Explaining barcodes

dinsdag 4 augustus 2020

13:21

**CELSeq2 (see figure a):**

PolyA-based scRNA-seq method: PolyT stretches used in the primers to fish out mRNAs by their polyA tail.

The first primers introduce a umi and barcode at the side of the polyA tail.

Barcode = cell specific (8bp long)

Illumina adaptor = makes sequencing possible (one adapter on each side of the read) (8bp long)

(T7 promoter = needed for amplification, outside of what will be sequenced)

Second used primers bind the transcript somewhere random and introduce the other Illumina Illumina adaptor.

What will be left after library production:

R1: Illumina adaptor – ?plate specific barcode? – UMI – cell barcode – TTTTTTTTTTTTTTTTTTTTTTTT

R2: Illumina adaptor – ?plate specific barcode? – Transcript sequence (– AAAAAA)

**Plate specific barcodes** (not shown in the image) = already demultiplexed by "seqAdmins workflow" (therefore no longer in the sequences of the R1 or R2 file, but in the FASTQ-title lines of each read (6bp long))

*Our cell-barcodes and UMIs set-up differs from the original CEL-seq2 protocol, therefore we need to specify in Kallisto | Bustools, the locations of these sequences. This runs with the expectation of the R1 and R2 coming from the sequencer already been reversed (R2 -> R1 and R1 -> R2.*

**Kallisto | Bustools method argument:**

-x 0,8,16:0,0,8:1,0,0

-x bc:umi:seq

Bc = cell-barcode; in R1, from bp 8-16

Umi = unique molecular identifier (unique mRNA molecule); in R1, from bp 0-8

Seq = sequence (mRNA coding region); in R2, starts at 0 with no limit.

**Explanation by Kallisto:**

"Additionally kallisto bus will accept a string specifying a new technology in the format of bc:umi:seq where each of bc,umi and seq are a triplet of integers separated by a comma, denoting the file index, start and stop of the sequence used. For example to specify the 10xV2 technology we would use 0,0,16:0,16,26:1,0,0. The first part bc is 0,0,16 indicating it is in the 0-th file (also known as the first file in plain english), the barcode starts at the 0-th bp and ends at the 16-th bp in the sequence (i.e. 16bp barcode), the UMI is similarly in the same file, right after the barcode in position 16-26 (a 10bp UMI), finally the sequence is in a separate file, starts at 0 and ends at 0 (in this case stopping at 0 means there is no limit, we use the entire sequence)."

Van <[*https://pachterlab.github.io/kallisto/manual*](https://pachterlab.github.io/kallisto/manual)>

**The whole line as run by KB-wrapper:**

nice -n 10 kb count -i ${INDEXFILE} \

-g ${T2G} **-x 0,8,16:0,0,8:1,0,0** -w ${BARCODEFILE} \

--overwrite --verbose --lamanno -t 40 \

-o ${output\_name} -c1 ${CDNAFILE} -c2 ${INTRONFILE} \

${r1} ${r2} >> log.out 2>&1;

