**Single-cell RNA-seq data analysis in Chipster 19.9.2018**

Maria Lehtivaara, Eija Korpelainen
**chipster@csc.fi**

In this tutorial you will get familiar with Chipster's tools for preprocessing Drop-seq data from raw reads (FASTQ files) to digital gene expression matrix (DGE). DGE contains an expression measurement for each gene in each individual cell. The data used in the exercises is originally the mouse retinal cell testing data from Dr.Seq tools: (http://www.tongji.edu.cn/~zhanglab/drseq/). We use a subset of the data to make things faster.

## 1. Open Chipster

Go to **chipster.csc.fi**, and **Launch Chipster.**

## 2. Open example session

We have put the data for you on the Chipster server. Click **Open example session** in the Datasets panel (top left) and select the session **course_single_cell_RNAseq_DropSeq**.
In this example session we have imported two fastq.gz files: read1 with the barcodes and read2 with the actual RNA sequence.

## 3. Quality control with FastQC

Select both FASTQ files and tool **Quality control / Read quality with FastQC**, and click the **Run for each** button. Select one of the resulting html files and the visualization method **open in external web browser**. Repeat this for the other html file, and compare the FastQC reports.

How many reads are there and how long are they? How is the base quality?

## 4. Preprocessing DropSeq FASTQ files

Select **both FASTQ files**, and the tool **Single cell RNA-seq / Preprocessing DropSeq FASTQ files**. Click **Show parameters**, and check that the input files are correctly assigned, and that the base range for the cellular and molecular barcodes are correct. **Run** the tool.
This tool combines several steps for preprocessing DropSeq FASTQ files. While waiting, click the **More help** button, and read the manual page.

What is the unaligned BAM file for? What kind of trimming & filtering is performed?

When the 4 result files appear, select **drseq_read_1_unaligned.bam** and open it with **BAM viewer**.

Can you spot read1 and read2 sequences?

Open **tagging_and_trimming_summary.txt** as text.

How many cell barcodes passed the quality filtering? How many reads were filtered out because they were too short?

Open **tagging_and_trimming_histograms.pdf.**

Were there lots of adapters and polyA tails?

Select the new **drseq_read_1.fq.gz** file and run the tool **Quality control / Read quality with FastQC** again.

How many reads are there now and has the length distribution changed?

## 5. Alignment

Select the **drseq_read_1.fq.gz** file created in the previous step, and the tool **Alignment / HISAT2 for single end reads**. Set the **genome** to **Mus_musculus.GRCm38.92** and run the tool.
Check the **hisat.log** and **drseq_read_1.bam**.

What was the overall alignment percentage? Can you see the alignment positions for the reads? Is the BAM header different from that of the unaligned BAM?

## 6. Save the session as cloud session

Click **File** and **Save cloud session...** In the File name field type a name for the session and your name (for example *Maria_dropseq_course*) and click Save.

## 7. Merge the aligned BAM with the unaligned, tagged BAM

Now we have two BAM files: the unaligned one with the cell and molecular barcode tags, and the aligned BAM which lost the barcode information during the alignment. We want to combine the alignment and barcode information, so we merge these two files. This tool takes only the best alignment for each read from the aligned BAM.

Select **both BAM files**. Select the tool **Single cell RNA-seq / Merge aligned and unaligned BAM**, make sure the files are assigned correctly, and set **genome** = **Mus_musculus.GRCm38**.
Examine the **drseq_read_1_merged.bam** file with **BAM viewer**.

Does the BAM now have both the alignment and the barcode information?

**8. Add annotations (tag reads with gene names)**

Next, we use **drseq_read_1_merged.bam** and run tool **Single cell RNA-seq / Tag reads with gene names** and set the **GTF** = **Mus_musculus.GRCm38.92**.
Examine the **merged_tagged.bam** file with **BAM viewer**.

Can you find a read with a GE field? Which gene is that read mapping to? Notice that there's also this XF-tag now -what info does it hold?

**9. BONUS EXERCISE: View BAM in genome browser**

Select **merged_tagged.bam** and the corresponding **.bai file**. In the visualization window, select Genome browser. Choose the **Mus_musculus.GRCm38.92 (mm10)** as genome and click **Go**. Browse to gene **Rp1**. Zoom in.

Are the reads evenly distributed across the gene? Which transcript isoform seems to be present?

**10. Estimate the number of usable cells -check the inflection point**

There can be hundreds of thousands of cell barcodes in the BAM file, and we need to extract those that correspond the actual STAMPs (beads with a cell), as opposed to the "empties" (beads with just ambient RNA). To estimate the number of usable cells, we plot a cumulative distribution of reads in cells. We should see a "knee": left side of the knee shows the actual STAMPs, and the right side of the knee are the "empties".
Select the **merged_tagged.bam** and the tool **Single cell RNA-seq / Estimate number of usable cells.**
Check the results in **inflectionPoint.pdf**.

Can you see a "knee"? What is the number of usable cells (=the inflection point)? What fraction of the total reads do they contain?

**11. Detect bead synthesis errors & create digital gene expression matrix**

Sometimes problems occur in barcode generation. Beads get stuck in some phase and thus miss some synthesis cycles. To get rid of those problematic barcodes, you need to have an estimate for the number of cells in the sample (Number of barcodes = 2 x the expected number of cells). Here we use 2000. Finally, we generate a digital gene expression matrix. We use the information from the previous step to limit the number of cells and thereby the size of the matrix.

Select **merged_tagged.bam** and the tool **Single cell RNA-seq / Create digital gene expression matrix.**
Set **How to filter the DGE matrix = number of core barcodes** and **Filtering parameter = 500**.

Examine the summary file (**synthesis_stats_summary.txt**).

How many beads are there in total? How many had some sort of errors?

Examine the **synthesis_stats.txt** file.

How many molecules were detected for the first cell in the list?

Check the **digital_expression_summary.txt** file.

How many genes and transcripts were there in the cell that had most genes?

Open the **digital_expression.tsv** file. (Click "Visualise", even though it will take some time to open the file.)

How many genes and cells are there? Why do you think the counts are so low? Are there any bigger counts (you can sort the table by clicking the column headers)?