Bi/BE/CS 183 2022-2023
Instructor: Lior Pachter
TAs: Tara Chari, Zitong (Jerry) Wang, Meichen Fang

Problem Set 1 Solution

Problem 1

(a) The role of the UMI is to avoid double-counting molecules after sequencing due to library amplification (PCR).

(b) It's a combination of cDNA fragments obtained from reverse transcription of RNAs in single cells.

(c) inDrops and 10x genomics.
The "sub-Poisson" describes a discrete distribution defined over the non-negative integers with variance less than the mean (since the variance of Poisson distribution is the same as the mean).
In those methods, close packing of hydrogel beads decreases the variance of the number of beads per cell. (In fact, it leads to an almost degenerate distribution where the number of beads per droplet is exactly one 98% of the time).

(d) 3' sequencing here refers to the procedures which determine the parts of the mRNA that end up being sequenced as cDNA (how the mRNA is captured). In the 10x sequencing system, a 3' sequencing method, this is done through polyA-tail capture by poly-T tails on the 10x beads. These transcripts are then fragmented and sequenced, thus only cDNA containing or near the 3' end of the transcripts will be amplified and sequenced.

(e) The sequence of events for 10x Genomics sequencing is:

1. Cell capture and lysis
2. 3' transcript capture and barcoding
3. Reverse transcription and amplification
4. cDNA fragmentation and size selection
5. Addition of sample index/label (sample index PCR)
6. Single- or paired-end sequencing

Problem 2

(a) The third cell. (The third entry has the highest expression level in the second column.)

(b) The third gene. (The third entry has the highest expression level in the second row.)

(c) i. One possible way to compute the rank is using Gaussian elimination:

$$G = \begin{bmatrix} 1 & 2 & 3 & 3 \\ 3 & 1 & 9 & 4 \\ 1 & 4 & 3 & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 & 3 \\ 0 & -5 & 0 & -5 \\ 0 & 2 & 0 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

There are only two non-zero rows. Therefore, the rank of G is 2.

ii. It suggests that the expression level of four genes are not independent and two genes are redundant in terms of expression information.

(d) i. Since we have three cells, $v = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T$.

ii. Use v above, the mean expression (M) is

$$\begin{aligned} M = v^T G &= [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}] \begin{bmatrix} 1 & 2 & 3 & 3 \\ 3 & 1 & 9 & 4 \\ 1 & 4 & 3 & 5 \end{bmatrix} \\ &= [\frac{5}{3}, \frac{7}{3}, 5, 4] \end{aligned}$$

iii. The two genes with the highest mean expression are the last two genes. Thus, we can truncate the identity matrix and only keep the last two columns:

$$P = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(e) i. For the L_1 distance, using above definition, we have:

$$\begin{aligned} D_{12}^{L_1} &= 2 + 1 + 6 + 1 = 10 \\ D_{13}^{L_1} &= 0 + 2 + 0 + 2 = 4 \\ D_{23}^{L_1} &= 2 + 3 + 6 + 1 = 12 \end{aligned}$$

$$\therefore D^{L_1} = \begin{bmatrix} 0 & 10 & 4 \\ 10 & 0 & 12 \\ 4 & 12 & 0 \end{bmatrix}$$

For the L_2 distance, we have:

$$\begin{aligned}
D_{12}^{L_2} &= \sqrt{2^2+1^2+6^2+1^2} = \sqrt{42} \\
D_{13}^{L_2} &= \sqrt{0+2^2+0+2^2} = \sqrt{8} \\
D_{23}^{L_2} &= \sqrt{2^2+3^2+6^2+1^2} = \sqrt{50}
\end{aligned}$$

$$\therefore D^{L_2} = \begin{bmatrix} 0 & \sqrt{42} & \sqrt{8} \\ \sqrt{42} & 0 & \sqrt{50} \\ \sqrt{8} & \sqrt{50} & 0 \end{bmatrix}$$

For the cosine similarity, we have:

$$\begin{aligned}
D_{12}^{c} &= \frac{3+2+27+12}{\sqrt{1^2+2^2+3^2+3^2}\sqrt{3^2+1^2+9^2+4^2}} = \frac{44}{\sqrt{23}\sqrt{107}} = 0.887 \\
D_{13}^{c} &= \frac{1+8+9+15}{\sqrt{1^2+2^2+3^2+3^2}\sqrt{1^2+4^2+3^2+5^2}} = \frac{33}{\sqrt{23}\sqrt{51}} = 0.964 \\
D_{23}^{c} &= \frac{3+4+27+20}{\sqrt{1^2+4^2+3^2+5^2}\sqrt{3^2+1^2+9^2+4^2}} = \frac{44}{\sqrt{51}\sqrt{107}} = 0.731
\end{aligned}$$

$$\therefore D^{c} = \begin{bmatrix} 1 & 0.887 & 0.964 \\ 0.887 & 1 & 0.731 \\ 0.964 & 0.731 & 1 \end{bmatrix}$$

ii. For all three measures, cell 1 and cell 3 are most similar to one another.

(f) Denote the constant contamination by ϵ.
The L_1 distance is not affected:

$$L_1(\boldsymbol{x}+\boldsymbol{\epsilon}, \boldsymbol{y}+\boldsymbol{\epsilon}) := \|\boldsymbol{x}+\boldsymbol{\epsilon}-(\boldsymbol{y}+\boldsymbol{\epsilon})\|_1 = \sum_{i=1}^{n} |x_i+\epsilon-y_i-\epsilon| = \sum_{i=1}^{n} |x_i-y_i|,$$

The L_2 distance is not affected:

$$L_2(\boldsymbol{x}+\boldsymbol{\epsilon}, \boldsymbol{y}+\boldsymbol{\epsilon}) := \|\boldsymbol{x}+\boldsymbol{\epsilon}-(\boldsymbol{y}+\boldsymbol{\epsilon})\|_2 = \sqrt{\sum_{i=1}^{n}(x_i+\epsilon-y_i-\epsilon)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2},$$

The cosine similarity can be affected. To show this, we can find one counterexample. For example, let $\epsilon = \boldsymbol{x}$

$$c(\boldsymbol{x}+\boldsymbol{x}, \boldsymbol{y}+\boldsymbol{x}) := \frac{(\boldsymbol{x}+\boldsymbol{x})\cdot(\boldsymbol{y}+\boldsymbol{x})}{\|\boldsymbol{x}+\boldsymbol{x}\|_2\|\boldsymbol{y}+\boldsymbol{x}\|_2} = \frac{\boldsymbol{x}\cdot(\boldsymbol{y}+\boldsymbol{x})}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}+\boldsymbol{x}\|_2} \neq \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2},$$

if $\boldsymbol{x}$ and $\boldsymbol{y}$ are not in the same direction, i.e. $\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} \neq 1$.

(g) For the L_1 distance,

$$L_1(a\boldsymbol{x}, a\boldsymbol{y}) := \|a\boldsymbol{x} - a\boldsymbol{y}\|_1 = \sum_{i=1}^{n} a|x_i - y_i| = a\|\boldsymbol{x} - \boldsymbol{y}\|_1,$$

For the L_2 distance,

$$L_2(a\boldsymbol{x}, a\boldsymbol{y}) := \|a\boldsymbol{x} - a\boldsymbol{y}\|_2 = \sqrt{\sum_{i=1}^{n} a^2(x_i - y_i)^2} = a\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

For the cosine similarity,

$$c(a\boldsymbol{x}, a\boldsymbol{y}) := \frac{a\boldsymbol{x} \cdot a\boldsymbol{y}}{\|a\boldsymbol{x}\|_2\|a\boldsymbol{y}\|_2} = a^2\frac{\boldsymbol{x} \cdot \boldsymbol{y}}{a^2\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}$$

For some questions in problem 2, you can write code to get the answers. Here is an example notebook.

Problem 3

Here is an example solution. All reasonable answers work.