**Functional Genomics**
**Professor S Ganesh**
**Department of Biological Sciences & Bioengineering**
**Indian Institute of Technology Kanpur**
**Lecture No 12**
**Application of Next - Gen Sequencing**

So welcome back to this course functional genomics. So in this lecture we are going to look into some of the applications of next generation sequencing. So in the previous lecture we have seen the principle behind the sequencing and the few examples like cancer genomics. So we are going to go beyond cancer and look at what are the other applications one could think of using this powerful tool of next generation sequencing.

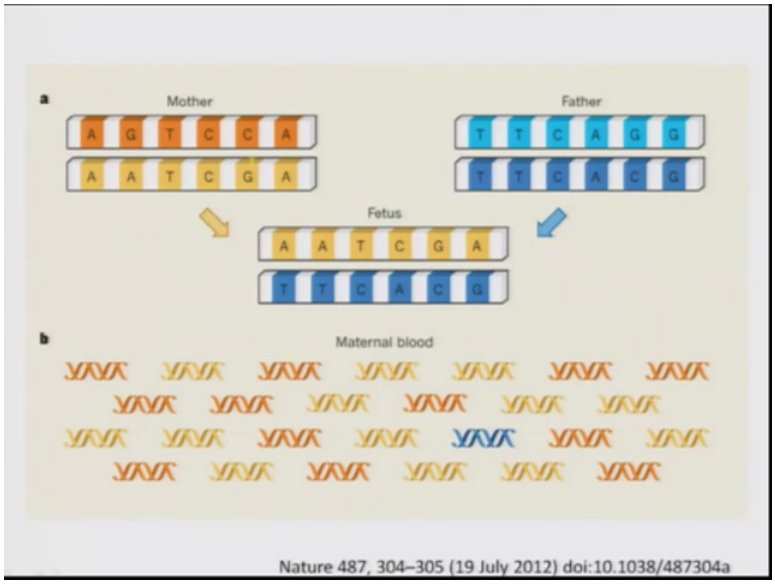
(Refer Slide Time: 0:45)


So here is a paper it is again a landmark paper which sought of established the power of next generation sequencing. Here they have developed or they have shown the sensitivity at which the sequencing can help as a screen for prenatal measurement of fetal genome. So you must have read about disorders that affect the you know children. For example mental retardation is one such disorder, there could be many other disorder like red syndrome and others, it happens because de novo mutation, meaning the gambit that led to the embryo had a mutation as a result the baby is having a mutation.

This is unexpected because the parents are normal but the baby could be and likewise in down syndrome and as the age of the mother at pregnant mother increases the chances of that developing embryo might have down syndrome increases. So the traditional method is to look at the amniotic fluid from the uterus and go for screening either it is a chromosome or DNA sequence analysis and so on. But that also puts the baby at risk because these procedure involved requires invasive procedure. So if you can come up with certain noninvasive screening methods to understand the genome of the growing embryo that is going to really help.

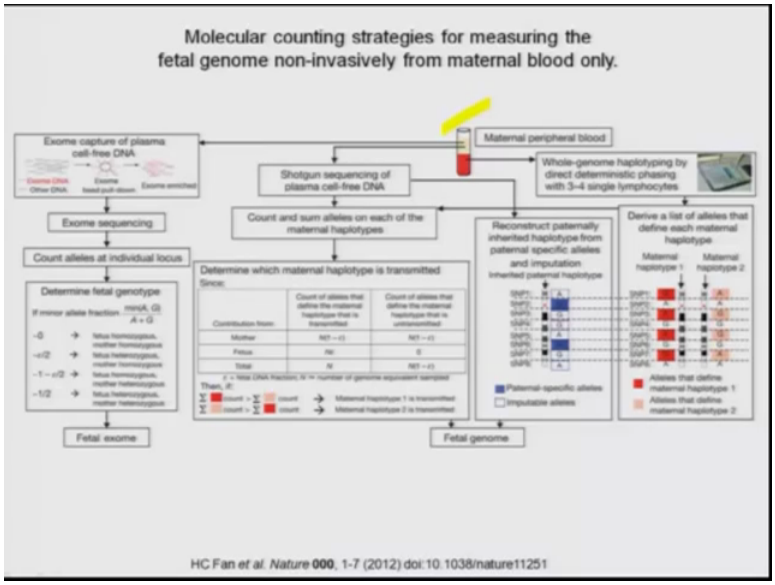
And with the cost of such sequencing coming down, it can be a routine screening for you know the developing embryo to see whether the embryo has any genetic defect before you know certain development stage where you know the parents could take a decision as to whether they would like to have the baby or not, which is legally allowed. So this paper talks about either noninvasive prenatal measurement of fetal genome, what is that.

(Refer Slide Time: 2:54)



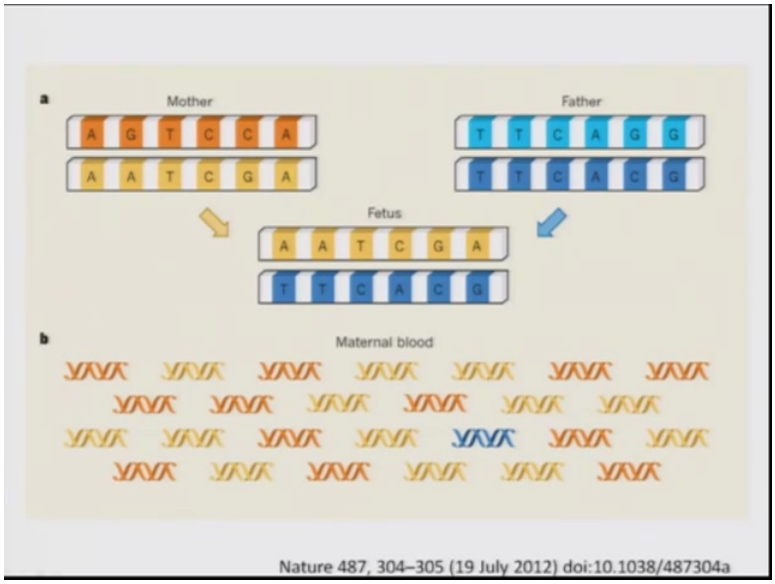

So what is known is that for long it has been suspected that you know some of the DNA or RNA of the fetal fetus, the growing embryo can be found in the circulating blood of the mother. So it can come even few cells can come because the umbilical cord where the exchange of blood takes place there could be few cells coming into or it could be only the DNA coming, only the RNA coming because there are process by which you know there is a now we know that RNA output into small vesicle they are transported and so on.

(Refer Slide Time: 3:35)




So what they have done is that they have used a method that is the tube maternal peripheral blood. And then captured the DNA that are present in the free DNA that are present in the plasma, liquid part of that blood and they have done exom sequencing, ok. And likewise they have done a shotgun sequencing for the plasma free DNA and then looked at the sequence.

(Refer Slide Time: 04:04)



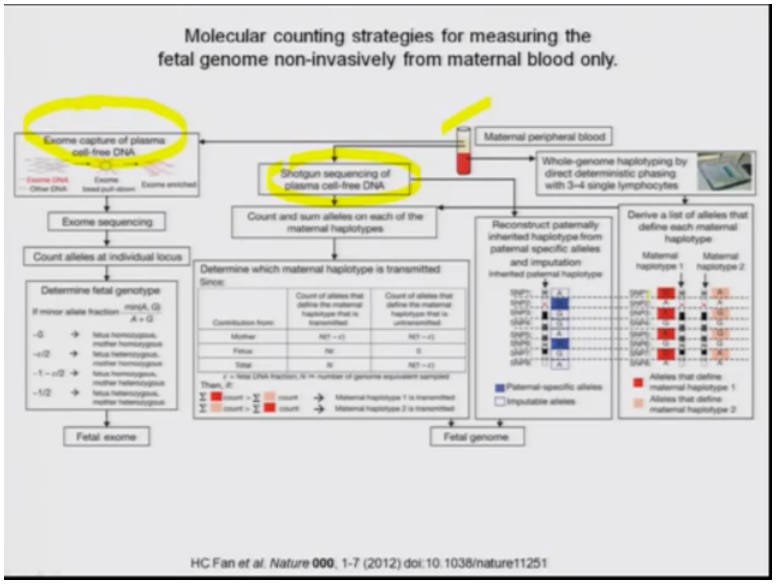

So what does it help, so this the idea is that if you have the blood carry some RNA or DNA from the fetus the growing embryo than by doing this sequencing where you are sequencing every

fragment that is captured and this is so sensitive because these are clonal, meaning that each cluster represent one copy of the DNA, we are able to sequence each cluster, get the sequence separately. You may find that you know if you know that the mother carries the genome from his father her father and her mother.

So you know that is one and of course is a complementary what is shown here but you can really type that what is shown here. So this is the mothers DNA which is representing her, for example father, this is the mothers DNA that represent the mother and these are the sequences which are different between father and mother these are polymorphic side which you are able to genotype and call from fomites come from and obviously the fetus would have DNA sequence that had come from her father, right.

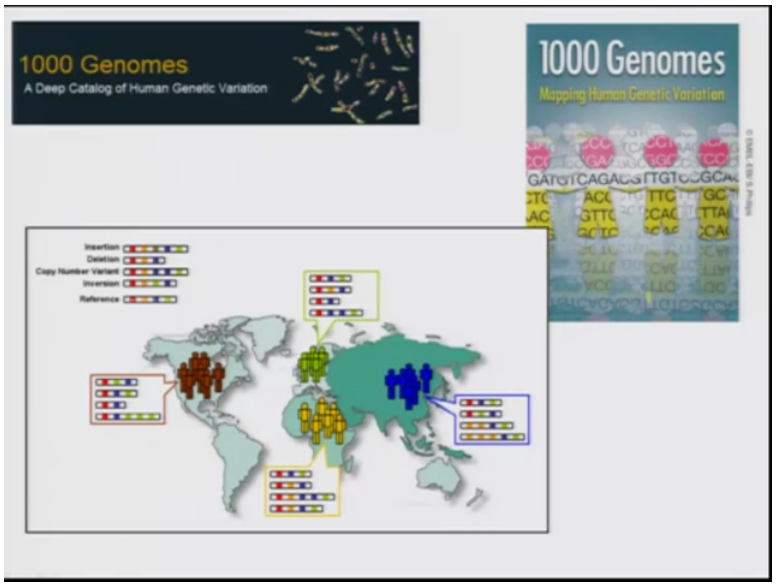
So one that has come from mother, the other copy that had come from father, so this is the information that you have you have the DNA sequence of the father, you have the DNA sequence of the mother which tells you what are the (())(5:17) that they have which has come from his father and his mother and so on. And then you have fragments of the DNA that are sequenced using this approach and all you have to look for is that you know you are there any genome that are representing the signature of you know father which clearly tells that yes that there are genome that represent the father that certainly should have been form the fetus because is coming out.

(Refer Slide Time: 5:58)



So that it tells that you know there are circulating you know free DNA derived from the fetus, so that is the approach that they have used and they are able to really show that you can do genotyping of the fetus by looking at the free DNA that is there in the serum, right and this is powerful tool because you can type you can look at the mutations and you can look at the copy number and say whether it is a down syndrome and so on. So that is really going to change even with some more improvement of the technique it can become a routine screen in the future, it is a possibility such is a huge advancement.

(Refer Slide Time: 6:24)

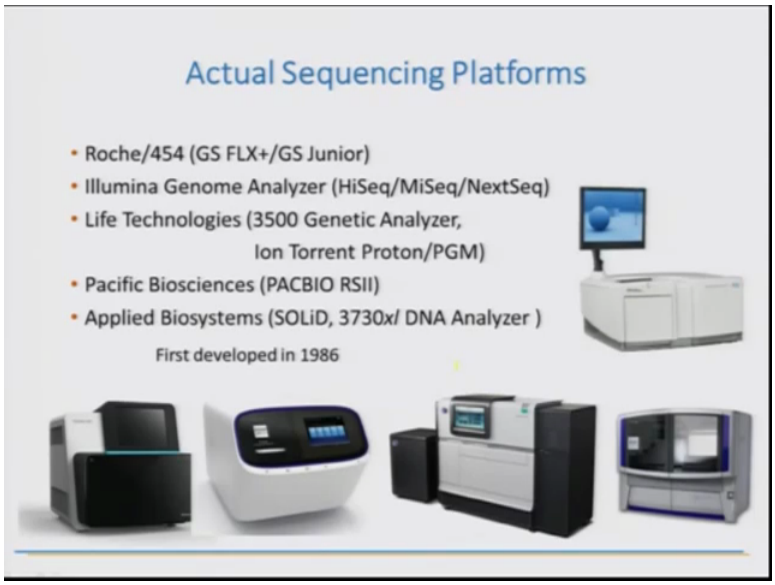



The other application that can you know that is of great interest is the thousand genome project, here the idea is to sequence complete genome of 1000 individuals and to look for variations, something that we have discussed about. And these are individuals that are not restricted to particular race they have selected from different continents like what is shown here and they are trying to compare and look at what variations exist amongst the population. The idea is to create what is called the reference sequence, say suppose a mice genome sequence has been sequenced.

Now when I when this sequence has been done I need to compare the sequence with some reference sequences to infer what is the variation whether the variation is something unique to me or present in the population. So for this to happen you should have enough number of samples already done and in present in the database. So that is the objective of the project. So you have at least hundred plus you know individuals DNA sequenced and kept in a database

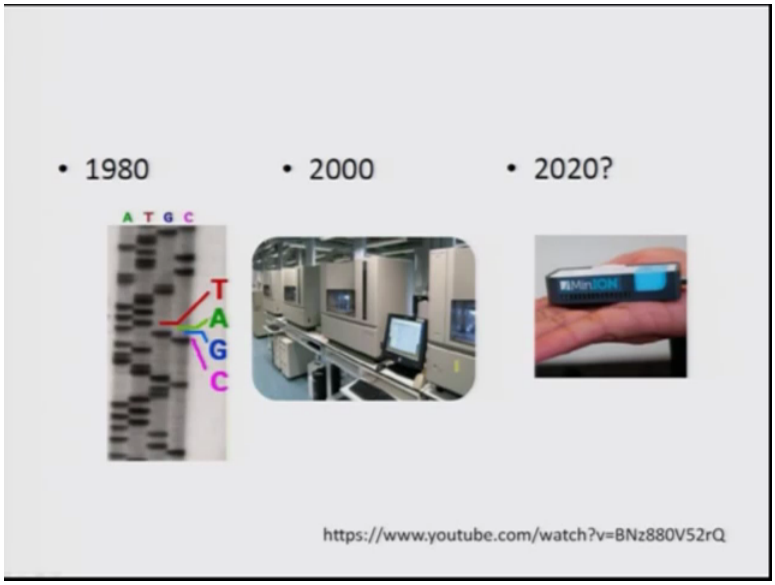
representing each population, so as in when such kind of technologies available for individuals then they are sequenced and then you can immediately match with the reference sequence and say whether the variant that was found is common or not common, whether which segment of the genomic represent, whether it could be deleterious and such kind of you know analysis can be done using this reference project.

(Refer Slide Time: 8:01)



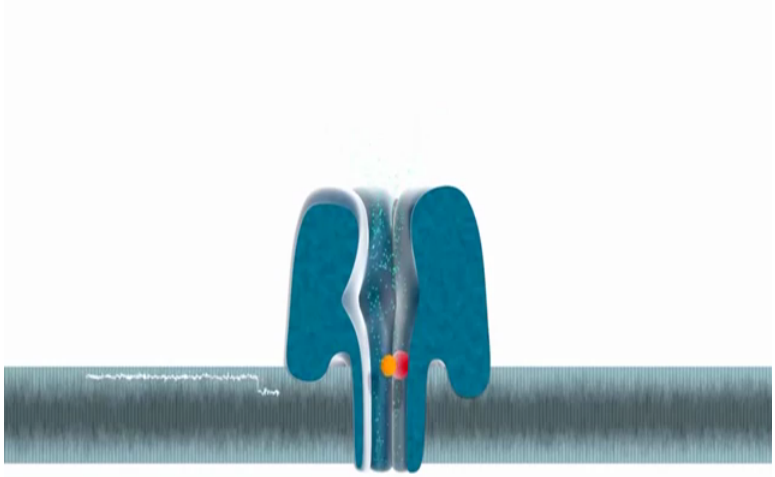

So these are the images representing the next generation sequencers, you can see they become very small table top models have come in. The different companies that are you know marketing these products and they have the chemistry and as well as the two you know the kits for preparing the DNA and ligating into adaptor and so on. So these are some of the popular model that you have.

(Refer Slide Time: 8:26)

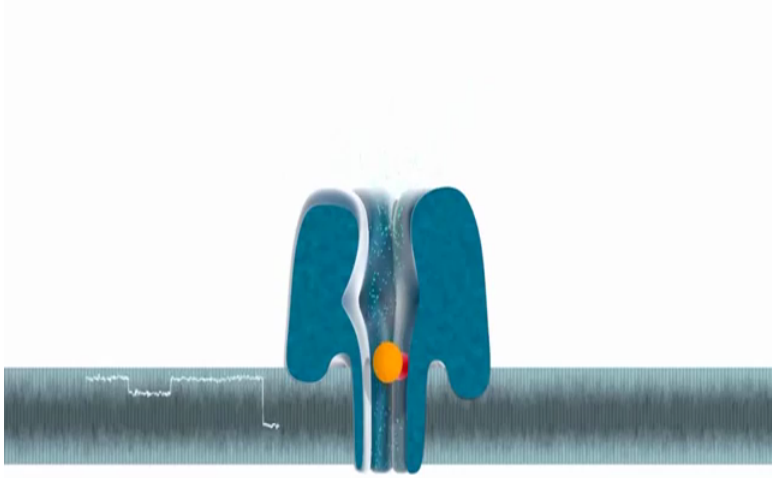



This is just to show you how things have changed, so you know we started with our discussion on the manual sequencing how the extra film was used to understand the sequencers to the high throughput capillary sequencers and now what we are talking about the next generation sequencing and what we are looking at very soon is a product called MinION this is all that you have, that is very small device that you can keep in your palm, it is a personal sequencer. So it is like just like your cellphone the USB cord you can connect to the computer and (())(8:57) little bit of your cell or your blood or whatever and it would complete your genome sequencing in few hours that is what we are expecting. And just to talk about that we have a video and I will play that in the next slide so where you can understand.

(Refer Slide Time: 9:22)

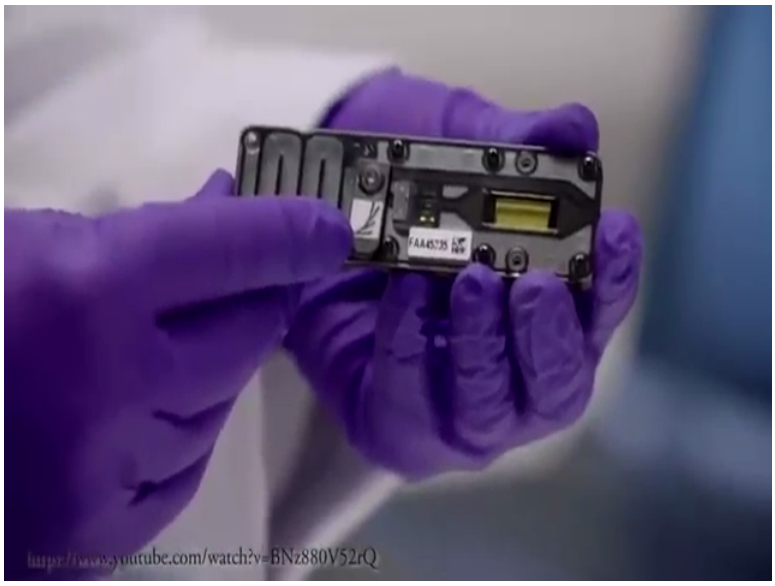
https://www.youtube.com/watch?v=BNz880V52rQ



https://www.youtube.com/watch?v=BNz880V52rQ

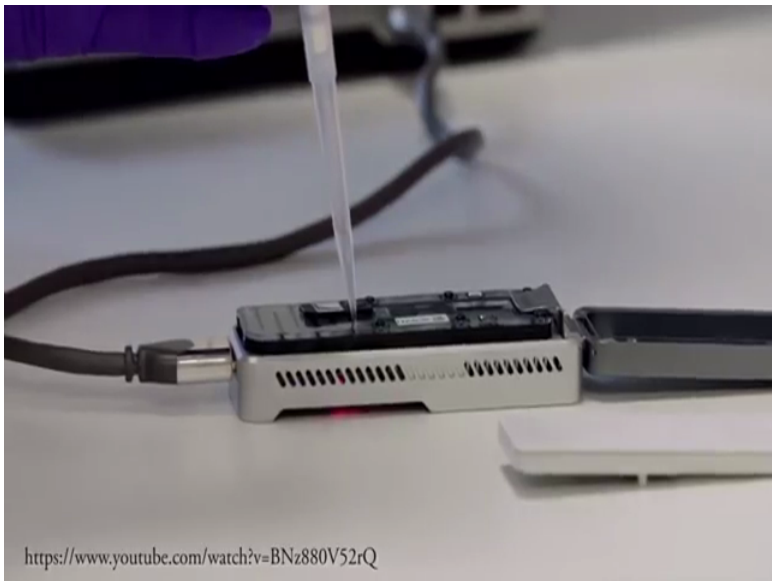
Oxford Nanopore Technoligies is developing Nanopore sensing technologies with the analysis of biological molecules. In oxford Nanopore Systems the Nanopore is inserted into a (())(9:38) membrane created from synthetic polymers. A potential is applied across the membrane resulting in a current flowing only few the aperture of the Nanopore. Single molecules that end to the Nanopore goes characteristic disruption in the current, this is a Nanopore signal, by measuring that disruption the molecule can be identified.

(Refer Slide Time: 10:07)

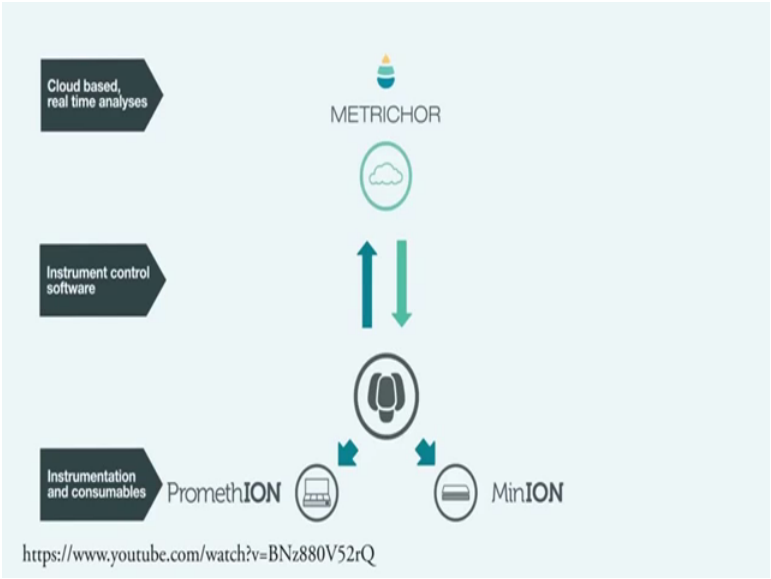

The MinION is a small portable device for the analysis of single molecules such as DNA, RNA and proteins. It has been designed for uses, few want it simple, fast, plug and play device that can generate real time data not confined to the laboratory. Inside the MinION device is a flow cell, this includes the censor array which is a collection of electrodes and micro supports, each one of which has an individually addressable electronic channel. Nanopores are built into membranes lying across the micro supports ad it is here that the single molecules are analyzed using the change in current passing through the Nanopore.

(Refer Slide Time: 10:58)



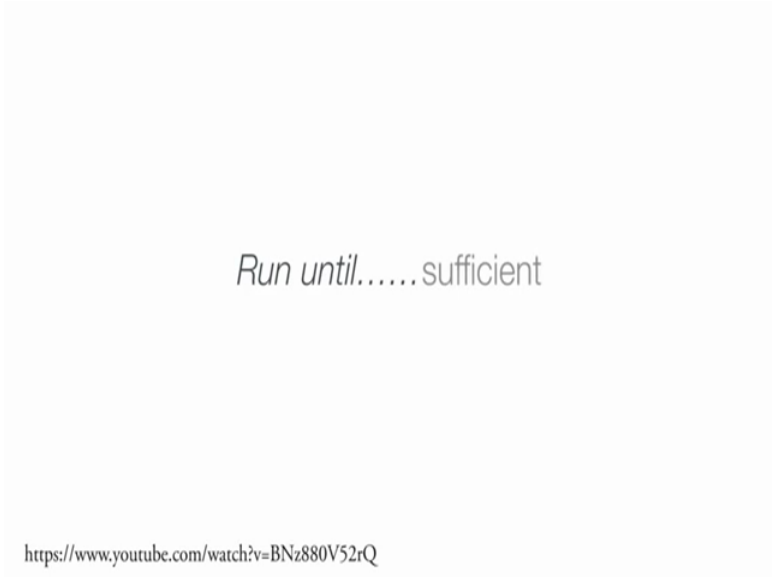

The MinION has a simple and fast workflow and its portability and versatility of experimental design of its opportunities in a number of different applications. Firstly the MinION is connected to a laptop through the USB free cable. The use of (())(11:06) is the instrument control software. After performing a calibration and quality control procedure, the device is ready for use. This sample is added to a port in the MinION and fluid flows across the surface of the censor array and the user begins the experiment. The Nanopore signal is measured by an application specific integrated circuit in the flow cell and processed by MinION the instrument control software.

(Refer Slide Time: 11:40)



This software carries out several code at a tasks and can be used to change the experimental workflows or parameters. The user receives real time feedback of k matrix, such is sample quality and number of events, combined with real time cloud based analysis from METRICHOR over the use of own systems analysis can start immediately and continue the experiment until the point where sufficient data has been generated to answer the biological question. The system may be active for minutes or days rather than being controlled by an arbitrary runtime and subsequent analysis, this is known as run until.

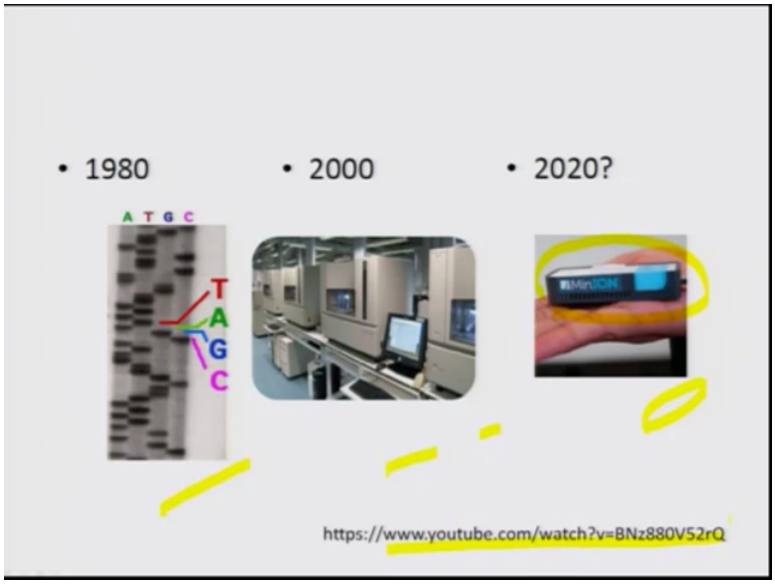
(Refer Slide Time: 12:23) 03





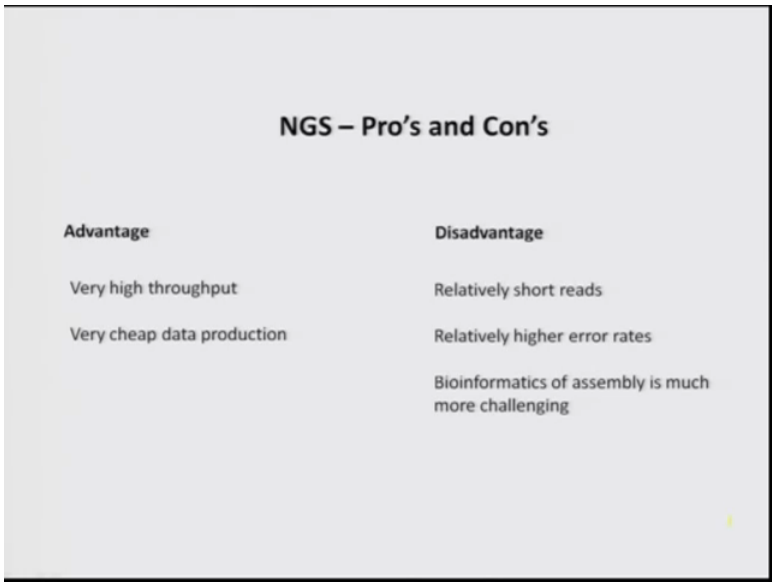
(())(12:22) systems produce full length, full read data whether it is DNA, RNA approach in real time. Data in real time as the machine runs. A running (())(12:36) is in real time as the number of advantages obvious one is which is compute cost and management overhead but it also means you can look for things in real time. If you can things in real time, you can then feed the back through system and you can move the whole system or part of it onto another sample or other experiment. Simple to use with minimum sample probe the MinION is the world's first portable device for the analysis of biological molecules. For more details please visit Nanoporetech.com

(Refer Slide Time: 13:18)



How this device can work you have to understand all these are commercial things they are not going to disclose everything that led to the development of this device by but it would give an overview as to what is the principle behind that. And one can buy this is not going to be expensive and if it becomes really because this has come like a cell phone and you can expect that this would flood if it is successful it will flood the market and sequencing would become just like the blood pressure or just like people measure the glucose level in your blood is going to be that easy so that is the future, right.

(Refer Slide Time: 14:01)



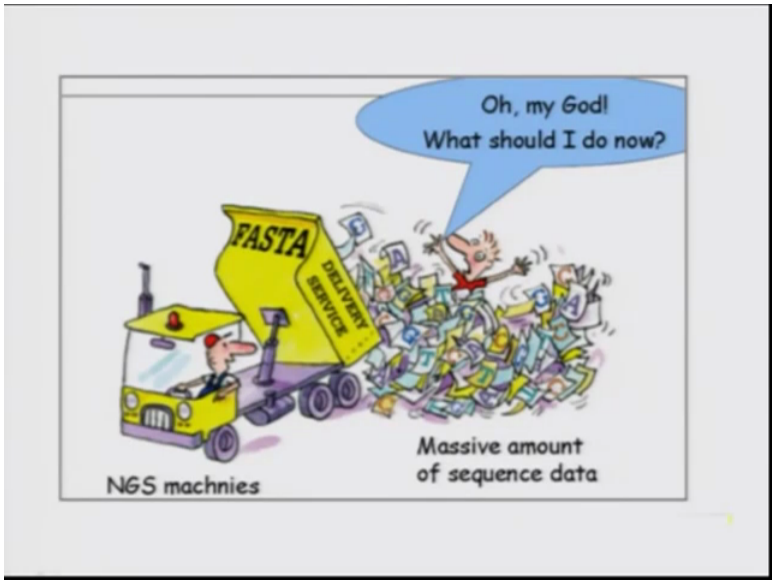

So having seen the video now let us look into some of the advantages and disadvantage of the next generation sequencing, so has is the case with any advancement any technique or any device, there are something good, there are something not so good. The advantages of course very high throughput, you know the number of sequence reaction it can do is anonymous, just put the entire genome it is able to sequence and you can scale up, you can multiply whatever, very cheap the data production but it has its own disadvantage.

The disadvantage is relatively short reads, so 50 bases, 100 bases that is all it can read. But what do you do is for the same segment there are multiple fragments overlapping fragments you are reading. Therefor you have to have more reads for a given segment to you know match, assemble and get the sequence. And because of this process is also relatively higher error reads because you know there are lot of amplification process.

So every cluster we spoke about that we spoke about that video we are shown that one DNA molecule stuck to the adaptor then you know you have PCR like amplification. The DNA polymerase goes over to make multiple copies and then that copies are sequenced. So when you are using this kind of polymerases they are not you know unlike our cells where the polymerase copies and there is a mechanism that looks at whether the copying process is accurate is there any mismatch.

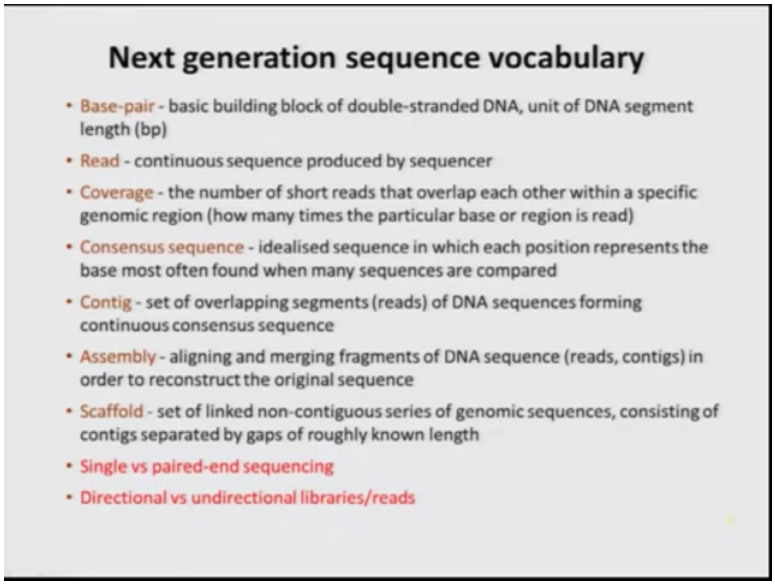
So that does not happen here therefore there could be errors, it is pretty high as compared to the traditional Sanger method. The third is that the you know the data that it generates, it is so huge that it really requires a very good bioinformatic tool to assemble the sequences to analyze is much more challenging. See in traditional Sanger method anybody should be able to analyze the DNA sequence but for the NGS platform the sequence that it generates requires really skilled person analyzing the data.

(Refer Slide Time: 16:14)



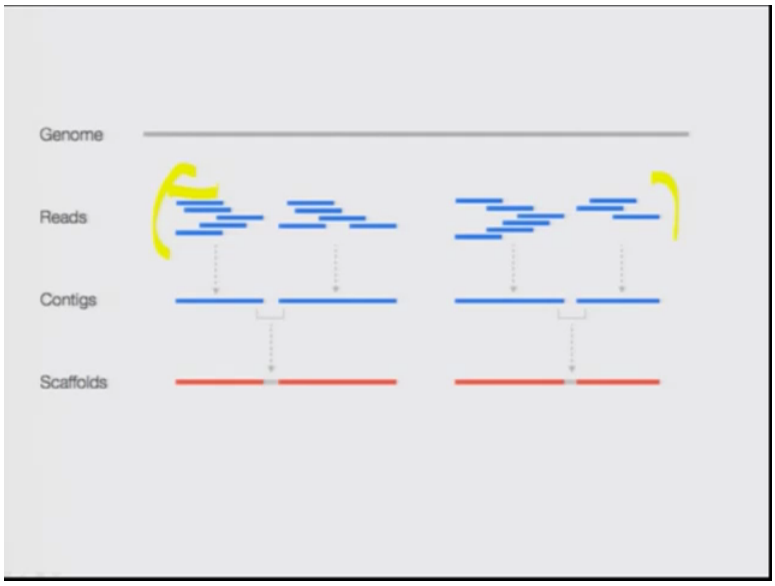

So that is one of the cartoon that talks about NGS machine that it is like dumps, you know it dumps, you know huge data sometimes you wonder what you do that unless you are able to mine and get the one that you want and others what is the noise, what is the signal, so that is something it comes out.

(Refer Slide Time: 16:33)



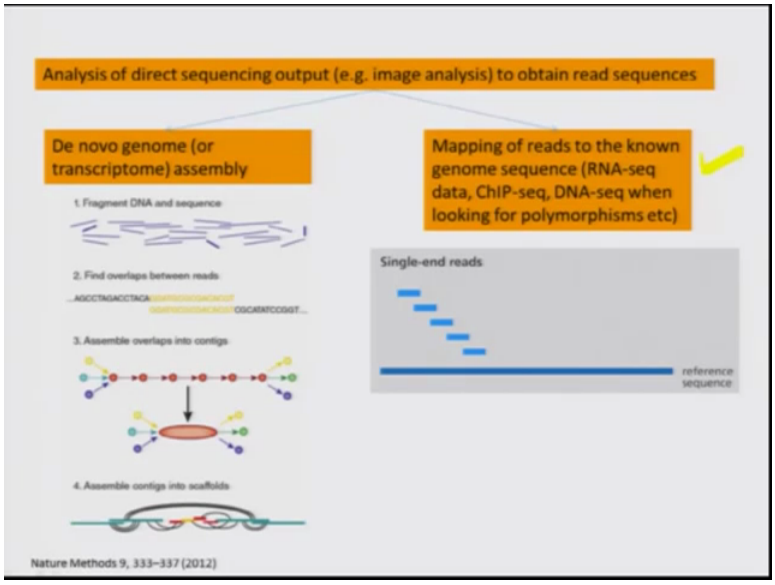

So there are many vocabulary people use, some of them in the next generation sequencing. So we will talk about some for example the read, they talk about how many reads. This is you know small a segment of the DNA how many times that has been sequenced in independent unit, independent clusters as a continuous sequence produce by the sequencer, right. Coverage, this is what we spoke about, you know how many short reads of overlapping segments, you know that that are present.

Consensus sequence, so when you have several sequences you know sequences you assemble to get you know consensus sequence. And this is a Contig Contig is nothing but you have assembled a small segment little larger than the reads but but they are not connecting to something else but represent much larger as compare to the reads. And then you know you have to align with the known sequence to you know get the region for example whether it is DNA coding region and all other things. So these are some of the issues that that come. So we are also talking about single versus paired end sequencing, directional unidirectional libraries. I will discuss it little later.




So this is what it is so you have the reads you know these are reads that represent small segments multiple of them you got the sequence and then each sequence is aligned with overlapping sequence to form, what is called as Consensus sequence that represent the Contig. But still the Contig you know there are gaps within that, but you are able to say that this Contig is located here in the the same you know physical orientation that you are able to do by aligning the sequence or comparing it with the genome sequence that is available. So this is how you are able to map, right.

(Refer Slide Time: 18:21)



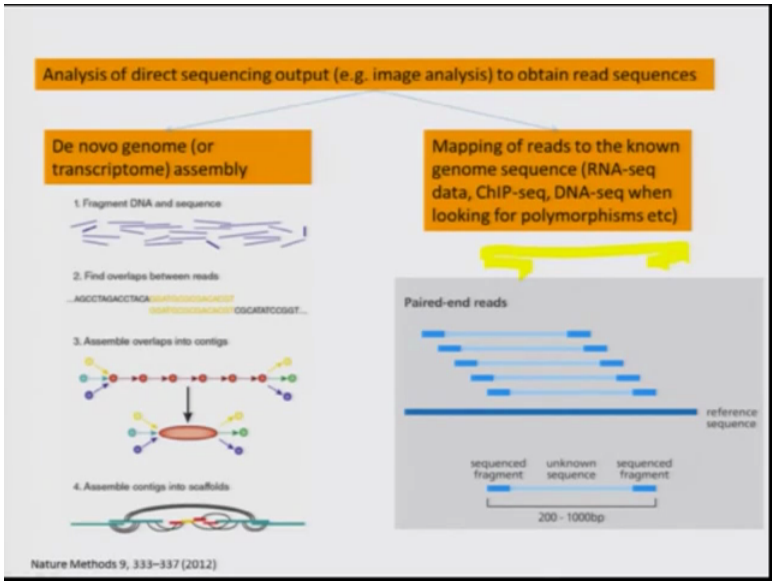
So this is the workflow the data that is generated you know how do you really study that, so the analyses of the the database leaves image analysis because every time (())(18:29) base is added in the cluster it is going to emit a particular light. So but in a chip we are going to have hundreds of thousands spots for every read. So you know there is a powerful image analysis tools that cause for every cluster as to what base has been update.

The way you are able to get put together this sequence, there are two different methods one is called as De novo genome, that is you are sequencing a genome and the sequence is not known previously, ok. So the way you analyze the sequence is very different as comparing to a genome which is already being sequenced but we are looking at what difference you have it in this particular genome. For example human genome is available I want to sequence my genome and identify what variants I got. So I am going to compare with some existing reference data, so the way you analyze is different. De novo meaning you are generating for the first time.

So you have of course the DNA fragment and then you have sequenced it and you are finding this overlapping segments and overlapping segments you know to make Contig and the Contigs are looked at for the sequences and then you make you know much larger (())(19:43) like what we discussed. Obviously this method is not going to give you the continuous DNA sequence, there will be gaps and one has to you know fill the gaps with library and other other methods that is for sure. But in the second method where you are talking about analyzing the sequence for example you want to look into the RNA sequence or you want to find where your transcription factor binds or polymorphism or mutations.

And basically you are reading you know single-end reads, one end of the DNA you are sequencing and then whatever sequence that come you are comparing with the reference sequence and that is where you are compiling as to whether sequence are identical or there any difference.

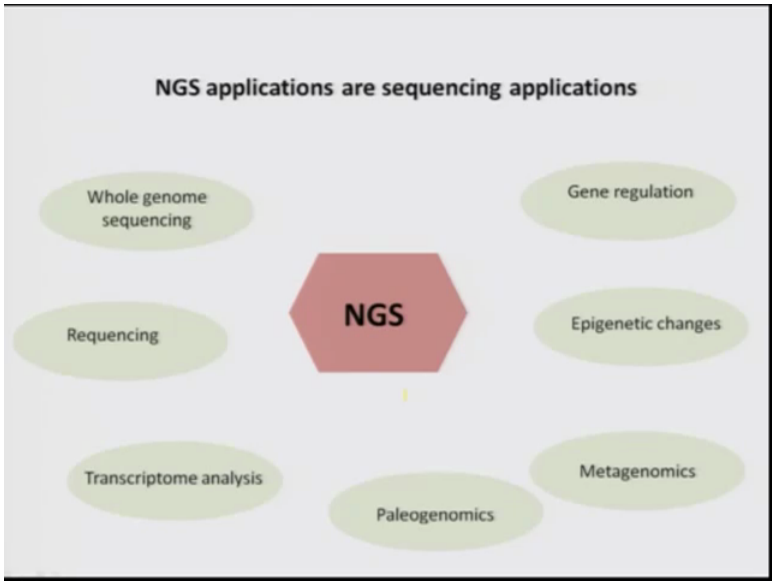
(Refer Slide Time: 20:27)




The other method is you have a longer DNA fragments as I told you the reads are you know the length that you can sequence is restricted here 50 bases, 100 bases, 150 bases whatever. So you are not going to sequence the entire DNA fragment but we are sequencing either end of the fragments and we are going to you know align in like this that gives you this called as paired-end reads, but in this we are going to have segments that are not sequenced, right.

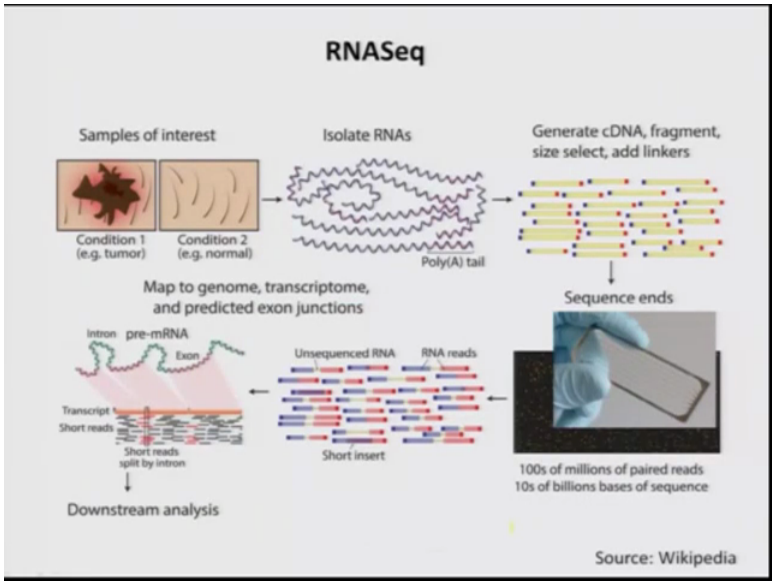
So that is a limitation but this is good enough for you to tell if you are sequencing this is an RNA, cDNA rather it only tells how many copies of the RNA is present. So still it gives you information likewise it you are talking about ChIP sequencing data what are the DNA elements on which the transcription factor bound. So how many times there is bound, so we are talking about quantitative measurements these parents reads are good enough because it tells you how many counts that are present.

(Refer Slide Time: 21:29)




So that is about the next generation sequence and sequence analysis to some extent. Now we are going to talk about much larger application beyond you know the mutation detection. Of course one is whole genome sequence, resequencing like what we have seen 1000 genome project and so on. We also looked at for example transcriptome analysis and ofcourse we can try to understand how the genes are regulated in terms of changes in the DNA, in terms of transcription factor binding to that and and and epigenetic changes for example methylation of the DNA or metagenomics we discussed where we are looking into large number of microbes we are sequencing, classifying them based on certain signatures or paleogenomics for example (()) (22:08) pieces you know you can sequence it quickly and understand the sequence and and so on.

(Refer Slide Time: 22:17)

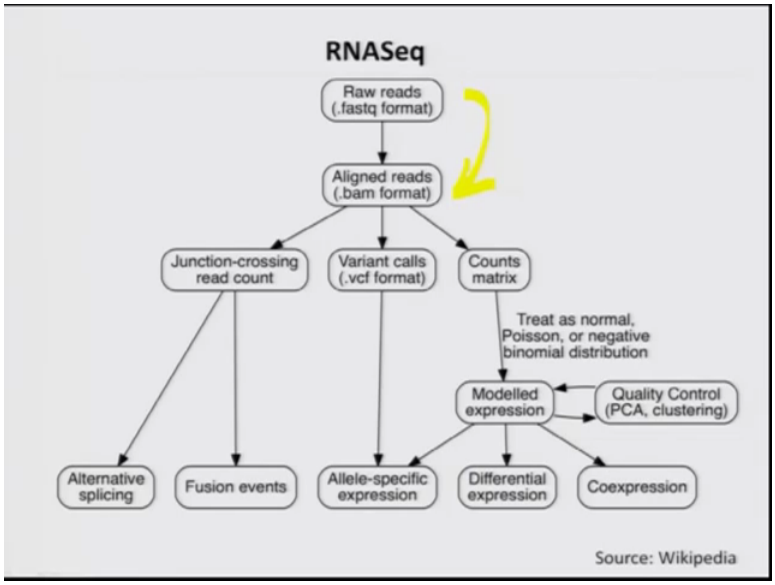



So these are examples one I am going to show you is what is called as RNAseq popularly known as for sequencing RNA, right. So let us say you are looking at two conditions, condition one is a tumor, condition two is a normal skin. Let us say the skin with the tumor and skin that is not having a tumor. You isolate RNA, exactly the same way like traditional method and then you are going to pull out all the MRNA if you are looking into the coding RNAs and then obviously the first step would be convert the RNA into a DNA what you call as cDNA fragment and make it double stranded.

And then just like the way you have done it for DNA, you add you know adaptors on either end of the DNA, this is what you do it, right. And then the application is exactly same like DNA sequencing you put through the small channels and then allow the DNA to go on bind and make clusters and then use one of the chemistry to you know read the sequence that is what shown here as a dot like the way it is explained in the YouTube.

And then you are going to analyze the sequence, so it could be you know short inserts or you have reads on either end of the DNA and once you get the sequences you are going to go and you know align with the the genomic data that is shown here for example that would tell you a given segment of you know transcript has come from representing exon and intron and and there is spliced product or unspliced product all this information can be obtained.

(Refer Slide Time: 23:57)




So this is how you do that further, so you have the raw data and then you have aligned them to make the Contigs or assembly. And then you you know compare them with the reference sequence that talks about for example Junction-crossing read count, for example alternate splicing, right. Exon one is joined with exon three or exon two or exon four. So by looking at the sequence in the exon boundaries you will be able to tell how many times, how many different ways the exon is spliced or you can talk about fusion events.

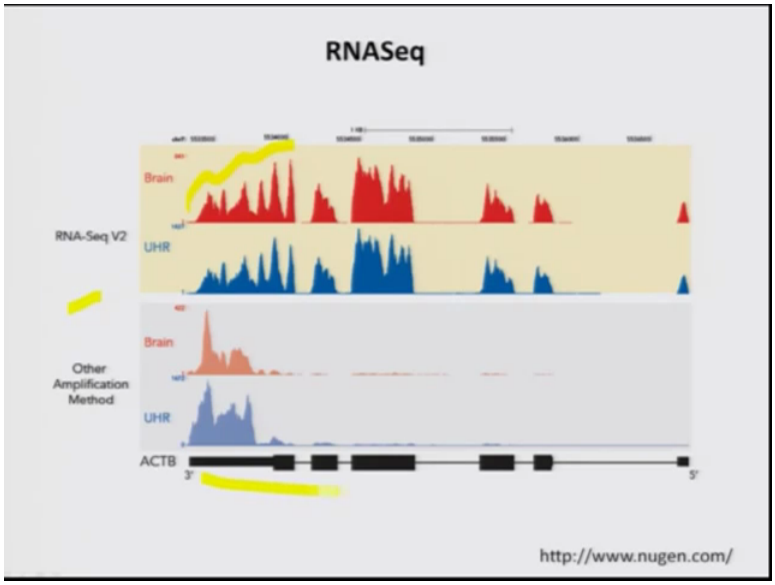
Sometimes your chromosomes are trans located as a result a two different gene is fused together therefore you have a what is called as a (())(24:31) transcripts have been different coding regions, easily you can look into that or you can look at even what is called as allele-specific expression, meaning you have you know for all the genes that are present in your autosomes, you have two copies one representing father, one representing mother.

For some of the genes it could be only the maternal gene copies expressed meaning the gene copy that you derived from mother is expressed not the gene copy that got from father. So how would you really do this, so you can sequence it and there are variants that are present where these variants allele could be different between father and mother. By looking at the sequence of the RNA we will be able to tell is there that both the paternal and maternal allele contribute or equal number of transcripts or there is a disparity, you have more of from the maternal one or more transcripts from the paternal.

If there is a difference then you can go back and look into the genome organization and sees there anything that affects the way the transcription is being done. So that is another powerful tool which is normally you miss out when you do a micro array or real time PCR that we discussed earlier. And then we can talk about other you know we can talk about count matrix meaning, how many reads you have got for you have gotten for a given run transcript of genes that talks about the quantity, the expression difference. So if you have a normal skin versus tumor skin then normally we talk about different ways (())(26:01) so on. But you can also talk about the differential expression and and then you will be able to calculate come up with that.

And then since you are looking at a large number of RNAs at the same time you can model it if gene A expression goes up what happens to other genes is it that there are some genes whose expression also goes up so you have (())(26:24) large number of samples we can do model and then we can we will be able to tell whether such kind of you know co expression patters you can identify which help you to predict more networks and so on, right.

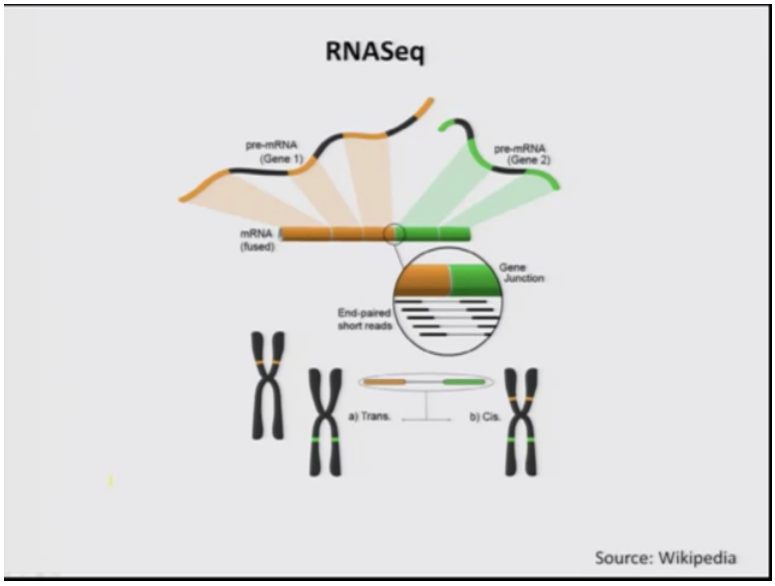
(Refer Slide Time: 26:42)




This is just to give you some example as to how people do this is RNASeq data, so these are nothing these mountains represent the reads how many reads happen for you know the region of the DNA you can five payment of the gene, three payment of the gene, exon 1, 2, 3, 4, 5, 6 and this is UTR. You find that the number of you know reads that you found for each segment. So with this way we are able to map it, not only that we can go on and talk about for example if we

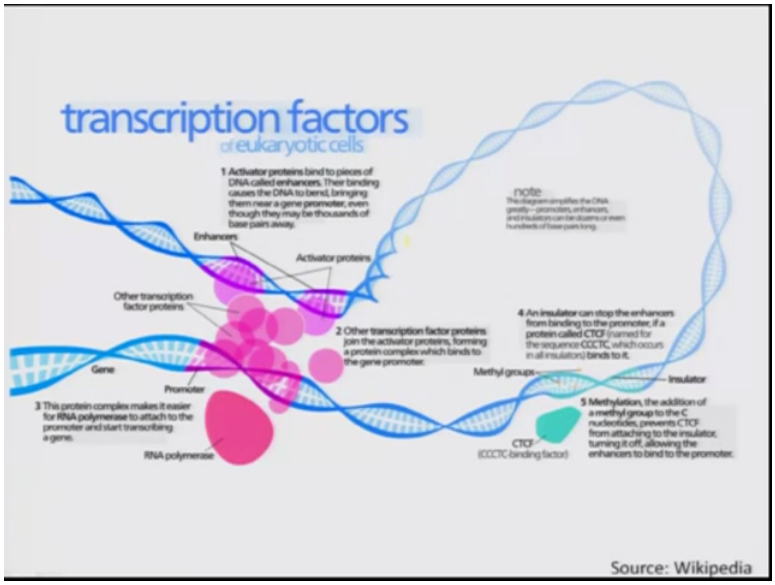
go for other methods for example sit in a library or other methods normally your analysis power more restricted to the three payment which sought of you know is taken care in case of the RNASeq. So it is one of the major advantages.

(Refer Slide Time: 27:32)




I gave you example of chimeric gene when there is a translocation it could so happen that to different genes fused together by looking at the transcript. Now here you can find out such kind of transcript where this pricing you know took place such that you know two different genes now transcript and joint together. So this is a powerful method that you have.

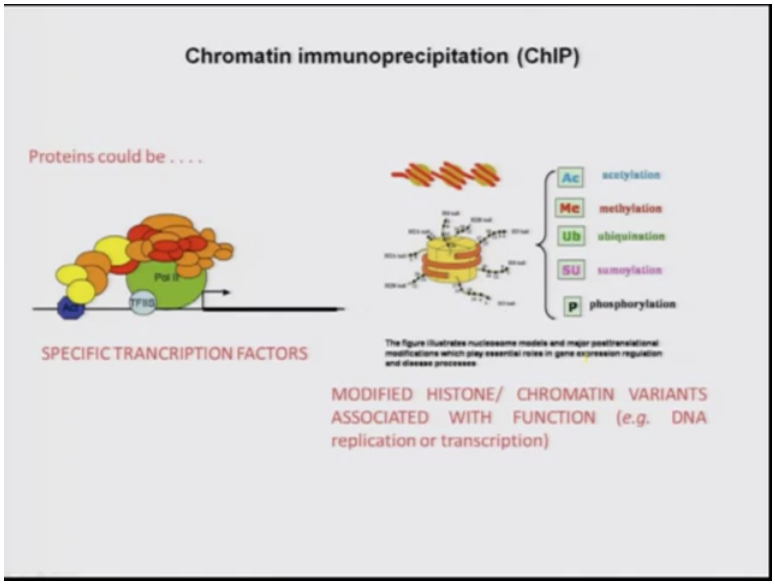
(Refer Slide Time: 27:56)




The last section that we are going to discuss is how the chromatin or the DNA modifications can alter the way gene functions, right. What are the way the transcription can be regulated, the cartoon represent here the DNA and then you have what is called as enhancer sequencers onto which there are some proteins comes and bind is called as enhancer activating proteins. And you have of course the transcription factors which bind to and then is all together is a complex (()) (28:22) to RNA polymer is for the gene to be transcribed.

So now whether a gene is expressed or not expressed depends on whether these transcription factors are present. And whether this region of the chromosome is open for the protein to come and bind. It happens because of many factors one of them could be the changes in the DNA which is called as methylation or at the chromatin it could be histone acetylation and so on. And how do you really look at such kind of changes at the global level.

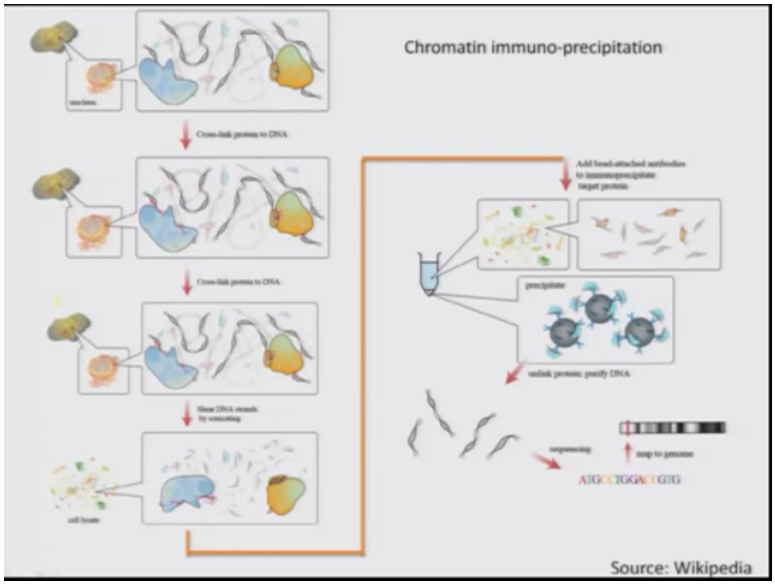
(Refer Slide Time: 28:59)




So one way people do is to do what is called as chromatin immunoprecipitation. So basically what you have is that you have the chromatin here histone shown that has tail coming out which can be modified. It could be acetylation, it could be methylation, it could be ubiquitination, it could be sumoylation, it could be phosphorylation. So if you have an antibody which recognize acetylated histone or methylated histone, or any other modification specific at antibody would now come and bind to this region.

So now if you can use this as anchor to pool the DNA that are present close to it then and then sequence the DNA it will tell you which was the regions that are in the chromatin acetylated or methylated, so this is the way. Or I could use an antibody against anyone of the transmission factor. Now and if I pull using this antibody the DNA that are bound to this transcription factor it will tell me as to which segment of that genome the transcription has bound to.

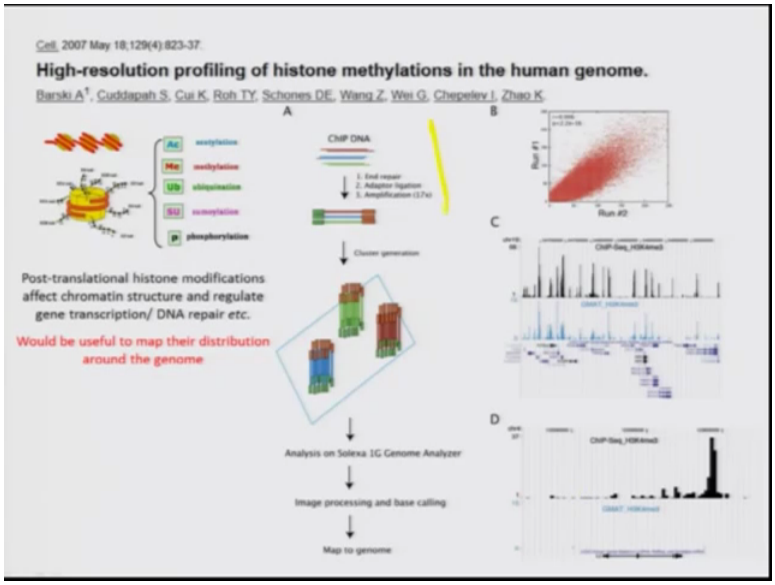
(Refer Slide Time: 30:03)




So this is a powerful method and this is what called as chromatin immunoprecipitation. The schematic is shown here, so what basically it do is there you isolate the chromatin and then cross link the proteins to DNAs. So can you chemical moieties which stabilizers what is shown here, so the DNA and protein is stabilized and then once you have done that what you do is you saw (())(30:21) of the DNA net, you know cut the DNA randomly therefore you have protein complex in which you have the DNA bound to it, so there are going to be.

Now if you use an antibody which recognize a given protein, for example I am looking at transcription factor A. Now this is the antibody that binds to transcription factor A, so it is going to pull wherever transcription factor is there in the protein it is going to pull if the transcription factor is bound to certain DNA that DNA also come along with that after which I extract the DNA and I sequence it and with the sequence I can go on and map the genome to see which region that sequence represent therefore I can tell this is a region onto which the transcription factor is bound to. So this is the way people do.

(Refer Slide Time: 31:23)



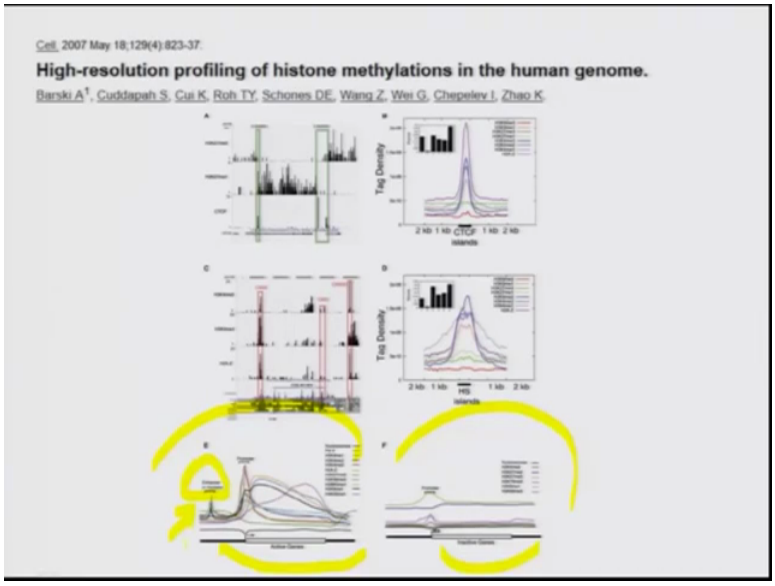

And this is one such landmark paper where they have looked at high resolution profiling of histone methylation of the human genome. So they have used this chromatin immunoprecipitation followed by deep sequencing to understand how it happens. So basically they have pulled the DNA using ChIP immunoprecipitation and then they went and did this NGS you know, you have small fragments that are derived from the ChIP DNA, you added the adaptors, you pass through the fluids the device and then we make clones, we make this clusters, several copies of it and then went through this sequencer to get the data.

So with that now we are able to tell for example, the regions in the genome where the methylation is very high, because you use an antibody now recognize for example methylation of the you know histone. Now it tells you for example this segment of the genome you have severally you know reached that means in many cells this is a segment onto which you know the antibody is able to bind which suggest this is region the histone is methylated. What is that region you know it is an upstream region or downstream to upstream to a particular you know gene.

So it tells you that this region is methylated and this is you know expressed or not expressed you can go and analyze later. And this is going to really help because we are talking about epigenetics. Epigenetics is I may have a normal DNA sequence but the way my genome functions is modified because the kind of environment I live in, for example, right. So if I can

understand now what are the regions that are you know of the chromatin that are methylated is going to help how the environment modify the genome its function without changing the base. So that is a huge huge advantage in understanding the function.

(Refer Slide Time: 33:26)




And some of other you know sequence shown here for example this is an active gene, this is an inactive gene or a condition in which the gene is inactive, a condition in which the gene is active. You can see that this is this is the number of reads that are denoted by the line diagram here you can see the enhancer region if you have used then you can find that the methylation is altered here, more reads come here likewise a promoter and so on. So this is the way depending on what you are looking at transcription factor and others, you are able to infer as to how that could possibly you know regulate the expression of the gene.

So that is the power of the next generation sequencing and and as you have seen the just you know these these technologies are coming up still you have long way to go and maybe next 5 years, 10 years you are going to see you know changes that that you cannot imagine beyond imagination and it could be anywhere in the city, town, there could be shop that can sequence the genome and may have methods to analyze the genome and predict as to whatever risk factors that you have or diagnosis that can you know become better just like what you have for blood tests and others.

So with that we end this section of the discussion that is next generation sequencing. The next classes we are going to discuss about the usage of these sequences to understand how we evolved and how we can analyze the genome sequencing.