Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis

Prof. Aritri Bir

Dr. B.C. Roy Multi-Speciality Medical Research Centre

Indian Institute of Technology Kharagpur

Lecture 38 : RNA Sequencing: Role in Infectious diseases I

Namaskar. Welcome back to the lecture series in NPTEL platform. So, we were in our lecture series of comprehensive molecular diagnostics and advanced gene expression analysis. So, in the week of molecular diagnostics in infectious disease now we are going to start the importance of RNA sequencing. So, today's class is on RNA sequencing and its role in infectious disease diagnostics. Here in the basic concepts we are going to cover the concept of RNA sequencing, what are the different steps of RNA sequencing, transcriptome analysis and finally, applications of RNA sequencing as a whole and as well as in the infectious disease platform.

So, the very basic concept of RNA sequencing is basically this is one technique which examine the quantity and sequence of RNA in a sample using next generation sequencing. So, basically this is the as a whole this is one analysis of the transcriptome which indicates the genes encoded by the DNA and those genes we need to understand whether the genes are on or off and if it is on to which extent it is working similarly to which extent it is silenced. So, this is what RNA sequencing is all about. So, in the RNA sequencing what is done the sequencing data after RNA sequencing the data comes as a short reads of mRNA.

Now, if we see the genome there are pre mRNA sequences there are exons there are introns, but when we get the mRNA there are exons only there is no intron. So, the short reads which we get after RNA sequencing are basically of exonic sequences and that is free of intronic non coding DNA. So, the reads are aligned back to the reference genome. So, if we get the reads in search see here you can see these are the short reads which we get after sequencing. So, the reads now need to be aligned over the reference genome to get the whole reading of the genome.

So, if we come to the RNA sequencing workflow it was initially based on the Sanger sequencing the older method, but then definitely NGS or the next generation sequencing has prevalent for over the sequencing technology. Now, the workflow which is followed in RNA sequencing is basically extraction of the RNA after extraction what to be done the RNA needs to be converted to cDNA by reverse transcription process after that what is done is ligation of one adapter molecule. Now, the regarding the adapter molecule what is the importance and why this is done that we are going to discuss soon. After the ligation of the adapter that specific part is amplified the commonest amplification method is PCR after this amplification RNA sequencing is done. So, this is the step by step over workflow of RNA sequencing.

Now, regarding the adapter ligation. So, what is basically done after the fragmentation of the cDNA into smaller parts adapters are added at each end of the fragments. So, what is this adapters these are the short pieces of DNA around 80 bases in length. So, these are the nucleotide pieces and they are added to the ends of the cDNA and these ends basically are required for identification to be identified by the primer. So, basically they attach to the target primers for amplification.

Now, there are three main components in these adapters one is flow cell binding sequence. So, this flow cell binding sequence is basically a platform specific sequence and as evident by the name it is it binds with the flow cell. Now what is the flow cell? Flow cell is basically the container the instrument where the sequencing is done. Now inside this flow cell how they bind this adapter oligos are basically complementary to the oligos which are attached in the sequencing flow cell. So, in the flow cell where the sequencing is conducted basically there are oligos attached over the flow cell and those oligos are complementary to the adapter nucleotide sequence.

Now the P 5 adapter is basically binding with the 5 prime end of the flow cell oligos whereas, the P 7 adapters bind to the 3 prime end. After the flow cell binding sequence there are another important region of this NGS adapter that is sequencer primer binding site. Now basically this sequencer primer binds with the target primer it can be single end sequencing where the binding is done or conducted from one end only whereas, there can be pair end sequencing where the binding is done on the both end. Apart from this there can be tags tags which act as sample bar codes. So, each specific mRNA population or the specific mRNA sequence is having its unique identity by this sample bar code and that is known as index or bar code region.

The regions which tags specific mRNA are basically the bar code region or index region for that specific pool of mRNA. Now attaching those tags basically what can be done in a flow cell in a flow cell chamber if there are multiple index bar code tags are located the multiple samples or multiple cDNA or rather multiple sequences can be pooled in the flow cell and can be sequenced in a single run. So, that is basically sample multiplexing. So, this is how the adapter helps in amplification by increasing the specificity and identification. So, this is the flow of RNA sequencing where you can see the mRNA is converted to the cDNA then there are tags which are attached after that we get the reads short sequence reads then what we need to do we need to read the reads or align the reads to get the complete sequence.

Now coming to one very specific type of RNA sequencing that is single molecule real time sequencing or SMRT sequencing. Now as we are talking about single molecule. So, basically we are directly sequencing individual single RNA molecule in real time using one specific technology that is zero mode wave guide technology. So, basically this SMRT sequencing is done over a chip which contains this ZMW technology. Now what is this zero mode wavelength wave guide technology? The ZMW technology is basically one nanophotonic confinement structure where there is a small hole or small well which is surrounded by metal film.

Now the hole is such typical that it is much smaller than the wavelength of the light. So, basically it restricts the entry of the light into the ZMW chamber. Now this confinement of excitation light and the fluorescent emission to a very small volume it is done in such a way that it enhances the signal to noise ratio. So, a single molecule can generate fluorescence and that can be detected by zero mode wave guide technology. So, this ZMW technology is integrated over a sequencing chip over which the sequencing is done and over this chip basically this sequencer nucleotide or this DNA RNA molecule the target nucleotides are sequenced in real time.

So, how it is done? Over this chip basically there is one single active DNA polymerase and one single molecule of single stranded DNA template both are immobilized over this ZMW well or confinement chamber. Now definitely there is reverse transcription to this cDNA now this cDNA molecule is ligated over the adapter sequence and immobilized in the ZMW well. Then that specific DNA polymerase which is active now binds over this immobilized cDNA molecule and begins the sequencing process. Now what happens the nucleotides here is basically phospholink nucleotide and during the sequencing process whenever a single nucleotide is added by DNA polymerase what happens the fluorescent tag on incorporation of a single nucleotide is basically detached or cleaved off. So, the fluorescent which where previously present is now off.

So, this loss of fluorescence is detected by the detector and this is how one single nucleotide incorporation or rather one single nucleotide addition is detected in real time at the level of single molecule. So, that is the single molecule real time sequencing is doing here. Then coming to single cell transcriptomics. So, previously in the single cell molecule sequencing we were talking about a single RNA. Here in single cell transcriptomics we are talking about a single cell we are talking about a cells transcripts.

So, basically here we are reading all the RNA profile or the pool of RNA from a specific type of cell. So, we have heterogeneous type of cell population from that we are accessing one specific type of cell and reading their transcriptome that is single cell transcriptomics. Now what is the advantage over the whole cell transcriptomic over this rather the advantage of single cell transcriptomics over the whole cell transcriptomics or the whole population transcriptomic. So, basically what happens there is cellular heterogenicity and this heterogenicity can be can come out with good resolution with this single cell transcriptomics. So, basically what is done gene expression analysis can be done in individual cell within the population and that helps in identification of rare cell type.

So, basically if there are multiple types of cell type multiple types of cells. So, if we get the individual cells transcriptome we can detect some abnormality or some new unique feature over the transcriptomes based on that we can analyze or identify rare cell types characterize that cell to cell variability and identify or detect subpopulation with specific and unique gene expression profiles. So, this is how cellular heterogenicity can be read via single cell transcriptomics. Then cell type identification and classification can be done following gene expression profile. Now cells residing different types of state they follow different dynamic changes on the basis in the basis of different stimuli or different diseases or different physiological state based on that cells response are different responses are different.

So, that cell state or different rare cell state or the transition state from the physiological to pathological series from the benign to malignant pattern all these can be read by a analyzing the single cell transcriptomic also different cell to cell interaction and communication can also be studied. So, the performance criteria of isolating one single cell depends on three things that is the throughput, purity and recovery. Throughput basically it indicates the number of cells that can be isolated per unit of time then purity refers to definitely no mixing means the number of cells collected after separation from the tissue. So, how much the they are separated the different population are separated from each other is the purity and finally, recovery is the final amount after this separation and separation how much amount the final amount of target cells we are getting in our hand is the recovery. So, these three factors basically decides the single cell transcriptomics.

Now, how this single cells are isolated there are various technique the very common one is FACS or fluorescence activated cell sorting where single cells are isolated based on their fluorescence and physical properties. These cells are basically labeled with fluorescent markers which target specific cell surface protein or different intracellular molecules. So, every cells are having their specific markers which are tagged by different fluorescent dye. Now, when this label cells passes through the flow cytometer the instrument where FACS is done flow cytometer they are sorted based on their fluorescent signal to individual wells or tubes and then they are analyzed further. So, this is how single cells are isolated based on their fluorescence.

Now, definitely there are microfluidic platform which is done based on droplet based technique. So, what happens the cells are basically encapsulated in droplets or micro chambers within the microfluidic devices which allows that individual cells to be isolated and the throughput here in microfluidic platform are very high where single cell isolation is a very pure technique. The purity here is very high along with the throughput. Then definitely there is manual picking where individual cells are visually identified and then isolated based on their physical properties using micro pipette under microscope. So, this is one micro pipette based technique done manually.

Then there is laser capture micro dissection LCM which allows isolation of specific cells from a specific region of interest. Here of course, we are using laser beam which basically cut and capture those individual cell or cell clusters from a histological sample and this cell then collected for transcriptomic analysis. Then micro well based platform, micro well basically are the array platform which consists of thousands of wells which can trap single cell from different heterogeneous cell population. Now the cells are distributed in the wells you again using microfluidic or gravity driven methods and then this multiple cells simultaneously can be processed for transcriptomic analysis. So, these are the different techniques by which we can isolate single cells in single cell transcriptomics.

After the isolation there is amplification of each transcript then there is sequencing of each single cell RNA. Now what is important is normalization of this RNA sequencing data because there is cell to cell variation, the variation in the efficiency library formation and sequencing. So, the normalization of the data is very very important. Now this normalization how it is done. So, it can be done via using internal controls.

Those internal controls are added from outside. So, those are the extrinsic RNA spike ins. So, what are these RNA spike ins? Those are RNA sequence of known sequence as well as known quantity. Those are added in equal quantity in each cell lysate and they act as internal control. So, basically when there is amplification along with the target cell the target cells RNA these extrinsic RNA spike ins are also amplified.

Then this normalization we do by using the scale the concept is basically scaling the expression values of endogenous gene based on the expression level of spike in control. So, we can check the ratios or the relative expression based on the expression of this spike in controls. Now where this spike ins are used? This spike ins are particularly useful for detecting technical variation or whether there is any biases or there is manual error etcetera. So, basically there is one comparison of the samples and their different conditions which are basically checked or rather compared with the expression of the spike ins. So, these are the extrinsic spike ins which are used as internal control.

Apart from that there is another thing that is unique molecular identifiers or UMIs. Now UMIs are basically short random nucleotide sequences which are added to the cDNA even before the library preparation. So, what it does? It basically distinguishes between the PCR duplicates there are PCR the same gene the same RNA is basically duplicated due to the PCR. So, there is a whole population of product which is coming from the same gene and they are having that same unique molecular identifiers. So, basically if there are two different RNA two different types of RNA or two different types of genetic code that can be differentiated based on this UMIs.

So, how they help? Here the normalization involves counting the number of UMIs associated with each gene and they are compared basically that is scaled the amplification values are scaled based on the total number of UMIs in each sample. So, how much a sample is having the UMIs basically how many types of UMIs are there is basically indicating the types of RNA transcripts present in that sample. Here the PCR amplification biases can be avoided using this EMI. Now both of these internal controls the extrinsic spike ins as well as the UMIs they can be used together in a combination to normalize the RNA sequencing data. So, in this class we have got the view of how the RNA sequencing is done the steps like RNA extraction after that there is formation of cDNA via reverse transcription then there is ligation of one adapter molecule then it is amplified by PCR and after that there is sequencing.

Now the sequencing are of different types apart from the total transcriptomic analysis there can be single cell RNA sequencing whereas, cells sorry the single molecule RNA sequencing where a single RNA can be read or studied following ZMW technology and also sequencing of the transcriptome in a single cell via single cell transcriptomic analysis. These are my references and see you in the next part of this RNA sequencing in the next class. Thank you.