A tutorial for the

single cell topological analysis (scTDA)

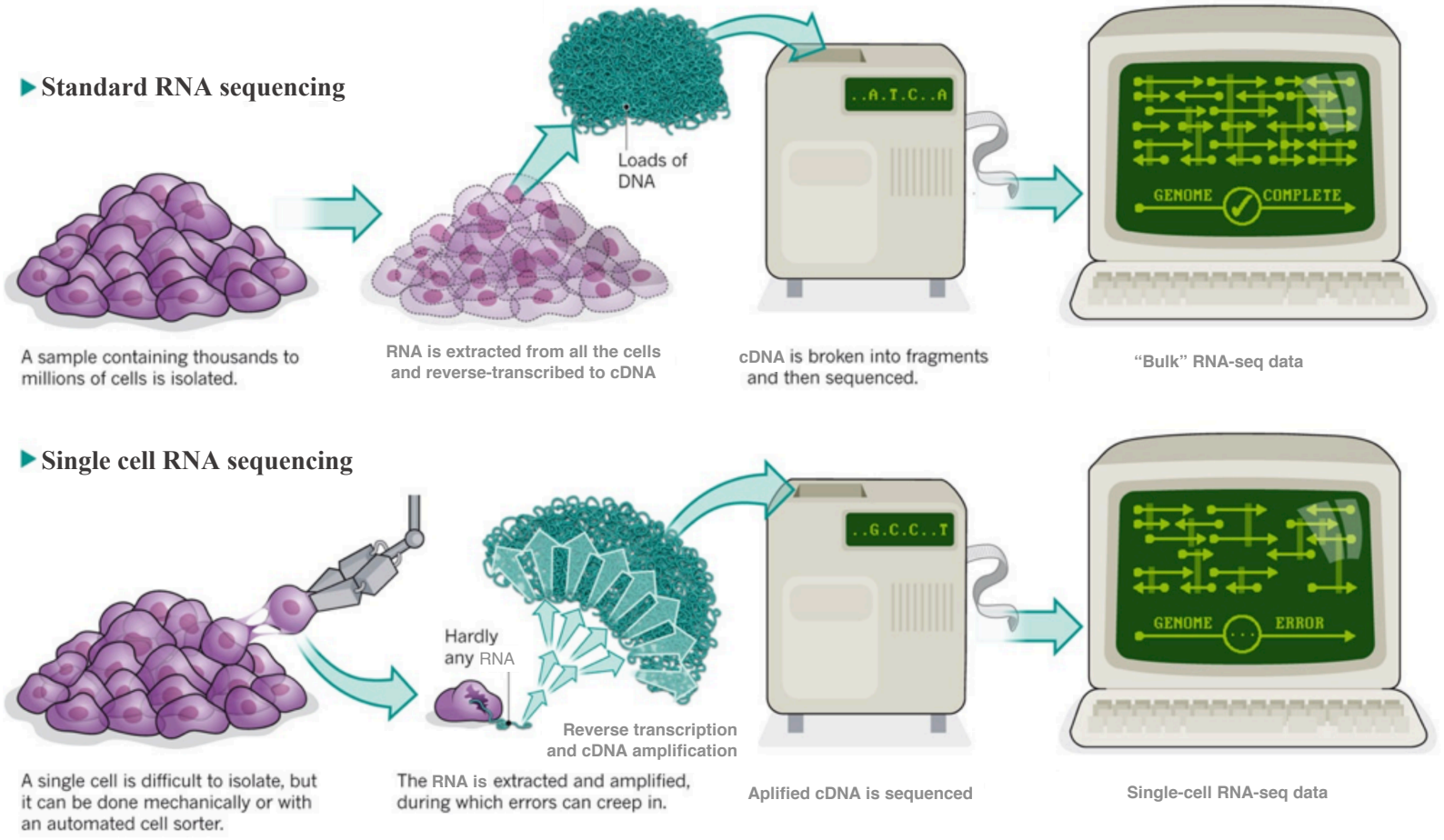
Quanhua Mu

Postdoctoral Fellow, HKUST

Outline

- Introduction to single cell RNA sequencing
- Introduction to scTDA (mapper)
- Case study: human embryo development at single-cell resolution

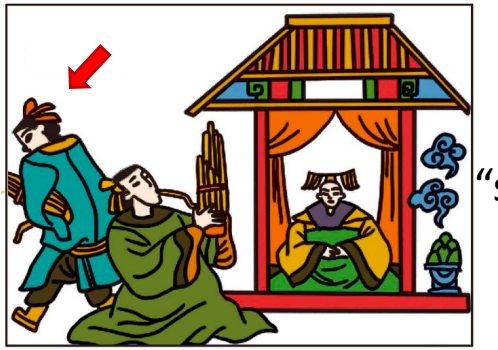
Single cell RNA sequencing (scRNA-seq)



Why single cell?

- Remove ensemble average
- Discover rare species
- Reveal mechanisms

"bulk"



"single-cell"

scRNA-seq data

1000s of cells

1000s of genes

Gene	Cell1	Cell2	Cell3	Cell4	Cell5	Cell6	Cell7	Cell8	Cell9	Cell10	Cell11	Cell12	Cell13	Cell14	Cell15	Cell...
A1BG	0	22	0	0	0	24	13	28	9	49	22	0	35	16	0	0
A1BG-AS1	0	0	0	0	0	0	0	0	0	6	0	0	0	2	0	0
A1CF	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0
A2M	0	0	0	0	0	0	0	0	0	1	4	4	30	0	0	0
A2M-AS1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2ML1	0	3	0	0	231	46	68	0	1	149	0	118	94	50	0	0
A2MP1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3GALT2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A4GALT	1	19	269	150	290	471	387	299	201	381	316	338	231	302	59	0
A4GNT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAAS	0	0	303	36	163	209	242	157	295	148	185	114	324	349	61	0
AACS	0	0	374	87	296	317	73	274	164	113	108	295	96	211	0	5
AACSP1	0	0	0	0	0	0	0	0	23	0	8	0	0	0	0	0
AADAC	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	1	0	0	0	0	7	0	0	1	0	0
AADACL2-AS	0	0	0	0	4	7	0	1	11	13	7	6	45	21	0	0
AADACL3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AADACL4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AADACP1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
AADAT	8	0	268	8	159	166	138	30	64	92	75	112	29	79	0	0
AAED1	3	0	251	1	134	92	73	59	83	73	91	0	17	67	0	1
AAGAB	13	52	650	539	1815	1233	1606	1217	1059	437	955	1013	1280	1799	97	79
AAK1	61	7	800	321	1164	1238	766	530	541	752	499	757	579	1520	65	41
AAMDC	0	0	283	41	192	54	72	47	44	6	59	45	1	31	0	0
AAMP	1212	274	2723	2433	4005	6759	7282	4315	4923	4632	3887	5046	5974	7274	22	367
AANAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAR2	261	1	42	147	848	940	600	538	625	546	459	860	555	989	53	17
AARD	0	0	0	4	8	3	13	0	5	0	0	0	0	0	0	0
AARS	213	1	1225	714	1160	1488	1228	928	1114	848	1001	969	1892	1459	85	4
AARS2	0	0	0	5	262	187	58	180	121	102	191	0	253	163	0	0
AARSD1	0	0	11	0	9	19	8	10	19	5	20	13	16	12	0	0
AASDH	5	0	1	123	165	249	444	133	146	212	227	143	218	142	19	4
AASDHPPT	147	21	560	1330	900	1211	1531	832	721	637	1116	776	1016	1098	44	27
AASS	0	3	125	619	680	434	404	523	493	381	787	727	755	250	208	7
AATF	998	85	494	455	2158	2345	2144	1401	1752	1765	1374	1535	1584	2328	168	52

A screenshot of the scRNA-seq data

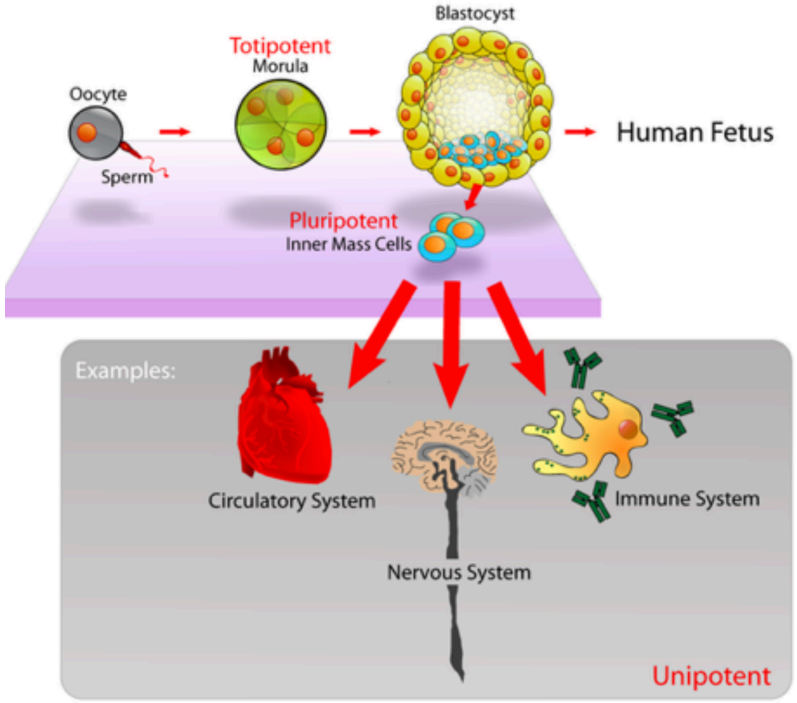
Cell	Source	SomeValue	Timepoint
Cell1	Embryo1	0.3220316	3
Cell2	Embryo1	0.2569401	3
Cell3	Embryo1	0.6377016	4
Cell4	Embryo1	0.5396607	4
Cell5	Embryo1	0.0386912	5
Cell6	Embryo1	0.188975	5
Cell7	Embryo1	0.0321679	6
Cell8	Embryo1	0.9965712	6
Cell9	Embryo2	0.5278399	3
Cell10	Embryo2	0.9762601	3
Cell11	Embryo2	0.802467	4
Cell12	Embryo2	0.8780193	4
Cell13	Embryo2	0.1046346	5
Cell14	Embryo2	0.2892319	5
Cell15	Embryo2	0.2540569	6
Cell16	Embryo2	0.828112	6
Cell17	Embryo3	0.9332562	3
Cell18	Embryo3	0.4744063	3
Cell19	Embryo3	0.0436245	4
Cell20	Embryo3	0.6666969	4
Cell21	Embryo3	0.984966	5
Cell22	Embryo3	0.1879344	5
Cell23	Embryo3	0.8628643	6
Cell24	Embryo3	0.4924577	6
Cell25	Embryo3	0.9701649	7
Cell26	Embryo3	0.2287696	7
Cell27	Embryo3	0.2206378	7

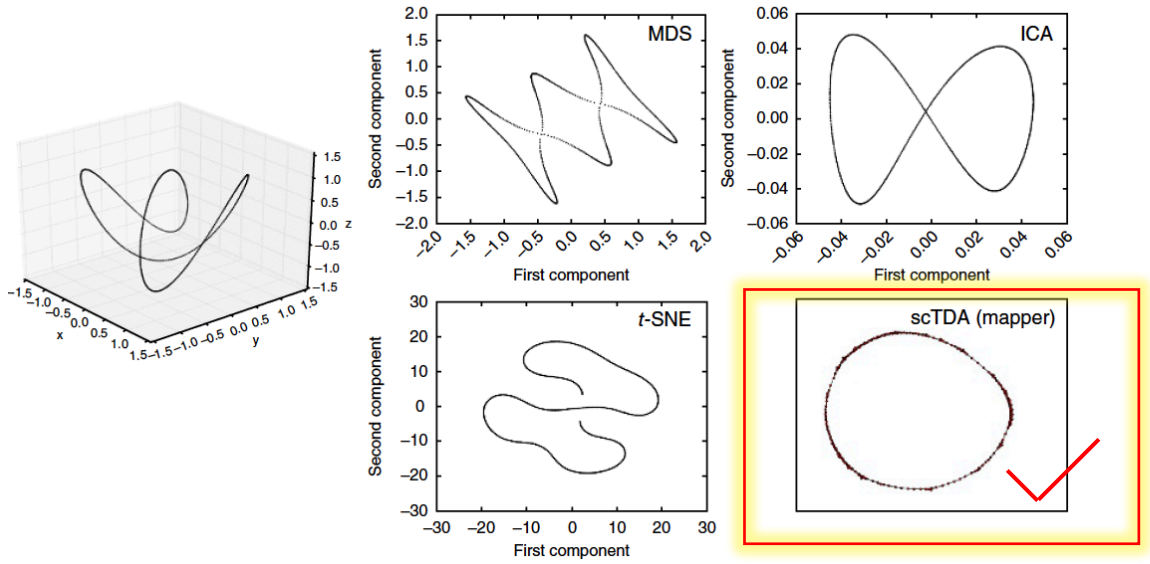
Annotation data of the cells

single cell topological analysis (scTDA)

- Biological problem: to reveal the developmental trajectories of the single cells

- Mathematical definition: to perform dimensional reduction while preserving the continuous relationship in high dimensions

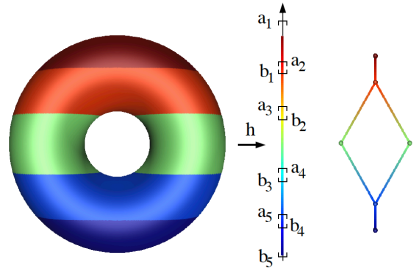


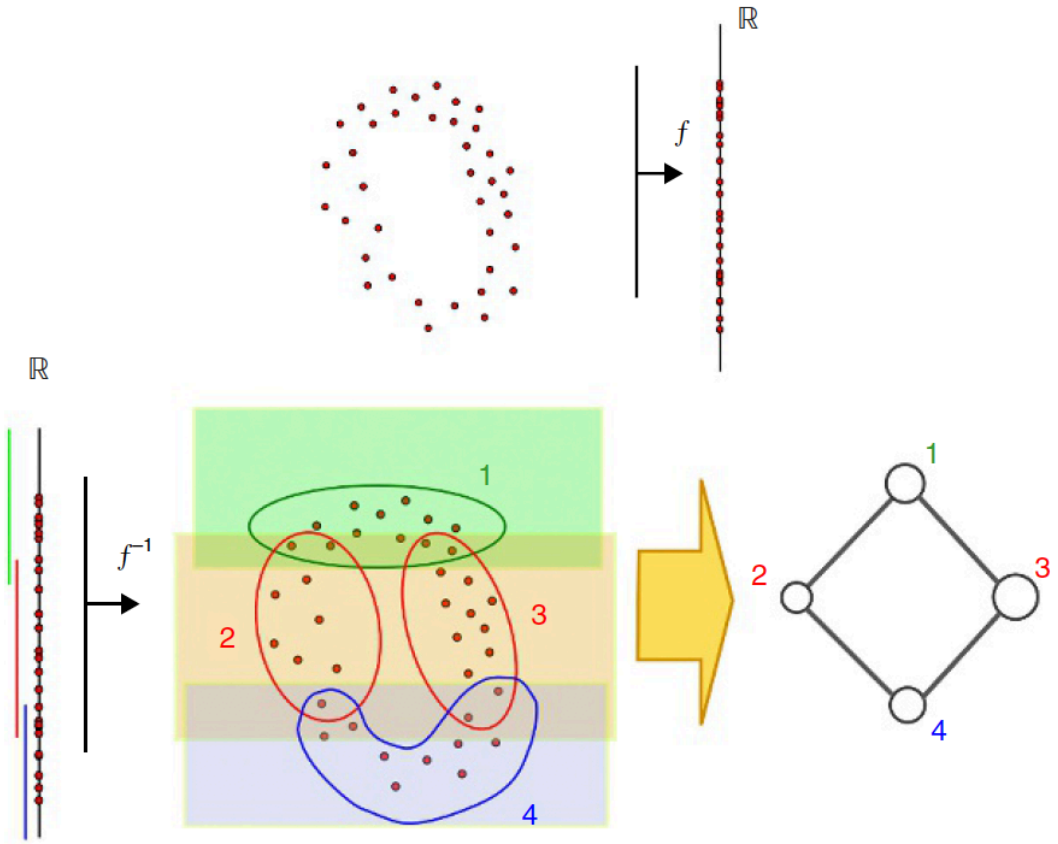


Rizvi et al, ***Nat Biotech*** 2017

scTDA and the Mapper algorithm

- scTDA is essentially based on the Mapper algorithm





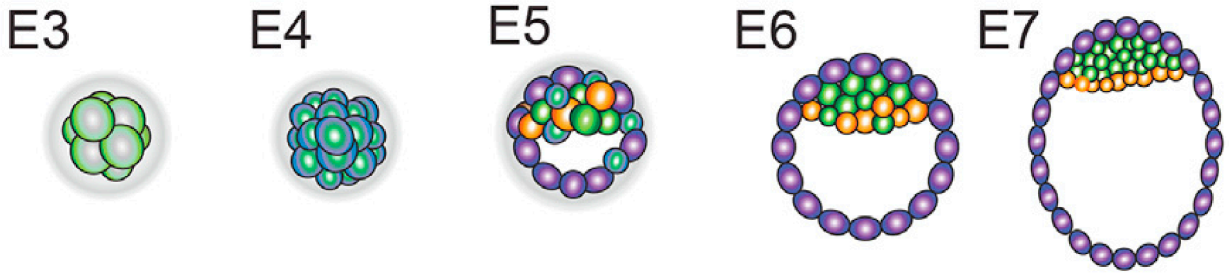
Mapper algorithm

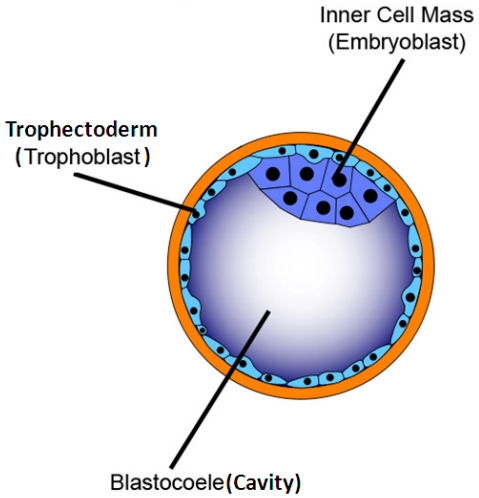
Top: Mapper builds upon dimensional-reduction function *f* mapping the high-dimensional single-cell RNA-seq point-cloud data into $\mathbb{R}^k$ ($k = 1$ in the figure).

Bottom: under the inverse function f^{-1} , a covering of $\mathbb{R}^k$ maps into a covering of the single-cell point-cloud data. Clustering is performed independently in each of the induced patches in the high-dimensional space. In the low-dimensional representation, a node is assigned to each cluster of cells. If two clusters intersect, the corresponding nodes are connected by an edge. Topological features in the low-dimensional representation are guaranteed to also be present in the original high-dimensional RNA-seq space.

Adopted from Prof Yao's course slides and Rizvi et al, ***Nat Biotech*** 2017

Case study: Human embryo development





						Total
Embryos	13	16	24	18	18	88
Cells	81	190	377	415	466	1529

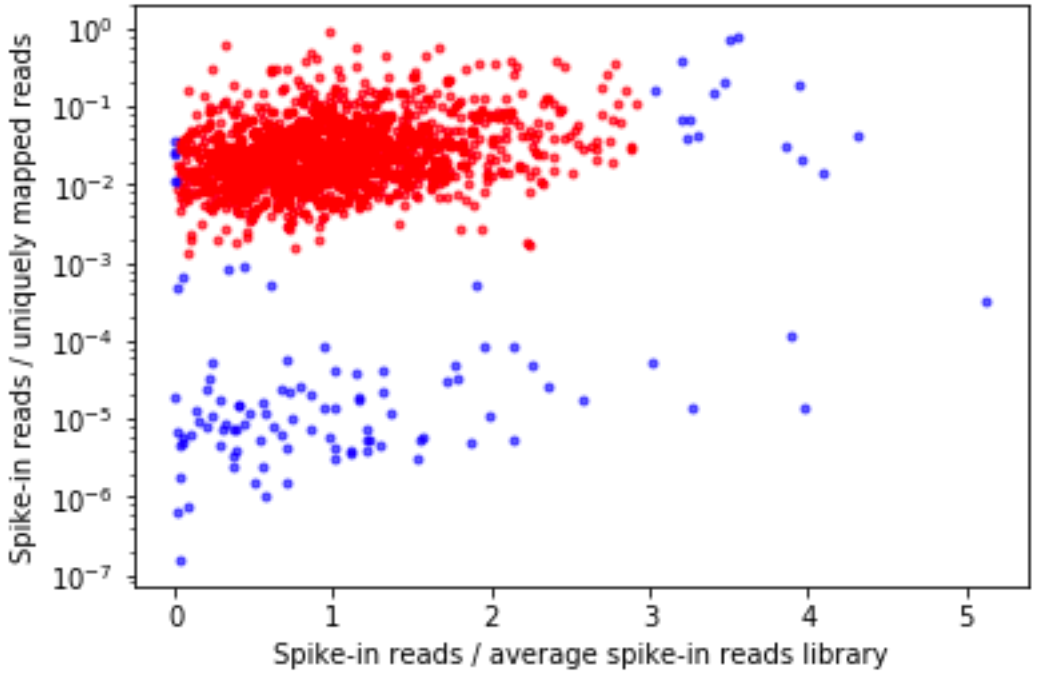
Petropoulos, Sophie, et al. "Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos." *Cell* 165.4 (2016): 1012-1026.

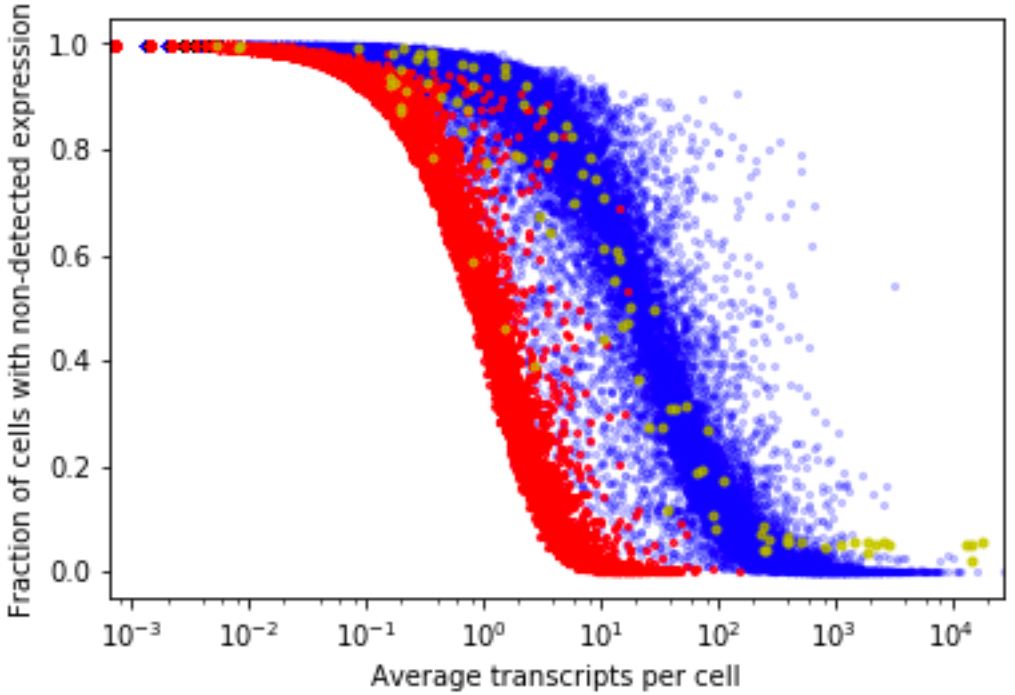
Case study: Human embryo development

- Data preparing and cleaning
  - Mapping to reference genome, counting mapped reads for each gene
  - Removing cells with low sequencing depth and low mappability

Case study: Human embryo development

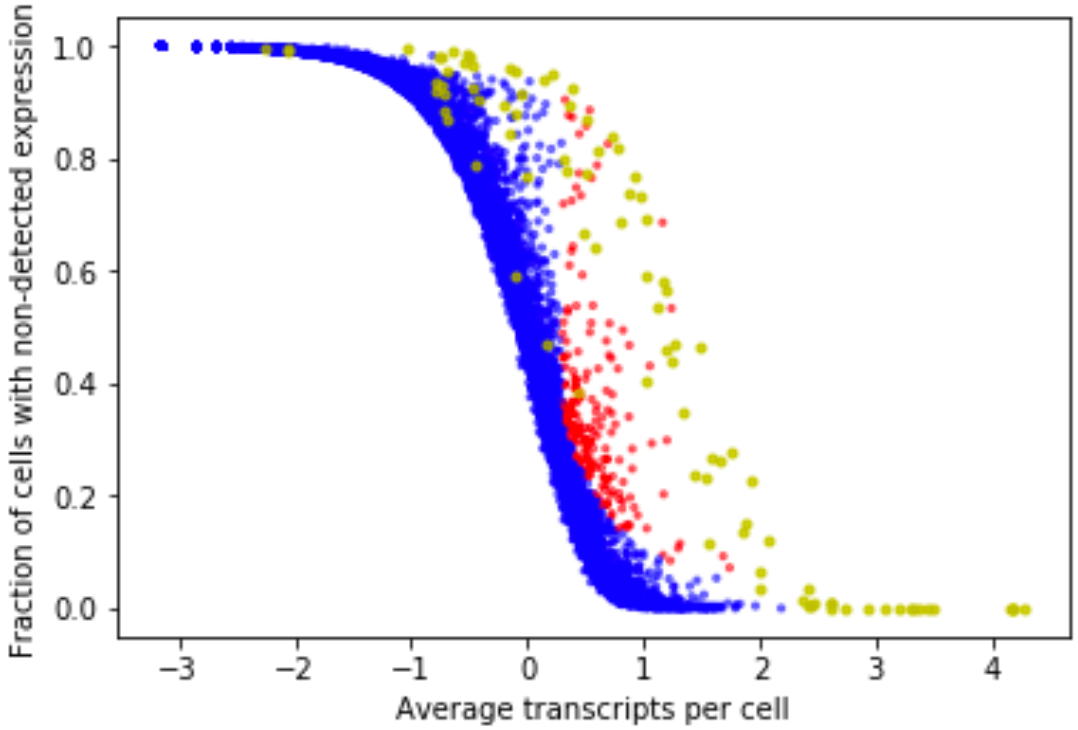
- Filter cells





Case study: Human embryo development

- Filter genes:
  - select genes with relatively high expression, and high variability

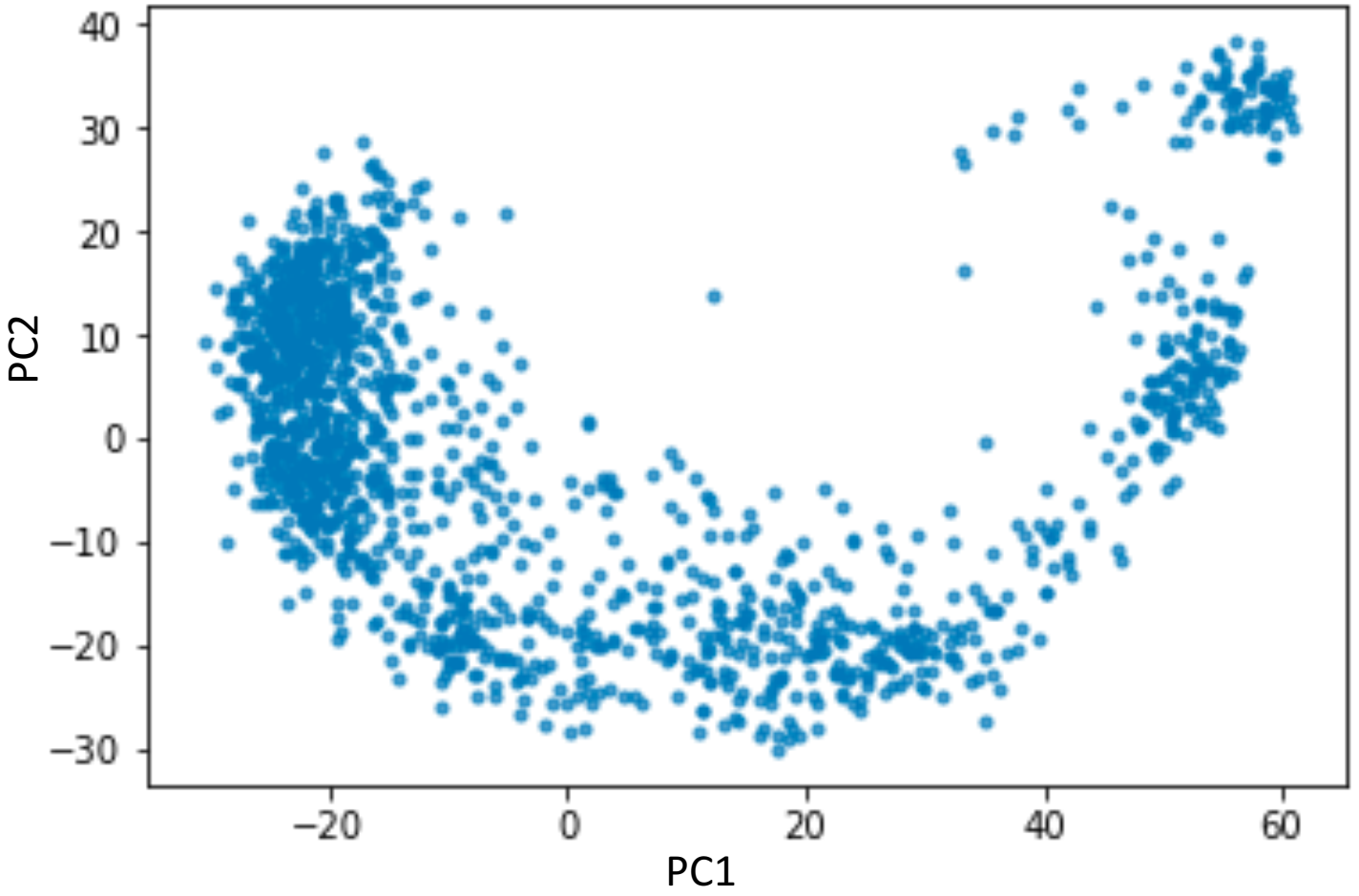


195 genes selected (**1%**)

p.select_genes(avg_counts=2.0, min_z=3.0)

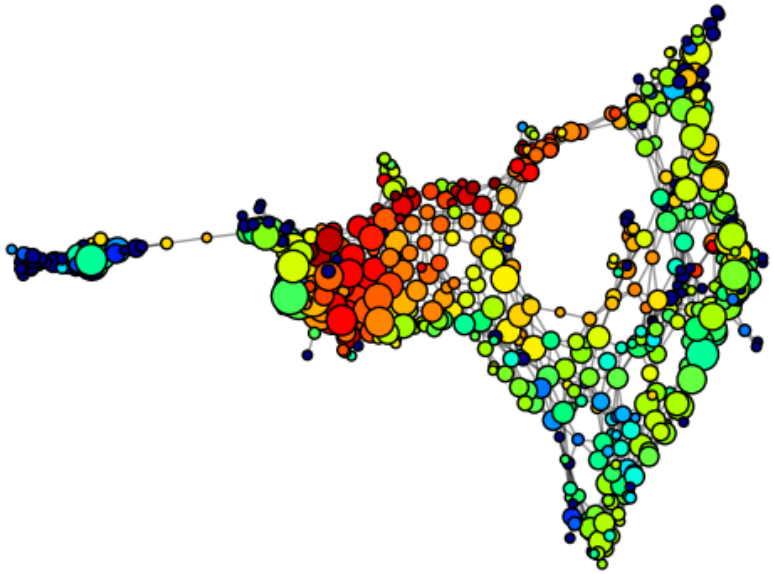
Case study: Human embryo development

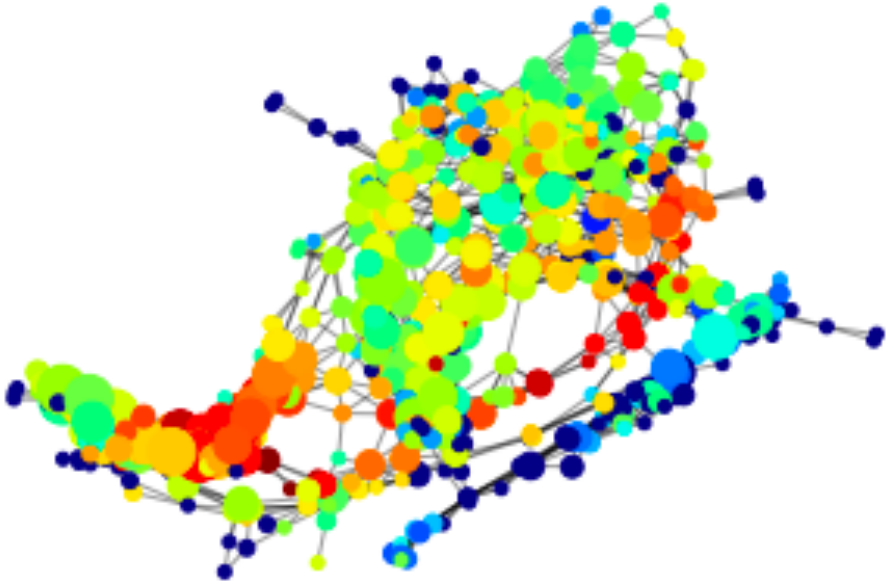
- Dimensional reduction by PCA



Case study: Human embryo development

- Topological representation based on PC1 & PC2 using Mapper
  - Parameter: 25 x 25 bins with an average of 40% overlap, unrooted

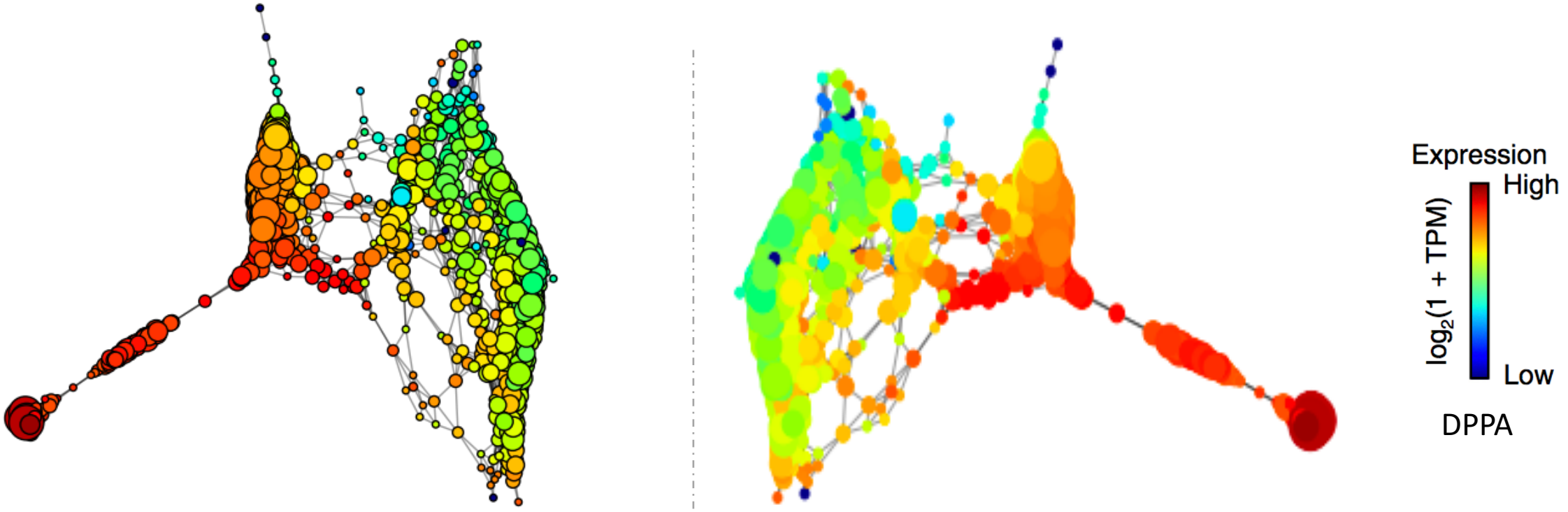




Visualisition of the network may change.

Case study: Human embryo development

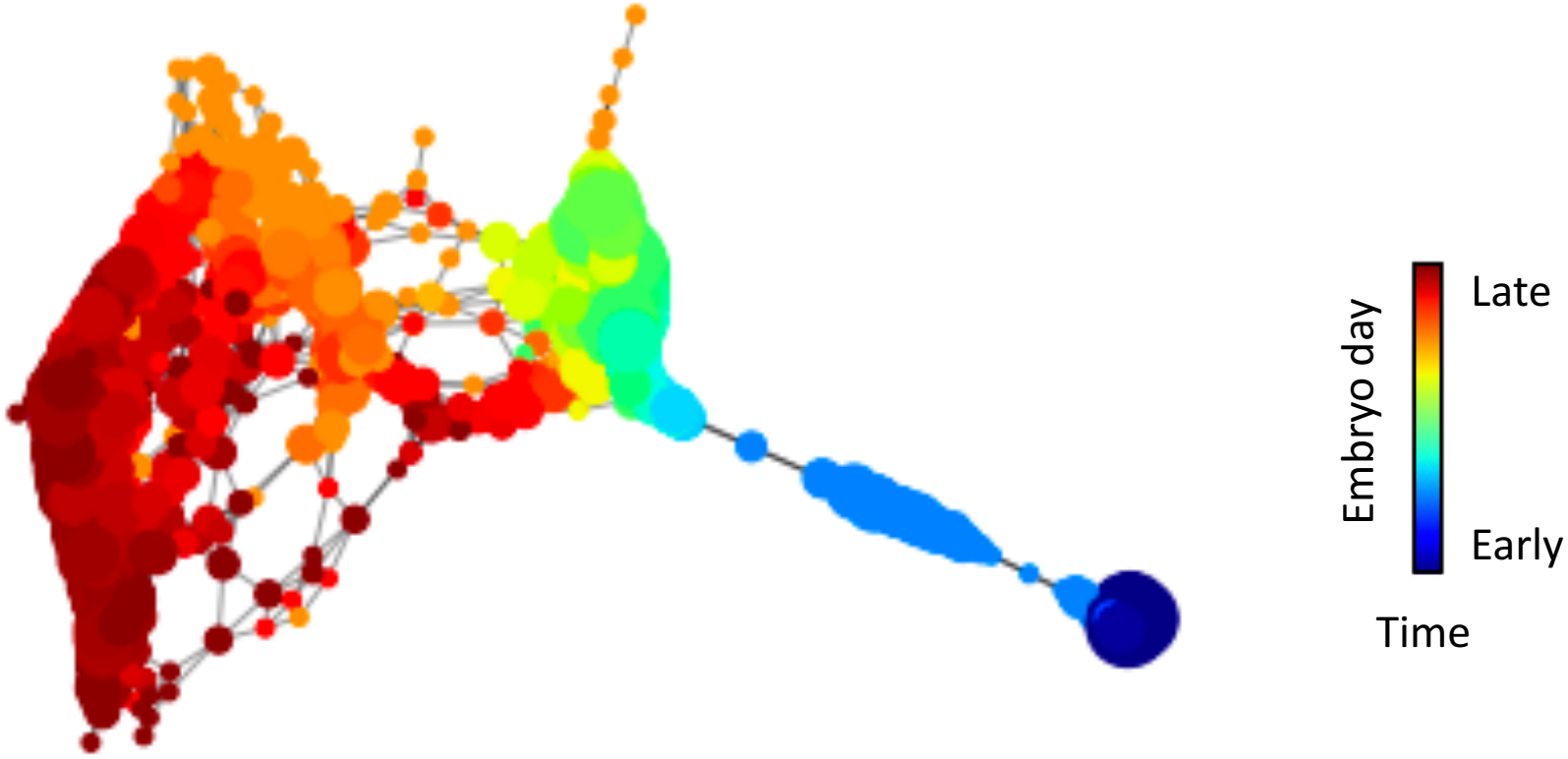
- Topological representation based on PC1 & PC2 using Mapper
  - Parameter: 25 x 25 bins with an average of 40% overlap, rooted



c = scTDA.RootedGraph('Embryo_mds', 'Embryo.no_subsampling.tsv', posgl=True); c.draw('DPPA5');

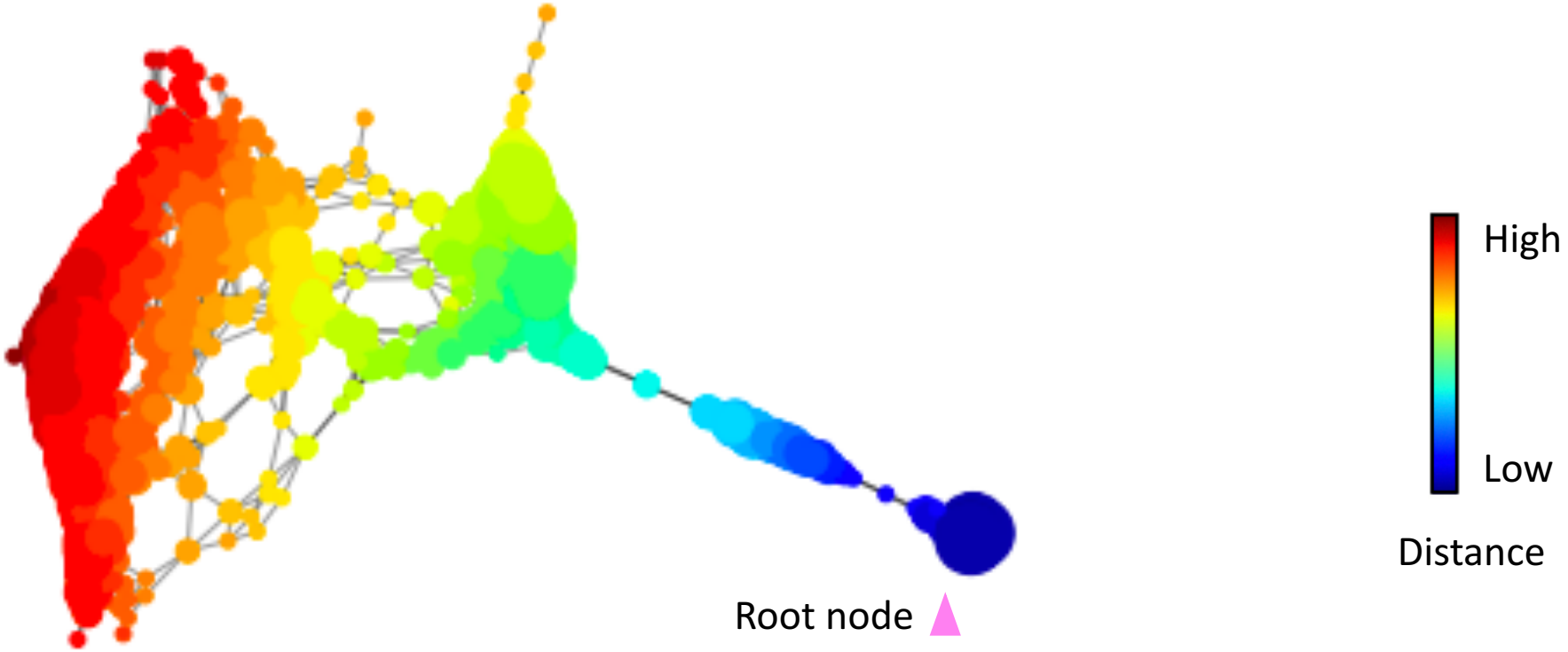
Case study: Human embryo development

- TDA correctly reproduces the differentiation time course



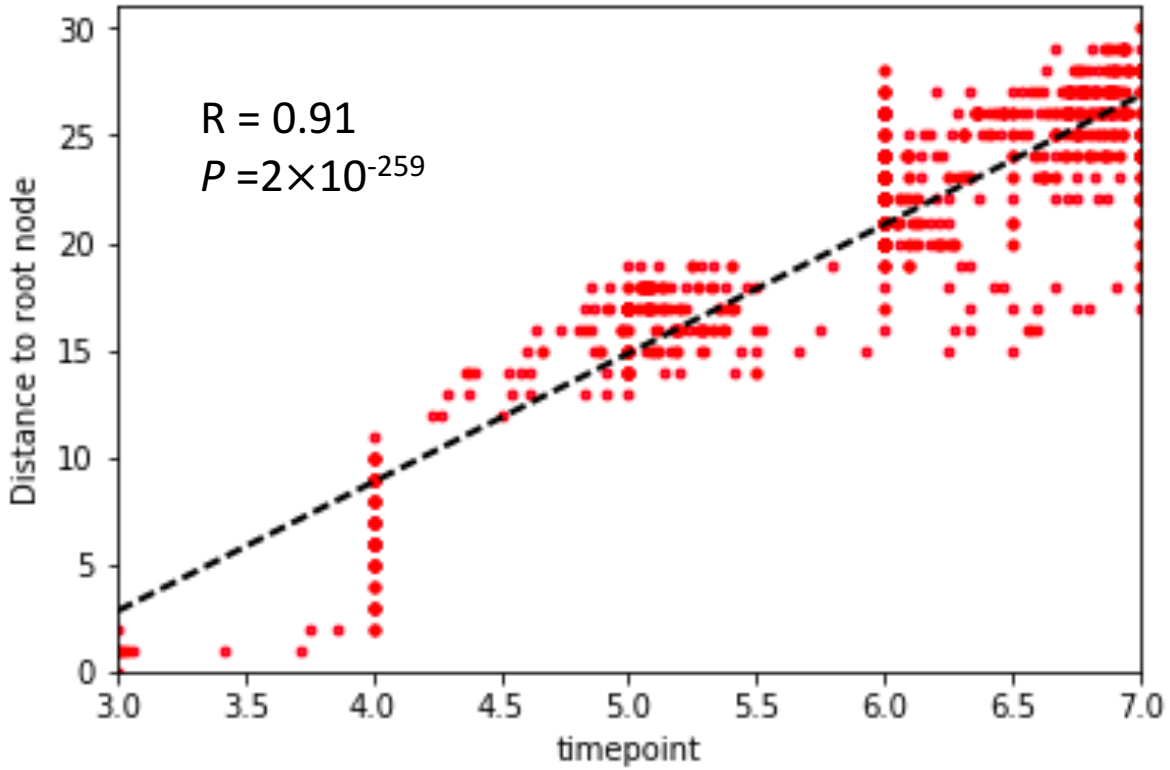
Case study: Human embryo development

- Pseudo-timing based on the topological representation



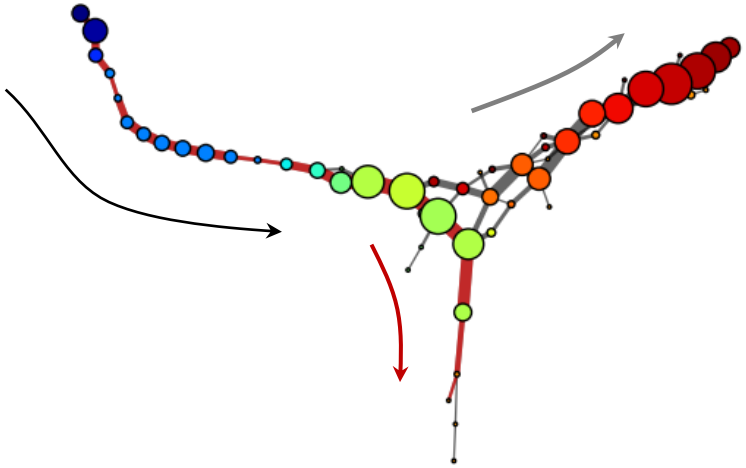
Case study: Human embryo development

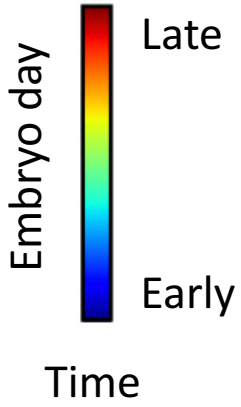
- Pseudo-time based on the topological representation



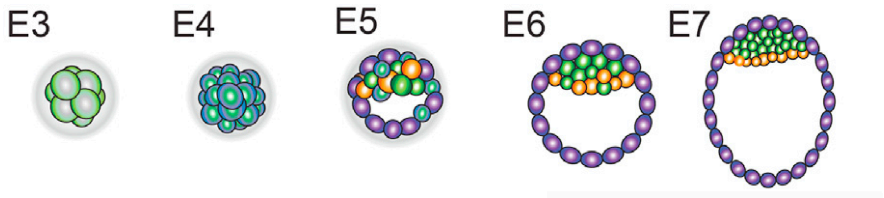
Case study: Human embryo development

- Skeleton of the network captures the differentiation process





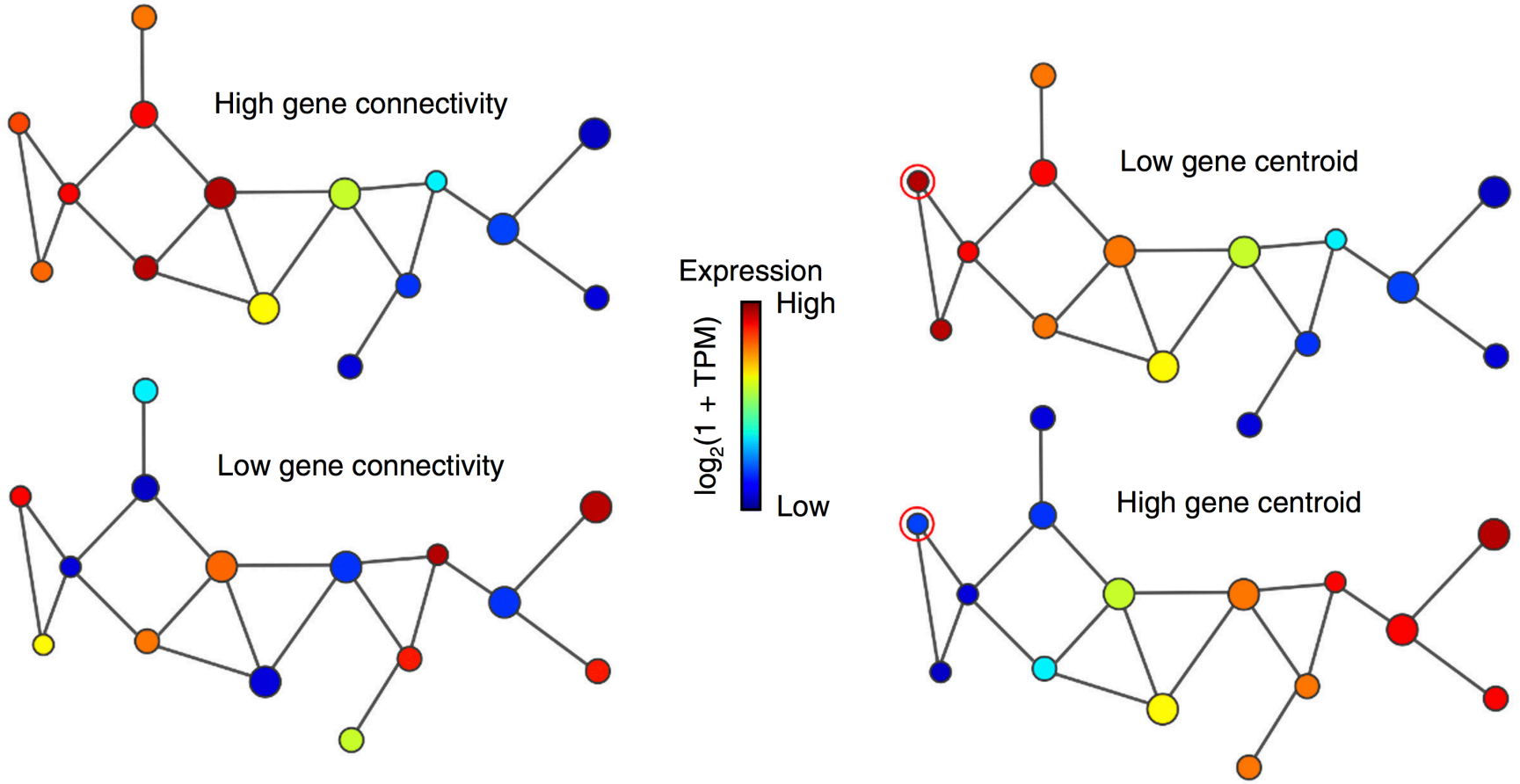
Red path: the inner cell mass
Gray path: trophectoderm




```
c.draw_skeleton('timepoint', markpath=True);
```

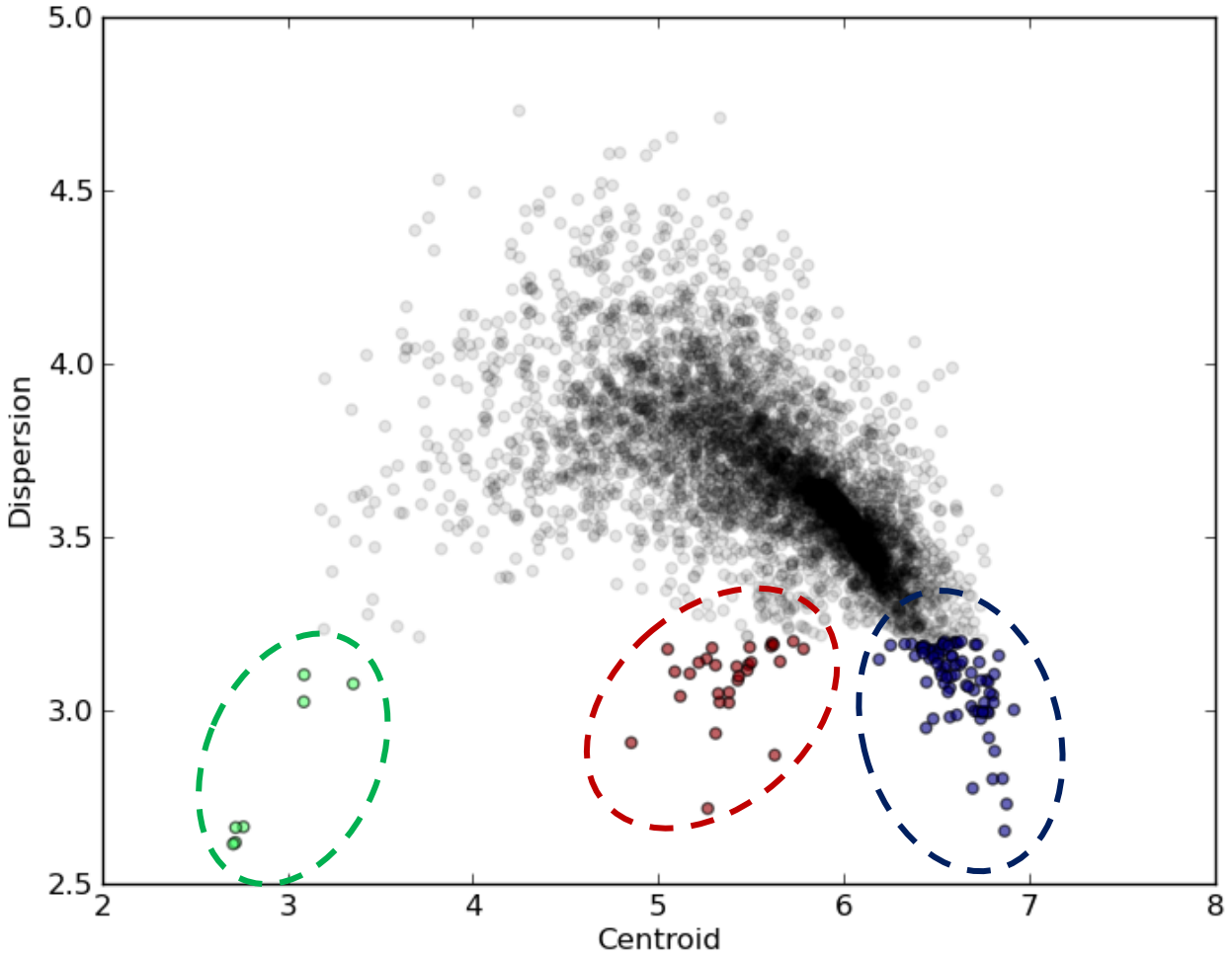
Case study: Human embryo development

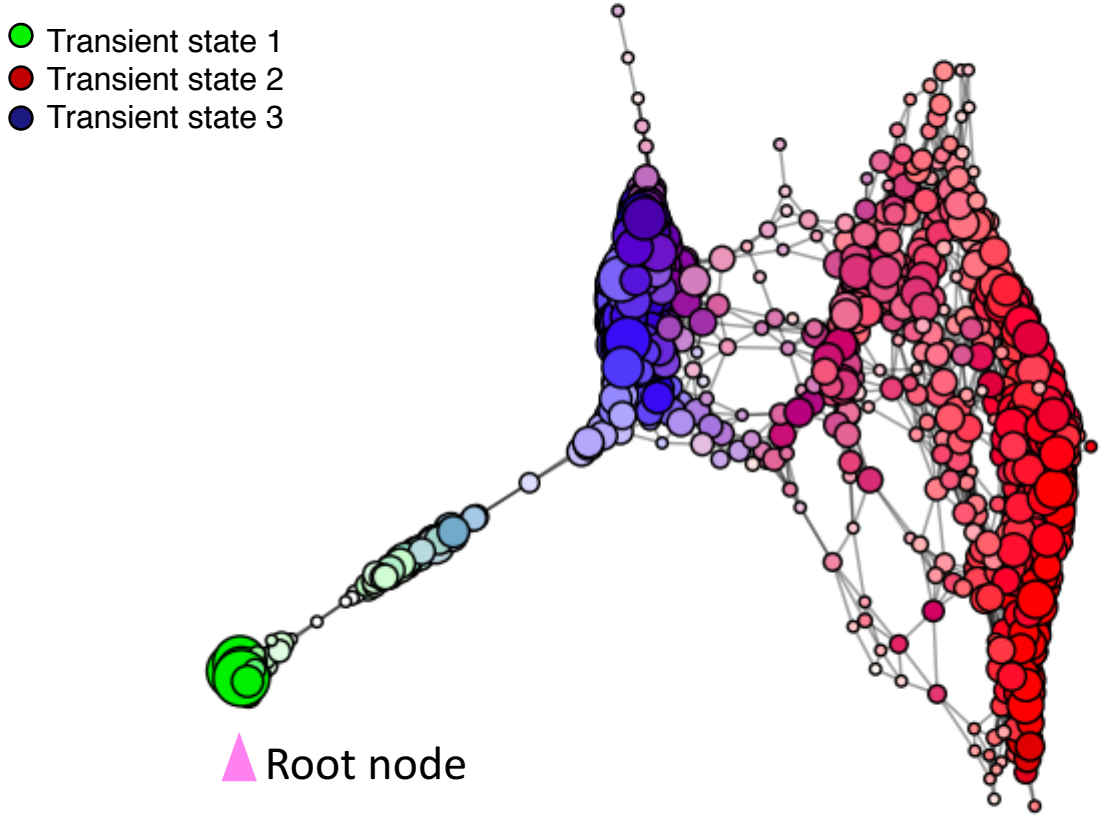
- Connectivity and centroid



Case study: Human embryo development

- Transient cellular states identified based on centroid and dispersion





Reference and links

- scTDA: https://github.com/RabadanLab/scTDA
- Tutorial: https://www.dropbox.com/s/ma80a641miteyxf/scTDA%20Tutorial.tar.gz?dl=1
- Methods: https://www.nature.com/articles/nbt.3854.pdf
- Dataset: https://www.cell.com/fulltext/S0092-8674(16)30280-X

Proposed data for course project using scTDA

LETTER

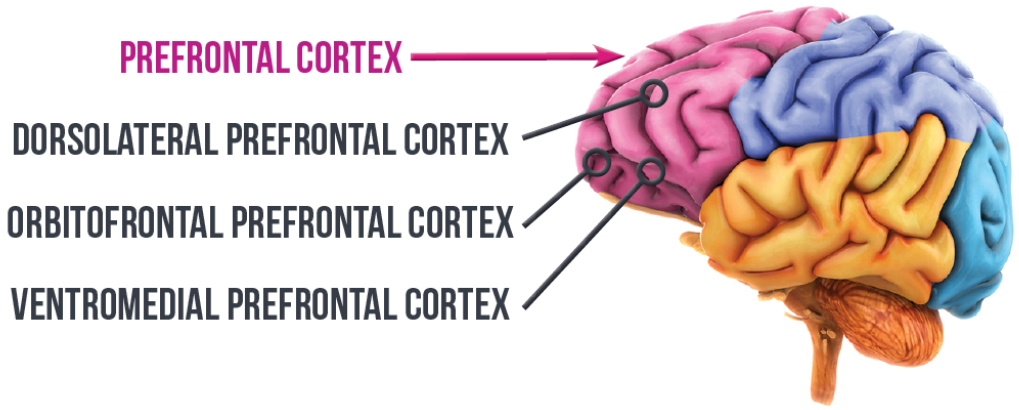
doi:10.1038/nature25980

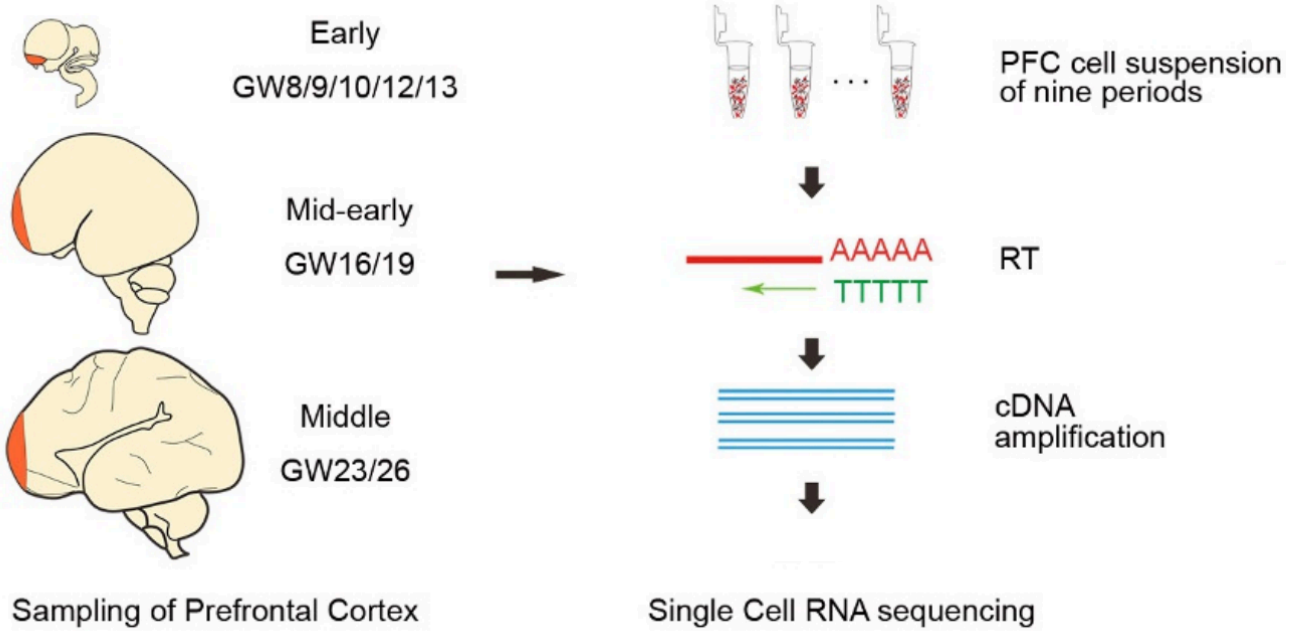
A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex

Suijuan Zhong[1,2]*, Shu Zhang[3]*, Xiaoying Fan[3]*, Qian Wu[1,2]*, Liying Yan[3]*, Ji Dong[3], Haofeng Zhang[4], Long Li[1,2], Le Sun[1], Na Pan[1], Xiaohui Xu[4], Fuchou Tang[3,5,6], Jun Zhang[4], Jie Qiao[3,5,6] & Xiaoqun Wang[1,2,7]

Zhong et al, ***Nature*** 2018

Sample and data information





PFC: prefrontal cortex; GW: gestational week; RT: reverse transcription

	GW8	GW9	GW10_01	GW10_02	GW10_03	GW12	GW13	GW16	GW19	GW23_01	GW23_02	GW26	Sum
Gender	Female	Female	Male	Female	Female	Male	Female	Female	Female	Male	Female	Female	
Sequenced cells	23	88	48	95	48	88	24	789	120	143	181	747	2,394
Filtered cells	23	88	47	92	47	85	24	776	120	132	176	699	2,309

Zhong et al, ***Nature*** 2018

Preview of the scRNA-seq data

Supplementary file	Size	Download	File type/resource
GSE104276_RAW.tar	36.9 Mb	(http) (custom)	TAR (of TXT)
GSE104276_all_pfc_2394_UMI_TPM_NOERCC.xls.gz	31.6 Mb	(ftp)(http)	XLS
GSE104276_all_pfc_2394_UMI_count_NOERCC.xls.gz	15.3 Mb	(ftp)(http)	XLS
GSE104276_readme_sample_barcode.xlsx	286.7 Kb	(ftp)(http)	XLSX

	GW08_PFC1_sc1	GW08_PFC1_sc2	GW08_PFC1_sc3
A1BG	4.54	0	0
A1BG-AS1	0	0	0
A1CF	0	0	0
A2M	4.54	0	8.87
A2M-AS1	0	0	0
A2ML1	0	0	0
A2MP1	0	0	0
A3GALT2	0	0	0
A4GALT	4.54	0	0
A4GNT	0	0	0
AA06	0	0	0
AAAS	68.1	270.63	97.62
AACS	95.34	0	0
AACSP1	0	0	0
AADAC	0	0	0
AADACL2	0	0	0
AADACL2-AS	0	0	0
AADACL3	0	0	0
AADACL4	0	0	0
AADACP1	0	0	0
AADAT	0	0	0
AAED1	0	0	0
AAGAB	0	0	0
AAK1	131.66	0	8.87
AAMDC	4.54	0	0

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104276

Acknowledgement

- The tissue donors
- The data generators