Interactomics: Basics and Applications
Prof. Sanjeeva Srivastva
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 48
NGS Technology – Bioinformatics and data analysis -I

Today is going to be the lecture in the very exciting areas of Next Generation Sequencing Technologies. I am sure you are all aware about the progress what we have made in genomic technologies. You know in the year 2000-2001 and specially, 2003, the draft human genome map was published and along with the draft human genome project was you know getting accomplished many other model or many other sequences were also finished that time.

And very first time we got glimpse of the you know possible genome sequences available for different model organisms. These projects were really long, you know just imagine it took may be you know 12 to 15 years time to accomplish, sequencing of you know one individual and one model organism. After that looking at its success looking at the impact of genomic technologies, a lot of innovation has happened.

This is one of the such an integrated area and I must say you know you can appreciate how biologists and technologists and clinicians can benefit from each other is one of this you know, area when genome sequencing information. Really you know triggered interest of engineers to come forward and make the new of technologies which are much more rapid, much more robust, much more reliable, much more reproducible. And those have resulted into series of next generation sequencing technology from first, second and the you know third generation technologies.

So, these technologies interface have really helped us now to move forward for what was accomplished in 10 years to maybe you know in in a day or 2, you can now do the sequencing. So, this is as I said you know these are the kind of technologies, sometime you know for a revolution to happen you have to wait for you know decades and you have to wait for centuries right. But this is kind of technology which actually happened in front of my eyes in during my career. And I am sure you know some of you would have also seen and

witnessed the kind of progression we have of you know the old (Refer Time: 02:16) sequencing based methods and moving toward the very fast rapid next generation sequencing platforms.

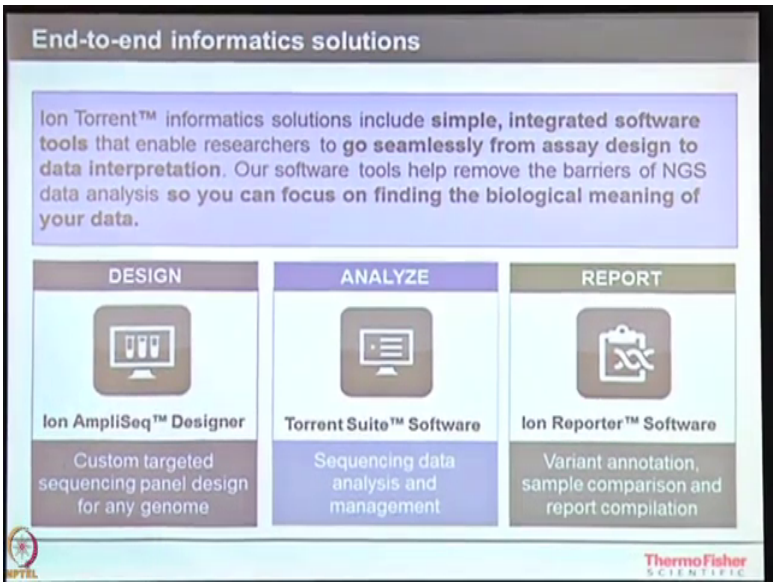
This is really interesting area. I think it is good idea for us together to really something that what are the current available technology platforms which can use to do the sequencing in a very high throughput manner. And let us also think about how best we could use these for many applications. So, in this light we thought to provide you couple of technology overview as well as some of the applications from series of application scientist. In this series, today we have Mr. Pravin Nilawe, a field application scientist from thermo fisher who is going to talk to you about the ion torrent informatics solution for NGS data analysis.

So, let me welcome Mr. Pravin to talk to you about this novel technology platform of ion torrent and he is going to then talk to you about the data analysis and applications.

Good morning everyone. So, you must have gone yesterday through the ion torrent technology right. You have seen about sequencing yesterday. So, we have got an idea how this system works, it is about signals, it is about recording the data in few minutes within seconds and then you are going forward with sequencing and recognizing which are the bases that are getting generated right.

So, you have seen lots of steps in doing sequencing. Today we will take it somehow through the steps of understanding those sequencing and how it is interpreted into the results ok. So, I will take it point by point what it is actually. So, we can move forward. So, what we are looking at into NGS technology or NGS sequencing.

(Refer Slide Time: 04:12)



We are looking into something like where we will have end to end workflows or end to end analysis to be completed right. So, you must have heard yesterday about something called as target sequencing. yeah It is something like your genomics is targeted for a particular region or it may be your genes that you are studying or targeted for a particular region through primers and you then pick them up and go forward with sequencing ok.

So, once you start sequencing, that is the first part what we look for; that is the design. So, you are targeting particular regions for particular variants. Now you must have heard why are we going to target those particular regions. So, important part is; you may be studying certain hereditary diseases. You must be studying something cancerous diseases right. You must be looking for something genetic disorders into it right.

So, those genes which are very much important for you, you like to know which are the regions where exactly those variations are happening or those are the mutations you want to capture in your study and know at what level those mutations are happening into your cells or into your samples right.

So, for that thing only you like to target those regions right and take them further for sequencing hm. So, what happens over here is? You have a technology called as ion AmpliSeq designer as the first stage, which helps you to design those regions which can be utilized for sequencing and NGS over here. So, and thus we have in three stage; one is the design, second is the analyze and third is something called as a report ok.

So, you try to design your regions of interest which you are looking for your diseases, say hereditary diseases. You take those design for particular genes, you run your samples onto your system called as torrent suite or as ion torrent and the software that does the analysis for you is called as torrent suite software. In this software, it will help you to understand what are the variants that you are getting. And at what level or which level of coverage are you getting in this data.

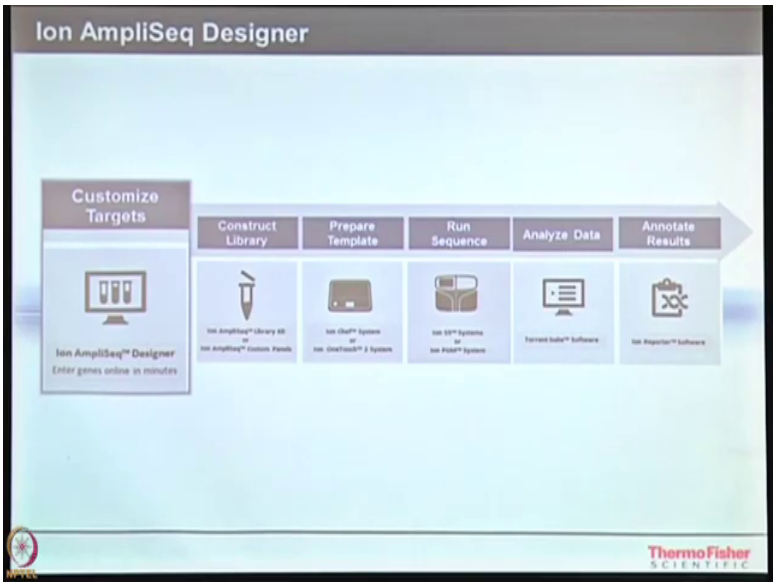
So, you are looking at torrent suite software where we are taking all the variants in your hands and trying to study them. So, with this variants where you are just getting to know that there is some change happening into your gene, you also want to interpret them in a way where you can understand it what exactly is happening to the protein level or what is actually happening at this sample level right. So, you will do is last part is to report that information through various databases. Have you heard about OMIM? OMIM.

Student: (Refer Time: 07:06).

Right. You have heard about db SNP; database of snips Single Nucleotide Polymorphisms. Have you heard about anything about cosmic databases which actually hold your cancerous variants ok. These are some words that are coming over here. So, these are very much important when you try to study certain different diseases or disorders into your analysis.

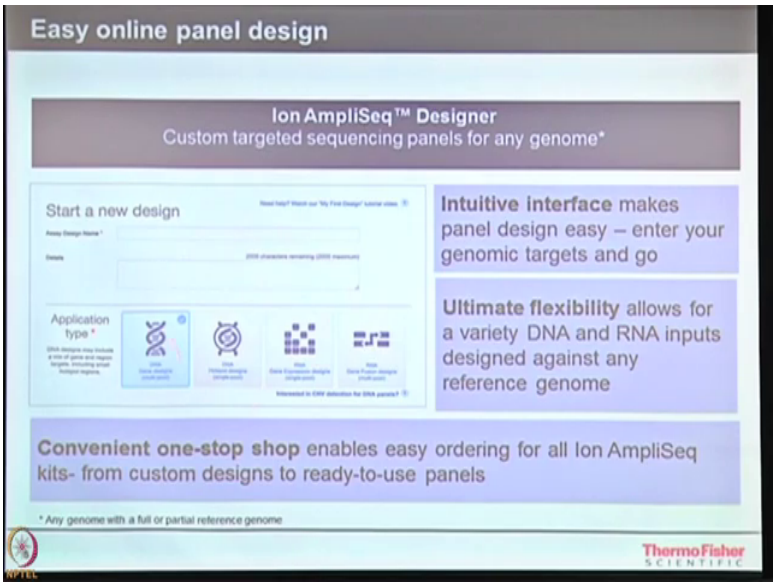
So, now what we are doing it over here is you have three stages I will take each stage individually and explain you as such. So, the first part is Ion AmpliSeq designer. So, what we have is a complete workflow of doing the analysis where you like to know which are the regions of interest of yours whether, I could design my regions in such a way that they could be sequenced on a NGS technology ok.

(Refer Slide Time: 07:55)



For doing that we had a tool we have a tool called as Ion AmpliSeq designer. It helps you to take all your required designed genes in as a list of genes or even the region of interest from your chromosomes ok. It takes that and helps you to design primers which could fit into technology of ion torrent sequencing ok.

(Refer Slide Time: 08:34)



So, you can design something of your own ok. Something like you what you do on your websites, normally you create your own accounts. You can create your own account over here, where you can also give the design name to it. And then what you can do is you have certain application types available. So, you can see you have the first option as DNA over here. So, this is like a gene base design that you can do, where you can provide just the HUGO Nomenclature genes over here.

So, you must be knowing about EGFR, if you have any idea. You know, knowing about some BRCA genes. BRCA 1, BRCA 2. So, these are some of the genes which are studied a lot into the entire world ok. So, they are studied for the different variants. They are studied for their different purposes where they also want to come down to a place where they can get to know which drugs are acting on regards of cancer in regards of your hereditary disease ok.

So, at the same time I have something like the gene design. I also have something called as hotspots. So, I was talking about something called as cosmics or db SNP s ok. So, these are variants which are already known or people have already studied, researchers have already known these are actually a deleterious nature variants that are coming out which are having effects on to your particular positions or group of patients ok.

So, if I know those variants and I would like to check it in Indian population I like to design a panel in such a way that I will use those hotspot information over here and I could apply it overall for all the population in India. So, I may have 100 to 500 samples, I like to test them overall and get to know whether the same variants are actually falling into your data or not or India into the Indian population or not right.

So, this could be very easy for one study, that is called as pharmacogenomics right. Your one gene can bring you results in such a way that your, it could easily tell you which therapy could be properly utilized for particular patients right. So, over here we have an option of DNA hotspots which takes such type of information. It may be a chromosome single location such as an SNP, it could be a deletion of a bigger range. So, in that way we can provide the information over here is hotspots.

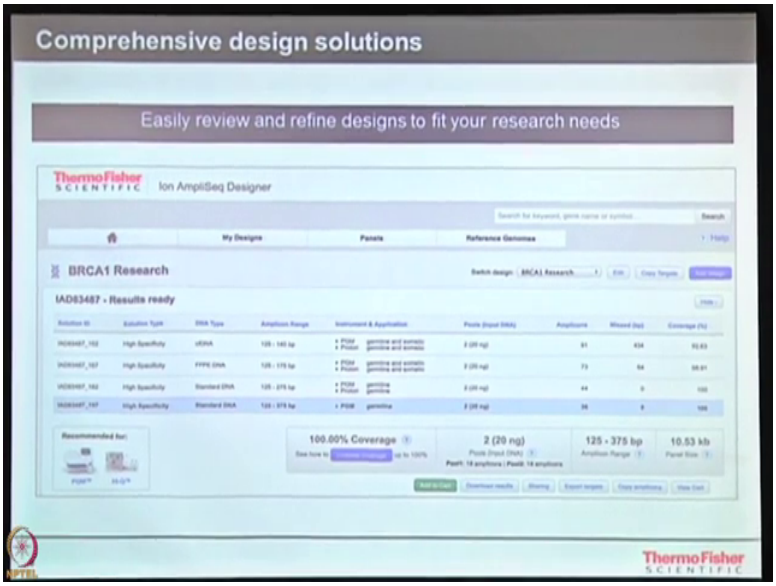
The rest other two things that are available one is gene expression. So, you must have studied about RNA seek. Have you heard about it anytime or something called as whole transcriptome. So, this is nothing, but a study of genes which are expressing into your samples or you can say if you are considering something like you have cancerous patient and as normal patient. So, you like to know which are the genes that are expressing into a cancerous patient which are very much different from what the normal patient is having right. So, such studies come into play when you talk about whole transcriptome sequencing ok.

There is a different technology. At the same time we can do same gene expression analysis into this place, where we like to target those genes which are actually into your expression studies ok. And we like to know which actually highly expressing which are actually low expressing. This could easily take you to the pathways which are affected due to these

regulations ok. So, this is one of the study and the last one is something called as gene fusion where two genes can fuse at a particular location at a point and there could be a protein change happening in to it.

So, such study or such type of designs could be made ready available which could taken up further for sequencing onto ion torrent suite technology ok. So, I will just go move forward. So, what happens with the design. So, just give you a small example of BRCA gene. BRCA 1 gene which has been designed over here.

(Refer Slide Time: 12:30)



So, you can see BRCA 1 gene and then it says what is the DNA types that we want to go for. Now what happens in to your data? You can get data primers design based on different types of samples ok. Now, whatever you do you can be using like you must be studying blood type of samples. Like, normally what your lab test does are you take only the bloods, you take the

blood samples you like to take all the genomic data out of it. Target your particular gene and then do the studies further right.

So, such type of analysis could be done. We can get a design ready for it and in that I get to know how many amplicons are getting designed for a particular gene oh my BRCA 1, I think the design looks like you have 36 amplicons that have been designed over here just for your DNA sample ok.

So, such type of samples or such type of variations is available over here just to know what can be done. There is one more type of sample that comes up into cancerous studies. Have you heard about FFP sample anytime? yeah These are actually formalin fixed samples ok. So, these are actually whenever you do cancer tests or somebody is detected with cancer these samples are actually fixed on a hm

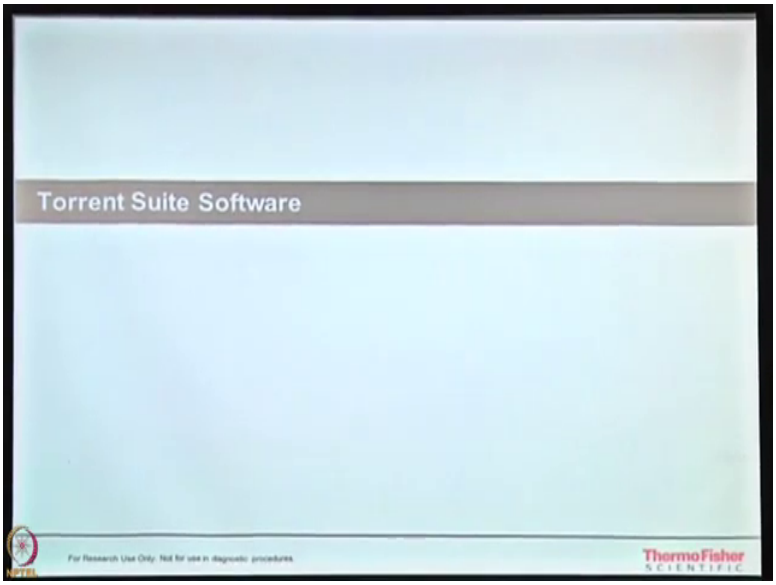
Student: Paraffin.

Paraffin right and then sent across to labs for testing. So, these can be easily picked up. So, you wanted to know what are the variations or mutations that you are getting into it. So, such studies could be done again using particular design over here. So, what happens this tools helps you to differentiate between what are the DNA types? What are the type of data that you want to go utilize for doing your studies? What type what type of targeting that you can do in such type of studies ok.

So, once this design is ready you get to know how many amplicons, what type of sequencing I could do like the technologies defined based on my sequencing range also. So, you have something like for FFP it is 125 to 175 base pair. So, I have utilizing a technology or sequencing where I could do 200 base by sequencing and there is also 400 base pair sequencing which I could take further with my blood samples also. So, this helps you to easily get to know what things has to be utilized, how could I run the data onto an ion torrent system and you do further take it further for the analysis level.

So, you got to know about the information how designing is happening right. How the panels are getting design, your genes are getting picked up and taking it further. Now once your genes are ready your panels are ready you like to take it further and sequence it and we will put it on to the system. You will first do all the amplification. Yesterday you must have gone through all the steps for whatever the amplification happening. You are taking it bar coding it and then taking it onto the chip and running it and your reports get generated finally, all right.

(Refer Slide Time: 15:20)



So, once your reports are generated the report analysis work is done by this software, torrent suite software ok. It understands whatever data is generated by signals ok. Whatever voltage chains are happening are recorded on to the systems and the software understands to understands it signals, clears the signals or filters this signals over there and then decodes them into particular bases. So, you have a chip has millions softwares and in million softwares you have millions of signals getting recorded. And the same signals are decoded

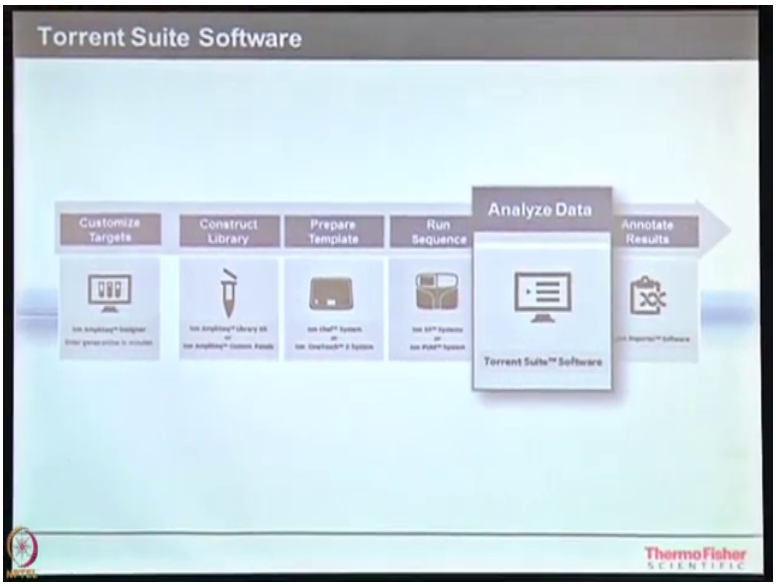
into your basis and giving you the entire to read length a bigger read length sequence over them ok.

So, you get sequences which are around 200 base pair 400 base pair or it may be only higher read length of 500 to 600 base pair also ok. So, once you have done sequencing, what should we do further? We have got the raw data right. See this is the raw data comes into various formats; one is fast queue, you must have not known about it or you may be having an idea about it.

Fast queue files, which is a raw data file containing sequences as well as the quality values for it and there is something called as bam file also. So, this is just for your knowledge bam file is one where if you do certain like aligning to the genome your data is generated and you are aligning to the genome you get these bam files which contains all the coordinates for those genomes.

So, you align it, get the coordinates which are the chromosome where is the actual alignment happening which position and whether it contains proper alignment or there are any mismatches or deletions into it insertions or deletions into it right. So, everything is recording through the software torrents suite software. The tool that does this mapping for you is called (Refer Time: 17:20) map ok.

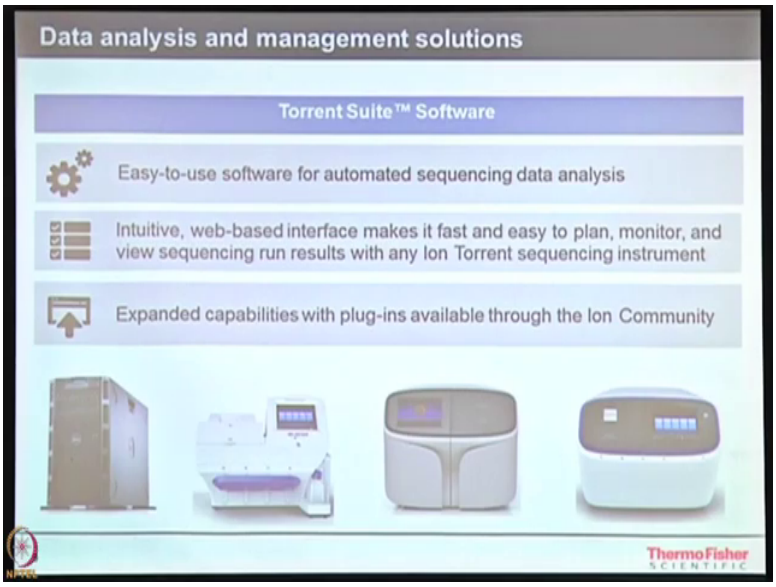
(Refer Slide Time: 17:24)



I have not put much into this, I will just take to this specific level of analysis that we do. So, once the system internally does the alignment with a reference you know see if you are running a particular BRCA sample and you do a run and you have a reference human genome we align it.

So, you like to know what is happening behind it right your genome is getting aligned, your data is getting properly aligned to it, but how? How much is the data getting aligned to it right? That would be your first question. So, if I have generated around 2 million reads for that sample. How many reads actually aligned to those region which are interest of yours right.

(Refer Slide Time: 18:02)

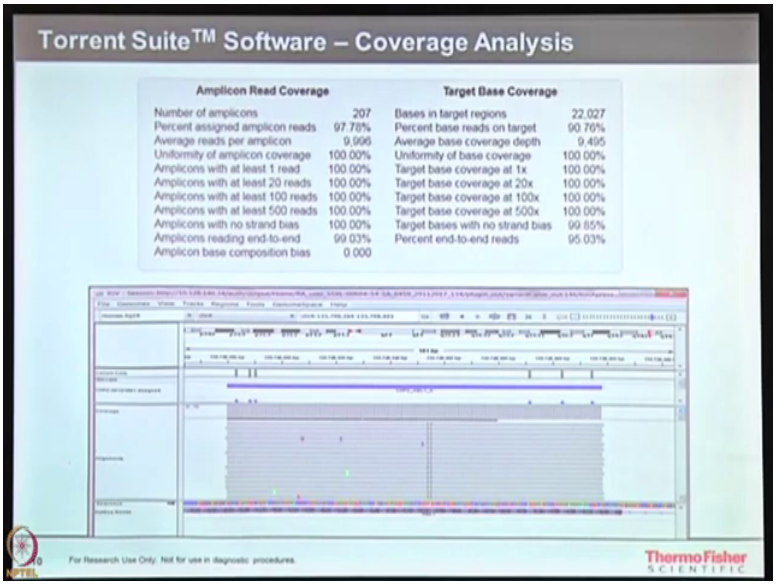


So, for that we have a tool called as sorry, I just explained a lots of thing of the torrent suite software just to take it a you know shorter way, you have something like automated sequencing data analysis, you have something like the interface is through a web based interface where you like easily can go into like a website and go through the runs, reports, download the reports in pdf run different plug ins that are available for doing analysis ok.

So, as I come back to what I was speaking; we have done data analysis onto a reference genome aligning your data to that reference, but you need to know something more about it. Whether it is actually representing your data or not. Actually, it is going aligning to your regions or not of interest right. So, that is my main important points. So, over there are helpful whatever plug ins are there those coming soon and to play over here ok.

So, we have certain plug ins which helps you to understand, if the region of interest that you have designed I could provide my information over there. Say I have BRCA 1 region, I have designed it. I wanted to know how many genes or how many reads are actually aligning to those regions ok.

(Refer Slide Time: 19:21)



So, how can I do that? So, I have a plug in called as coverage analysis ok. This coverage analysis helps you to know if any region is getting aligned with number of reads, it let us you know how many reads are actually mapping to it.

So, what happens? So, it says, if I have a region, say this is an example for a cancer hotspot panel which has around in the target region has around 22 k basis, 22000 bases into it. So, I just wanted to know how many percentage of bases are like, how many reads of mine or how

many bases of my reads are actually on target aligning to my BRCA region or like CHP panel region.

So, it is 90 percent of my bases are actually aligning to my region of interest in con hotspot panel ok. With that it let us me know which is the base depth coverage across it. This is like if I m aligning my reads or my sequences to a particular region of interest, how many reads are overlapping in those particular region.

So, what is the mean depth across those particular region right. So, I need to know if a read has been covered by how many reads are covering a particular base over there. So, in that you have around 9495 mean depth ok. It is an average depth covering a particular base over there. So, you may be having a range of around 8000 to 10000 into your CHP panel that have you designed or it may be your custom panel that you have designed and then you get to know this particular depth.

Now, this gives me an idea whether my panel properly having all the reads or not right. If I am having a depth of 1000, 2000 it is still good. So, now, over here as an example, it is very high and that I will also like to know whether my region of interest say I have 22 kb of my region of interest. In that whether if I look for 1x coverage, how many bases are covered by 1x coverage? Ok.

At least one read covering each base, that is what I could say is as one x coverage. So, in that you have around 100 percent bases which are covered at 1x. At the same time I like to know how many bases have covered at 20 x that is 20 times a particular base is covered properly or not right. So, 20 times a rediscovering a particular base. So, at the same time its 100 percent. Till the 500 x you can see its 100 percent.

So, this shows me whether my design is fine or not, whether my each base is getting covered properly or not, whether this data could be taken further for my variant analysis or not. Where my variant of interest could would be properly picked up or not. So, this gives me an overall idea. At the same time I have something like end to end reads. So, if I am designing something like your gene ok, I am taking my gene, designing a particular primers for your

exonic regions one end to the other end. So, whether it is covering end to end or not. So, how many reads are actually covering end to end sequencing into it right. That gives you a confidence for your primers also.

So, your primers which have been designed. So, it shows me 90 percent of my total reads are actually having end to end alignment to my region gene of interest ok. So, that gives me more confidence whether my data is coming good or not ok. So, any questions till now? Yeah.

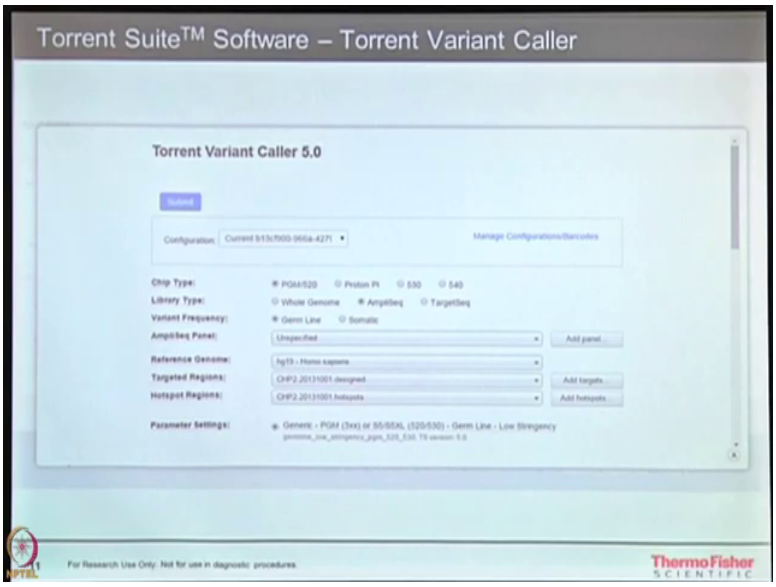
Student: So, does this 1 x coverage essentially means that single base resolution.

It is not single base resolution, it is like when I when you are taking a reference genome you are aligning a sequence I am just trying to know whether a single base is covered by one read or not ok. So, I am aligning alignment is nothing but you have a reference, you have a read getting aligned to it ok. There may be a thousand reads that are getting aligned. So, I am just trying to get to know whether that particular base in genome is getting covered by how many reads.

So, if I am having something like one read passing to it. So, it is covering that particular base ok. So, that is the first stage actually, once that is done it is taken for then aligned to the reference genome. So, in alignment what happens we like to see, how many reads are aligning to a particular region and the same thing I like to calculate over here. We need some statistics right. We cannot go and visualize the data every time. There is a way to visualize it the tool is called as IGV. So, this is called as IGV; Integrated Genomic Viewer. It helps you to visualize your data how much data has been aligned to that region, how many data is for other regions.

So, you can go scroll through genes to genes and get to know that ok. This is just an example that I of put forward one of the gene which has been aligned by the reads it has the forward reads as well as the reverse reads into it. So, in this way we like to know whether the regions are getting covered or not. Whether there is a proper coverage coming at my gene of interest or not ok. So, after this, what I will be going to do forward. So, I have another plug in called as torrent variant caller ok.

(Refer Slide Time: 24:46)



The torrent variant caller is optimized for to calling all my variants. So, my variants could be SNP indents right. So, these variants could be called using this software, torrent variant caller. It haves features into it such as I could give the chip types since the system over here right now is I nx ins 5. We still have two more systems available; one is proton and one is PGM ok. So, after that you have almost similar of workflows. So, you can go for CHP panels while as I was saying you have different panels available for cancers, you have panels available for hereditary diseases.

So, with that what I get to know as there is something called as design files ok. So, whatever design that you create on AmpliSeq those files could be downloaded in a particular format called as bed format and these files could be uploaded for doing the analysis in variant caller ok.

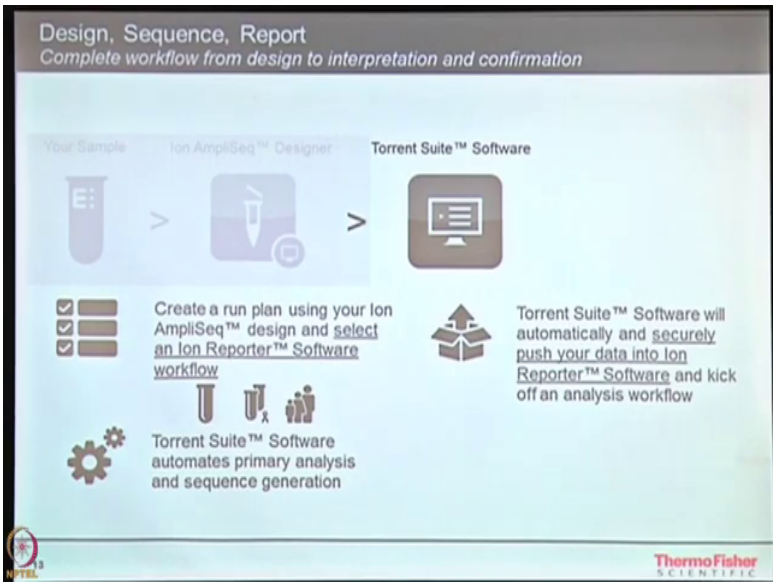
Once my design is uploaded variants would be only called into those regions which we have designed, it will not go and look for some other regions ok. Once those designed region is given it also has something called as hotspot I had spoken about earlier about designing hotspot regions also where I could like to know whether my any of my known variant is present or not.

So, I could give a file called as hotspot bed file in a format such a way that it will recognize that particular position into my data, into my variants. And it will then represented that yes this is the hotspot that I was looking at. It will provide you a final results in an excel sheet ok.

So, what you do is further I give forward my designs my hotspot regions of interest and I could give certain parameters to call the variants. So, variants could be called base (Refer Time: 26:37) nature or else my (Refer Time: 26:39) nature. 50 percent frequency or else 5 percent frequency and then the variants could be called at that stage ok. And once I run this plug in I could submit the data and once I run this data I will get all the variants data generated into it ok.

So, what happens? The variants could be downloaded for all the SNP s index into an excel sheet entirely. So, you can take that and study it further.

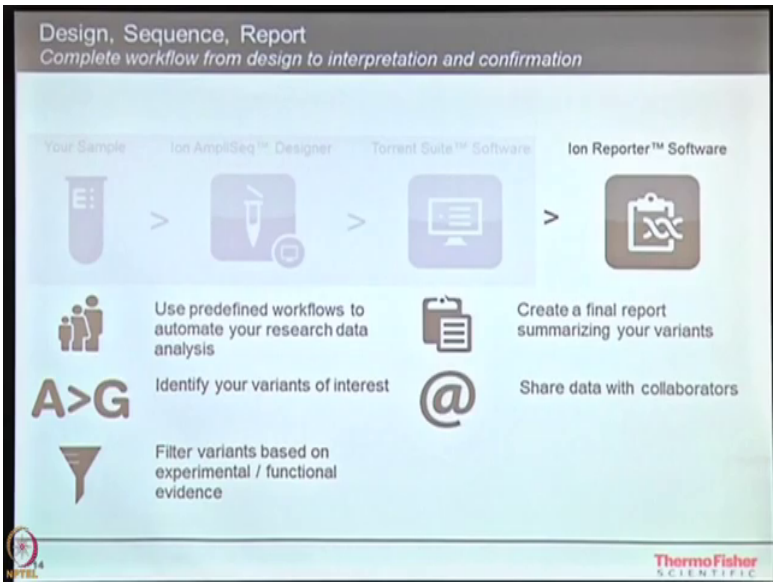
(Refer Slide Time: 27:20)



So, now what is happening over here is. So, I will take a step by step mode, I first did designing on AmpliSeq right. Then I did the torrent suite software where it decodes all the bases and provides you the sequences for it, it provides you alignment with the reference genome, it does the coverage analysis for you which looks for the regions which have been designed your interest gene of interest and then take it further and do variant calling.

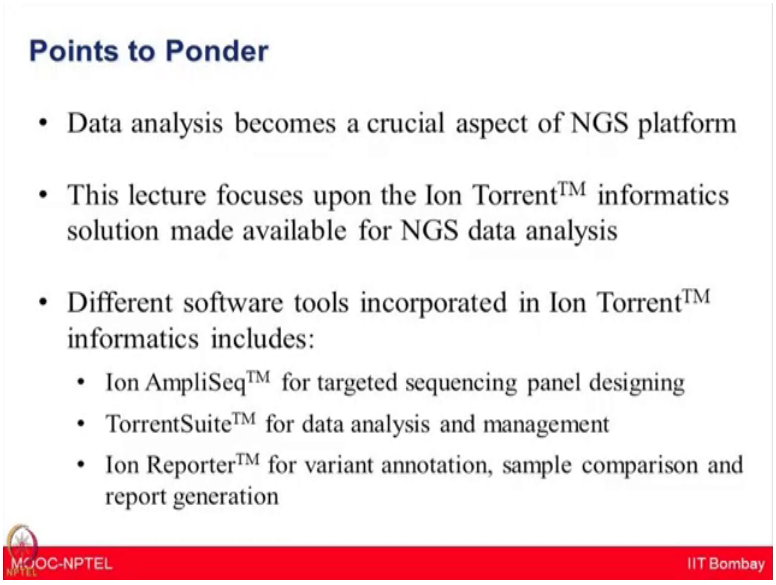
So, once we have done with the variant calling, you just have variants with you right. You have just the SNPs giving you the change from a to t c to t or just the deletions a is deleted or t is deleted. But you still need to know something more about it where exactly this is happening with gene it is happening, whether it is actually having deleterious effects or not, whether it is having any effects with the patients or not right. So, you need to know something more about that.

(Refer Slide Time: 28:09)



So, for doing that there is one more tool called as Ion Reporter Software. So, it has lots of information in built into it. So, this helps you to correlate your variant with the information that is already stored into databases.

(Refer Slide Time: 28:24)



So, today you learn and at least got a glimpse and some understanding the basics of NGS platform and data analysis. He also introduced to the Ion AmpliSeq designer, the Torrent Suite software and the Ion Reporter software. Well, some of these are very specific to a given technology which is not the mandate of the course to teach you how specifically these softwares work, but rather you know by showing these kind of available platforms intention is to give you an overview and a good understanding that how these technologies and these softwares could be used for your applications.

So, in today's lecture; you also got understanding about how to visualize the data and interpret the NGS data. Usually, when we are able to obtain this kind of big data set I think it is really important to look at data in (Refer Time: 29:38) systems wide manner right. And the

big data being generated and your intention is also to integrate the data and compare data from other systems as well.

So, this light you know having dedicated server and high computing systems can definitely help to do the analysis in a much more rapid manner. And, that could also provide us to do lot more things which one could try to do now comparison aligning with the reference genome and many other you know type of multi omics analysis can also be performed.

So, I must say that you know one thing which is limitation is our computing power the way we can process the big data simultaneously for you know large number of samples as well as different type of information obtained at the gene level and the protein level or MRNA level and trying to correlate all that information together we need really the highly computing power and lot of you know space to do these kind of analysis.

So, next step is to do variant analysis which can be also done using a cloud based tool. And I am sure you know in the as we go along in the next lectures you will specifically study how to use this data form or application orientation. Especially, in the context of cancer. So, we will continue more on this sleight of NGS and its revolution and we will talk to you again in the next lecture.

Thank you.