Data analysis, lesson 2 – Prof. Calogero

Bulk RNA sequencing

Bulk = a mix of cells that are sequenced all together. Piece of cDNA that is blunted on the 2 sides → addition of adapters. Finally, I obtain a library that I can sequence. When I sequence short reads, the longest sequence that I get from RNA ranges between 100 and 500 nucleotides. The methods based on 2^{nd} generation sequencing require RNA fragmentation immediately after a cDNA creation.

Library structure:

- Read primer 1.
- If the piece of DNA is reversed, I obtain the read primer 2. In fact, in sequencing, I can work forward only or forward and reverse; this second approach is mainly performed, in order to have the maximum coverage. However, as sequencers have a higher sensitivity, the sequencing of a read only (made of 75 nucleotides) is enough to identify the piece of DNA of interest.
- Index primer allows to sequence more than one sample: when I prepare my sample, I put an index with a different length for each sample. Then, sequencing is done all together, but samples will be divided according to the length of the index. Multiple samples can be sequenced all together and divided, sequently, by the sequencer according to the length of the index.

These sequences are stored into fastq files, which are similar to fasta files (which have the header and the sequence); fastq files have a header, the sequence, the separator (a +) and the quality information (used for alignment).

Quality = during nucleotide sequencing, it is the probability that the nucleotide is wrong. Small probability = good quality. This probability is encoded in ASCII codes, which are characters present in American rider machines. The quality score (Phred score) derives from the score that was used to sequence the human genome; it was designed for Sanger sequencing, but then it was readapted for new sequencing technologies.
$Q = -10 \log_{10} p$

The score will be a small number (a lot of 0 will be present before the first number) due to the p value, but when the log10 is calculated, the number will be bigger, allowing a better representation of the quality. When the quality goes to 9, there is a good alignment between the nucleotides, while at values >9 I lose the linearity between the sequence and the quality. Today, the highest Phred score is 40, which has an accuracy of 99.99%. in order to have the linearity between the sequence and the quality score, we use the ASCII table for a type writer: as we can see from the first column of the table, 0-31 characters are not written in a conventional ASCII term (they give a low quality). 33 is associated to "!", while 73 is associated to "I". In order for a character to represent data in a good way, a Phred score 0 is equal to 33, while when it is equal to 40 (the highest value), it is represented by the "I" character (associated to 73). From "!" to "I", I represent all the quality scores that are assigned to nucleotides during sequencing. For each base, there is a character associated: in the slide, the first letter is a G and it is associated to "!", so it has a very bad quality (0), which is related to a high probability that the first nucleotide is wrong. Quality is used by the software to associate a character to a position of the genome during alignment.

These sequencing data are generated as FASTQ files, whose general scheme is called bcl. The files are divided based on the index (which uniquely describes the sample): from it, other different files are created for each sample. The sequences that I get are completely randomly positioned, because they derive from a big "container" of many sequences (made of many millions of reads). When bulk RNA sequencing is performed, sequences are made of 20 million reads per sample, which is enough to detect the location of GENES (not transcripts!) in the genome. If also the non-coding genes have to be investigated, the number of reads should increase around 800 million, because of a higher complexity of the non-coding RNA. During sequencing, if I

sequence 5 million reads (for example), I will however obtain a result, which is related to the most expressed genes only (low resolution). In order to obtain a reasonable representation of my sample, I need a sufficient amount of reads that represent what is in my sample. Moreover, absolute evaluations are never performed! Comparisons are done, which measure the efficiency of retrotranscription of different genes in different samples (example: comparison of the expression of GAPDH in sample 1 and sample 2). What is the output that I obtain from a sequencer? The output contains the name (in the slide, NIST…), the barcode (associated to the name of the sample), L (which represents the lane of the sequencer from which data come. Multiple lanes are used, and this helps to track the quality of each sample) and R (which indicates the reads: read 1 = forward read, from 5' to 3'; read 2 = reverse complement read). For each read, there is a @ indicating its name, followed by some parameters that are used to check the quality of the sequencing. The last character is 1 for read 1 or 2 for read 2. Given a fastq for read 1 and a fastq for read 2, they have the same length but a different last character of the name.

QC raw data

If the sample is bad, it will not be improved by the bioinformatic analysis; fishing expeditions = procedures made from people having a lot of money, which consist in sequencing bad data. It does not work! Experimental balancing is important, and noise must always be considered: it is due to technical variations of the procedure.

Evaluation of the sequence quality = evaluation of the results coming from the sequencer. This can be done with a software, called fastQC: it is a Java software, which can be used for DNA or RNA sequencing. In the slide there are 2 panels, which derive from the analysis of 2 reads. Both qualities are pretty good, but the 2nd one starts decreasing in the reverse read. However, the medium values stay in the green region.

In the next slide, the analysis is not so good; these problems can be solved by eliminating low quality reads from the analysis. Instead of analysing everything, I analyse reads from the 1st to the 15th, because the following ones are bad quality. Both reads are analysed if more complex and sophisticated phenomena (like splicing events) must be analysed, otherwise the analysis of one read only is enough, if the read has a length > 50 nucleotides. Every bar, in the graph, is 1 nucleotide of the read.

Another plot that is obtained from FASTQ analyses is the one present in the slide: this is a description of the changes with respect to the mean expected quality in the different lanes of the analysis. This plot should be always blue = the variation with respect to the mean is very small. If this value increases and becomes red, this means that some lanes are very bad.

Another output that is obtained from sequencing is the ratio of the incorporation of the different nucleotides into the sequencing points (?): if there is a distribution like the one that is obtained in humans (where each nucleotide is equally represented), everything should be around 35%. However, depending on the sequencing procedure that I use, thing change. For example, in the slide, Ribo-seq is used: polysomes are purified, the RNA that is uncovered by the ribosomes is cut and the pieces of RNA coming from them are obtained → retrotranscription → DNA. The overall distribution of the fragments is around 25%. The reason why there is a strange distribution in the initial part of the plot is that not all the random examers have the same affinity for the target: the probability that 5G and 1A are present in the same read is higher than than 5A and 1G; some sequences are ineffective during the annealing, so there is a discrepancy regarding the sticking of the examers on the read.

Slide: this is a kit that sequences only the 3' end, it is used for scRNA-seq. In eukaryotes, 3' ends are non-coding and have a different composition with respect to the coding exons → different distribution in the graph.

fastQC also provides the distribution of the GC content, whose distribution, in RNA sequencing, can be different from the expected sinusoidal one. Even if I get multiple peaks, this is not important; what is important is that the behaviour of the samples is the same, so that the distribution of the data is the same. In fact, when data are put together, the scheme of the curve is the same.

Sequence duplication = during sample preparation, PCR (12-15 cycles) must be done in order to have enough material to work with → sample homogeneity, but some sequences can be artifactually over-represented. This could be an issue in the analysis of the differentially expressed genes. Whenever I do bulk RNA seq, if the removal of PCR artifacts is essential to guarantee a good analysis. During bulk RNA analysis, a typical thing that is done is ion fragmentation: Mg is put at 65°C, and RNA is fragmented in different positions (double-stranded or secondary structures of RNA are less sensitive to fragmentation than single strands). In scRNA-seq, artifacts have to be eliminated, but this is not required during bulk analysis.

During the removal of PCR duplicates, the ability of having few false contaminants increases.

After RNA degradation, the adapters have to be eliminated. To do that, there are specific tools called trimming: they allow to trim the adapters or to remove sequences below a certain threshold (example: 20. Each sequence in the 3' end will be trimmed, and only the ones having a quality = or > than 20 will be maintained, while the others will be removed). Skewer, for example, allows to perform trimming in single and paired-end data.

scRNA seq

In a library, there are 2 adapters (P7 and P5) bound to a piece of DNA. New sequencers allowed to load many samples at the same time, but indexes are required on the 2 sides of the sequence: they are used to discriminate different samples. This is what happens in bulk RNA sequencing, which is more consolidated than scRNA seq. Most of scRNA procedures are based on the 10X genomics approach.

Read 2 contains the piece of DNA that I want to sequence; R1 is mainly used to discriminate between 2 things: 10X barcode tells me which is the cell I am analysing (different cells have different barcodes), so I can discriminate cells. UMI (Unique Molecular Identifier) is 6 random nucleotides long and it changes among sequences coming from the same cell, while it does not change if it derives from PCR artifacts from the same cDNA that has been amplified for many times. The UMI allows to count how many transcripts are there in different cells. In the slide there are bad data, while in the other slide the quality is higher.

<u>Exercise</u>

- Download fastQC and run it on the command prompt
- Download "lesson 2 dataset 1" and "dataset 0" (this dataset will be used for fastQC). Dataset 1 is a real scRNA experiment.

The experiment is based on ALK+ lymphoma cell lines. An inhibitor of ALK is called Crizotinib, and it gives different effects: sensitivity, intermediate response and resistance. There are 14 sample, derived from the same cell line, that is treated and not treated with the drug. Which are the genes affected by Crizotinib over the various experimental groups (sen, int and res)?

fastQC is a quality control for fastq. What are the steps for the analysis? Fastq file → I generate the QC to see if the quality is good → I do the trimming (= I remove the adapters from the sequence) → alignment, which consists in taking my sequence and the genome, to see where my sequence maps on the genome. Once I do the alignment, I get a file (called gtf) which has the position of the known genes on the genome; reads dropping in specific regions of the genome will be associated to a specific gene. The next step is annotation and counting: when I do RNAseq, I am interested in seeing how many reads drop on a specific gene. So, I will have to convert all the fastQC of each sample into a vector of numbers that indicate how many reads drop

on a specific gene over the genes that are annotated on the gtf. At the end, I have counts for each gene of each sample. In the exercise there are 40 samples, so there will be 40 columns = samples and 20.000 rows = genes. The number of counts will indicate the number of reads that drop on a specific gene. Counting is done sample by sample, and it is compared later on.

Trimming and alignment are evaluated by multiQC. Starting from a folder, this folder will contain as many folders as the number of my experiments. In RA there is the fastq that I have to manipulate (QC → trimming → alignment → counting). These steps have to be performed for each folder. In this exercise, I will look at 2 different quality controls: a quality control performed on a single sample and another one performed on all the samples (which is done by multiQC). multiQC is a Python program, which can be installed locally (in this case it should work) or using docker. If docker runs and, on the command prompt, I type "docker", the output is a series of things that I can do.

- Docker ps gives a docker ID, the name of the docker image and other things.
- Docker ps -a shows me what dockers have run and in which container. Knowing the ID of the container I can see what's inside.
- Docker logs shows the logs coming out as the container is executed.
- Docker run is needed to execute a docker locally. -t makes it interactive: docker runs at the time it starts. -v indicates that a connection between the hardware and the docker box is created: the docker folder sees a file or a folder of interest that is present on my computer. -w indicates which is the working directory in which I execute the docker.

First, I have to move to the folder where my dataset is (I do that with cd) → pwd tells where I am. Next command: docker run -t -v "pwd" : "pwd" -w "pwd" ewels/multiqc (I use the command that describes the folder where I am to connect it with the docker. The first pwd indicates where I am, the second one tells the docker which folder look at). -w is the folder where the command has to work.