Genomics at a glance – Part 2/2

NGS Workflow

First step in NGS workflow is to obtain samples for the sequencing. Either DNAs or RNAs need to be extracted from samples. Unfortunately, sequencing technology is not at the stage that we could load samples directly onto a sequencer. The samples have to be sheared and become random 'short' fragments. Also, depending on NGS platforms, DNA or RNA short fragments have to be amplified in order to obtain sufficient amount of short DNA/RNA fragments before sequencing. This non glamorous traditional laboratory work is still very critical for obtaining high quality sequencing results.

There are small number of different NGS platforms available currently, but Illumina sequencers are the industry leading platform. More or less, these different platforms are based on sequencing by synthesis (SBS) technology except Oxford Nanopore. The fundamental of SBS technology is to capture signals when DNA polymerases add labeled complementary nucleotides on target sequences.

After a sequencer generates short sequence reads, these reads need to be mapped onto a reference genome to figure out the origins since a sequencer generates nothing, but millions of short DNA/RNA fragments. Adding meaningful labels on these short sequences, so called aligning, is the beginning of NGS data analysis. This enables downstream analyses. There are many different flavor of aligners available, but not all aligners are RNA-Seq analysis ready.

As illustrated on Figure 2, there are many ways to apply NGS in different studies. However, in terms of target sequences, those applications can be organized into two groups, DNA-Seq and RNA-Seq. Only about 9.2 percent of human DNA does something, and little over 1 percent of human genomes codes

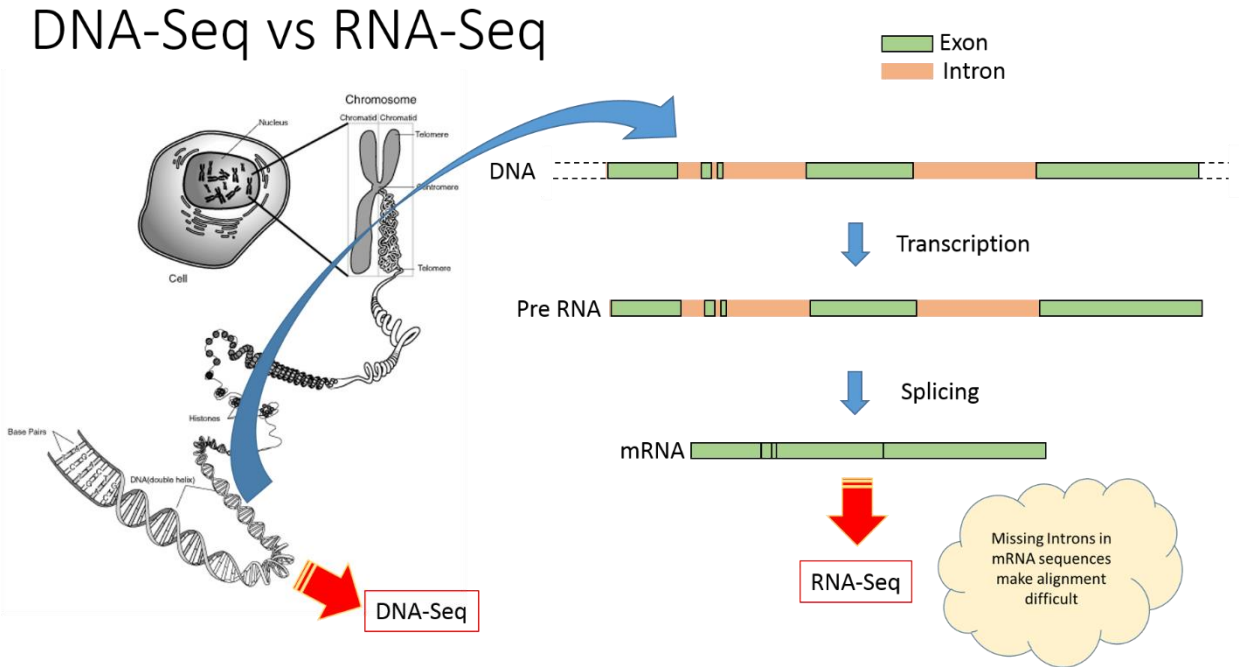

Figure 1 DNA-Seq vs RNA-Seq (The original image was obtained form https://upload.wikimedia.org/wikipedia/commons/9/91/Chromosome.gif and modified.)

proteins. And, these coding regions are shared by about less than 20,000 genes. Hence, a small fraction of chromosome sequences transcribed to mRNAs in the nucleus of cell as shown in Figure 5. Once pre-RNAs are synthesized, splicing mechanism process removes introns and rejoins exons to create messenger RNAs (mRNAs). These matured RNAs are called messenger RNAs since they are transported out to cytoplasm to deliver genetic information to protein synthesis apparatus located in cytoplasm. This makes mRNA sequences (transcriptome) quite different from gene sequences on chromosomes.

DNA-Seq

DNA sequences tell scientists which part of chromosomes carries genetic information (locations of genes) and which stretch carries regulatory information. More importantly, scientists can spot the changes in genes or regulatory regions that may cause disease. Unfortunately, these gene information do not come with sequence itself. There has been tremendous effort to label these sequences according to their functions such as ENCODE project. Without having a proper annotation for genome sequences, sequences are just very expensive garbage. Nonetheless, the ability to sequence genomes quickly and cheaply allows doctors to identify the particular type of disease a patient has. It will still takes some time that sequencing becomes a routine in the doctor's office, but the cheaper and faster NGS technology creates vast potentials for diagnostics and therapies.

Preparation of DNA sequencing sample is called as 'library preparation'. There are many different version of reagent kits commercially available. All these kits rely on a series of standard molecular biology reactions. A common procedure includes purifying genomic DNAs, making random DNA fragments and adding adaptors (known short DNA fragments attached to both ends of purified-fragmented DNAs). Genomic DNA library is then amplified through PCR (polymerase chain reaction) to increase the amount of target sequences to meet the requirements for various sequencers. Once sequencing is completed, sequence read files need to be mapped onto proper reference sequences with annotations. This is the so-called 'aligning' process. Since genomic DNAs were sheared randomly, the final sequence read files do not contain any information of their origin. Aligning process is used to map the locations of sequence reads on chromosomes and finding what genes are on that locations by looking up the annotation comes with a reference sequence. After alignment is done, further statistical analyses reveal where nucleotide variations are located on genes. This is an example of variance analysis, and the procedure is subjective and differs for different types of studies.

RNA-Seq

Not all genes are expressed at a given moment in a cell, and the distribution of mRNAs is cell or organ specific. Although all cells have the same copy of genes in their chromosomes, cells in different organs exhibit distinct gene expression patterns. There is not a direct correlation between the distributions of mRNAs and proteins, but mRNA profiling provides a better clue to estimate protein level activities.

Library preparation for RNA-Seq is nearly identical to the one for DAN-Seq. Extracting RNAs is the first step, but due to the instability of RNA molecules, these RNA sequences have to be converted to complementary DNAs before sequencing. The purpose of RNA-Seq is to obtain the information of which genes are expressed and the amount they are expressed at a given moment. Normally, RNA-Seq data is much harder to automate since there are many potential workflows. There should not be an expectation of uniform results for all RNA-Seq like what you expect from DNA-Seq because gene expression is changed over the time. However, this variability provides a unique opportunity to compare differences

between two cells under different conditions. For example, RNA-Seq is useful to compare gene expressions between normal cells and cancer cells from a patient.

Limitations of NGS

Neither DNA-Seq nor RNA-Seq alone can provide a complete answer for any biological question. This is not only NGS's problem, but also any experimental process or single technology in life sciences. A conclusion drawn from a controlled experiment will not actually reflect the real conditions in a life form. There is always the possibility that the same experiment with similar conditions might end up with a completely different result. Besides these innate difficulties NGS comes with errors, like any other technology; hence it requires a careful interpretation. Again, NGS is currently the best technology we can use, but nothing magical.