Subtype Identification and Clustering of single-cell RNA-seq data by nearest neighbor searching

Zhixin Li

July 13, 2018

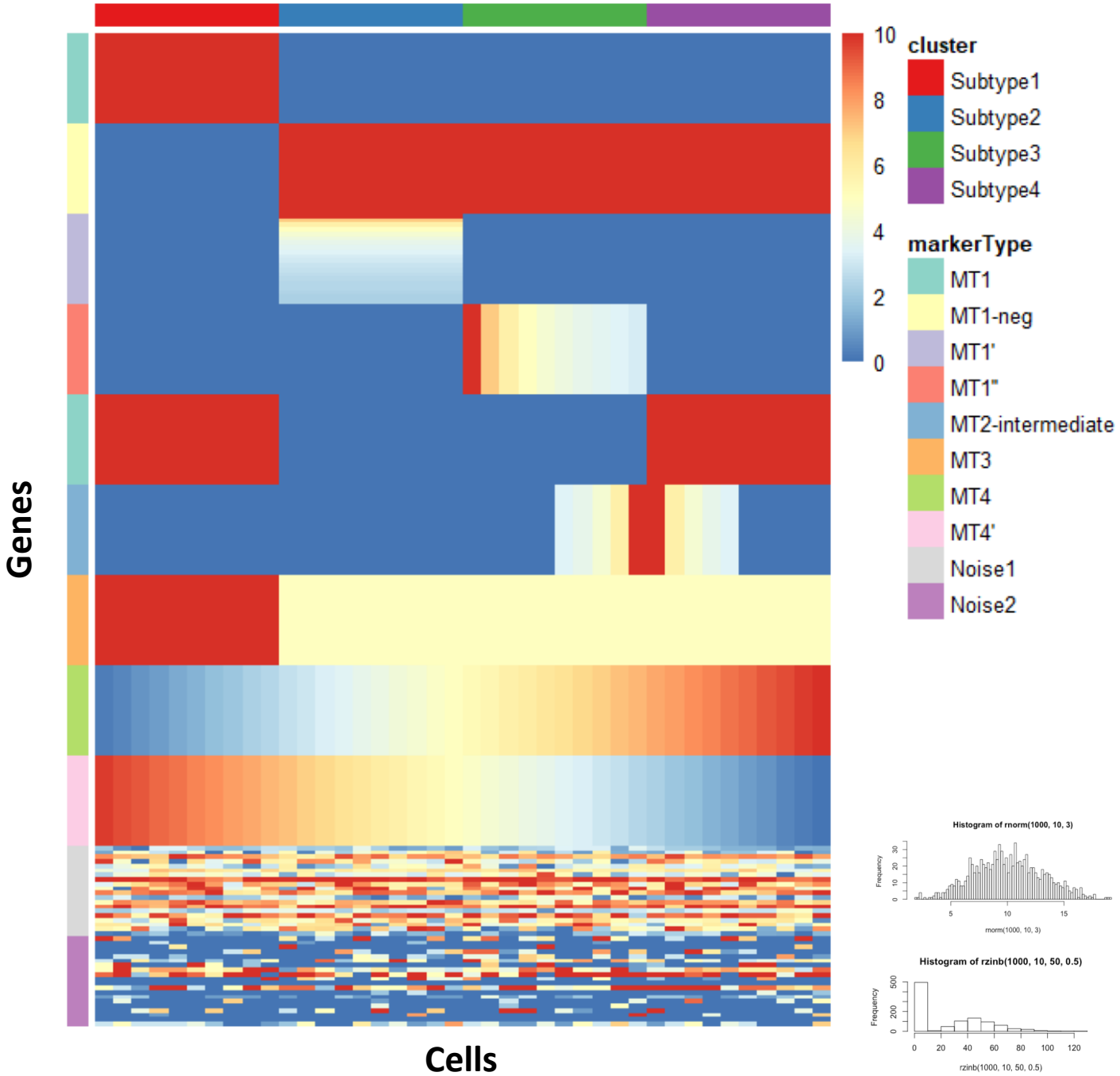
Clustering and the role of marker

- **Heterogeneity**
  - developmental/pathogenic mechanism

- **Before scRNA-seq:**
  - Morphology
  - Surface protein

- **Marker gene:** uniquely/highly expressed gene for certain cells.
- **Cell identity:** marker can reliably distinguish between two or more clusters.

- The main purpose of clustering is to find subtypes and corresponding markers.
- Crucially, marker genes can be used for experimental validation.



Simulated dataset

- A subtype can be defined by a few robust markers (marker module) .
- The more markers, the stronger the reliability of the subtype.

- **Marker types (MTs)**
- **MT1** – uniquely expressed genes with the same expression level
- **MT1-neg** – negative markers of MT1
- **MT1'** – uniquely expressed genes with different expression levels
- **MT1''** – uniquely expressed genes which change across cells
- **MT2-intermediate** – intermediate genes between two subtypes
- **MT3** – highly expressed genes in certain subtypes
- **MT4** – genes with gradually rising/falling expression
- **Noise** – randomly expressed genes

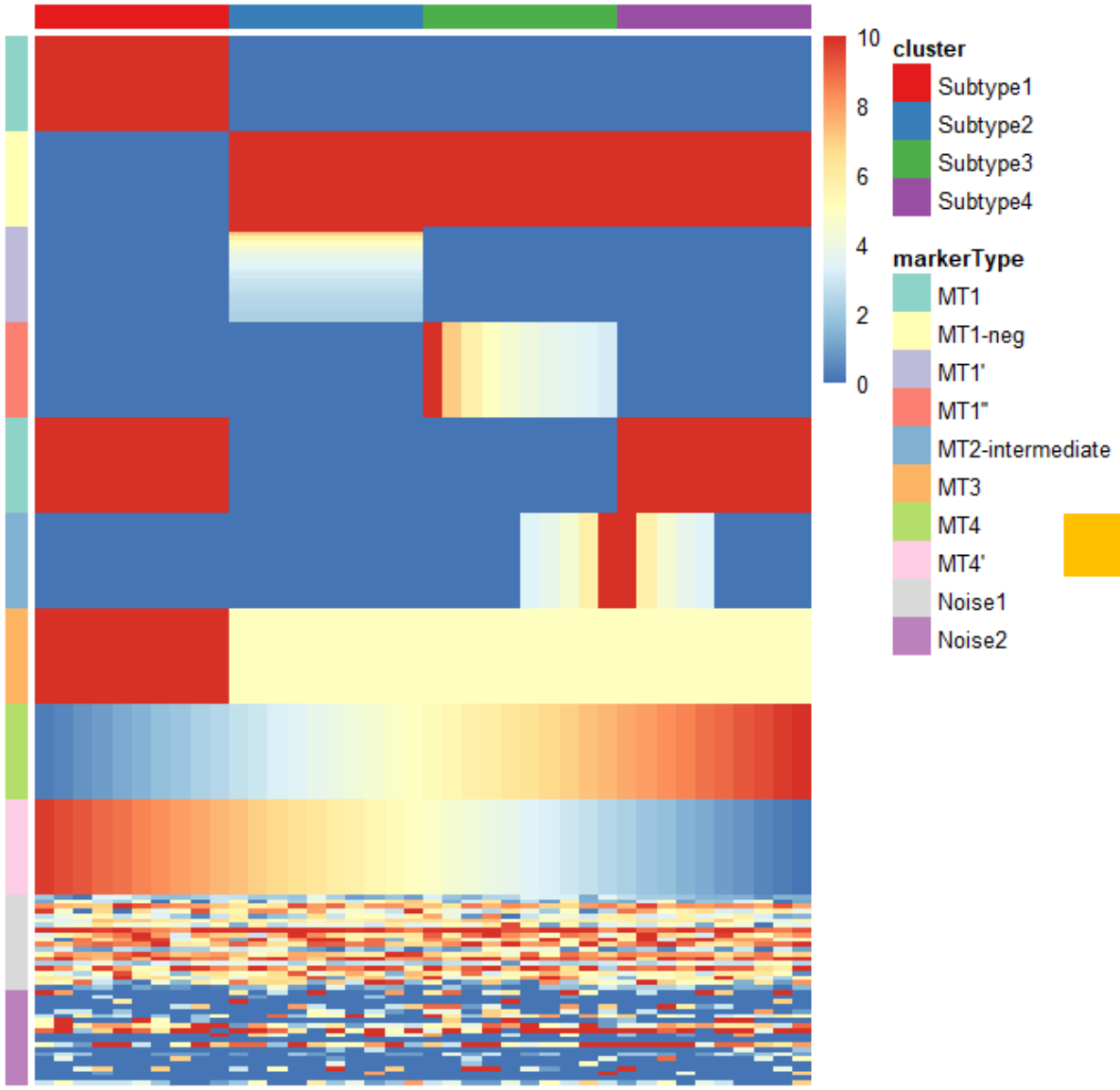
Current clustering method and their limitations

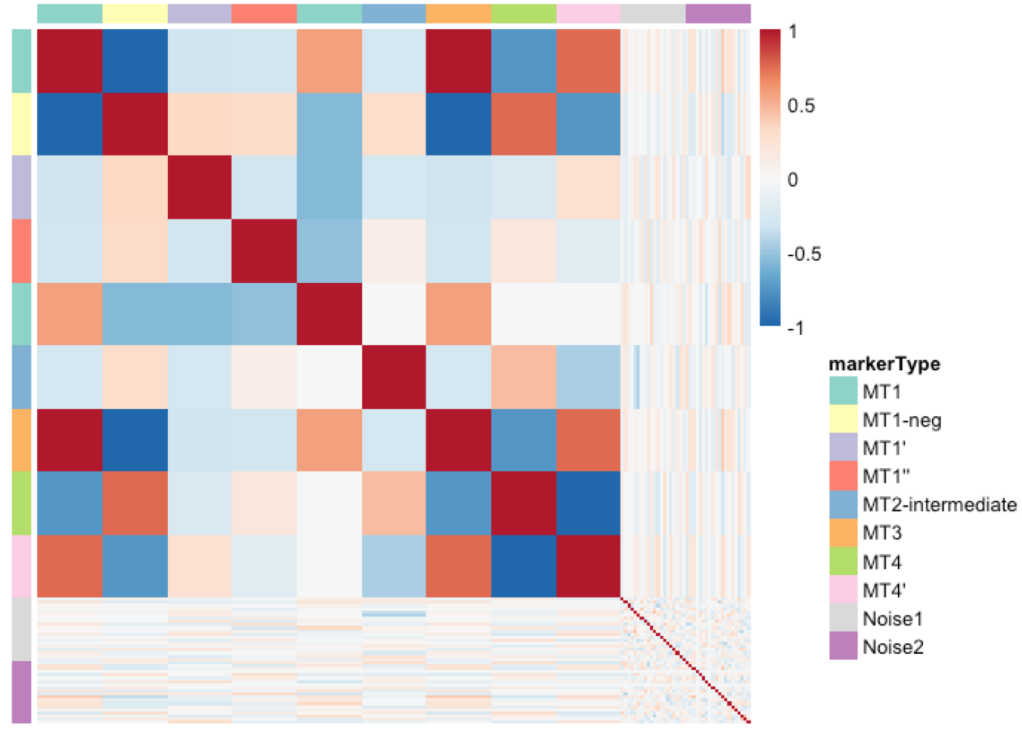
- SC3 (Kiselev et al., Nature Methods 2017) is based on PCA and spectral dimensionality reductions.
- SIMLR (Wang et al., Nature Methods 2017) is a kernel-based similarity learning method.
- Seurat (Butler et al., Nature Biotechnology 2018) is a community detection approach similar to SNN-Cliq
- t-SNE + k-means (10x Genomics official method)
- SINCERA (Guo et al., 2015) is based on hierarchical clustering
- pcaReduce (žurauskienė and Yau, 2016) combines PCA, k-means and “iterative” hierarchical clustering.
- SNN-Cliq (C. Xu and Su, 2015) is a graph-based method.

- Find the relationship (e.g. distance) between cells based on all the features (gene)

- Limitations:
  - Inevitably need to choose a k for clustering (subjectivity)
  - feature selection is crucial
  - Some confounding genes (cell cycle, apoptosis) may largely affect the clustering
  - Risk of overfitting, because of p >> n (p denotes feature count, n denotes sample size)
  - Computationally expensive, especially for large dataset (>1000 cells)

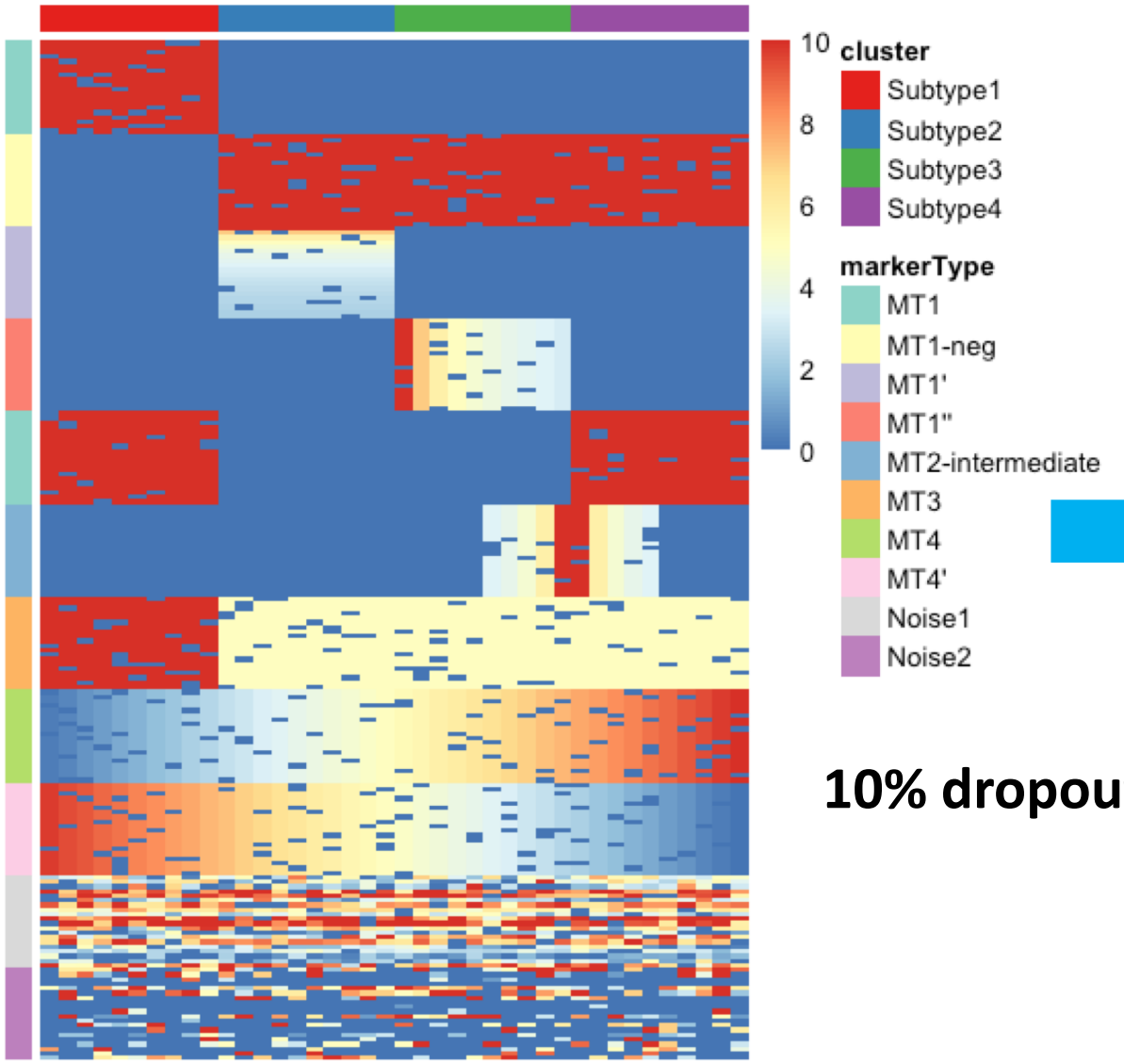
Inspiration



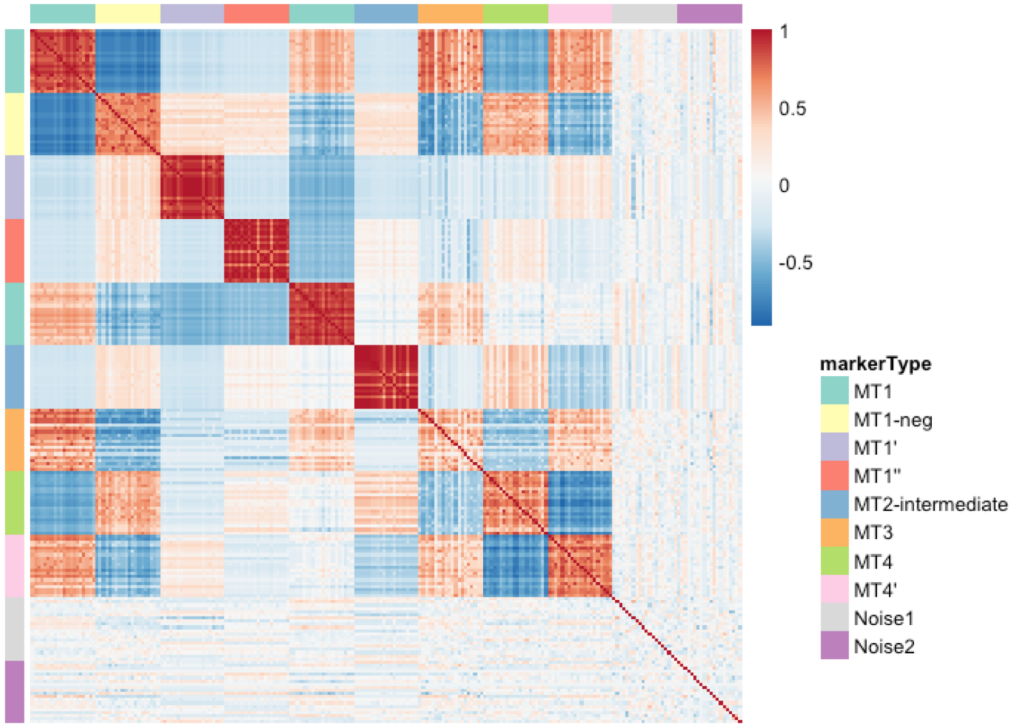


Correlation between the genes in the same marker module is significantly higher.

Can we identify all the marker modules first and then using them to do the clustering?

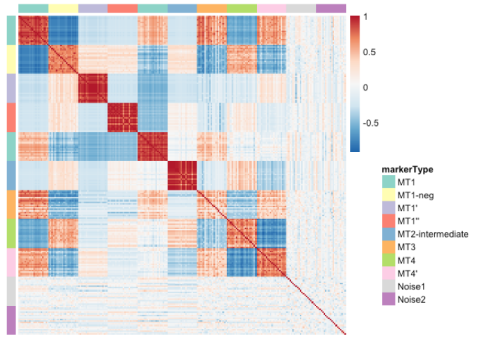


10% dropout rate

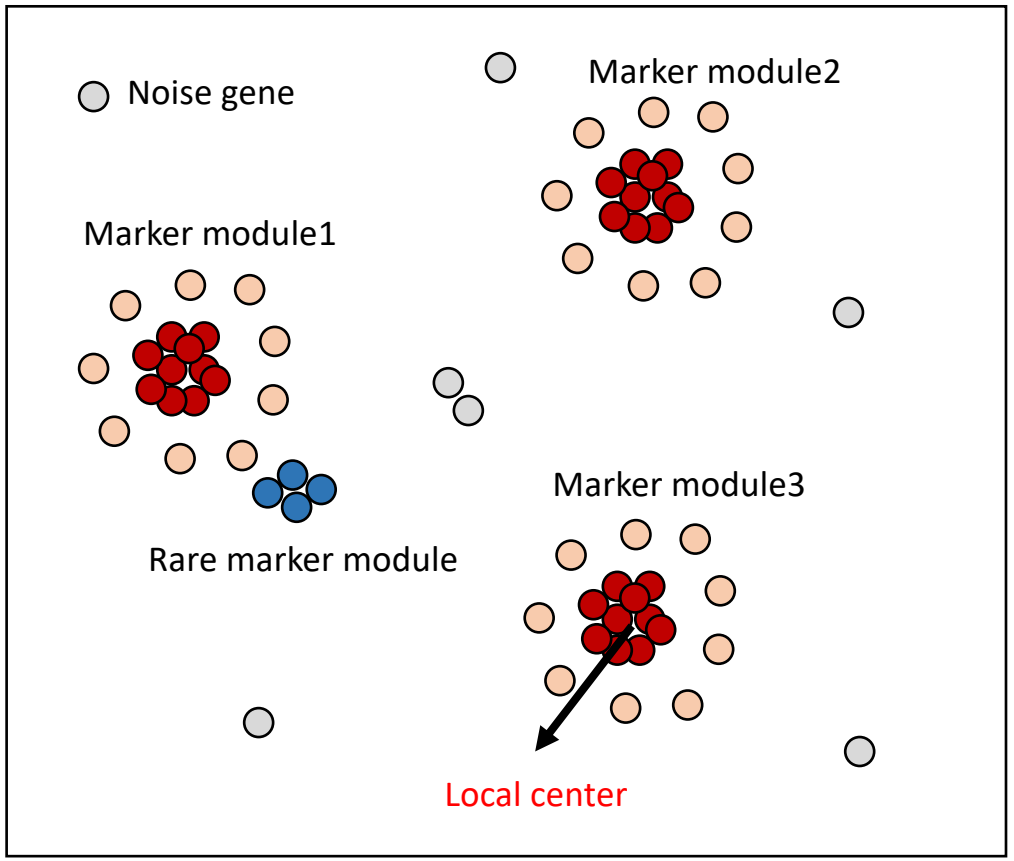


Marker modules still can be captured by correlation relationship.

Dropout events: high chance of missing nonzero entries as zero, due to the technical and biological noise. (low number of RNA transcriptomes and the stochastic nature of the gene expression pattern)



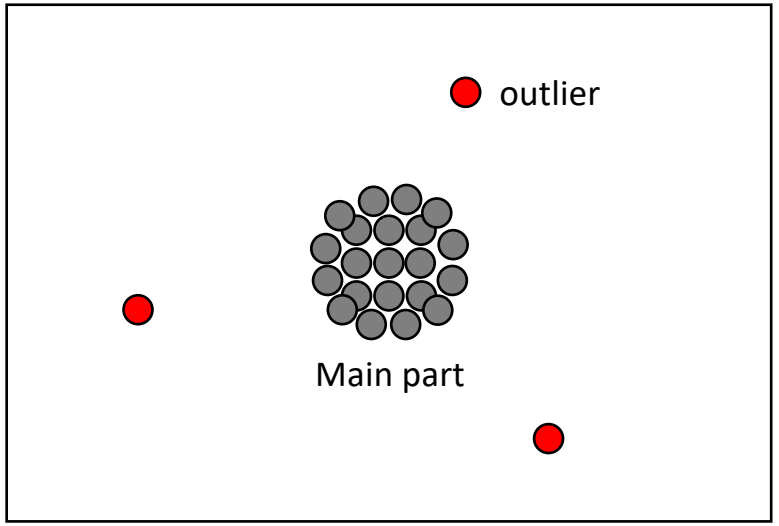
$D_{ij} = 1 - |r_{ij}|$
D, distance;
r, correlation;
i and *j*, *i*th and *j*th gene
$0 \leq D_{ij} \leq 1$.



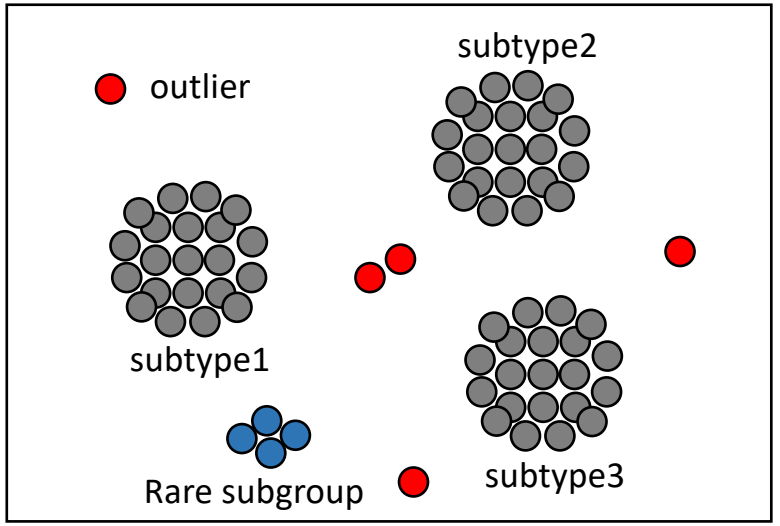
Gene-gene distance network

Algorithm overview

- A robust subtype of cells can be defined by a few (5-10) core markers.
- Core markers locate in the **local centers** (gene clusters with high density) in the gene-gene distance network.
- Core markers identification is a question of local center identification.
- **"Slicing and scanning"** method can effectively detect the local center of a network.
- **KNN** can be used for full marker module identification.
- Subtypes can be inferred by the core markers .



outlier

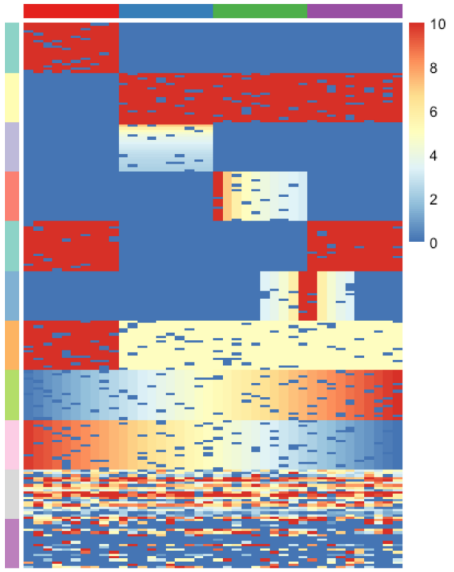


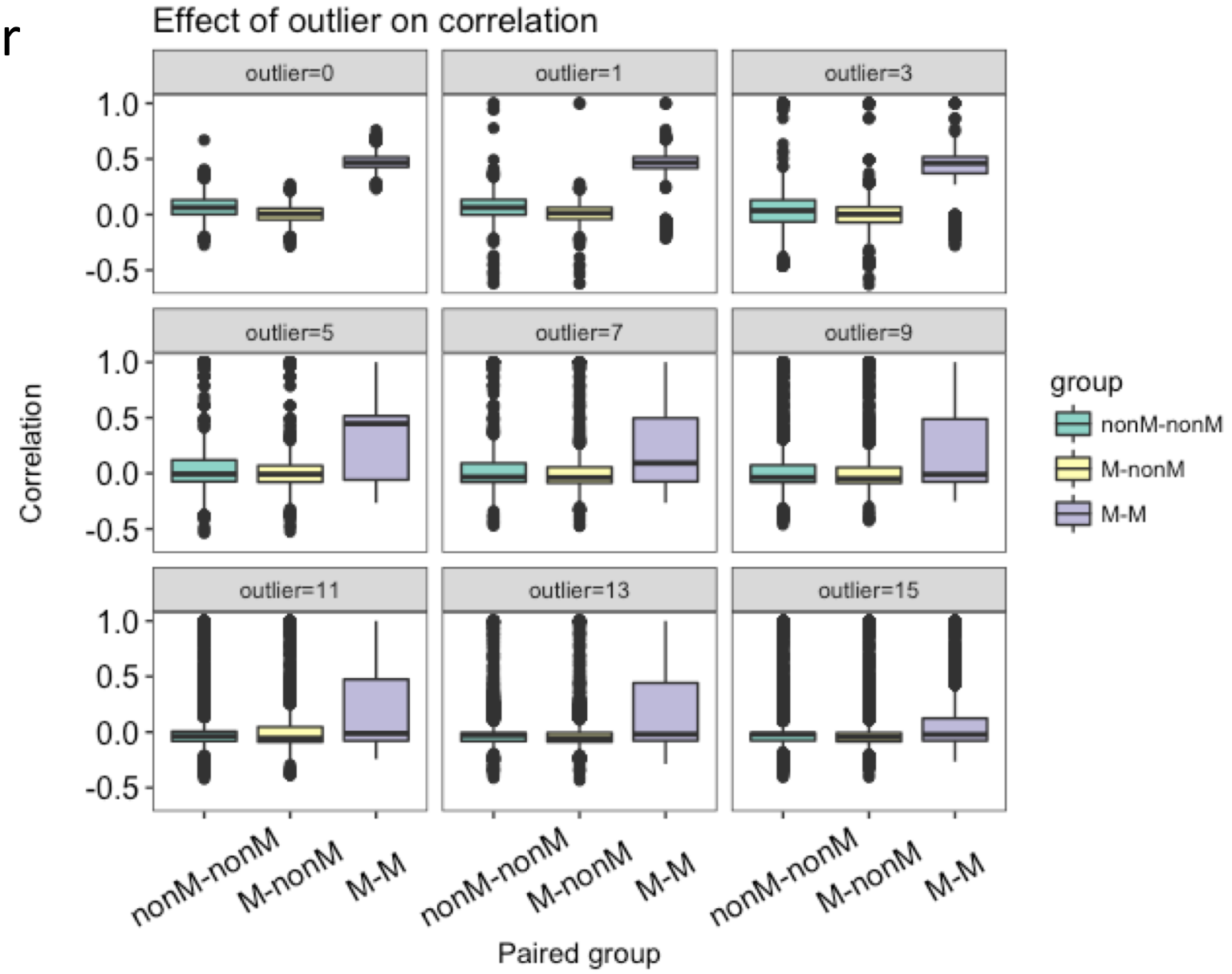
- **Classical version**: observations lying “far away” from the main part of a dataset and probably not following the assumed model (Becker and Gather, 1999).
- **Single cell version**: cells far away from any subgroups in a heterogeneous cell population.
- **Purpose**: removing the extreme outliers far away from any cells and keep the rare cell populations.

Quantify the effect of outlier

Public dataset:
Pollen et al.,2014, NBT
Cells: 249
Genes: 6982

Ground-truth:
Clusters
Markers and non-markers



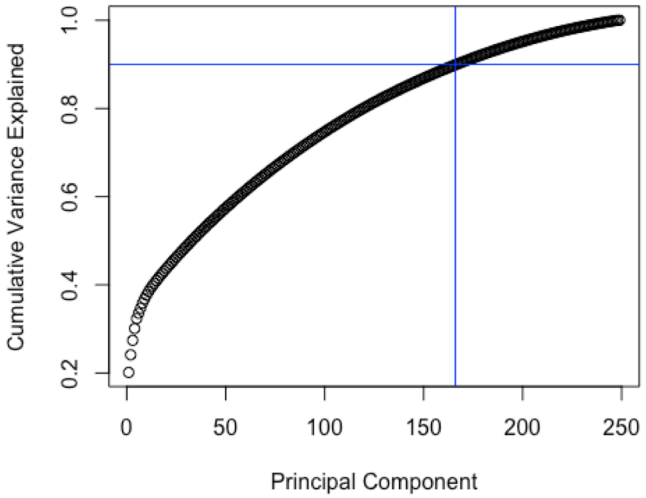


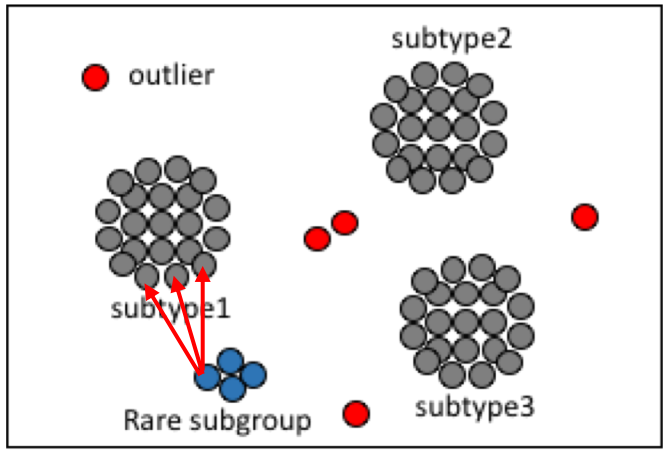
Conclusion: extreme outliers will seriously affect the correlation between markers.

Outlier detection algorithm

- **Dimension reduction**: top components which explained 90% total variance were left. removing noisy features.
- **Mahalanobis distance** were calculated based on the PCA result.
- $(D_{\text{kth nearest neighbors}})_i = \frac{\sum_{j=k}^{j=k+n} (D_{\text{Mahalanobis}})_{ij}}{n}$
- where *i* is the *i*th cells, *n* is the sample size to collect, *j* is the *j*th nearest neighbor.
- *k* is the least cells needed to define a subgroup (e.g. *k*=4, then clusters with 2 cells were defined as outlier)

- Why using kth? Remove far clusters contain less than *k+1* cells.
- Cell has top 5% largest $D_{\text{kth nearest neighbors}}$ will be considered as outlier.



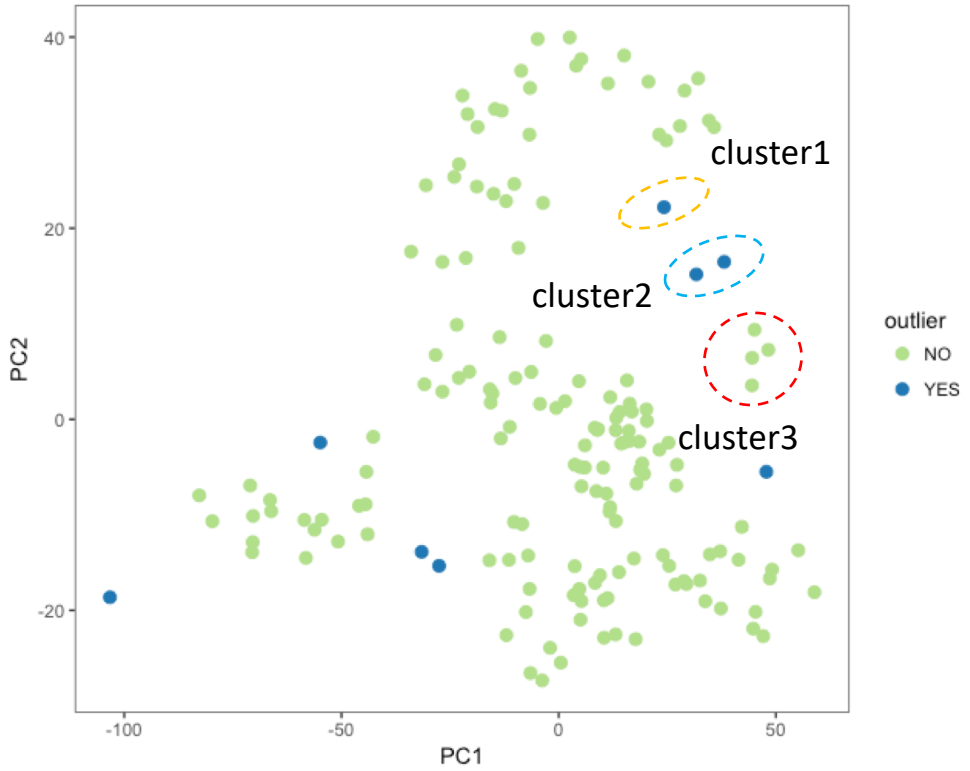


K=4, n=3

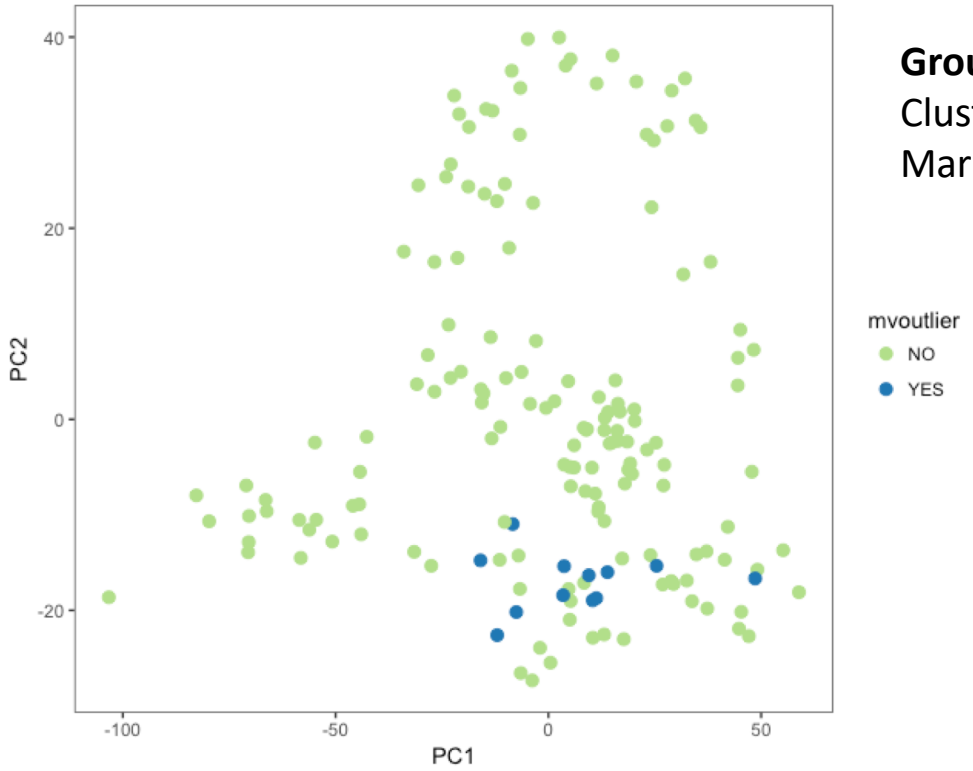
Performance (2 dimensions)

Public dataset:
Deng et al.,2014, Science
Cells: 268
Genes: 22431

Ground-truth:
Clusters
Markers and non-markers



our tool: *k*=4, *n*=5



Compare to mvoutlier

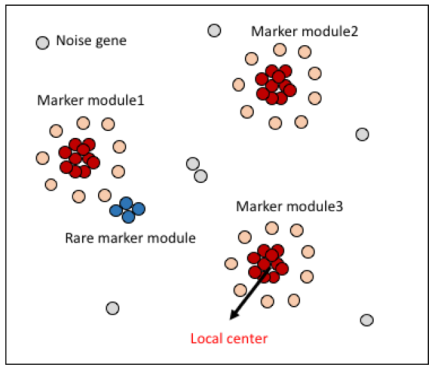
Simulation: only 1, 2 and 4 cells were retained for cluster1, cluster2 and cluster3 to simulate some outliers.

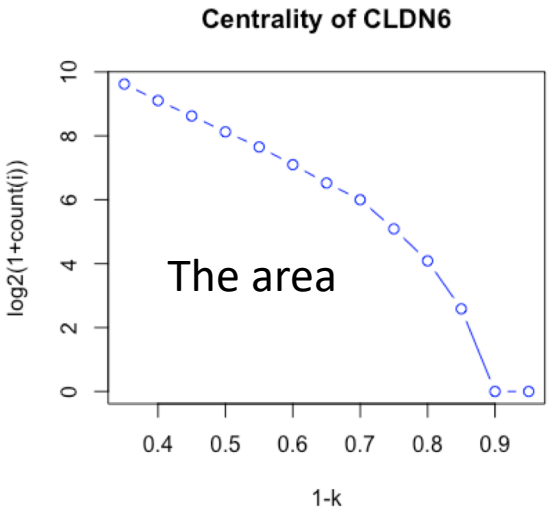
Our method can accurately identify the outliers (cluster1 and cluster2) and keep the rare cluster3.
mvoutlier failed to detected the simulated outliers.

“Slicing and scanning”

- Gene which connects to most other genes with high correlation could be local center.
- $(Score_{centrality})_i$
- $= \int_{k=0}^{k=C} (1-k) \cdot log_2\{1 + (count_i)\}$

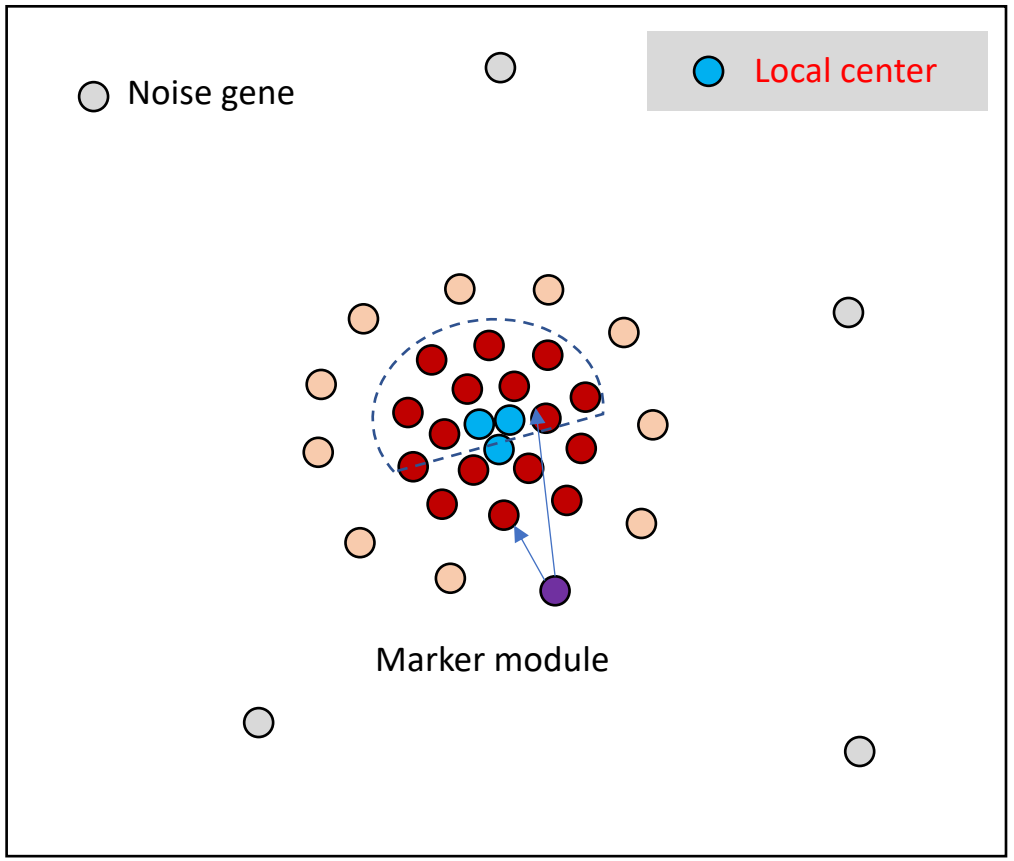
- Where *i* is the *i*th gene, *C* is the max correlation threshold, $count_i$ is the connected gene number under the threshold *k*.





	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	centrality
CLDN6	0	0	5	16	33	63	91	136	200	278	393	549	786	23.12
CLDN7	0	0	5	18	30	54	85	134	196	268	376	551	770	22.99
KRT18	0	0	3	14	34	53	90	125	183	266	379	543	767	22.85
CLDN4	0	0	2	16	28	52	88	129	192	265	373	530	750	22.77
ANXA3	0	0	1	12	30	57	96	129	187	274	381	531	761	22.76
EPCAM	0	0	2	15	29	49	81	127	184	258	372	519	744	22.66
ARHGAP29	0	0	2	13	27	52	85	124	177	262	370	520	729	22.62
KRT19	0	0	1	10	26	53	84	125	179	261	371	553	776	22.59
KRT8	0	0	2	9	29	47	80	118	176	252	363	535	755	22.50
TACSTD2	0	0	1	8	23	49	79	118	177	250	362	503	718	22.30
SPINT2	0	0	0	6	26	48	80	115	176	251	362	525	745	22.21
WNT6	0	0	0	4	23	45	78	114	172	243	364	515	737	22.05
ACSM3	0	0	0	4	17	40	72	115	161	238	340	483	695	21.76
CDH6	0	0	0	3	19	45	75	111	162	234	331	479	691	21.75
ANXA2	0	0	0	8	14	37	58	100	148	221	319	482	677	21.55
CDH1	0	0	0	3	18	37	64	98	156	221	324	471	687	21.50
PLP1	0	0	0	1	10	33	69	102	153	234	353	515	727	21.39
EFNA1	0	0	0	2	16	35	65	93	150	223	322	467	671	21.35
SEMA3D	0	0	0	2	18	37	64	102	144	225	309	447	646	21.34
VAMP8	0	0	0	2	17	34	63	100	144	215	316	459	671	21.32
CD24	0	0	0	4	14	33	62	103	148	212	307	454	622	21.30
DMKN	0	0	0	1	11	32	67	101	148	211	331	482	700	21.23
CYP26A1	0	0	0	2	11	31	58	95	146	213	304	433	631	21.04
MYL12B	0	0	0	2	11	28	59	96	144	213	301	436	646	21.03
PDLIM1	0	0	0	1	14	31	59	95	143	213	304	427	633	21.01
TSTD1	0	0	0	0	12	34	66	94	144	207	304	454	641	20.93
LY6E	0	0	0	3	10	30	51	89	137	203	291	416	603	20.87
TINAGL1	0	0	0	0	12	31	55	95	136	205	291	426	622	20.73
PERP	0	0	0	0	9	27	58	82	135	198	291	458	654	20.64
RPS2	0	0	1	7	13	35	54	68	108	161	242	328	510	20.56
TDRP	0	0	0	0	8	29	50	90	132	202	281	417	619	20.51
SLC2A3	0	0	0	2	5	24	48	87	129	183	287	411	590	20.46
RPS3	0	0	1	6	14	36	57	73	101	152	215	305	473	20.39
EDNRB	0	0	0	1	3	20	50	84	122	183	293	438	656	20.32
ATP1B3	0	0	0	0	1	25	51	91	139	201	300	463	675	20.28
SEMA3A	0	0	0	0	4	22	44	74	121	185	285	424	621	20.10
SPP1	0	0	0	0	3	20	45	86	121	178	273	425	624	20.05
KRT7	0	0	0	0	6	20	42	68	117	179	257	391	609	19.96
TPM1	0	0	0	0	4	25	44	75	116	175	251	393	557	19.91
PTN	0	0	0	0	2	16	43	78	116	176	265	408	606	19.77

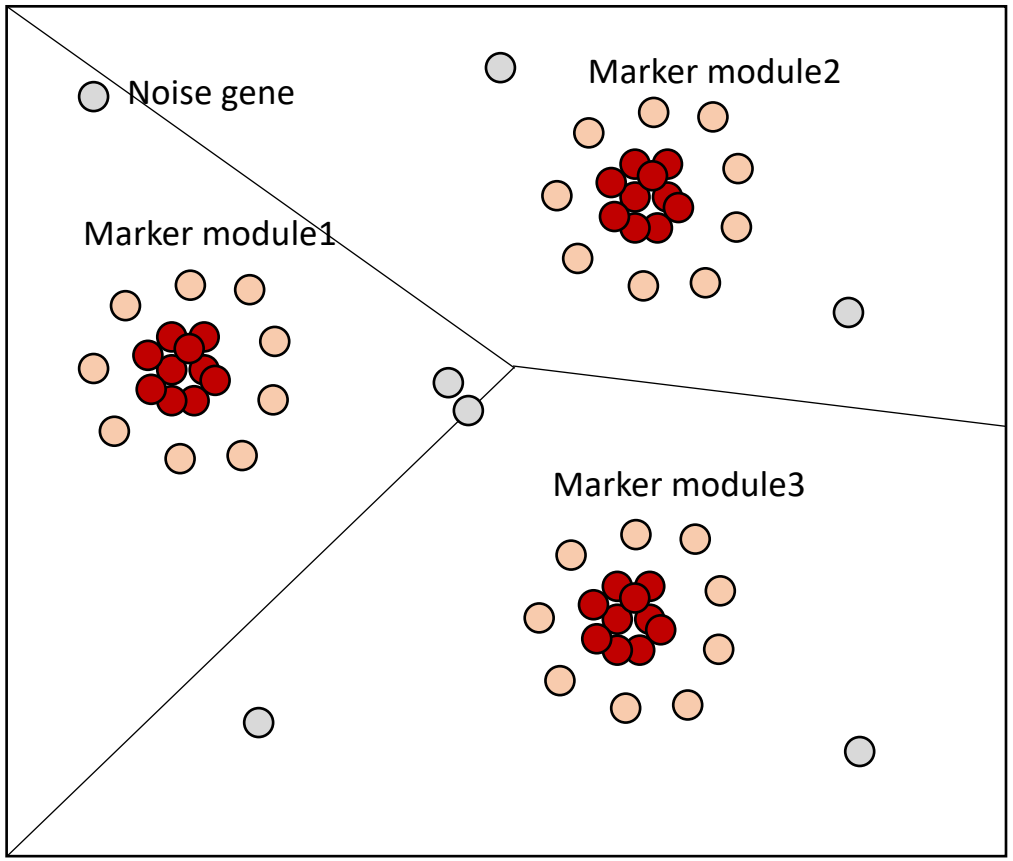
Nearest neighbors searching (core markers)



Gene-gene distance network

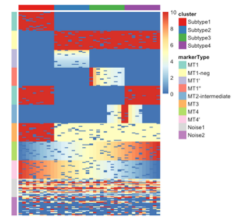
- Step1: define the local centers, initialize the marker module
- Step2: Iteratively find the nearest neighbor of the marker module
- Step3: update the marker module
- Step4: until the new nearest neighbor far away from half of the genes (or center) in the marker module under a certain threshold

Full marker module identification (KNN)

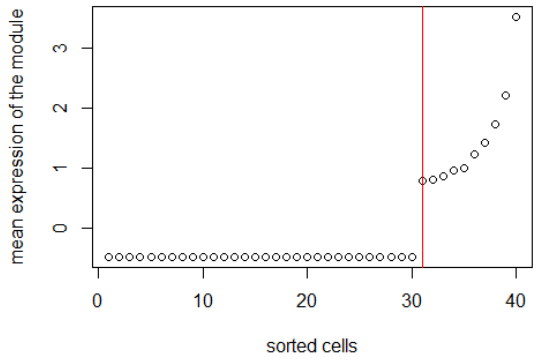


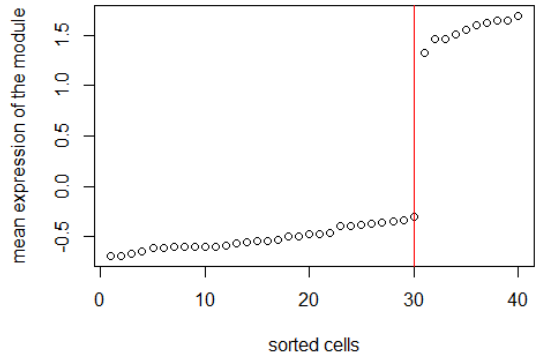
- A new gene is classified by a majority vote of its neighbors, with the gene being assigned to the marker module most common among its k nearest neighbors.
- K was chosen by cross validation.
- Kd tree was used to improve the searching efficiency.

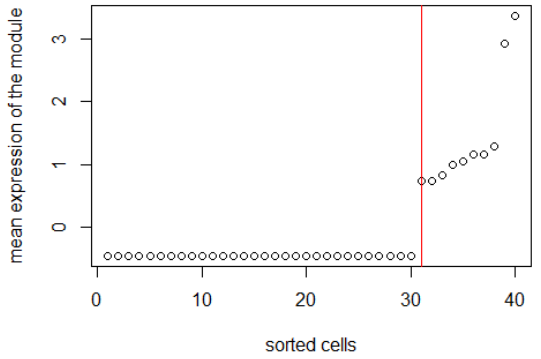
Performance on simulated dataset

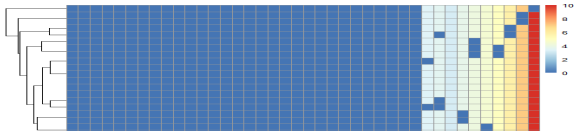


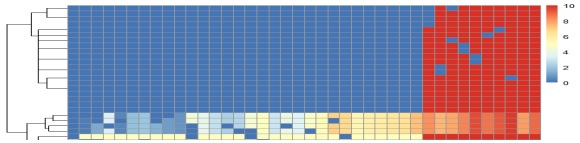
breakpoint

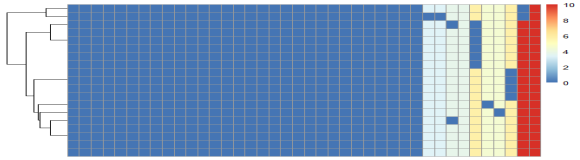




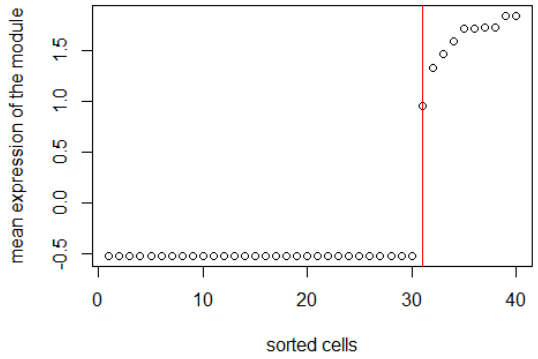


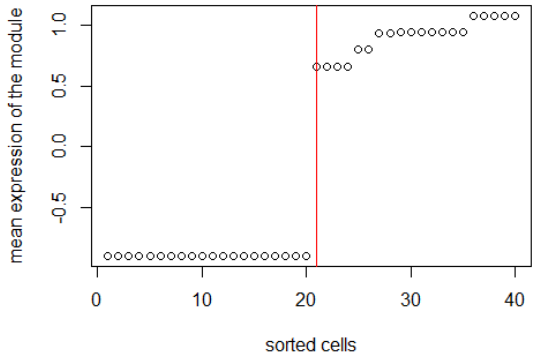


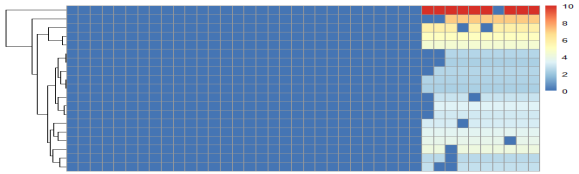


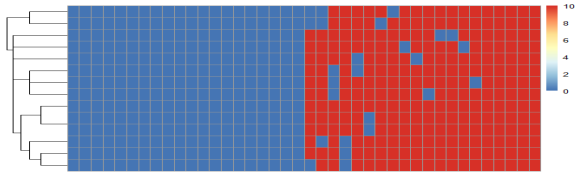


Marker module

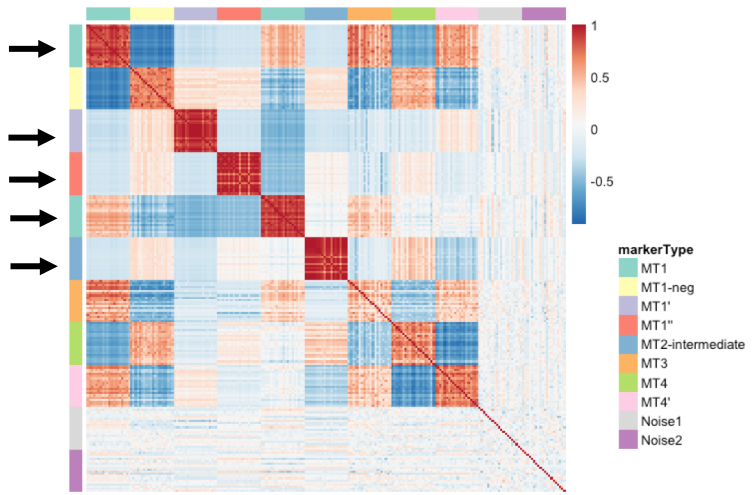








High correlation

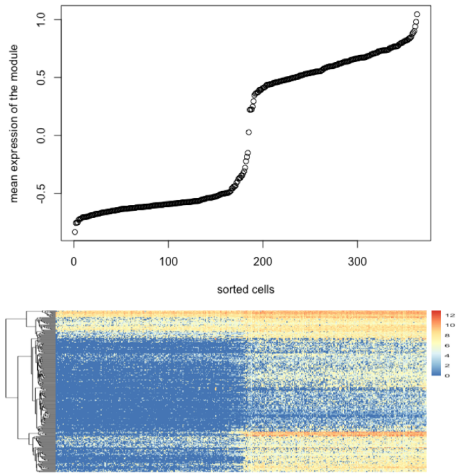


Performance on real dataset

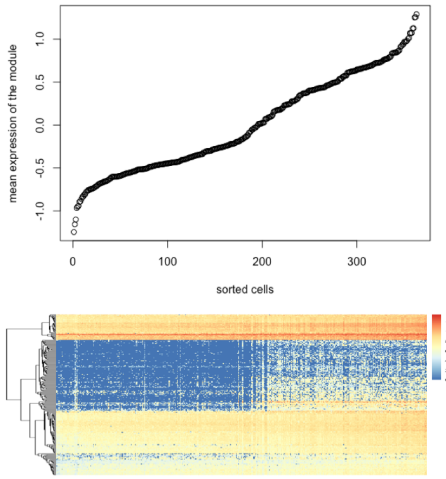
- **Smart-seq(2)**
  - Expensive
  - 100-1000 cells
  - 3k-10k expressed genes
  - Low dropout rate
- **10x genomics**
  - Relative cheap
  - 1k-10k cells
  - 500-3000 expressed genes
  - High dropout rate

estimate the dropout rate

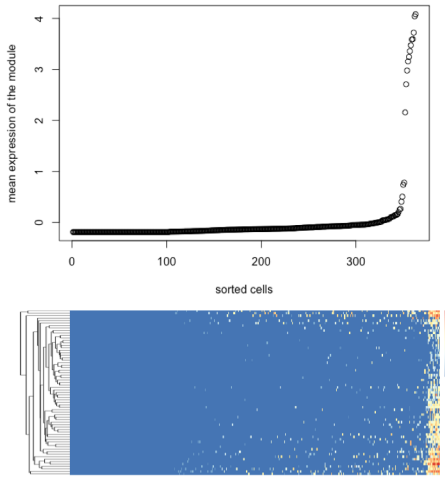
Smart-seq dataset (our HSCR cells)



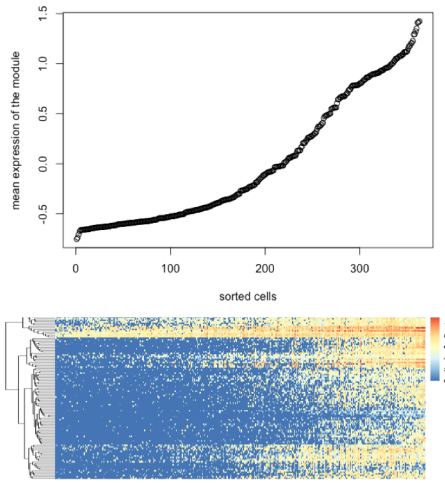
cluster1



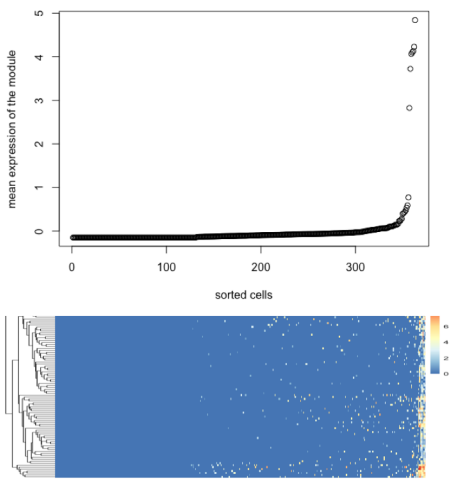
cluster2



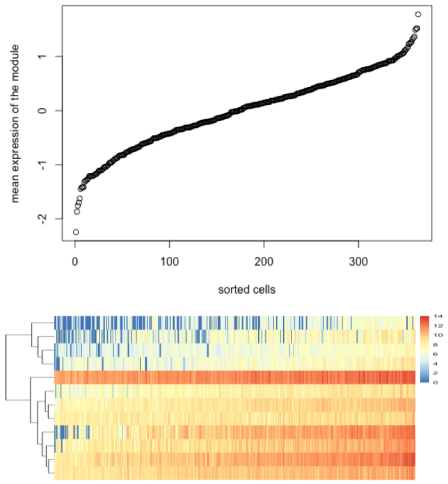
cluster3



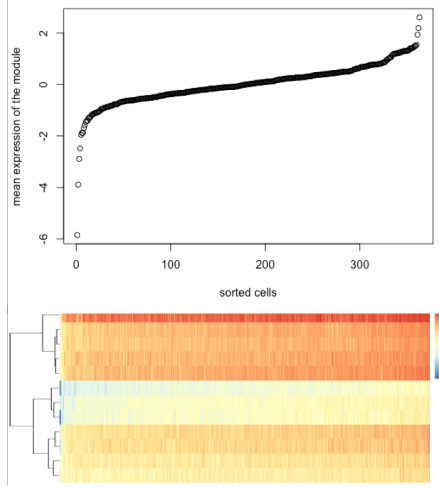
New module

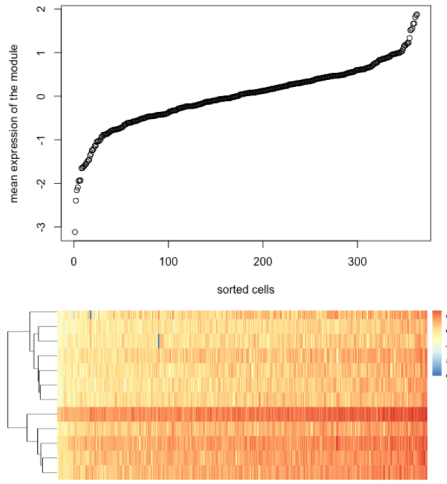


cluster5

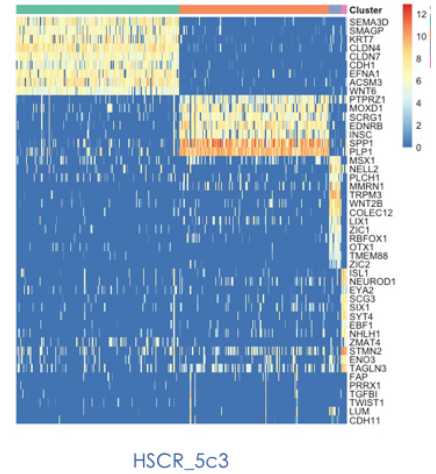


New module





New module



Ground truth

All marker modules in ground truth can be detected by MSIC.

Three new modules are detected.

Co-expression network analysis

- MSIC is Similar to WGCNA in some part.

- WGCNA
  - No outlier detection
  - Hierarchical clustering (poor for core marker identification)
  - Only for full module identification
  - Cannot be used in scRNA-seq

- MSIC can also be used for bulk RNA-seq samples anaylsis (>100)

Execution time comparison

- Genes: 30,000

Time/min	400 cells	1200 cells	2000 cells	3000 cells	4000 cells
MSIC	~5	~10	-	-	-
SC3	~10	~120	-	-	-
SIMLR	1.66	20.40	-	-	-
Seurat	-	-	-	-	-
…					

Low algorithm complexity

The most time consuming step is correlation calculation.

More tools and datasets will be tested in the future.

Advantages of this method

- No need to choose a k for clustering
- No need to do feature selection
- Can identify negative markers, intermediate subgroup and rare subgroup
- Greatly reduce the computing resource
- Great benefit for downstream analysis (pathway)

Discussion

- Dropout rate may largely affect the result with the increasing of total cells.

Any suggestion will be appreciated!