Exploring Single Cell Transcriptomes

Brett Kiyota Kieran Maheden

CPSC 547 221214

Background

- All cells have the same set of DNA and genes
- How do different cell types arise from the same set of blueprints?
- Not all genes are *expressed* (turned on) in every cell
- Expression entails *Transcription* (DNA to RNA) and *Translation* (RNA to protein)
- RNA is typically considered an inert intermediate between long term information storage (DNA) and the functional form of the information (protein)



How do we learn about gene expression?

"Adhfe1"	60	50	15	21	47	35	48	49	16	18	49	44	
"2610203C22Rik"	0	0	1	0	2	1	0	0	0	0	0	0	
"3110035E14Rik"	0	0	0	0	0	0	0	0	0	0	0	0	
"Mybl1" 229	290	471	505	430	694	289	394	585	742	613	629		
"Vcpip1"	1071	1072	839	890	1015	1304	1162	1406	866	1141	1201	1066	
"1700034P13Rik"	4	1	1	2	2	4	7	2	2	4	6	1	
"Sgk3" 213	187	484	496	666	660	168	230	459	573	688	525		
"Mcmdc2"	17	18	39	39	30	42	12	14	24	33	44	27	
"Snhg6" 256	297	234	261	348	416	244	280	248	313	325	321		
"Snord87"	5	1	3	3	12	6	4	4	8	3	7	2	
"Tcf24" 34	44	67	84	155	157	36	55	75	96	135	112		
"Ppp1r42"	0	1	1	0	0	1	0	0	0	0	0	1	
"Cops5" 1895	1503	1623	1729	2014	2120	1670	1705	1599	1879	1897	1559		
"Cspp1" 707	728	790	775	1027	1196	673	862	861	912	1141	962		
"Arfgef1"	1309	1387	1511	1615	1646	1935	1397	1642	1566	1946	1719	1450	•
"Cpa6" 0	0	0	0	0	1	0	0	0	0	0	0		
"Mir467e"	0	0	0	0	0	0	0	0	0	0	0	0	
"Prex2" 896	1003	1227	1204	1379	1741	1123	1327	1232	1650	1667	1388		
"A830018L16Rik"	155	123	111	150	235	289	138	196	115	162	234	220	
"Mir6341"	0	0	0	0	0	0	0	0	0	0	0	0	
"Gm17644"	0	1	0	0	0	0	0	0	0	0	0	0	
"Sulf1" 1687	1327	569	511	813	689	1235	1413	660	637	774	511		
"Slco5a1"	66	59	120	109	155	177	60	82	95	127	201	142	
"Prdm14"	938	781	1536	1746	1092	1366	812	848	1438	1864	1070	1038	
"Ncoa2" 1228	1152	1089	1257	1484	1787	1094	1301	1048	1358	1563	1414		
"Tram1" 2143	1840	1733	1880	2286	2418	1741	2020	1779	2211	2240	1748		
"Lactb2"	2190	1964	2959	3270	3528	3474	1972	2220	2971	3371	3573	2659	

Example count data of 12 averaged cell populations (columns) for a few given genes (rows)

- Often researchers are only interested in a single gene
- However when performing exploratory experiments or trying to characterize a new system, a more full view is more valuable
- Based on how the technologies have developed over the past two decades, RNA sequencing is the leading method for profiling gene expression
- Classically, cells are taken as a population, and RNA is isolated and sequenced, producing a count for each gene
- However this approach misses heterogeneity in the averaged sample, erasing the evidence for rare cell types or states

Single-cell RNA sequencing (scRNAseq)

14 H C 4 H	50	45	-	47		25			10				10		10						
"Adhfe1" 60	50	15	21	47		35	48	5	49		16		18		49		44				
"2610203C228ik" 0 "21 "Adhfe1" 60	50		15	21	47		35 ^µ	48		49		16		18		49		44			
DI Hactoneocoopti H o	0		1	0	2		1	0		9		0		0		9		0			
	60	50	15	21		47	35		48		49		16		18		49		44		
"Vc "102 "2610203C22Rik"	a	a	1	a		2	1		a		a		Ø		a		a		a		
"Vc "My "2610203C22Rik" "17 "My "31 "Adhfe1" "Sg "Vc "My "2610203C22R	60	-		15	21	4	7	35		48		49		16		18		49		44	
"17, "19" "31, "Adhfel" "Sg" VCI "261 apacropp "Mci "17" "Myl "261 apacropp "Snl "Sg" VCI "31" adhfel" "Snl "Sg" T7T (Myl "26102030 "Sn' "Sg" VCI "31100350 "Sn' "Sn "Sg" VCI "31100351 "Tc" "Sn "Mci "17" (Mybl1" 2 "Tc" "Sn "Mci "20" (Mybl1" 2	ir" a	60	50	1 15	a	21 2	47	1	35	a	48	a	49	a	16	a	18	a	49	a	44
"Mci "Sg "Vci 31 "26102030	22Rik"	0	0	1		0	2		1		0		0		0		0		0		0
"Sn "Mg "17("My "21100255	1/D46	0	0	0		0	ő		0		0		0		0		0		0		0
"Sn "Sg "Vc "Mubl11"	220	290	471	505		430	694		289		394		585		742		613		629		0
"Sni "%ci "31 ("My") "3610203 "Sni "%ci "31 ("My") "3110035 "Sni "Sgi "%ci "3110035 "Tc: "Sni "Sgi "%ci "3110035 "Pp" "Sni "Sgi "Vcpip1" "Ppi "Sni "Sgi "Vcpip1" "Coi "ppi "Sni "Mci "370034 "Coi "ppi "Sni "Mci "310034 "Coi "poi "Tc: "310" "Sgk3" 22 "Csi "coi "Tc: "310" "Sgk3" 22	229	1071	1072	839		430 890	1015		1304		1162		1406		866		1141		1201		1066
		4	1072	1		2	2		4		7		2		2		4		6		1
		4	484	496		∠ 666	2 660		4		230		459		2 573		4 688		525		1
"Cs "Tc. "Sn Sgk3	215	17	18	39		39	30		42		12		459		24		33		44		27
	250	297	234	261		39 348	30 416		244		280		248		24 313		325		44 321		27
"Coi pp : Sni "Sgi "1700034f "Coi pp : Sni "Mci "Sgk3" 2 "Cs : "Coi "Tc: Sni "Sgk3" 2 "Ar "Coi "Tc: Sni "Mcmdc2" "Ar "Coi "Tc: "Snhg6" 2 "Mi "Ar "Coi "Tp: "Snord87" "Pr. "Mi "Ar "Coi "Tcf24" 3 "AR "Ar "Coi "Tcf24" 3 "AR "Mi "Ar "Coi "Pp1r42" "A8 Mi "Ar "Coi "Sp1742"	250					348													321 7		2
"Mi "Csi "Pp Shord8/		5 44	1	3			12		6		4		4		8		3				2
	54		67	84		155	157		36		55		75		96		135		112		
"A8 "Mi Ar" "Cs "Ppp1r42" "A8 "Pr' "Cpi "Cs "Cops5" 1 "Mi "Ar "Mi "Ar "Cops1"		0	1	1		0	0		1		0		0		0		0		0		1
		1503	1623	1729		2014	2120		1670		1705		1599		1879		1897		1559		
"Gm "A8 "pp, "Cp "Cspp1"	101	728	790	775		1027	1196		673		862		861		912		1141		962		
"Mi "Pr' "Cops5 1 "Gm "A8 "Mi "Cp "Cspp1" 7 "Su "Mi "Pr' "Cp "Cspp1" 7 "Su "Gm "A8 "Mi "Arfgef1"		1309	1387	1511		1615	1646	•	1935		1397		1642		1566		1946		1719		1450
"Mi "AR "Mi "Copos 1 "Gm "A8 "Mi "Cp "Copp1" 7 "Su "Mi "Pr "Cp "Arfgef1" "Su "Gm "A8 "Mi "Cpa6" 6 "Si "Su "Mi " "Mird67e"	3	0	0	0		0	1		0		0		0		0		0		0		
"Gm "Mi "Pr" Cp, "Cspp1" ; "Gm "Mi "Pr" Cp, "I "Arfgef1" "Su "Gm "A8: "Pr" (Cpa6") "Su "Mi "Pr" (Cpa6") "Pr" (Su "Gm", A8 "Mir467e") "Pr" (Su "Gm", A8 "Mir467e") "Nc" SI "Gm" A8300181 "Tr" Nc" SI "Gm" Mar6341] La "gr" Pr" Su "Gm" Tacada		0	0	0		0	0		0		0		0		0		0		0		0
"Nc "S1 "Mi "Prex2" 8	396	1003	1227	1204		1379	1741		1123		1327		1232		1650		1667		1388		
"Nc "S1 Su "Mi "Prex2" 8 "Tr "Pr "Su "Mi "A8300181 "Tr "Nc "S1 "Gm "Mir6341" "La "Nc "S1 "Gm17644" "La "Nc " Su ff1" 1	_16Rik"	155	123	111		150	235		289		138		196		115		162		234		220
"La "Nc "S1 GM "Mir6341" "Tr "pr' S1 "Gm17644" "La "Nc "pr "S1 "Gm17644" "La "Nc "S1" Slo5a1" "La "La "rc "Slc5a1" "La "tr "Prdm14"		0	0	0		0	0		0		0		0		0		0		0		0
		0	1	0		0	0		0		0		0		0		0		0		0
"La "Nc("Pr "Sulf1" 1 "Tr "Nc("Slco5a1" "La'"Tr "Prdm14" ""Ncoa2" 1	1687	1327	569	511		813	689		1235		1413		660		637		774		511		
"Nc,"Slco5a1		66	59	120		109	155		177		60		82		95		127		201		142
La'"Tr. "Prdm14"		938	781	1536		1746	1092		1366		812		848		1438		1864		1070		1038
"La [,] "Nc, "SICOSAI" "Tr."Prdm14" "La [,] "Ncoa2" 1 "La [,] "Tram1" 2	1228	1152	1089	1257		1484	1787		1094		1301		1048		1358		1563		1414		
TT CHILL 2	2143	1840	1733	1880		2286	2418		1741		2020		1779		2211		2240		1748		
"Lactb2"		2190	1964	2959		3270	3528	5	3474		1972		2220		2971		3371		3573		2659

- One recent method that addresses heterogeneity is single cell RNA sequencing
- Rather than averaging the RNA across all cells, each cell is isolated first, then RNA is collected
- This method generates orders of magnitude more data
- Rather than each sample or population average having a count per gene, each cell has a count per gene
- Experiments can include >20 000 genes and >100 000 cells with multiple "batches" of cells

Cell type labelling task

- Cells are dynamic entities that are constantly changing in response to internal and external factors.
- Assigning a discrete label to cells (often based on gene expression) is common practice to ease the process of interpreting sequencing data
- Large efforts focused on linking changes in gene expression to the behaviour of certain cell types
 - Typically, disease states only appear in specific cell types
 - Connecting cell type to disease to potential intervention is a common goal



https://www.geeksforgeeks.org/structure-and-types-of-animal-tissues/

Dataset



Multimodal single cell data integration challenge: results and lessons learned

🐵 Christopher Lance, 🐵 Malte D. Luecken, 🐵 Daniel B. Burkhardt, 🐵 Robrecht Cannoodt,

Dia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, Qiao Liu, Nikolay Russkikh, Gleb Ryazantsev, Uwe Ohler,

NeurIPS 2021 Multimodal data integration competition participants, 💿 Angela Oliveira Pisco, Jonathan Bloom, 💿 Smita Krishnaswamy, 💿 Fabian J. Theis





- RNA expression data is provided as a flat table with 70,988 items (cells) and 22,050 attributes (genes)
- The value for each item-attribute pair is a quantitative value representing the normalized and transformed RNA counts (quantitative information)
- An additional 5 attributes encode:
 - cell_id (*categorical*): Unique alphanumeric string that is assigned to each cell
 - day (sequentially ordered quantitative): the time point at which sequencing was performed
 - donor (*categorical*): a unique identifying number that is assigned to the 4 healthy adult donors
 - cell_type (categorical): inferred cell type label

doi: https://doi.org/10.1101/2022.04.11.487796

Goals of our analysis project

- Explore a scRNAseq dataset using common and accessible analysis packages and analyze how the different visual encodings can affect interpretations and guide downstream analyses
 - Seurat (https://github.com/satijalab/seurat)
 - Scanpy (https://github.com/scverse/scanpy)
 - Additional softwares: trajectory inference (Monocle3), graphs (PAGA)
- 1. Evaluate how well are overall cell type differences visualized between and within timepoints.
- 2. Evaluate how well are changes in specific genes visualized between and within timepoints.

Cell types between and within time points with DR



- Dimensionality reduction is a common tool for global analysis of scRNAseq data
- Each point represents a cel
- Three common DR tools
 - PCA
 - o t-SNE
 - UMAP
- Clear differences between linear (PCA) and nonlinear DR viz
- Some noticeable differences between UMAP and t-SNE for finer details or clustering
- Qualitative interpretations
- B-cell progenitor
- Erythrocyte progenitor
- Hematopoietic stem cell
- Mast cell progenitor
- Megakaryocyte progenitor
- Monocyte progenitor
- Neutrophil progenitor

Cell types between and within time points: subsets of genes



- Three different representations of the same heatmap
- Left: each row is a cell
- Middle: each row is the aggregation of all the cells of that cell type
- Right, each row shows the distribution of expression values for that cell type

Comparing between within timepoints with density plots



Looking at individual genes

- Each cell type is represented by the density of expression values
- Each column of the graph is a selected ene
- lote that while this gives information on a single gene, it also gives us an idea of the distribution of that gene
- Note the large number of cells that have no expression level of a given gene

Trajectory inference



- UMAP dimensionality reduction
- Learn a trajectory (pseudotemporal ordering) that fits the cells' lower-dimensional coordinates (principal graph embedding algorithms)

Take away from standard pipeline

- Many options for examining expression within a single time point
- No inbuilt tools for looking at genes that are dynamic across timepoints in a given cell type
- Comparisons with DR tools also tend to be very qualitative and many key comparisons cannot be made between timepoints
- All of the tools tested were very computationally intensive issues for accessibility
 - DR plots took ~30 minutes to run per plot
 - Unable to run many existing trajectory inference softwares given our hardware (laptops)
- Dichotomy between available visual encodings provided by these widely used toolkits.
 - Attempting to capture global variation in the dataset where metrics are purely qualitative (dimensionality reduction)
 - Selection of a few specific genes for more quantitative analyses
- Future direction of visual encodings: interactivity, combine dimensionality reduction methods with visual encodings that can lead to quantitative interpretations