Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis

Prof. Arindam Ghosh

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

Indian Institute of Technology Kharagpur

Lecture 24 : Next Generation Sequencing (Part 2)

Hello everyone. Welcome back to your lecture series on Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis. We are in module 5 and today we are continuing with our discussion on next generation sequencing. This is a discussion on next generation sequencing, which we will be discussing in three parts. In the last class, we discussed what are the various steps and methods of next generation sequencing. We mainly discussed about Illumina sequencing by synthesis and we also briefly touched on comparative technologies like Oxford, Nanophore and Pacific Biosciences right.

So, we will be again discussing other varieties in our next class, but in today's class we will briefly review what we discussed in last class and then we will go into very important step that we did not discuss elaborately in our previous lecture that is library preparation for a next generation sequencing. We will be discussing both DNA and RNA library preparation methods. Those are mainly the samples that we load into the sequencer. Thereafter, we will also be discussing how to clean up those samples, how we can quantify those samples, what are the quality control measures for next generation sequencing samples and we will also be discussing the concept of whole genome and whole exome sequencing.

So, we have a lot to start. So, just briefly reviewing last day's lecture, you can see this is our sample of interest, this one. This DNA read in grey colour and we need to prepare this sample that is we need to attach something in the sequencing primer and then need to attach the adapter, these coloured sections. We need to attach on both the ends, so that these can be loaded into the sequencer the flow cell which has got many oligos, so that they can bind and then they can undergo the synthesis of the new strand and they can repeat the whole thing by breach PCR, you already know that. And then what happens since multiple copies thousands and thousands of copies of the same template are formed, they are read simultaneously using CCD, Fluorescent Chain Terminator chemistry and just focusing one single cluster in the left hand side depending on the colour of various fluorescent terminators and their corresponding bases that are attached

that we know of since each batch has got different colour, we can easily detect the sequence, alright.

So, this was the concept of sequencing by synthesis in Illumina platform. Now, one important step remains that you cannot I mean we cannot load the isolated DNA directly into the sequencer, it would not work. So, we need to prepare the sample that is we need to add those adapters and important things to both ends. So, how do we do that? So, today we will be discussing that step. So, adding adapter sequences to the desired molecules.

So, we are now discussing DNA library preparation mind it, ok. There are mainly two ways that are widely used number one we will discuss the TrueSeq style and as well as there is also another style that another method that is known as Next Terra or transposon based methods and then we will also discuss enrichment strategies how we can improve those sample for loading, right. So, in the TrueSeq style of DNA preparation what happens the first step is we have got an intact DNA big strand. We need to fragment that strand, we are we can do fragment by many methods for example, sonication with the help of ultrasonic sound or by bead shearing any method of choice we can break the DNA. Now, we should know that the breaking is actually not even there can be process where some region is extending I mean the two complementary bases do not exactly match because the fragments are random they are randomly done with the help of mechanical or chemical forces.

So, what we need to do? First we need to make sure the unequal fragments are cut in exact shape with the help of some chemical treatment. I am not going to those details because these concept is important what we are going to discuss. So, next when the both the strands are equally cut with the help of enzymatic methods we add one single A overhang to each base one single A overhang, ok. Next what happens we need to add a T overhang. We are adding a T overhang to the next adapter molecules ok the adapter molecule already has got T overhang and then these two will match ok.

So, now, these two are complementary sequences and they have matched and then this is actually ready, ok. This is ready to be loaded into the sequencer. However, if the starting amount of DNA was less what we can do? We can make copy of that using PCR we can add a primer which will prime of the adapter sequences and we can get multiple copies. So, in true 6 style we actually do not need PCR in order to load the adapter sequence, but if we need more copies then of course PCR can be employed. The next data style of DNA preparation we use transposase enzymes.

Now you I hope you know transposons are known as jumping genes and these transposase enzymes actually have got enzymes with adapter sequences they can actually

jump into they can accommodate themselves inside the DNA fragment and thus they can actually add adapter sequence and fragment them both in a single step, right. So, here prior fragmentation of DNA is not needed based on area of interest we can design the transposase enzyme. So, fragmentation and adapter primer sequence are added together. Now this is still not ready ok. What do we need? Now we need to amplify this.

Now this sequence, primum sequence can betargeted by another set of PCR primers which can now amplify and grow these colour sequences. So, now this is ready. So, mind it in next era style of library formation there is no PCR free method. In the first year only the first colours are added. It is only after one cycle of PCR when the primers target prime of these colour sequences that the next adapters can be added, ok.

So, now after learning those two method of library preparation let us discuss about some enrichment strategy. What are enrichment strategy? It may be so that when you are fragmenting the whole DNA we are only interested in some regions, ok. For example, in this whole grey DNA we are interested in only this pink regions. This pink regions might be the one that are actually coding for the messenger RNA and ultimately coding for the protein and this concept is actually the thing which iswhich we will highlight in later section during whole genome and whole exome sequencing, ok. So, that actually can be done intwo ways.

Number one suppose once the library is prepared we can do is what we can do is we can design some probes which are having the complementary sequences to these pinks, but not the grey ones, ok. So, once the library material is prepared on the sample is prepared we have added the primers and adapters sequences we can put these probes which has got a biotin handle, ok. And then we can extract or pull using those handles using beads or enzymatic methods we can give some washes and finally, we will have the our desired library material of interest. We can also design another method by which we can design specific primers which are actually complementary to only the pink strands, but not the grey ones, ok. These primers have got priming sequence, priming sequences attached to them.

So, after one cycle we will have this one which is still not readyas a library material, but after two cycles which will use a second primer which is complementary to this priming sequence we will ultimately get the total library material, ok. So, these are actually the conceptual I mean the concept should be clear how we can employ enrichment strategies in order to only focus on our material of interest. Why this is important? Because it will save sequencing dollars very very very important since the sequencing process is much expensive if we put those materials into the sequencer which will have no interest or which might not give any important information then we are simply using money and resources are always limited. So next we will be changing our rhythm and we will be

discussing some RNA sequencing library preparation, right. Mainly we are interested in messenger RNA which comprises of only 3 percent and the rest is actually ribosomal RNA which consist of 97 percent, ok.

So, if we do not have the perfect strategy we might waste our money. So, here also we will apply those similar concepts. So, the strategy is first depletion and enrichment of the RNA samples, ok. Thereafter RNA will be converted to complementary DNA then we will add sequencing adapters, ok. And the methods are true seek style library preparation and there is also another method we will be discussing that is known as smart seek.

And there are many many other methods of RNA library preparation which we will not be discussing but know this there can be many methods but today we will be discussing the true seek and the smart seek methods. So, during enrichment we all know the messenger RNA has got a poly A tail, right. And we will use this property, we will use this property by which we can design some beads which has got oligo DT, right. And then they will capture only the messenger RNA. So, these beads will only be captured by the messenger RNA and then we can simply wash those beads and after this capture the rest of the sample which contains ribosome RNA can be washed out and only the enhanced sample the beads can be washed out and ultimately we will get purified messenger RNA which is ready for sequencing.

We can use another method in which we can deplete. So, this is depletion or enrichment. We are first one is enrichment where we are only selecting the mRNA, messenger RNA. In the second method we will let go of the rRNA so that only mRNA will remain. So, how we can do that? Again we can design some complementary probes to the rRNA and again these probes have got a biotin handle and we can pull using those biotin handles by using multiple enzymatic methods for example using streptavidin.

And then with some washes those rRNA will be left gone from the sample and the final elute will only have messenger RNA, alright. So next, so this is the enriched mRNA, a big strand. We need to fragment the mRNA, ok. So mRNA is fragmented then with the help of primers you know we have already discussed how RNA can be converted to complementary DNA using reverse transcriptase enzyme. There can be oligo-dity primers and you already know the drill.

So using reverse transcription a single strand of complementary DNA synthesized and using second strand synthesis the phenomena by which DNA polymerase will use the first strand as a complementary strand to form a double stranded complementary DNA, so dsDNA. And now it is ready to be sequenced. Then we can actually apply one of those methods that we will use for example, true-seq method in which we can add one A overhang and then we can add this adapter molecule which has got one P overhang and

ultimately they will form the library material. And if we want more we can always use this adapter sequences as primer complementary bases and we can prime those sequences and we can have multiple copies of these library material which needs to be sequenced.

So this was true-seq method. So what happens in smart-seq method? In smart-seq method we use primers which have got oligo-dity which will target this poly A tail of mRNA. And we will use a special type of RNA dependent DNA polymerase which has got template switching activity. So what is template switching? Now see when targeting this poly A tail the DNA, RNA using I mean using this complementary strand of RNA a new sequence of DNA will be synthesized and that polymerase has got a special property by which it can add some C bases at the end, ok. So in the next cycle what will happen? This C bases you can see in the first cycle the C base has been attached, ok. So in the next cycle what is happening? The this Cs the enzyme will now add complementary Gs to this C which is at the end of the complementary DNA, right.

So when ultimately we have the C DNA after two cycles we have got a C DNA which has got poly A tail, ok, the primer sequence as well as handles on both the ends one A handle and one C handle, right. So now we can actually make multiple copies of this complementary DNA using PCR and thus we can have DNA library. Mind it again I am explaining if it is not clear a primer which is using this poly A sequence, ok, we have primer has oligo dt. Now we will synthesize a sequence and we will design the polymerase in such a way so that at the end it will add some C bases on its own, ok. Next what happens? In the next cycle the polymerase will add complementary G, ok.

So after two cycles we have got known handles on both the sides of the DNA. So in this sequence we will have a poly T and poly C and in the complementary sequence we have poly G and poly A, ok. So like that we can make sequences with known adapters and once we have known adapters we can always design the oligos in the flow cell so that they can bind with those adapters. Next we will be discussing pooling of samples using barcodes and indices, ok. So what is this barcoding? You know always when we are discussing or we are uttering the term barcode it means adding some unique property to a sample by which we can sort them later.

So let it be done once we need sampling we can just enter the barcode and we can get information about all the samples all the batches of the individual sample like that. So what happens over here? In this DNA, right, we are inserting a reading primer, ok. We know the DNA adapt library sequences, we know how to prepare this DNA, right, TrueSeq, NextEra, right. So what we do? We add a read primer, ok. This is known as I5 primer at one end, ok.

And at the other end we add another primer whose sequences are known, ok, that is known as I7 index, ok, this one, ok. Mind it, this is done by adding a complementary sequence to the known sequence of interest, ok, or the known sequencing primers, ok. Now, so this is normal phenomena. So how this is already happening in Illumina extension sequencing, but how we can differentiate our pool samples. So in order to separate samples from different origin what is done actually the two different primers for two different samples are actually reversed.

For example, in sample number 1 we can add the I5 primer on top and I7 primer in the bottom. The names are like that, ok. And in the another sample we can add the I7 primer on top and I5 primer at the bottom, right. So these are inserted in addition to the DNA library sequencing and adapter primers, ok, adapter sequences. So these are added specially so that these sequences can be captured later on.

Now this whole thing goes into the sequencer and we have a big read sequences of both the sequence of interest as well as these priming sequences, this information is stored. However, this read primer of I5 and I7 will be unique for this sequence, however this will be reversed if it is another sample, ok. So for example we have got two samples, one is treated and one is untreated, one is from control and one is from for example one is control sample and one is treated sample, right. So one is diagnostic sample and one is normal control patient. So all of them can be sequenced at once by inserting just a barcode and we can do it for many samples, right.

So next when we are sampling them we can just get hold of the barcode sequences and then we can pool or segregate all the samples just from sequencing data. So that we do not have to do a sequencing all over again with once with control sample, once with treated sample so on and so forth, it saves much time, right. Next we will be discussing sample cleanups, ok. So what do we mean by a sample cleanup? You see this sample cleanup is actually done so that we do not have unwanted sequences in our sample, ok. Number one was sample enrichment that is also true, but in spite of that there might be some unknown sequences or databases which may contaminate our result or which may give some false positive reading.

So how it is done? It is usually done with the help of magnetic beads and it actually I mean though this method is known as amperus tri that is solid phase reversible immobilization. It is actually the concept the name might be very intimidating, but the concept is very easy. It is based on the fact that nucleic acids can be precipitated by treating with a solution of polyethylene glycol that is PEG and salt that is sodium chloride, right. So what is the advantages? This bead based cleanup actually do not require columnar centrifuge, ok. They can be isolated using I mean they can isolate certain sizes of nucleic acid even prior to the sequencing.

So we do not need gels, right. We can do it. We can easily separate the samples that are known fragment using gel I mean cutting of the gel and treating extracting from the gel that is cumbersome which is much cleaner. So this is based on a finding which was actually discovered in the 70s by scientists which a group of researchers which who showed that using some specific concentration of polyethylene glycol different I mean a specific length of DNA fragment can be precipitated. So here you can see the experiments that they did is from the paper. So 15 percent polyethylene glycol is actually precipitating the entire sequence when we are I mean entire fragment across all ranges. When we are decreasing the concentration suppose a fragment sample has got multiple fragments of DNA mixed.

They treated the same sample to different concentration of polyethylene glycol. So the least concentration of polyethylene glycol is only precipitating the higher molecular weights. But the higher concentration of polyethylene glycol is separating entire fragment almost the entire fragment of DNA, right. So they use this principle to separate what they did? They first use a lower concentration of polyethylene glycol, ok.

So a large segment of DNA was precipitated. They took up the supernatant and they increased the concentration of polyethylene glycol. Then some bigger I mean other than the most big segment the remaining biggest segment was precipitated. So they went on so on and so forth to increase the concentration. Ultimately they showed that by increasing the concentration using serial treatment they were able to separate different strands of DNA. This is a very good observation, but mind it this method actually needed centrifugation because every step they still need to centrifuge, take up the supernatant and again treatment, keep it centrifuge, right.

But if you are using this same principle, but we are implementing magnetic beads then the process is much more streamlined. So a sample a beaded solution, magnet solution of magnetic bead is added which has got a and we add a known concentration of polyethylene glycol, right which will only precipitate the bigger segments for example, right. So once they are precipitated they are captured using the beads, right and when we are removing the supernatant. So the the segment that are precipitated might be the ones that are of interest, might be the ones that are not of interest. So depending on your need and what size of sequence you need or what size of the element you need for sequencing you can design it in such a way so that our sequence of interest may be precipitated.

Then we can capture that with the help of a bead, we can remove the supernatant and then we can simply wash the beads with ethanol and then dry and elute them in water. So that one then again the precipitated DNA from the beads can come up in our solution, ok you get my point. So we are using the principle of precipitation, but in order to capture

the beads we do not need to use the centrifuge because magnetic beads can easily separate. Now see when we are placing the magnetic beads, ok the magnetic beads can be placed in a base where the beads will get separated, ok and the other I mean the the liquid I mean the elute after removing the supernatant can be treated with water. So basically the thing is when why we do supernatant, we do centrifugation so that the beads will be collected or there will be a pellet at the bottom of the tube, right.

So when we pull out I mean when we try to pour the sample there will be we can make sure that the material from the pellet is not driven out or not thrown out in the supernatant extraction. What magnetic beads done? After addition of magnetic beads if we place them in a magnetic base they will catch hold of the beads which has got our sequence of interest. So even if we remove the supernatant we can make the tube fully dry, right the beads will still be hold on held on to the magnetic bases, ok and then you can remove the base and we can add the water and I mean you can say even after elution we can elute it with water and then what will happen the magnetic beads will be held on to the surface, ok and then only after water washing the DNAs from the beads will be captured in the liquid you can collect the liquid and then we can remove the magnetic beads. So you get my point very important concept alternative to centrifugation use of magnetic beads, you add magnetic beads make sure the magnetic beads are held in the base, remove the supernatant then the magnetic beads are still holding the sample we elute them with water, remove the elute which has got now the dissolved fragments of interest but the magnetic beads are still held on to the magnetic base. So how we do quantification of sample and quality check? It is very important the sequencing is as I told it can be expensive and it is very important to check the library.

So mainly it is done by checking the size distribution of the sample there are multiple commercial platforms that are available for example, Agilent Bioanalyzer, Tape Station, Fragment Analyzer there is one equipment from QIA Genre and QIAxel there is also another commonly used instrument from Parkin-Elmer that is known as Labchip, ok. And mainly the quantification methods are even pooling using either quantitative real time PCR or digital PCR. There are also fluorometric approaches like Qubit and Pigogin. We will not be discussing details of this method but it is important for you to know these methods so that mainly for MCQ purposes and also for your knowledge. Now sample quality control, how do we do quality control? It is very important prior to sequencing you will test your libraries to see if they are accurate.

Like any quality control the plot is done where you we are plotting the fluorescence mainly the fluorescence of the sample is checked across multiple wavelengths, ok. So, fluorescence and size of the bases are plotted, ok. So, generally a standards are put with known molecular weight and known fluorescence, ok. So, they will spike when in the sample, right. Now when they are mixed with the sample generally in case of true seek

style well prepared library will have this sort of distribution.

See actual library has got bases of multiple sizes, ok. So, it is not a single spike there can be this, but still if you think that it has got multiple non specific sizes you can again use these bit-bliss cleanup methods to narrow down the curve. One thing to note that in true seek style of DNA preparation since these adapters have got T over ranks they can actually adapt on to themselves, ok. So, generally in 130 base pair if you are getting a spike these are actually adapter dimers. So, we do not want that because they will also be sequenced we can use bit-bliss cleanup method prior to sequencing to remove these adapter dimers, ok. There can be other observations during primer for example, in primer contamination we get a smearing in the lower marker zone.

We need to again use bit-bliss cleanup methods. Again for next error library preparation the distribution actually looks like this it it is a wide plateau ok, but know this in case of next error sequencing this is actually ok we can use this for sequencing, ok. Next there is also another thing that you need to note for example, other than the product of interest we can get a second bubble product, ok. So, what is this bubble product? It mainly it is that we have over amplified the our sequences. For example, there is a bridge PCR that is done. So, what happens the now these products actually tries to bind to themselves, but you know this products are not actually complementary to each other.

However, only the adapters are actually complementary complementary. So, what will happen? They will adapt like this where only the adapters the terminal sequences of the adapters have actually bind bound and the in in between there is no binding, right. So, this is actually what happens in case of over amplification, but note this this is also actually ready to be sequenced, ok. So, this is the workflow summary of sample preparation we need to first isolate the DNA in an RNA sample that needs to be sequenced, then we add adapters using multiple molecular biology enzymes and then we do the quantification and check for libraries. Finally, the terminal part of our discussion that we already covered the concept that is whole genome versus whole exome sequencing.

So, naturally whole genome means sequencing the entire DNA, right and whole exome means encompassing and it actually includes both introns and exons. We already know what is the entire genome and whole exome sequencing mean targeting only the protein coding regions that are known as exons. For example, targeting this pink regions from the grey DNA. Now, what information do we get in from the whole genome sequencing we get a comprehensive analysis of all structural variation, copy numbers, variation, regulatory elements and therefore, when since we are getting all those information they can actually take complex disorders who have got non-coding regulatory elements, regulatory elements in the non-coding regions those rare disorders can be diagnosed,

right. And mainly they are suitable for identifying novel variants in non-coding regions.

Mainly when we need information about non-coding region we target we go for whole genome sequencing. Other than that we always use prefer whole exome sequencing because they are most focused on diagnosing any protein function disorder because proteins are coded by mRNA and they are only happening from the coding regions of the genome that are known as exon and they are used for identifying mutation responsible for Mendelian disorders and definitely it is much more cost effective alternative because it provides a focus solution whereas whole genome sequencing is much expensive. So, this is summary for today's class we have covered many things, we have discussed brief review, we have discussed DNA RNA library paper, proper library preparation properties for methods for all the processes, we have discussed how we enrich the samples, we have discussed sample quantification, QC as well as whole genome and whole exome sequencing. So, this is the reference for today's class and I thank you for your patient hearing.