A telescope for the RNA universe : novel bioinformatic approaches to analyze RNA sequencing data

Pulyakhina, Irina

Chapter 6

General discussion and future developments

As can be appreciated from this thesis, RNA sequencing is used for a variety of applications [252, 253, 254]. Bioinformatic analysis of mRNA sequencing data can reveal the effect of different treatments via the analysis of differential gene or transcript expression (see **Chapter 3** for the examples of mRNA analyses in the context of aging). It can be used to explore alternative splicing, perform functional studies via pathway analyses, detect allele specific expression (see **Chapter 5** for the examples of ASE in context of breast cancer) [255, 256, 257, 258]. Nuclear RNA data analysis makes it possible to study intermediate splicing events and explore intermediate products and their impact on the splicing mechanism (see **Chapter 2** and **Chapter 3** for more details).

However, regardless of the progress in data analyses, one should not forget that the analysis is performed on *biological* systems. The analysis will produce artificial results when the sequencing experiment does not adequately capture the required data and the technical procedure to obtain the data introduces biases [259, 260, 261]. One solution is to improve the analyses, however, a more solid solution is to address and eliminate the steps introducing the biases. Note that technical artifacts will always be present in any dataset, however, it is important to recognize the biases and correct for them, which becomes easier with less biases from other sources.

6.1 Direct RNA sequencing

For RNA sequencing, the main source of biases comes from the cDNA synthesis step. The majority of RNA-Seq techniques does not sequence RNA directly. A possible solution to avoid technical artifacts which are incorporated during the cDNA synthesis step is to skip this step [266] – such techniques are called *direct RNA sequencing*, or *DRS*. Researchers started developing DRS after failing to elucidate numerous biases that cDNA synthesis introduces (see the **Introduction** for more detail).

6.1.1 Possibilities of direct RNA sequencing

The first DRS technique that became available was Helicos True Single Molecule Sequencing, or *tSMS*. Helicos tSMS [263, 264] works via the direct hybridization of RNA to the surface of an ultra clean glass slide containing poly(dT) oligonucleotides covalently attached at their 5' ends [265]. During sequencing, the terminating nucleotides with a fluorescent label are incorporated one per cycle. The terminator prevents incorporating multiple nucleotides in one cycle and the fluorescent label (specific per nucleotide) is used to identify the incorporated base [266]. Helicos tSMS single-end short reads are 30-35 nucleotides long with a relatively high sequencing error rate (the frequency of substitutions is

0.2%, insertions – 1.5%, deletions – 3.0%), which makes it hard to analyze such reads. Another common technical issue that biases the data and complicates the bioinformatic analysis is the "dark" nucleotides. A fraction of nucleotides remains unlabelled and therefore is not detected upon the incorporation. Such dark nucleotides will appear as deletions in the sequenced reads. The problem of unlabeled nucleotides is common for other sequencing techniques, but we only detect them with single molecule sequencers.

Bioinformatic analysis of the Helicos data revealed that, even though DRS has great potential and opens a wide range of opportunities, a new technique comes with new challenges. Helicos tSMS was the first NGS technique to sequence single molecules without the necessary PCR step, which was the main advantage of the technique. Another advantage was that very short sequences could be used as the input material, which would be beneficial when dealing with fragmented DNA, i.e., ancient DNA which is often partly degraded. The main difficulty for the analysis in Helicos tSMS technology is raised by the presence of dark nucleotides. Bioinformatic pipelines that were designed, tested and adjusted on the Illumina data cannot be directly applied on Helicos data, and new tools for the analysis of DRS need to be developed.

6.1.2 Future of direct RNA sequencing

Along with Helicos tSMS, more techniques are currently being developed and are expected to be available soon. One of them – DRS by Oxford Nanopore – will be discussed.

Oxford Nanopore[1] sequencing is a new technology for nucleic acid sequencing [267]. The technique uses a protein nanopore incorporated into a polymer membrane (Figure 6.1). The membrane has a very high electronic resistance and a potential is applied across it. A DNA molecule is sequenced as it passes through the nanopore. Every nucleotide passing through the nanopore gives a disruption of the membrane potential, which can be measured and associated with a base [268]. In addition to DNA sequencing, Oxford Nanopore is currently developing direct RNA sequencing and adapting the nanopore to distinguish RNA nucleotides. Protein sequencing using nanopores is also potentially possible, as proteins have been shown to move through nanopores [269, 270, 271]. The major challenge of unfolding a protein – denaturating the tertiary protein structure and making the amino acids pass through the nanopore – has recently been successfully addressed [272], which promises a wide potential of protein sequencing using nanopores.

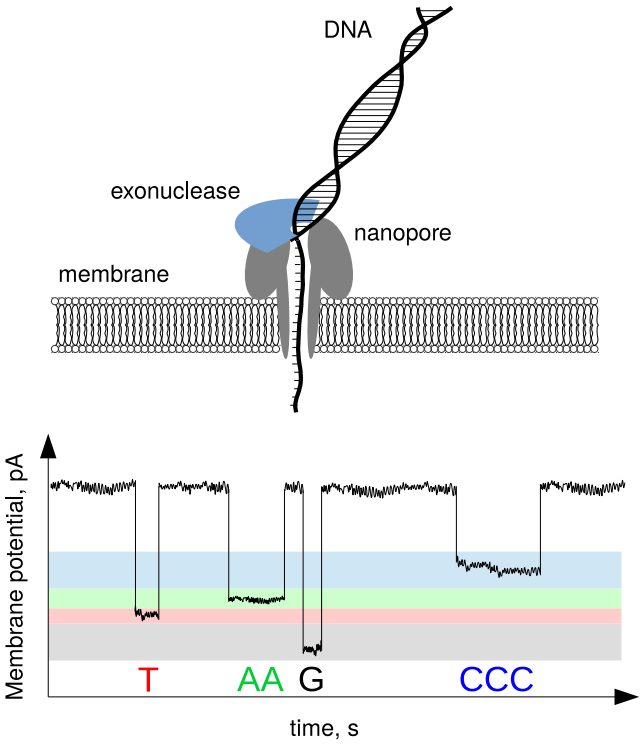

Figure 6.1: Main principles of the Oxford Nanopore sequencing. An exonuclease attached to the protein *nanopore* situated in a hydrophobic lipid layer (*membrane*) cleaves nucleotides on one of the DNA strands (left panel of the figure). The second, untouched DNA strand passes through the nanopore. Every time a new nucleotide going through causes a disruption of membrane potential (right panel of the figure). Based on the intensity and the duration of the potential the number and sequence of passing nucleotides are identified.

As mentioned before, together with the opportunities and expanded applications of direct RNA sequencing come new technical and bioinformatic challenges. One of the technical challenges can be RNA secondary structures. When researchers are interested in full length RNA sequencing, regions of RNA forming semi-stable structures might be more

[1]https://nanoporetech.com/

challenging to sequence and will potentially have lower coverage. Another challenge comes from RNA modifications (see Section 6.2 for more detail), as we do not know how chemically modified nucleotides will contribute to the process of direct RNA sequencing. For example, using the Oxford Nanopore system, incorporating modified nucleotides will result in a different membrane potential. However, it might also come as an extra complication, as not all types of RNA editing are known at the moment, and not all changes in the membrane potential can be discriminated. This can result in misclassified nucleotides due to unknown changes in the membrane potential. Therefore, current pipelines and tools can rarely be applied directly on the sequencing data generated with novel sequencing techniques.

6.2 Contribution of RNA editing

When RNA-Seq data is generated with as few technical biases as possible, the analyses results can still contain misinterpreted observations due to biological mechanisms that are not well studied or known at all. One of such mechanisms that has a direct influence on the interpretation of certain RNA-Seq analyses and which has been underestimated for a long time is the phenomenon of *RNA editing*. RNA editing is a chemical modification of mature RNA (not necessarily messenger RNA). A number of chemical modifications [273] has been discovered, although the function of the majority of these modifications remains unknown. Some types of RNA editing are introduced enzymatically, while others may occur due to chemical instability, damage or free radical mediated adduct formation [81]. RNA editing events, or *REEs* happen with different frequencies, the most frequent and well studied being editing by deamination and methylation.

6.2.1 Types and potential function of editing

One of the most frequent RNA editing events – adenosine deamination – is performed by *ADAR*, *A*denosine *D*eaminases *A*cting on *R*NA. As the result of this hydrolytic deamination, inosine is created in place of adenine. Inosine is recognized as guanine by the cellular machinery and will be represented as **G** in the output of Illumina sequencing (as the DNA polymerases used in the PCR reactions prior to sequencing will introduce a **C** at all positions complementary to **I**) [82]. **A**-to-**I** editing is known to affect a large amount[2] of adenines [3] and effects mainly (but not only) double-stranded RNA. Since structural RNAs tend to form double-stranded structures, they are extensively undergoing **A**-to-**I** editing. This type of editing can be specific or promiscuous; in some cases a particular position is edited in multiple molecules and in other cases different positions within a certain region are edited in different RNA molecules [274]. Studies on RNA editing in cancer [275, 276, 277, 278] point towards the instability of RNA editing (meaning that the same position will not always be edited in all the transcripts) and go hand in hand with the potential importance of RNA editing.

The second most frequent type of RNA editing event, detected in up to 0.5% of isolated RNA molecules, is RNA methylation [279, 280] performed by the methyltransferase enzyme. Methyltransferases bind to RNA regions with a certain sequence and methylate adenosines [281, 282, 283]. RNA methylation is enriched near stop codons, in 3' UTRs and long internal exons, and methylation was discovered to happen during or after splicing [284].

RNA editing and its functions have not been studied extensively yet. Position-specific RNA editing can be used by cells as a mechanism to recode genomic information and increase functional protein diversity, as editing can take place in exons and alter the coding sequence [285, 286]. Promiscuous RNA editing is considered to play a regulatory function, i.e., occur in miRNA targets or protein binding sites, as it is also known to often take place outside the coding region [287]. Editing happening immediately

[2]Note that the frequency of **A**-to-**I** editing can be low, even too low to be distinguished from sequencing errors, therefore estimating the amount of edited positions is challenging.

after transcription prior to splicing might have an effect on splicing progress, as edited positions can be located around splice sites in the regions that the spliceosomal proteins bind to.

6.2.2 Bioinformatic analysis of RNA editing

RNA editing is a fundamental biological process which deserves thorough exploration on its own. However, as can be appreciated from **Chapter 5**, it is also crucial to detect REEs because they can interfere with RNA-Seq bioinformatic analyses. For instance, **A**-to-**I** REEs can be misinterpreted as **A**-to-**G** genomic variants in samples with unknown genotypes. Since editing does not happen in 100% of the transcripts, it might cause false-positives in the detection of allele-specific expression. There is currently a scarceness of bioinformatics methods to reliably detect REEs [288, 289]. At the moment, all RNA positions different from the reference genome and not present in the list of genomic variants are often considered REEs or sequencing errors and are discarded from the downstream analysis [290]. It is a considerable challenge to differentiate between REEs and sequencing errors. As mentioned before, it is performed by the editing enzymes that introduce the chemical modification at a single RNA molecule level, therefore the frequency of RNA editing varies widely and is mainly very low [278, 291]. There is a strong need in new bioinformatic pipelines to reliably distinguish between genomic variants and REEs [292].

6.3 RNA and proteins

Bioinformatic analysis of mRNA sequencing data is challenged by both technical and biological issues, but when the analyses results are clean and reliable, the next step is to study the consequences of these results at a protein level. Various events that are detected in the RNA-Seq data may have an effect on the sequence and thereby the structure, stability, expression level and even function of proteins. Note that this has limited effect in the experiments, when two samples with similar protein profiles are compared, and has a greater effect on the experiments when no controls or references are used (i.e., one sample or sample group).

6.3.1 RNA and protein expression

Integrative -omics analysis is a class of methods that link RNA and protein expression [293]. RNA-Seq and mass spectrometry data is combined to identify expressed genes and proteins/peptides and to get a reliable list of expressed proteins (Figure 6.2). This is a powerful approach, since not everything that happens with RNA will be translated and will lead to a functioning protein [294, 295, 296], in other words, $RNA \neq protein$. Not all RNAs will be translated, nor when present, with the same efficiency. It has been shown that the correlation between protein and gene expression is not high, which indicates either biases in the process of sequencing and mass spectrometry [297, 298], or the bioinformatic analysis [299, 300], or the combination of both. One of the main pitfalls in the integrative -omics is that the protein expression is correlated with the *gene* expression, while the real RNA molecule that leads to a protein and should be correlated with protein expression is a transcript [294]. Since more than one transcript is produced from the majority of human genes, and proteins have different stability, correlating gene and protein expression becomes artificial [301, 302].

A logical step to correct for the biases would be to use transcript expression instead of gene expression. Gene expression is measured as the cumulative expression of all transcripts produced from it and expressed in the cell. However, different proteins can be translated from different transcripts of the same gene. Unfortunately, the latter is more difficult to analyze, as to measure transcript expression actual transcripts need to be identified. With currently used Illumina sequencing as one of the main

sources of the RNA-Seq data and its relatively short reads of hundreds of bases, reliable detection of full-length transcripts is challenging and often impossible [303]. This task becomes harder when the difference between the transcripts is minor, i.e., in such cases as allele specific gene expression [304]. The structures derived from the two chromosomes may differ due to allele-specific alternative splicing which can be tagged only with one heterozygous SNP [305]. Due to ASE, a certain allele coding a deleterious or a toxic protein (unlike the other allele) might be higher expressed, which might lead to the overexpression of a deleterious protein [306]. However, it is very complex to study its effect on protein level. When two transcripts differ in only one nucleotide, current limitations of Illumina read length (hundreds of nucleotides) will fail to distinguish between the two transcripts on a whole transcript scale. Using sequencing techniques producing longer reads, such as Pacific Biosciences or Oxford Nanopore, might become a pivotal point in the integrative -omics analysis. However, even transcript and protein expression does not always show a good correlation. E.g., as mentioned above, some RNAs can be translated more efficiently than the others. A current study [167] shows a decoupling between transcript and protein expression occurring with age.

6.3.2 RNA and protein structure

Whereas tools for linking RNA and protein expression are developing quite fast, the link between RNA and protein structure remains understudied. Genetic variants and REEs can cause non-synonymous mutations resulting in altered protein sequence, structure and function. A mutation in the human dystrophin gene in the locus Xp21 of the X chromosome can lead to the development of a muscular disorder called Duchenne Muscular Dystrophy. The dystrophin protein (encoded by the dystrophin gene) is responsible for connecting the cytoskeleton of the muscle fiber to the extracellular matrix. Known out-of-frame mutations result in a transcript from which it is not possible to produce dystrophin. This leads to the loss of the dystrophin's function and the development of a muscular dystrophy [307, 308, 309]. A mutation in *LMNA* gene – a gene encoding lamin A, a protein providing structural support to the nucleus – creates a 5' cryptic splice site within exon 11 resulting in an abnormally short mature mRNA transcript [310]. This transcript yields an abnormal isoform of prelamin A – precursor of lamin A – which is not able to provide the necessary support to the nuclear lamina, which leads to reduced cell division and premature aging (also known as progeria, or Hutchinson-Gilford disease). Another example of a mutation which causes the creation of non-functional protein and leads to a disease is a mutation in human *CFTR* gene (*C*ystic *F*ibrosis *TR*ansmembrane conductance Regulator) located in the q31.2 locus of chromosome 7. The proteins encoded by *CFTR* functions as a channel for chloride ions, which can move in and out of cells. The most common mutation, which is a three nucleotide deletion, leads to the loss of an amino acid and a non-functional protein [311, 312].

Apart from single nucleotide changes, alternative splicing events might lead to a whole protein domain removal or incorporation. Studies of protein structures and structural alterations are challenging, as structural biology and exploring protein structure and molecular dynamics involves computationally heavy, intensive tasks. These tasks include protein modelling, molecular dynamics and quantum mechanics calculations (to be precise, short cuts based on knowledge derived from quantum mechanics), i.e., to study the changes in the binding between the substrate and the amino acids in the enzymatic active site, which can be caused by a genomic variant [313]. Apart from the need in extended computational power and resources, the task of linking transcriptomics and structural biology is challenged by the limited amount of available 3D protein structures. This makes the task even more complex, especially when considering linking transcriptomics and structural biology genome wide. Unfortunately a possibility of local modelling of certain protein domains does not give a lot of insight, as even very conserved and closed domains often behave differently in different circumstances [314, 315]. These circumstances can be the surrounding amino acids, the electrostatic potential around them, or the level of the domain's exposure to the surface.

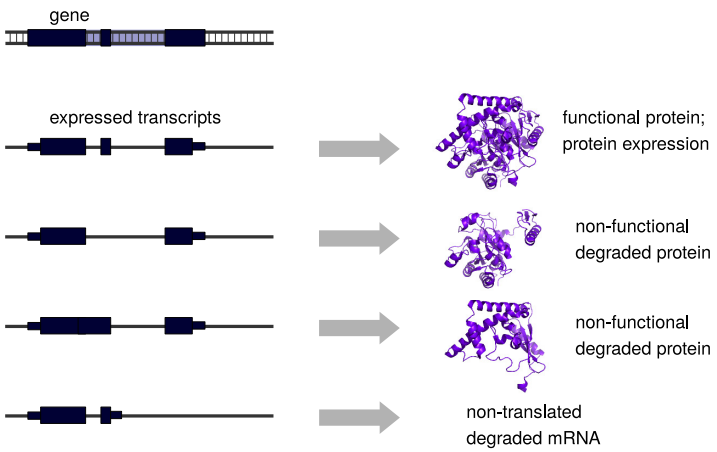

Figure 6.2: The Figure shows that multiple mRNAs can be transcribed from one gene (left panel of the figure). Some of the mRNAs lead to a stable functional protein, while some result in non-functional protein and some are not used as a protein template at all.

A developing field of *structural genomics* might be the bridge connecting transcriptomics and protein structures and functions [316]. The principle of structural genomics is to identify as many protein structures as possible using all currently available resources and techniques. These techniques include *de novo* structure determination using X-ray crystallography and Nuclear Magnetic Resonance, modelling based methods such as *ab initio* modeling, sequence based homology modelling and domain modeling based on the fold similarities rather than the sequence identity [317, 318].

The intermediate goal of the field is to provide as many protein structures as possible, and the end goal is to identify and describe all proteins within the genome of interest. Structural genomics has already successfully been applied on bacteria [319] and extending the initiative to higher organisms, such as humans, is hopefully a matter of time.

The initiative of structural genomics has a great potential, as it will expand our understanding of protein folding, structure conservation among and within species, provide a better and broader understanding of protein function, and make protein stability and function prediction easier and more reliable.