Introduction to single cell RNA-seq

Welcome to the first lesson of the course **Mastering Single Cell RNA-seq Data Analysis with Trailmaker™.** This comprehensive course, provided to you by Parse Biosciences, is designed to equip you with the knowledge and skills needed to analyze single cell RNA-seq (scRNA-seq) data using Trailmaker™. Trailmaker™ is a software tool developed by Parse Biosciences. From data uploading and processing to integration, exploration, and visualization, we will cover everything you need to know to analyze single cell RNA-seq data.

In this first lesson of the course, we're going to briefly introduce single cell RNA-seq. First, we will see what it is used for and what are the main steps of a typical scRNA-seq workflow. We will then focus on the distinctions between traditional bulk RNA sequencing methods and single cell approaches. Lastly, we will talk about how split-pool combinatorial barcoding and droplet-based methods work, and how they compare.

What is single cell RNA-seq?

Single-cell RNA sequencing is a cutting-edge technology that allows researchers to investigate gene expression at the level of individual cells.

A typical scRNA-seq workflow involves several steps, including cell isolation, RNA extraction, library preparation, sequencing, and data analysis. Each of these steps

has its own set of challenges and requires careful optimization to generate high-quality data. Let's see the steps in more detail.

- **Cell isolation** is the process of capturing high-quality individual cells from tissue, a crucial step in scRNA-seq as it determines the quality and quantity of RNA that can be extracted from each cell. This typically involves processing a tissue through dissociation into a single-cell suspension. Isolation can be achieved by isolating whole cells, cell-specific nuclei, or cell-specific organelles. Methods for isolating single cells include fluorescence-activated cell sorting (FACS), magnetic-activated cell sorting, laser microdissection, and microfluidic systems.

- Once cells are in a suspension, they are **pre-processed for single-cell sequencing**. For droplet-based technologies, this involves encapsulating individual cells into droplets along with barcoded beads, allowing for the unique labeling of each cell's RNA. In contrast, split-pool combinatorial barcoding follows a different workflow where cells are not physically isolated. Instead, cells undergo multiple rounds of pooling and splitting, and each cell receives a unique barcode through this process without ever being physically separated. This innovative method enables the unique identification of RNA from each cell.
  The primary goal in both cases is to ensure that transcripts from each cell are uniquely identified, enabling accurate downstream analysis, we will explore both methods in detail in the last part of this lesson.

- The **RNA is extracted** from each cell and then reverse transcribed into complementary DNA (cDNA) using reverse transcription. The cDNA is then

amplified and fragmented, and adapters are added to the ends of the fragments to generate a sequencing library. This step is called **library preparation**.

- **The library is sequenced** using high-throughput sequencing technologies to generate millions of short reads.

- Last step is **data analysis**, an important step in scRNA-seq, as it requires the use of various bioinformatics tools to process and analyze the millions of reads generated by sequencing. These tools enable the identification of differentially expressed genes, cell types, and cellular states. And here, the Trailmaker™ tool comes to the rescue! In this course, you will see how easy it is to analyze your data using this very user-friendly graphical interface.

The advantages of scRNA-seq are numerous. It allows researchers to identify rare or novel cell types, study the dynamics of cellular processes such as differentiation or development, and investigate the molecular mechanisms underlying disease states. In addition, scRNA-seq can help to identify biomarkers for diseases and develop new therapeutic strategies.

Although scRNA-seq is a relatively new and still evolving technology, when paired with the proper knowledge and tools, it is an incredible discovery tool for researchers.

How does single cell compare to bulk RNA-seq?

Unlike traditional bulk RNA sequencing methods, scRNA-seq enables researchers to explore the heterogeneity of gene expression across different cell types, states, and developmental stages. Let's compare the two methods in more detail.

Starting from a tissue, in a typical bulk RNA-seq experiment, all the cells that are inside a tube are lysed, RNA molecules are captured, and a cDNA library is created and sent for sequencing. In the end, the output will be a measure of the average gene expression of every gene in that population of cells. However, the limitation of this method is that it cannot differentiate between gene expression from individual cells or subpopulations from a single sample. For example, all the cells in a population could be transcribing a particular gene at low levels, or there could be a small sub-population of cells that uniquely expresses the same gene. Although the average gene expression level across the entire population would be the same in both scenarios, it would be impossible to determine from the average alone which situation is actually occurring.

Therefore, the idea with single cell RNA-seq is to be able to differentiate between these situations. The way this can be achieved is by isolating the cells and sampling the RNA molecules in each single cell individually. In this way we can look at the gene expression in every cell. In the case reported in figure 1, we can see clearly that we have only one sub-population of cells, the yellow one, that are transcribing the red gene. This allows us to study the variability inside of a cell population.

Overall, we can say that bulk RNA-seq measures the average gene expression of a population of cells, which can mask cellular heterogeneity. Conversely, scRNA-seq enables the investigation of gene expression at the level of individual cells. This can reveal cellular heterogeneity and dynamics that would be difficult or impossible to obtain using bulk RNA-seq.

scRNA-seq can also reveal rare or novel cell types that may be missed by bulk RNA-seq. In addition, scRNA-seq can be used to study the dynamics of cellular processes such as differentiation or development, which can be difficult to capture using bulk RNA-seq.

However, we must keep in mind that scRNA-seq also has some limitations compared to bulk RNA-seq. The data generated by scRNA-seq is noisier than that generated by bulk RNA-seq, and technical artifacts can introduce biases into the analysis. In addition, scRNA-seq requires more cells and is more expensive than bulk RNA-seq.

Now that we know how single cell RNA-seq differs from bulk RNA-seq, we understand that the main point is that individual cells must be studied. So, let's take a step back. The question is: how do we capture RNA molecules from individual cells? There are different ways to capture and label cells, as we saw in the previous section. In the next section, we're going to briefly review two commonly used methods: droplet-based scRNA-seq and the split-pool combinatorial barcoding method for scRNA-seq.

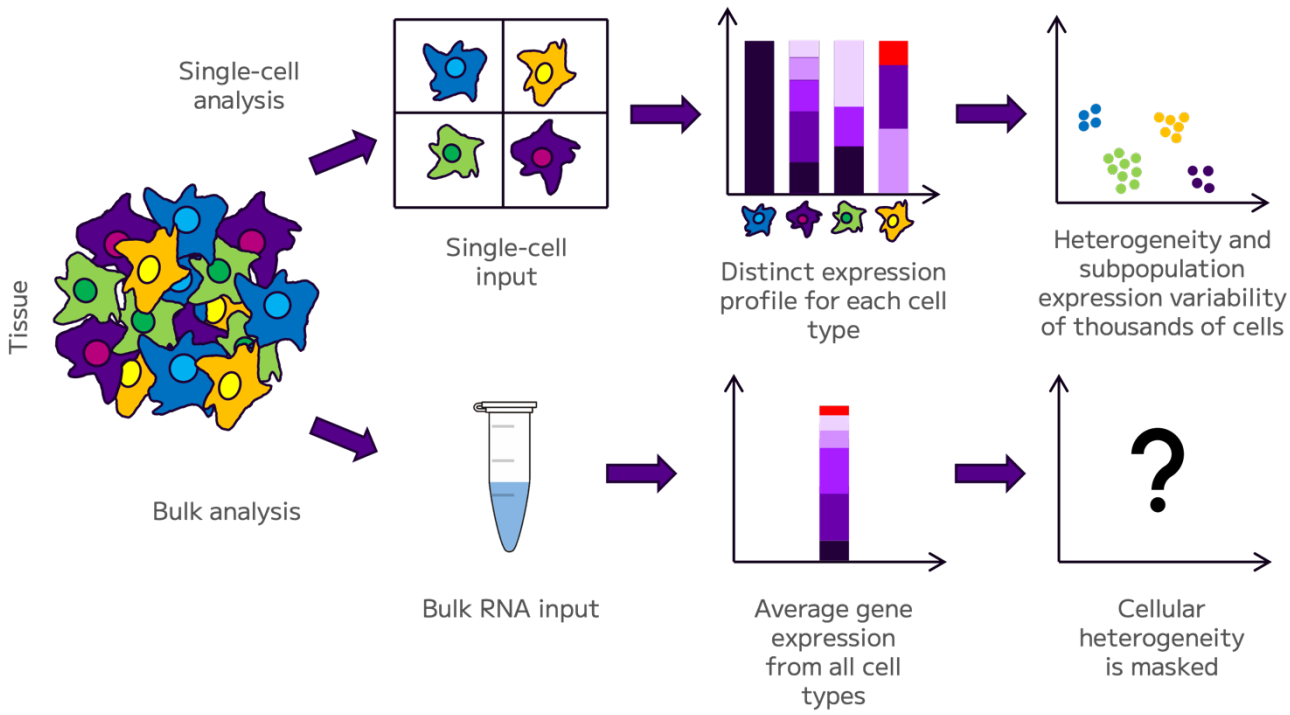


Figure 1. Bulk RNA-seq analysis compared to single-cell RNA-seq analysis.

How does droplet-based single cell RNA-seq work?

Droplet-based scRNA-seq is a popular method for generating single cell data. It uses microfluidics to encapsulate individual cells into droplets containing lysis buffer, reverse transcription reagents, and barcoded beads. The cells are lysed, and the RNA is reverse transcribed into cDNA, which is then amplified and sequenced using high-throughput sequencing technologies.

Specifically, a droplet, which is represented in figure 2A, is simply a small sphere of a solution immersed in an oil emulsion, in which there should be only one cell and something that is called a barcoded bead. If we take a closer look at the barcoded

bead, represented in figure 2B, we can see that they are coated with oligonucleotides. These oligos contain different sequences inside. At one of the ends there is the PCR handle, a sequence common to all the oligos that allows PCR amplification. On the other end there is the Poly-dT sequence, present in every oligo, it is used to capture the polyadenylated mRNA molecules. In the middle there are two very important sequences: one is the cell barcode, which is a unique sequence for every bead, you can think of it as the name of the bead. These barcodes allow each captured mRNA molecule to be mapped back to the individual cell from which it was transcribed, enabling the generation of a gene expression profile for each cell. For example, the first bead contains only blue barcodes, while the second contains only purple barcodes. This is something unique for every bead. Lastly, there is a unique transcript, sometimes referred to as UMI (Unique Molecular Identifier), which is a sequence that is used to uniquely identify the RNA molecules. They are represented in figure 2B with different colors in every bead, meaning that this sequence is unique for every oligo in each of the beads.

Taking a step back, we saw that these beads are inside an isolated volume together with the cell, and all the processes previously described (cell lysis, reverse transcription, PCR) happen in this closed space (figure 2C). In other words, every cell is lysed inside here, all the RNA molecules that belong to a cell are captured using one of the beads, and for each cell, all the RNA molecules are tagged with these oligonucleotides. In this way we can reconstruct the gene expression of every cell in a population.

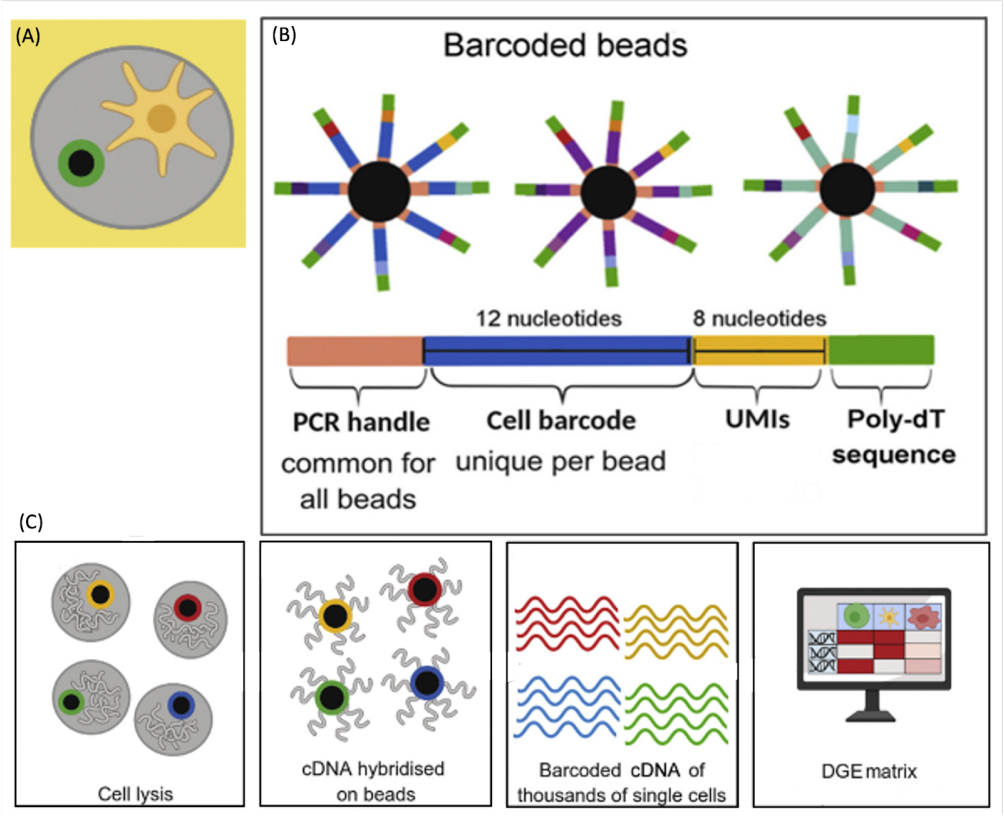


Figure 2. (A) Droplet encapsulating a cell, a barcoded bead, and lysis buffer. (B) Barcoded beads. (C) Process of cell lysis, RNA reverse transcription to cDNA, sequencing, and quantification as count matrix.

Disadvantages of droplet-based scRNA-seq

Despite its wide usage, droplet-based scRNA-seq also has notable limitations. The encapsulation process isn't perfect; some droplets may contain multiple cells or no cells at all, leading to noise and dropout in the data. Additionally, the PCR amplification steps involved can introduce biases and distortions, which may affect the accuracy of the gene expression profiles. We will see how to discard this low-quality data when we talk about the Data Processing module of Trailmaker™.

Moreover, the scalability of this method is limited. Processing large numbers of cells can be time-consuming and expensive. Additionally, droplet-based approaches

require specialized fluidics instruments to generate the droplets, adding to the complexity and cost of the experiment.

Split-pool combinatorial barcoding is a commonly used approach that circumvents many of the limitations of the droplet-based approach. This method, used in Parse Biosciences assays, offers a more scalable instrument-free approach. We will talk about this method in more detail in the following section.

How does split-pool combinatorial barcoding work?

Split-pool combinatorial barcoding allows for scalable single-cell RNA-seq without the need for specialized instrumentation. It leverages the unique labeling of cells or nuclei through multiple rounds of barcoding, without the need for physical isolation of single cells in a reaction chamber, enabling the processing of millions of cells in parallel, while maintaining unmatched data quality by reducing technical noise and biases.

One of the key advantages of split-pool combinatorial barcoding is that it enables large-scale single-cell projects without requiring dedicated microfluidics equipment. Indeed, cells (or nuclei) are fixed and permeabilized, so they act as individual reaction compartments, and the need to capture individual cells in droplets or microwells is eliminated. Therefore, if your lab has a centrifuge, thermal cycler, and some pipettes, you're ready to go. This method dramatically scales up the number of cells and samples that can be processed per experiment, allowing up to 1 million cells to be processed together. Because of the fixation step, samples

can be collected on different days, stored, and processed together for single cell sequencing, thus reducing batch effects.

Split-pool combinatorial barcoding works by repeatedly splitting a pool of cells, barcoding them, and then pooling them together again. This process is done in several rounds, each time adding a unique barcode to the molecules from each cell (fig. 3).

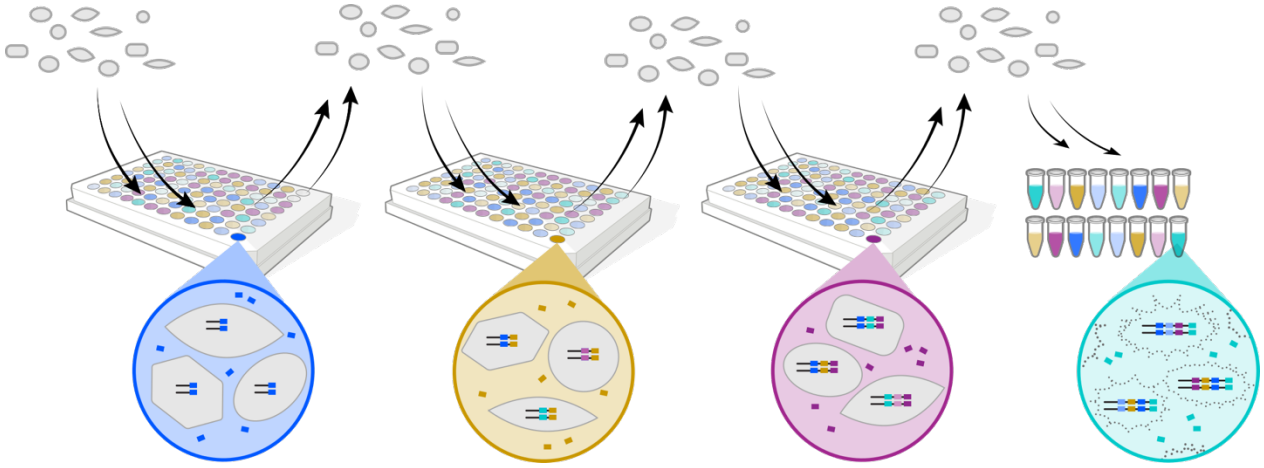
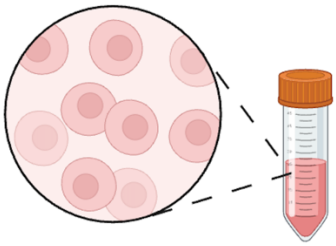


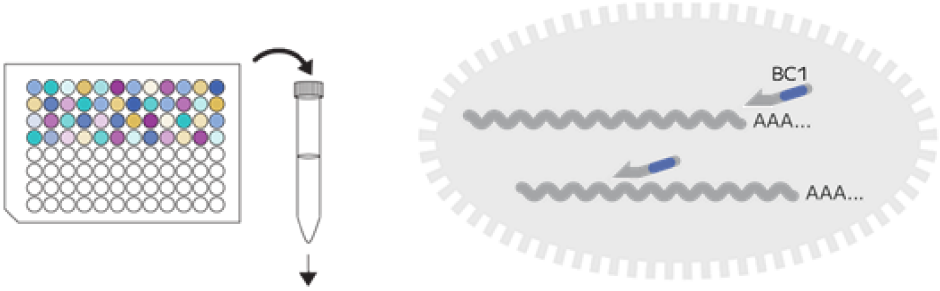
Figure 3. Diagram explaining the split-pool combinatorial barcoding process from Parse Biosciences experimental protocol.

Here's a step-by-step breakdown:

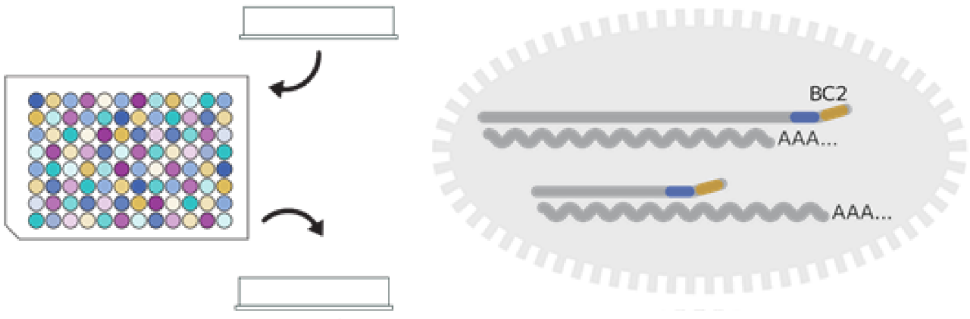
1. **Fixation and permeabilization**: Cells are first fixed and permeabilized, effectively turning each cell or nucleus into its own reaction compartment. This step is crucial as it eliminates the need for microfluidic isolation, allowing the process to scale up efficiently.



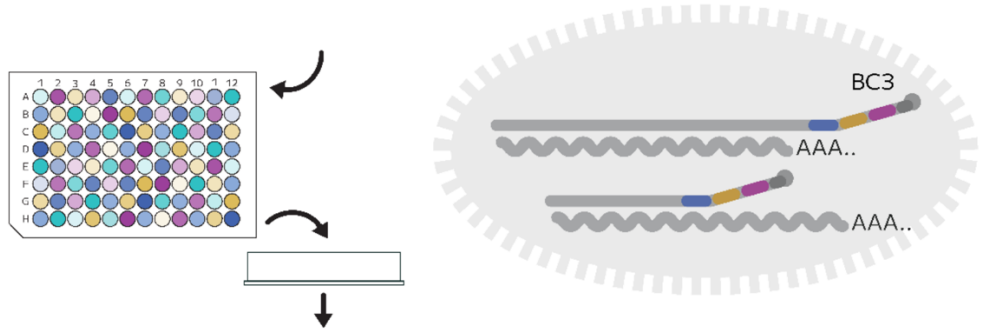
2. **First round of barcoding**: The fixed cells or nuclei are loaded into a plate with multiple wells, where the first set of well-specific barcodes, associated with specific samples, is applied. During this step, an in-cell reverse transcription reaction takes place, where RNA is transcribed into cDNA, incorporating the first well-specific barcode.



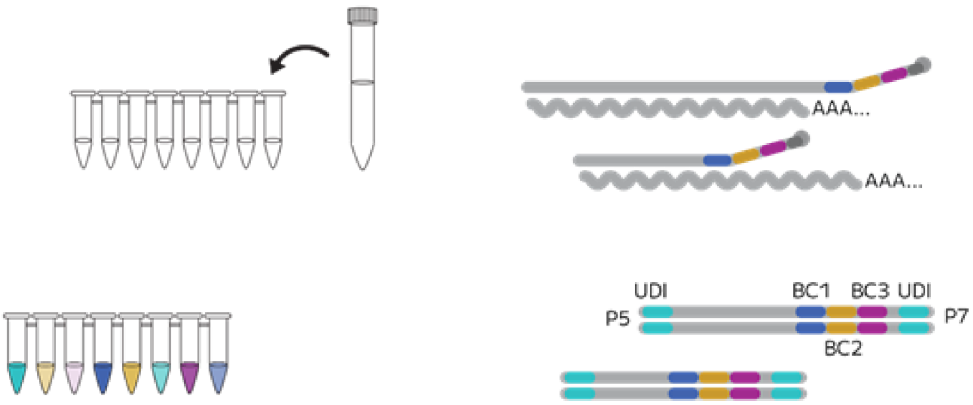
3. **Pooling and second round of barcoding**: After the initial barcoding, the cells or nuclei are pooled together and then split again into a new set of wells. Here, an adapter with a second well-specific barcode is ligated to the first barcode. This step further distinguishes the RNA from different cells.



4. **Pooling and third round of barcoding**: The cells or nuclei are pooled and split once more, and a third barcode is ligated to the cDNA.



5. **Fourth round of barcoding and sublibrary generation**: The cells or nuclei are then pooled and split across several sublibraries; they are lysed, a fourth barcode is added through PCR, and the barcoded cDNA is prepared for sequencing. This involves fragmentation and end repair of the cDNA, adapter ligation, and final amplification.



The four rounds of barcoding generate an exponentially large number of possible barcode combinations, more than sufficient to uniquely label up to 1 million cells or nuclei, while avoiding doublets. The combination of barcodes from each round allows to identify which RNA molecules came from the same cell, enabling the reconstruction of gene expression profiles at the single-cell level.

What are sublibraries?

It's important to clarify the distinction between a sample and a sublibrary in the context of in the context of split-pool combinatorial barcoding technologies. After the third round of barcoding and before cell lysis, the cells or nuclei are pooled together and then split into distinct populations known as sublibraries. Each sublibrary is assigned a unique fourth barcode (the Illumina index), allowing them to be sequenced separately.

At the end of sequencing, you will obtain a pair of FASTQ files for each sublibrary. These FASTQ files, which we will discuss in more detail in an upcoming lesson, do not correspond to individual samples. In other words, each pair of FASTQ files that share the same Illumina index from a single Parse Biosciences assay belongs to the same sublibrary, each of which contains cells from all samples.

So, why are the FASTQ pairs termed "sublibraries" and not "samples"? In a Parse Biosciences combinatorial barcoding assay, each FASTQ pair encompasses multiple samples due to an additional layer of multiplexing. Typically, in many next-generation sequencing (NGS) experiments, the Illumina index acts as the sample barcode. However, in Parse Biosciences data, each sublibrary (FASTQ pair) contains cells from all the samples that were processed together. This is because, during the

split-pool barcoding process, different cells from various samples are combined into the same sublibrary.

In an upcoming lesson, we will explore how the Trailmaker™ pipeline works, detailing the steps involved in going from sublibraries to individual samples, ensuring that the data from different samples can be accurately separated and analyzed.

Advantages of split-pool combinatorial barcoding

Split-pool combinatorial barcoding offers several compelling advantages over droplet-based single-cell RNA-seq methods, which not only simplify the workflow but also enhance the quality of the data by reducing technical noise and biases:

- **Avoidance of ambient RNA contamination**: One of the major challenges in droplet-based single-cell sequencing is the presence of ambient RNA. Ambient RNA refers to free-floating RNA molecules that escape from broken cells during sample preparation. These molecules can be incorrectly barcoded and assigned to the wrong cells, leading to inaccurate data. Ambient RNA can impact the ability for cells to be accurately clustered, e.g. into cell types, and may prevent you from answering your biological question.

  While in other scRNA-seq technologies the free-floating molecules in solution are encapsulated alongside intact cells in the microwells or droplets, split-pool combinatorial barcoding significantly reduces the risk of ambient RNA contamination. This is achieved with two mechanisms. First, by using a combination of barcodes to uniquely label each cell's transcriptome across multiple rounds of barcoding. For an ambient RNA molecule to be incorrectly assigned to a cell, it would need to follow the exact same barcoding path as the cell it is mistakenly associated with - an extremely unlikely event. Additionally,

the workflow includes wash steps that physically remove free-floating molecules from barcoded cells before they are lysed and processed for sequencing. This reduces the likelihood of ambient RNA being grouped with a cell in downstream analysis.

- **Lower doublet rate**: The doublet rate in split-pool combinatorial barcoding is lower compared to droplet-based methods. This is due to the sequential barcoding approach, which uniquely labels each cell across multiple rounds of barcoding. In droplet-based methods, doublets occur when two cells are encapsulated within the same droplet, leading to mixed signals. However, the split-pool approach's reliance on multiple distinct barcoding steps significantly reduces the chance of two cells being incorrectly combined, thereby lowering the doublet rate and improving data quality.
- **No physical isolation required – higher throughput**: Unlike droplet-based methods, the combinatorial barcoding approach does not require physical isolation of individual cells, making it easier to scale up to large numbers of cells. Therefore, this method can handle millions of cells in parallel, providing high-resolution data for large scale projects.
- **No need for specialized equipment**: Unlike droplet-based methods, split-pool combinatorial barcoding doesn't require expensive microfluidic instruments. Instead, the experiment can be conducted using standard lab equipment like centrifuges, thermal cyclers, and pipettes, making it more accessible to a wider range of laboratories.
- **Low cost**: This method dramatically reduces per-cell costs thanks to its efficient use of reagents. This reduction in reagent waste makes the technology not only cost-effective but also sustainable for large-scale studies.

- **Unmatched data quality**: The split-pool method enhances the detection of lowly expressed genes while avoiding issues such as ambient RNA contamination and doublets, that are common in droplet-based approaches. By fixing and permeabilizing cells, which then serve as individual reaction vessels, the method minimizes technical noise and improves the accuracy of gene expression profiles.
- **Sample multiplexing and reduced batch effects**: Split-pool barcoding also allows for efficient sample multiplexing, where distinct samples can be processed within the same experiment. During the first round of barcoding, each well is assigned a unique barcode, which identifies the sample type of origin. This feature enables demultiplexing of samples during data processing, effectively reducing batch effects - a common challenge in integrating single-cell data from different experiments.

In this lesson, we learned about single cell RNA sequencing (scRNA-seq), a cutting-edge technology that allows researchers to investigate gene expression at the level of individual cells. We also discussed the main steps of a typical scRNA-seq workflow. And we compared scRNA-seq to traditional bulk RNA sequencing methods, highlighting the advantages and limitations of each approach. In addition, we saw in detail how the droplet-based and the split-pool combinatorial barcoding methods work.

If you have any questions, don't forget to ask them in the dedicated section of the community forum at this link: https://community.trailmaker.parsebiosciences.com/