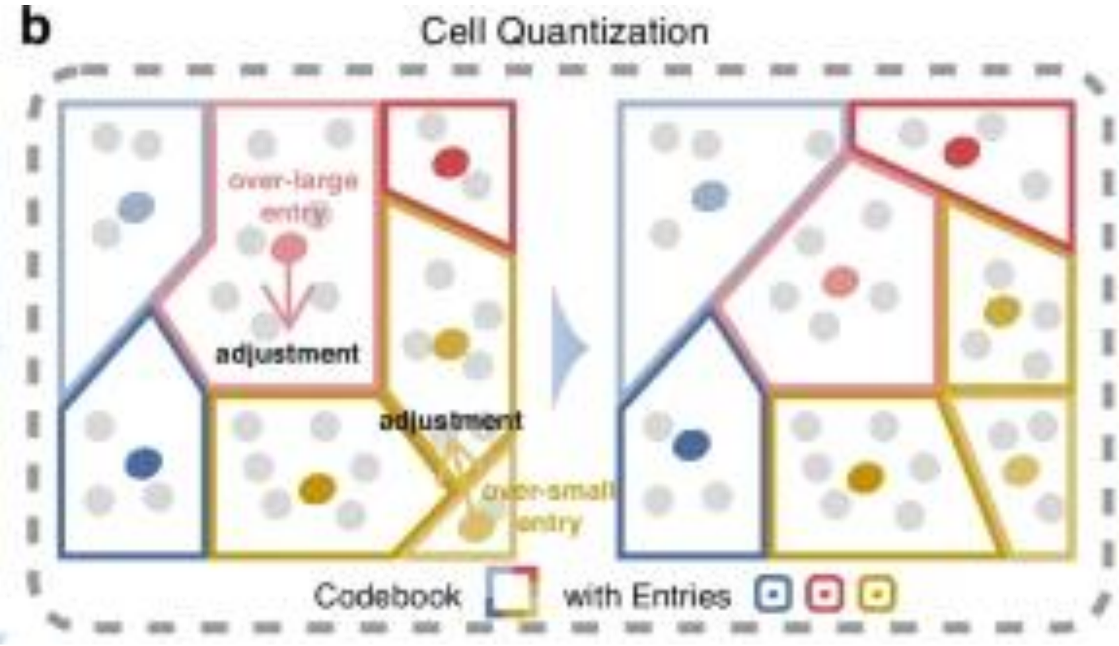






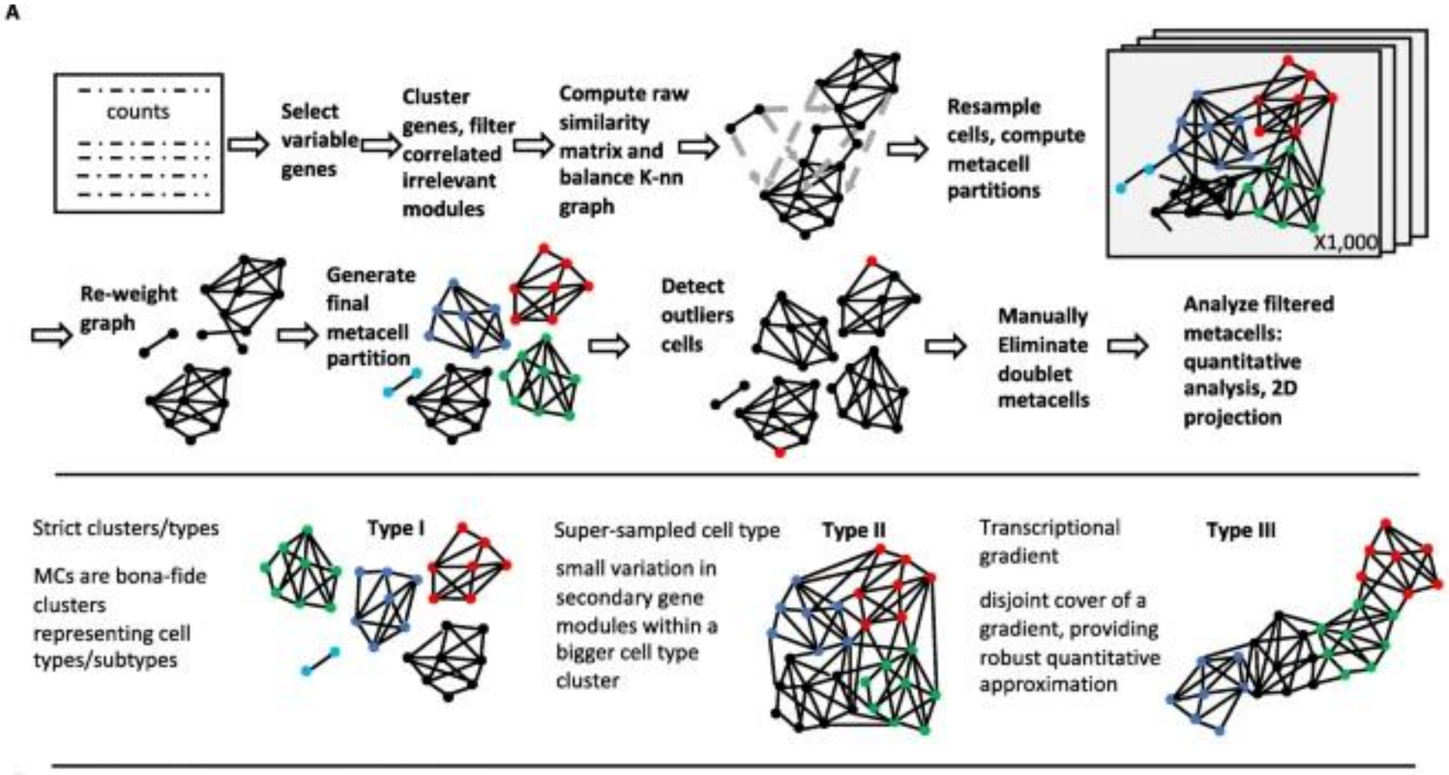
- Cell-pooling

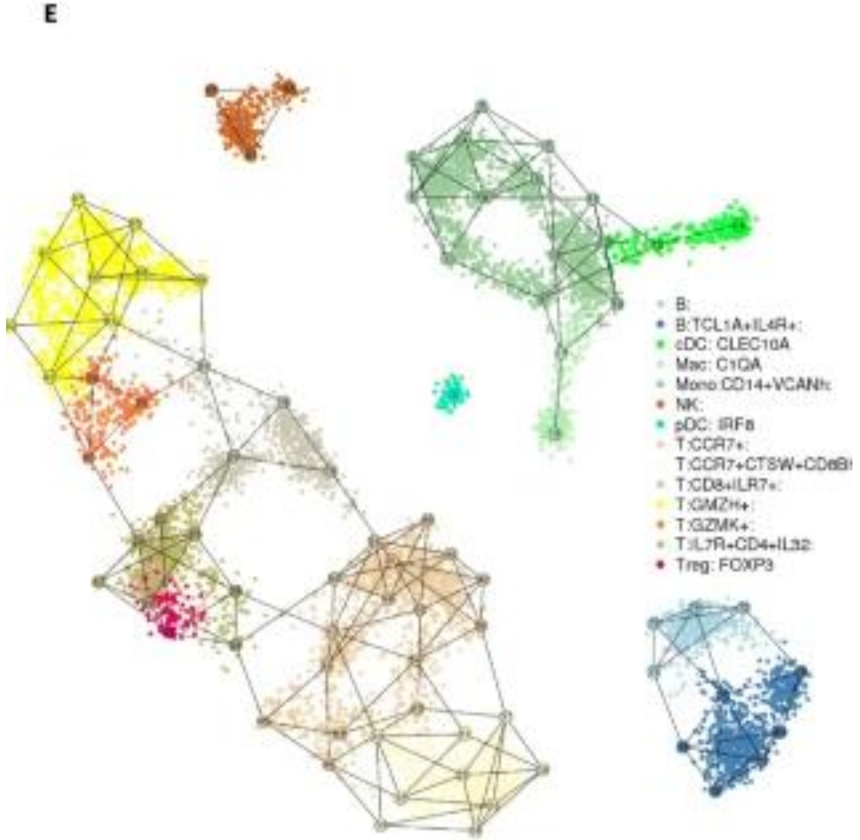
	Cell1	Cell2	…	CellN
Gene1	3	[redacted]	[redacted]	[redacted]
Gene2	[redacted]	3	.	[redacted]
Gene3	[redacted]	14	[redacted]	18
…	.	[redacted]	.	.
…	[redacted]	[redacted]	[redacted]	[redacted]
…	.	.	.	[redacted]
GeneM	25	0	[redacted]	0



-High drop-out rate: zero count ↑
-merge cells → pseudo cell → averaging → overcome drop-out!

- Metacell

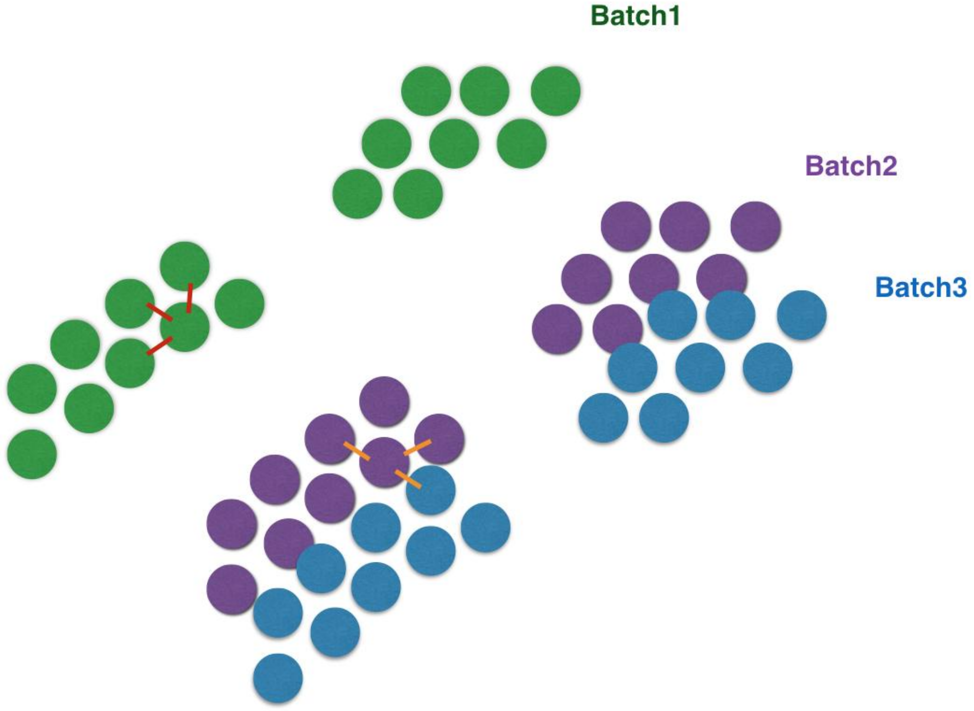

A
counts
Select variable genes
Cluster genes, filter correlated irrelevant modules
Compute raw similarity matrix and balance K-nn graph
Resample cells, compute metacell partitions
X1,000
Re-weight graph
Generate final metacell partition
Detect outliers cells
Manually Eliminate doublet metacells
Analyze filtered metacells: quantitative analysis, 2D projection
Strict clusters/types
MCs are bona-fide clusters representing cell types/subtypes
Type I
Super-sampled cell type
small variation in secondary gene modules within a bigger cell type cluster
Type II
Transcriptional gradient
disjoint cover of a gradient, providing robust quantitative approximation
Type III

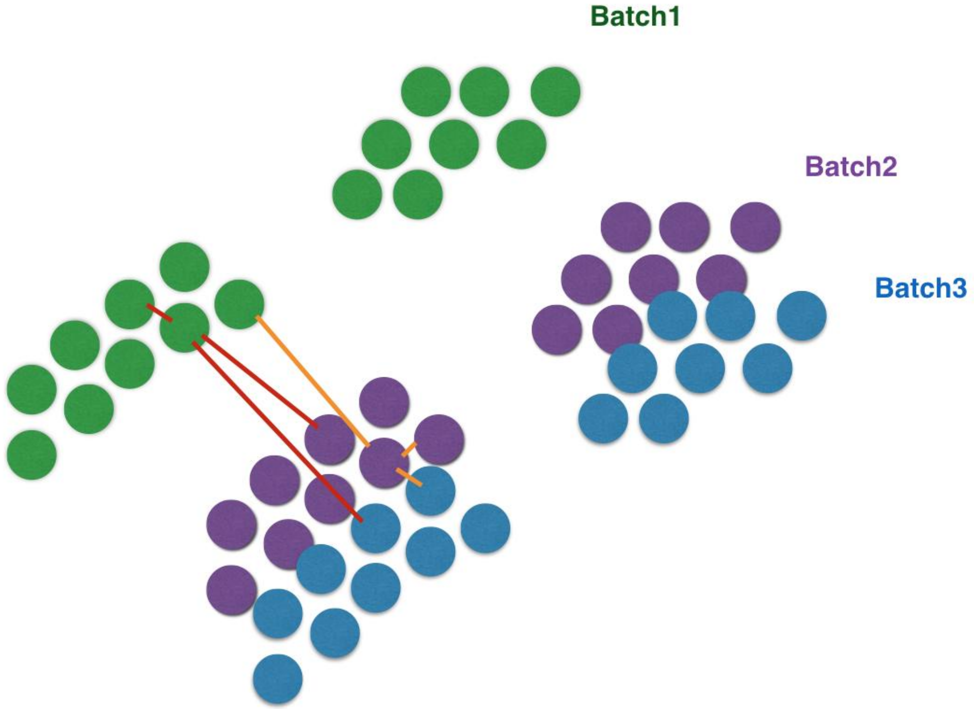


-balanced KNN graph construction
-resampling → consensus-based partitioning
-remove outliers

- # BBKNN

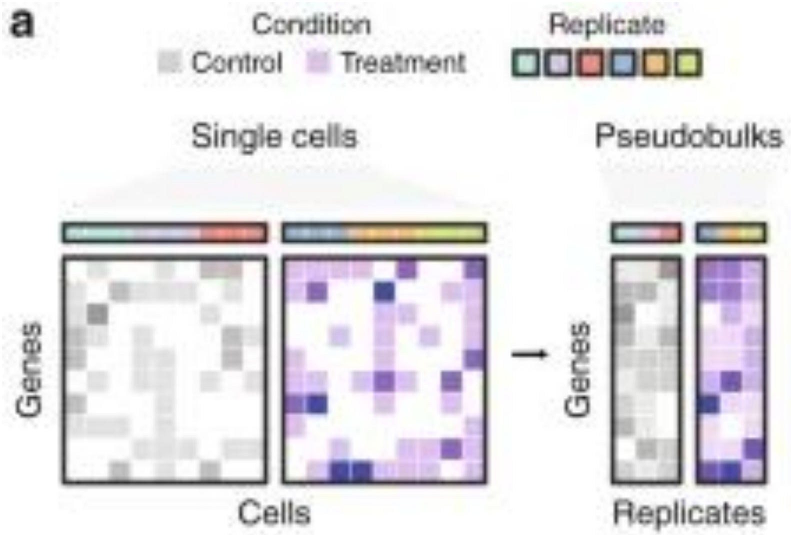
Difference by cell type > difference by batch in a given cell type





- KNN for whole data
- KNN across batch with smaller k
→ Batch corrected pooling

• Pseudobulk DE analysis

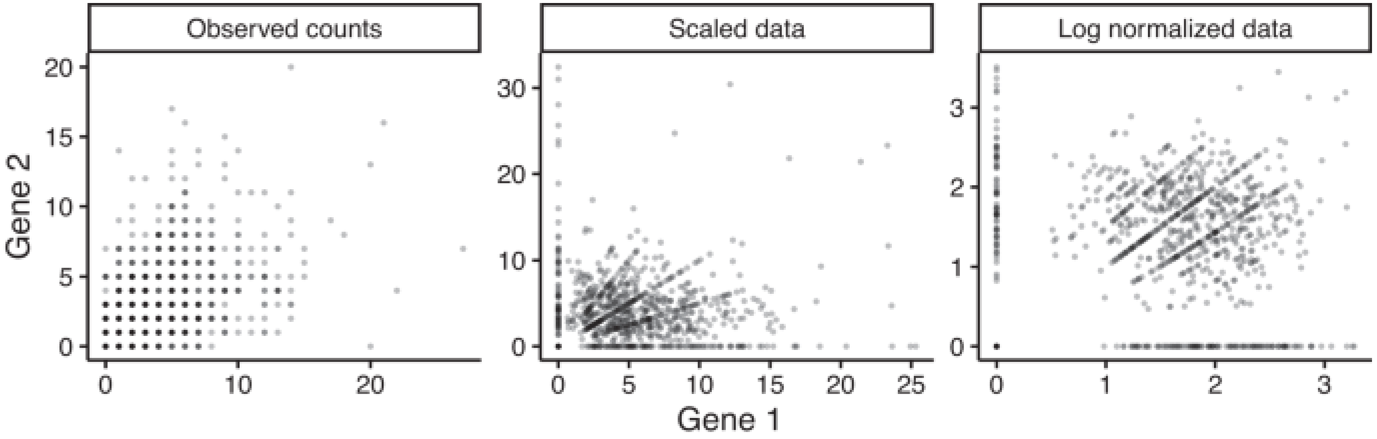


-Extreme case
-Generate a pseudobulk for each sample (each cell type)
→ Perform bulk DE analysis (DESeq2, Limma, edgeR …)

-Overcome high dropout
-susceptible to outlier cells
-cannot account for expressing cell ratio

Confronting false discoveries in single-cell differential expression

• Network analysis in scRNA-seq

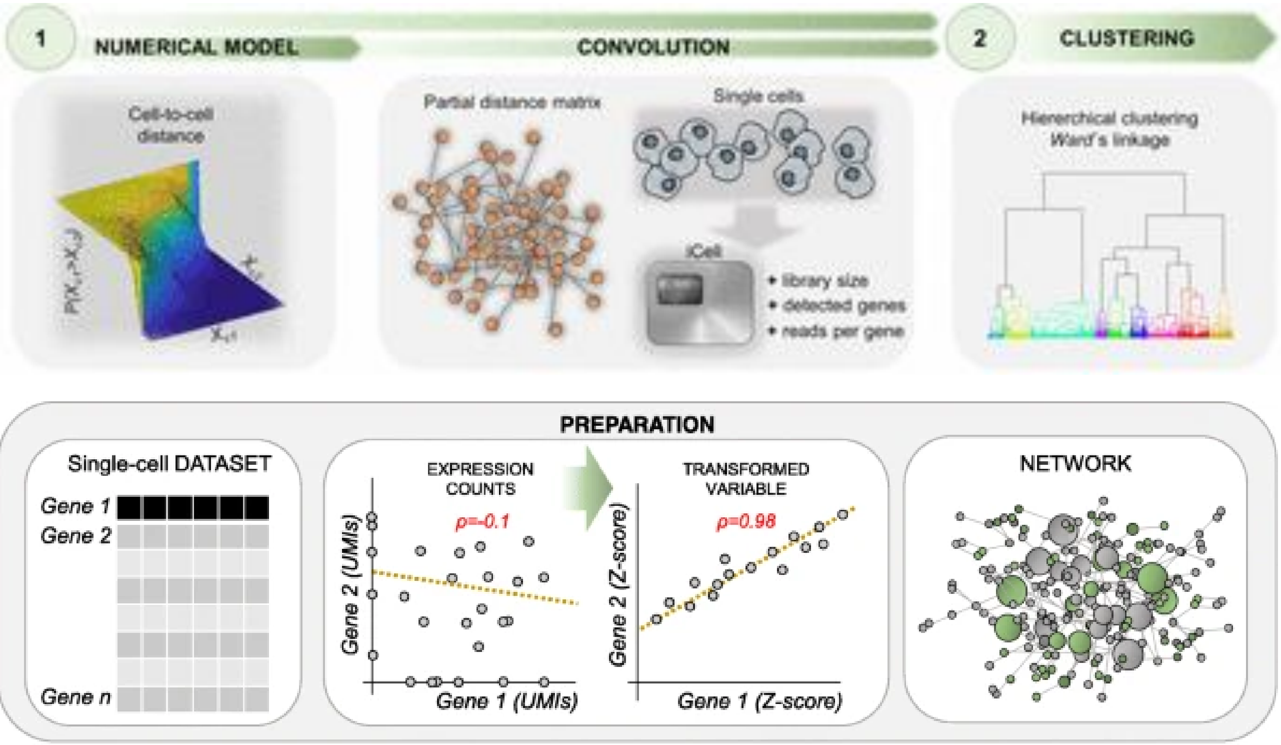


-Skewed too much to zero-counts

→ Hard to obtain a suitable correlation

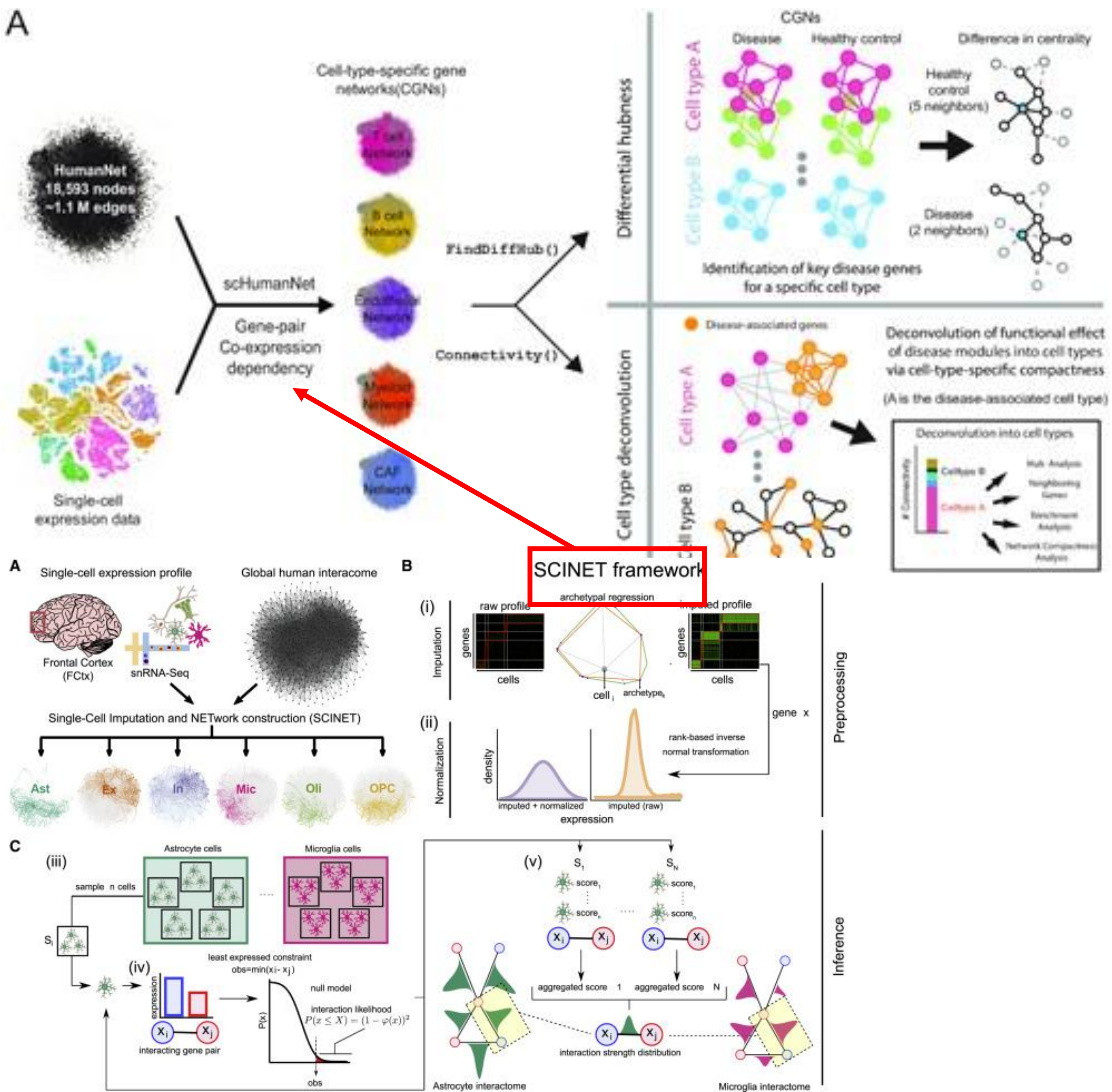
Cell-type-specific co-expression inference from single cell RNA-sequencing data

- BigScale2



-High granularity clustering: Recursive clustering (Hierarchical clustering)
-all pairwise comparison → DE → measure Z-score for each gene
→ Correlation (similar effect of cell-pooling)

- scHumanNet



-Filtering out the “cell-type-specific” network from the reference network

-HumanNetv3, String

*SCINET framework

-clustering → Archetype → transcriptome interpolation (smoothing)

-gene expression transformation

→ Better distribution

-subsampling (per cell type)

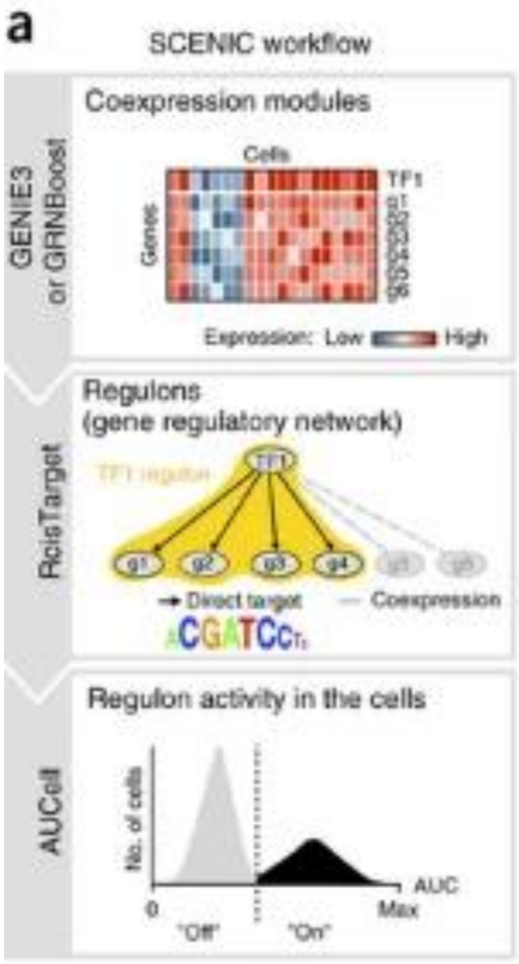
-p-value for interacting gene-pair vs Null

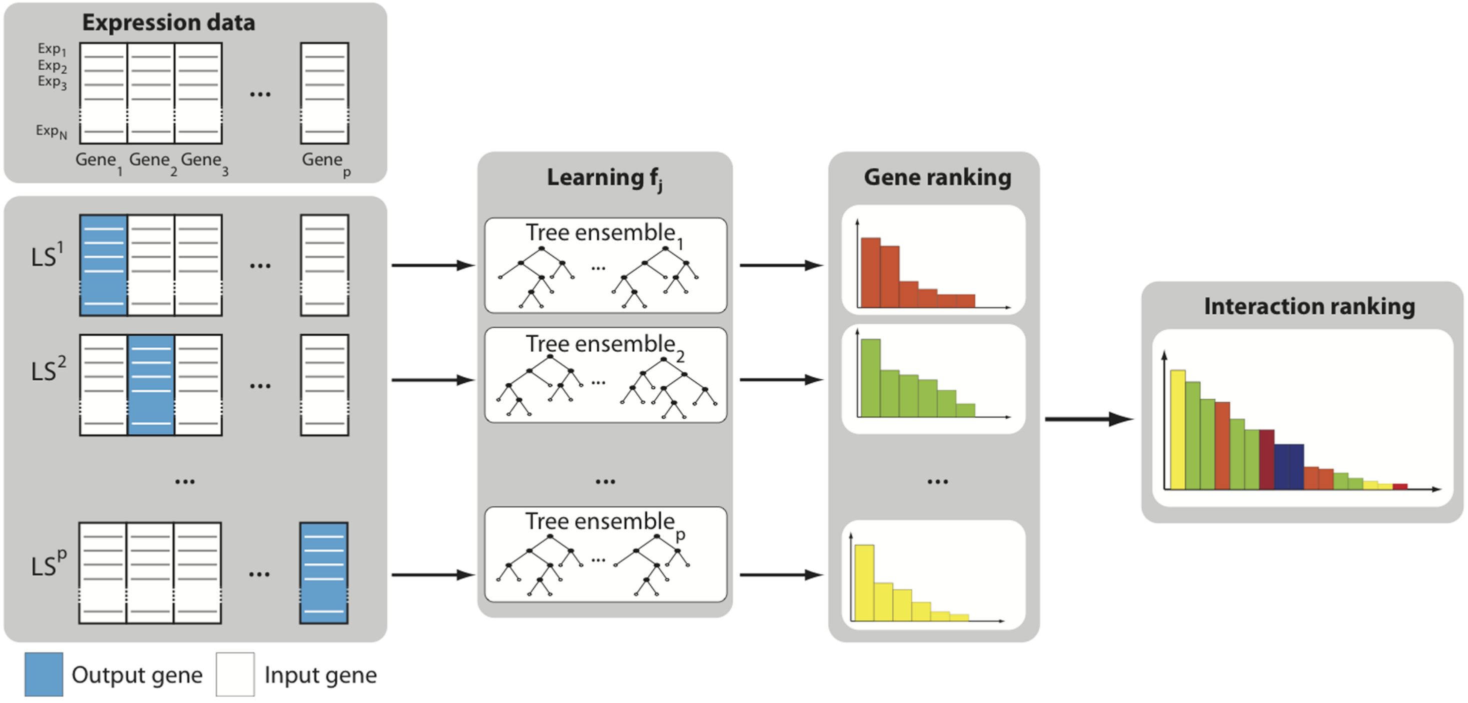
-aggregate p-values by Fisher’s method

Or

Just take the edge and use the original edge score

- SCENIC (genie3)





*Genie3

-Constructing gene-regulatory network (or Transcription Factor regulatory NW: TRN)

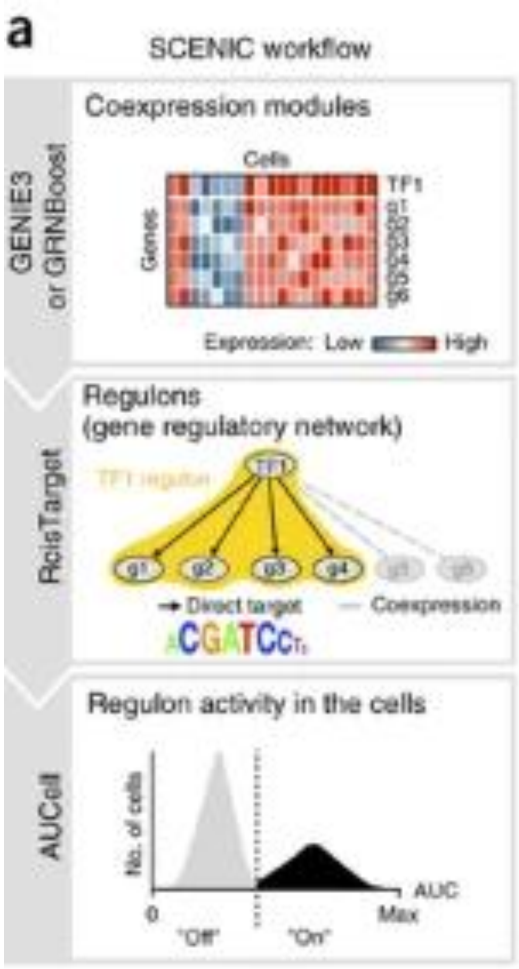
-Based on Genie3 (developed for bulk RNA data)

-output gene exp ← explained by input genes (random forest) [coexpression pattern] → TF filtering

-output gene (i) ← input gene (j1, j2, j3) score → ranking (by importance)

→ GRNBoost for speed

- SCENIC (genie3)



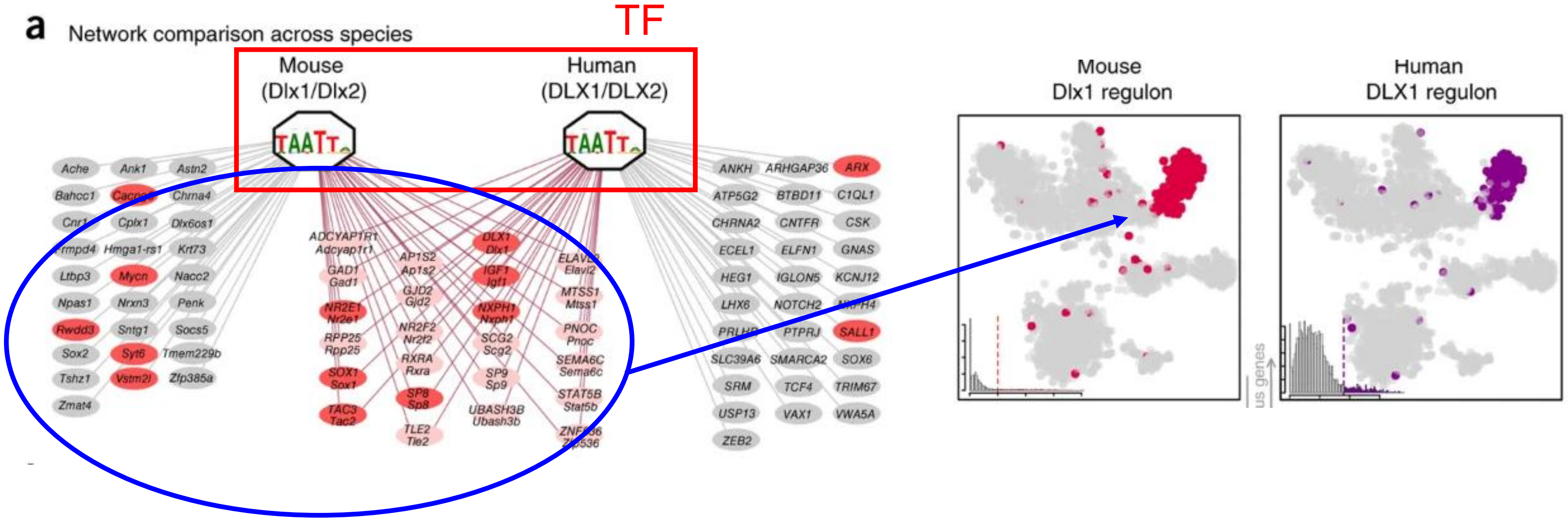
-We obtained TF- target gene (correlation or association)

→ Not all the TF binds to the target gene

→ RcisTarget: cis-regulatory motif analysis

→ Only enriched TF can bind to the promoter of a given genes

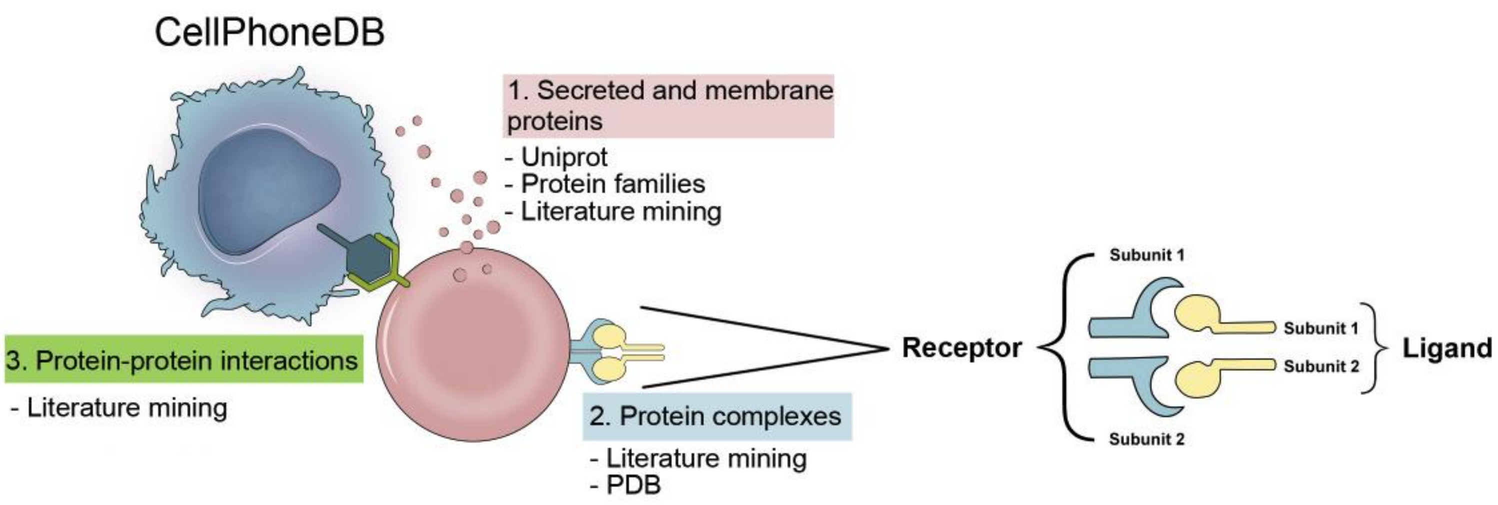
*Regulon activity: AUCell



TF

Regulon

Cell-Cell interaction (CellPhoneDB, CellChat)



-Cluster-to-Cluster interaction

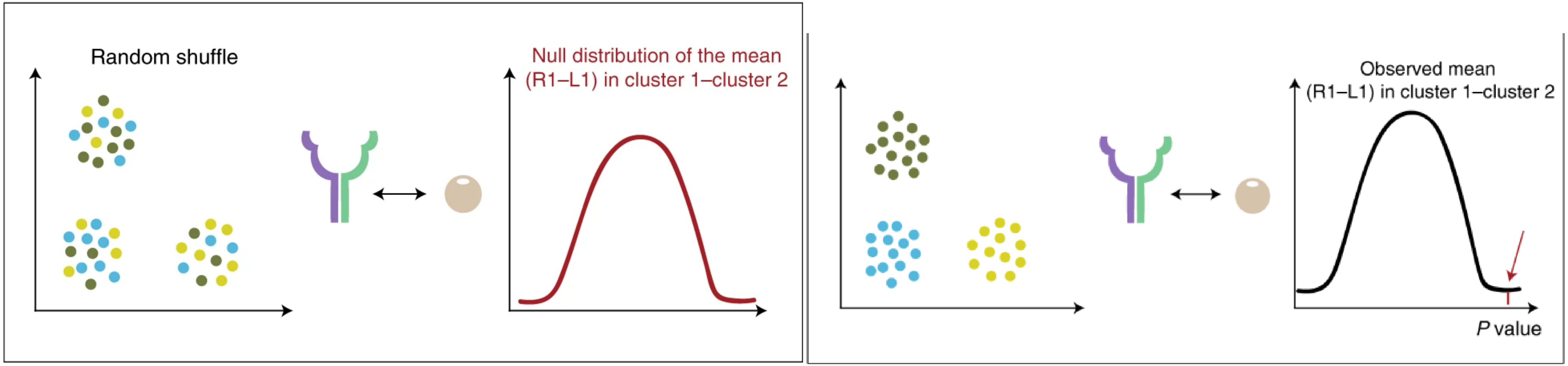
→ Measure the mean expression of receptor-ligand pair

→ Shuffle the cells: Null distribution

→ P-value measurement for each pair

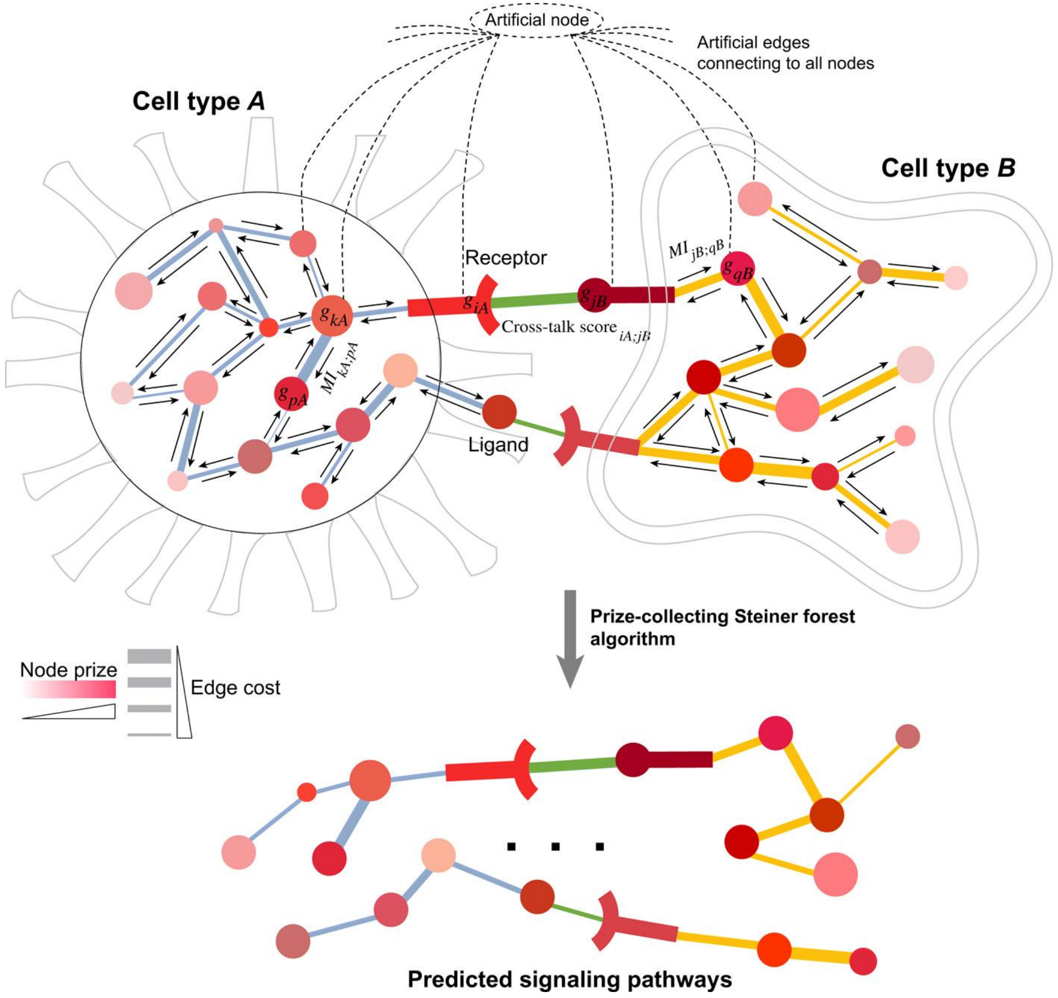
→ Strength: mean expression

+ complex: mean expression



Random shuffle
Null distribution of the mean
(R1–L1) in cluster 1–cluster 2
Observed mean
(R1–L1) in cluster 1–cluster 2
P value

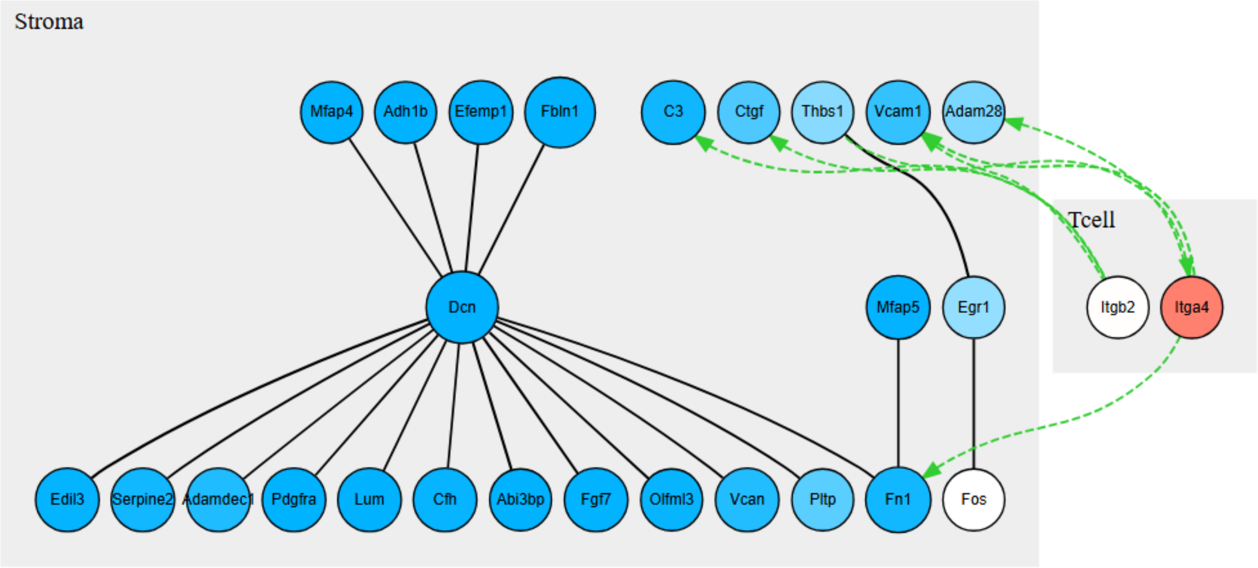
- Cell-Cell interaction (CytoTalk)



What is the downstream of interaction?

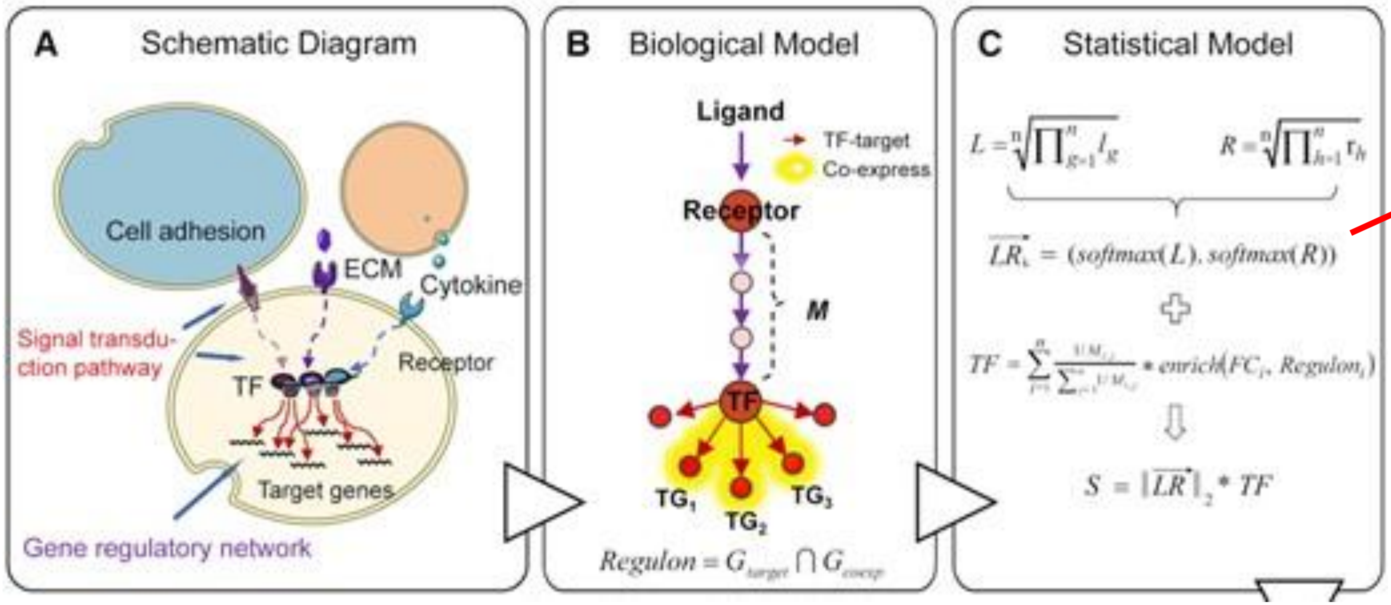
-Cell-cell interaction: intercellular interaction
-Intra-cellular network: mutual information (co-occurrence of gene expression across cells)

-Network propagation algorithm
→ Remain only the significant edges



Cell-Cell interaction (CellCall)

Framework of CellCall

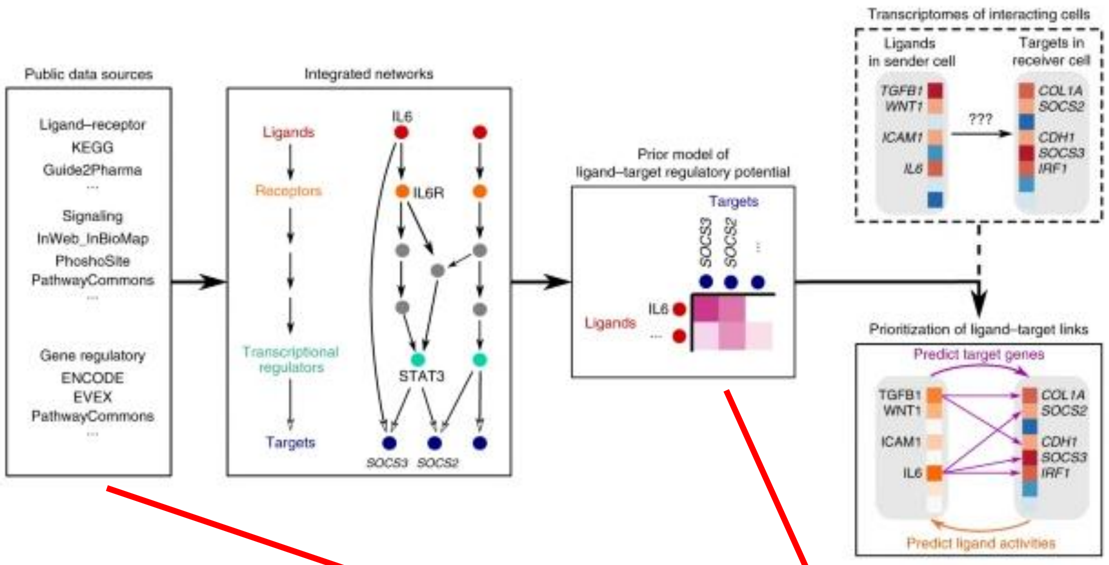

A
Schematic Diagram
Cell adhesion
ECM
Cytokine
Signal transdu-ction pathway
TF
Receptor
Target genes
Gene regulatory network
B
Biological Model
Ligand
TF-target
Co-express
Receptor
M
TF
TG_1
TG_2
TG_3
$Regulon = G_{target} \cap G_{coexp}$
C
Statistical Model
$L = \sqrt[n]{\prod_{g=1}^{n} l_g}$
$R = \sqrt[n]{\prod_{h=1}^{n} r_h}$
$\overrightarrow{LR_i} = (softmax(L), softmax(R))$
$TF = \sum_{i=1}^{n} \frac{1/M_{i,j}}{\sum 1/M_{i,j}} * enrich(FC_i, Regulon_i)$
$S = \|\overrightarrow{LR}\|_2 * TF$

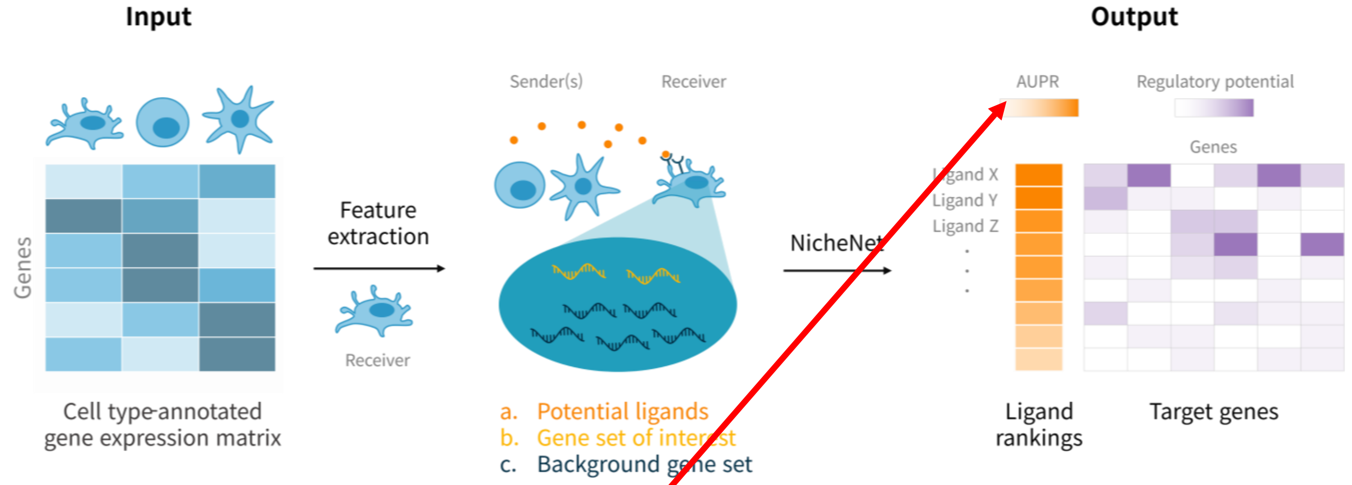
LR: conceptually mean expression

Only looking at the gene expression from the ligand-receptor is not enough
It should show some **perturbation of target genes** due to cell-cell interaction!

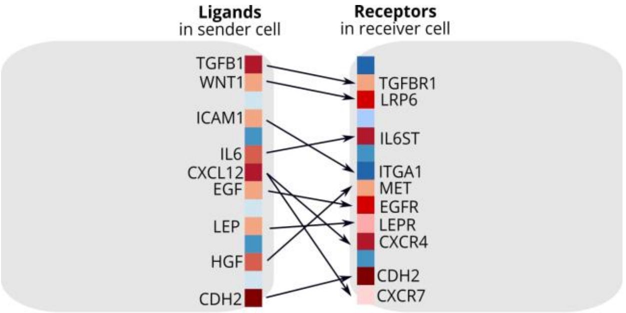
Interaction score = LR * TF_k
TF_k: regulon activity of TF
-LR → TF: KEGG, etc
-TF → regulon: known DB (TRRUST …) & coexpressed with TF
-TF_k: GSEA for those regulon
-Multiple TF_k : weight sum (number of node; LR → TF)

• Cell-Cell interaction (NicheNet)





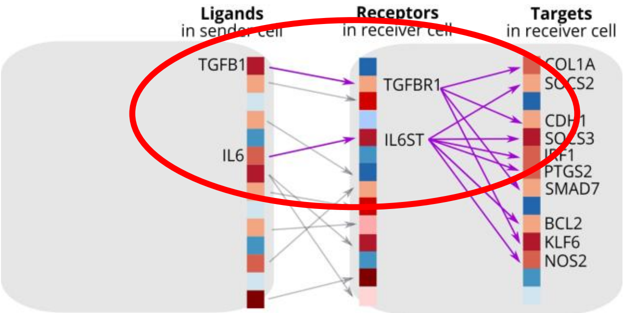
Typical CCC



-Merge each path from DB: weighted network (prior model)
Ligand – receptor – TF – target genes

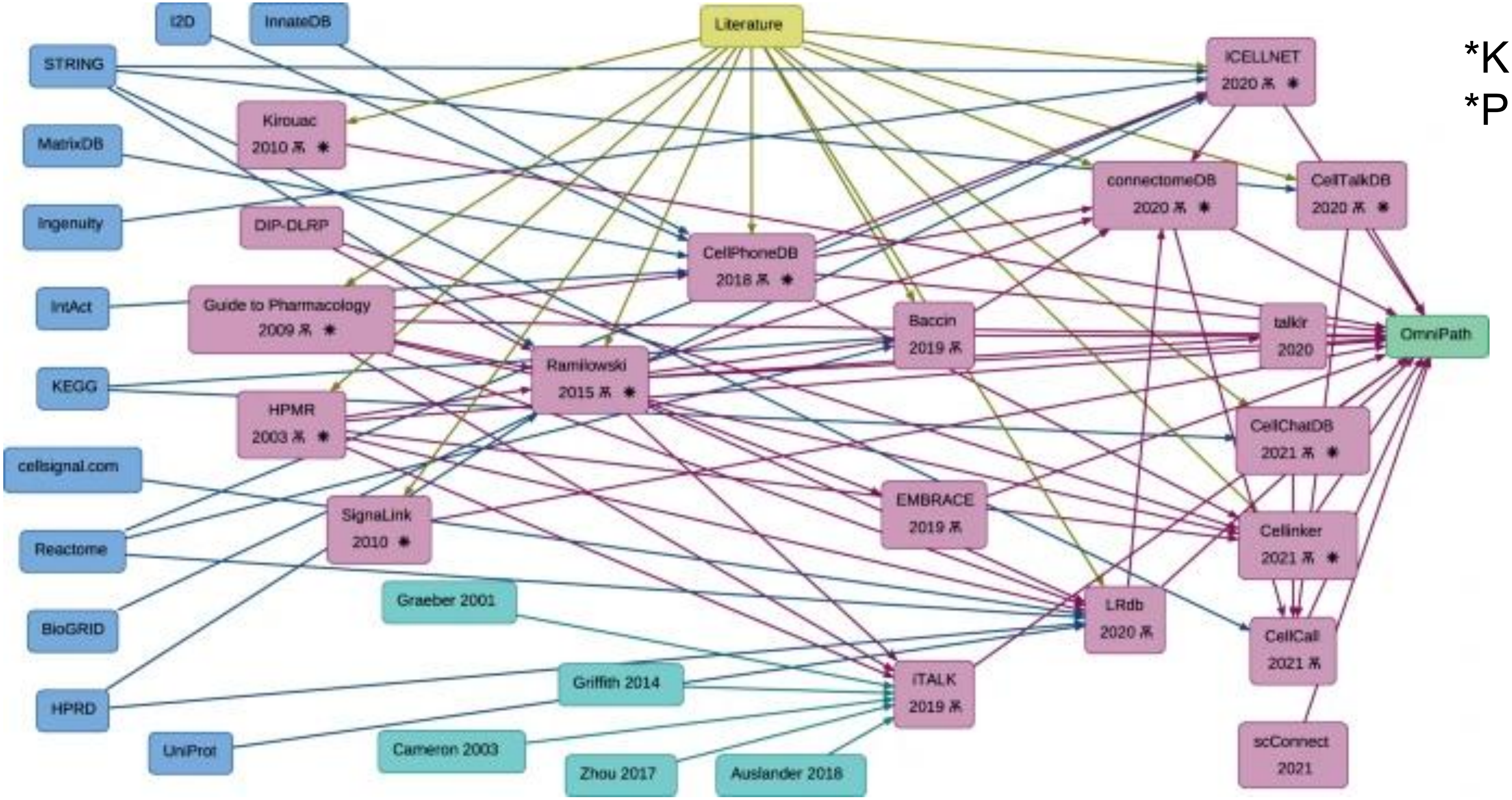
-ligand ~ potential target genes vs bg genes
-ligand ranking: ligand [exp] ~ predefined targets [gene exp]
(how much ligand expression can differentially express target genes)
-target genes are selected by a predefined ligand-target link

NicheNet



- Cell-Cell interaction comparison

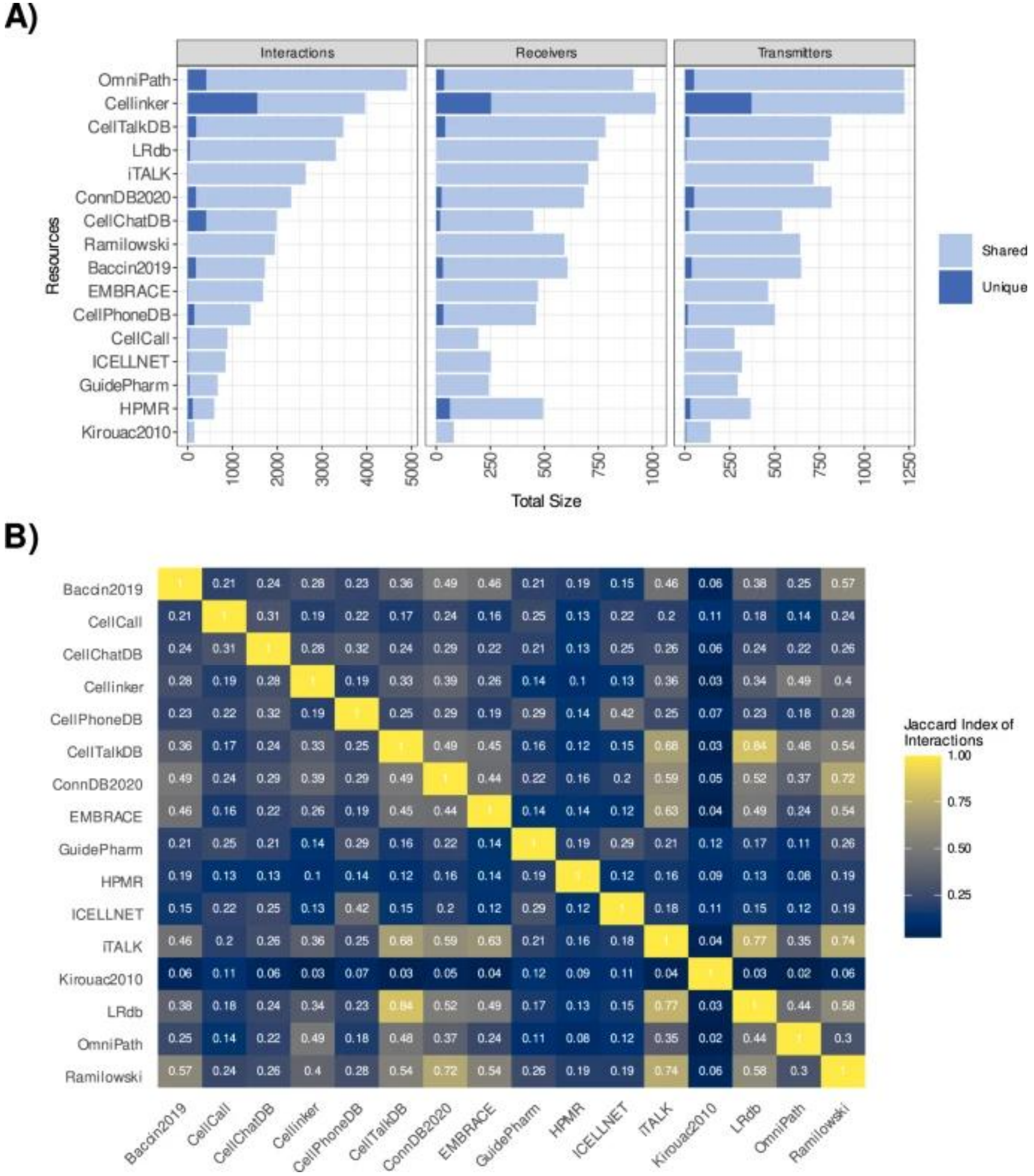
-Heterogeneous DB



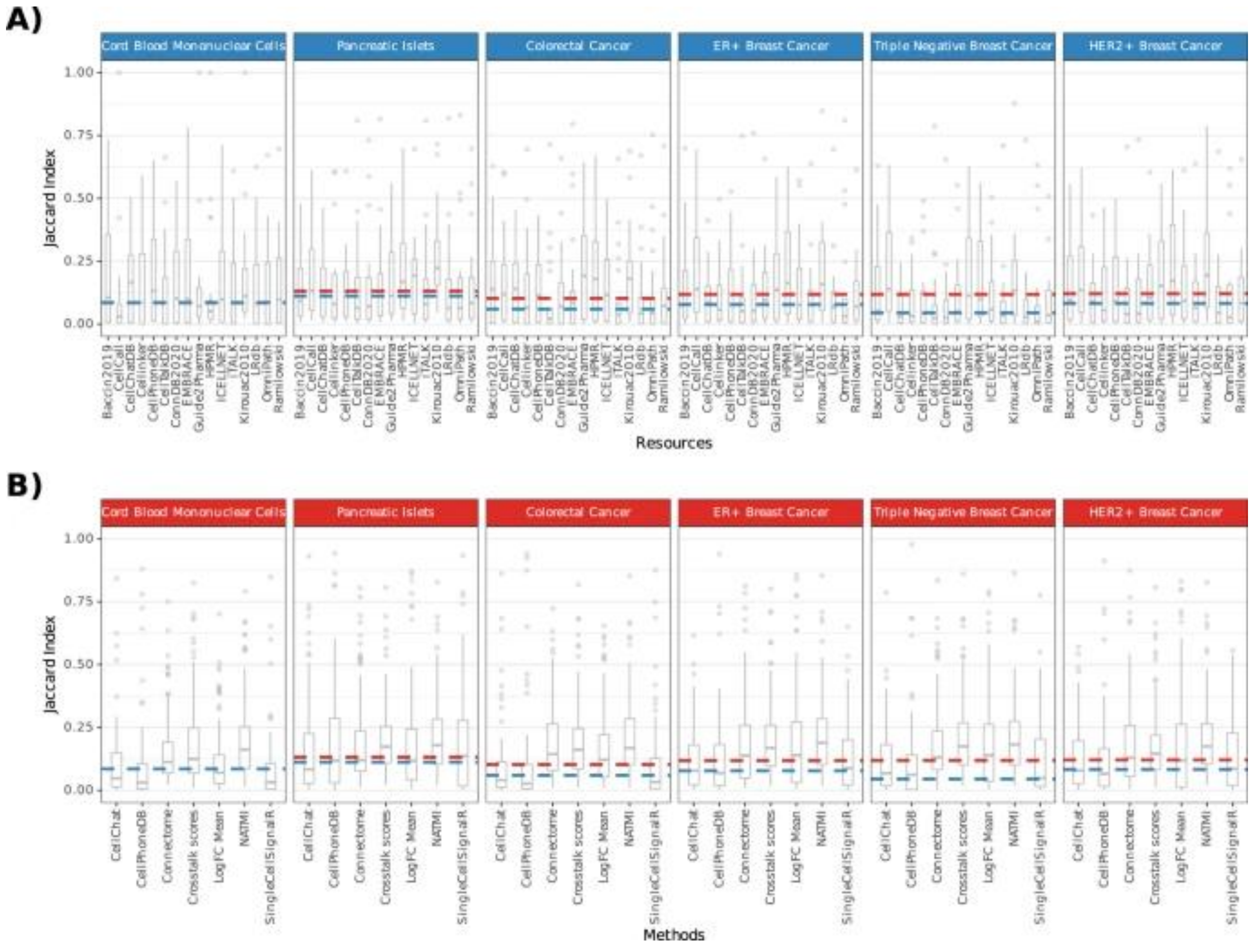
*KEGG, Reactome, STRING
*Published literature

Cell-Cell interaction comparison

-Different CCC methods are too different

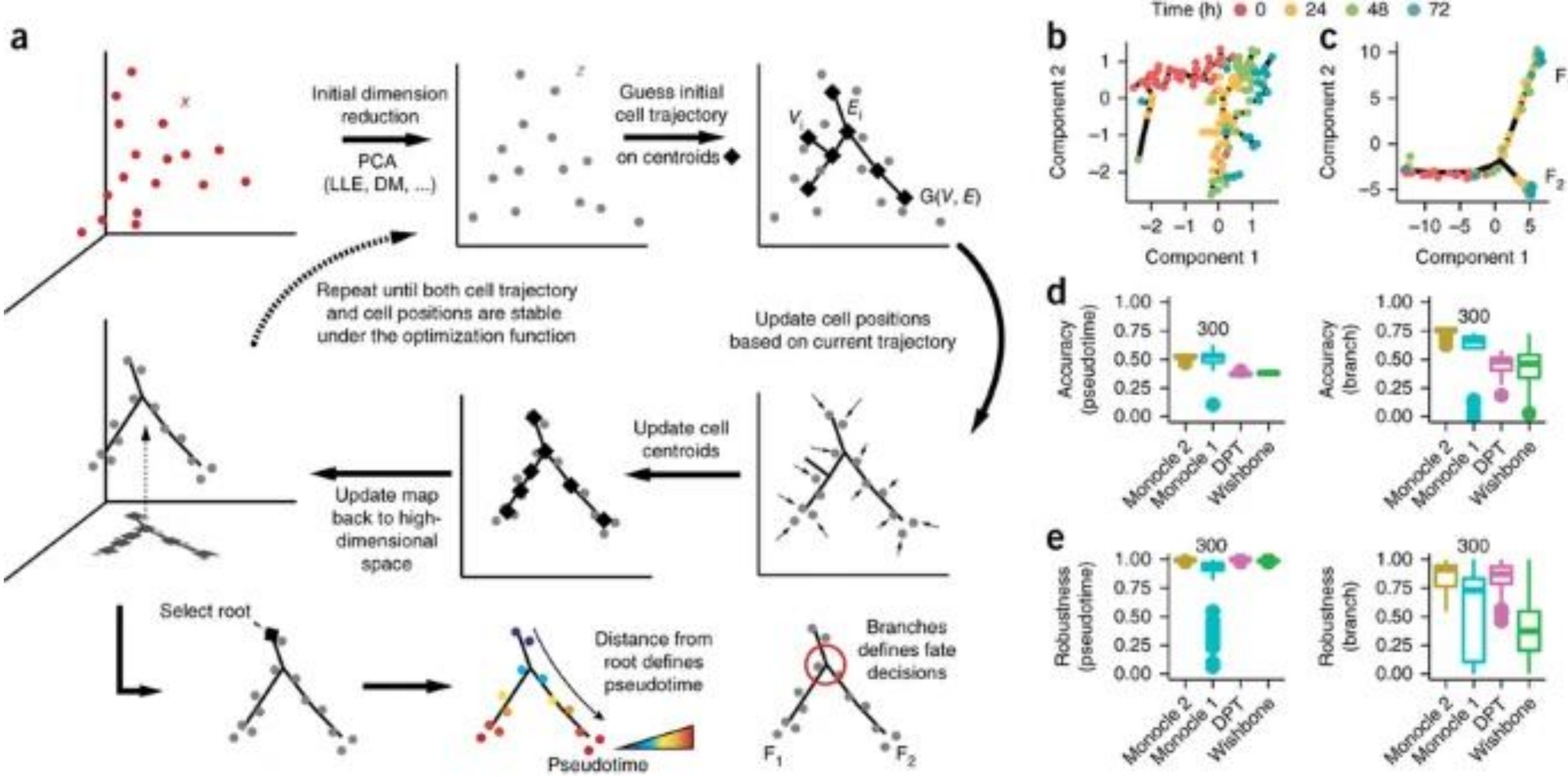


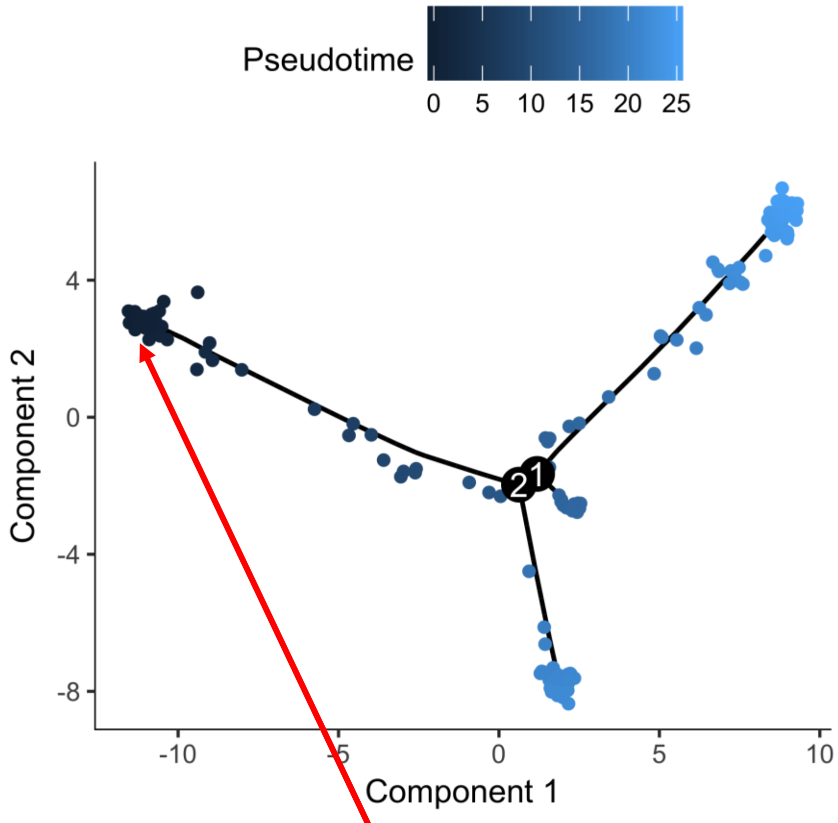
	Bacdn2019	CellCall	CellChatDB	Cellinker	CellPhoneDB	CellTalkDB	ConnDB2020	EMBRACE	GuidePharm	HPMR	ICELLNET	iTALK	Kirouac2010	LRdb	OmniPath	Ramilowski
Bacdn2019	1	0.21	0.24	0.28	0.23	0.36	0.49	0.46	0.21	0.19	0.15	0.46	0.06	0.38	0.25	0.57
CellCall	0.21	1	0.31	0.19	0.22	0.17	0.24	0.16	0.25	0.13	0.22	0.2	0.11	0.18	0.14	0.24
CellChatDB	0.24	0.31	1	0.28	0.32	0.24	0.29	0.22	0.21	0.13	0.25	0.26	0.06	0.24	0.22	0.26
Cellinker	0.28	0.19	0.28	1	0.19	0.33	0.39	0.26	0.14	0.1	0.13	0.36	0.03	0.34	0.49	0.4
CellPhoneDB	0.23	0.22	0.32	0.19	1	0.25	0.29	0.19	0.29	0.14	0.42	0.25	0.07	0.23	0.18	0.28
CellTalkDB	0.36	0.17	0.24	0.33	0.25	1	0.49	0.45	0.16	0.12	0.15	0.68	0.03	0.84	0.48	0.54
ConnDB2020	0.49	0.24	0.29	0.39	0.29	0.49	1	0.44	0.22	0.16	0.2	0.59	0.05	0.52	0.37	0.72
EMBRACE	0.46	0.16	0.22	0.26	0.19	0.45	0.44	1	0.14	0.14	0.12	0.63	0.04	0.49	0.24	0.54
GuidePharm	0.21	0.25	0.21	0.14	0.29	0.16	0.22	0.14	1	0.19	0.29	0.21	0.12	0.17	0.11	0.26
HPMR	0.19	0.13	0.13	0.1	0.14	0.12	0.16	0.14	0.19	1	0.12	0.16	0.09	0.13	0.08	0.19
ICELLNET	0.15	0.22	0.25	0.13	0.42	0.15	0.2	0.12	0.29	0.12	1	0.18	0.11	0.15	0.12	0.19
iTALK	0.46	0.2	0.26	0.36	0.25	0.68	0.59	0.63	0.21	0.16	0.18	1	0.04	0.77	0.35	0.74
Kirouac2010	0.06	0.11	0.06	0.03	0.07	0.03	0.05	0.04	0.12	0.09	0.11	0.04	1	0.03	0.02	0.06
LRdb	0.38	0.18	0.24	0.34	0.23	0.84	0.52	0.49	0.17	0.13	0.15	0.77	0.03	1	0.44	0.58
OmniPath	0.25	0.14	0.22	0.49	0.18	0.48	0.37	0.24	0.11	0.08	0.12	0.35	0.02	0.44	1	0.3
Ramilowski	0.57	0.24	0.26	0.4	0.28	0.54	0.72	0.54	0.26	0.19	0.19	0.74	0.06	0.58	0.3	1



Low Jaccard index (low overlap)
-across DB
-across method

- Pseudotime analysis (Monocle2)


a
Initial dimension reduction
PCA (LLE, DM, ...)
Guess initial cell trajectory on centroids ◆
V_i
E_i
G(V, E)
Repeat until both cell trajectory and cell positions are stable under the optimization function
Update cell positions based on current trajectory
Update cell centroids
Update map back to high-dimensional space
Select root
Distance from root defines pseudotime
Pseudotime
Branches defines fate decisions
F_1
F_2
b
c
Time (h) 0 24 48 72
Component 1
Component 2
F_1
F_2
d
e
Accuracy (pseudotime)
Accuracy (branch)
Robustness (pseudotime)
Robustness (branch)
300
Monocle 2
Monocle 1
DPT
Wishbone

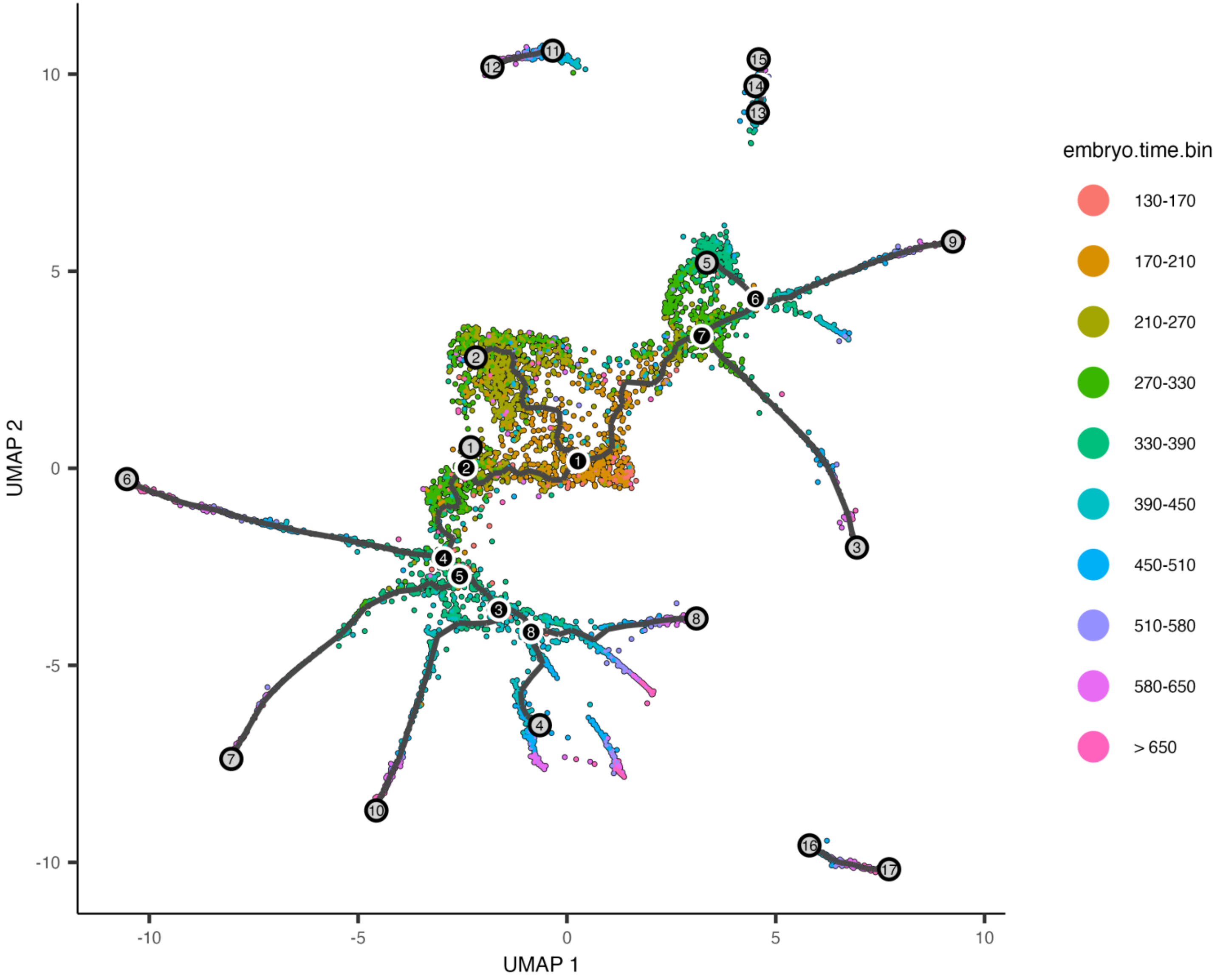


*Require root cell or cluster

Aligning cells into a virtual embedding
-Dimension reduction → initial centroid (k-mean)
→ update cell position
→ High dimension → re-do until convergence

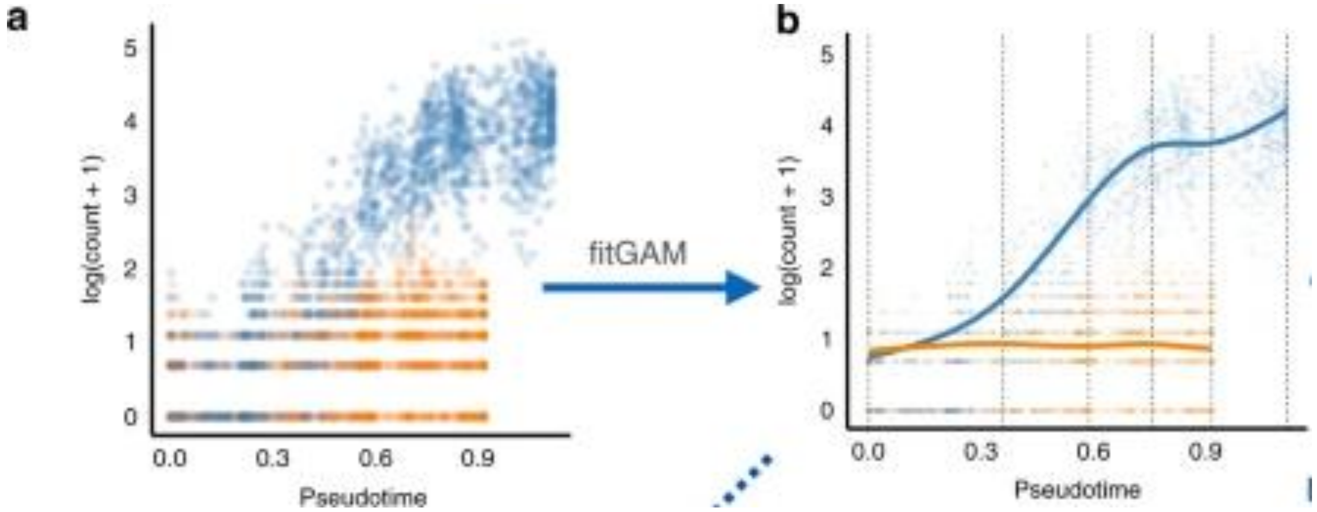
Conceptually: clustering + network construction
Gene: DEG (across differentiation), HVG …

- Pseudotime analysis (Monocle3)



Adapt dimension reduction space into a familiar UMAP projection

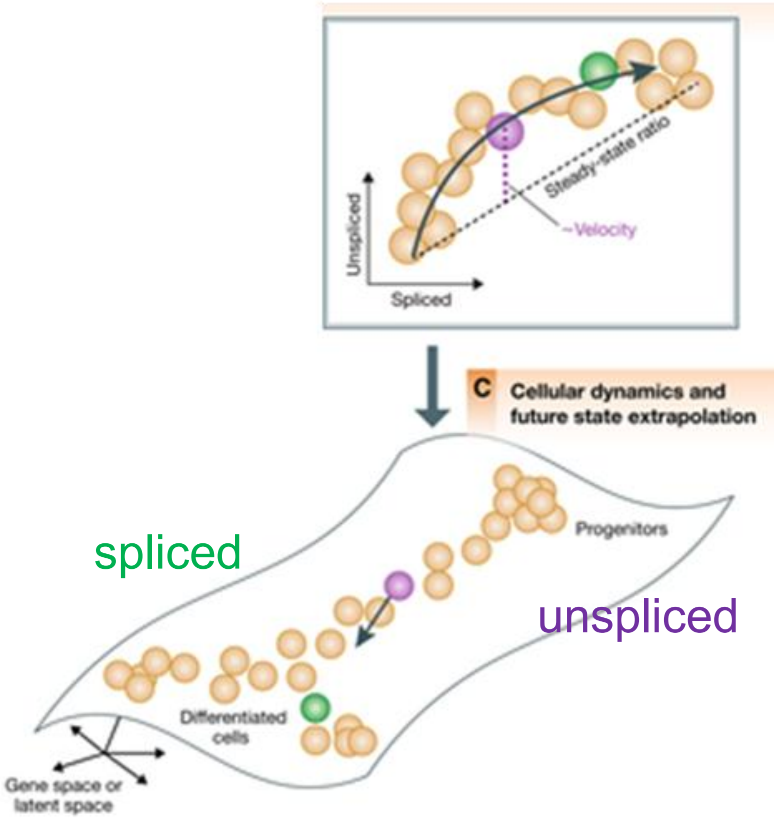
- TradeSeq



What kind of genes are associated with a given trajectory
-Pseudotime ~ Gene expression (correlation) → likely to be poor
-Negative binomial generalized additive model (nonlinear approach)

→ but, super slow …
→ Maybe, binning??

- Velocyto, scVelo

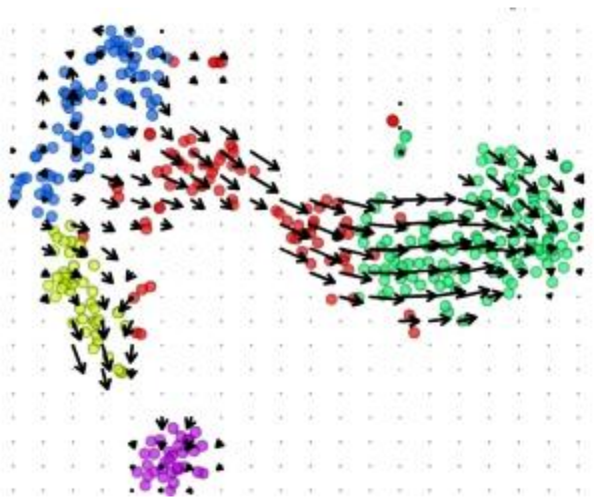


Assumption: during the differentiation, progenitor may have more unspliced RNA while differentiated cell may have fully spliced form

-Compare between unspliced/spliced ratio! (RNA velocity) for each gene
-Merge all the velocities (from all the gene) → project on the user embedding space (UMAP)
-Sum of transition P = 1 for each cell

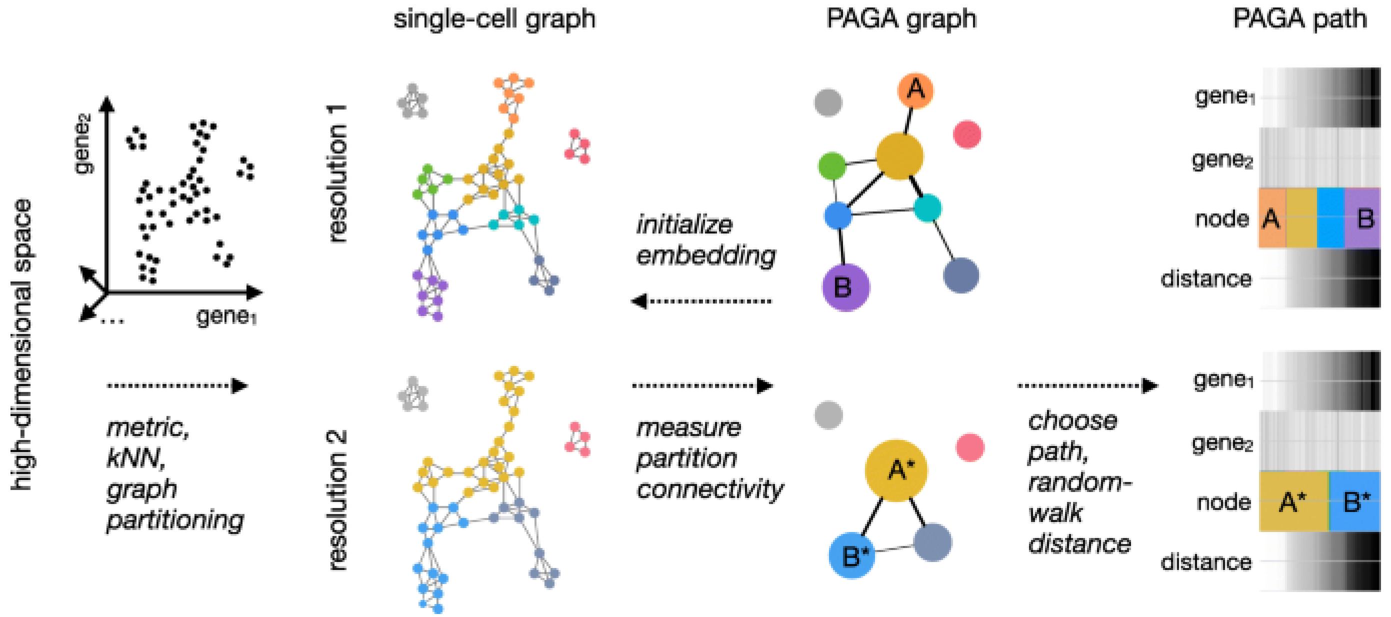
*RNA velocity: steady-state approximation
spliced RNA deg speed = splicing rate * unspliced RNA
- Spliced RNA * degradation rate = 0

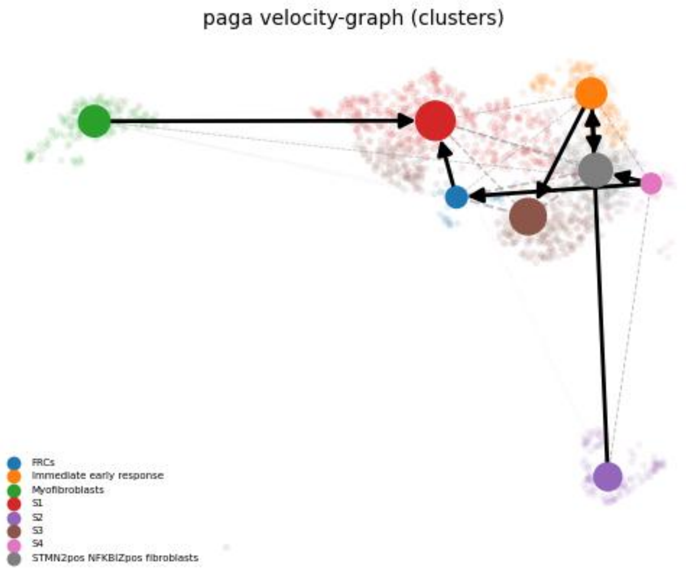
$$\frac{ds}{dt} = \beta u - \gamma s$$



*scVelo: those parameters change across time and cell states

- PAGA (Representation)





-Abstraction of trajectory (scvelo) result
Using graph-based cell-cell similarity (→ connectivity calculated by trajectory)

-PAGA transition confidence score: actual / random expected model

The confidence should be interpreted as the ratio of the actual versus the expected value of connections under the null model of randomly connecting partitions.

- ImmuneDictionary