An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Atima Agarwal
Prof. Harsh
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Supplementary - 05
Applications of NGS - Ion Torrent

Welcome to MOOC course on Introduction to Proteogenomics. After understanding how one can sequence gene by using Illumina platform. We will now move on to another NGS platform by thermo fisher. Our today's speaker is Dr. Atima Agarwal who has already explained us about how the thermo NGS platform works in her previous lecture.

In today's lecture she will talk briefly about how the relate files look and how this platform has increased the coverage. She will also talk about various applications of this NGS platform especially in the areas of metagenomics, genomics and multiplexing sequencing. She will also talk about thermo database of 16 sRNA from more than 9000 organisms. Dr. Atima Agarwal will then talk about already available panels of well study targets which one could use directly their own research.

Another colleague from thermo Mr. Harsh will also joined her and talk about the work flows how one could analyze the sequencing data. He will also talk about the Torrent Suite software to load the basic softwares and requisites required for using the sequencer. So, let us now welcome Dr. Atima and Mr. Harsh from Thermo Fisher.

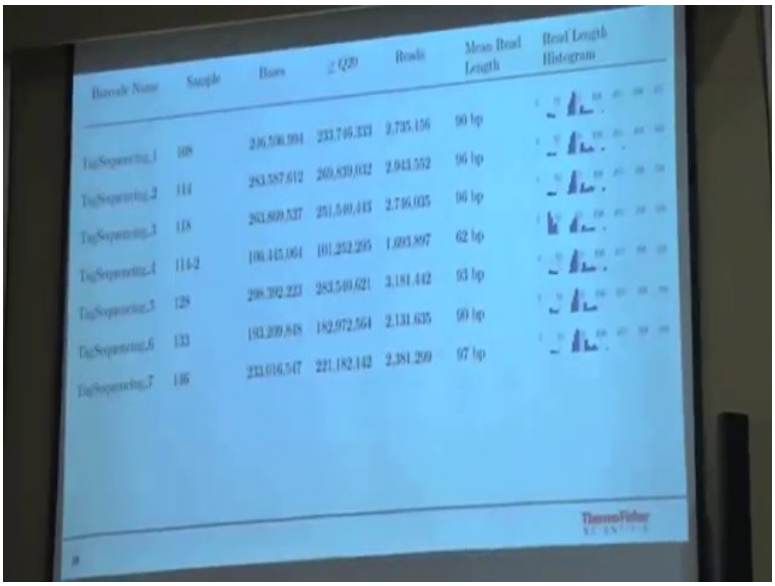
We all know it is tumor sample is heterogeneous. So, what sample we are actually taking the DNA out whether that is representing the actual tumor position, tumor condition, we do not know. So, liquid biopsy that ways is a better way of handling and more so because it is a non invasive way. So, it becomes 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 wherein you are just not having that flexibility of taking any biopsy, so there these liquid biopsy solutions are

coming very handy. This again is a panel which has some 22 genes or things like and these are all multiplexed in one tube, and they are all getting amplified at one go.

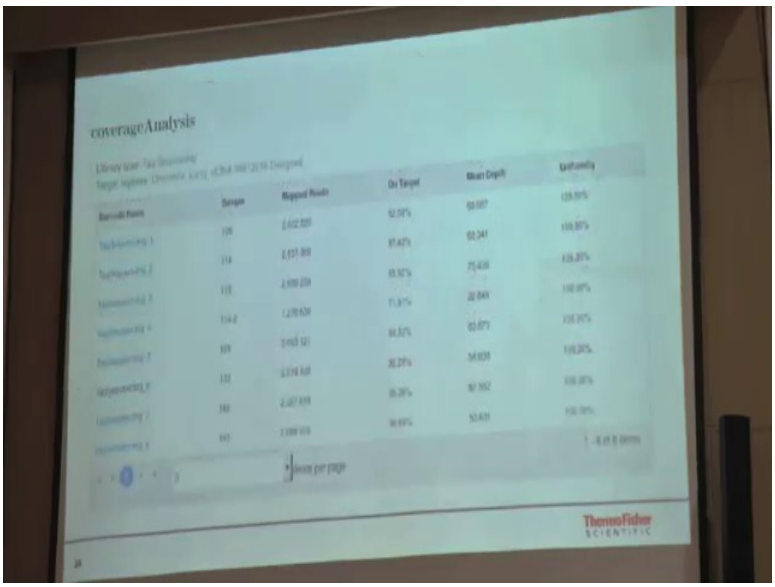
So, this is. So, you generated so many million reads, and these are the read lens, because it is a cell free DNA as it is by nature it is a fragmented DNA. You do not get DNA fragments more than 120 to 170 base pairs. So, you would like your primer pairs to be amplifying very short regions, so that you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

(Refer Slide Time: 03:09)



Since we are barcoding many number of samples. So, this is just the statistics of how many bases every sample has produced, and these are the basic read length histograms for every sample.

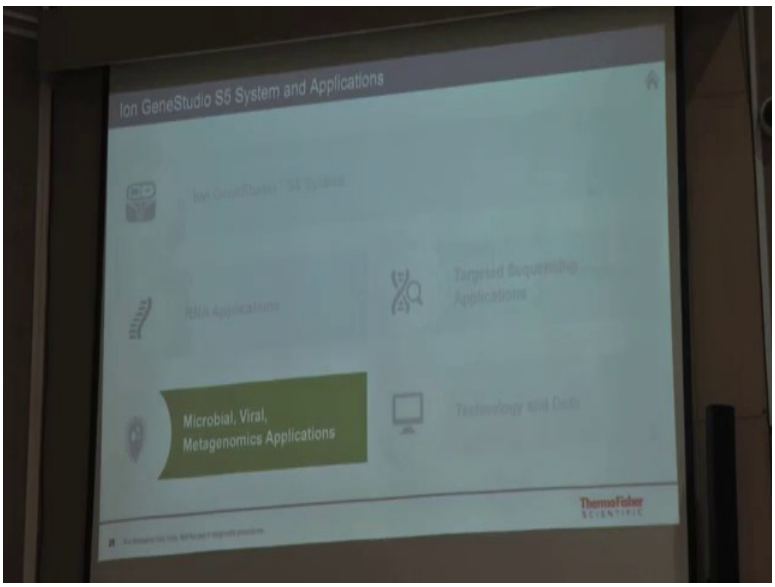
(Refer Slide Time: 03:19)



Then it shows that now you were targeting prob for example, 52 genes within the whole 3 GB human genome, yeah. So, how well that is covered; so, there is no point of doing an NGS if out of those 52 genes you are not getting good data for 7 genes, 8 genes. So, the whole point of doing an NGS is that you can do a multi biomarker study with very little amount of sample, and all with the same efficiency which is which is kind of a very specific highlight for our system. So, this is what is being shown in this coverage analysis that how many reads were on target, what is the depth.

So, in NGS we talk a lot about the depth like in Sanger what we use to do so supposingly you have now amplicon you would sequence it with a forward primer, you would sequence it with a reverse primer. So, that is what you are getting a 2 x covering that is all. In specially with the samples with the oncology samples these samples being heterogeneous in nature there could be a norm lot of normal cells or wild type DNA fragments which are coming. So, you tend to sequence them at a larger depth. So, this is the depth which is shown at times it is in hundreds, thousands also, and this is how well that those 52 genes have been covered.

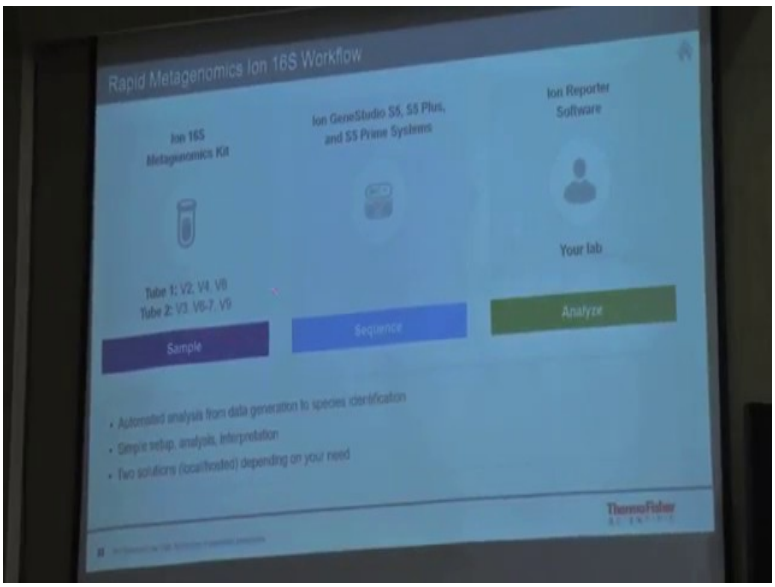
(Refer Slide Time: 04:47)



So, how well all those 52 genes at one go has been covered. So, these statistics all have been will be generated once the run is over within 1 or 2 hours. All these statistics are already generated and then you have an array of applications which can be done on the system. You can do RNA-Seq, you can do microbial, viral, or metagenomics applications wherein you are trying to sequence these genomes, you can do targeted sequencing.

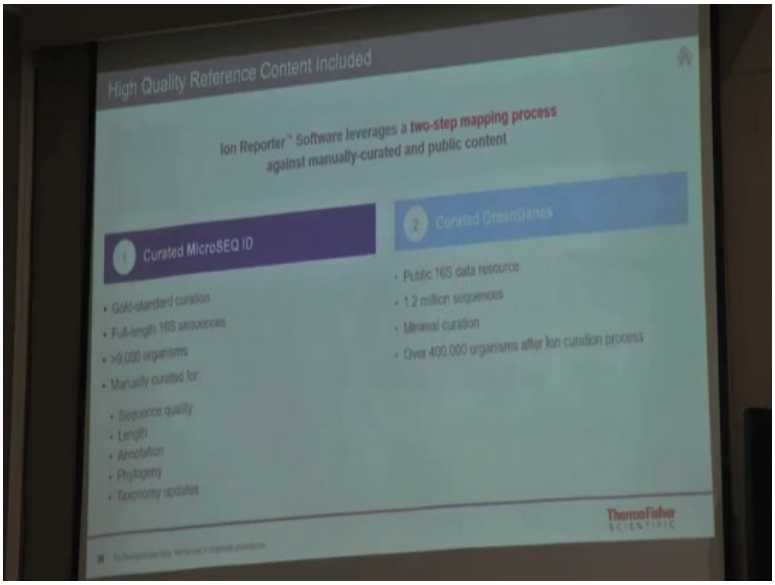
Now, these targeted sequencing is generally done by various kinds of panels. These panels can be targeting various kinds of inherited diseases, these panels can be targeting various tumors, so various markers which are known to be relevant in various kinds of tumors and then once you have the data, we have softwares in place to give you very well annotated variance which can be directly reported.

(Refer Slide Time: 05:40)



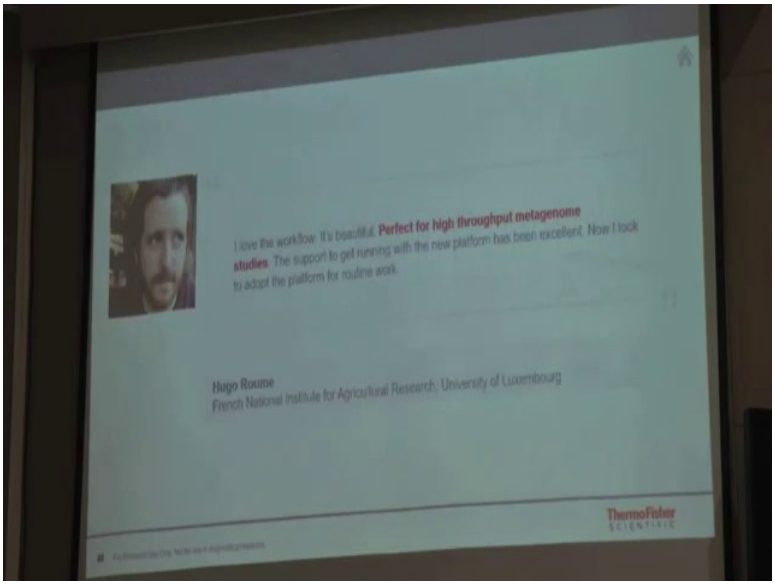
So, this is a metagenomics kit which sequences 7 hypervariable regions of the 16S rRNA at one go and you can multiplex depending on the number of region now amount of data which you want to generate per sample, you can multiplex these samples. And finally, once you have amplified you ligate adapters, then it goes on to chef on to One-Touch 2 and ES and you sequence them and you analyze the data on ion reporter software, wherein you are basically blasting these sequences to various databases.

(Refer Slide Time: 06:20)



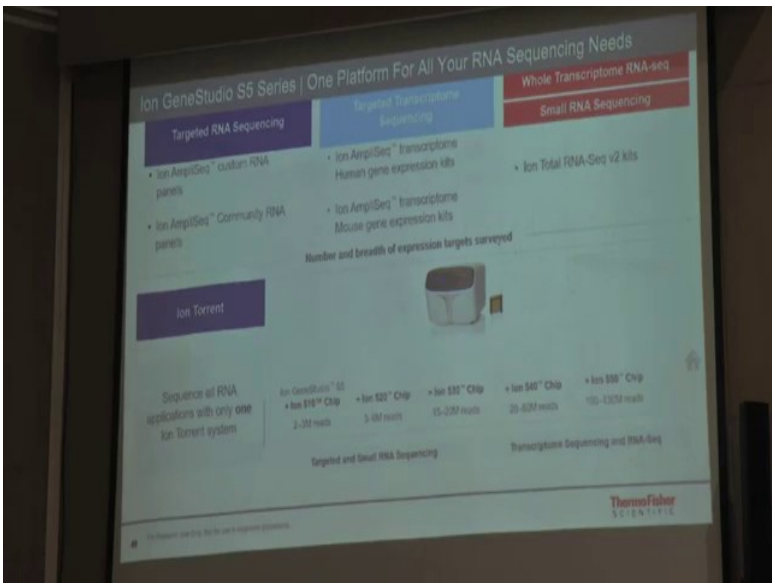
So, we have a Curated MicroSEQ ID which has full-length 16S sequences more than 9 coming from more than 9000 organisms. This is a database which has been which is maintained by Thermo Fisher. Then there is a Curated GreenGenes database which has with this is a publically available 16S data resource which has around over more than 400,000 organisms after curation.

(Refer Slide Time: 06:48)



So, these are some comments from some users we have plenty of them, we have plenty of publications, we have plenty of clinical trials going on. So, going to RNA applications again we have various solutions.

(Refer Slide Time: 07:07)



Starting from targeted RNA sequencing, so if you want to sequence certain specific transcripts or certain specific RNA within the within your sample, you can design a custom panel. If not that then we already have a readymade kit which is AmpliSeq transcriptome, human gene expression or AmpliSeq transcriptome, mouse gene expression kits, these are these gene expression kits are targeting more than 20,000 RefSeq transcripts and the beauty is that you can get all this data with as little as 10 nanograms of RNA.

So, this is one thing then one another way of planning a RNA-Seq experiment is that you just want to sequence anything and everything whatever RNA is there in your sample, you want to sequence the mRNAs, you are going to sequence the long term coding RNAs, you want to sequence small RNA. So, there what you do is you enrich your sample based on the fraction which you are interested in.

So, supposingly you are interested in sequencing all the mRNAs which are lying in your sample. So, what you will do is, so now, we all know that all the mRNAs are adenylated. So, they have a A tail. So, you will use a oligo dT beads and fish out all these A tail fragments and then prepare a library by almost similar way what we have discussed.

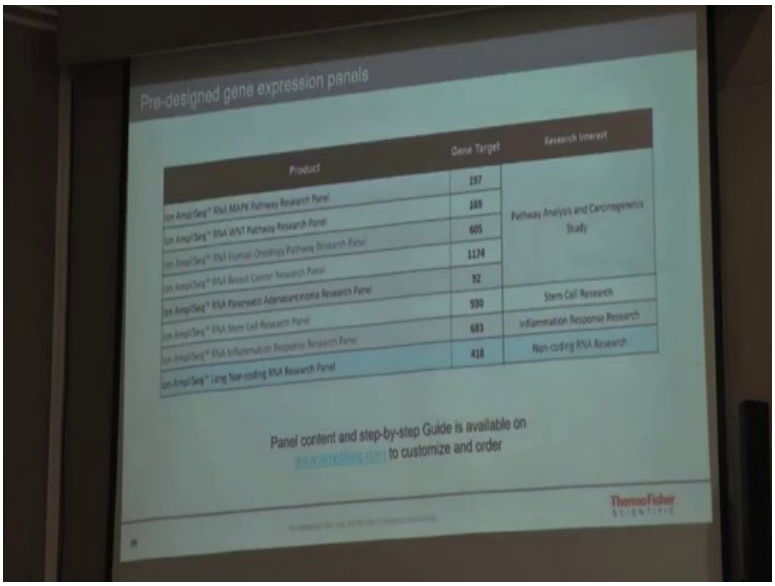
Now, if you are interested in the total transcriptome but not the ribosomal portion, because we know that 90 percent of the cellular RNA will have ribosomal portion which is quite constant and you really do not want to waste your reads being mapped to ribosomal portion.

So, there are kits from Thermo Fisher itself which something like ribominus which will deplete this ribosomal fraction and now what you will have is all the RNA fragments coming from your sample which are apart from the ribosomal fragment fraction. So, you sequence all of that at one go.

Then small RNA sequencing, small RNAs we know that they are generally in the range of 18 to 40 base pairs. So, how to enrich these small RNAs, based on the size exclusions. So, there are certain bead based exclusions or column based exclusions which you do and you enrich these fragments and then again you ligate it with adapters, and then you sequence this fraction.

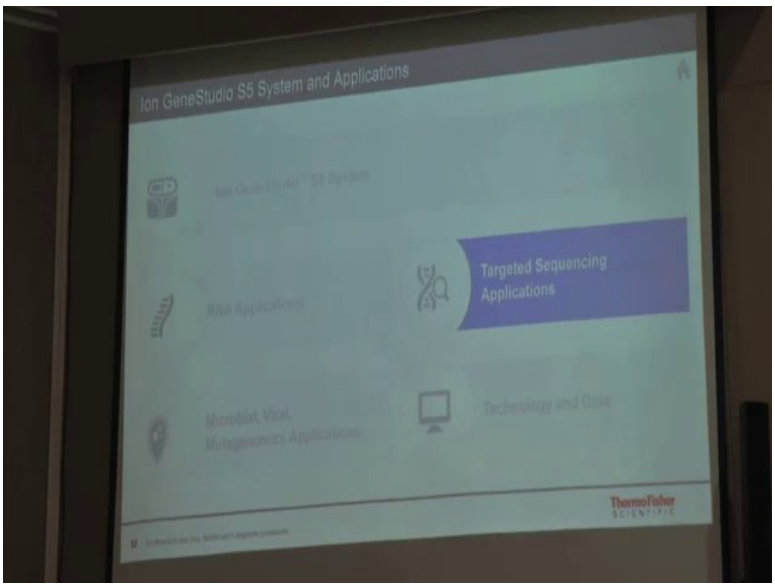
And again since the system is quite versatile, it will give you from anywhere from 2 million reads to 150 million reads. You can plan all these experiments on a single instrument just by changing the consumable like by changing the either you use a 510 chip or a 540 chip or a 550 chip base. So, the box remains the same which is the main I would say main expenditure in terms of having an instrument. It is just the consumables which are changing and which are giving you variable throughputs, and which are allowing you to do multiple different kinds of experiments.

(Refer Slide Time: 10:10)



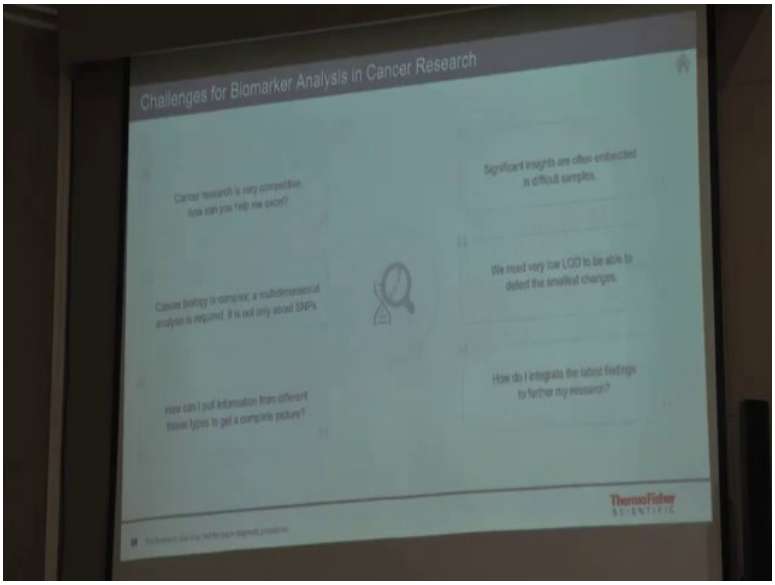
We already have certain RNA panels which are there off the shelf. So, if you are working on any of these kind of pathways, you can just take these panels and you are good to go.

(Refer Slide Time: 10:23)



And then coming to targeted sequencing applications.

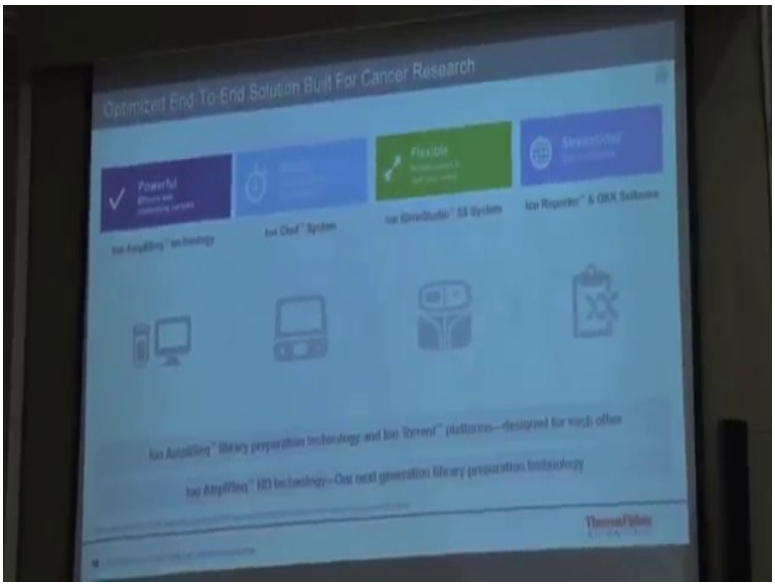
(Refer Slide Time: 10:26)



So, again as we discussed that some there are some specific challenges for in cancer research that I do not want to sequence only single nucleotide variations, I want to sequence all different kinds of aberrations which are going on in the genomic content which includes fusions, which includes various kinds of a CNVs and all. So, and then you want a very these panels need to be very specific, they need to have that capacity that all the targets are being

amplified with the equal efficiency and then you at the same time you need to have various efficient LOD so as to not compromise on the sensitivity of the acid. So, how do we integrate all these things into one thing?

(Refer Slide Time: 11:28)

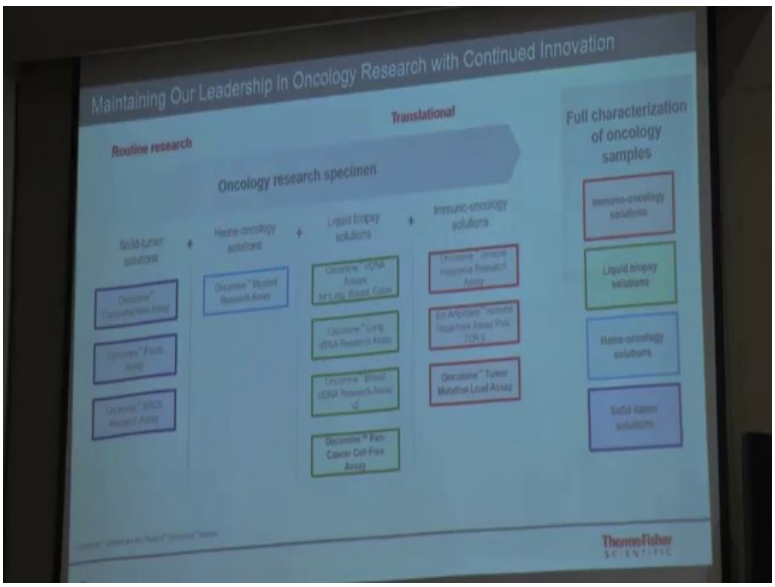


So, we discussed about the power of AmpliSeq technology so, which is a kind of boon to cancer research wherein all the templates or all the targets are being amplified with equal efficiency. Then these are being templated on Ion Chef System, sequenced on GeneStudio S5 and Harsh is going to talk about ion reporter and OKR Software briefly to tell you that how these variations are then annotated, and then finally, a report is given out.

Very recently like not very recently but sometime during the beginning of this year, we had come up with even better technology than AmpliSeq. This is AmpliSeq HD technology which is specifically designed for samples which are highly degraded and samples; so, these would be samples or DNA or RNA which is coming from FFPE blocks or coming from liquid biopsy samples.

So, this AmpliSeq HD technology really helps you in get not compromising on the LOD, and still amplifying all the targets at the same efficiency.

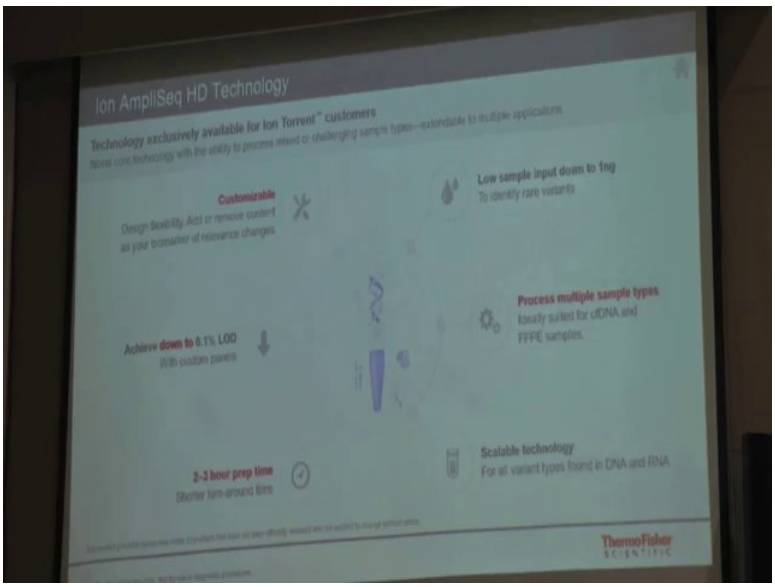
(Refer Slide Time: 12:42)



So, this is our various difference, so basically with all these panels, we are enabling full characterization of oncology samples. These are various panels which are targeted for solid tumor cancers, these are for heme-oncology, these are for liquid biopsy if you are working on liquid biopsy and we all know that immuno-oncology is being discussed a lot. This is basically; the idea here is that you are harnessing the own body's immune system to fight against these cancers.

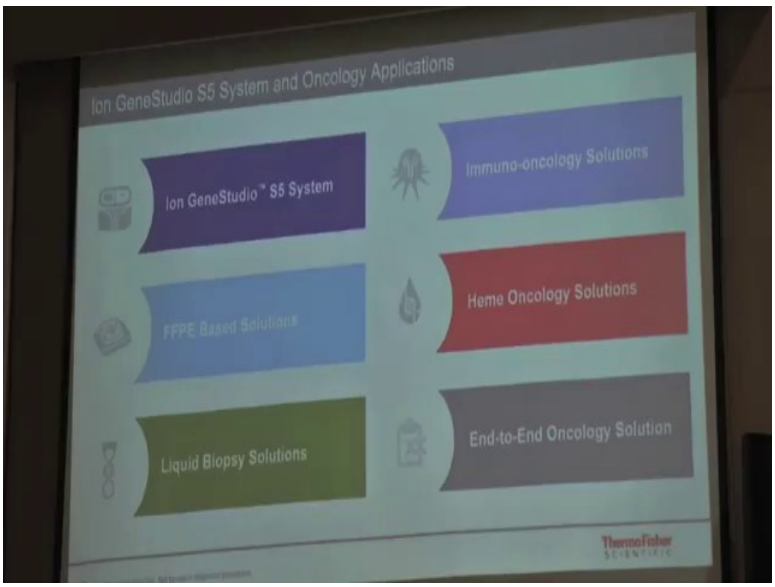
Now, when these kind of therapies are being used, it is very important to know that whether a person is going to respond to these therapies or not. So, to make a well informed decision that whether these persons are going to respond to these therapies, these are those various assays which will help you define that whether you are working with the active tumor microenvironment or a suppressive tumor microenvironment.

(Refer Slide Time: 13:46)



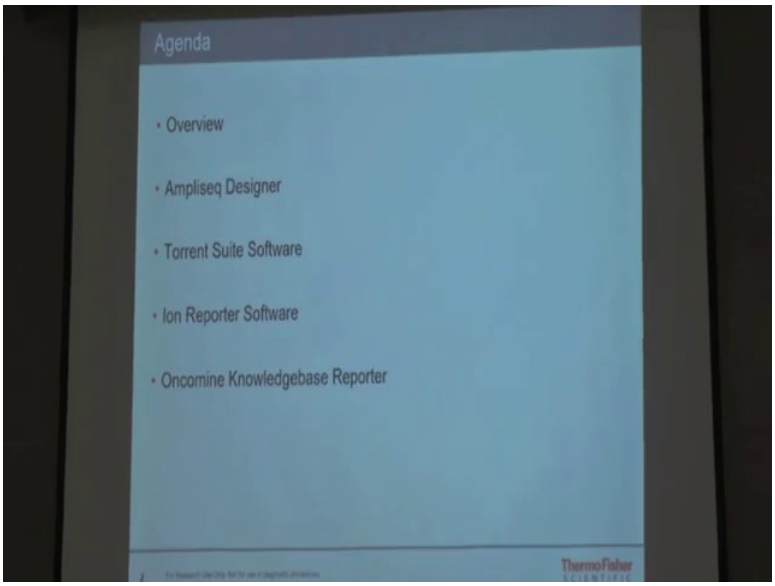
So, basically I just spoke about AmpliSeq HD technology, what it is enabling us to do is that you can use as low as 1 nanogram of sample. You can process multiple different kinds of samples together cell free DNA FFPE samples. This is a scalable technology which works for all different kinds of variations found at DNA and RNA level and then you have a shorter sample prep time, and the limit of detection of the panels which are used to which are being designed using AmpliSeq HD technology at the back end is as low as 0.1 percent and the beauty is that if at all you are not happy with the panels which are already there on shelf, you can design these panels.

(Refer Slide Time: 14:35)



So, Harsh is the core bioinformatic bioinformatician. So, he is going to briefly talk about these softwares, how they work.

(Refer Slide Time: 14:52)

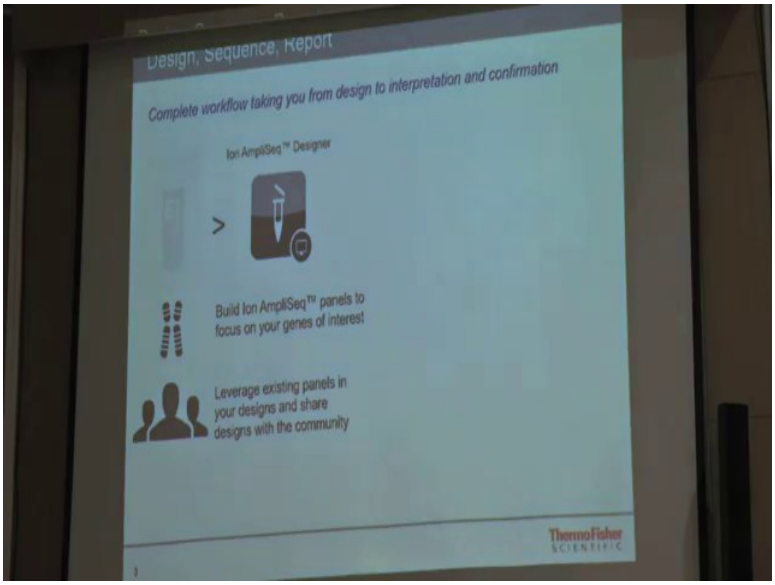


I will briefly talk about the overview, how we are positioned. I will talk about the AmpliSeq Designer software which is the tool for designing an assay. I will talk about the Torrent Suite Software which comes along with the instrument. So, whenever you have any instrument in

place and in your lab, this particular Torrent Suite software will be preloaded on that, and it will do all the basic jobs for you.

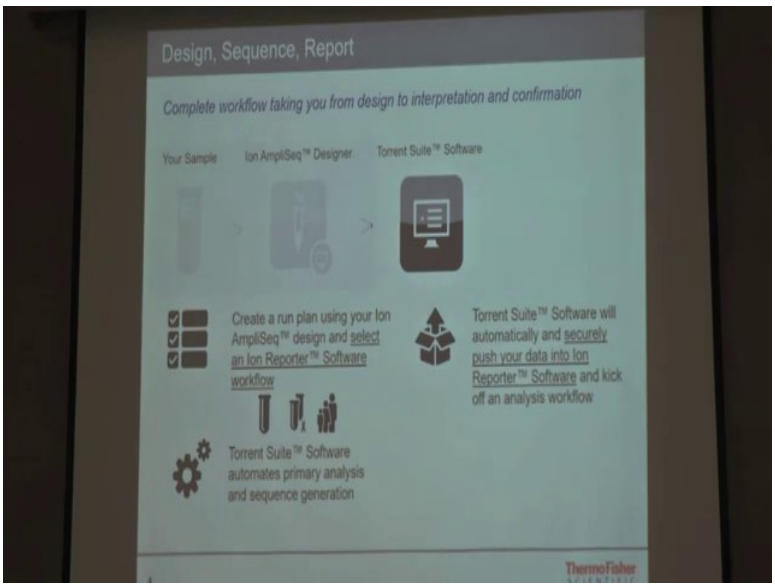
So, irrespective of whether you have H 5 or H 5 plus H 5 prime or PGM the software are going to remain constant, fine. Then we have a Ion Reporter Software that is basically for annotating your variants, finding out the meaningful information about your mutations present in the sample, and then lastly Oncomine Knowledgebase Reporter software for finding the relevant therapies or drugs available against a particular mutation.

(Refer Slide Time: 15:51)



So, quickly so we will start with the AmpliSeq designer. It is mainly for a targeted sequencing. So, what do you do is either you we already have a lot of panels available of the shelf. So, what do you do is; you go to AmpliSeq designer; see the list of panels available for you. If any panel is suitable for your requirement, you can directly go ahead and order those panels. If not, when you create your own panel; panel is basically your own assay right.

(Refer Slide Time: 16:12)

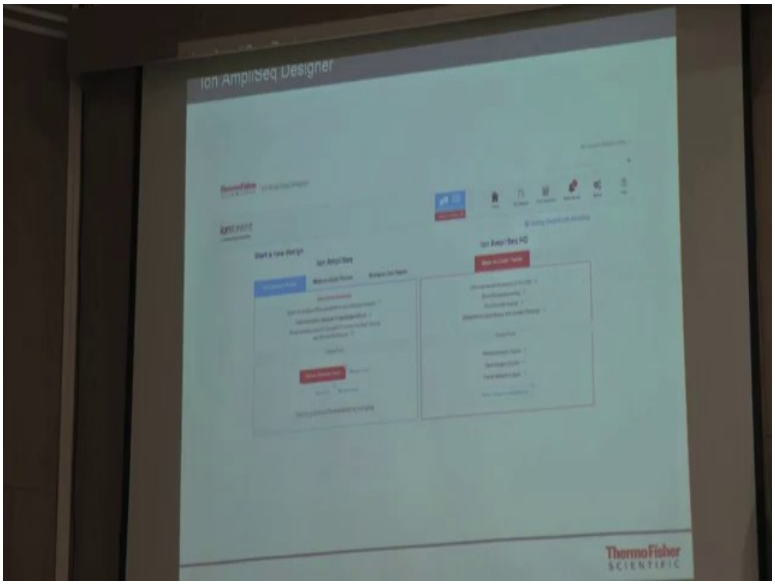


Then you sequence on the instrument; now machine comes with a software called Torrent Suite software. So, it will basically convert your raw signals into ATGC sequences. It will do the preliminary analysis, it will align the data back to the human genome or any other organism that you are working with and it will find the mutation. Now, once you have the mutations available you go to a software called Ion Reporter for annotating the mutations Now, this is the workflow is you design your assay, you sequence your samples and then you finally report ok.

(Refer Slide Time: 16:54)

Now, to begin with the AmpliSeq designer. This is how the user interface looks like. It is a web based portal the link is www dot ampliseq dotcom. The registration is completely free. It should not take more than 5 minutes for you to sign up there. Once you sign up for AmpliSeq designer, you are automatically signed up for the Thermo Fisher cloud as well as the Ion Reporter software. So, you do not have to have multiple login credentials for Thermo Fisher softwares.

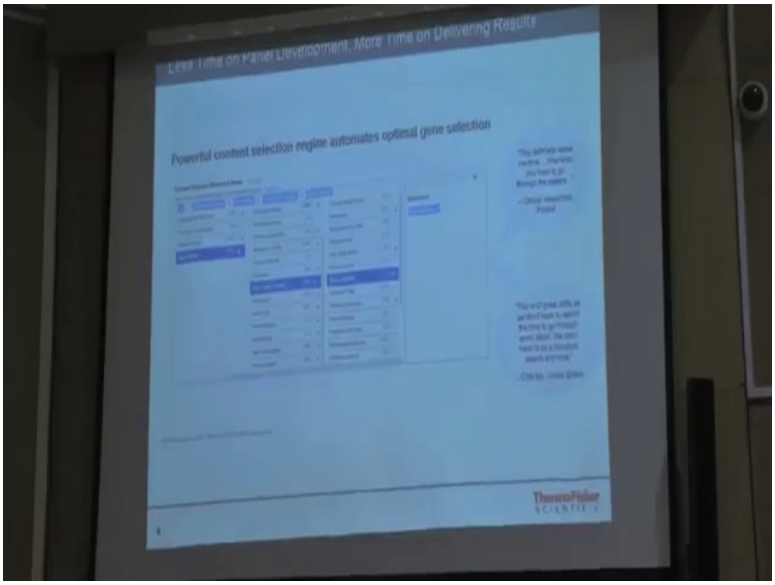
(Refer Slide Time: 17:18)



When you log in this is how you see. So, on the left hand side, you have a ion AmpliSeq; on the right hand side, you have an AmpliSeq HD. On the left hand side, you have on demand panels, made to order panels and ready to use panels. So, ready to use panels are the panels which are created by Thermo Fisher. Made to order panel has basically two sections whether you can create your own panel or there are community panels which are designed by some other users and which are made publically available for the other users and then third is on demand panel.

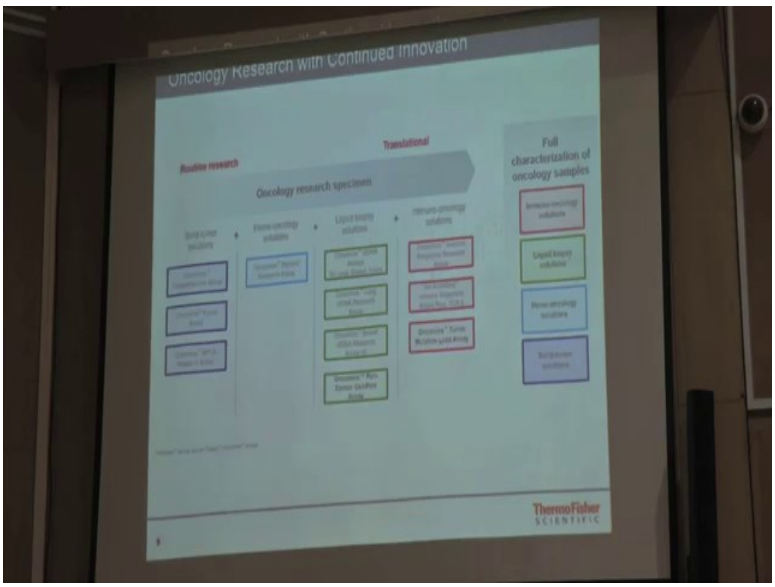
So, for on demand panel we have a 5000 odd genes. We have already designed the primers, and you can choose which genes you want to sequence out of those 5000 odd genes. So, there are two category is that either you can select the genes by the disease research area or you can include your own gene list as well. So, let us say if you are beginning with an experiment where you do not know what are the genes associated, what with a particular disease; then you simply go to the on demand panel.

(Refer Slide Time: 18:23)



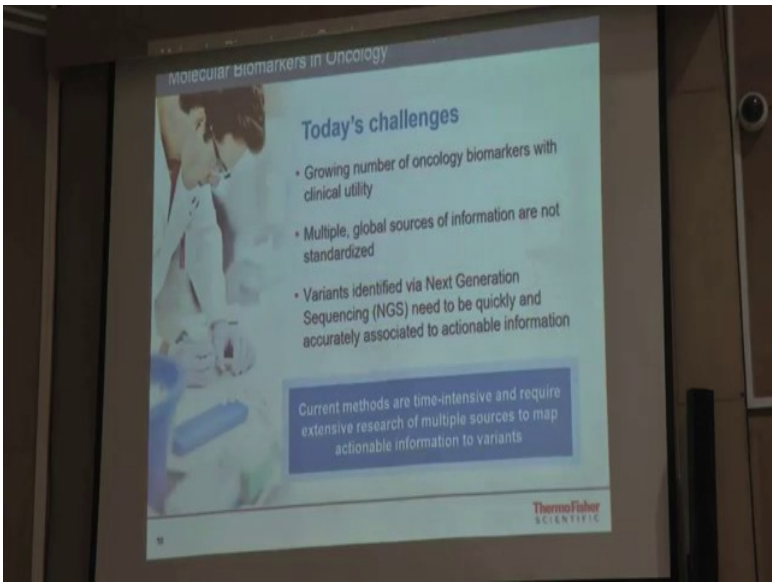
And then there are list of categories available. You keep on further bringing it down till you it that is your requirement and then the finally, design a panel and then proceed further.

(Refer Slide Time: 18:33)



But these are the basic oncomine panels that we have available with us. So, if you use any of these panel, we have a complete solution available. You can go to our softwares and do the analysis for those samples.

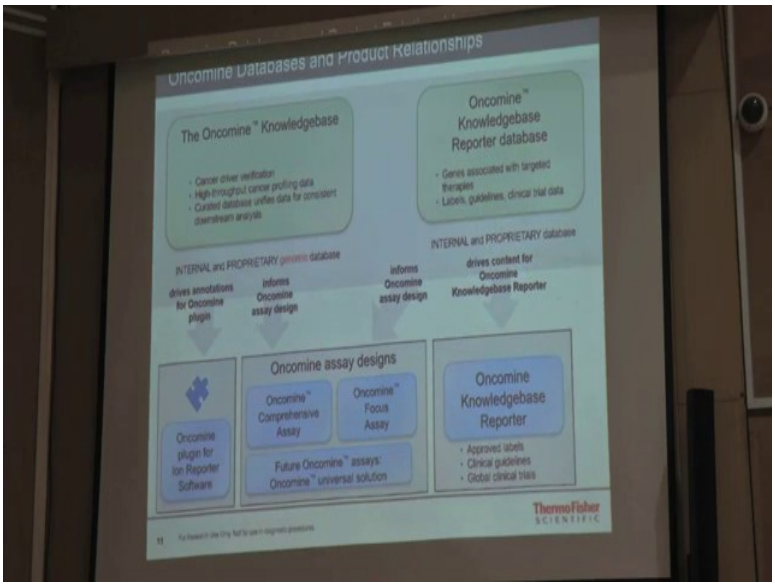
(Refer Slide Time: 18:46)



Now, the current challenge that we have is for specially for cancer samples is once you have identified the mutations of interest or rather mutation, how do you actually go forward, because the ultimate aim is to find the drug or therapy or find the cure for that or medicine for

that. So, right now what we are doing is and these biomarkers are increasing day by day right, so there has to be some software or there has to be a mechanism where you have a database of all the driver mutations and keep them updated, and then utilize in an efficient manner.

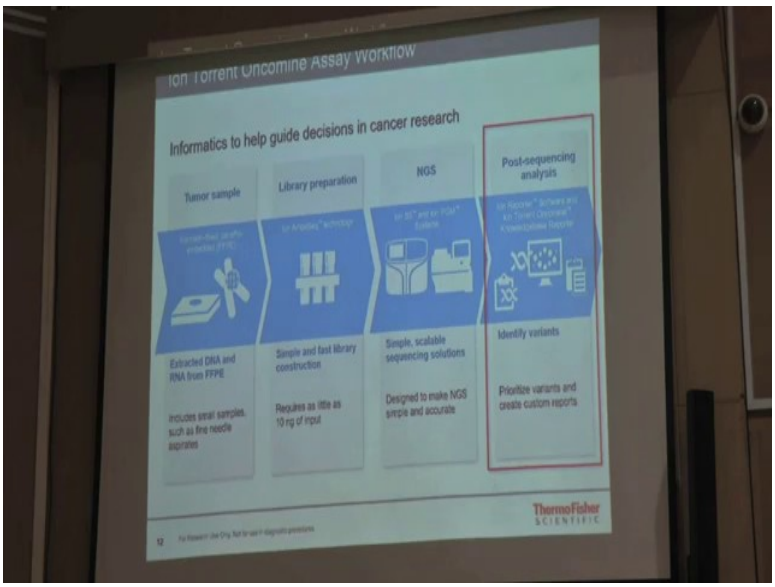
(Refer Slide Time: 19:26)



Now, we have acquired a company called Compendia a long back and we have created a strong database of oncomine knowledge base that has all the details about the genes associated with targeted therapies and labels and guidelines right. Now, we have developed the software called Oncomine Knowledgebase that will basically annotate a variant in your particular sample and it will say whether the particular mutation is driver mutation or not.

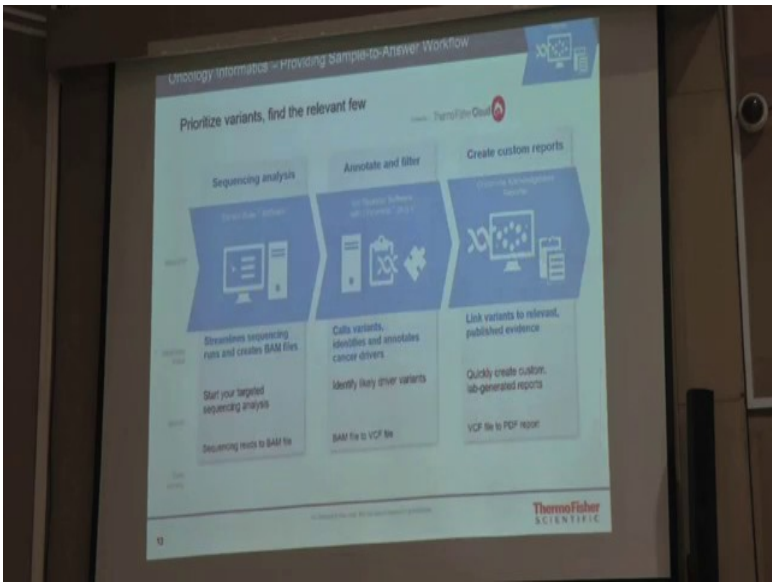
Now, with the use of both oncomine knowledgebase an oncomine reporter database, we have created our oncomine panel. So, the oncomine panel has a list of genes for which we have therapies available or some information available in clinical trials or FDA or Sonova nccn so that.

(Refer Slide Time: 20:18)



When you sequence so you take your tumor sample or cell free sample or your blood sample, you prepare your library, you do sequencing and then you finally, go to the analysis part.

(Refer Slide Time: 20:31)

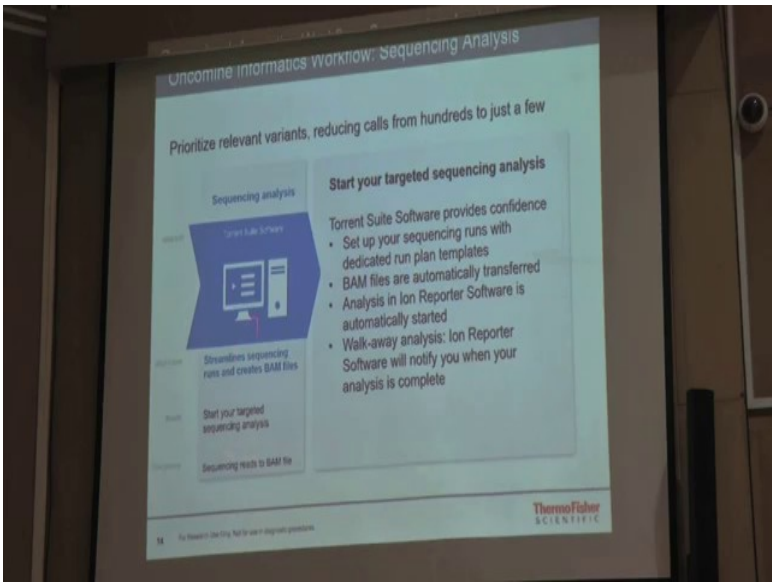


Now, in analysis you do the basic analysis on Torrent Suite software. So, Torrent Suite software will produce a BAM file which is a standard output file for any aligned data. Then we have a software called Ion Reporter. Now, you can do the variant calling on the Torrent

Suite software, or you can do the variant calling on Ion Reporter software that is a choice which user has.

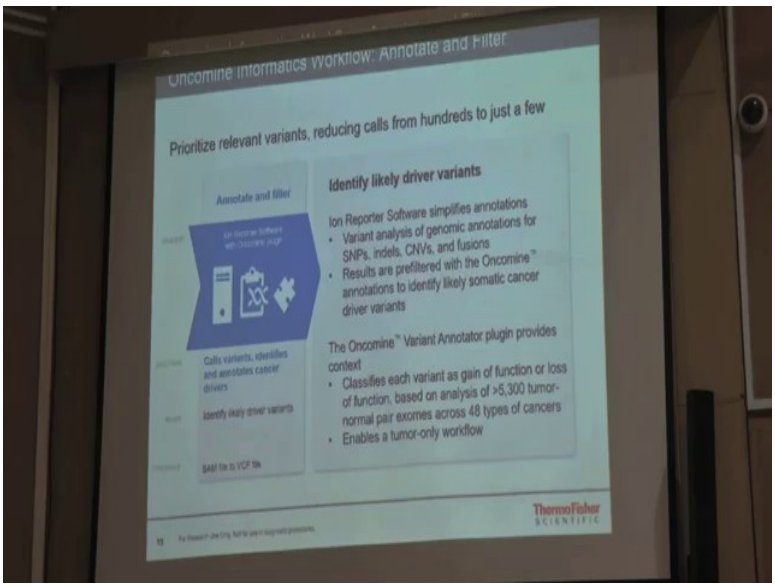
So, variant caller software will take the BAM file, and it will produce the VCF file which is the standard variant file format and then once you have the VCF file, you put into a software called Oncomine Knowledgebase reporter and it will give you a PDF report with all the therapies available for that particular target.

(Refer Slide Time: 21:16)



So, this is the information about the Torrent Suite software.

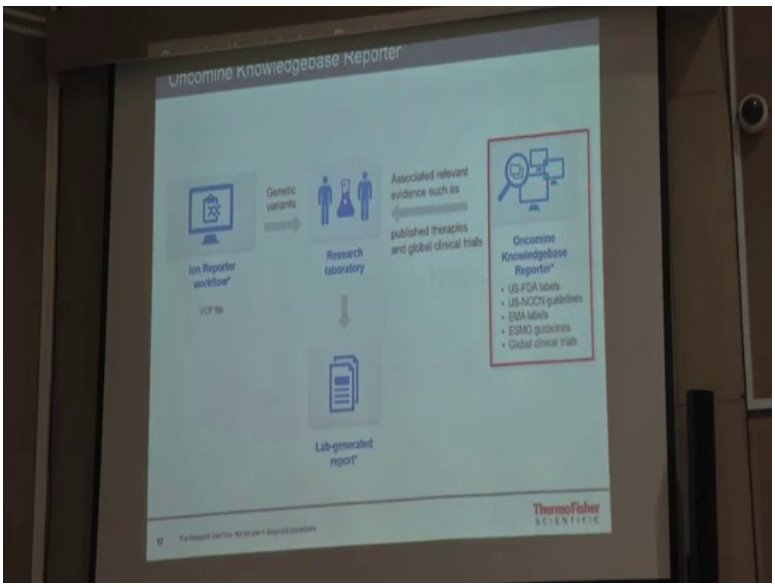
(Refer Slide Time: 21:22)



Same way regarding the Ion Reporter and now this is basically the Oncomine Kowledgebase Reporter software. So, on an Oncomine Knowledgebase Reporter, we have basically labels and guidelines which includes US-FDA, US-NCCN, ESMO, and EMA and then we also have global clinical trials data available in that.

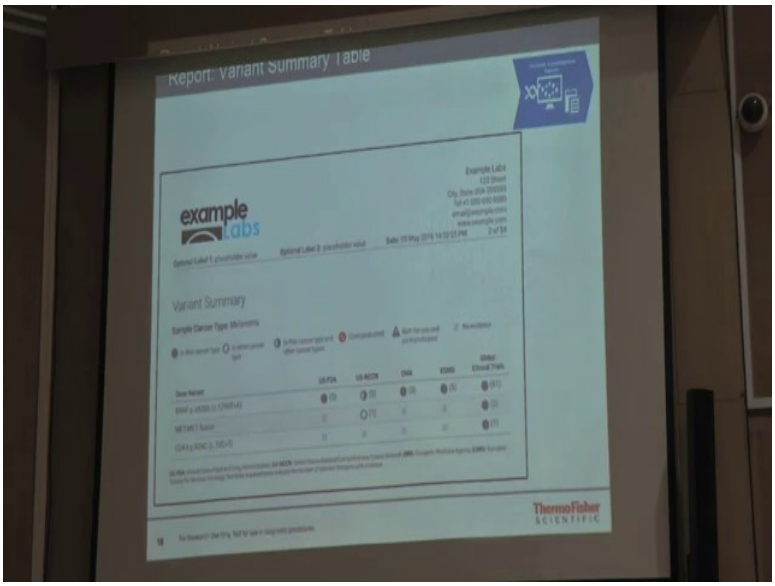
So, this software we routinely upgrade as and when the new therapies and when new levels become available, we update the database so that user has the most up to date database available for the analysis.

(Refer Slide Time: 22:00)



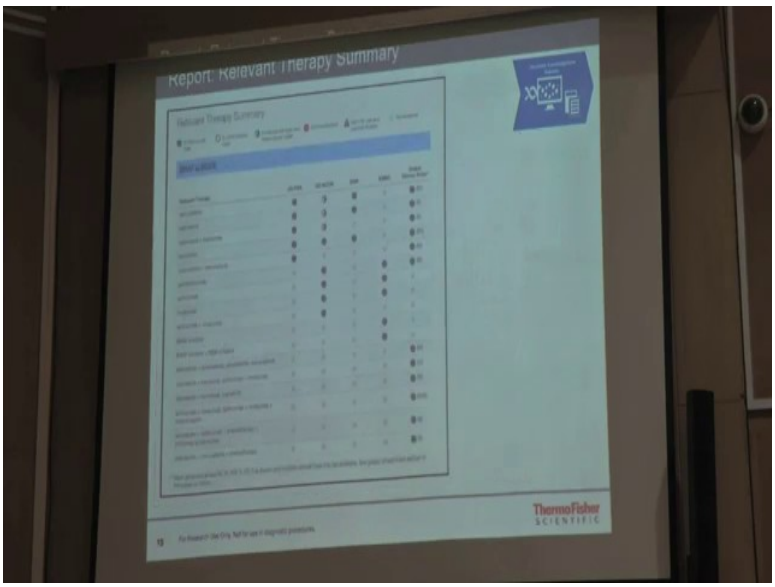
Now, so you have the VCF file from Ion Reporter. You have the Oncomine Knowledgebase report a software, you put them together and then you finally, generate a report.

(Refer Slide Time: 22:13)



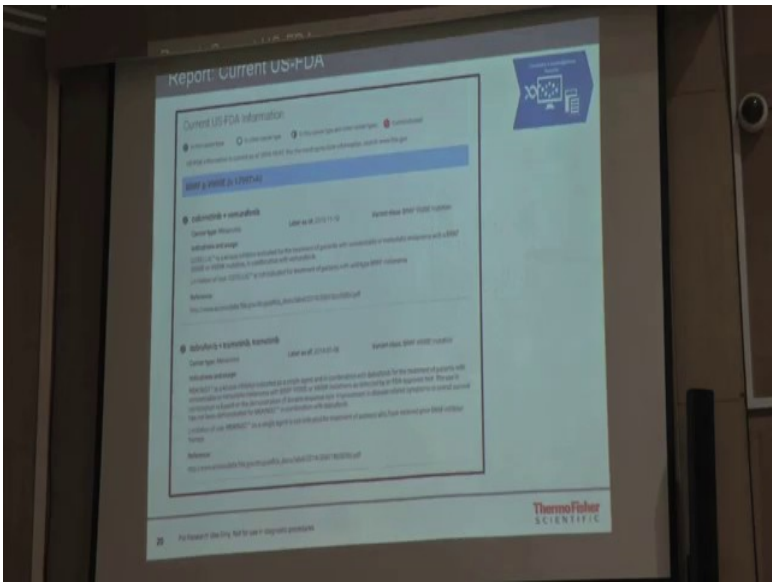
Now, this is an example how the report looks like. So, when you put the VCF file into the software, this is how it looks like. So, this is a mutation, it is a BRAF. How many therapies, labels; therapies are available in FDA, how many of them are available in NCCN, how many available in EMA, ESMO and clinical trials.

(Refer Slide Time: 22:35)



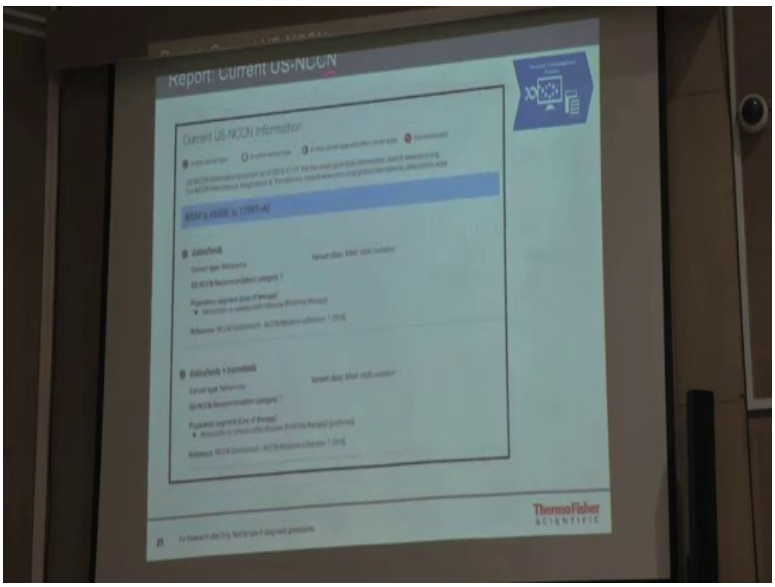
Now, if you go scroll down further, this is how it looks like; there are lot of therapies available for different mutations or driver mutations.

(Refer Slide Time: 22:42)



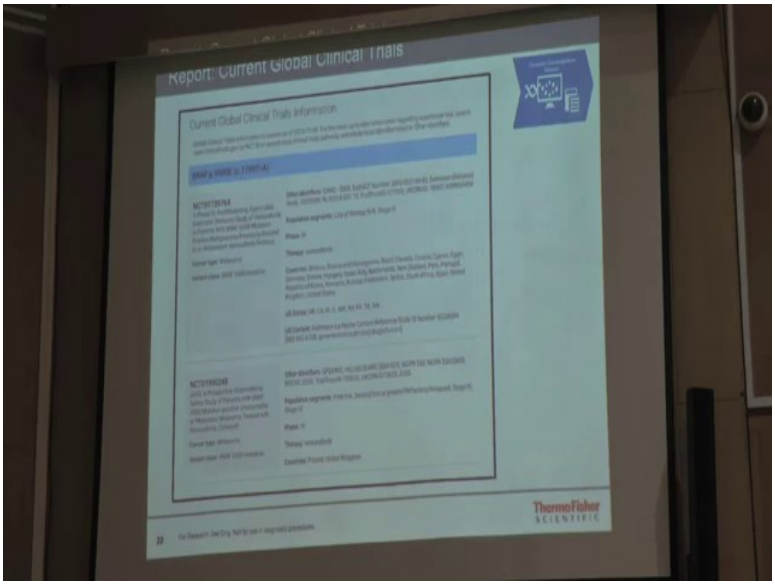
Now, if you scroll further, you will have a detailed view of that. So, under US-FDA, this is how it looks like; for BRAF, V600D mutation. This is the information available.

(Refer Slide Time: 22:57)



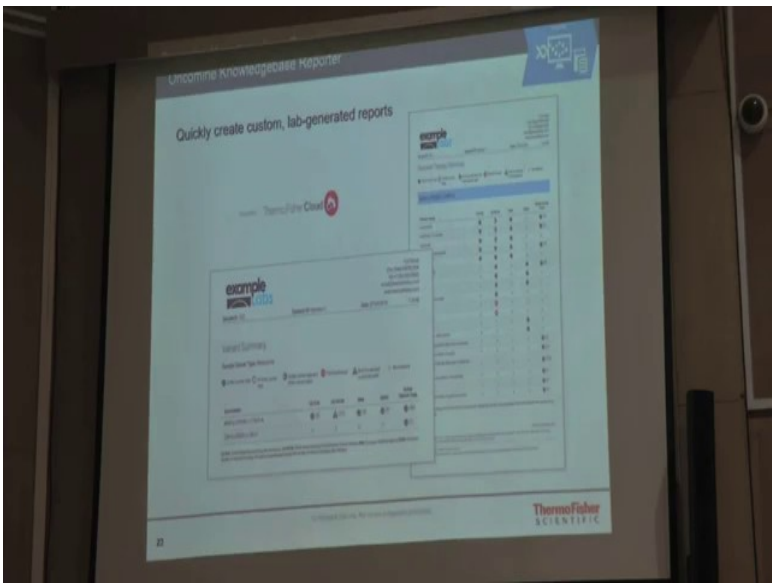
Under NCCN, it is represented as this.

(Refer Slide Time: 23:01)



Under clinical trials, it is represented like this. So, for some of yeah, it is last side. So, at the end for some of the clinical trials, you also have the contact details available at which centers these clinical researchers are; sorry, clinical trials are going on.

(Refer Slide Time: 23:22)



So, this is the; this is a report which normally every clinicians wants to look at or the researchers wants to look at who are working in the diagnostic area. So, this is the entire informatics workflow in a brief for ion. If you have any other specific questions regarding any specific applications we will be happy to answer if you have 2 3more minutes.

Student: Sir, there is a doubt in the workflow. You said that when there is base pair and hydrogen ions will be released.

Yeah.

Student: Hydrogen ions are not released when DNA base pairs

They do.

They do because there is a bond being formed between the

Student: Hydrogen yeah.

OH, and H and that is how a phosphodiester bond is being formed I would have in some of the slide. a phosphodiester bond is being formed and then that is how hydrogen ion is released, it does.

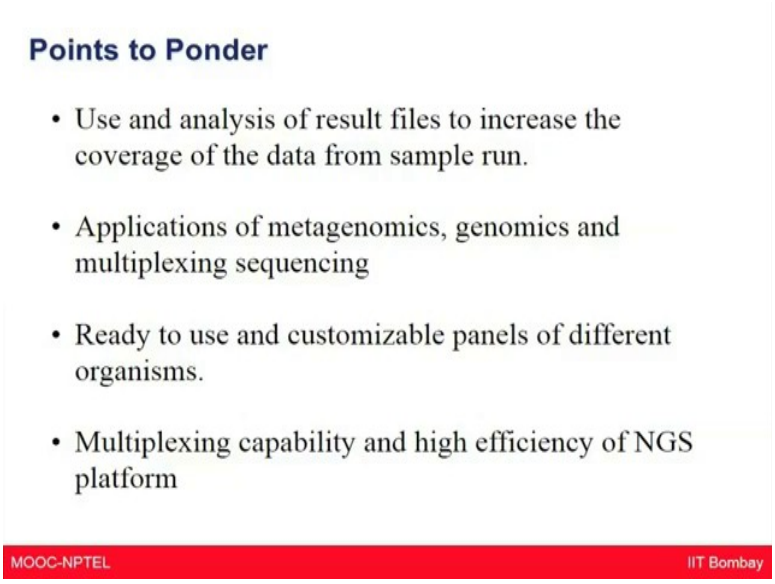
Student: So, you said that for mRNA sequencing you repeat ribosomal RNA.

You did

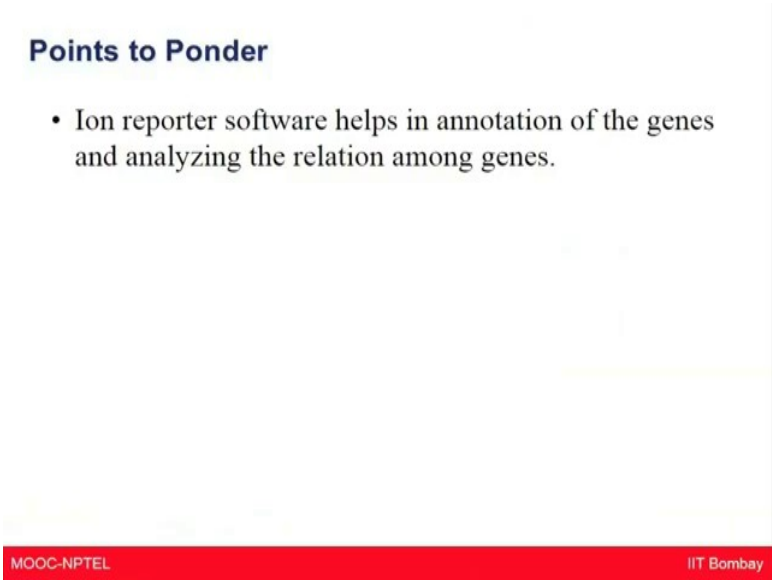
Student: What

You enriched the adenylated portion of the RNA. So, all the RNA fragment, so because mRNAs these are the coding RNAs right. So, and these are known to have A tails. So, what do you do; generally people do is they take a oligo dT column or take a oligo dT beads, and they just fish out all these RNA fragments which have an A tail, so that is how now what the sample which you are sequencing will only have mRNAs.

(Refer Slide Time: 24:55)



(Refer Slide Time: 25:15)



Today, in conclusion, I hope you learnt about thermo NGS platform which has multiplexing capability and high efficiency. The AmpliSeq has capability to sequence from low amount of DNA even up to the 1 nanogram of sample. The LOD of sequencing could be up to 0.1 percent. We also heard the Ion Reporter software which helps in the annotation of the genes and analyzing the relation among the genes. The next supplementary video will be on another application of genomic analysis using droplet PCR.

Thank you.