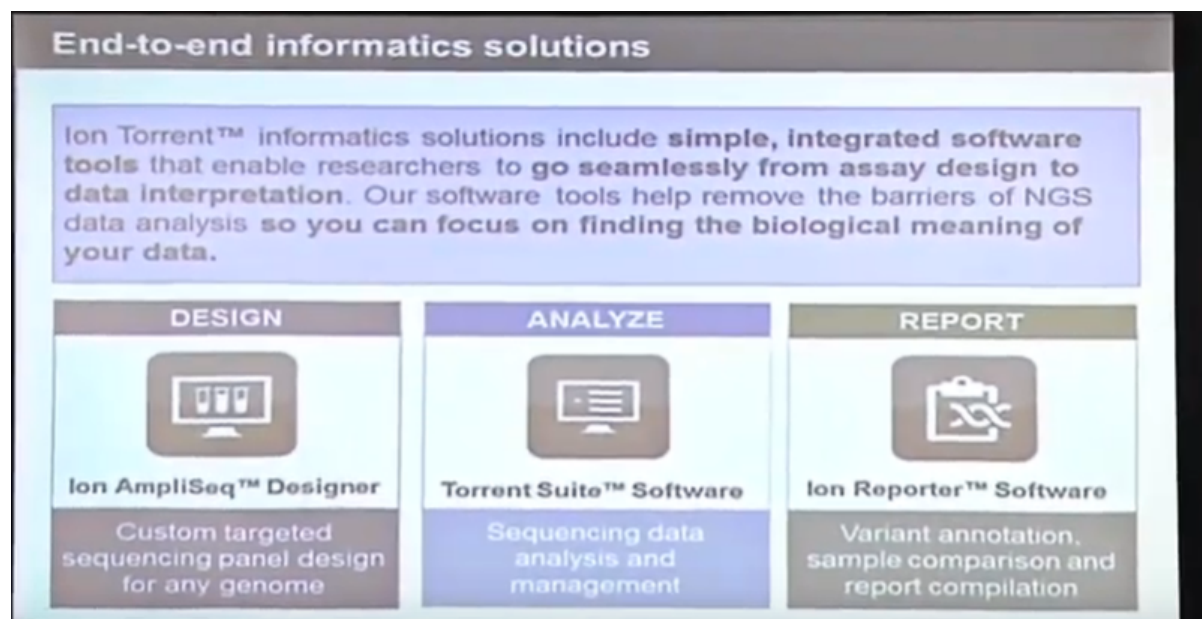


## **Lecture 33**

# **NGS Technology- Bioinformatics and data analysis-I**

Welcome to MOOC codes, on applications of Interatomics, using genomics and proteomics technologies. Today is going to be the lecture, in the very exciting, areas of next-generation sequencing technologies, I am sure, you're all aware about, the progress that we have made, in genomic technologies, you know, in the year 2000, 2001 and special 2003, the draft human genome, map was published and along with the draft human genome, project was you know, getting accomplished, many of the model organism 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 me, 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 such an integrated area, then 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 in two series, of next-generation sequencing technology from first, second and the third generation technologies. So, these technologies interface, have really helped us now, to move forward for, what was accomplished in ten years? To maybe you know, in a day or two, 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'm sure, you know, some of you would have also, seen and witnessed, the kind of progression we have, of you know the Sanger sequencing based methods and moving toward the very fast rapid, next-generation sequencing platforms, this is an interesting area, I think it's good idea for us together, to learn something, that what are the current available, technology platforms, which can use to, do the sequencing, in a very, high throughput manner and let's also think about, how best we could, use these for many applications. So, in this slide, 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. Praveen Nilawe a field application scientist, from thermo Fisher, who's going to talk to you about? The Ion Torrent, informatics solution, for NGS data analysis. So, let me welcome, Mr. Praveen, to talk to you about this novel technology platform of Ion Torrent and he's 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 you have got an idea, of how the system works, it's about signals, it's about recording the data, in few minutes, within seconds and then your 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. Okay? So, I will take it point by point, what it is actually? So, we can move forward, so what we are looking at in to ingest technology, ingest sequencing,

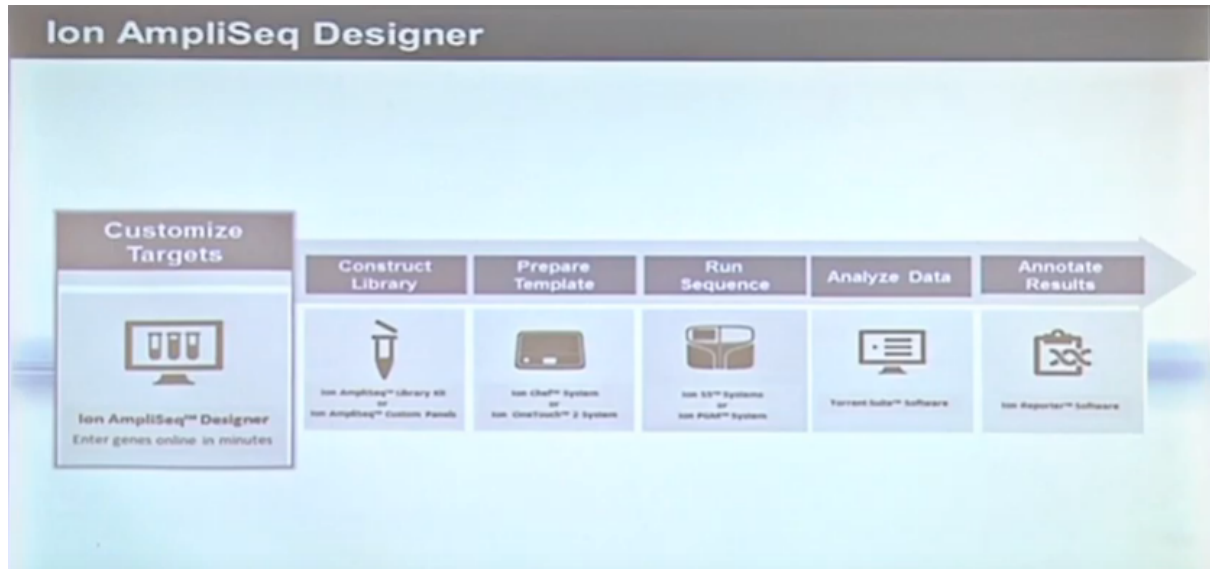
Refer Slide Time :( 4: 31)



we are looking into something like where we'll have, end-to-end workflows or end-to-end analysis to be completed. Right? So, you must have heard of yesterday about something called a targeted sequencing? Yeah! Does something like your genomics, is targeted for particular region or it may be your genes: that you are studying, are toggle it, targeted for a particular region, through primers and you then pick them up and go forward with sequencing. Okay? So, once you start sequencing: that's the first part what, 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. 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 in NGS, oh, yeah! So, in thus, we have in three stage. One is the design, second is the analyze and third is something called as a, 'Report'. Okay? So, you try to design your regions, of interest, which you are looking for your hair diseases, say Herod Italy diseases, you take those designed for particular genes, you run your samples onto your system, called a, 'Torrent Suite' or else ion torrent. The software that does the analysis for you, is called a, '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're looking at, Torrent Suite software, where we are taking all the variants in your hands and trying to study them. So, with this variance, where you are just getting to know that, there's some change happening in pure 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 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. Okay? 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?

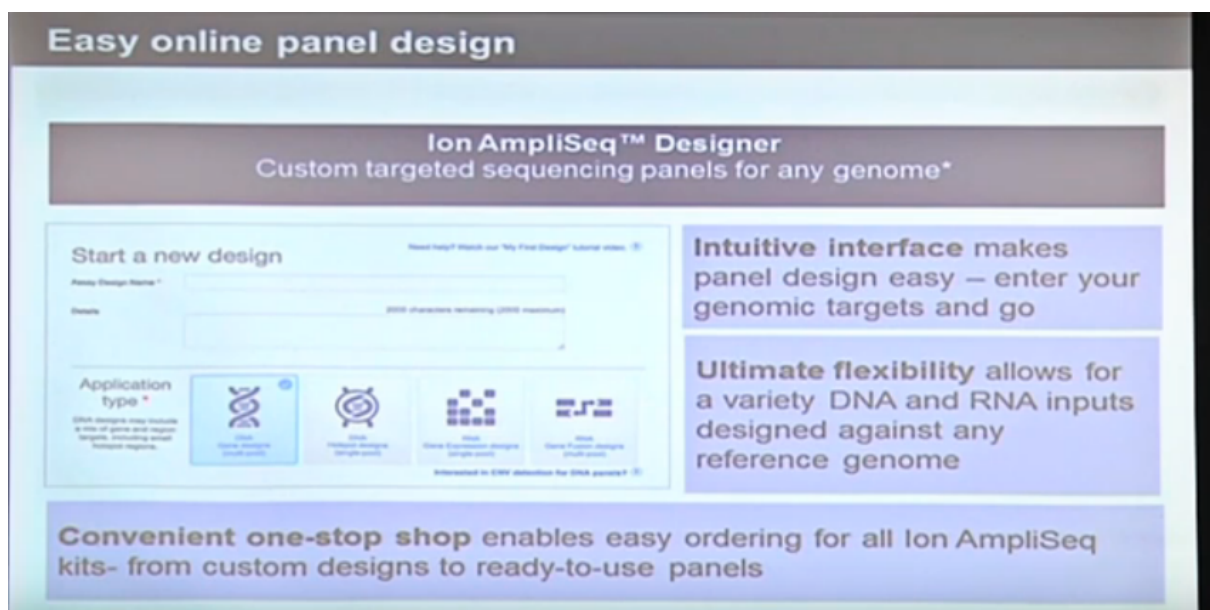
You have three stages, I take each stage individually and explain you as such. So, the first part is an implicit designer, so, what we have is a company workflow of doing the analysis, where you like to know, which are the regions of interest of yours.

Refer Slide Time :( 8: 14)



Whether I could design my regions, in such a way that they, could be sequenced on a NGS technology. Okay? For doing that, we had a tool, we have a tool called as, ‘Ion Ampliseq Designer’ it helps you, to take all your required design genes, in as a list of genes or even the region of interest from your chromosomes. Okay. It takes that and helps you to design, primers which could fit into technology of ion torrent sequencing. Okay. So, you can design something of your own, okay.

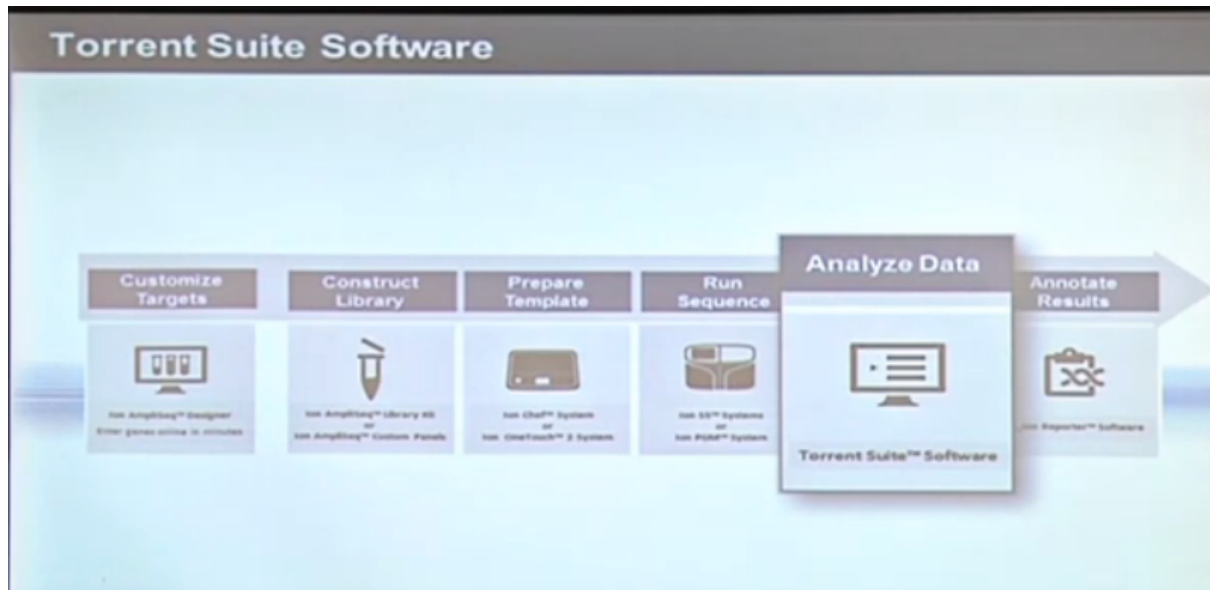
Refer Slide Time :( 8: 54)



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 based design that you can do, where you can provide just the Hyuga nomenclature genes over here. So, you must be knowing about EGFR, if you have any idea, you know knowing, knowing about some Brca gene, Brca one, Brca two. So, these are some of the genes, which are studied a lot, into the entire world. Okay. 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. Okay. 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, 'comics' or DBSNPS. Okay. 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 patients or group of patients. Okay. 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'll use those hot spot 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 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 the results, in such a way: that you're it could easily, tell you which therapy, could be affair properly utilized, for particular patients. Right? So, over here we have an option of DNA hotspots, which take 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 as 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 I expressing into your samples or you can say, if you are considering something like, you have cancerous patient and ass 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. Okay? There's 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. Okay? And we like to know, which are actually highly expressing, which are actually low expressing, this could easily take you to the pathways, which are affected due to these regulations. Okay? 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 into it. So, such study or such type of designs, could be made ready available, which could taken up further, for sequencing onto Antorrence technology. Okay? So, I'll just go move forward, so, what happens with the design? So, just give you a small example, of braca gene, braca one gene, which has been designed over here.

Refer Slide Time :( 12: 51)









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. Right? So, once your reports are generated, the report analysis, work is done by this software, torrent suit software. Okay. It understands whatever data is generated by signals. Okay. Whatever voltage chains are happening, are recorded on to the systems and the software understands, to understands its signals, clears the signals or filters the signals, over there and then decodes, them into particular bases. So, you have a chip, has millions of Wells and in millions of Wells, 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. Okay. 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. Okay. So, once you've done sequencing, what should we do further? We have got the raw data. Right? See this is the raw data comes in to, various formats, one is fast queue, you must have not known about it or you may be having an idea about it, first queue files, which is a raw data file, contingency as well as the quality values for it and there's something called as, 'Bam File' also. So, this is just for 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 a, 'Ste Map'. Okay? I have not put much into this, I 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 braca sample and you do a run and you have a reference human genome, you align it. So, you're not 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 a first question. So, if I have generated around two million reads for that sample, how many reads actually align to those region, which are interest of yours right,

Refer Slide Time :( 18: 22)

## Data analysis and management solutions

### Torrent Suite™ Software

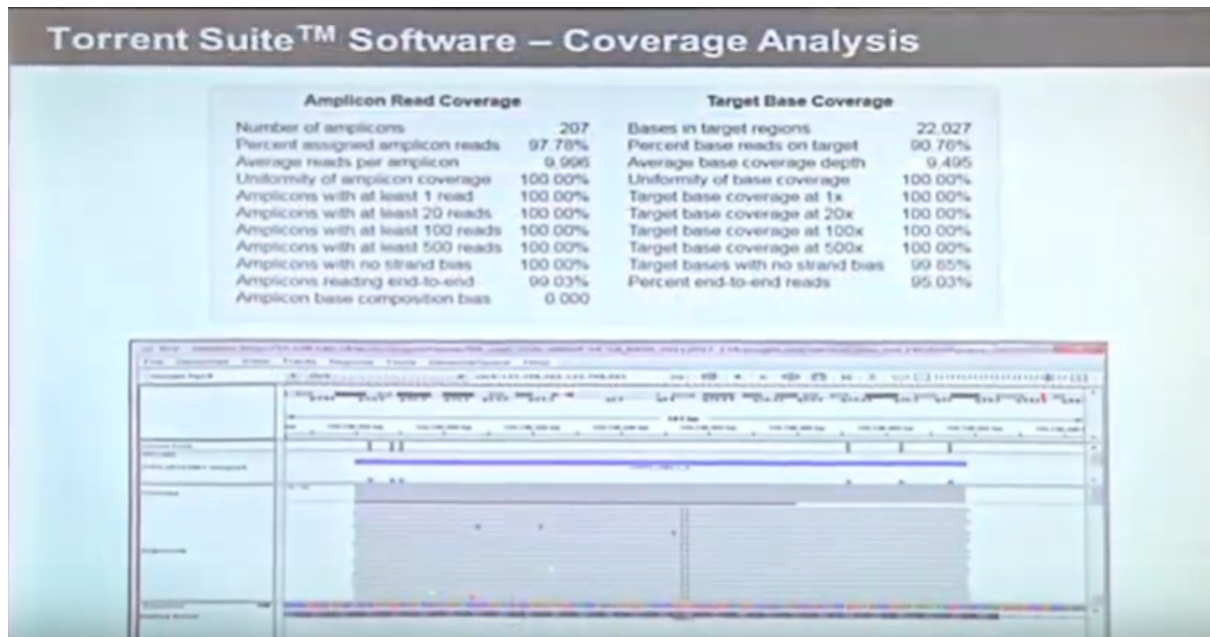
-  Easy-to-use software for automated sequencing data analysis
-  Intuitive, web-based interface makes it fast and easy to plan, monitor, and view sequencing run results with any Ion Torrent sequencing instrument
-  Expanded capabilities with plug-ins available through the Ion Community



so for that we have a tool called as, 'Okay'? 'Okay', sorry, I just explained a lots of thing of the torrent suite software, just to take it in a 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 plugins that are available for doing analysis. Okay? So, as I come back to, what I'm speaking? You 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 protect aligning to your regions or not of interest. Right? So, that's my main important points, over there, are helpful, whatever plugins are there, those coming soon and to play over here. Okay? So, we have certain plugins, which helps you to understand, if the region of interest that you have designed, I could provide my information over there, say I have braca one region, I have designed it, I wanted to know how many, genes or how many reads are actually aligning to those regions. Okay? So, how can I do that?

Refer Slide Time :( 19:41)

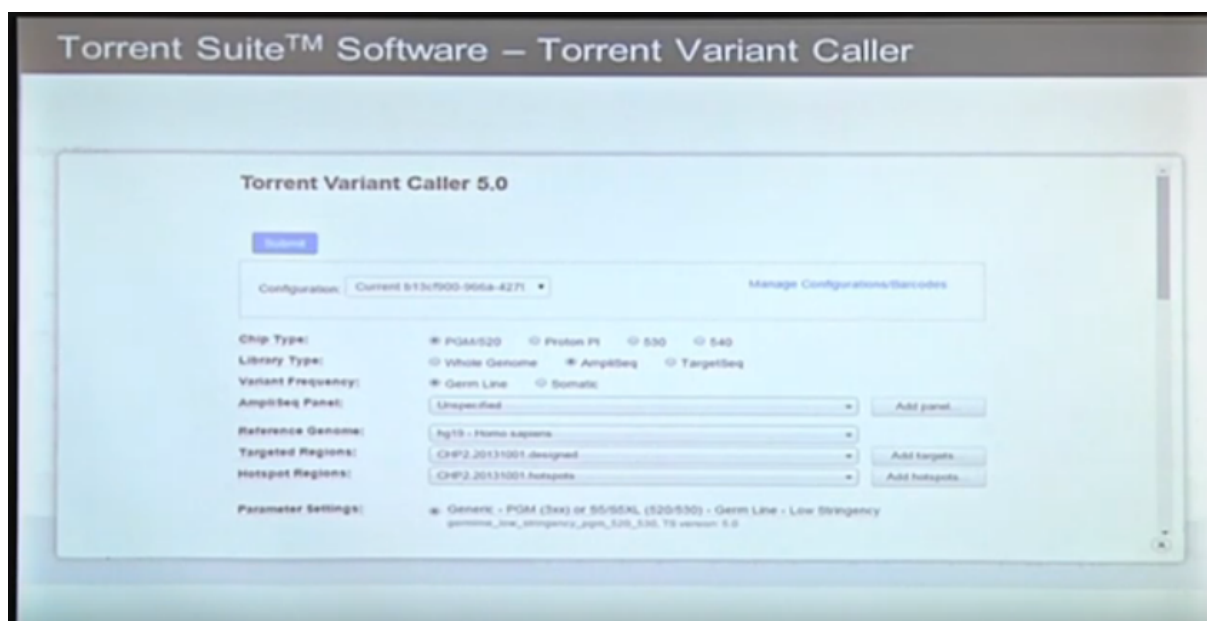




So, I have a plugin, called as, 'Coverage Analysis'. Okay? This coverage analysis, helps you to know, if any region is getting aligned, with number of reads, it lets you know, how many reads are actually mapping to it. So, what happens? So it says, if I have a region, say this is an example for a, cancer hotspot panel, which has around, in the target region has around 22 K basis, 22,000 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 braca region or like CHP panel region. So, it is 90% of my bases are actually, aligning to my region of interest, in consort spot panel. Okay? With that, it lets 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 9,000 495 mean depth. Okay? It's an average depth, covering a particular base over there. So, you may be having a range of around, 8,000 to, 10,000. Okay? Into your CHP panel: that have you designed or it may be a custom panel that you have designed and then you get to know, this particular depth. Now, this gives me an idea, where the map panel properly, having all the reads or not. Right? If I am having a depth of thousand, two thousand it's still good. So now, over here as an example, it's very, high and that, I'll 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 one X coverage, how many bases are covered by one X coverage. Okay? At least one read covering each base: that's what I could say as one X coverage. So, in that you have around 100%, bases which are covered at one X, at the same time, I'd like to know, how many bases have covered at 20 X: that is 20 times so particular bases covered properly or not. Right? So, 20 times are you discovering a particular base, so at the same time it's 100%, till the 500x you can see it's, 100%. 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, 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'm designing something like your gene. Okay? I'm 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 95%, of my total reads, are actually having end-to-end alignment to my region, gene of interest. Okay? So that gives me more confidence whether my data, is coming good or not. Okay? So, any questions till now? Yeah! It's not single base resolution, it is like when I do only, when you, are taking a reference you, you are aligning a sequence, I am just trying to know whether a single base is covered by one read or not. Okay? So, I'm aligning, alignment is nothing but, you have a reference, you have a read getting a line to it? Okay? 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, if I am having something like one read passing to it, so it is covering that particular base. Okay? 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 aligned into a particular region and the same thing, I like to calculate over here, we need some statistics. Right? We can't 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 move it, helps you to visualize your data, how much data has been aligned to that region, how many data is for other region. So, you can go scroll through genes to genes and get to know that. Okay? This is just an example that I've 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. Okay? 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. Okay? So, after this, what I will be going to do forward? So, I have another plugin, called a, 'Torrent Variant Caller'. Okay?

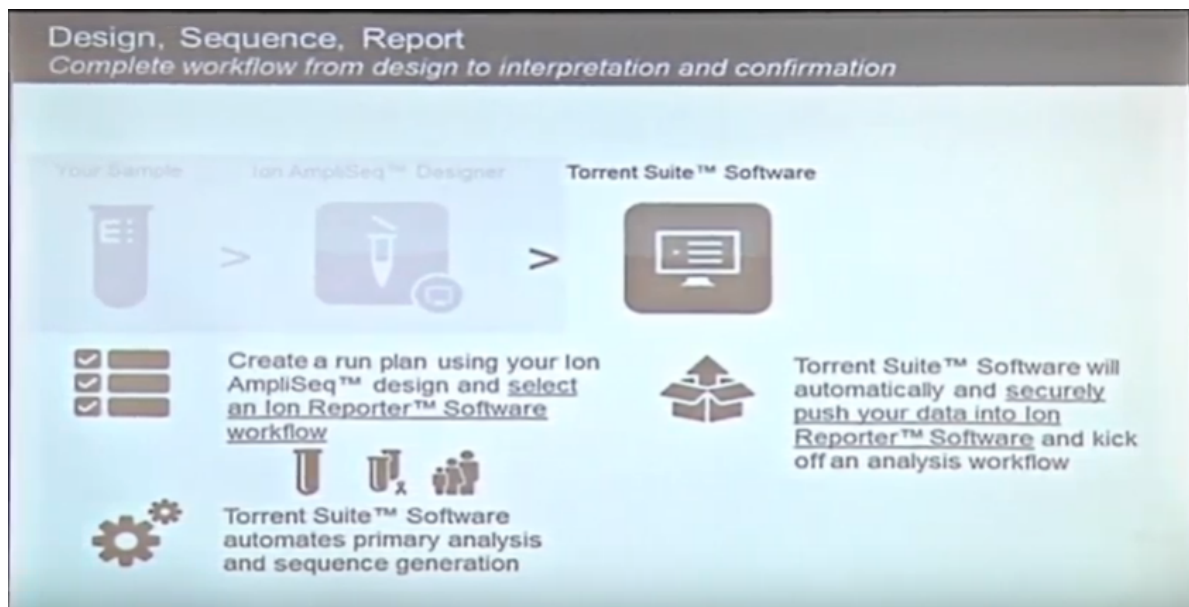
Refer Slide Time :( 25: 03)



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 has, features into it, such as I could give the chip types, since the system over here right now, is an X and s5, we still have two more systems available, one is proton and one is PGM. Okay? So, after that, you have almost similar the workflows, so you can go for CHP panels, 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 is, there's something called as, 'Design Files'. Okay? So, whatever design

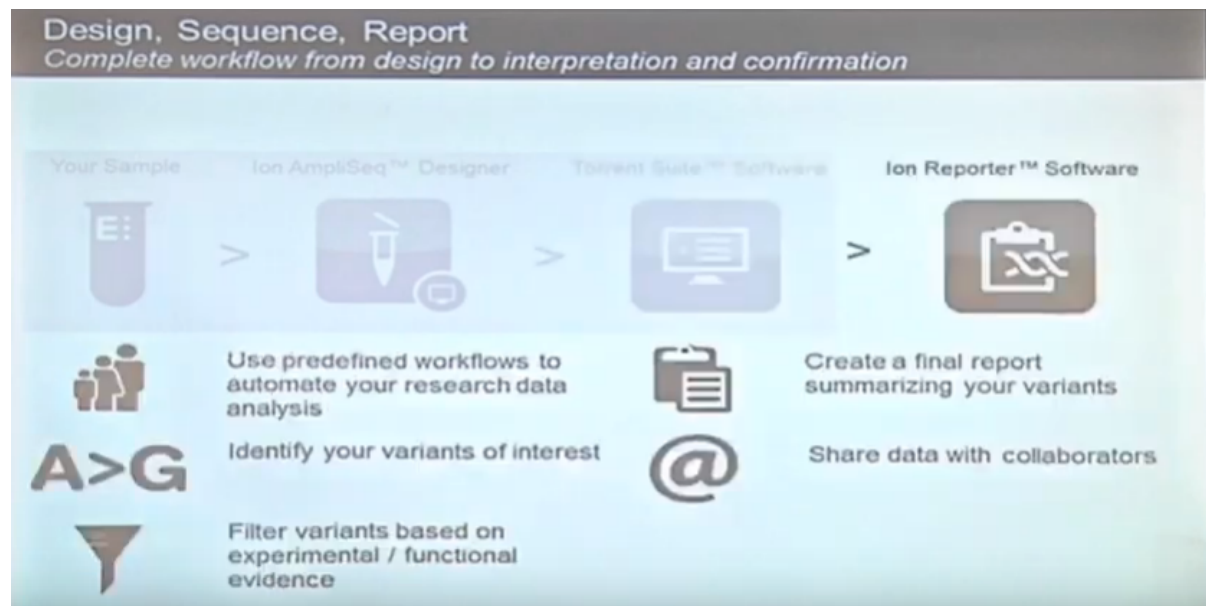
that you create on MPC, those files could be downloaded in a particular format called as, 'Bed Format' and these files could be uploaded for doing the analysis Variant Caller. Okay? Once my design is uploaded, variants would be only called into those regions, which we have designed, it will not go and look for some other regions. Okay? Once those design 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'. But, 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 hot spot that I was looking at, it will provide you the final results in an excel sheet. Okay? 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, 'Variance'. So, variance could be called base or mathematic, nature or else my genre in nature, 50% frequency or as 5% frequency and then the variance could be called at, that stage. Okay? And once I run this plug-in, I could submit the data and once I run this data, I'll get all the variants that are generated into it. Okay? So, what happens? The variance could be downloaded, for all the SNPS index, into an excel sheet, entirely. So, can take that and study it further. So now, what is happening over here is? So, I'll take a step by step mode, I first did designing, on ampliseq. Right?

Refer Slide Time :( 27: 40)



Then I did the torrent suit software, where it decodes, all the bases and provides you the sequences for it. It provides you alignment, with the reference genome, 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 you 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: 28)



So, for doing that, there is one more tool, called as, ‘An 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. Okay?

Refer Slide Time :( 28: 43)

## Points to Ponder

- Data analysis becomes a crucial aspect of NGS platform
- This lecture focuses upon the Ion Torrent™ informatics solution made available for NGS data analysis
- Different software tools incorporated in Ion Torrent™ informatics includes:
  - Ion AmpliSeq™ for targeted sequencing panel designing
  - TorrentSuite™ for data analysis and management
  - Ion Reporter™ for variant annotation, sample comparison and report generation

So, today you learn and at least got a glimpse and some understanding the basics of NGS, platform and data analysis. Here also introduced, to the ion ampliseq designer, the torrent suite software and the ion reporter software. While some of these are very specific to a given technology, which is not the mandate of the codes, to teach you, how specifically these software's work. But, rather you know, by showing these kind of available platforms, intention is to give you, an overview and it could understanding: that how these technologies and these software's, could be used for your applications. So, in today 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's really important, to look at data, into the 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 there is different type of information, obtained at the gene level and the protein level or amaranone level and trying to correlate, all that information together, we need really the highly computing power and lot of inner space, to do these kind of analysis. So, the next step, is to do variant analysis, which can be also, done using a cloud-based tool and I'm sure you know, in the as we go along, in the next lecture, you will specifically study, how to use this data, for more application orientation, especially in the context of cancer. So, we'll continue more, on this slight of NGS and its revolution and we'll talk to you again, in the next lecture. Thank you.