NGS-CN summer school SPATIAL TRANSCRIPTOMICS

NGS-CN

Q&A Hybrid Day 1 (online)

Talk 1 – Sequencing technologies at a glance - Maximilian Krause

Q: What about IonTorrent sequencers and limits?

A: Thank you for the question. I am sorry I did not inform you about the IonTorrent technology. It was not in the scope of my talk about high-throughput sequencing.

While IonTorrent is considered a “NGS machine”, its lower throughput (2-130Mio reads) makes it not competitive against the huge machines that spit out billions of reads these days.

It still has its value, especially in diagnostic labs. They profit from faster sequencing, often require less throughput, and the smaller dimension and investment footprint make it easier accessible. However, the lower throughput prevents it from being a versatile machine.

Actually, the IonTorrent technology was deprecated recently and is no longer supported by the company.

https://www.thermofisher.com/de/de/home/life-science/sequencing/next-generation-sequencing/instruments/genexus-system.html

Q: What would you recommend as an RNA sequencing platform and kit for sequencing RNA extracted from lung cancer FFPE samples with low QC? Are there any methods to improve the quality?

A: Here we have to separate the question into multiple sub questions:

Q1: How to get RNA out of FFPE samples
Q2: How to improve quality of extracted RNA
Q3: Which method fits best on FFPE RNA samples to create libraries
Q4: Which Sequencing platform to use

To give you the necessary background for most answers, I want to start with this: FFPE crosslinks RNA and proteins together. To get RNA out of your sample, you need to reverse crosslink - which often damages the RNA and leaves chemical moieties on the extracted RNA. Therefore it is virtually impossible to extract “fully intact” (so high quality) RNA from FFPE samples. You will always only extract broken RNA pieces - no matter how good your extraction is.

(Q1) However, you still can train and validate on different extraction methods to choose the most gentle and fastest method to extract high amounts of RNA. There are plenty of providers and kits for such methods. Usually though you have to balance between “gentle method” and “high amounts of RNA” and find a good middle ground.

(Q2) As always with RNA - the cleaner and faster you work, the better the RNA quality - after being destroyed by fixation. And the more gentle the treatment the better.
We refrain from mentioning specific companies or kits - most of them are close by and it

depends more on your training and machine availability which one fits best your needs.

(Q3) As the RNA is already broken into pieces, we would recommend NOT to use standard "polyA based RNA Library prep". Instead, a random priming of your cDNA synthesis is required. Since random priming will also work on ribosomal RNA (rRNA), these molecules have to be depleted first. And again- you cannot use polyA enrichment on degraded RNA - otherwise you only catch the very last bits of gene transcripts.
Therefore, we recommend using rRNA depletion methods for your Library preparation. All big vendors sell kits for these directions (Takara, NEB, Illumina, Watchmaker…) and they are mostly on par. We have our favorite kits, but again it is up to you to use your method of choice.

(**Q4)** Again, as we already have degraded RNA, you only get short pieces of RNA into cDNA, and thus only short pieces available for sequencing. Additionally, due to expected degradation and tough extraction from FFPE tissue, your amount of RNA will usually be limited. It therefore makes no sense to think about longread sequencing or fast methods. We recommend going for "standard" shortread sequencing that is either easily available for you, or the cheapest.

BONUS: Since your sample is already fixed, think about going towards Spatial Sequencing by designing a probe-based panel for your assay, and get the transcriptomic data in a spatially resolved fashion.

Q: Do you think 454 sequencing will still have its niche applications?

A: 454 Sequencers were cancelled in the mid 2010s. It was fast and efficient - and potentially cheaper as it used no fluorescent molecules. Still it did not scale up to the demand that was needed. Since the device is deprecated and not under maintenance, I think the answer is NO.
The IonTorrent method works on a similar principle. It has its niche in fast diagnostics with low machine investment costs. However, this will be challenged with the newer generation of fluorescent-based sequencers.

I think the main advantage was electrical readout of sequencing and with this faster speed. The new generation for this niche is the handheld Oxford Nanopore Sequencer that also delivers super long reads (but less quality), and the very new ROCHE sequencer based on Expandomers. These technologies will sit in this niche and make it useless to bring back the 454 Sequencer.

Q: What are the limitations and benefits of on machine-Analysis of RNA-seq data?

A: Either the sequencing machine itself, or the vendors nowadays offer "standard pipelines" for RNA-Seq analyses. These can be valuable for many customers as they are standardised and need little knowledge to further analyse your data.
However, the downside is that the output and customization is standardized. Meaning either you are happy with the results you get (in format and in analysis depth) or you anyways have to re-analyse this data. Any special requests (specific filtering needed for your data, other normalisation, addition of other datasets…) are not easily possible.
Additionally, often the final output of these standard pipelines is not the endpoint of your

analysis. You anyways have to get trained on the data analysis to get the interpretation and figures you need.
Therefore it is advisable to get trained (or at least understand) analysis pipelines and run each step on your own.

Q: What is the difference between PolyA enriched and rRNA-depleted libraries in RNA-seq data? Is there a measurement that can tell which type the RNA-seq data is?

A: There is a big difference. polyA enrichment only catches RNA that has a polyA tail. Mostly this is mRNA and some lncRNA. But not all organisms have long polyA tails. Plus you need good quality RNA otherwise you only sequence the 3" end of RNA and not the whole body

rRNA depletion data is richer. It is not only mRNA but also lncRNA, miRNA, snoRNA and other RNA species. Furthermore, not fully spliced mRNA will also be sequenced. Additionally, you do not only sequence from the 3"end, but you target the whole RNA body. By this you get a bigger representation of the total RNA being present. You detect more. And it is more species agnostic.
The Metadata should specifically mention which kit was used. If that is not the case, one should check for intron retention amount and expression of small RNAs (snoRNAs, miRNAs etc). These should only be present in high amounts if rRNA depletion was performed.

Q: What are your experiences with different expression patterns depending on the sequencing technology? For example, we have seen that we get different marker genes depending on whether we use 10x or PACBIO short-read sequencing.

A: Yes this is possible. Also as mentioned, even the sequencer you sequence on can have an effect. So you need to be aware of this and correct for it in analysis. I hope a bioinformatician can answer more specifically on how to correct for this.
These differences come from Library prep biases, Sequencing biases and technical biases in the protocol. How strong these effects are is different too - but usually the biological effect is bigger than technical effects.

Q: Which of these methods are categorized as droplet and non-droplet methods?

A: The most known Drop-Seq: single-cell RNA sequencing (scRNA-seq) technology. separates cells into droplets, enabling the profiling of thousands of cells in parallel.

Non-Droplet Methods:

- Sanger Sequencing (Chain Termination Method): determines DNA sequences by adding labeled nucleotides one by one, stopping the reaction when a fluorescently labeled nucleotide is incorporated.

- Next-Generation Sequencing (NGS):

NGS encompasses a variety of techniques like pyrosequencing, sequencing by ligation, and sequencing by synthesis, which all sequence DNA fragments in parallel using high-throughput platforms.

- Other Non-Droplet Techniques: includes methods like Illumina sequencing, Ion Torrent sequencing, and e.g. PacBio RS sequencing, all of which utilize different principles to sequence DNA without using droplets.

Talk 2 – From Cells to Reads: Introduction to single-cell techniques - Susanne Reinhardt

Q: I did not understand perfectly how non-barcode containing molecules are discarded or are selected not to be sequenced

Question was answered verbally

Follow Up: Yes, it helped to confirm what I understood. So, from what I got, the fragments of non-barcoded molecules are just flushed away during PCR, either because p7 primers ligate to each other and, thus, the molecule does not get amplified, or because there is an active selection process for barcoded molecules. Am I understanding it right?

Follow up answer: Yes, that is correct. The constructs with both ends P5-P7 get "actively" exponentially amplified, whereas the P5-P5 constructs do not amply and thus get "diluted" (in relation to all available transcripts).

They may contribute to a small percentage in the final library, but they cannot form clusters on the sequencer because they need P5 and P7 tail for this process.

Q: Can you comment on how single cell sequencing methods are adapted for bacterial cells? Given the absence of a small polyA tail and would have lesser RNA content/cell compared to eukaryotes.

A: Bacterial single cell sequencing is not very widely used. We tested it a while ago with the plate-based Smartseq2 and we could enrich cDNA. It is a bit more difficult because the transcripts are longer than eukaryotic RNA because of the Operon-model. Furthermore, bacteria are quite small - so the capture efficiency with fluidic-based methods is potentially not good. But I have no experience here - so cannot really give a clear answer.

Q: Are you planning to have CITE-Seq at DcGC as well to profile the transcriptomic as well as proteomic signatures of cell types?

A: We already established CITE-seq. I did not mention it in the talk to decrease the complexity of the talk. So just approach us in case you plan an experiment. :)

Q: What is the reason one should select between single nuclei sequencing vs single cell sequencing? Or is it mostly based on the biological question?

A: The RNA from nuclei is roughly only 10% of the RNA of a complete cell. So you lose quite some information when sequencing nuclei. So, if you have the choice, I would recommend to sequence cells. But this is not always possible - for example when starting from difficult tissue (frozen material, lung, ...) or huge, sensitive cells (like hepatocytes) ending up in a low percentage of viable cells. Then it is better to get good quality data from nuclei instead of background noise from disrupted cells.

Q: What about the limitations of single-cell transcriptomics applied to plant tissues regarding the sorting of single plant cells without degrading the cell wall? Is it feasible?

A: Plant cells are a bit tricky. You need to remove the cell walls, otherwise the cell lysis does not work. As far as I am aware, researchers prefer to isolate nuclei from plants (instead of cells). This is also advantageous to avoid sequencing a lot of chloroplast-RNA.

Q: In single cell approaches you are capturing RNAs coming from nuclei, right? What about if I am specially interested on capturing also RNAs in the cytoplasm, e.g, derived from cell condensates?

A: Most single cell experiments focus on cells (thus including cytosolic RNA). Sometimes the cell quality is not sufficient because you may need to freeze the tissue before further processing. When you do not reach sufficient viability, then you should consider sequencing nuclei instead to improve your data.

Q: Is there a possibility to pre-save cell death/ to save cells from dying ? With a specific solution or method ?

A: Unfortunately, there is not really a good solution. Otherwise everyone would use it.

Try to use more gentle enzymes (e.g. Papain over Trypsin), keep dissociation times as short as possible, be gentle with the cells (slow pipetting to reduce shearing forces)...

Q: Is there also funding for spatial transcriptomics or only single cell sequencing?

A: The funding is for all kinds of sequencing is supported by the DFG (could be bulk RNA seq, whole genome sequencing, single cell or spatial approaches). The only requirement is that you will need a certain budget (0.1-1 mio. €)

Q: thank you for the insightful presentation, my question is about Funding opportunities, is it only for Deutsche researching or also for international institutions?

A: The DFG funds projects only within Germany. But if you have a German research group with whom you can collaborate with, you can apply together. This is then possible. :)

Q: In my group we are interested in doing single-cell transcriptomics of intact plant cells due to our RNAs of interest are located in the cytoplasm inside membraneless organelles (cell condensates) like Processing Bodies or Stress Granules among others. This is why this is our special need and my question about compatible single-cell approaches… Do you have experts working on plant systems to know their opinion? Thanks in advance.

A: Thanks for explaining more in detail. In the NGS-CN we have unfortunately little experience with plant cells. We had a test run (and failed because of the cell wall...), if you contact us via e-mail, I can do a little research to give you more information..

Talk 3 – From Reads to Counts: Introduction to single-cell RNA sequencing data - Andreas Petzold

Q: I'm currently working on a circRNA project using TCGA RNA-seq data. They provide BAM files with single-end reads (mostly).

What I'd like to do is filter out samples that are polyA-enriched, as they are not suitable for circRNA detection. If I can classify the samples as polyA-enriched, rRNA-depleted, or RNase R-treated, I can proceed directly with circRNA analysis.

A: I would suggest converting the BAM files into FastQ files (samtools fastq). Then you can identify reads with polyA using cutadapt, fastp or bbmap. These tools typically tell you which reads they modified. Get the read names that have too much polyA. Then you should be able to filter the original BAM files using this list with samtools (or bamtools).

Q: What do you mean by p5 and p7 in the library construct and what are the significance.

A: P5 and P7 refer to so-called grafting parts that are used to attach the construct to the flowcell. The P5 is the grafting part at the 5' end of the construct and the P7 is at the 3’ end. For sequencing R1, the construct is attached with the P7. In paired-end sequencing, the fragment is then turned around and attached using the P5. Then sequencing of the R2 starts.

Checkout the Illumina webpage or https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html .

Q: how do you handle multi mapped reads or ambiguous transcripts especially for genes with overlapping isoforms or pseudogenes ?

A: Cell Ranger discards them unless one of them is an exonic locus. Ambiguous transcripts do not pose a problem because Cell Ranger just checks whether there is one compatible alignment (to one isoform). However it does not do counts by transcript isoform but only by gene. Overlapping genes (that are on the same strand) however are a problem.

As far as I know, pseudogenes are removed from the gene annotation before Cell Ranger runs the analysis. So they are not included

Q: If my data has translocation genes, what are the parameters that I can adjust during the alignment step and I would appreciate it if you have any recommendations in general for analysis of data like this? (sorry I mean fusion genes)

A: I do not think that you can do this with Cell Ranger. However there are for sure specialized pipelines for this. A google search "fusions gene rna seq single cell" brought up scFusion which might be what you are looking for.

Alternatively, you could try to run a fusion gene tool developed for RNA-seq data. First, you would run Cell Ranger and then extract the reads with their cell barcode and umi from the BAM file (samtools fastq). These reads would be the input for your fusion gene tool. Once you have called fusion events, you would need to find out for which reads and which cell barcodes support fusion events (so to say get counts per cell barcode and fusion event). Then you would count UMIs

Q: Is Cell Ranger compatible with data from platforms other than 10x Genomics?

A: Andreas: No it is only compatible with 10x data. Other platforms typically have their own pipelines which should do similar stuff. Or you can use STARsolo, kallisto+bustools, alevin..

Q: To what extent these barcode errors can be considered in the filtering step? One nucleotide error or two or three?

A: Cell Ranger corrects at most 1bp.

Q: Are barcodes susceptible to e.g deletions or insertions of nt's? Is Cellranger whitelisting taking this possibility into consideration? Should we adjust the sensitivity?

A: Since the barcodes are at the beginning of read R1, they should have a high sequencing quality (which then typically declines towards the end of the read). However, when your library is sequenced with other, problematic, libraries (often low-complexity libraries), this could lead to Ns (not so much insertions and deletions). No, Cellranger does not take into account indels because that would make correction too complicated.

Q: Could you explain the two-step mapping one more time? thanks.

A: In the first step the reads are mapped to the genome. Reads that are unmapped, mapped to regions between genes (intergenic) and map to multiple regions are discarded.

In the second step the reads are matched/mapped against the transcript annotation to make sure that they overlap a known transcript. All reads that do not, or map to the opposite strand, or could map to multiple genes are discarded.

Q: How can I run cell ranger if I do not have a linux system?

A: There are a few options:

- 10x offers a cloud service (might cost, privacy issues)
- You can contact your local computing centre to set up an account there and then analyse remotely
- There might be some pipelines that are R/python based. You can install R/python on a non-Linux computer
- You actually can install Linux on your Windows PC
- Cell Ranger may work on Mac (do not know)

But I would encourage you to get a Linux system. For example, Ubuntu is very similar to Windows and user-friendly. These systems just give you much better performance (often have more cpus and RAM) which is crucial for single-cell analysis steps.

Q: I know that some people use some quality trimming using tools, e.g. fastp, for bulk RNASeq (I am aware this is not fully consensual). Is it recommended to use these tools in a single cell? And does Cell Ranger have tools or are there other tools?

A: When you work with 10x data and use Cell Ranger, read trimming is not needed (because Cell Ranger takes care) and should not be done. If you use one of the other tools for 10x data, it may be helpful but you need to be careful to trim the correct part of the construct.

However, in our experience, read trimming helps with SmartSeq2 data. The SmartSeq2 adapter sequence can be mapped apparently quite easily (or the mapper tries really hard). This leads to spurious counts for genes that are not really expressed.

Talk 4 – Analysis of single-cell RNA sequencing data - Katrin Sameith

Q: What do you mean by shallow sequencing?

A: Shallow sequencing refers to a lower sequencing depth, i.e. sequencing less reads per cell

Q: Do we lose any transcripts in scRNA-seq compared to the bulk RNA-seq? If so, is there any data to know the average loss of transcripts in ScRNA-seq and SnRNA-seq compared to the bulk RNA-seq?

A: You will likely lose very lowly expressed genes. I don't know if there is a systematic comparison out there, but it might be worth checking.

Q: What if we have cells that are enriched in mitochondria? How can we filter critically?

A: It depends on your research project and what you expect. Sometimes we exclude all cells with % mitochondrial reads above 5%, sometimes 40%.

Q: Could you please explain in detail how we would set the thresholds for low nfeature and high mitochondrial percentages?

A: In R, I exclude all cells where the % of mitochondrial reads is above a threshold. I determine the threshold visually based on the QC plots that I showed.

Q: This may be a very specific question but, I would expect more metabolically active cells (e.g. cells from brown adipose tissue) have more mitochondrias and, thus, have more mitochondrial reads. Is there any suggested change in the analysis for the first step (filter low quality cells)?

A: Yes, you need to leave the cells in :) ... I worked on a project where we set the threshold to 40%, because my collaborator expected stressed cells. You just have to loosen the standard thresholds a lot I think.

Q: Which one has more impact on the cluster number- PCA dimension or cluster resolution?

A: The number of PCs influences the whole analysis including the UMAP. You need to find a good number. The cluster resolution influences the number of clusters, i.e. the granularity of clusters.

Q: What about clusters 3 and 4?

A: If I was to analyse the data properly, and not just for the talk, I would have a closer look. Do different resolutions merge them? Are known marker genes expressed in both or just a part of them? Are the cluster marker genes expressed in both clusters, or parts of the cells?

Q: Can we consider low/absence of expression in a set of genes for criteria to identify a cluster (e.g. cluster 2 has almost no expression of genes expressed across the other clusters)?

A:Yes, I do that. For demonstration purposes, the dot-plot focussed on genes that are higher expressed.

Q: What is the ideal number of marker genes to annotate a cluster?

A: I am not sure there is an ideal number. I rank the genes based on foldchange and adjusted p-value.

Q: Which species does Enricher support for cell type assignment?

A: I don't know by heart. Check out the website, they have a range of supported species!

Q: How did you do the mapping? Is it a deconvolution method (maybe integration or deep learning prediction)? And what about the other clusters with overlapping markers?

A: This was a published dataset, where mapping was done already. We did run integration to harmonize the two samples.

Q: How can we do filtering for SnRNA-seq, which doesn't have mitochondrial genes?

A: You can still look at the number of counts and detected genes per cell.

Q: If I found a contaminated cluster and I want to remove them, do I have to start again from the count matrix then normalize and so on to have the new clustering?

A: Unfortunately yes :) ... you start again, remove the cells in the beginning, and start the analysis all over.

Q: Thanks for the informative lecture, I would like to ask, which tool should we use for quality control? I run my fastqc data on cellranger and get the output of "filtered_feature_bc_matrix.h5", is this file already after the quality control filtration step?

A: This file contains the gene counts table for all barcodes called as cells. All barcodes not called as cells are removed. This means you still have to do QC in your analysis.

Q: I have cells that fuse and have a lot of mitochondria. How can I set a cutoff or threshold for mitochondrial percentage in this case?

A: I would look at the distribution on the UMAP and violin plots, to make sure the cells with high mitochondrial % are your fused cells. In this case, I think you might have to skip the filtering.

Q: What is the relationship between PCAs and UMAP? Are the PCAs used as input for UMAP, or is it only the highly variable genes?

A: Yes, the PCA results are the basis for the UMAP.

Q: Is there any "Gold Standard" to choose for the best resolution?

A: There are some recommendations and defaults, e.g. I often use 0.5. But really you have to optimize it for each and every dataset.

Q: This may be too technical but, is there any information that log-normalisation (especially when one gets reads per 10k reads) leads to compositionality issues?

Since this is very technical, we can discuss this after the course :)

A: Ok, happy to do that! :) I am afraid I never looked into this.

Q: What makes a cloud of cells separate from others - number of genes, or number of variable genes, or number of unique genes?

A: The UMAP is based on the PCA results, hence the similarity of gene expression profiles between cells in general. UMAP looks at similarities in high-dimensional space and tries to project it into low-dimensional space.

Q: If I have too many clusters, how do I reduce the number in R?

A: Well, start with the resolution parameter :), if I understand your question correctly.

Q: For what has been discussed, could it be combination of gene levels (e.g. low expressed gene A and high gene B) a solution to characterize clusters more clearly

A: Yes, sure. This is a process, and you will look at many genes. Cluster marker genes on the one hand, and known marker genes on the other hand. Usually my collaborators know their genes of interest.

Q: Is there any way to reduce the mitochondrial content during cell preparation, if the cells of interest are very small?

A: I am not sure about the experimental part, and what the influence of cell size is on the analysis results.

Talk 5 – Decoding the Human Striatal Landscape: Leveraging snRNA-seq and Data Integration Strategies in Multi-center studies - Shobbit Agrawal

Q: Do you think the best method for single-cell data integration should prioritize correcting batch effects over preserving biological variation between cells, or vice versa?

A: When integrating single-cell data, neither approach should be absolutely prioritized over the other - it's about striking the right balance for your specific biological question.

The ideal integration method maintains a crucial balance: removing technical batch effects while preserving true biological variation. This balance depends entirely on your research question. If you're studying subtle cell state differences, you might prioritize preserving biological signals even at the cost of incomplete batch correction. Conversely, if you're establishing a reference atlas from multiple datasets, stronger batch correction might be warranted.

If you have very strong batch effects, prioritize batch correction first and then tease apart the biological aspects using post-hoc methods on clusters and other downstream analyses. For advanced approaches, consider DRVI (Disentangled Representation Variational Inference), which is an unsupervised deep generative model that learns nonlinear, disentangled representations of single-cell omics.

You can also apply methods like CNA (Co-varying neighborhood analysis) to identify cell populations associated with phenotypes of interest from single-cell transcriptomics after batch correction. CNA helps identify biologically relevant signals that may have been partially obscured during strong batch correction by analyzing the co-variation patterns in local neighborhoods of cells.

The best approach is often dataset-specific - comparing multiple integration methods (Harmony, Seurat, scVI) on your particular data and evaluating both batch mixing metrics and biological signal retention. Look for methods that allow parameter tuning to find the optimal balance for your specific biological question.

Q: Given that inter-subject variation dominates over inter-site technical variation, how can one statistically distinguish subtle technical artifacts from genuine biological differences when integrating multi-site spatial transcriptomics data?
A: Use a process of elimination approach—first remove clear quality control failures and identify outlier subjects. Technical variations that don't alter your scientific conclusions can often be tolerated. The impact of technical artifacts depends heavily on your specific research question.

Employ spatial context as a validator—genuine biological signals typically show consistent spatial patterns while technical artifacts appear randomly distributed. Anchoring analysis to known anatomical landmarks can help separate real biological variation from site-specific technical effects.

Consider specialized methods like MRVI that explicitly model and account for both technical and subject-level variation simultaneously. For spatial data specifically, examine localized regions to check whether patterns respect anatomical boundaries (likely biological) or follow processing artifacts (technical).

The distinction becomes clearer with biological replicates across sites—consistent patterns despite technical variation strengthen confidence in biological findings.

Talk 6 – Ready, Set, Go: Considerations for setting up your experiments - Antje Becker

Q: During the cell preparation via MACS (2 times), the cell population I am looking for is not so clean, it has debris and other cell populations also. Would it be still fine to do the 10x scRNA-seq? Do I need to go for FACS in this case?

A: Depends on the amount, maybe FACS would make sense

Q: What is sensitivity and specificity in the context of spatial resolution or visum platform?

A: The graph this is referring to (I believe), shows increased UMI counts per area in FF tissue versus FxF or FFPE samples.