A Conceptual Introduction to Single Cell RNA Sequencing

By Eduardo da Veiga Beltrame
Assistant Professor of Computational Biology at MBZUAI
https://mbzuai.ac.ae/study/faculty/eduardo-beltrame/
Adapted from Caltech thesis: *Stories in Single Cell RNA Sequencing*
https://doi.org/10.7907/4kgh-8420
Draft date: 2024-08-29

Aside from viruses, all living things are made of cells, self replicating bags of molecules. There are many types of molecules in the cell, but here we will focus almost exclusively on only three important kinds: DNA, RNA, and protein. They are polymers, chains of a few kinds of molecules that serve as building blocks, called monomers. Different monomers, being different molecules, have different properties (size, charge, how flexible they are), and their sequence determines the properties of the polymer. This is why sequencing is such an important tool in molecular biology: it allows us to identify what molecules are present in a sample, and what their properties are.

In proteins the monomers are amino acids. There are 20 of them. All you need to remember is that amino acids can have very different properties and are really versatile, enabling proteins to do all kinds of things in the cell, such as chemical reactions. Proteins that perform chemical reactions are called enzymes.

In DNA (deoxyribonucleic acid) the monomers are nucleotides: adenine (A), thymine (T), guanosine (G) and cytosine (C). The backbone of DNA contains a sugar molecule called deoxyribose that has an oxygen atom making it very stable and rigid. This stability makes DNA an excellent medium to store the genetic information of the cell. When a DNA molecule is paired with another containing a complementary sequence, it forms the famous DNA double helix structure.

In a cell the DNA molecules with the instructions for everything the cell does are called the genome. In all multicellular organisms the cell genome is tightly tucked away inside the nucleus, a compart-

ment from which molecules cannot easily get in or out of. Having a nucleus is the defining feature of eukaryotes. Many single celled organisms do not have a nucleus, they are prokaryotes and archaea, and their genome is floating all around the cell in as one or more more big pieces of DNA.

In RNA (ribonucleic acid) the monomers are adenine (A), uracil (U), cytosine (C), and guanine (G). The information encoded in the sequence of bases in a piece of DNA can be copied into an equivalent sequence in an RNA molecule. This process is called transcription, it is done by an enzyme called RNA polymerase. RNA molecules produced from a DNA template are called transcripts. A major function of RNA in the cell is to serve as a template for copying some information from a stretch of DNA and taking it to other places in the cell to make proteins. The kinds of RNA that are specifically being used to make protein are called messenger RNA (mRNA). Stretches of the genome that contain information that encodes transcripts are called genes.

The backbone of RNA is a ribose sugar. It similar to deoxyribose, which forms the backbone of DNA, but without an oxygen atom. That makes RNA floppier so it can fold in many kinds of different structures and perform other useful things in the cell. The best example of this is the ribosome, the molecular machinery that makes protein by reading the sequence of amino acids to add from a molecule of mRNA.

To summarize: The process of going from DNA to RNA is called transcription, done by RNA polymerase enzymes. These RNA molecules are transcripts. The control of this process by the cell is transcriptional regulation. The segments of DNA containing the information to make transcripts are called genes. Going from RNA to protein is called translation. Translation is done by ribosomes, large molecular machines made of RNA and protein. The process of producing the molecules that are used by the cell (which sometimes are the RNA molecules themselves, sometimes proteins) is called gene expression.

When RNA is first transcribed it is all located in the nucleus, and to perform most of its functions it must be exported out of the nucleus into the cytoplasm (the rest of the cell). However, RNA that will become proteins, called messenger RNA (mRNA) must first be processed before being exported. That is because not all of the content of a gene may be an exon, a region which encodes a part of a protein. So the pre-mRNA undergoes splicing, a process where introns, the parts that don't encode protein, are removed and the ends are joined together, leaving a mature mRNA made entirely of exons. Splicing is done by the spliceosome, a molecular machine located in the nucleus

that, like the ribosome, is made of both protein and of RNA. This is summarized in Figure 1.

This flow of information also reflects how hard it is to study RNA, DNA and protein in cells. Between 2007 and 2010 with the introduction of next-generation sequencing technologies (NGS) the costs for DNA sequencing plummeted by about ten thousand times to about ten cents per million base-pairs sequenced[1]. Now that cost is at about one cent per million base-pairs. Decreasing costs made our ability to read DNA a commodity tool, and makes it very convenient to have sequences be the information output of many biological experiments. If the experiment to answer a biological question can be turned into a sequencing problem, it can be done cheaply, and it can be done at scale.

[1] The NIH tracks the cost of DNA sequencing and briefly discusses it's evolution here: `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`

Because we can convert RNA to DNA using reverse transcriptase enzymes, it is possible to use DNA sequencing to study RNA molecules. We can't do the same thing with proteins, and so it is harder to study protein molecules than it is to study DNA or RNA.

The DNA of an individual is essentially the same across all cells. DNA is a very stable molecule, and the genome has to be very stable so that it can be reliably copied and passed on to the next generation of cells. With RNA it is a different story. Every cell at each moment will have a different composition of RNA molecules, which reflects what the cell is doing at that point in time. Transcriptional regulation is how the cell controls how much and what kinds of RNA molecules it makes.

Understanding transcriptional regulation is key for understanding, controlling, modifying and engineering biological systems. To study transcription it is necessary to measure the RNA in the cell. Experiments that use sequencing to look at the composition of RNA molecules in a sample are commonly referred to as RNA-seq experiments[2].

[2] If interested, the best resource I know of to learn all the important aspects of RNAseq experiments (most of which also apply to scRNA-seq) is the RNA-seqlopedia, written by the Cresko Lab of the University of Oregon: `https://rnaseq.uoregon.edu/`

Because the RNA in a cell changes all the time, there is an endless amount of RNA-seq experiments that could be performed in a single species - even in a single individual. One of the most informative things to look at are the differences between individual cells and populations of cells, or tissues. A tissue sample will contain multiple cell types, and each cell might be in a different stage doing different things. The average of the RNA contents of all the cells in a tissue is usually going to be very different from the contents of each cell, because usually cells in a tissue will have very different content. This heterogeneity of a population of molecules in cells is what makes it important to measure their RNA content many times and under multiple conditions.

Not all types of RNA are heterogeneous. For many types of RNA

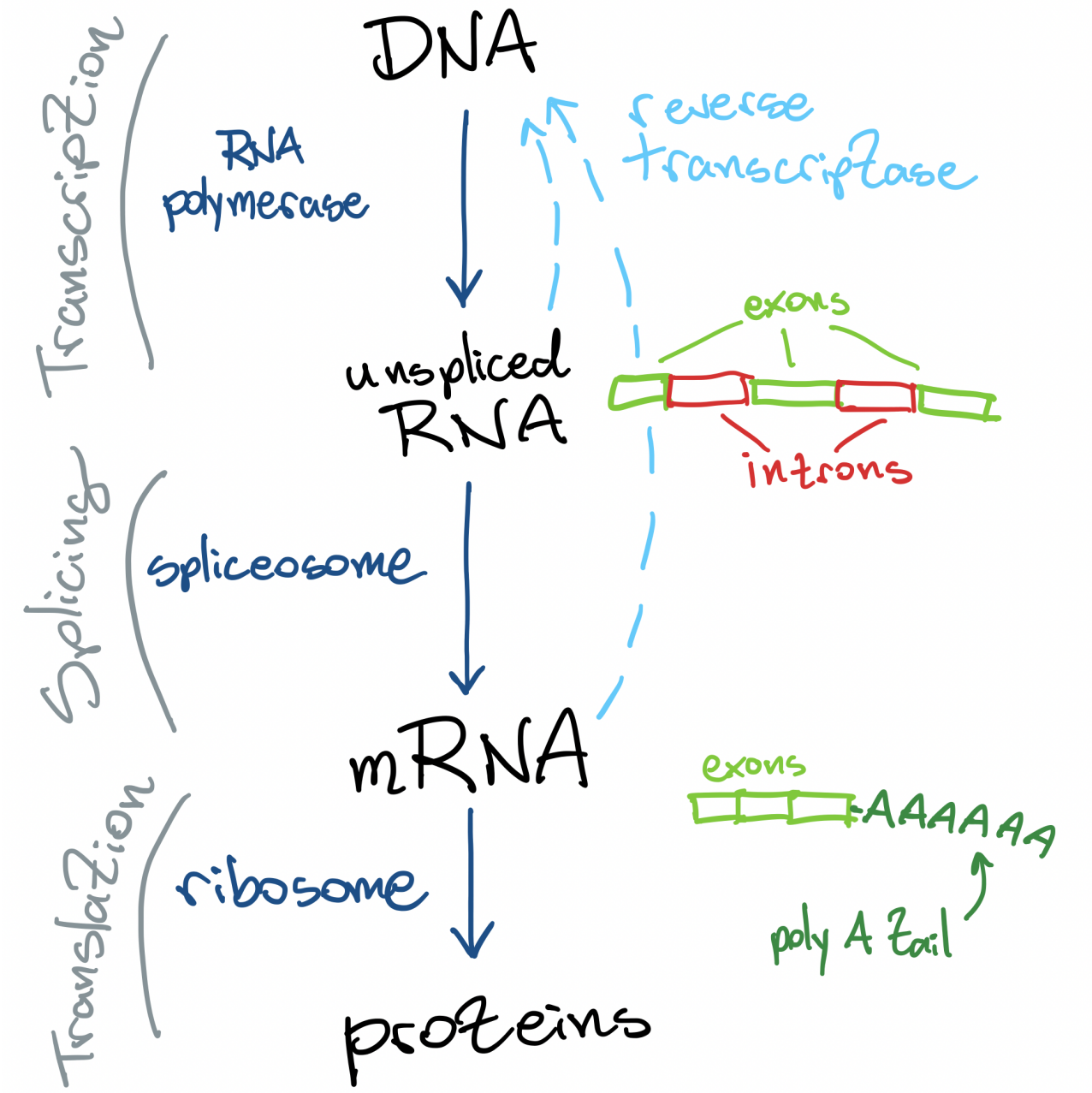

Figure 1: An overview of the processes that happen inside the cell when going from DNA to protein.

the composition across cells is largely the same, and it doesn't change much over time - these kinds of RNA are very homogeneous across cells. Ribosomes, which are made of a few several dozen molecules of RNA and protein are the prime example of RNA homogeneity in cells. Cells need to make a lot of protein very quickly, so they have a lot of ribosomes, and sometimes up to 80% of the RNA in a cell is ribosomal RNA, which will always have the same composition. Thus it is not interesting to measure the ribosomal RNA of cells multiple times under multiple conditions, because it doesn't change much

Fortunately eukaryote organisms evolved a mechanism to add a special tag called polyadenylation tail to mRNA that allows for capturing and sequencing only mRNA molecules, which change all the time and are very interesting to measure. As part of pre-mRNA processing, dozens to hundreds of A letters (adenine molecules) are added to the end of every mRNA molecule. This long AAAAAAA... sequence makes it straightforward to capture, amplify and sequence the mRNA of a sample by using a single probe that is complementary to the poly A tail - a poly T probe: TTTTTT...

Because of this quirk of biology, it is possible to look at all the mRNA in cells without the need to capture everything else, which would include all of the homogeneous ribosomal RNA, and which would significantly increase sequencing costs and make many experiments impractical. There are other heterogeneous RNA molecules, such as long non coding RNAs (lncRNA) that do not have poly A tails. As a casualty of convenience, these other kinds of RNA that do not have a poly A tail do not get studied nearly as much as mRNA. It is possible that we are missing fundamental parts of the puzzle by focusing too much on only mRNA, but only time will tell.

An alternative strategy to capturing everything is to use specific probes to capture only things you already were interested in ahead of time. This latter approach is what was used in microarrays[3], which were invented in the 1980s and widely used into the 2010s, when RNA-seq became a very popular technique due to lower requencing costs and the ability to capture all mRNA without deciding what to look for ahead of time.

There are multiple variations of experimental procedures for performing RNA-seq, but it typically involves the following main steps.

1. A sample containing cells is homogenized, meaning the cells are broken using reagents such as detergents, and the RNA is released in solution. Typically the sample will be enriched for mRNA and then purified.

2. Using the reverse transcriptase enzyme, a strand of DNA that is complementary to each RNA piece is created. This DNA is

[3] For the interested reader, here is a nice historical review of the invention of microarrays:

Michael C. Pirrung and Edwin M. Southern. The genesis of microarrays. *Biochemistry and Molecular Biology Education*, 42(2):106–113, 2014. doi:10.1002/bmb.20756

referred to as cDNA (complementary DNA).

3. The long cDNA pieces, that can be several thousand of base pairs long (many transcripts are very long!) are broken into smaller pieces of no more than a few hundred base pairs each, as the sequencing platform usually requires short pieces.

4. Extra sequences are added to the ends of each cDNA molecule so that they can be sequenced and the cDNA is amplified (meaning many copies of each cDNA molecule are produced). The exact steps vary depending on protocol. The process of preparing a collection of cDNA molecules for sequencing is typically referred to as library preparation.

5. After library preparation the sample is sequenced, and a list of the sequences in each cDNA fragment is created. Processing and analyzing this list of strings containing the four letters ATCG comprises most of the discipline of bioinformatics.

Single cell RNA sequencing (scRNA-seq)

As sequencing costs decreased it became possible to process more RNA-seq samples in a single experiment. But the samples for "bulk" RNA-seq samples are essentially a smoothie of a piece of tissue, and this makes it hard to look at the cellular heterogeneity. For example, if a tissue consists of many different cell types, like the brain, then it is hard to obtain a pure sample that has only one cell type in it, such as neurons or glial cells, because different cell types are physically intermingled and it is hard to separate them.

Additionally, because cells of the same type can be in different states doing different things, looking at heterogeneity at the individual cell level requires a way of separating and sequencing the RNA from individual cells. For example, a cell undergoing division (in the mitotic phase, when cell growth stops) will be expressing different genes than a cell that is quiescent or growing (in the interphase).

This is the fundamental motivation behind single cell RNA sequencing (scRNA-seq) and all other single cell techniques: to be able to look at heterogeneity and understand how the cells in a sample are different from each other it is necessary to measure each cell individually. If a tissue does not have a lot of cellular heterogeneity, then there is not much more to be learned from looking at individual cells than from an average aggregate. The difference between single cell RNA sequencing and bulk RNA sequencing is that between drinking a smoothie and tasting individual berries. While on average the smoothie and berries will look the same, the smoothie taste masks

the heterogeneity of individual members of the berry population.

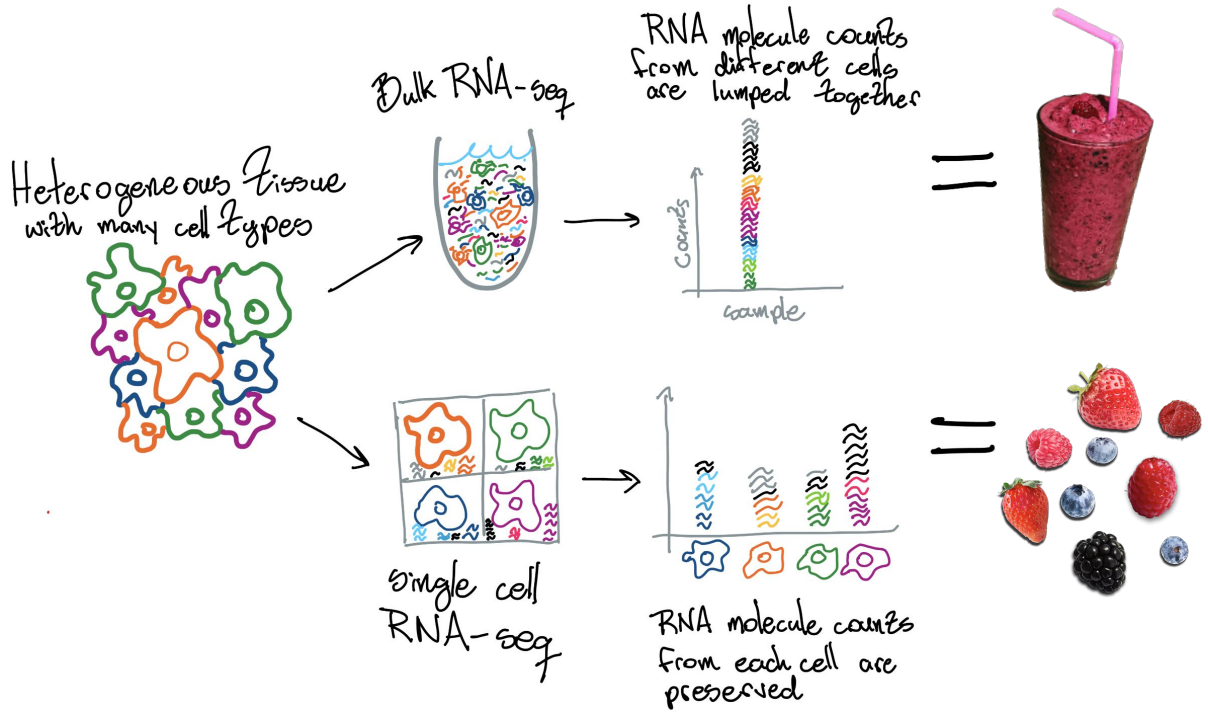

Figure 2: The difference between single cell RNA sequencing and bulk RNA sequencing is that between drinking a smoothie and tasting individual berries. While on average the smoothie and berries will taste the same, the smoothie taste masks the heterogeneity of individual members of the berry population.

The output data of a scRNA-seq experiment is a gene count matrix, containing one row for each cell, and one column for each gene from which an mRNA came from. Across all cells usually between 10.000 to 20.000 genes are seen, and a typical experiment will survey a few thousand cells, although it is currently possible to survey tens or even hundreds of thousands of cells in larger experiments.

Steps in scRNA-seq experiments

A Typical scRNA-seq has three main steps: cell isolation, barcoding, and library preparation. Chemically speaking, scRNA-seq is very similar to (bulk) RNA-seq with very little starting material - a typical mammalian cell has on the order of a hundred thousand mRNA molecules[4]. Nevertheless, a now typical scRNA-seq experiment sequences between about 500 to 10.000 original mRNA molecules per cell, which suggests we are surveying 1-10% of the mRNA molecules in each cell.

Cell Isolation: Cells are physically dissociated and usually isolated physically, being separated into different containers, for example, in the wells of a -well or 384-well plate. Three kinds of approaches are used for the physical separation of cells: plate based, droplet based, and split-pool based methods. Methods are frequently

[4] A mammalian cell typically has 10-20pg of RNA and the average mRNA molecule has 2200 bases, corresponding to about 500 daltons, so 10pg of RNA corresponds to about one million RNA molecules. It is estimated that 2-5% of the RNA in a human cell is mRNA, thus 20,000-50,000 mRNA molecules is a reasonable estimate for the number of mRNAs to expect in a typical cell. This estimate shouldn't be taken blindly, as different cell types can vary wildly on their sizes, and mammalian cells tend to be larger compared to other species. But the take home message is that scRNA-seq seems to be capturing a large fraction of the mRNA in cells.
See the three relevant BioNumbers entries for this estimate at: `https://bionumbers.hms.harvard.edu/bionumber.aspx?id=111204`, `https://bionumbers.hms.harvard.edu/bionumber.aspx?id=101469&ver=1`

grouped according to the cell isolation method, which we discuss more below.

Barcoding: Upon reverse transcription, two short sequences (10-20bp) are added in addition to the mRNA cDNA sequence: a unique sequence corresponding to the original cell, plus a random sequence corresponding to the original molecule, called a unique molecular identifier (UMI). Most scRNA-seq methods have UMIs, as their absence makes accurate quantification of original molecules significantly harder[5].

Library preparation: Cells then undergo the same procedure performed for bulk RNA-seq samples, but on a much smaller scale of individual reactions.

[5] That's because when individual molecules are copied via PCR (polymerase chain reaction) different molecules have different numbers of copies made. If there are no UMIs, it is impossible to tell how many mRNA molecules (transcripts) of each gene there were originally, because two transcripts of the same gene often have the exact same sequence. In order to quantify original abundances without UMIs it is necessary to make estimates of how much each transcript gets amplified based on their sequence (which can cause amplification biases), and this is a hard challenge.

Main kinds of scRNA-seq cell isolation methods

There are now hundreds of studies describing different methods and protocols for performing scRNA-seq. Often these methods are tweaks and improvements on existing methods. Broadly speaking, there are three main ways in which cells may be isolated and barcoded: in physical containers (plate methods), in microfluidics emulsions (droplet methods) and via sequential split-pool barcoding.

Plate based methods: Cells are manually or robotically isolated in physical compartments such as 96 or 384 microwell plates, with a typical throughput of hundreds or thousands of cells. Barcoding happens by adding a distinct DNA barcode to each well.

Droplet based methods: Cells are encapsulated in a droplet using a microfluidic device. In addition to a cell, each droplet also encapsulates a DNA coated bead, and this DNA has a unique sequence for each bead. Upon lysis, the cell releases it's mRNA which is captured by the bead DNA and reverse transcribed so that the barcode is added to the cDNA pieces.

Split-pool methods: Cells are not all physically isolated at once. Instead, the sequence and the transcript sequence are now on the same piece of DNA. Cells are manually or robotically split into a few dozen or few hundred separate compartments. Within each compartment a partial barcoding happens, adding a common barcode to all cells in it. Cells are then mixed back together, and then split again, repeating this procedure. The number of potential barcodes that can be created is given by the number of compartments to the power of the number of rounds. By making the number of potential barcodes much greater than the number of cells (e.g. a million barcodes with ten thousand cells) the number of cells with the same barcode can be made very small, and thus each cell can be considered to receive a unique barcode.

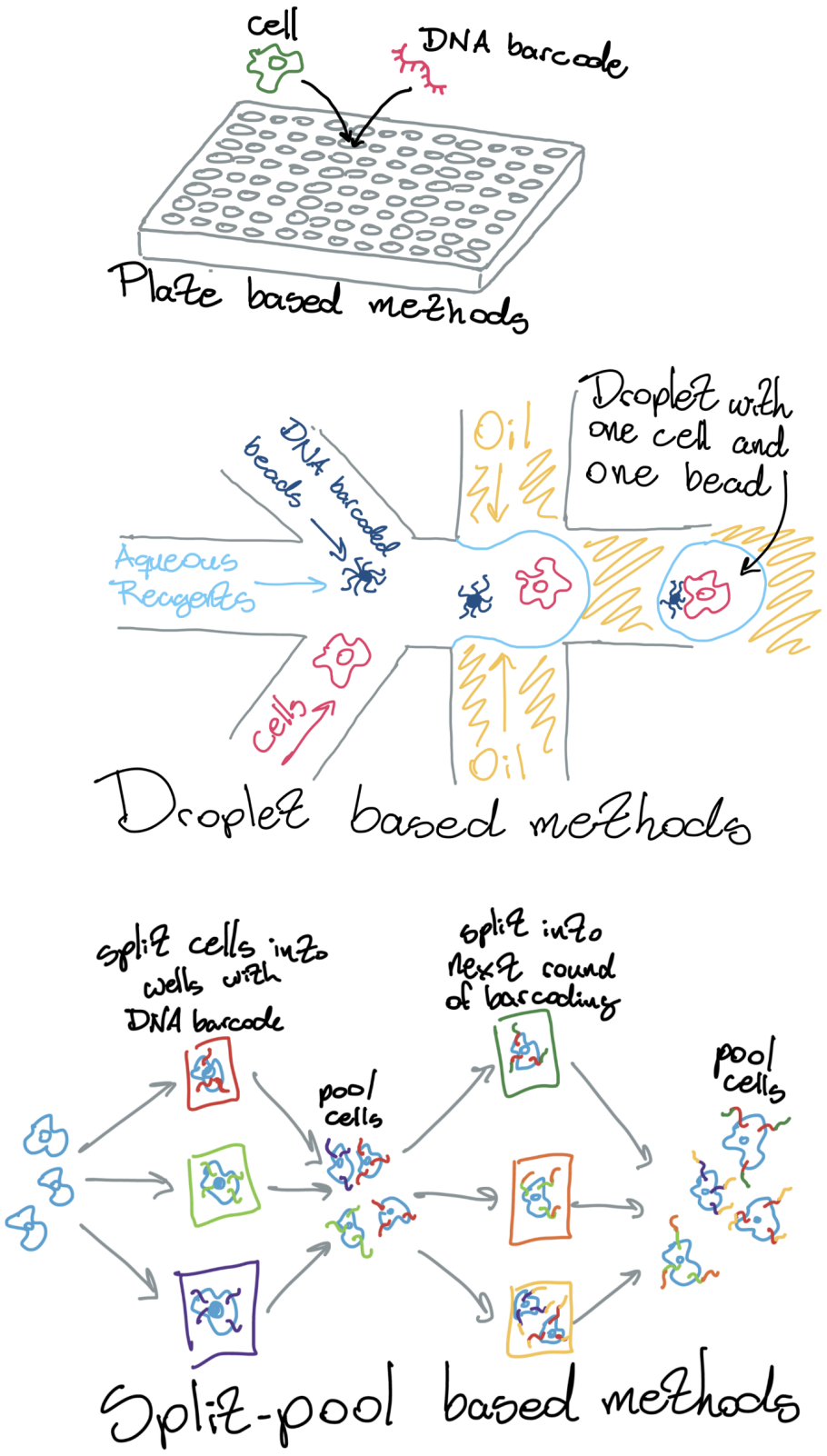

Figure 3: Conceptual illustration of the three barcoding strategies: plate based, droplet based, and split-pool based methods.

Annotating scRNA-seq data

The gene count matrix produced by a scRNA-seq experiment contains the information of how many mRNA molecules from each gene were seen in a cell. Identifying what types of cells exist in a sample based on the transcripts seen is a task commonly referred to as annotation.

Annotation is fundamental, because stratification of data in a sample is the whole point of scRNA-seq: if you don't annotate you can't stratify groups to compare, for example neurons vs glial cells. Without identifying the subpopulations present in a sample, scRNAseq yields information equivalent to bulk RNA-seq, where the transcripts of all cells are averages.

Annotation is perhaps the biggest challenge when datasets are the first of their kind: when doing an experiment on a new species, a new tissue, or a new technique for the first time. In such cases direct comparison with existing datasets may not be possible. These novel datasets typically need to undergo exhaustive manual annotation, and the annotated dataset forms a stepping stone for subsequent studies in related systems.

When a biological system has been extensively studied using other methods, it is often possible to annotate cell types by carefully reviewing the literature for marker genes, which are genes that are only expressed in one or a few cell types. With a list of marker genes for each cell type of interest in hand it is possible to annotate them in a sample.

Some systems such as blood have only a few cell different cell types, all already well studied and with distinct marker genes. This enables annotation of cell identity with confidence. However, when there are dozens or hundreds of potential cell types, and when not all of them have well established markers (or when the known markers cannot delineate different states), assigning cell identity is a more subjective task.

In these cases the common approach is to try to cluster (group) the cells based on some measure of similarity, compare each cluster with the remaining cells, and then based on the genes that are different between the two groups (differentially expressed genes), attempt to identify a subset of them matching the profile of a known tissue.

This process has many caveats. There are many different workflows that could be used to define clusters. The degree to which clusters should be broken down or combined is also subjective. Ideally each cluster should correspond to a single cell type, but how to know when that is the case? Even when cell types are well defined, if there are many of them, multiple rounds of clustering, inspection,

and sub-clustering may be needed.

One strategy to address some of this is to use a hierarchical taxonomy instead of a flat one. In addition to reflecting the fact that certain cell types can have subtypes and sub-subtypes, a taxonomy also naturally reflects the uncertainty in our classification. For example there might be clear neuronal markers that distinguish them from glia and other brain cells, but some neuron types might be better characterized than others, meaning that as we go down in the taxonomy the uncertainty increases. It is fair to say that annotation of cell types in a new biological sample is the most challenging and time consuming step of scRNA-seq analysis.

scRNA-seq and machine learning

Over the past few years, the amount of data being generated with scRNA-seq techniques has scaled exponentially, with the number of cells surveyed in a single study going from dozens to millions[6]. Each year hundreds of studies are published, often containing dozens of experiments and tens of thousands or hundreds of thousands of cells each. This is because for a few thousand dollars, it is now possible to profile tens of thousands of single cells in a standard experiment[7].

An interesting feature of scRNA-seq data is that it is very standardized: a big sparse matrix of cells by gene counts, and at the moment most of it is produced with one technology commercialized by one company, 10x Genomics (the same way that almost all sequencing is done with technology commercialized by Illumina).

Because the data generated by a scRNA-seq experiment captures data from all mRNA in the cell, data generated to answer one particular biological question may lend itself to answering other questions that look at different aspects of biological variation.

As we develop better techniques to integrate, annotate and compare cells, old data gains new value. It may be reanalyzed together with new data, and the number of questions and experiments one could conceive from the large datasets being created is much greater than any individual lab could pursue, and the great boon of sharing data is enabling reanalysis and asking new questions.

This stands in contrast to the way most experimental science is conducted: with experiments carefully designed to gather the right data to answer a specific question. As the amount of public data continues to grow exponentially, we will have to learn to come back to old data armed with new questions, principled methods, and a lot of scruples so as not to end up fixating on artifacts. Perhaps not fixating on artifacts will be easier to do with old data, since with new data there is typically a much greater motivation to find "something".

[6] Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, April 2018. doi:10.1038/nprot.2017.149

[7] A standard scRNA-seq experiment from 10x Genomics (`https://10xgenomics.com`) will cost $1000-2000 for 10-20k cells plus another $1000-2000 for sequencing.

It is interesting to note that this is not a new phenomena in the biological sciences - very much the same kind of thinking was spurred by microarrays, and long after writing the preceding paragraph, I ran into the following passage in this 2006 article on microarrays by Jörg D. Hoheisel [8]. Nothing new under the sun, as they say:

Microarray technology has initiated an experimental approach that is based on unbiased sample screening and accumulation of data, preceding the formulation of hypotheses. To an extent, it has placed data production before intellectual concepts, although of course further and more detailed studies are required to confirm and refine the hypotheses that result from such studies. In this respect, biology is becoming more similar to physics. Although the value of this approach in biology is still a subject of debate, physics has clearly demonstrated its power. However, even those who are used to microarray technologies sometimes still need to dissociate themselves more fully from a hypothesis-driven view, as it is not data production but data interpretation that is still often biased by pre-existing ideas.

[8] Jörg D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, 7(3):200–210, March 2006. doi:10.1038/nrg1809

At the same time that this tidal wave of biological data started, driven primarily by ever decreasing sequencing costs, another revolution started unfolding in statistics and computer science, driven primarily by ever decreasing computing costs. In the past decade the broad field of machine learning saw very rapid development as ever larger neural networks were successfully applied in all kinds of ways to all kinds of data, such as image classification, speech recognition, and control systems.

The current wave of excitement got started in 2012 with the success of AlexNet, a neural network for image classification that performed significantly better than everything else at the time[9]. This caused an ongoing flood of attention, investment, and research developing all kinds of extremely clever algorithms for dealing with data. Often these ideas are tested in toy and benchmarking scenarios (such as standard sets of images or texts used in benchmarking), because that's where a lot of curated data is readily available.

[9] The story of how AlexNet kicked off the current excitement wave is chronicled in this 2018 Quartz article: `https://web.archive.org/web/20210301025354/https://qz.com/1307091/the-inside-story-of-how-ai-got-good-enough-to-dominate-silicon-valley/`

But as time goes on and the dust settles, people start applying the most promising ideas and algorithms to new domains, and seeing what works well and what doesn't. This is where machine learning really impacts science, when expert domain knowledge is coupled with judicious application of suitable algorithms for the system at hand.

Single cell RNA sequencing in particular is at a really interesting point[10], because the data is all in a standard format (a matrix), there is a lot of it publicly available, and there is a lot of interest and potential for doing something useful with this data. At the same time, there aren't yet enough people thinking about it from the machine learning side. It has only been 9 years since AlexNet, and indus-

[10] The other really interesting high-throughput standardized kind of data being created in biology is imaging data, for which translation of machine learning methods should be even more straightforward, since many of them are already developed for images.

try has so far soaked up most of the machine learning trained researchers, and I think the landscape will change dramatically in the coming years as machine learning researchers turn their attention to other kinds of standardized data.

This bonanza of computational methods is a blessing (and a temporary headache) for biologists drowning in data. While we now have efficient methods for dealing with high throughput scRNA-seq data, there are *many* of methods[11]. These tools range from unpublished scripts, preprints, to entire frameworks maintained by several people, and it is not at all clear which one is the "best". Rapid development also means that careful benchmarking and comparison ends up on the back burner. It will likely take a few more years for the dust to settle.

For example, in the beginning of scRNA-seq, many methods for doing differential expression and visualization were directly taken from bulk RNA seq. Although the experimental methods are very similar, the data from single cell and bulk RNA seq are different, and linear algebra techniques should not be judiciously applied[12].

Even though they are commonly applied, techniques like normalization and log transformation are prone to introducing artifacts and false variability on scRNA-seq data. Given a matrix of gene counts X, normalization is the practice of taking the counts X_{ig} of cell i and gene g and dividing them by a cell scaling factor S_i to obtain a new normalized value X_{ig}/S_i . Log transformation usually means adding a *pseudocount* c, (where usually $c = 1$) to the original count value and taking the log of that, to obtain a new value $log(X_{ig} + c)$, but could also mean performing that operation with the normalized values; $log(X_{ig}/S_i + c)$. For a discussion on systematic errors caused by normalization and log transformation, see Lun 2018[13]. For another discussion on these issues and problems that might arise due to normalization and log transformation, as well as a proposed alternative to standard PCA that does not rely on the Gaussian assumption see Townes 2019[14].

On the next chapter we will talk about one way to deal with these challenges: using bayesian generative models, particularly in the context of the scvi-tools framework (scvi-tools.org), which offers an extensible collection of generative models tailored for scRNA-seq data.

[11] `https://scrna-tools.org` currently counts 1059 tools developed for dealing with scRNA-seq data.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology*, 14(6):e1006245, June 2018. doi:10.1371/journal.pcbi.1006245

[12] For example, the commonly used principal component analysis method models data as coming from a continuous multivariate distribution, and optimizes a gaussian likelihood. The fact that normal distributions are not appropriate for dealing with low count values (where discreteness becomes apparent) and that poisson or negative binomial distributions should be used was already discussed by Anders and Huber in 2010 in the paper where they introduced the popular DESeq package for performing differential expression on bulk RNAseq data.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, March 2010. doi:10.1038/npre.2010.4282.1

[13] Aaron Lun. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. page 404962, August 2018. doi:110.1101/404962

[14] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295, December 2019. doi:10.1186/s13059-019-1861-6

Challenges in academic bioinformatics software development

Many common development issues with bioinformatics tools are a direct consequence of the common incentives in academia, where the publication prestige incentivizes people to claim to have done something novel often leads people to try to reimplement the wheel instead of adding onto existing tools, so that they can be better framed as something new and made from scratch. In a "hot" and rapidly field like scRNA-seq this causes the following issues which are frequently witnessed in academic software development:

Software is often developed by a single person. This increases the odds of bugs and problems with the implementation, because no one else reviews the code. Peer reviewers will not typically review someone's code (it is a lot of work) and just because code is open source and available it does not mean other people will check it. In practice having multiple people working on a codebase is the surest way to decrease the occurrence of errors, while at the same time making the code more readable for others.

The codebase is frequently not developed further after publication. Often this is because the academic incentives diminish after having a publication accepted, or because the person who wrote the code graduates and leaves the lab.

Workflows and methods are developed as ad hoc tools. Frequently new workflows are not thought through and many steps incorporate arbitrary (unjustified) choices. Doing ad-hoc things is an integral part of science and experimentation, but once something works, or seems to work, people tend to forget it was done ad-hoc. So people will extrapolate the context in which it is supposed to work, while disregarding the need for benchmarking or validation because everyone else is doing it. This is akin to developing a "superstition" or a "myth", and happens in science occasionally.

Software is hard to discover, deploy, and compare. There are so many tools, often annoying to install and run, that nobody can feasibly do an exhaustive search. People just use what their lab friends are using.

Software is rarely benchmarked outside original study. Benchmarking is a lot of work, and there is no incentive for further validating the software after publication. Additionally, most benchmarks are often only done in humans and mice because those are the most popular organisms, but people tend to assume that a method will work on all other scRNA-seq data. For example, one aspect of mammals is that they have large cells, while many other organisms such as *C. elegans* have small cells.

Some suggestions

I highlight these problems here because they are really pervasive - anyone working with scRNA-seq bioinformatics will have encountered them - and they merit more attention and discussion. I don't have a simple solution to prescribe, but I do have a few things I think people should have in mind when developing bioinformatics software.

Have independent code reviews. If developing code that will be used by other people, or that will be part of a publication (and thus potentially used by others), always ask a friend to do a code review. The ideal scenario would be to have your friend developing the software with you, so that if something is broken or hard to understand, it will be brought to your attention.

Be scrupulous about your own work while it is still being developed. We tend to want to do everything as fast as possible, but good work takes time. In reality it is always a compromise, but I think that starting from a mindset that good work takes time is helpful.

Consider how your software is going to be maintained after publication. If the software being provided is just meant to reproduce the publication this is less of an issue, but if you claim that other people should be able to use your software, then it is imperative that at least one person be responsible for maintaining the codebase after publication, and it is really important to discuss this with other authors.

Consider extending existing software rather than developing from scratch. Carefully assess what the landscape is, and whether there is already a codebase or framework to which your software workflow could be added. It is not always possible to do this, since frequently there will not be a good match, but if it is possible it's a win-win.

- More people will be looking at your code.
- It will make your code better, since there will likely be certain guidelines on how the contributed code should be structured.
- It will make your software easier to find and use, since there are other people already installing and using the framework.
- It may reduce the burden of maintaining the code, and will help keeping it up to date with other software dependencies make it easy to install, since the other developers will likely already be focusing and longer term support of the codebase.