EMBL-EBI

Single-cell RNA-seq analysis with Python
17-21 February 2025

This Q&A document will be used interactively for the duration of the course. The document is split into days and sessions therefore please add your question in the correct day and session. Any general questions can be asked on Slack in the '#course_discussion' channel.

Pre-course challenges

If you have any questions whilst working through the pre-course challenges/materials, please add them below.
Q:
A:

Q:
A:

Day one, Monday 17 February 2025

Silvie Korena Fexova, Introduction to single cell technologies, Handbook link

Q: How many transcripts would be detected with the new 10X GEM-X? Similar to v3.1?
A:

Q: Do the election of the single cell preparation protocol also rely on the expected cell type to analyse? i.e. neuron-related vs. cancerous
A:

Q: What could be best sc technology for prokaryotes?
A: Same tech could work, depends what you're looking for but not all prokaryotic mRNA is polyadenylated and a lot of the current protocols depend on oligodT priming. Some wet-lab protocols might need optimisation as well due to cell walls present etc

Q: HI! Can you say something else why spec method is more sensitive to improve gene detection (less introns, rRNAs?) thanks!
A:

Q: Do you think that sorting cells by Flex sorter for example before doing ScRNA sequencing is better than doing ScRNA directly?
A:

Q: In one of the first slides, the percentage of mitochondrial /nuclear DNA obtained with 10x (I think) is shown. If we work with cells with a Chloroplast, would the % of Chloroplast DNA be equivalent? Or is it different for different organelle?
A:

Q: Are there specific single-cell platforms (e.g., Smart-seq2 vs. 10X Genomics) that are better suited for lncRNA analysis? And do you typically see lncRNA transcripts being picked up?
A:

Q: How can estimate the cost of a single cell experiment for a human sample ? what considerations should i have when I contact sequencing facility?

A:

Q: I didnt get why some single cell protocols are not recommended for alternative splicing?
A:

Wendi Bacon, Wet-lab overview, Handbook link

Q: Is the video that was showed at the beginning of the lecture available somewhere?
A:

Q:
A:

Wendi Bacon, Dry-lab overview, Handbook link

Q:
A:

Q:
A:

Tallulah Andrews, Experimental Design and Q&A, Handbook link

Q:
A:

Q:
A:

Day two, Tuesday 18 February 2025

Tallulah Andrews, Designing your analysis, Handbook link

Q:
A:

Q:
A:

Anna Vathrakokoili Pournara, Getting set up: infrastructure terms, Handbook link

Q:
A:

Q:
A:

Jiawei Wang, Raw reads to expression matrix, Handbook link

Q: So, for the threshold - since the method is adaptive, what we are saying is that it is almost completely reliable? And if so, then, the sort of in-between area that we end up getting would correspond to cells with a very little signal?
A: The **adaptive threshold** in single-cell RNA sequencing is not completely reliable, but it provides a flexible way to distinguish cells from background noise. Since the method adjusts based on the dataset, you can retain more potential cells and validate them later in **downstream analysis**. Instead of strictly filtering out barcodes with low RNA counts, you can include more cells and look into **clustering results** to identify non-cell populations later. These non-cell clusters often exhibit **distinct gene expression patterns**, making them easier to detect and filter out later.

Importantly, **low gene expression levels do not necessarily indicate weak biological signals**. The true signal in single-cell RNA sequencing is measured by **gene expression patterns across cells**, rather than just the total expression per cell. Even cells with relatively low expression can be meaningful if they follow distinct expression trends.

Q: Just to confirm: barcode and UMI are parts of bead, right? So Barcode together with UMI makes a bead?
A: A bead carries barcodes and UMIs, but a barcode + UMI does not make a bead. The bead itself is a solid-phase support (e.g., hydrogel or magnetic bead) loaded with these unique oligonucleotides.

A bead typically has oligonucleotides with the following structure:
(Poly-dT) – Cell Barcode – UMI – Adapter Sequence

Where:
Poly-dT: Captures mRNA (binds to poly-A tail).
Cell Barcode: Links transcripts to a single cell.
UMI: Uniquely labels each RNA molecule.
Adapter: Required for sequencing.

Q: Is there a difference in reference genome maps that we can use for CellRanger and StarSOLO for example? Can you also have multiple reference genomes and map to all of these separately?
A: There are differences in the reference genome formats used by **Cell Ranger (more standard)** and **STARsolo (more flexible)**. Cell Ranger requires a **pre-built reference** specific to its pipeline, which is generated using `cellranger mkref` from **FASTA** and **GTF** files. This reference is optimized for 10x Genomics data and is splicing-aware to ensure accurate UMI counting. In contrast, **STARsolo** uses a standard **STAR genome index**, which is created with `STAR --runMode genomeGenerate`. This index is more flexible and can be used for a variety of single-cell RNA-seq protocols but doesn't require the specialized format that Cell Ranger needs.

You can map to multiple reference genomes either **separately** or by using a **combined reference**. For separate mapping, you would create individual genome indexes for each species and run alignments independently, which is useful for analyzing distinct species. Alternatively, a **combined reference** merges the FASTA and GTF files from different species (like human and mouse) into one index. This approach is beneficial for **xenograft** experiments or when working with mixed-species samples. The choice depends on your experimental setup—separate references work best for pure species, while a combined reference is ideal for mixed-species analysis.

Q: What would be an appropriate follow-up when you have to many red crosses in FastQC?
A: When you see too many red crosses in FastQC, it often indicates quality issues with your raw FASTQ files, such as low-quality reads, adapter contamination, or overrepresented sequences. In this case, the appropriate follow-up is typically to perform some preprocessing steps on the data to clean it up before moving on to downstream analysis. **Remember**: you always have the flexibility to **remove part or all of the low-quality reads** based on your needs either by trimming low-quality ends or removing low-quality reads entirely.

More specific follow-up actions for your reference:

1. Low-Quality Reads Removal

    - Use tools like *Trimmomatic*, *Cutadapt*, or *fastp* to trim low-quality bases from the reads, especially at the ends where the quality is usually lower.
2. Adapter Removal
    - If adapter contamination is flagged by *FastQC*, you can use *Cutadapt* or *fastp* to remove adapter sequences from the reads.
3. Duplicate Marking
    - If duplicate reads are an issue, tools like *Picard* or *Samtools* can help mark or remove duplicates to avoid overestimation of gene expression.
4. Filtering
    - You might also filter out reads with too many low-quality bases or reads that are too short (if they are below a certain threshold of length or quality).

Q: Is it recommended to perform some preprocessing of fastq files? I mean low quality reads removal, duplicate marking etc?
A: Good point! Preprocessing is highly recommended to improve data quality and ensure more accurate downstream analysis (e.g., alignment, gene expression quantification). By removing poor-quality reads, correcting for adapter contamination, and handling duplicates, you make sure the data you're working with is as clean as possible for downstream tools like alignment and clustering. Please refer to the answer to the above question for more details.

Q: Interpretation of result tabs for FastaQC: Per Base Quality vs Per Tile Quality
A: Per Base focuses on the quality of each base in the sequence, showing the overall quality distribution across all the reads. Use Per Base Quality to assess the overall sequencing quality at different positions across all reads.

Per Tile focuses on the quality of reads within each sequencing tile on the flow cell. Use Per Tile Quality to identify specific regions of the sequencing flow cell that might have quality issues, which could suggest problems with the sequencing process or the machine itself.

This is a great example of how tools like ChatGPT can be used to quickly generate comprehensive and well-organized explanations of technical concepts. I apologize for not being able to demonstrate this during the course due to time constraints, but I'm confident that after completing the course, you'll develop a solid understanding of the overall workflow, which you can easily adapt to suit any specific details or needs (with the help of Google and ChatGPT)!

Q:

A:

Q:
A:

Yuyao Song, Hugo Tavares, Iris Yu, QC, pre-processing and normalisation, Handbook link

Q: Sometimes you do not have the raw file, how can I run it (CellBender) in that case?
A: CellBender should be run on the raw data, because CellBender will need the whole data to properly model the background to distinguish empty from non-empty droplets. If you, for example, used an already filtered file, then the output from CellBender may not be accurate. Cellranger itself can filter cells, using emptyDrops method described here (I believe - sometimes they update cellranger, and I haven't checked the very latest version. Edit: here are the details of how cellranger calls cells). It's not necessarily bad to use the filtered matrix from cellranger, followed by further QC as is being demonstrated in the session.

Q: Can pybiomart work for all species with an annotated genome in ENSEMBL? What happens to the genes that are not annotated? Do they still get recognised as genes or are they excluded from the output file?
A: Yes, pybiomart can work with anything that's on BioMart. If a gene doesn't have an annotation in ENSEMBL, then you can still keep the unannotated genes in your data. When we do the merge operation - this bit of code `.merge(h38_genes, how="left", on="gene_ids")` - you can use we chose the option how = "left", which means we want to keep everything from the dataframe on the "left" of the command, which in our case is the original genes in our anndata. This way, if there is an annotation on ENSEMBL, we will merge bring that information with the original table, if it wasn't present, then the merge function will fill the cells with missing values.

Q: Accessing the files and notebooks for the demonstration
A: Step 1 is to copy the files from penelopeCloud to your Desktop:

```
cp -r ~/Desktop/penelopeCloud/scRNA-seq_python_course_2024/ ~/Desktop/
```

Then change into your copied folder (this is where you'll work from the rest of the course):

```
cd ~/Desktop/scRNA-seq_python_course_2024/
```

Then launch a jupyter notebook from that folder by running the command:

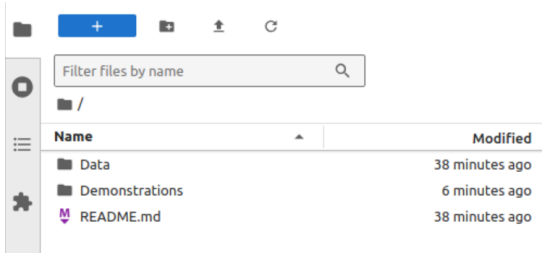
```
singularity exec /home/training/scrnaseq2024.sif jupyter lab
```

This will print a bunch of stuff on the terminal, but towards the bottom you will have a link that looks a bit like this:

http://localhost:8888/lab?token= … etc…

Ctrl + Click on that link and it will open Jupyter Lab on the browser.

Within jupyter, you should see the following in the file explorer on the left:



You can then navigate to the “Demonstrations” folder and open the notebook 02_iris_qc_normalisation_yuyao_mod_Feb2025_demo.ipynb
And hopefully everything should work from there.

Q: Would you do the same QC for Nanopore sc datasets?
A: Could be different, do not have experience, the principles would apply, probably be careful with mitochondrial count.

Q: Would you consider including ribosomal counts in the filetring?
A: personally no but for some applications maybe could play a role.

Anna Vathrakokoili Pournara, Feature selection, dimensionality reduction, clustering, Handbook link

Q: How do you set the resolution for clustering using Leiden algorithm? When to use 0,25 or 1?

A:

Q: Can you elaborate on the differences in results you would expect between the two methods to assign highly variable genes? Or would you expect the same outcome?
A:

Q: If you choose to go with DE analysis for annotating your clusters I assume it's quite important to have the right resolution. If not you might start looking at DEGs between clusters of the same celltype while watering down the actual informative DEG describing celltype A from B. How would you go about this?
A:

Q: Q: for dimension reduction speciall PCA and t-SNE, for sc-RN-seq prospectives. can i combine first PCA followed by t-SNE (local/ global data preservation?
A:

Q: Sometimes when I used "cellranger aggr" and "ad.concat" with "outer" optionl got different results. Have you ever had this issue?
A:

Q:
A:

Yuyao Song, Batch correction and data set integration, Handbook link

Q:How would one decide what exact integration method to use since there are quite many. Does it depend on what data you have? Or are there any more specific requirements? Or do people use several ones?
A: check benchmark (scIntegration Benchmarking or scIB), but yes, it depends on the data and others.

Q: So it better to use an data integration method that would take into account "batch differences" or would you still first to a batch correction and then an integration method afterwards?
A: pure batch correction method when you are sure you have technical replicates (eg when the only difference is they were in a different plate) then you can do a batch correction first.

Q: Are there any cases when there is 1 batch only? Which considerations should we take when integrating this data?
A: **No need for data integration in this case.**

Q:If I understand correctly, batch effects are corrected only for clustering, so when we come to DE expression analysis, we will also have to correct for batch effect with other methods (normalisation etc), or not?
A: Yes, you understand correctly.

Q: Can you repeat what you said about low/high expressed genes in terms of recovering high variable genes? In bulk RNA-seq we discard those genes with low counts because they tend to produce more dispersion.
A: The HVG selection algorithms in single-cell analysis try to reduce the effect of gene mean expression on gene variance, given that some genes might be lowly expressed but highly specific to some cell types. So they try to get highly variable genes per bin of genes with different expression ranges.

Q: If you can correct for the batch effects having technical replicates, and then you want to perform data integration. In this last case, you can use the corrected-pseudo counts as input of data integration methods?
A: It is possible to use the corrected counts for data integration, however in this case, you can just do one data integration, specifying integration_key and technical_replicate_key together as the batch_key, such as pasting them to create a new column to be the batch_key. Depending on the algorithm they might also have options to specify if there are true technical replicates.

Q: When you have different samples all processed in the same way but coming from different tissues labeled with "tissues" in adata (with two

replicates for each tissue from two different individuals for example) and you want to integrate over each pair of replicates to remove individual variations, is it possible to perform data integration among the replicates of each tissue separately and use sc.tl.rank_genes_groups(adata, groupby="tissues", method="Wilcoxon") to find differential expressed genes between different tissues (using a non-parametric Wilkoxon test)?
A: The issue here is that the two tissues could be processed differently and sequenced differently and so on. It is hard to tell how much technical batch effect is there, but there almost certainly is, then Wilcox could find these as the DEGs. I would suggest first checking individual-integrated, tissues-non-integrated data and see if it matches your known facts, such as whether some populations should be overlapping between the tissues. Usually, you will need to integrate between tissues again to try to remove the global expression shift, then you can just directly integrate once by a column created from pasting tissue and individuals to account for both.

Q: Regarding the two questions above: at that point, if I paste the two columns tissue + individual, it is equivalent to integrate on individual variations right?

Q: Can you define again the difference between batch integration and data integration?
A: Please can you find it in the recordings and slides.

Day three, Wednesday 18 February 2025

Yuyao Song, Feature Selection, Dimensionality Reduction, and Data Integration, Handbook link

Q: Can you go back to the UMAP interpretation again and re-explain it? Thanks

A: In general, the principle is that UMAP is a 2D visualization method for you to “see” your high-dimensional data. However, it inevitably loses some information and misrepresents some structures, so don’t take it for anything more than giving you a relatively good picture and for you to show the cell types after you annotated them. To detect cell populations, do this by clustering using a neighbour graph calculated from PCA. Try to avoid any interpretation or statement regarding distance such as “the two clusters are far from each other on the UMAP so they are not similar”. This kind of statement is not correct, because if you look at the space from another angle they might overlap with each other.

Q: if you have 4 groups instead of 2 (4 conditions) is it only the .concat method that changes? The list gets 4 elements instead of 2 or are there any more important changes in that case?
A: The only change is that there are 4 elements in the list. With more objects, it is more possible to have columns not merged due to different values or duplications in records and so on, so it might require a bit more checking beforehand.

Anna Vathrakokoili Pournara, Feature selection, dimensionality reduction, clustering, Handbook link

Q: Can you explain why we calculate silhouette_score based on the PCA embeddings ("X_corrected")?
A:

Q: is recommended/possible to check cluster consistency/coherence using gene expression profile, like GO enrichment, similar to bulk RNA seq protocol?
A:

Q: I didn't understand at which point we apply normalization due to "read sequencing coverage".. I mean how do you know a gene is more expressed in a cell and it is not a matter of sequencing coverage
A:

Q: What's the best method for subclustering? Like the B-cell sub-clustering you mentioned, do you just adjust the resolution? Or are there other methods?
A:

Andrian Yang, Differential Expression and Differential Abundance Analysis, Handbook link

Q:From which experimental step is derived the difference in abundance/coverage of transcripts) I can't see
A:

Q:
A:

Q:When studying a rare cell population, which method do you recommend? Pseudobulk or Single Cell specific (Latent, etc)
A:

Q:Makes sense to use pseudobulk when we are trying to study as a cohort and assuming each sample is a replicate in the group. But cases where there is a lot of heterogeneity in samples within a group, would pseudobulk still be advised?
A:

Q: this is specific to one of the project I'm working on combining WES and Single-Cell data. Using WES to improve cancer cell fraction in Single cell for better downstream. And also studying the samples individually instead of cohort-wise. Since we have variant information as well. What method would be recommended in such case? Or does this design impact the selection of method in any way?
A:

Q:**As I understand it, we create a pseudobulk from the raw data (which is a common and recommended strategy), then we use some DE method (ex. edgeR). Is it possible to get normalised data but without DE (previous this step), the usual, complete list of reads per gene per cell (all genes and all cells included not only DE ones). I am thinking here of something similar to TPM, FPKM or RPKM as in the case of bulk RNA-seq.**
A:

Q:
A:

Q:
A:

Q:

A:

Q:
A:

Day four, Thursday 20 February 2025

Nadav Yayon, Spatial transcriptomics, Handbook link

Q: You mentioned that there is no way to resolve the multiple nuclei in cardiomyocytes. Are there any specific tips you would take into consideration to analyse the heart?
A:

Q: What might be the limitations on using ICC, antibody-based stainings or cell painting protocols vs. H&E to capture morphological features?
A:

Q: Very cool. I can imagine you could also use a consecutive slice to perform multiplex imaging to incorporate more protein information into your morphology predictions or to validate. Do you have plans to follow up on this?

A:

Q: How could such staining affect cell's transcriptome? I guess the fixation, permeabilization might be quite harsh for RNA integrity, it would be great to have your thoughts on this! :D Super cool talk
A:

Expression Atlas team, Single Cell Expression Atlas and Single cell data submission, Handbook link

Q:

A:

Q:
A:

Q:
A:

<u>General question:</u>

Q: Can we treat singe cell sequencing as bulk rnaseq?
A:

Q: Is possible to use single cell for detecting variants?
A: