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 25 : Next Generation Sequencing (Part 3)

Namaskar. Welcome back to your lecture series on Comprehensive Molecular Diagnostics. Today we will be discussing next generation sequencing. We are continuing with our discussion on next generation sequencing. This is the third part of NGS where we will be covering few other varieties of next generation sequencing that we have not discussed in earlier two lessons we have covered a lot. We will be there after for example, pyro sequencing, ion torrent sequencing, single molecule real time sequencing.

We will be next followed we will be discussing the importance of bioinformatics in NGS data analysis. We will also be covering shotgun sequencing, we will be discussing about the human genome project and lastly we will be discussing the role of I mean we will compare Sanger sequencing with next generation sequencing, ok. So, strap yourselves now to start with pyro sequencing. So, pyro sequencing is based on principle where you see let us forget pyro sequencing just focus on formation of a phosphodiester bond with the DNTP this is the big chain and DNA polymerase is adding a new DNTP.

So, this 3 prime hydroxyl bond is attacking or attaching with the 5 prime phosphate bond of an incoming DNTP this is a DNTP, ok. So, when this happens a pyrophosphate molecule is released, ok and when this pyrophosphate molecule is released this pyrophosphate molecule can be treated with some substrate in order to produce light and that light will be detected by a computer in order to detect the sequence. Now, you might be asking well any DNTP will be producing light pyrophosphate because this reaction is irrespective of this base it can be adenine guanine thymine or cytosine, right you are absolutely right. So, how we can know the sequencing? Well we can know the sequencing if we know what base is actually being added by the polymerase, right. So, with this very basic concept let us move forward.

So, in pyro sequencing basically the steps are these it can be broadly divided into 4 steps. The 3 basic steps were first library preparation are done you already know how DNA libraries are prepared. Next is digital PCR, droplet PCR, emulsion PCR that we have already discussed, next loading of samples and the proper pyro sequencing

reaction. So, in preparation of libraries that we discussed in the last class we are adding some primers, priming sequences and adapters, okay. Next step is digital PCR.

Now this is a familiar slide from our variations of PCR lecture where we already discussed about digital PCR where what happens there are multiple emulsions, small droplets and each droplet. So, you see since we are adding oil in water and oil in water actually forms emulsion this is also known as droplet PCR or emulsion PCR, right. And since the reading is done digital it is also known as digital PCR. So, everything all the ingredients of PCR are already added, right. And what happens on the droplet on the single droplet we we design the droplet in such a way that it has got multiple primer sequences on its surface.

So, our cap region of interest can bind, okay. And then it can undergo multiplication it is much similar to that what happens in NGS on the flow cell, okay. So, the first strand binds it forms another complementary sequence then again this complementary sequence can bite to another priming oligo on the surface and then this cycle goes on repeating so that a multiple copies of the same target DNA is formed. Now everything is inside the droplet inside the ampoule, ok. Now each of these beads now the next is the loading step.

So, each of these are now loaded onto the pyrosequencing chip. So, there are so each actually these micro slots can add I mean each of the content can be added in loaded in this micro slots, right. And we can now flow them or we can load them or flood them with known dNTPs in sequence, ok. Now each of these reaction slot have got multiple components. For example, these are the ones sulfurylase, luciferase, apyrase, APS that is adenosine, phosphosulfate, luciferin, DNA polymerase, primer everything, ok.

So, now what happens? Suppose we are first flowing A or we are first flowing P or we are first flowing C or we are starting with G. So, now each and every dNTP will be reacting with them, they will produce a light and that can be detected. If it is not then will flow in the other dNTP. So, this is what is exactly happening. So, first we flow in dATP.

If the reaction is not happening, if the polymerization has not occurred, this dATP will be removed from the reaction using the enzyme apyrase. Apyrase is an enzyme that can convert dNTP to triphosphate to monophosphate, ok. Then we move to dTTP, ok. Again in dTTP we see the reaction the it is not complementary therefore, triphosphate is not generated, ok. Next we give dGT.

So, dTTP is again treated by apyrase and it is removed. Next we treat with dGTP. Now what happens? If this g dGTP is complementary, if this is actually added reacted upon by DNA polymerase, what will happen? There will be a spark of light, not a spark, there

will be a production of light and that will be detected by the computer. So, what is happening? See, so for the first three cycles, ok let us give an example. So, for the first three cycles we have actually treated with A, D and G, no reaction has happened, right.

Now the base which has got a complementary G, we need a C for that reaction to happen, ok. So, now since we have added C, it is actually complementary to G, there will be a light detection, right. So, what is actually happening? Since C is being added, pyrophosphate is being liberated. This pyrophosphate is actually acted upon by APS and with the help of sulfurylase enzyme to form ATP. And this ATP is then acted upon by luciferin from luciferase and which actually breaks down to be oxyluciferin in light, the same mechanism that is used by fireflies.

Fireflies produce a lot of ATP and that thus they emit light, ok. So, the now since we can control what bases are going in, each of the reaction can be easily tracked. There are cameras, we can actually track this light, there it is actually run by luminometer and we can easily determine the sequence. This is the principle of pyro sequencing, ok. Next, we will be discussing ion torrent sequencing, very much similar concept initially again provided by the fact that when a dNTP is added to the base,pyrophosphate molecule isI mean let out along with a proton.

In previous pyro sequencing, we are only targeting using the reaction mechanism with this pyrophosphate, but here we will be targeting this H plus ion. So, how it does? So, a change in pH because if the reaction happens, there will be an emission of H plus ion. Again that change in pH can be detected and if we know what dNTP is flowing in, we can now easily detect the sequence. Again after this is carried out on a semiconductor chip, but mind this pyro sequencing, this can be done in a flow cell. This is not I mean this one actually is a wet reaction.

There are multiple enzymes that isacting here, ok this platform. Whereas, the ion torrent sequencing is loaded on a semiconductor chip, this is much dry and there is much more electronics that is involved. So, the first two steps are exactly the same. We will prepare library, we will use droplet PCR and then we will load the samples on the semiconductor chip. Next, we will flow in the dNTPs.

It is much similar the semiconductor chip and next the change in proton will be detected. So, it will give out a signal. So, this is actually the magnified view where there is an ion sensing plate, ok here at the lower layer. So, whenever there is an addition of a proton, there will be a signal that is detected by the galvanometer, ok. One question, I will definitely expect you to answer me in the live session what is the full form of ISFET, ok.

So, I will give you some brainstormingwork for you to do that is a part of your homework. So, when we are using the sensor plate, this IS first part is ion sensing, ion sensing, ok. Anyway, so this will be detected and thus based on what dNTP is flowing, we can actually detect the signal. Now one thing each and every I mean reaction I mean each and every dNTP when they are flowing, if they are not participating in the reaction, they are automatically washed out of the sample and only the complementary base will trigger the ion and it will be sensed by the galvanometer, it is electronic signaling, ok. The earlier one was in pyro sequence, it was light detection by luminometer.

Anyway, so it when the complementary sequences are added, we get both protons and we get the pyrophosphates and attachment of these protons actually gives the signal and they can be sensed. Now, what happens if successively two same nucleotides are added, then there will be a double p, ok. So, this is actually a standardization which is actually done at the I mean during the calibration of the machine, so that exact so whenever we get a spike, we already know, ok. But mind it, it is not like the pacific biosciences phenomena which we will be discussing very soon. We will be again recalling that where each and every different nucleotide was giving a different signal.

Here, the signal intensity is the same because the reaction is same. However, since we know which dNTP was flowing in, we can actually backtrack and calculate the sequence. Just one thing to note, if you get a double peak, it means two same nucleotides have been added and they were actually complementary. For example, the strand has got 2A. So, naturally two Cs will be added and they will give two I mean a big spike, right.

So, next we will discuss single molecule real time sequencing. Now, this should be very familiar. Yes, I have already told you this is basically the same phenomena as pacific biosciences, I mean same technology that is being used by pacific biosciences where we are adding the circular or hairpin adapter at both the ends and then polymerase is actually tracking the whole sequence and this amplifying, right. So, one whole amplification is known as one pass, right. You remember this diagram where there was a and how it is being detected? Then they are using fluorescence, ok, coloured fluorescence which is unique for each and every dNTP and the whole dNTP is actually flowing around and whenever the polymerization is occurring, the fluorescence molecule is actually staying there which is attached at the end of the pyrophosphates.

We can you can recall from the very first class of next generation sequencing lecture and thus with time, since there might be a background intensity right, but whenever the polymerase is coming and attaching for example, if it is green then we recall I mean we can corrobert that is A that is being added, red for G, blue for C, orange for T so on and so forth depending on the colour of the fluorescent terminator, ok. Now, the beauty of it is whenever and now for the whole pass, this is the whole read, ok and based on that a

consensus circular sequence is determined. We already discussed this in the Pacific Biosciences technology and this is very important to detect mutation, this is just a recap of previous, but one thing to note that technology is actually known as SMRT or Single Molecule Real Time Sequencing, it is also a variety of NGS. So, if this comes mind it, this is nothing but the technology that is used by Pacific Biosciences. So, if a mutation is being consistent in all the reads, it means that is there whereas, I already discussed in that class this method has got some random error 10 to 15 percent, but those random errors can be nullified using multiple consensus sequences, ok.

So, next we will be discussing the role of bioinformatics in next generation sequencing. So, what does bioinformatics do? Mind it, if we need to strictly answer this question, what are the steps of next generation sequencing, ok. Then the first step will be library preparation, the second step is sequencing proper, I mean the sequencing reaction, ok and the third step is actually analysis of the data using bioinformatics. So, without this we are going nowhere, ok. So, what does bioinformatics do? It is a final step of next generation sequencing workflow, it involves processing, analysing of large scale data, enabling variant identification, everything basically we are getting a whole lot of sequencing data.

You can see this is the, for example, if this is the last step of Illumina sequencing where we are actually getting multi reading multiple samples or multiple sequence of the same cluster, if there is a lagging, one step is lagging behind, we discussed why we do choose short read sequences. It is actually the bioinformatics which is actually trying to get hold of this overlapping sequences and thus they can finally match with an existing sequence which will give the final result. So, where from we got the existing sequence, we got it from the Human Genome Project, ok. So, this is the importance. I will again be discussing this in the later part.

So, this is about NGS, we are not discussing how the bioinformatics analysis is done because that demands a, that warrants a course of its own and basically there are multiple modules, there are multiple programs in which you can just feed in these big millions and millions of sequencing data and it takes a lot of time. Mind it, the library preparation, the sequencing, reaction, they do not take up much time, but the most important and the cost daunting task is the bioinformatics analysis and that actually gives us the correct information. Next, we will be discussing shotgun sequencing, ok. Now, so we are actually technically done with next generation sequencing at this point, ok. So, we are discussing some allied concepts.

So, shotgun sequencing is actually a high throughput DNA sequencing method that breaks DNA into random fragments and sequences them and then arranges the data to form a complete genome, right. Now, why shotgun sequencing? The analogy is like with

shotgun if you shoot, you do not have any control. So, they are, the DNA is fragmenting uncontrolled way with the help of any chemical, shearing force, for example, sonication or magnetic beads, the random, we cannot control where the break has happened, right. And now, we will sequence each of these randomly. We can use any sequencing method, for example, in the fragmentary short, we can use Illumina sequencing, we can use Sanger sequencing.

Earlier, before next generation sequencing, before Illumina sequencing was invented, we only had one option to do Sanger sequencing, right. So, suppose these three fragments are analyzed, ok. Next, what we need to do, so how to know which fragment comes after one another, right. So, when DNA, so multiple samples for same sample in multiple situations are randomly fragmented, we can easily match them by targeting the overlapping sequences, ok. So, you can see this fragment in one sample actually has got this part, which also has got a overlapping sequences in this part.

Another fragment has got an overlapping sequences. So, if we just try to align them, we can do easily do it by the help of computer software, again, precursor of bioinformatics and this is actually the bioinformatics. So, we will get the complete genome, ok. Now, this is how the human genome project was actually done, but there are some actually disadvantages. The main disadvantage is incomplete genome assembly.

Why? For example, since the fragment is, I mean the, yes, DNA fragmentation is uncontrolled, there might be loss of some sequence, ok. When the sequence information is actually lost, when the terminal sequences are not overlapping, right, then we will not be able to align the sequences, right. If we do not have overlapping sequences, there is no way of aligning. So, genome information may be lost, ok. And suppose there is one sequence that comes in repeats, ok.

So, one sample is repeated, again, I mean one sequence is repeated, again another sequence is repeated. So, if we even determine the sequences individually, it becomes very difficult to determine which sequence comes after what. So, specially for these, we need methods that actually can read the DNA in long stretch, right. Again, there are computational complexities, very important loss of special information, the exact thing regarding repeats, which sequence comes after where very difficult to understand. And the last thing changes in haplotype phasing, for example, which all these were actually discussed when we are comparing long read sequences, the advantage of long read sequences versus short read sequencing methods.

For example, one chromosome is coming from parent and one chromosome, I mean father, and one chromosome is coming from mother, but when fragments are uncontrolled, there is no way of determining in short read, for example, so many jigsaw

puzzles, which is derived from what chromosome, ok. So, it has got some shortcomings, but this was only this was the method that was only available during human genome project back in 2001. Hence, human genome project is actually incomplete, ok. So, let us discuss the human genome project in brief, what is human genome project? It was, ok, it is a historical event, very important event that has changed the face of genomics altogether, right. So, it was an international scientific research initiative to map and sequence the entire human genome, identifying and determining all the DNA base pairs in the human chromosomes, alright.

So, how it was done? Mind it, it was not easy to start with because human DNA contains 3 billion base pairs, it is about 10,000 genes, ok. And it started in 1984, I discussed very early as soon as there are two methods that were available with the scientist, one is Sanger sequencing and another is polymerase chain reaction, it was quite possible to map the entire genome, right. So, this idea came into, I mean it was initiated by the NIH, National Institute of Health and Department of Energy in USA in 1990, ok, where the project was under, I had already said it was 3 billion dollar project, right. And it was headed first by James Watson and later Francis Collins succeeded him, ok. Now, let us discuss the timeline, let us see how it actually happened, it is very interesting.

So, by 1990 the 2 percent of the entire genome was sequenced, ok. Mind it, Sanger sequencing was automated, ok. Applied Biosystems actually discovered a machine which can on based on capillary methods, ok. We discussed what machines were discovered, right.

If you remember the first NGS class, right. But the whole ball game changed, I mean till 1998 even only 6 percent was done, mind it in span of 8 years only 6 percent was done because the technological advancement has not taken place, the things were costly very much, right. So, from the first 8 years only 4 percent advancement and now you will see what magic happened, ok. In 1998 Applied Biosystems came out with the ABI PRISM 3 3700 that actually had 96 capillaries and it could actually detect so many samples, right. And they Applied Biosystems in the year 1998 they collaborated with TIGER that is the Institute of Genome Research in order to form Celera Diagnostic. It was a non-profit organization headed by Craig Venter and this Celera Diagnostics actually purchased 230 ABI PRISM 3700 and their aim was to sequence the entire genome faster than the human genome project.

So, what was their aim? Why? Why they were in such hurry? Because the non-profit organization actually aimed for profit, ok and their aim was to sale the sequence data also to patent the genes that were useful for disease treatment. So, mind it had they sucks I mean we will see. So, in this situation every time you need to diagnose genetic disorder you need to pay them money. So, there was a whole big potential to earn multi-billion

dollars with the advancement of science and technology, right. So, they decided to commercialize and it was controversial definitely.

So, a lot of scientists made a buzz in the community that this patenting of genes is not right, right. So, since then the race of public and private partnership not partnership race between private and public started for advancement of science and technology and ABI PRISM 3700 played a big role. Mind it the NIH was also forced to procure the ABI PRISM 3700. They were using older machines to sequence them that is why the progress was so slow and this played a huge role in sequencing because it could as I discussed it could sample read 96 samples in the span of less than two and half hours with read length of 800.

This thing can do the one sequencing in 15 minutes. So, with 1536 samples per day in 24 hours it actually reduced the cost of the base. We discussed how the inflation has happened I mean deflation actually if you can say like that the cost actually dipped simultaneously across the events of human genome project, alright. And what was only 6 percent in the year 1998 cellular diagnostics actually published the completed the task that was available at hand using Sanger sequencing at 2001 in 2001 in only the span of 3 years, ok. Well it was thought to be 100 percent, but later it was found it was not, but one good thing is human genome project also published their data at the same time. So, that did not give an opportunity to cellular diagnostics to patent the data gain financially, but it definitely helped I mean it fast send hasten the actual progress of work at human genome project libraries, right.

And ultimately the project was completed in April 2003 with the foundational reference of understanding of human genetics and advanced biomedical research. And now you can see this diagram is known I have already shown you this diagram, ok. See in 2001 the total amount of sequence was actually incomplete only late almost 2022, right that the entire genome was sequenced, ok. So, this April 2000 2023 marked the 20th year anniversary of human genome project completion. So, in last part of our discussion we will be discussing the so after you know we started with Sanger sequencing and this Sanger sequencing actually help human genome project, right.

And then we discussed NGS next generation sequencing which took the technology in leaps and bounds, ok. So, do you think that, ok. So, let us first see that how do Sanger sequencing hold up today with next generation sequencing, right. So, Sanger sequencing in terms of accuracy is very accurate 99.

9 whereas, NGS there is a scope of error. So, it is 99 to 99.9, right. Sanger sequencing is very cheap when we consider less than 20 samples, ok whereas, in case of more than 20 samples Sanger sequencing becomes expensive and cumbersome, ok. And then we have

to use we do generally go for NGS. Against if we are considering less than 20 samples the Sanger sequencing is much faster whereas, you know NGS involves library preparation, adapter sequence, adding sequence proper bioinformatic analysis. So, considering less than 20 samples that hassle is not what.

So, if it is more than 20 samples then NGS is good to go, right. But where NGS excels is in the sensitivity sorry, the sensitivity of Sanger sequencing is 15 to 20 percent and the sensitivity of Sanger sequencing is 1 percent this is actually the sensitivity in making error, ok. So, definitely these are good for NGS, ok. You see you might be asking the accuracy I mean when Sanger sequencing can determine a sequence, ok the final result we can trust them with 99.9 percent accuracy, right. But the initial detection and the sensitivity of the sample in the machine is this, ok.

So, do not confuse if you if you want to remember one thing for I mean regarding the efficacy you should remember the accuracy of the Sanger sequencing is much more compared to NGS, right. Next sample coverage, ok. We already told you regarding the read length we can use 800 to 900 base pair in Sanger sequencing, but we generally prefer 300 base pair for NGS. So, one another advantage of Sanger sequencing for long reads compared to NGS it can read I mean long sequences, right. However, regarding the amount of read per sample it is quite obvious with the NovaSeq platforms that have come out recently billions of sample can be read.

So, the thing that was done in a span of so many years, ok now in one run 128 human genomes can be diagnosed, ok. So, you might be thinking what if NGS was discovered. So, one thing so less than 20 samples it is very cost effective it is cheap. So, Sanger sequencing is still the gold standard of sequencing, ok. Mind it if there are less than 20 samples these are the situation less than 20 sample when we prefer Sanger sequencing to NGS.

Now, the we are getting back to the question NGS was discovered much later, right. So, what if NGS was discovered back then would not be so that human genome project could be completed in one day, right, right, wrong. Because the final step of NGS, Nags generation sequencing the modern sequencing method uses bioinformatics computer software where they actually tally whatever fragment of sequence reads are I mean coming out of the computer or sequencer to tally them with the known sequence where from they got the known sequence it is from the human genome project. So, without human genome project or without a prior mapping of the sequence NGS would not have worked, right. So, it is a paradoxical situation had NGS been there we I mean the final step it would have not been so easier, right.

Because the final step of NGS the way it can accumulate billions of data together by

matching with the first time known sequence. And during human genome project the first time known sequence is not there. So, everything was unknown the scientist had to map each and every sequence manually with the help of shotguns shotgun method and then with the help of Sanger sequencing by overlapping the data much more I mean there were many regions that were lost and with modern methods those areas were filled up and they were detected. So, to summarize we in our all the three sequencing lecture classes not all the three all the last this module we have covered Fanger sequencing, we have discussed all varieties of Nags generation sequencing technologies that are important for current laboratory as well as clinical use.

In this class we have covered SMRT ion torrent as well as pyro sequencing. We have I mean we discussed how bioinformatics plays a big role in NGS data analysis. We have discussed what is shotgun sequencing. Shotgun sequencing as I told you not exclusive to any type of sequencing, but it is a method by which we can organize I mean map the entire sequence using any method of our choice. We discussed what are the historical events and the progress of human genome project and lastly we discussed even the old method that was discovered in 1970s still holds good ok compared to all the modern methods of Nog generation sequencing in certain situations. So, these are my references for today's lecture and I thank you for your patient hearing.