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 39 : RNA Sequencing: Role in Infectious diseases II

Namaskar. Welcome back to the NPTEL lecture series of Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis. So, we are in the week of molecular diagnostics in infectious disease. In the last class we read about RNA sequencing the very basic concept. Here we are at the second session or second part of RNA sequencing. So, in the last class we read the concepts different steps.

In this class we are going to cover the transcriptomic analysis and the application of RNA sequencing in infectious disease. So, what is done that in the workflow of RNA sequencing if you remember after isolation of the RNA they are converted to cDNA, then those cDNA are basically after ligation of the adapters they are amplified in then following the next generation sequencing principle they are sequenced. So, at the end there are short reads of sequence we need to read or we need to study and align to get the whole genome. So, the analysis of data workflow follow that we after the sequencing is done we get a file which contains the read sequences from the NGS platform, then those read sequences are aligned to the reference genome and finally, quantification of those expressions or quantification of those genes expression are done.

So, here you can see the workflow after we are getting the sequencing reads we need to align it to the genome, then we need to assemble those transcript after that we can quantify their expression levels. Now, let us go a bit in details. So, we are basically getting short reads and we need to map them over the reference genome. Now the problem is that as we discuss that the mRNA are having only the exons there is no introns and those introns are basically spliced of via splicing mechanism after that the exons are joined together. So, in the reads we can get those exon exon joining site or the splice junction.

So, we need some reference genome which can supplement or the reference genome must have those exon exon splice junction site over which we can map the reference genome. So, basically if there is one reference genome which contains the exon exon splice site. So, if we get one read which has this splice site the mapping is easier.

Similarly, we need some technology some computational software which can identify this the presence of this splice junction. So, splicing aware aligner is required there are available splicing aware aligner like GASnap, MapSplice, Rum, Star, Top Hat they are the computational software which can identify that yes there is a splice junction.

And they can recognize the difference between a read aligning across and exon intron boundary and can read with a short insertion. So, basically consider this is an exon and this is one intron. So, these splicing aware aligners they can read this junction also they can read the two different exon splice junction. So, these are the splicing aware aligner software programs which are required for mapping this short segment. Basically mapping is aligning the short segments of sequence RNA sequence or reads according to the reference genome.

Now while assembling here you can see we are doing the after the reads are we we are going to map the reads and this map can be assembled over the transcript following two mechanism that can be a reference based mechanism or that can be a de novo assembly. Now using the reference mechanism when we are trying to map the transcript it basically indicates the location and structure of the known transcript over the reference genome. So, basically if we are having one reference genome. So, this is our reference genome over that this maps this short reads are aligned fine. Then what we can do we can quantify the abundance of the known transcript and also alternative splicing or different isoform can be identified.

So, basically this reference based transcript assembly is totally dependent on the reference genome. Reference genome or reference genome must be available that should be accurate in the sequence and the annotations are are easy to read or rather we can use the annotation to read the genes. So, basically using the reference genome we can locate different sequences over the genome. So, this is the reference based technology. Now we are coming to the de novo assembly.

De novo assembly is where we are not using any reference genome or annotation. So, the reads are aligned de novo and this is first assemble into a longer conti or longer contiguous sequences and that uses one de novo assembly algorithm. So, algorithms are there which can read the sequences based on their contiguity. So, if we are having a read and another read see here you can say that these sequence are overlapping. So, these are contiguous reads.

So, after these we are having this part fine. So, here we do not need any reference we are not rather using any reference genome or based on their overlapping or contiguous sequences the alignment is done here. And this is how the full length of the transcript is reconstructed or the partial transcript sequence is done by overlapping region. And

finally, what we get is a transcript which is not known initially. So, basically when the reference genome is not available one unknown organism one unknown genetic sequence where we do not have the reference genome here this de novo reconstruction or de novo transcription transcripts assembly this method is utilized.

What we get is one novel transcript or some transcriptomic diversity we can dig into. So, this is the two different approach by which transcript can be assembled to get the complete or sometimes partial sequence which we have targeted. Then what we need to do we need to quantify or estimate the gene expression level. So, there is pool of computational softwares again which can basically estimate the gene expression level like cuff links flux capacitor then MISO these are the computational software which can count the number of reads and map the full length transcripts. Of course, then there is another type of alternative computational software like HT set which can quantify expression without assembling the transcript just by counting the number of reads that map to an exam.

Then again this estimation also needs normalization this normalization can use different types of indices one such is RPKM or reads per kilo base of transcript per million mapped reads. This is basically scaling the raw read counts by the total number of mapped reads in each sample. So, the raw reads and after alignment the total number of mapped reads in each sample they are scaled. So, this is one example of normalization. Similarly say there are different other indices by which we can normalize the quantification data of gene expression.

By RNA sequencing we can check the differential expression of different genes. Now what is this differential expression the differential expression means there can be two condition a diseased versus a normal one a treated versus a not treated one benign versus a malignant one. So, how they are differentially expressed that can be checked via using the RNA sequencing and there are multiple tools which can detect this. So, basically what is done two simultaneously two types of cells are sequenced two type of pools are sequenced one is a treated pool of RNA one is a non treated pool of RNA one is a control RNA versus one is a diseased RNA they are run and sequenced simultaneously and there are software which are simultaneously comparing these two types and that is the differential expression of RNA. Then alternative splicing can also be detected.

Now alternative splicing are basically arranging the exons in different way. Now the first thing which is required is to identify the splice junction. So, you all know splicing is basically if there is exon and there is intron and again there is exon. So, by the mechanism of splicing the intron is spliced off or cut off and the two exons are joined this is the splice junction. So, as we have discussed there are different computational software which can identify this splice junction.

So, that is splicing our aligners and this splice junction are identified by those computational algorithm after that this exon exon junction are estimated from different samples from different regions to identify the splice site. Now after the detection what is the type of splicing that can be detected. So, alternative splicing can be of multiple type one is exon inclusion type another is exon skipping type. So, basically how much and or the whether the exon should be one exon suppose there are three exons 1 2 3. So, whether this two exon number two exon will be included or that can be excluded.

So, the product can be 1 2 3 or the product can be 1 3. So, here basically this exon is skipped. So, these type of alternate splicing can also be detected in RNA sequencing. So, what can be done there can be quantification of the relative abundance of different mRNA isoform and the changes can be detected based on exon inclusion and skipping. And how that is done by analyzing the read coverage across the exon and splice junction.

So, how if we covered the whole reference genome then in that case we can detect the how much the exons are covered. So, the exon coverage can be done again differential exon usage analysis softwares are there which can detect how much exon inclusion or exon skipping means alternative splicing how much is present there. Again introns can also be retained that is intron retention it is another type of alternative splicing now that can also be detected by capturing the reads spanning intron exon junctions. So, if we have once again a reference standard which where we have refer where we have exon and intron junctions and if we have softwares which can identify the intron exon junction in that case we can identify another type of alternative splicing that is intron retention. Now, this intron retention identification helps in identifying the genes and the regulatory factors which are responsible for alternative splicing regulation.

Similarly, fusion gene detection can also be done using RNA sequencing. Now, this fusion genes as evident by the name that there are polymorphism or there is some alteration of the exon exon junction or there is some alteration or mutation in the sequence. So, there the final product is a fusion gene. Now, again if the short transcriptome is aligned over the reference genome what will happen there are long reads and there are short reads. Now, the long reads are basically indicating the very huge part of genome rather very long part of genome whereas, short reads they are basically the part of exons.

So, if we have such exon and there are short reads they can be incorporated into the exon, but if there is some fusion gene what will be what will happen suppose there is a fusion gene there. So, they will be unmapped they cannot be aligned over the reference gene or genome. So, this unmapped short reads are further analyzed to identify if there is any fusion point. Now, the reads that map to exon exon junction involving different

genes are indicative of fusion genes. So, consider these are the two exons fine.

So, in these read we are having these exon exon junction. Now, if we check backwards that when we are sequencing the short reads we are getting like this. So, basically to this reference genome we have aligned the reads. Now, if there is one fusion gene present suppose this part is basically fused with a new gene. So, these this short read is not aligned or not mapped.

So, now, we are analyzing this in these gene we can see that exon exon splice site is present, but yet it is not mapped over this reference gene. So, there might be a possibility that there is a new sequence which is added over this that is the fusion gene. Now, in case of paired end sequencing from the both ends we read the sequences. So, here in that case we are reading from this as well as from this side. So, the detection of fusion gene is much easier because we are reading it from the both side.

Coming to e q t l s or expression quantitative trait loci. Now, here basically there is integration of this RNA sequencing data with the data genetic variation data. So, it helps to identify genetic loci which correlate with gene expression variation. So, here we are identifying genetic variants. Now, what can be a variant? Suppose there is a single nucleotide polymorphism and identifying them is very important because these single nucleotide polymorphism or different genetic variants can be rather are associated with different pathologies.

Now, this e q t l analysis basically empty identify those variants which are spread over the population. So, the population study the population data of polymorphism can be obtained by e q t l s. Now, the e q t l s can be of two types one is cis e q t l another is trans e q t l. So, the cis e q t l are the variants which are located within or near the regulatory genes. They typically can affect the expression level of the genes on the same chromosome.

Those are mostly located near the promoter near the transcription start site like that whereas, trans e q t l are located elsewhere in the chromosome, but can regulate the gene expression over different chromosome as well. So, trans e q t l are basically away from the gene which is regulated by them. So, they are basically weaker, but can unveil new gene regulation pathways. So, from distant if it is regulating some other genes they have importance in they are important in identifying different new pathology or new physiological pathways. So, this is how e q t l can help in identification of variants via contributing in studying the gene regulation as well as if those genes are regulated or how this variation are bringing the consequences over the population.

So, different genetic basis of different diseases and traits can be studied via identify this

genetic variants. Different candidate genes can be identified which has been expressed via this trans e q t l or cis e q t l also it helps in raising population data. So, functional annotation of Gawa's finding can be much much help by this e q t l's. So, what it does it studies this genome wide association studies and link them to different types of disease associated polymorphism or S N P's and this is how they can identify target genes as well which are manipulated by the biological mechanism. So, e q t l helps in identify the variants and then their consequences.

Now, if we come to the importance or role of this RNA sequencing in infectious disease diagnosis definitely via doing this via conducting the RNA sequencing we can identify a wide range of infectious pathogens or agents starting from virus bacteria fungi parasites and what not. Then there are RNA virus which are very much known for their genetic diversity and rapid evolution in the strains like influenza virus, coronavirus, HIV. So, detecting their sequence genetic sequence or RNA sequence they can be identified and characterized easily via RNA sequencing. Then RNA sequencing can detect viral quasis species and different genetic variants different drug resistant strains their mutations those can be studied. Then the signature gene expression profile associated with different specific pathogens gene expression profile which decides the clinical outcome expression profile which is associated with different stages of the disease that can be studied via RNA sequencing.

Then based on this RNA sequencing different diagnostic platforms can be invented one such as RT q p c r based technologies or biosensors which can rapidly identify the pathogens. Then differential gene expression analysis it basically distinguishes the infected and uninfected individual by simultaneously studying different types of or different types of people or different types of target population. Also the RNA sequence sequencing helps in epidemiological surveillance and outbreak monitoring by tracking the spread of infectious disease. Identify the transmission cluster who is the contact we can trace by via contract contact tracing based on that quarantine measures can be developed vaccines can be developed by identifying specific target genes or specific single nucleated polymorphism or target genes which is causing the pathology. Again the reconstruction of the transmission network can be done and based on that different outbreaks can be traced and the molecular mechanism to monitor the evolution of the pathogen over time can be studied.

So, this is the importance of RNA sequencing in infectious disease diagnostics. Coming to the summary the RNA sequencing analysis basically involves multiple steps including read alignment transcript assembly then estimation of the expression and finally, differential analysis of each or different types of target population. Now, mapping RNA sequencing reads request splicing error aligners which are available which can read different types of splice junction. Then transcript assembly method can be based on

inferring transcript model where reference is used and where reference is not used that is Genovore construction. There are different estimation tool which can rather quantify how much a gene is expressed.

Differential gene expression analysis differential expression analysis where RNA sequencing is used to distinguish between two types of target population one is treated another is uncreated or one other is diseased another is the control one is the not infected one versus the infected one. So, simultaneously they can be studied. Then EQTL or expression quantitative trait loci which is basically important for the genetic variant identification. Then in infectious disease RNA sequencing is unparalleled because it provides the comprehensive insights into pathogen identification it helps in studying host response dynamics and it is also helpful in epidemiological study of infectious disease. So, this is all about the RNA sequencing and its implication in infectious disease diagnostics these are my references. Thank you and see you in the next class.