An Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 10
An overview of NGS technology

Welcome to MOOC course on Introduction to Proteogenomics. In the last few lectures, the emeritus speakers Dr. Kelly Ruggles and Bing Zhang talked about genomics, gene mutations and its effects on phenotype. So, now, we have better understanding of the importance of mutations in a clinical study and level of impact it can make in the efficient treatment. Therefore, now you know the next logical way is how to do these experiments using the latest technologies available, especially how to use various type of NGS instruments to do these kind of experiments yourself.

In this light today we have invited one of the industry leaders Dr. Atima Agarwal from thermo fisher scientific. And she is going to talk about how one could use the next generation sequencing technology and sequence target genes in the genome. In today lecture, she will talk about NGS technology evolution and development over the time, and then there will be hands on session which we are also going to show how to use these instruments for generating data. So, let us welcome Dr. Agarwal to talk to us about recent advancements in the field of NGS and data analysis with a short demo session.

Basically we are going to talk about the recent advancements which are happening in sequencing field. Like pre in 1990, it was a very steady field wherein people were using radioactive label dNTP, ddNTPs and that is how things were getting sequenced. And then it was quite some time then like more than a decade wherein Sanger sequencing was the only sole sequencing technology wherein people were using fluoresce ddNTPs and they were sequencing various fragments.

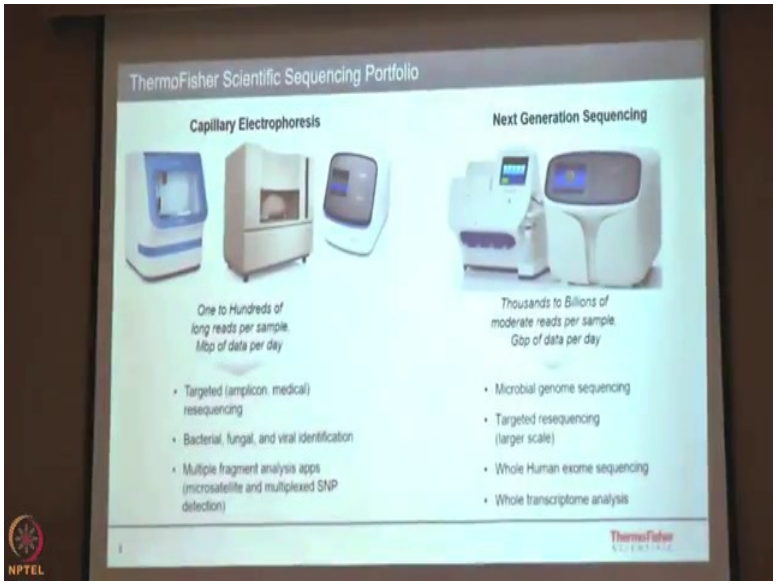
So, past decade there had been many advancements in terms of sequencing, it is not only Sanger now, there are many technologies which has come up in terms of next generation sequencing, and how they primarily differ from a Sanger sequencing or capillary sequencing which all of us were used to for quite some time was that earlier it was more or less it was that one well was sequencing one fragment. Yeah, so you could have

loaded a 96 well plate or a 384 well plate, and you could have got 96 sequences or 384 sequences.

Now, within this past decade what has happened is there is a term called massively parallel sequencing which has come up. So, massively parallel sequencing is something where in your sequencing millions of fragments at one go. So, that is what you are trying to do. And because you have been able to do that those millions of sequencing together, now you are trying to solve a lot of bigger problems faster. So, that is how like you would know that a human genome for the first time when it was being done on Sanger, it took around 10 to 11 years. Now, things are being done in 3 days, so that is how the technology has advanced.

And likewise even the cost at which you were sequencing the genome earlier and the cost at which we are doing now that is also tremendously decreased. I coming from thermo fisher, I am very, I would say I am very proud that I was all these decades my company was associated with these sequencing technologies and we had been pioneer in this arena.

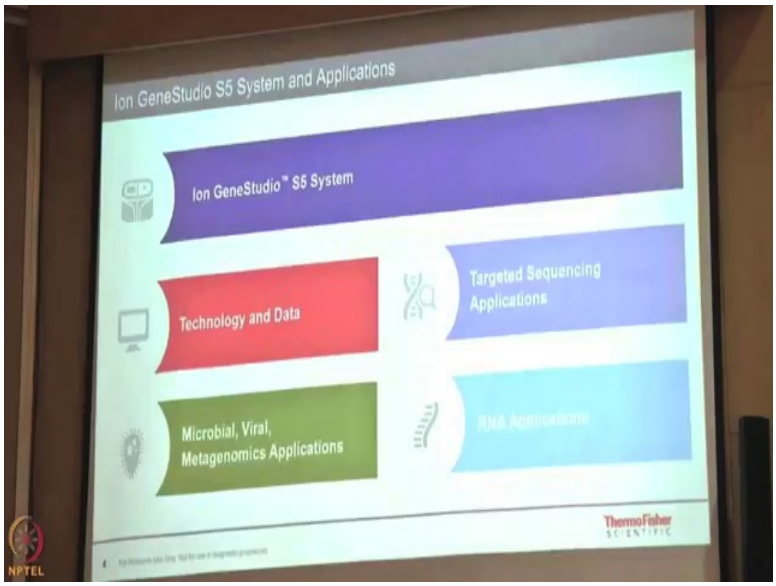
(Refer Slide Time: 03:57)



So, I would just like to show you very, so this is this is how our Sanger sequences look in they have where you are basically plating either a 96 well plate or a 384 well plate and you generate sequences. Here also we have come up with various advancements where in sequences on our cartridge base. So, the user has only has to put in a cartridge, and he

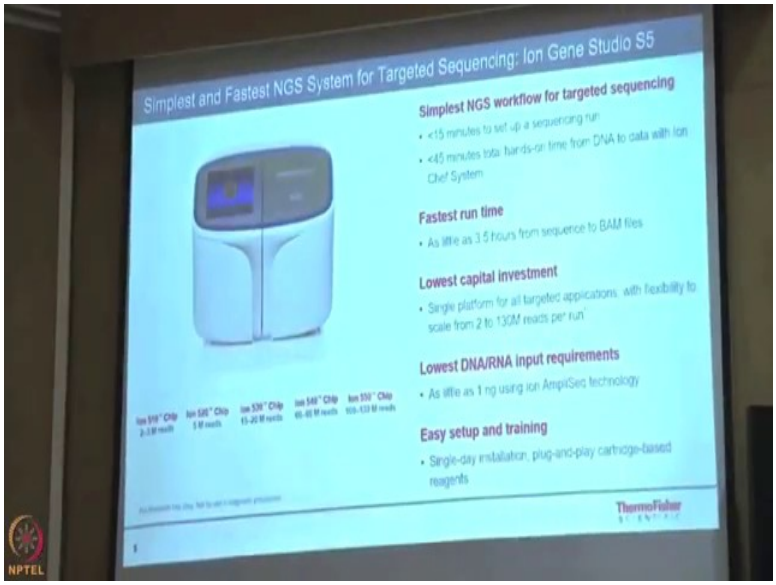
is good to go, and he just has his prepared samples. This is what is we are here to talk about today next generation sequencing wherein you have thousands to billions of reads being produced a per run and the data which you were producing is in terms of Mbs and Gbs.

(Refer Slide Time: 04:36)



These sequences can be like applied to various many different fields like metagenomics, microbial sequencing, RNAseq, exome sequencing, transcriptome sequencing, targeted sequencing.

(Refer Slide Time: 04:55)

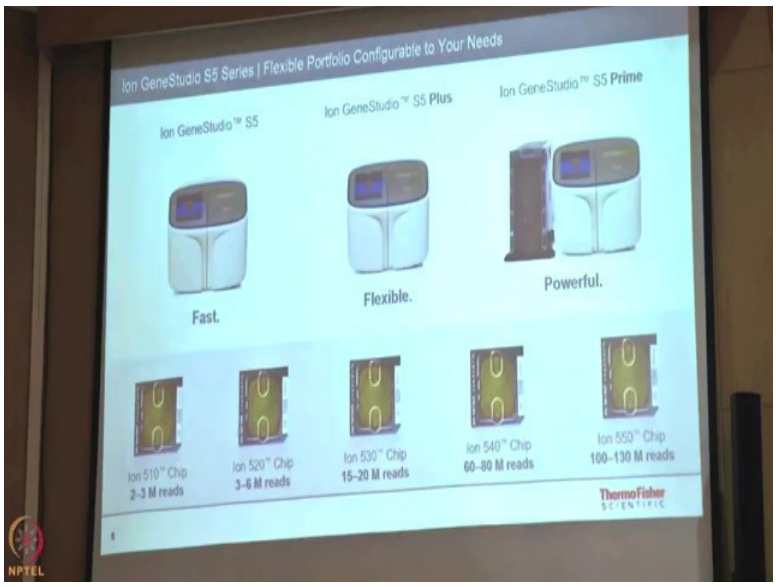


So, all of that a basic workflow what is involved is I will just briefly explain this slide. So, this is the sequencer which we are here to which we are here displaying. And this is Ion Gene Studio S5, which has five different chips, which gives you a versatility from producing anywhere from 2 million reads to 130 million reads that is a very big like very important feature. So, that like if you have this instrument in your institute, a person who wants to do microbial genome sequencing or who wants to just sequence a meta genome, he can also use the same system, and a person who wants to do a full exome or a transcriptome, he can also use the same system.

The other big highlight of our system is that one is it gives versatility that it gives you various like it gives you huge span in terms of the number of million reads. Another big advantage specifically for oncology is that within oncology, a very big problem is in terms of the amount of sample you are dealing with. So, generally dealing with tissue blocks which yield very very less amount of DNA or RNA as a starting material, so that ways we are we have a very robust technology called AmpliSeq technology and this technology has the capacity to multiplex many hundreds and hundreds of targets at one go, and amplify all of them at similar efficiency, and then sequence them all of them at similar efficiency.

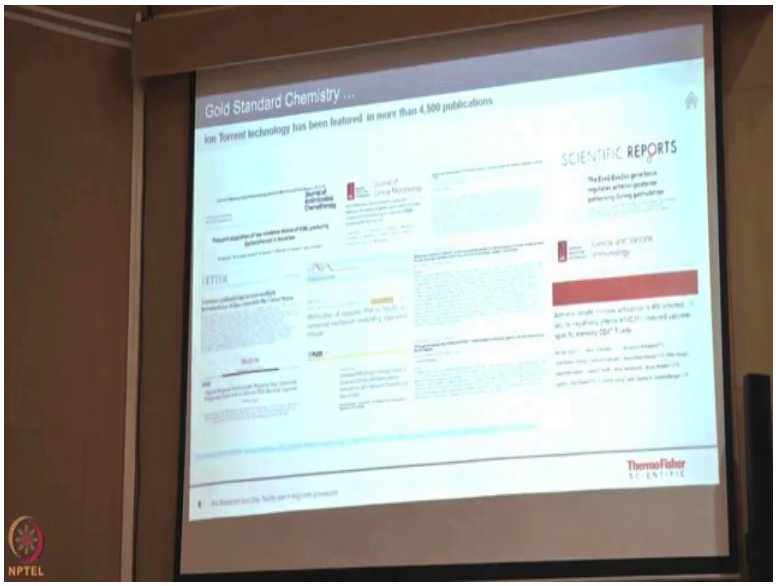
So, then you are not losing on so supposing you want to study 100 targets out of a tumor which is some very few mm, and you gets hardly 10 nanograms of DNA or nucleic acid material. Now, you want to study 100 targets. So, this Ampliseq technology is basically making it is such a robust technology that all these primer pairs which are working in one tube they are so efficiently being mixed that all these targets are being amplified that it is similar efficiency and that is how you are sequencing them at a similar efficiency. So, that is what we are if we have pointed it out here that this sequencer along with its technology has the lowest DNA or RNA input requirements.

(Refer Slide Time: 07:12)



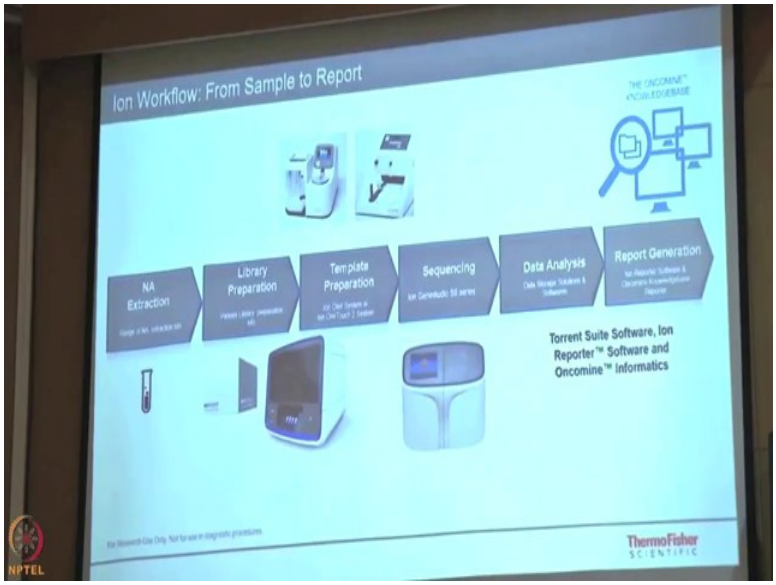
Then these are the three systems which Thermo Fisher provides GeneStudio S5, S5 plus, S5 prime. And basically they differ in their compute capabilities in terms of their throughputs and in terms of the compatibility of chips, they differ very in a very minor way, but other than that you have this 510 to 550 chip which can be used with these systems. And the difference in compute power as a user what it is giving you is faster runtimes. So, a user a very new user is good probably with S 5 wherein run times really do not matter to him much, an established cruiser might we would like to go for an S5 plus who has some planned experiments one by one. And then he just wants he does not want to get held with the time of run, and then this is S 5 prime wherein you have back to back runs. And this is the technology which has the capacity to finish the run within as little as even 2.5 hours.

(Refer Slide Time: 08:18)



So, this had been on the in this research arena or clinical diagnostic arena for since around 2010 this technology, and it has around 4500 publications all over. And since we have been focusing a lot into clinical oncology and all so, maybe for us it is more of application notes and all lately, and certain studies which have been done and published in different formats other than scientific publications. So, the workflow is what we want to show here is so a typical NGS workflow would start from nucleic acid extraction. So, this nucleic acid can be a DNA can be an RNA. So, once you have extracted this nucleic acid, then you prepared library.

(Refer Slide Time: 09:09)



Now, library preparation is nothing whatever DNA or RNA you have extracted, basically it depends on what kind of application you are working on. If you are working on microbial sequencing, then probably you are starting sample is a viral DNA or RNA or a bacterial DNA and RNA, and you want to sequence that order at one go. So, what you are going to do in that library preparation is that you are going to shear this genome either by enzymatic shearing or by mechanical shearing. You break down this genome into small pieces, and then you ligate adapters. These are double stranded fragments which are ligated on both the ends and that is what you do like, basically what you are trying to do is you are shearing down the interested genome into smaller parts and ligating it with adapters. The sequence of these adapters is known, so that is how the primer can sit and it can sequence the region in between, so that is what you do in a library preparation.

Now, for a oncology sample wherein you are trying to look for different markers specific for a particular cancer or particular type of cancer, what you are going to do is you isolate DNA and RNA, and we have wonderful panels which are like spanning from you we have panels for solid tumor, we have targeted panels for human ecology, we have panels for cell free DNA or RNA that has liquid biopsy which is gaining a lot of attention these days. Then we have panels for addressing the immune oncology also.

So, you use any of these panels, you amplify the targeted portions. And then once you prepared this library, what you do is, this is your template preparation. During this part what you are doing you are amplifying these library molecules onto the ions spheres.

(Refer Slide Time: 11:09)

This is being done onto this system, where in what you are doing is this these are two tubes. Well in one tube there is oil and in one tube there is a recovery solution or some kind of a detergent. This you place your plate here, and you place a filter where in your sample is kept with the ions spheres. This passes through this instrument and goes onto this plate. This plate is Peltier controlled, yeah, on both the sides, and that is how all these library molecules are then getting amplified onto the ion spheres. And from here from on this tube they are passing into a centrifuge which has two tubes.

(Refer Slide Time: 11:52)

So, all this amplified ion spheres are getting passed onto these two tubes. And finally, you are taking your amplified ions spheres out of this. Then you have an enricher. We have subsequent slides wherein I can tell you the biology of it.

(Refer Slide Time: 12:10)

You have an enricher wherein you have an 8 well strip kind of thing placed. And what you do is, that you are amplified ion spheres because not probably not all of the ions spheres have amplified. So, what you are doing is that you are placing that mixture here. When you were amplifying, you were using a biotinylated primer. So, whatever has amplified, whatever ions spheres of amplified, they have now biotin label on them. So, what you do? You place the streptavidin quoted beads in one of the well, and this is again a semi automated way this system this is called an enrichment system. This will go on its own for half an hour. And then finally, through this biotin streptavidin binding, it will fish out all the amplifier ions spheres, and it will leave the non-amplified ions spheres.

And once that is done, I will just cover that part in the slides first. So, we have two solutions; one is the semi automated way which we have presented here. We can show you the videos for the fully automated way of sequencing wherein you use a ion chef. So, wherein you do not have to deal with one touch two and enrichment, what you do is you put this ion chef this serves two purpose. One is it can prepare libraries view, and

second is it can do the all the amplification on these ion spheres, and you get ready to load it ready chips which are already being loaded.

(Refer Slide Time: 13:36)

These chips these are basically silicon wafers on which there are millions of wells, these millions of wells are basically enabling these millions of fragments to be sequenced all at once, so once. So, what happens is on the system so, this is the once you have your sample prepared, if you are using ion chef, then your sample is automatically loaded onto these chips. If you are using this semi automated system, then you have to load the sample with a pipette.

(Refer Slide Time: 14:15)

And once it is loaded, this is the sequencer which takes not even 15 minutes to set it up. So, there is a there is a lot of apprehension that a next generation sequencer is very hard to kind of handle, and it is it is difficult to manage that kind of a system that is what we want to show you that this is the chip clamp, wherein the chip goes and sits.

(Refer Slide Time: 14:45)

These are already bottled reagents, wherein if you see this is a sequencing cartridge wherein all the 4 dNTPs are being placed. Then behind this there is the waste cartridge. This is the wash solution which goes onto the chip every time sequence polymerization

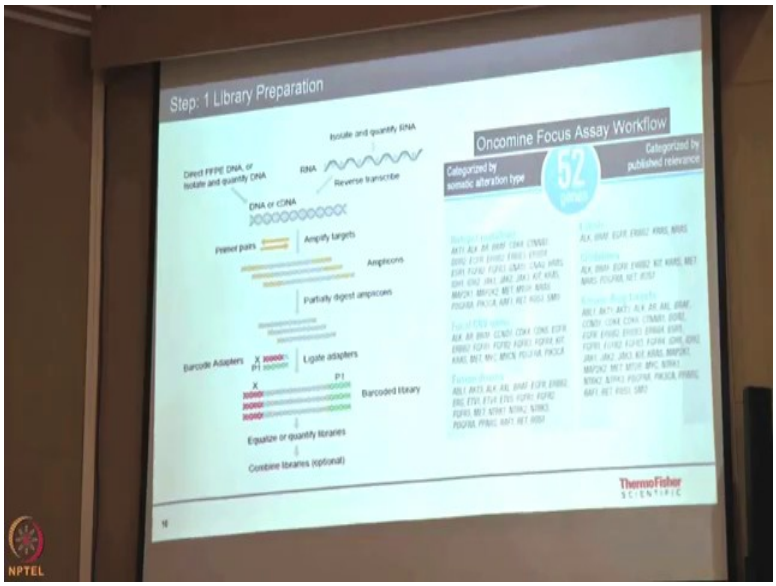
has happened. And this is the solution; this is the wash solution after every sequencing run. This wash solution goes onto the into the fluid lines and that is how it cleans the system.

So, now you can imagine how much time, it would actually take you to pull these cartridges out of a aluminum foil and place it into the instrument. The instrument on itself, once you have placed all these reagents on the instrument, so the instrument takes around 40 minutes to initialize itself and then you already have your loaded chip here, you are good to go. And this is the fastest system which can produce bam file which can produce data or which can finish a sequencing run within as little as 2.5 hours. So, we have we I would shown you various models of the chip starting from 510 chip to 550 chip, these chips have different capacities for an ion chemistry you can in terms of read lengths, you can go from a 200 base pair to 600 base pair, yeah.

So, why these read lengths are important? You consider different read lengths for different kinds of experiments. So, supposing your intention is to sequence certain markers from a tumor sample. So, now, a tumor biopsy as it is whatever DNA or RNA you if it is not a fresh frozen biopsy, it is something like FFP, it is coming from an FFP block that DNA or RNA itself is highly degraded. So, then you will use a smaller sequencing chemistry to sequence these small portions. If it is from if your intention is to sequence a meta genome, then your you are trying to; you are trying to deduce various kinds of bacteria or viruses present in that sample whether it is a sample which is coming from a cancer biopsy, whether it is a sample coming from an environment, whether it is the sample coming from the soil.

So, if you study metagenomics, you would like to read as long as possible. Then probably you would go in for a 400 or a 600 base pair chemistry, so that is how things differ in terms of read length. And what I want to point out is my colleague Harsh is going to talk about the data analysis part very fast. So, we are not only leaving you at the sequencing part, we have full-fledged solutions in terms of data analysis which takes you so in and further to for clinical people it also helps in generating reports. So, reports which can be directly given out to patients, and we told that this is the therapy which you could be using for treatment.

(Refer Slide Time: 17:41)



So, going to the next slide just these are just some figures of the workflow. So, this is the library preparation what I was talking. So, supposingly, I am just taking an example of one of the assay this is oncomine focus assay which has 52 genes these are all oncogenes. And in the form of hotspot mutations, CNV gains, fusion drivers.

So, it is a very well-known fact that if you are working with oncology and if you are working with cancer genomics, you could not be only looking at single nucleotide variations, or you could not be only looking at insertions and deletions. You have to look if you are looking at genomic aberrations you have to look them in totality. So, you have single nucleotide variations, you have CNV gains or losses, you have fusion drivers everything. Now, if these are the kind of mutations you are looking at, your starting sample has to be RNA as well as DNA, because these two parts of the markers are come going to come from the DNA, and fusion drivers are going to be reported from the RNA.

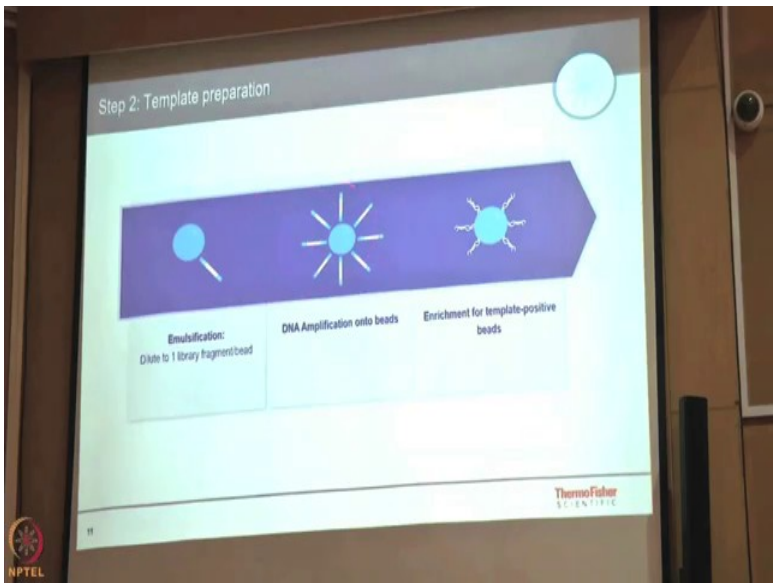
So, what you do is you isolate DNA and RNA you reverse transcribe the RNA, and then you have this pool of DNA or c-DNA, you take a primer pool. Now, this is this is what is the primer pool coming from. So, oncomine focus assay which is dealing with 52 genes. So, this primer pool will be targeting all these 52, all the hot spots and CNVs and fusion drivers in this 52 genes, and this is going to be one vial.

And your sample which is anywhere from even as less as 1 nanogram goes, and into this one vial, and all these targets all these targets, so these are basically 52 genes which are

actually addressing as they are addressing around 1000 biomarkers. So, all these 1000 biomarkers are getting amplified in a single tube and so this is that amplification happening. Once you have generated these amplicons, because these primers are coming from conserved regions, you partially digest these. And then you ligate adapters this is what is we discussed about library preparation that you want to ligate the sequence of interest to adapters, where in the sequence is known, so that the primers can come and sit here and sequence the region in between.

So, and we can use bar coded adapters why we are using bar coded adapters, so that we can use more than one sample on a chip. So, you have seen the chips. And so basically you would like to sequence more and more sample. So, it would depend we already tell you that on this particular chip for this particular panel, you can how many samples you can multiplex that comes with our literature. So, that is how you are using these bar coded libraries.

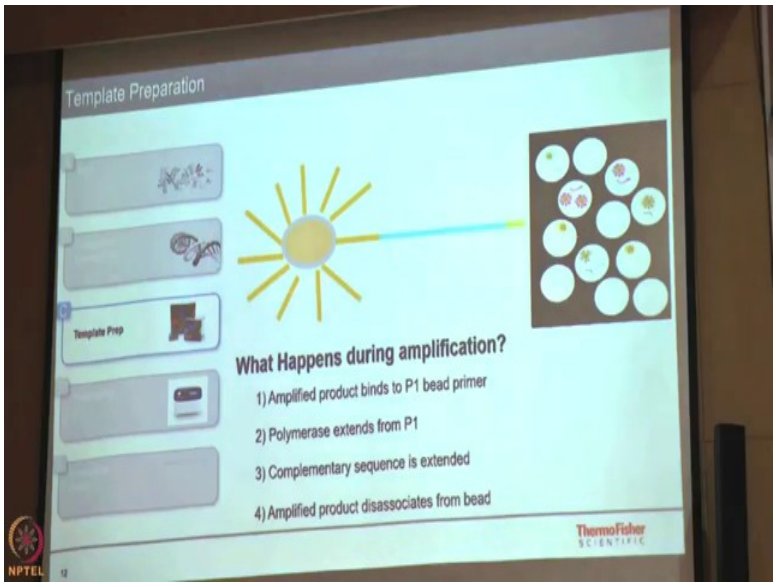
(Refer Slide Time: 20:40)



So, this was the second step which is template preparation which can be done on a semi automated way you would use in OT 2 and an enricher. When you want to opt for a fully automated way, it is ion chef which we could not get here, but we have a video will show you that video. So, what is primarily happening is these are your hundreds and thousands of library molecules, which are coming from the amplified portion which you have amplified from the sample. And then you like the idea here during template preparation

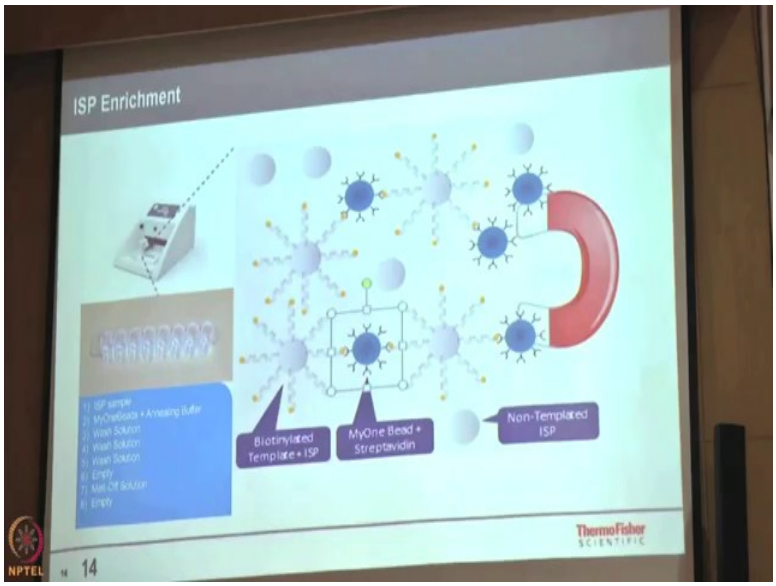
is that this one library molecule gets clonally amplified onto the ionosphere. So, that when this molecule is being read, the signal is enough that it can differentiate noise and the signal. This is what is happening.

(Refer Slide Time: 21:20)



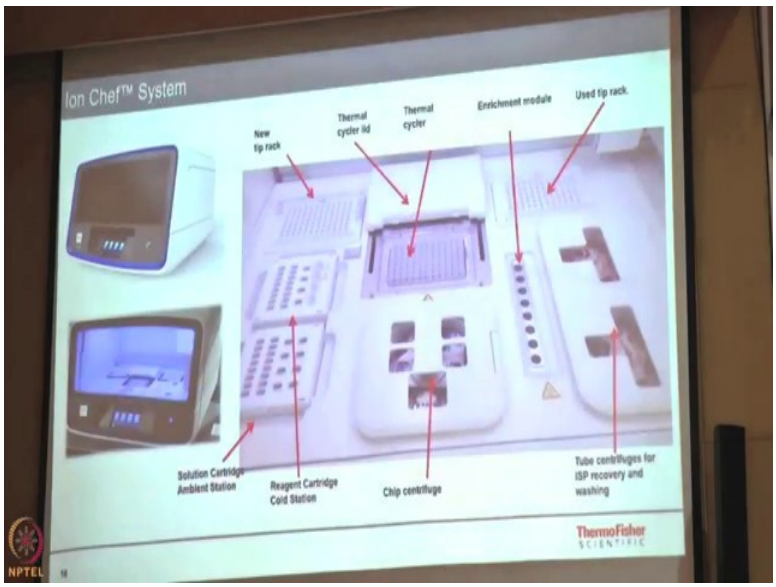
So, this is an ionosphere which has an oligo which is complementary to one of the adapters of the which are linked to the library. So, this is your library molecule which on both sides is linked with an adapter. And this is the primer and the primer extents so, this reaction is on this plate, and this is peltier control. So, all these small droplets which are being formed during the emulsion PCR, they were they are working as individual PCRs. So, these library molecules are then getting amplified and what finally you are getting is, you are getting this particular library molecule all over amplified onto the ionosphere.

(Refer Slide Time: 22:09)



Now, you are using streptavidin coated beads on this system and that is how you are fishing out the amplified beads with non-amplified beads.

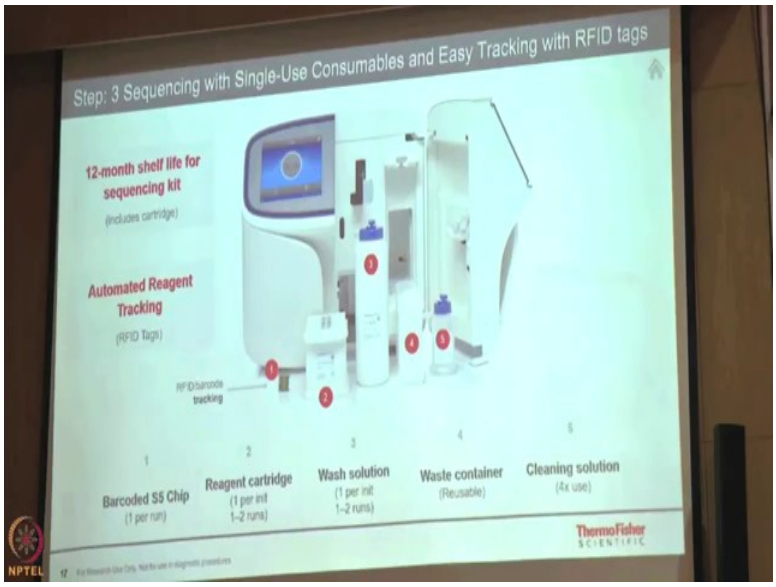
(Refer Slide Time: 22:23)



And this is the ion chef system, which is a fully automated way into how you prepare your libraries. So, when you are preparing libraries using ion chef system, you prepare your DNA samples, and you load them into these cartridges. And there are various other cartridges which are put in different for different sections of the system and that is how the final you get a final library pool into one of these tubes.

Right now what is being shown is the template preparation step on ion chef, you again load these cartridges, and then the system does the what is being done on this plate it is being done in a 96 well plate here. The emulsification and the amplification of that emulsion is happening, and then enrichment is happening through this cartridge, and finally, this is the centrifuge wherein you have your chips loaded. So, the chips are also loaded automatically, and then these chips are ready to go directly onto the sequencer.

(Refer Slide Time: 23:24)

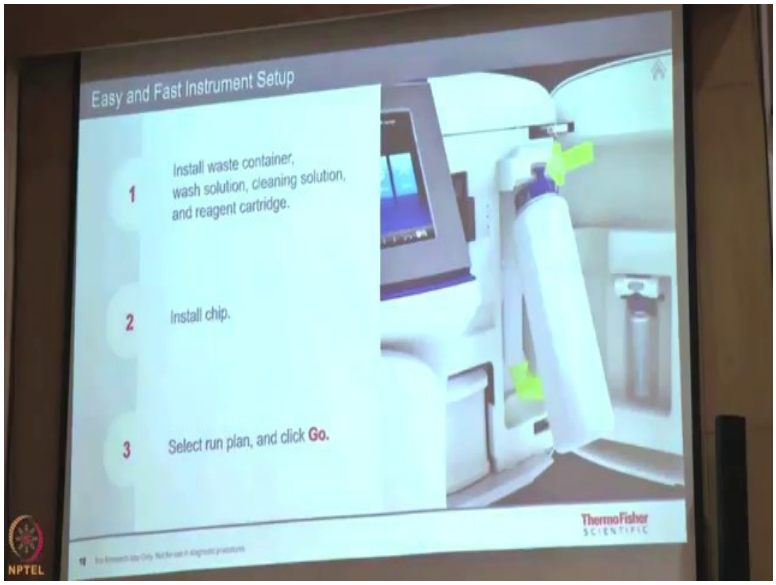


So, this is what we had seen. So, in the sequencer, you have first the chip which goes into this clamp which is a temperature control clamp. And you use polymerase and this polymerase because it works at a specific temperature that is why this clamp is temperature controlled. So, you already have your sample loaded onto this chip. Then this is the sequencing cartridge which has all the four nucleotides dNTPs, which are flown onto this chip in a sequential manner.

This is the wash solution. This is the discard which goes at the back of the wash solution. This is the wash solution which goes onto every time there is a nucleotide being sent to the on the chip. Then this wash solution goes and it washes any extra dNTPs left on the chip. And this is that wash solution which is placed here which you can see which cleans the system automatically after every run. So, it practically does not take a user to more than 15 minutes to set this sequencer up and get going with the run, so that is all. So,

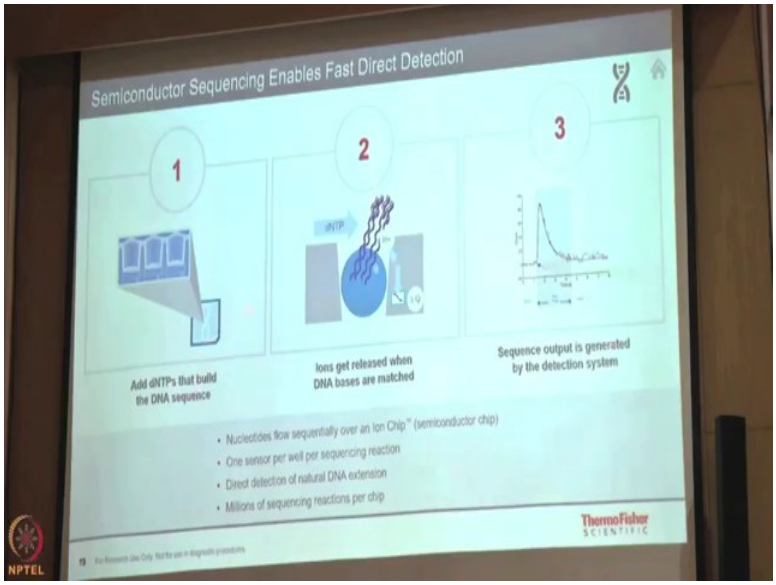
that's all. So, I practically shown you, you install the waste container, wash solution, cleaning solution and reagent cartridge.

(Refer Slide Time: 24:42)



You install the chip, and you select the run plan and you are good to go. This system also does not need any standalone computer with it. It can be connected with LAN and then you can operate on the system from anywhere.

(Refer Slide Time: 24:58)



So, what is happening in terms of sequencing? So, as you have seen this chip has millions of wells. And in each well what is happening is now this is supposingly an

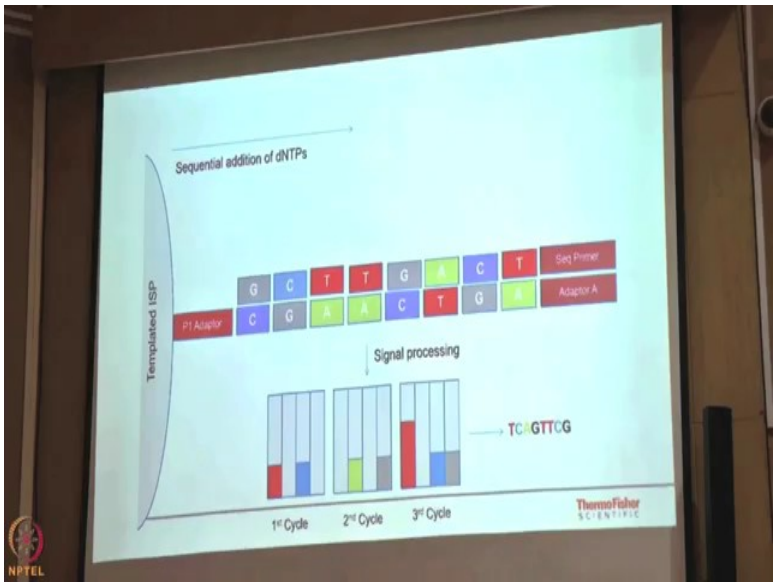
ionosphere which has amplified molecules onto that. We have already added primer which is against the adapter. Then the sequencer what it is doing through this cartridge, it is pumping various dNTPs in a very sequential manner which is already known to the algorithm of these software.

So, what happens is the moment there is a dNTP flown in onto the chip. So, supposingly we have flowing in ATP as the first nucleotide. So, now, this particular action is taking place on millions of wells on the chip. So, all those wells which have an ion sphere which has the T as the first base, they there and this ATP will go and bind. The moment there is a phosphodiester bond being formed is a hydrogen ion release this, and because we did clonal amplification. We amplified one library molecule all over the ionosphere.

So, there is not one hydrogen ion release. There is going to be thousands of hydrogen ion release which is good enough to bring a change in pH for this particular well. Each well has its own sensor. So, now, the system now the software will calculate or while the run is going on the software algorithm is also going on, and it is see that from the first; from the first run of dNTP when that was flown onto the system, how many wells showed a polymerization. And in how many there was a change in pH which got translated into a change in voltage and that is how we get a signal.

This signal is very well differentiated from noise and that is what gives you the accuracy in data. So, what is happening is the nucleotides are flowing sequentially over a chip, there is one sensor per well. There is direct detection of natural DNA extension, and then this phenomena is happening in millions of wells at one time.

(Refer Slide Time: 27:10)



So, just a cartoon to show this in a better way, this is your ISP. These are those two adapters which you have ligated during the library preparation process. And this is the sequence in between which is which you want a sequence. So, what is happening is that the primer which we have added which the system is already added, before the sample is being loaded onto the chip. The primer comes in and anneals here. There is a sequential addition of dNTPs.

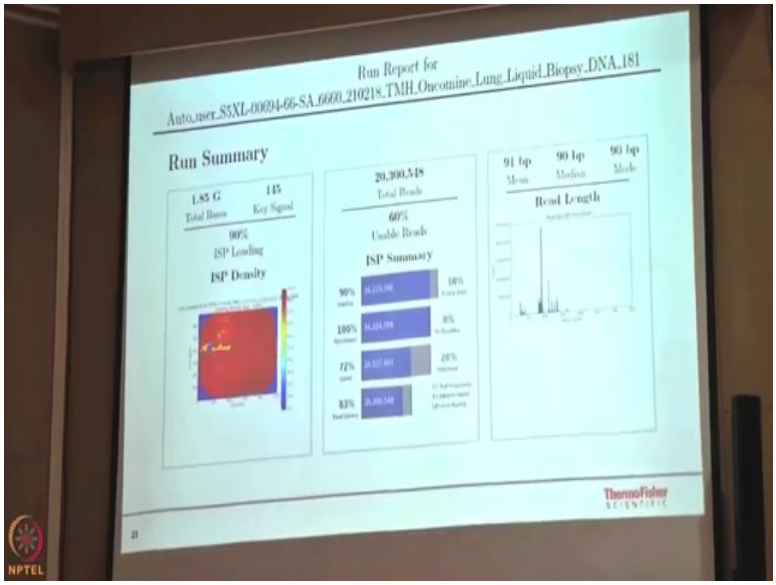
And so, a T will come in bind, there would be a change in pH which will be converted into change in voltage and that would be read as a signal. Then again now what happened there was, now this system has flown in a dATP. Now, the second base in question is g because they do not have complementarity that is this particular base will get washed off. Then this thing this sequential addition of dNTPs will go on, and signals will be captured. What do you think will happen at this stage? There are 2 As and we are flowing in a T.

Student: Signal will be doubled.

Signal will be double and that is how the system will know that they were 2 A's yeah. So, the signal will be exponential if they were had been 3 A's. So, that is how it will work. And based on these signals you get finally a sequence which is then a sequence coming from that particular well out of those billions of wells. So, after a two and a half

hour run, you have probably 80 or 100 million such sequences which now need to be analyzed.

(Refer Slide Time: 29:04)



So, this is how our typical run report looks that this is your chip this is the area where in you have loaded the sample. And this I will not go into these parameters, and this red basically this is the gradation in terms of how good the chip was loaded. This is 90 percent loading. And finally, it will give you some statistics. Now this is a 530 chip run. And what we have done is, we have run a liquid biopsy panel for lung which is targeting the variations which are there in DNA, so that will include single nucleotide variations and insertions and deletions.

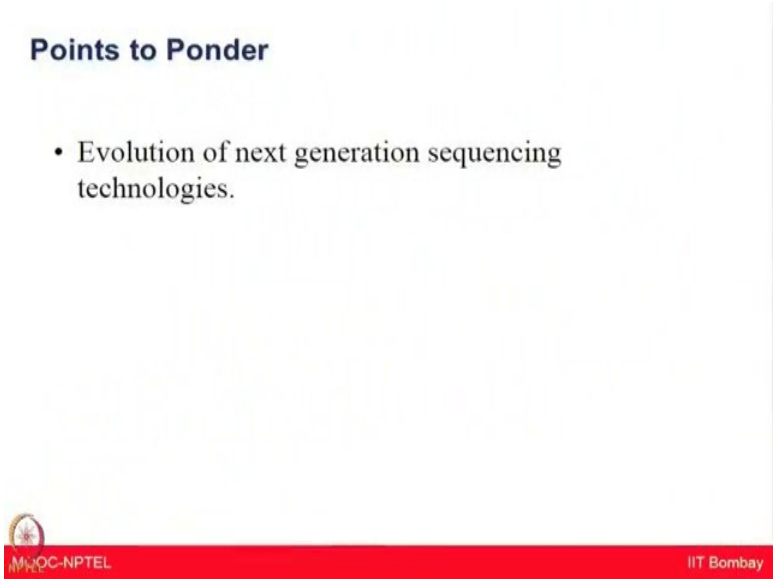
And this is the typical read length, because so as you would all know that liquid biopsy is gaining a lot of attention because of its non-invasive nature. Because it gives the way the it is being sampled, it gives a better representation of the tumor. Because if you are taking a solid tumor, it could we all know its tumor sample is heterogeneous.

So, what sample we are actually taking the DNA out whether that is representing the actual tumor position, tumor condition we do not know. So, liquid biopsy that ways is a better way of handling. You can more so because it is a non-invasive way. So, it becomes a very good prognostic marker. So, especially when you are dealing with lung cancer cases, you are not allowed to do biopsy many times. So, after a point of time or at times, there would be patients where in you are just not having that flexibility of taking

any biopsy. So, they are these liquid biopsies solutions are coming very handy. This again is a panel which has some 22 genes or thinks like, and these are all multiplexed in one tube, and they are all getting amplified at one go.

So, this is so you generated so many million reads, and these are the read lengths because it is a cell free DNA as it is by nature it is a fragmented DNA. You do not get DNA fragments more than 120 to 170 base pairs. So, you would like your prime appears to be amplifying very short regions, so that you do not skip on that.

(Refer Slide Time: 31:17)



So, today we have learned about how sequencing genome or a specific target gene sequencing has now become very easy accessible and affordable to many of the researchers which could facilitate them to dig deeper, and solve many clinical challenges. After understanding about genomics and looking at some of these hands on sessions about you know various technology platforms used from the leading industry leaders from the Thermo Fisher or Illumina Scientific, now we are going to move on to the next module which is on proteomics.

So, in the next module, I will talk to you about fundamental concepts of proteomic technologies, and then we will have other eminent speakers who will talk to you about current workflows using mass spectrometry based proteomics, and how to do different type of tools and databases for proteomic analysis.

So, let us stop here for the genomics module. And of course, you know you got excited you go you have learned lot of new things here, but you have also now got exposed to various type of publicly available databases and repositories which you can download yourself. And now start implementing those using various tools available and showed in this workshop from which now you can address many specific questions, and you can probably do many experiments even without having non-availability of even the technologies or the NGS system in your lab or in your own place, you can still do many of these experiments and obtain some very meaningful information.

Thank you.