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 23 : Next Generation Sequencing

Namaskar. Hello students, welcome back to the lecture series on Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis and in today's lecture we will be starting with next generation sequencing which will also be the topic of discussion in subsequent lectures. This is the part 1 of next generation sequencing where we will be recapping in brief about the traditional sequencing which you have encountered in the last lecture class. We will be discussing in details regarding the Illumina sequencing by synthesis is the major player in the dominant player in the NGS field. We will also be discussing some competitive technologies specially for long read sequencing method that is Oxford, Nanopore as well as specific bioscience sequencing methods and we will also be discussing in brief about the future prospect of sequencing ok. So, regarding traditional Sanger sequencing we discussed about what were the modifications.

So mind it the latest next generation sequencing catches up after the latest modification of traditional Sanger sequencing. So, last we in the last class we stopped our lecture by discussing the modification that was done to traditional Sanger method was using fluorescent chain terminators. And these fluorescent chain terminators can be incorporated randomly as billions of template undergo extensions and they are then read pass through a capillary electrophoresis and they can be interpreted using the chromatogram. So, this is traditional Sanger sequencing.

So, this is how it is done ok. So, this whole thing should be familiar to you. So, each chain has been terminated by a fluorescent chain terminator which has got a specific colour and using capillary electrophoresis we get a this type of chromatogram which has got different colours in different peaks. Now one thing we should we discussed regarding the traditional Sanger sequencing that Sanger sequencing led to the whole development of human genome to the phenomena of human genome project, but is actually the we can also give credit to human genome project that human genome project actually assured in the need for development of newer methods also to decrease the cost of sequencing. Now see human genome project was actually a 20 year method or 20 year

phenomena in which the entire almost the entire sequence of the human genome was processed you can see there was a huge gap and only recently almost the 100 percent of the sequence has been determined ok.

So, it was finished in 2001 ok and in from 2001 to almost 2022 there was a big gap this gap remained ok, but compared to when it started it got a big leap during the 1990s to 2000s due to advancement of multiple such technology right. We should keep in mind the human genome project actually used traditional Sanger sequencing method ok. So, this area we are talking about where it was finished in 2001 it used traditional Sanger sequencing. Next generation sequencing came after that and human genome project actually involved 3 billion US dollar to the 3 billion US dollar initiative. However, it I mean compared to the starting it now that it has reduced the cost of sequencing to 100 dollar per genome 100 million dollar per genome ok.

Now considering the throughput of traditional Sanger sequencing it can perform up to 384 sample and considering the latest machines that does that is based on Sanger sequencing there are basal 700 base pairs that can read it can capable of processing a maximum of 1 million genetic bases each day ok. So, you might think there is a lot ok Sanger sequencing machine the automated Sanger sequencing machines can do a lot, but if you consider the number I mean human genome consist of 6 billion bases. And the fact that you need to in order to remind the sequence of human genome you sample one genome multiple times almost 6 to 7 times if you calculate that then a single Sanger sequencing machine would take essentially 100 years to achieve this right, but actually human genome project was completed much earlier than 100 years this was done using factory of sequencer. So, this is one such factory where multiple such high throughput sequencing machines are running continuously 24 7 to achieve this goal ok. So, actually we need what was the need of so after human genome project the development still needed development was still needed because we saw human genome project was completed in 2001, but the technology still NGS came much later.

So, if it was goal was achieved what was the need the need being the genetic material used for human genome project came from very small people the sample size was small right. So, and most of the genome is still not understood. So, the thing is there are still a requirement to sequence thousands to millions of genomes to understand the function of genes and various genetic diseases right. If we need to make consensus sequence considering globally from all varieties of human beings it is very difficult to do it Sanger sequencing only ok. So, that is the need why next generation sequencing came to play.

Now we can see the cost of sequencing one single base 1 million genome actually sorry once million genome actually came from 10000 dollars to now it is actually 1 cent. This is the it is almost a 1 million fold decrease in the inflation you can say like that and there

are major dips you can see in 2007, in 2010, 2015 where the cost has reduced drastically why this is the NIH diagram ok why this is because of the advent of newer technologies that increase the throughput that reduce the cost of sequencing right. So, now the cost has reduced drastically compared to what it was when human genome project was completed ok. So, we will discuss all of that. Now you see the first and the major variety that all of us should know when discussing next generation sequencing are Illumina sequencer sorry Illumina sequencing by synthesis SBS and these are the equipments of Illumina sequencer that are available.

So, this is the mini instrument that is known as MI-Seq ok. This is the relatively high throughput instrument the middle one this is the Hi-Seq and this is the latest one that is that uses the maximum capacity that is the most powerful this is the NOVA-Seq equipment SEQ SEQ stands for sequencer right. Now if we compare the ability of these equipments you can see the reads in millions it is actually 30 million read is possible by MI-Seq instrument MI-Seq, Hi-Seq gives 3000 right and I mean these are millions ok and it is the NOVA-Seq actually gives 13000 million reads per day ok. So, you can see compared to Sanger sequencing which actually gives almost 400 reads ok and if you compare to the number of giga bases per day it gives 4 trillion ok compared to 100 ok. So, that is the leap and bound I mean leaps in technology which has happened over the years.

Now how this is possible? The Illumina sequencer actually dominates 90 percent of the sequencing images. So, whenever maybe in any lab sequencing machine is not available they can outsource mostly it is done by this Illumina sequencers ok it uses imaging based techniques they use the number of reads from many they they can range from millions to billions as I told you depending on the instrument the number of reads can widely vary the each read consist of 300 to 600 bases they are very very very accurate so almost 1 in 1000. So, 199.9 percent right and now the cost of sequencing an entire genome has been reduced to 1000 dollars and the process can be completed in 2 days. So, just imagine so the entire 23 billion dollar initiative of 20 years are now been drastically reduced.

So, you can actually visualize the leap in the technology ok. The basic architecture that is I mean that is the principle of Illumina sequencers are based on flow cells. So, what are flow cells? These are hollow glass slide with separate lanes through which all the reagents and template DNA are flown all right here the lanes have been shaded that you can understand and the cross section of a single lane has been shown over here ok. So, this is the reagent is flown from here and the reagent is flown out ok and this is a polyacrylamide coated surface in the interior where many oligos are attached ok. We will just get we will magnify all of this and we willsee what is happening in the molecular level in much details just wait.

Now comparing the 3 equipment that we just discussed this is theflow cell of a mycic equipment and we have placed itI mean it is being shown near a by comparing to standard append of tube which is which are routinely used in laboratories for centrifugation for storage of samples multiple things. And you can see it is the size almost comparable actually the append of tube is much bigger when we compare it to the flow cell of the high cic equipment it is much much much bigger right compared to the MIC and that is why that the power the space it can provide such many throughputs such many reads ok. And the novacic takes it to a all new level which is actually much much much bigger and by the virtue of that it can be 13 billion reads ok. Anyway so what exactly it is to be done I mean prior to the sequencing process. Now mind it the sample preparation will is actually a detailed procedure and will be there are multiple methods of sample preparation for NGS we will be discussing that in much detail in our next class.

So, for today you should note that the first step is to extract the genetic material it can be DNA if the genetic material is RNA it needs to be converted to DNA all right. So, this is the sequence of interest. So, next what needs to be done primers are primers needed needs to be attached the sequencing primers need to be attached to the two ends and to that capture sequences that are known as adapters are attached. So, these are the procedures that are that needs to be done prior to our sequence of interest being channeled to the sequencer ok. Again we will be discussing these in details.

So, what happens when we send our template of interest into a flow cell ok this template is captured by an oligo you can see I I hope you can appreciate there are two colours of oligos that are present in here green and orange colour. So, as we flow the DNA the templates will be attached to the end of the oligo ok this is attached by the hydrogen bonding. So, this template is first captured and then we flow in polymerases through as the see this is actually using the PCR method only ok. So, if this is the situation what this is the template strand and this is the primer that has been attached now if we flow in polymer as it will undergo amplification and a newly synthesized strand will be generated which is actually an add on to the oligos that is actually present over here ok. So, in the next step what happens this is a one cycle of PCR that happens with the help of the sequencing primers the nucleotides and the DNA polymers that we continuously flow through the flow cells.

Next what happens the old strand or the original strand original template is washed away using reagents and the newly synthesized strand bends over and it forms a bridge with another oligo ok it is designed in such a way that another oligo captures the other end ok and then it goes another cycle of PCR. So, it will amplify like this. So, this step is actually the bridge PCR. So, this is not a template strand this is the primer and then again another RNA I mean DNA polymerase will act and it will amplify. So, ultimately the

two strands will be generated ok one is now attached to the green and another is attached to the orange.

We can flow in specific reagents so that the and this process can actually occur many many many times it goes over and over again. So, it forms another bridge and another strand is synthesized and it forms over and over again. So, that after many cycles thousands of copies can be made mind it this all came from a single strand single template. So, they will have same strand same sequence ok. Now we can flow in specific reagents.

So, after they are opened we can flow in specific reagents so that we can cleave the bonds that are attached to any specific colour. For example, now we have freed the green colour green oligos have been freed and we have got only source template strand that have been amplified and that are only attached to the orange oligos that are already present in the flow cells ok they are engineered in such a way. Next they are bound to a sequencing primer. So, this is our template that means to be sequenced and we bind a sequencing primer. So, initially oligos are captured on the flow cell multiple sequence of bridge PCR ok this this is the bridge PCR that is happening this also term this bridge PCR.

So, it is happening we get multiple such sequence of interest and we now add a sequencing primer. So, all of this are actually done in the sequencer machine prior to the actual start of the sequencing. So, you can see these are the components of a sequencing machine. So, example incorporation buffer there is a reagent card tree there are pooled mixture with multiple DNA DNTP libraries there are flow cells. So, actually you can see this is the flow cell if you can pause the video you can visualise this is the flow cell ok that is actually placed over here we will place it over here ok and this is the manifold ok we can actually close it just like this.

Over here you can see these are multiple pipes we can place this pipe on top of here ok and connect these pipes with reagents you can see now the from last step all things have been put into place the flow cell has been the flow cell cluster have been placed over here and the reagents over here these are the samples and these are the reagents everything is already placed we need to just cover and operate via touch screen. So, that all of these processes that I just discussed are done automatically all the bridge PCR depending on time. So, all the polymers is the DNTP etcetera over here. So, based on these tubes there are multiple pumps which aspirates the reagent and flows them through the flow cell they are float and they are collected over here to the wash drain and this all happens and this is to be done prior to the actual sequencing process ok. So, now the flow cell clusters and then moved to the sequencer ok you see for clustering we are using a machine also from Illumina as you know C-bot ok specifically over here C-bot the

model of the machine.

So, now, the as is shown in the picture this flow cell cluster from the C-bot this is a C-bot machine are transferred to the high seek equipment ok. And the high seek equipment in this case we can place the flow cell cluster over here we there are multiple reagents over here these are the refrigerated components and this is the touch screen through which the we can give I mean specific commands and this sequencer is actually that is not shown here that is also attached to a microscope. So, for direct microscopy and visualization ok. So, this is how the hands on thing is done it is actually not that much possible to show you the entire thing laboratory tour in this course, but you can have some idea, but we are most focusing on the theoretical part of the generation sequencing and it is much easier to understand just by seeing the machines that you can correlate how things are happening in the hardware side of the things ok. So, a sequencing primer is now bound to the template.

So, what they will do? Now, compared to traditional Sanger sequencing the next generation sequencing Illumina sequencing by synthesis sequencing by synthesis means again templates I mean a complementary standard will be synthesized and simultaneously sequenced. So, sequencing by synthesis. So, they are also using fluorescent terminator, but a big difference is reversible. So, what do we mean reversible? Suppose this is the terminator. So, when this suppose G is incorporated the reaction will be terminated and depending on the color which fluorescent I mean yellow for A, blue for G, green for T and red for C.

So, that is the traditional interpretation that we can understand till now, but how these molecules vary that they are reversible means after one reaction we can use a chemistry step we can treat it with some chemical. So, that the OH group is resumed ok and now it can. So, after the signal. So, it is terminated once ok we get the image then we add the chemistry step. So, that a 3 prime hydroxyl group is formed the next terminator can be added then we can again take a picture and so on and so forth.

So, once terminated in the next step reaction can continue. So, this is the beauty of next generation sequencing. So, you can see over here. So, once the one nucleotide is captured it is suppose the first when the image is taken it is A. Next again the chemistry step is done now the reaction can continue.

So, again the next time it will capture another sequence for another nucleotide for example, green we take another image determine ok this is T again this is treated with the chemistry the further extension can occur. So, the next base it incorporates is C. So, this way we can stop the reaction at multiple phases we can take an image after each addition is done and thus we can get the whole sequence with the help of multiple

images in easily in one go ok.

So, this. So, considering. So, these are the clusters ok. So, these are the clusters and we are focusing on the top left and the bottom right clusters. So, just by multiple phases of the camera I am phase of images after 5 cycles depending on the colour we can easily determine that the sequence was A G C C T in case of the top left and in case of the bottom right it is G T A A C ok. So, thus multiple colours using multiple such I mean four different colours and using reversible chain terminator chemistries billions of sequences are possible at a time and this gives the power to the machine ok.

I hope you understood the basic concepts. However, the thing is we also told you that the N G S machine is limited ok for long stretch I mean for up to 800 to 900 bases we actually preferred traditional Sanger method why not N G S? N G S is very good it is very fast, but the read length is limited to 300 bases. The reason being all these chemistry steps are the enzymes and the chemistries that are used are not 100 percent efficient it means some strands do lag behind. So, in one strand we can see it has it is lagging behind where in the other strand it has gone further ok. So, in one strand it is lagging behind one strand after three steps these are the efficient steps of imagine two nucleotides have already been added, but in the one chain even after two steps the two subsequent bases have not been added right. So, this error it is not that bad if you just look like this, but over time it is adds up and the signal actually moves quite far from the original signal and it becomes very difficult for the machine to I mean calculate what was the original signal using real time images.

So, based just due to the fact that these chemistries are not 100 percent efficient. So, we always try to limit the read to 300 bases because it has been found that up to 300 bases whatever lag length limit is there it is actually since we are using multiple sequences are studied at the same time it can be the error can be minimized ok. So, in Illumina sequencing my synthesis what happens one image is taken in each colour. So, one image in blue colour one image in green yellow and red ok. However, you can see the image actually looks like this is not not actually clear to I mean no dot is a perfect circle there is no signals cannot be separated from each other and it is actually very cumbersome to visualize how it looks like it specially it is true if suppose two sequences adjacent have only the same colour for example, over here G.

So, it basically looks like one single blob and the machine cannot separate. However, in the next run it is very clear that one is blue and one is yellow. So, it means one is A and one is G and subsequent in the subsequent cycle the machine can actually highlight the special I mean distinct cluster of emission spectra which is actually determining the two different bases ok. So, the Illumina sequencing by synthesis technology might seem very easy, but the thing is not easy I mean might seem very error proof because we have

got four distinct colours, but in reality the emission spectra is actually quite overlapping. You see red is well separated, but the blue yellow and green emission spectra is a very overlapping.

So, the machines in order to differentiate the machine has to undergo much colour compensation colour compensation which actually lowers the accuracy of the ribs ok. So, there was a need for further development. So, they I mean Illumina actually came up with two channel chemistry red and green. So, how this two colour sequencing works? Two colour mind it two colour sequencing are used to work I mean analyze four bases. So, how is it possible? Let us see they are using green for T.

So, when we are using green camera T is visible when using red camera C is visible, but when using A it is visible in both the cameras both the frames and G it is actually not visible in any frame ok. So, this is the differentiator, but how it is achieved? You see the logic is very simple. So, using two colours we are actually visualizing four bases. So, if both it is available in both the channels it is an A, if it is none in both if we are in dark in both the it is G, single in green light it is T and in red light it is C. So, naturally if the machines are cheaper the reagents are cheaper because they have to measure only two colours compared to four colours right.

So, how far is it possible? What do you think? Is it possible to determine the sequence of four bases using single colour? Yes the answer is yes. So, single colour chemistry equipment have been developed now they are using only green colour. Why? Again the reason being when using multiple colours the colour compensation the emission spectrum might be overlapping. So, in that regard two colour chemistry is much better compared to four colour chemistry and theoretically single colour chemistry should be further better right. So, what is happening over here? You can see both A and T are visible using green light, but C and G are not visible using green light.

So, they are taking two images and in between the two images there is one chemistry step. So, you see A has got the green fluorescent chemistry I mean fluorescent dye that is attached using a cleavable bond right and C has got a available bond where something else can attach the fluorescent dye can attach. So, on treatment with the chemistry step what happens this green gets detached from and is attached to C right no changes happen in T and G. So, in the second image A will not show any signal, C will show green signal, T will continues to be detected and there will be no detection of G. So, when you are using a single colour based on what bases are visible I mean what bases are actually dots are visible using the green colour using the two images we can compare when one is on and one is off and both are on when both are off the four different sequences ok.

Now again these machines are much cheaper because we need to utilize only one

colour, but initially the accuracies of these machines were not high compared to the four colour chemistry when two colour chemistry came out, but with modern development these two and single colour chemistry machines they are developed by Illumina only they are actually rivalling four colour chemistry and hence they are much more preferred because the cost of sequencing has gone much lower compared to four colour chemistry when compared to two and two colour chemistry when compared to one colour chemistry alright. So, now we will briefly discuss the methods of long read sequencing the competitive. So, we have discussed about Illumina sequencing by synthesis. The next competitive technologies oxford nano four technology where they use a lipid by membrane there is a nano pore ok nano pore which is 1.8 nano meter and which has got a specific electric current ok over here and through the nano pore the multiple strands of DNA are traded ok and now each base have got specific charges and when each base actually flows through the nano pore the change of current is actually detected and it is plotted over here ok and thus depending on the amount of current we will get different peak.

So, one high peak for A medium peak for C a bit higher peak for C and low peak for G. So, this is a standardised thing that they have come up with and in reality you know this does not look very clear compared to I mean as it is appearing over here because in here there are six bases at a time in the pore and with four possibilities per base there is actually a 4000 possibilities that can happen in the pore, but with modern technologies they have developed yes there are errors actually the error rates are actually quite high. So, 10 to 15 percent and there are biased errors ok this is a trend of errors if error starts to happen it will have a bias it is not a good thing. However, they can give us really long reads which are very very very useful in certain user case scenarios which we will be discussing very soon and they can also directly sequence RNA you can just thread RNA molecules into that and into the nano pore it will be detected and I mean oxford nano pore technology are promised maybe protein sequencing can also be done using this nano pore technology in future which is a very good thing on its own right. The important advantage is very portable ok see the device is very small you can see the scientist here using the device is she is in the wild I mean wherever I mean in the periphery in the field and this machine is actually powered by the computer and they just need to put one drop of sample and within few I mean short amount of time the entire sequence will be done very very very innovative.

And lastly we will discuss specific bio science sequencing technology over here which actually is very similar to the Sanger sequencing or the Illumina sequencing over here actually see Illumina sequencing used fluorescent chain terminator chemistry over here. Pacific bio sciences the thing they use they are using a fluorescent group that is attached to the phosphate group there is a long special type of nucleotide that they have designed they use that is multiple poly phosphate group and at the end of phosphate group of

fluorescent chain terminator is I mean a fluorescent dye is attached. So what happens there are this is a very small glass pore ok which has got a 100 nanometer thick I mean girth or the thickness through which there is a I mean where DNA polymerase is there which has got a template strand that is needs to be sequenced and we have got these multiple nucleotides which have got a long phosphate tail and a fluorescent dye that is attached to it ok. Now what will happen all of these will be continuously flowing through this pore ok and there is a camera which can detect fluorescence ok. Now when multiple nucleotides are flowing together it will give a baseline color like this jitteriness, but when a actual complementary nucleotide will come and attach with the help of polymerase it will be extended using the template strand what will happen it will stay there for a while and after hence it will give a spike of pulse because it will be staying there right and after the reaction is done the phosphate group will be cleaved and then again the signal will come down ok.

So this is the beauty so when suppose this shows a blue color then we can easily comment that the nucleotide that was added was G and thus this type of technology is used over a long stretch of nucleotide to be sequenced and we can get the target sequence this is the proprietary technology that are used by Pacific Biosciences. So lot of labs are also now buying these equipments specially for long read methods. Now this thing is actually used for 100 kilobase which is not as high as the oxford bioscience oxford bioscience give to millions but is much higher compared to sanger, sanger can give maximum 900 and illuminia can give 300 bases this is 100 kilobase much higher and the error rates are higher I agree 10 to 15 percent may be with newer iterations the rates will decrease but the error rates are random that is a good thing it means if we need to detect a mutation we can run it over and over again to minimize all the random errors. So for example see this is a sequence ok this is a long sequence ok which are I mean two sequences ok. So complementary sequence is done and they add a circular adapter like this in the two ends and then the whole target is sequenced over and over again.

So there might be errors that are random errors or there might be mismatch that is actually due to mutation. Now since this machine only shows random error so if we sequence a gene over and over again and if the error occurs in the same place over and over again in all the consensus then we can easily conclude that this is actually the mutation ok whereas there may be errors that are possible in other regions but since they are nullified in other cycles those were random errors. So this is very very very important. So yes though there are downsides of long reads they are harder to prepare they costs more but the benefits are they are easier to assemble. I mean imagine a jigsaw puzzle the one you have got a big stretch of DNA and you can have 1000 breaks and you need to arrange them 1000 breaks means the analogy of the Sanger sequencing whereas if the big strand of DNA is only broken in three places you can easily I mean if you can read higher sequence you can easily reassemble the target sequence ok.

So it is very important to identify structural variations as well for example if there are some structural variation in gene by specifically targeting a small sequence it is very difficult to understand what are the structural variation. For example in multiple mutation in cancer there are multiple translocation for example in various cancer where retrovirus can lead to change of sequence over a long region right. Hence in those areas if we used small I mean traditional or next generation sequencing which can read only three bases it is very difficult to visualize the entire thing wherever if we use specific biosciences or oxford nanofold technology which can visualize long regions we can easily visualize what are the structural variation. And next we can also determine phase variation for example we know for example maternal mitochondrial DNA. So, one mitochondrial DNA is obtained from our mother there are also specific type of chromosome in which one chromosome is obtained from mother and one chromosome is obtained from father.

So, all these type of things so when using smaller reads it is very difficult to visualize what chromosome and what is the source whereas long reads will easily help us determine what are the chromosomal variants that have been inherited both in case of autosomal as well as in case of mitochondrial DNA ok. So lastly I would like to conclude that there are multiple medical application of NGS which we will be actually discussing in our subsequent modules but just to give you an idea earlier prenatal testing ok it used to be done by the invasive amniocentesis method. But now it has been largely replaced by next generation sequencing we just need to simply collect DNA blood from mother because the fetal DNA is also circulated in the mother's blood ok. Like that liquid biopsy etcetera everything will be discussed transplant rejection what specific genes are leading to transplant rejection can be rejected by next generation sequencing cancer detection detection of multiple pathogens as well as determination of treatment of cancers were what are the various regions that are responding to any cancer treatment what are the changes of guardian genome mutation etcetera everything can be easily determined using next generation sequencing. So, what are the prospects you can see the cost has been reduced to 100 dollar compared to it was 1000 dollar but what are the consequences maybe with the cost going lesser and lesser maybe very soon the genetic sequencing genome sequencing will be a part of routine examination where just like parameters like blood glucose you can create in it maybe genome sequencing will be available for all patients.

But again how will we be utilizing this data and there are multiple safeguards and ethical consideration what are the noble formulation what are the personalized medicine applications everything will be exploring what are the possibilities in our later parts of the lecture series. So, with that I conclude that in summary we discussed about recap of traditional sequencing we discussed the Illumina sequencing by synthesis we discussed

the long read sequencing and the competitive Oxford and Pacific Biosciences technology as well as we also discussed in brief about the future prospects of sequencing. So, these are my references and I thank you for your patient hearing.