An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Prof. Kelly Ruggles
Department of Biosciences and Bioengineering
New York University
Indian Institute of Technology, Bombay

Lecture - 03
Introduction to Genomics - I
Gene sequencing and mutations

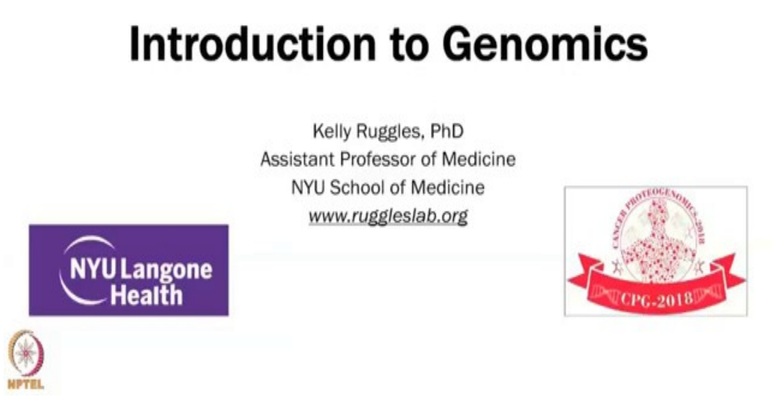
Welcome to MOOC course, on Introduction to Proteogenomics. In previous lecture, you heard Dr. Henry Rodriguez; he gave you very nice and broad overview of the field of proteomics. Especially, cancer proteogenomics, the major milestones which have been achieved and what kind of challenges which lies ahead for the community.

Now, we are going to talk in much more detail in depth about first genomics module, then proteomics and then we will try to integrate the big data and proteogenomics to make meaningful insight. In this light the first lecture today is going to be by Dr. Kelly Ruggles. She is assistant professor at New York University in USA and she will going to talk about genomics, one of the major aspect to understand the comprehensive view of any disease or any system.

Dr. Kelly will talk to you about the diversity of omics in biomedicine and especially, the milestone which have been achieved using genomic technologies. Various type of mutations like hereditary or acquired mutations, which affect different type of diseases especially, in context of oncology. What are the publicly available where one could try to study and find the mutation studies and little datasets, which are publicly available. Along with that the basics of gene sequencing and methodologies will be covered as well.

Hi so, I am Kelly Ruggles, I am an Assistant Professor NYU and I am going to give you an Introduction to Genomics.

(Refer Slide Time: 02:09)



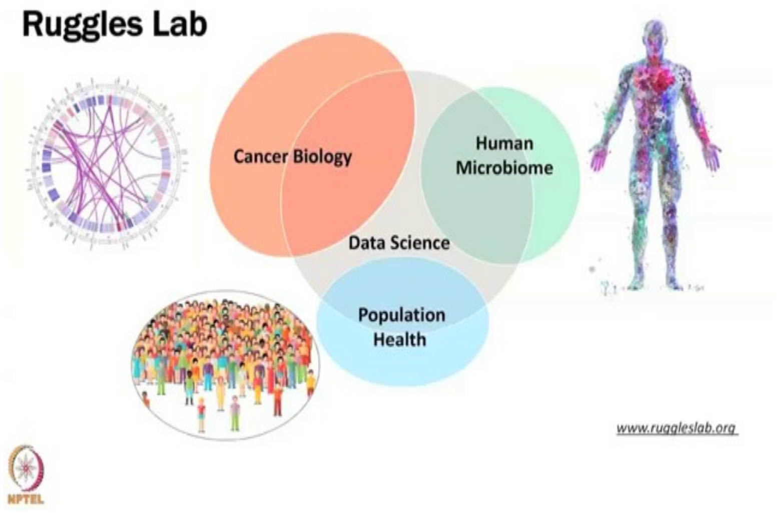
Introduction to Genomics

Kelly Ruggles, PhD
Assistant Professor of Medicine
NYU School of Medicine
www.ruggleslab.org

NYU Langone Health
CANCER PROTEOGENOMICS 2018
CPG-2018
NPTEL

Just to start, I wanted to just give you an idea of what my lab does, to give you little context about who I am and what I focus on. So, our lab really is interested in applying multi omics integration methods across a lot of different questions and diseases.

(Refer Slide Time: 02:28)



Ruggles Lab

Cancer Biology

Human Microbiome

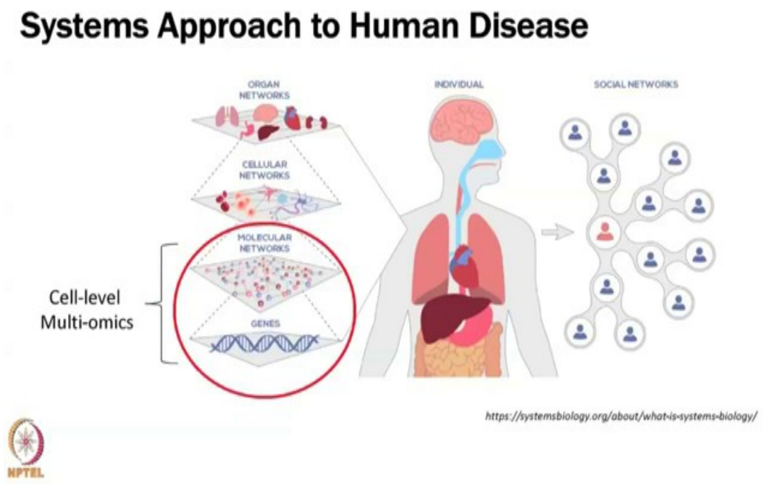
Data Science

Population Health

www.ruggleslab.org

This includes cancer, the human microbiome and I do a little bit of work with pop health. I am obviously, just going to talk about the cancer stuff today, but we do sort of think about the integration across not just cancer, but all lots of different scientific questions.

(Refer Slide Time: 02:44)

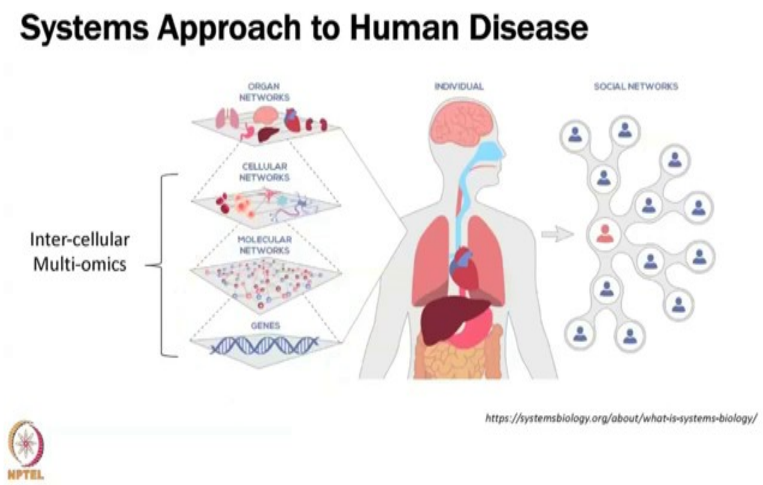


Systems Approach to Human Disease

https://systemsbiology.org/about/what-is-systems-biology/

And, really what are interested in is looking, taking a systems approach to human disease. So, looking at combining omics at the cellular level so, understanding and how proteogenomics within the molecular networks interacts and how that impacts disease.

(Refer Slide Time: 03:04)



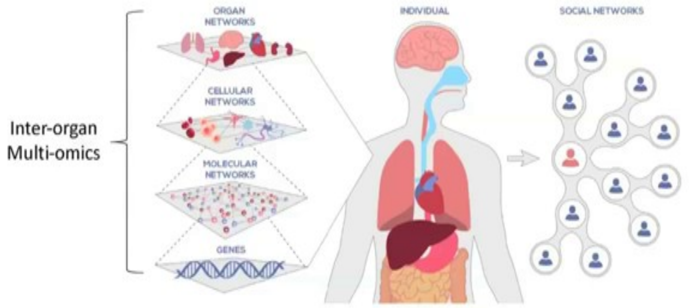
Systems Approach to Human Disease

https://systemsbiology.org/about/what-is-systems-biology/

Also thinking about intercellular multi omics so, not just within one cell, but across many different cells.

(Refer Slide Time: 03:12)

Systems Approach to Human Disease

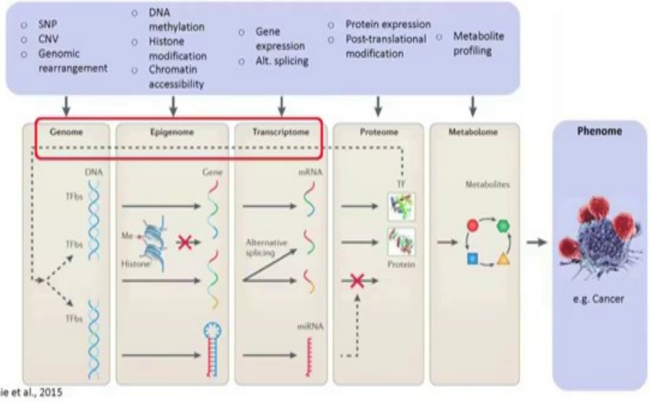


https://systemsbiology.org/about/what-is-systems-biology/

And, then in some cases if we have the data which is of course, the always the limitation, if we can look at inter-organ multi omics. So, if you have many different organs and many different omics you can really have a very large and comprehensive view of human disease. And, I think this is really the goal and we are starting to get there as we are able to generate more and more data and so, I think everybody here is going to eventually, touch on many of the things in this slide.

(Refer Slide Time: 03:43)

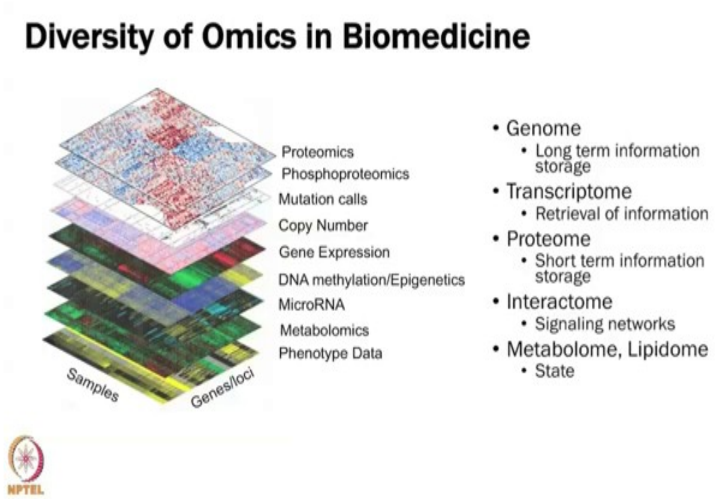
Integrating 'Omics Data



Modified from Ritchie et al., 2015

So, I wanted to bring it up from the beginning. So, really we are going to be talking about all levels except not metabolomics. So, I do not think anyone here is going to be talking about that, but in terms of genomics, epigenomics, transcriptomics and proteomics and there is many different data types that and we can data type that we can gather from these different omics data levels. And so, today I am going to touch specifically on the genomics, epigenomics and transcriptomics data and go through an overview of how we collect the data and then how we analyze it.

(Refer Slide Time: 04:21)



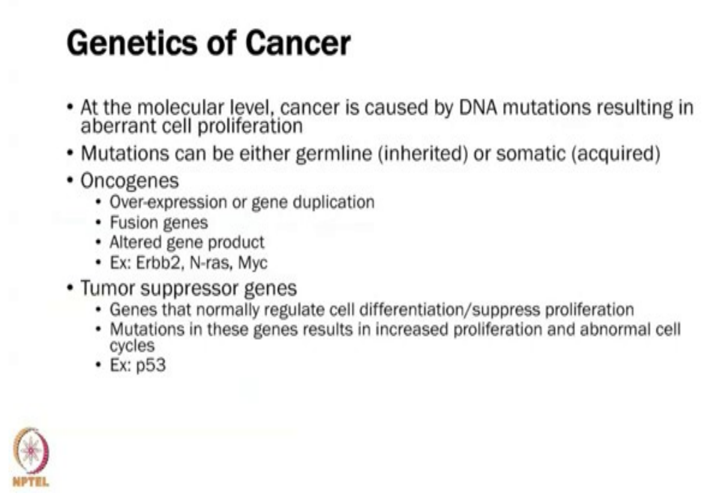
Diversity of Omics in Biomedicine

- Genome
  - Long term information storage
- Transcriptome
  - Retrieval of information
- Proteome
  - Short term information storage
- Interactome
  - Signaling networks
- Metabolome, Lipidome
  - State

So, as I mentioned there is a large diversity omics that we are currently studying in biomedicine. So, this slide was taken from the TCGA, which I am going to talk a bit about. And, then I added some extra slides to represent proteomics and phosphoproteomics here, but there we can look at the genome which is really the long term storage information of the cell. The transcriptome which is the retrieval of this information: the proteome which is the short term information storage and then how the signaling networks interaction, the interactome which we will also touch on.

So, here is just a lot of the different levels of data that we can gather and at this point and so, then in terms of next gen sequencing, we are really covering all from the micro RNA up through the mutation calls and that is what we are going to be; we are going to be focusing on.

(Refer Slide Time: 05:11)



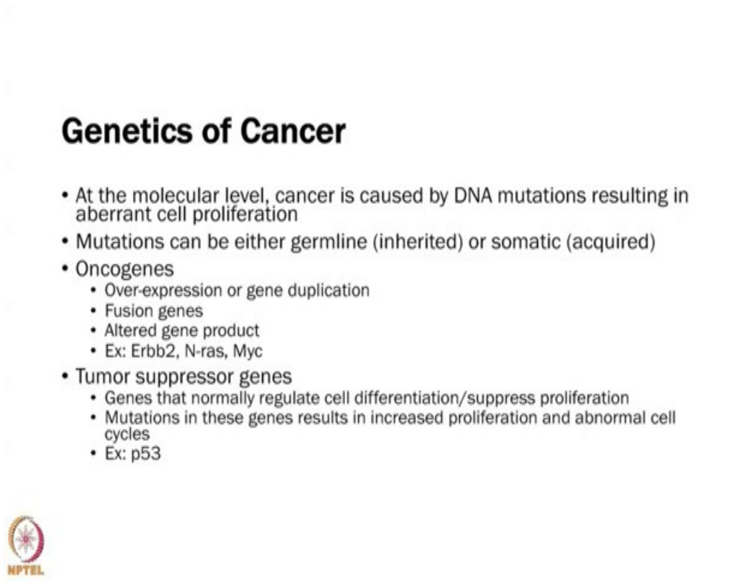
Genetics of Cancer

- At the molecular level, cancer is caused by DNA mutations resulting in aberrant cell proliferation
- Mutations can be either germline (inherited) or somatic (acquired)
- Oncogenes
  - Over-expression or gene duplication
  - Fusion genes
  - Altered gene product
  - Ex: Erbb2, N-ras, Myc
- Tumor suppressor genes
  - Genes that normally regulate cell differentiation/suppress proliferation
  - Mutations in these genes results in increased proliferation and abnormal cell cycles
  - Ex: p53

There has been a long history of studying to the genetics of cancer, which predates the proteogenomics of cancer and so, the reason why genomics and cancer have become such an enormous field is, because its known that at the molecular level cancer is caused by mutations that results in aberrant cell proliferation.

And, these mutations can either be germline, meaning that they are inherited or they can be somatic meaning that they are acquired at some point in life. And these mutations, if they are occurring in oncogenes they can cause over expression of these oncogenes, oncogenes meaning that they increase oncogenesis, increasing tumor proliferation, they can be fusion genes, they can be an alter gene product.

(Refer Slide Time: 05:57)

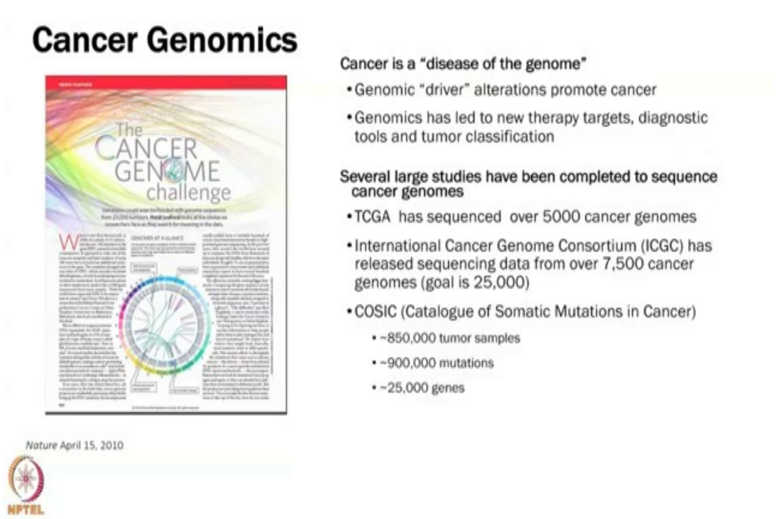


Genetics of Cancer

- At the molecular level, cancer is caused by DNA mutations resulting in aberrant cell proliferation
- Mutations can be either germline (inherited) or somatic (acquired)
- Oncogenes
  - Over-expression or gene duplication
  - Fusion genes
  - Altered gene product
  - Ex: Erbb2, N-ras, Myc
- Tumor suppressor genes
  - Genes that normally regulate cell differentiation/suppress proliferation
  - Mutations in these genes results in increased proliferation and abnormal cell cycles
  - Ex: p53

So, there is many examples of this, that we will cover some of them in the hands on. So, such as Erbb2 it is just one we are going to talk about, specifically in the genomics hands on session and then there is tumor suppressor genes. So, these are genes that normally, regulate cell differentiation and suppress the proliferation of a cell. So, if those so, if there are mutations in these genes, there is an increase in that would reduce them as an increase in proliferation. So, you can we look at both tumor suppressor genes and oncogenes in terms of the mutations in the cell.

(Refer Slide Time: 06:43)



Cancer Genomics

The CANCER GENOME challenge

Nature April 15, 2010

Cancer is a "disease of the genome"

- Genomic "driver" alterations promote cancer
- Genomics has led to new therapy targets, diagnostic tools and tumor classification

Several large studies have been completed to sequence cancer genomes

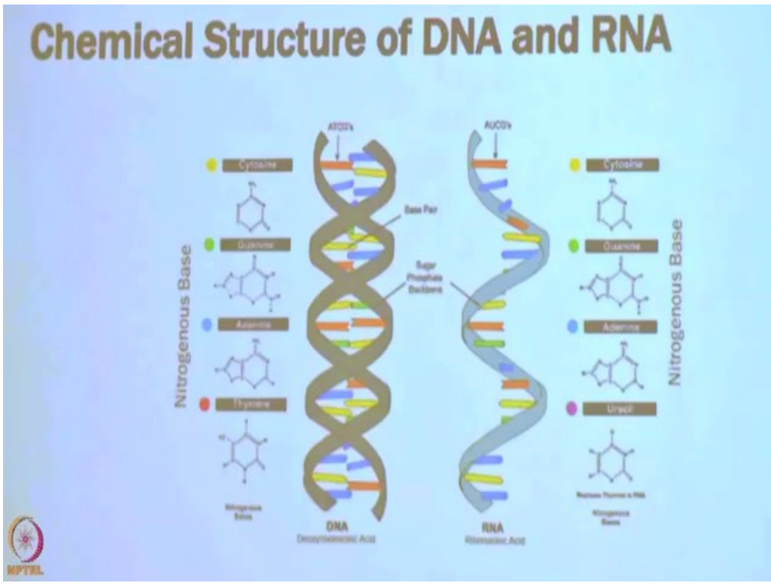
- TCGA has sequenced over 5000 cancer genomes
- International Cancer Genome Consortium (ICGC) has released sequencing data from over 7,500 cancer genomes (goal is 25,000)
- COSIC (Catalogue of Somatic Mutations in Cancer)
  - ~850,000 tumor samples
  - ~900,000 mutations
  - ~25,000 genes

And as I mentioned, there is many genomic drivers, this was in 2010 there was a I think this is a Nature, yeah so Nature. There is there was an article about the cancer genome and the TCGA, which is the cancer genome atlas which I am going to cover in a bit of more detail later, really was one of the projects that spearheaded this.

So, they sequence actually, this is an outdated number I think it is more like 11,000 cancer genomes over 33 different kinds of cancer. So, it is a really wonderful data set that everyone has access to and there is also the International Cancer Genome Consortium, which is another large project that has worked to collect lots and lots of genomics data for cancer that you can also access.

And, then there is what let us call it oh there is a typo there and I apologize COSMIC, which is the Catalog of Somatic Mutations in Cancer. So, this is just cataloging all of the mutations that have been found to occur in tumors as well. So, this is another really good resource that is publicly available. So, now I am going to jump into genomics, I think a lot of you probably know this, but I really I should probably have a slide in there just, because you know just in case otherwise you will be completely lost after this.

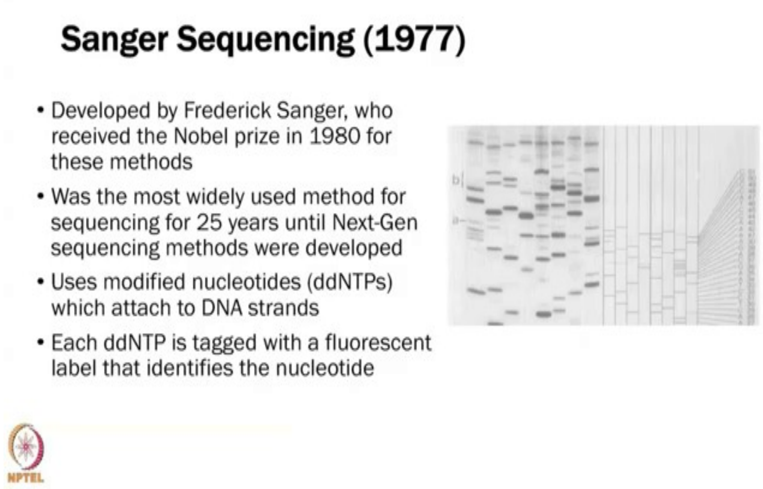
(Refer Slide Time: 07:54)



Chemical Structure of DNA and RNA

So, the chemical structure of DNA and RNA. So, DNA is double stranded, RNA is single stranded and there is these nitrogenous bases that make up nucleotides. They differ between DNA and RNA and that there is a thymine in DNA and uracil in RNA, triplets of these bases encode different amino acids. So, that is something that we are going to go into. So, just two slides on the history of sequencing just, because I think it is important to start there.

(Refer Slide Time: 08:30)

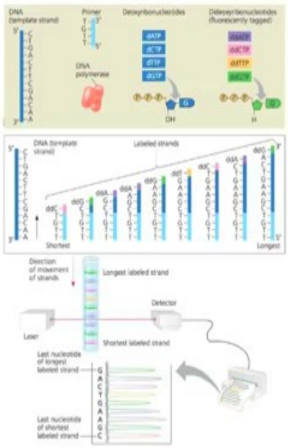


Sanger Sequencing (1977)

- Developed by Frederick Sanger, who received the Nobel prize in 1980 for these methods
- Was the most widely used method for sequencing for 25 years until Next-Gen sequencing methods were developed
- Uses modified nucleotides (ddNTPs) which attach to DNA strands
- Each ddNTP is tagged with a fluorescent label that identifies the nucleotide

So, Sanger sequencing, which how many people have heard of Sanger sequencing? Ok, great wonderful. So, this was developed by Frederick Sanger, who received the Nobel prize in 1980 for these methods and it was the most widely used method for until Next-Gen sequencing came around and we are going to spend most of our time talking about Next-Gen sequencing. So, what this method did was use these modified nucleotides, which attached to the DNA strands and each of them was tagged with a fluorescent label that identified which nucleotide it was. So, I will show a schematic of this.

(Refer Slide Time: 09:08)



Denature DNA + ddNTPs + 3′ primer

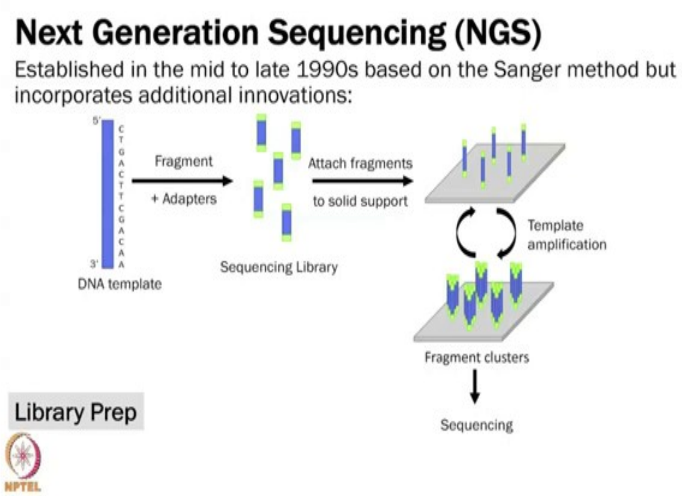
Strand is synthesized until a ddNTP is inserted at random, preventing further elongation

Labeled strands are passed through a polyacrylaminde gel to separate strands by size and then a fluorescent detector can sense the color of each strand

So, here we have our DNA template. So, the double stranded DNA was denature to get a single stranded DNA. And, then you add in these nucleotides that have either a non-fluorescent or fluorescent tag on them and then you have a DNA polymerase. So, it allows the strands to grow and becomes a double stranded, because DNA polymerase is adding on these deoxyribonucleotides and whenever it gets to one where there is a fluorescent probe, it stops.

So, it keeps growing and then these fluorescently tagged ddNTPs are randomly added. So, as soon as they are added on it stops and then so, you know that the length this 6 nucleotides up is a C and 7 is a G and etcetera and then you run this on a poly acrylamide gel, separate out the strands and then you can use the florescent detector to figure out at what length each of the different nucleotides is at. So, this worked really well, I just took a really long time.

(Refer Slide Time: 10:19)



Next Generation Sequencing (NGS)

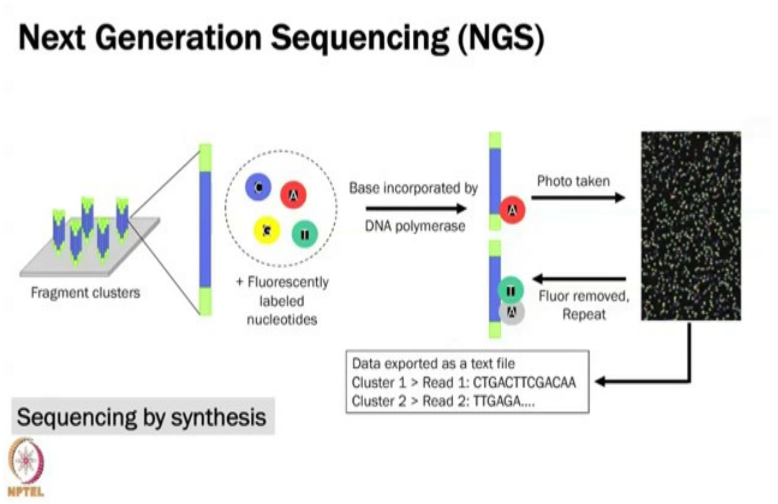
Established in the mid to late 1990s based on the Sanger method but incorporates additional innovations:

Library Prep

So, eventually Next-Gen sequencing came along. So, this was established in the mid to late 90s and it was based on this Sanger method, but it incorporated some new innovations. So, how many people are you have used like Illumina sequencing for example? So, there were more Sanger sequencing people here, than Illumina yeah. So now, what happens is you still have this DNA template. So, it is a single stranded DNA template and it is fragmented and there are adapters that are put on the template.

And, we are going to talk a lot about these adapters, because you can do some cool stuff with them and then what happens is the these fragments are attached to some a solid support like a flow cell and we will talk about that as well and then there is an amplification that occurs. So, that you stick your fragment on to this, flow cell and then you amplify it. So, you end up getting a cluster of the same, of the exact same sequence that is stuck in a certain part of your flow cell. And then you so, this is what is called library prep. So, this is how the library is prepped from your DNA template.

(Refer Slide Time: 11:39)



Next Generation Sequencing (NGS)

Fragment clusters
+ Fluorescently labeled nucleotides
Base incorporated by DNA polymerase
Photo taken
Fluor removed, Repeat
NPTEL

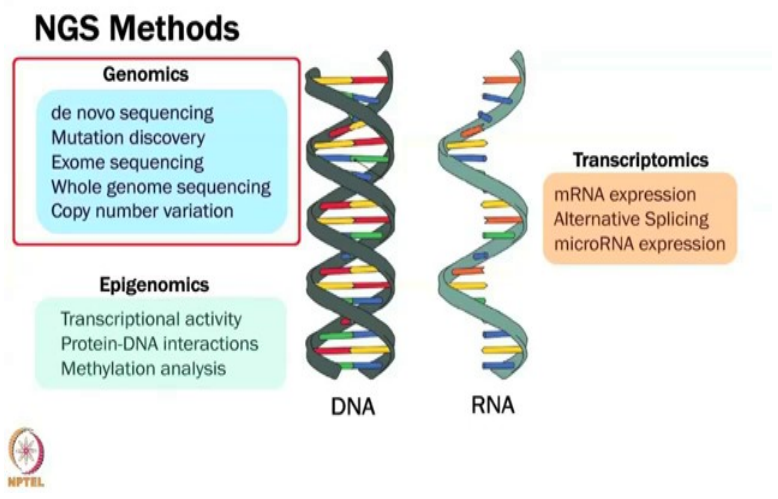
Data exported as a text file
Cluster 1 > Read 1: CTGACTTCGACAA
Cluster 2 > Read 2: TTGAGA....

Sequencing by synthesis

And, then what happens as we do what is called sequencing by synthesis. So, it is similar to what I mentioned in the Sanger method except now, these fluorescently label nucleotides, you can actually you can you when they bind to the DNA they stop the synthesis, but then you can remove the fluorescent probe so that you can continue to grow it. So, you do not in the Sanger sequencing once it hit there it stopped. Now, you can take that floor off. So, you can keep growing instead of having to do all of the different lengths and then running it through a gel.

Now, you can just keep adding more and more on and just taking a photo on every time the floor is put on, you take a photo, you remove it, etcetera etcetera until you grow it all the way up the strand and then the data is exported as a text file. So, it reads out what cluster it is, what read it is and then the actual sequence. So, this is a much faster and more efficient method than the Sanger sequencing. What we are going to talk about is a lot of these different methods ok.

(Refer Slide Time: 12:49)



NGS Methods

Genomics

de novo sequencing
Mutation discovery
Exome sequencing
Whole genome sequencing
Copy number variation

Epigenomics

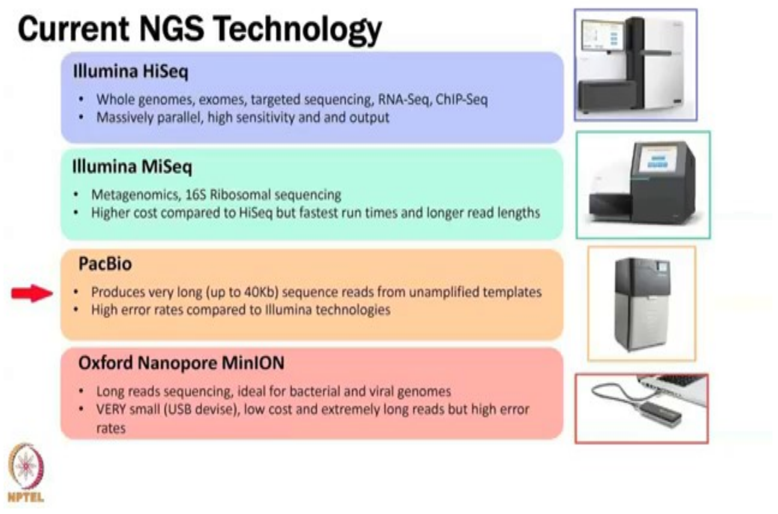
Transcriptional activity
Protein-DNA interactions
Methylation analysis

Transcriptomics

mRNA expression
Alternative Splicing
microRNA expression

So, genomics we have de novo sequencing, mutation discovery, exome and whole genome sequencing and copy number alteration or variation detection. Epigenomics we have assays to look at transcriptional activity, protein DNA interactions, and methylation analysis. And, in this transcriptomics level, we can look at mRNA expression, alternative splicing and micro RNA expression and there is more, these are just some of the basics.

(Refer Slide Time: 13:20)



Current NGS Technology

Illumina HiSeq

- Whole genomes, exomes, targeted sequencing, RNA-Seq, ChIP-Seq
- Massively parallel, high sensitivity and and output

Illumina MiSeq

- Metagenomics, 16S Ribosomal sequencing
- Higher cost compared to HiSeq but fastest run times and longer read lengths

PacBio

- Produces very long (up to 40Kb) sequence reads from unamplified templates
- High error rates compared to Illumina technologies

Oxford Nanopore MinION

- Long reads sequencing, ideal for bacterial and viral genomes
- VERY small (USB devise), low cost and extremely long reads but high error rates

So, I included some current next-gen sequencing technology that is being used at the moment and sort of the newest instruments. So, the most common one at this point is

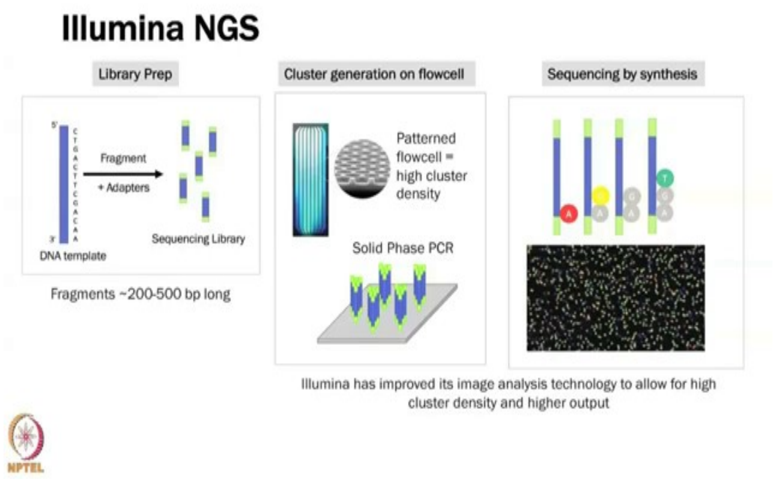
Illumina HiSeq. So, the this one, I am going to talk about why it is the most commonly used and what technologies, they have come up with that have made it the most popular. So, Illumina HiSeq it is commonly used for the whole genome and exome sequencing, targeted sequencing, RNA sequencing, ChiP-Seq.

And, it is really you can parallelize a lot of samples at once, it is pretty fast and it is highly sensitive. It has a high output and then there is the Illlumina MiSeq, which I wanted to introduce, but we will not talk about it much, because it is not used much for cancer. It is mostly used for metagenomics so, microbiome data, small bacterial sequences.

So, or 16S ribosomal sequencing and so, the cost is higher compared to the HiSeq, but it is faster and has longer read lengths, which makes it good for these, these other smaller genomes that are less annotated. There is PacBio and I am going to go into each of these in more detail, which produces very long sequence reads, but it has very high error rates. At this point frequently used in combination with Illumina, because they offer different strengths and then there is the Oxford Nanopore MinION which is super cool. Has anyone seen one of these?

They are like USBs which makes them pretty exciting and they are also really good for long reads, they are ideal for that reason for this bacterial and viral genomes. They are very small, they have historically had very high error rates, but they are trying to improve this; we will talk about that a little bit more too.

(Refer Slide Time: 25:24)



Illumina NGS

Library Prep

Cluster generation on flowcell

Sequencing by synthesis

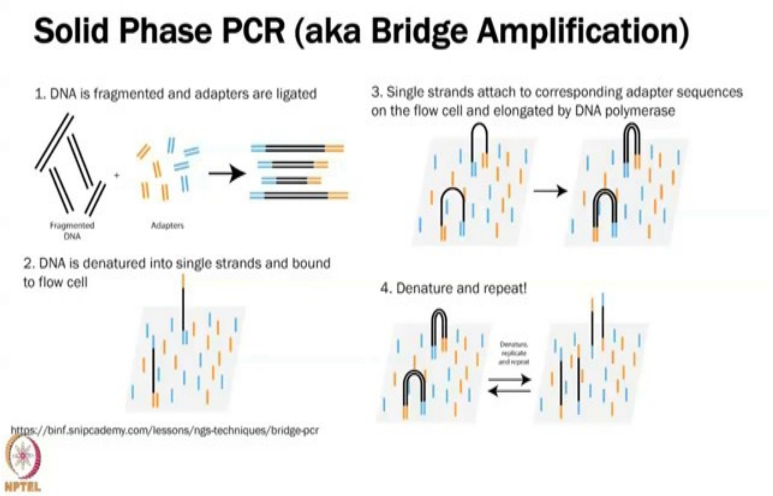
Fragments ~200-500 bp long

Illumina has improved its image analysis technology to allow for high cluster density and higher output

So, for Illumina we talked about the library prep; so, it happens in the same way that we discussed before where you fragment them. So, here I found it a couple of different; here it says 200 to 500, but I have seen different quote numbers for how long these fragments are. And then so, you then do this cluster generation, on this flow cell and the flow cell is really what I think made Illumina the top of the market for this. So, they created this pattern flow cell; so, you could get really clustered high cluster density and this increased the output of the actual sequencing analysis.

So, once you do this the cluster generation on the flow cell and I am going to talk a little bit about what this solid phase PCR is on the next slide. You do the same sync with the sequencing by synthesis that we talked about before, where you just keep adding on different nucleotides and just taking the picture of them as we go ok.

(Refer Slide Time: 16:28)



Solid Phase PCR (aka Bridge Amplification)

1. DNA is fragmented and adapters are ligated

2. DNA is denatured into single strands and bound to flow cell

3. Single strands attach to corresponding adapter sequences on the flow cell and elongated by DNA polymerase

4. Denature and repeat!

https://binf.snipcademy.com/lessons/ngs-techniques/bridge-pcr

So, the solid phase PCR which is also known as bridge amplification. What happens is you have your DNA that is fragmented and then you have these adapters. So, again the adapters are put on so, that you can attach it to the flow cell and then the DNA is denatured and it is a single strand. So, you break it into single strands and then you attach it to the flow cell and then the single strands, they have these other adapters on that are sitting on the flow cells.

So, what happens is the single strands end up bridging over to attach to the complimentary adapter sequence. So, it creates this bridge and then it is elongated by using DNA polymerase. So, it creates a double strand and then it is denatured again. So, then it creates single strands again and then it bridges over again and it keeps creating these more and more of these clusters.

Student: Killy.

Yes, yeah

Student: From this image what it appears is that the adapters are for the 5 prime and pre prime image.

Correct.

Student: So, when you put it on the plate why do not you put just the 5 prime on so that all the orientation will be in one end, because here it seems from the diagram that both the 5 and the pre prime are plated.

So, you are saying why not just put like all 5 prime adapters on and all. So, I do not think you can control which ones are going to stick, but.

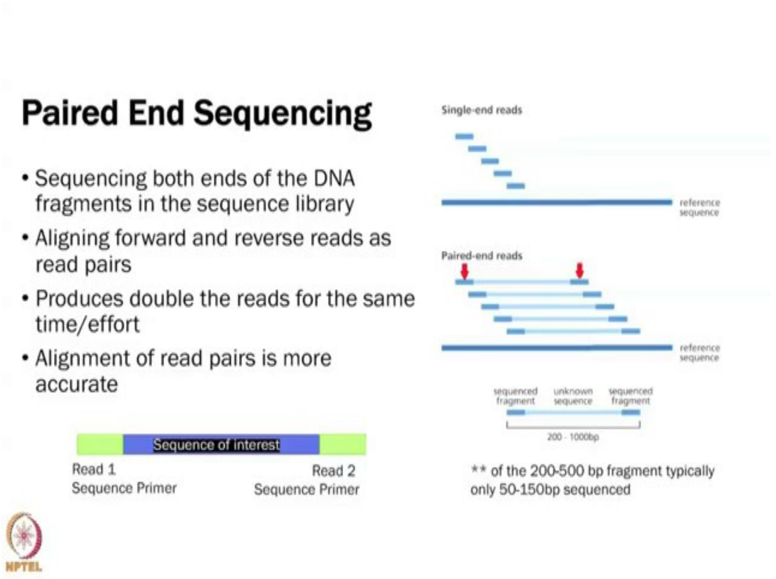
Student: If the 3 prime adapters are on the plate only the 3 prime are attached.

Yeah.

Student: So, they are all in the same orientation.

But the adapters themselves attached to the plate regardless oh, I see what you are saying why you not I think it is actually, because you want; it is a good question, I will look into it more, but my guess is that, because you want to get both. If you do single stranded you want to get it from both directions. So, paradin versus single reads which we will talk about; I have a feeling that is why, but I do not, I will actually look into it. So, thank you that led me perfectly to my next slide.

(Refer Slide Time: 18:44)



Paired End Sequencing

- Sequencing both ends of the DNA fragments in the sequence library
- Aligning forward and reverse reads as read pairs
- Produces double the reads for the same time/effort
- Alignment of read pairs is more accurate

** of the 200-500 bp fragment typically only 50-150bp sequenced

So, paired end sequencing versus single end sequencing, paired end sequencing is now, really the norm for sequencing, but I think some people likely still use single end reads as well. So, I figure we should talk about it. So, what it means is that you are sequencing both ends of the fragment. So, as I mentioned the fragments are let us say they are 500 base pairs long, typically when you based on the reagents that are used for Illumina, you only measure 50 to 200 of those base pairs. So, you are not measuring the full fragment, you are just measuring the ends of the fragment. So, if you do it just from one end right, you are only going to be measuring, let us say those 100 base pairs on one end of your fragment, one of the first let us say the 3 prime end.

But, if you read both the forward and the reverse right, 100 on both sides then you have a 300 base pair gap, but you have 3 2 you have 100 and 100 on both sides and you know that they are from the same strand. So, you know that that gap exists, but you can get both. You can get more information from that cluster than you would normally, if you just did it from one direction. So, really you are producing double the reads for the same amount of time.

So, it is a little more, it is more expensive which is the limitation, but if you can do it is a better use of your sample and your time and the alignment is much more accurate and we will talk about alignment. So, when you actually take these reads and you try to align it to a reference genome, you have a lot more information on what, how to align it, because you have even though you have that gap, you know exactly what is on both sides of that gap.

So, how this is done and you sort of alluded to this in your question, there is a sequence primer that is specific to the 3 prime and the 5 prime end and that is included in this adapter yeah. So, that is how the pair agree and reads work. Any other questions on this? Yeah.

Student: So, if you must know what is genomes sequencer whatever generation to the next generation yeah, yes and specifically designed datasets like primers, like 100 or 150 gap.

Say it sorry, say that again.

Student: So, how do you specifically produce, specifically designed adapters.

Yeah. So, the adapter is.

Student: Only 100 or 150.

Oh no. So, that the fact that it only reads a 100 or 150 is based on that actually based on how much of the it is like the Illumina gives you a certain amount of DDMTPs to actually make the read. So, it is more about the reagents you use than the primers.

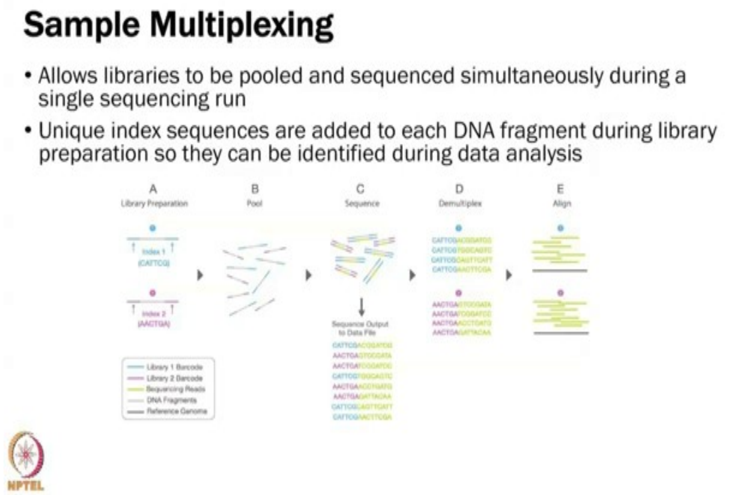
Student: So, it is not random, it is not random.

It is not random no, but no it is not random the 150 is not random, because you tell the Illumina how many times to the instrument how many times to take a picture right. You are going to say take a 150 pictures and then stop, because I know that my reagent can get me to 150.

Student: Ok.

Yeah, that is not random you will always have 150, if you decide to do a 150.

(Refer Slide Time: 22:35)



Sample Multiplexing

- Allows libraries to be pooled and sequenced simultaneously during a single sequencing run
- Unique index sequences are added to each DNA fragment during library preparation so they can be identified during data analysis

So, another thing you can stick in the adaptors is a way of actually multiplexing your samples, which means you are pooling lots and lots of samples together. So, you can run them all at once instead of just running one sample at a time. So, what is done is there is a unique sequence that is added to the adapter that indicates, what sample that that sequence came from.

So, during the library prep you add these unique sequences on and then that way when you mix them all up and then you do all of your library, the library preps done and then into your sequencing. After you have your data, you have this sequence that identifies. So, the blues came from you know sample 1 and the pinks came from sample 2 and then you can just in terms of the bioinformatics you can go in and pull out which ones came from which sample. So, this is a very common technique at this point as well yeah.

Student: So, during video sequencing versus your targeted sequencing library sequencing paired end would be better compared to paired end would be better?

Paired end would be better for targeted. I think paired end is always better right, if you can afford it. It is always the limitation for these things right, is that your question yeah.

Student: Yeah together paired end will enhance signal.

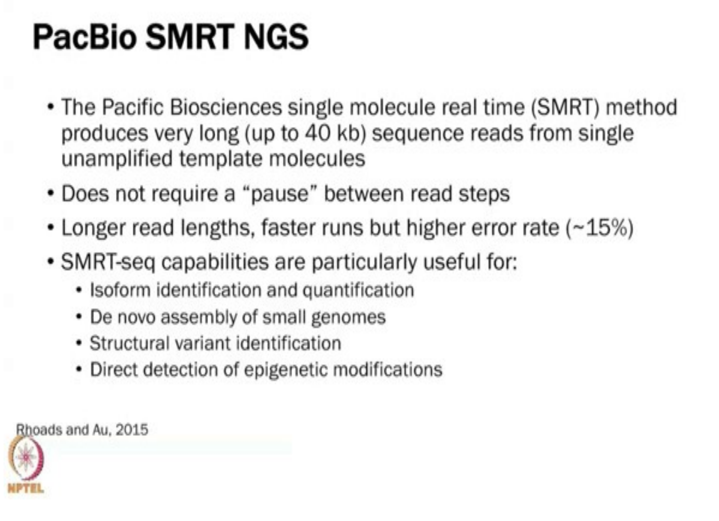
Yeah, paired ends is always better if you can swing it.

Student: You are doing targeted or DDM.

Yeah, yeah.

Student: Thank you.

(Refer Slide Time: 24:18)



PacBio SMRT NGS

- The Pacific Biosciences single molecule real time (SMRT) method produces very long (up to 40 kb) sequence reads from single unamplified template molecules
- Does not require a "pause" between read steps
- Longer read lengths, faster runs but higher error rate (~15%)
- SMRT-seq capabilities are particularly useful for:
  - Isoform identification and quantification
  - De novo assembly of small genomes
  - Structural variant identification
  - Direct detection of epigenetic modifications

Rhoads and Au, 2015

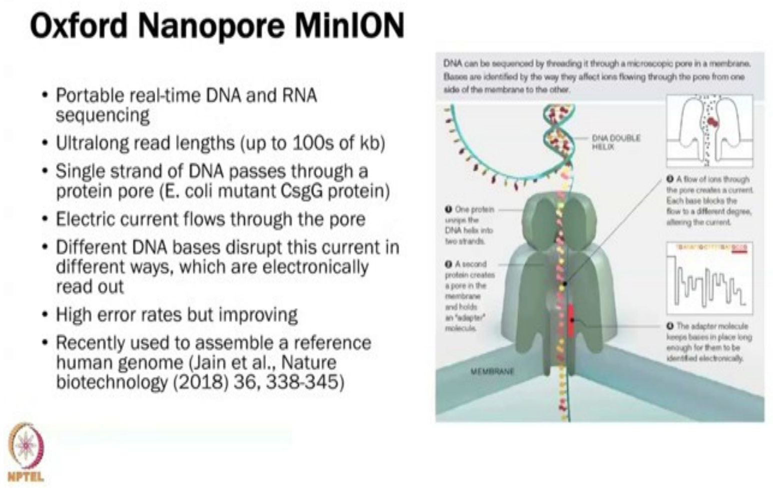
So, I did want to touch on the two other sequencing platforms that are not as commonly used, but I think they are going to become more and more present. So, I wanted to talk about them a little bit. So, there is this PacBio, may use what’s called a SMRT method, that can produce really long reads. So, the limitation with Illumina is as we mentioned that it will only go up to about 200 base pairs. So, when you are aligning and you are trying to figure out where those sequences belong on the genome; you do not have a lot of information, you only have those base pairs.

But, if you have you know 40 kb then you have a lot of information and you can do de novo sequencing, you can get a lot of information about let us say you have an unknown species that you know nothing about, you do not have a genome. So, you can gather these long reads and then try and do de novo alignment and figure out the actual reference genome of a new species. So, that is where these things typically come in.

And, another this is I am not going to go into all the methods behind this, I did include if you are interested there is this paper that goes into a lot of the details on it, but it does not require this pause between the reading of a step. So, in Illumina you know you have the floor is added on and you have to take a picture and you have to take the floor off and then you put a new one on and you take a picture. So, there is this pause step every time that you add anyone on.

So, in this method that is not that does not happen. So, it is a lot faster so, unfortunately though it has a very high error rate. So, 15 percent is pretty bad. So, I think they are likely working on lowering this, but it is something to keep in mind. But, it is still really good for de novo assembly of small genomes, structural variant identification and you can actually, directly identify epigenetic modifications without having to do another completely different type of nomex analysis, which is very cool.

(Refer Slide Time: 26:25)



Oxford Nanopore MinION

- Portable real-time DNA and RNA sequencing
- Ultralong read lengths (up to 100s of kb)
- Single strand of DNA passes through a protein pore (E. coli mutant CsgG protein)
- Electric current flows through the pore
- Different DNA bases disrupt this current in different ways, which are electronically read out
- High error rates but improving
- Recently used to assemble a reference human genome (Jain et al., Nature biotechnology (2018) 36, 338-345)

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

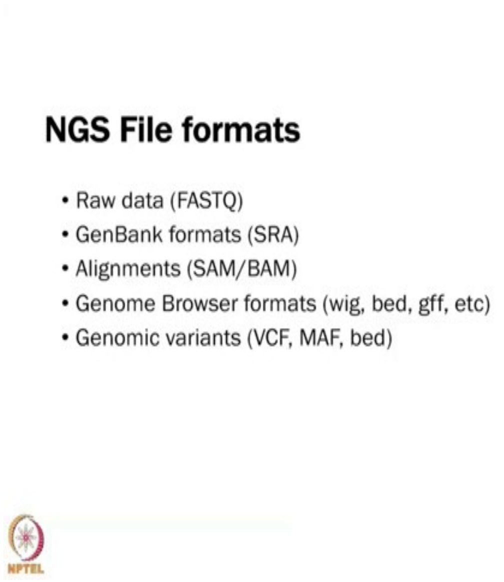
And, then the other one which is actually quite similar in terms of some of its benefits is this Oxford MinION which again is this USB science sequencer and it can do really ultra long read lengths. So, up to 100s of kb and what happens is the way that they do this is, there is this protein pore that is actually a E.coli mutant of a CsgG protein. And they figured out that if you put this on a membrane you can actually, read the DNA strand through this pore.

And, the different bases, depending on what base is coming through the pore, it disrupts the current in a specific way that you can read electronically. So, I think this is super cool, I think that the problem that I have heard from people who are more involved in the actual instrumentation of the field is that it actually goes too fast. So, the reason why there are high error rate is, because it is moving too quickly and they are trying to figure out how to slow it down. And, I think it is from what I have read it has got a lot better and there was a recent paper, I could hear that it came out this year that actually used this to assemble, a reference a human reference genome.

So, it was just kind of a proof of concept paper to show that you could use this to do some of the stuff that we have previously, just used Illumina for. So, I think that this is a very exciting development and that in Next-Gen sequencing especially, because you can just hold it and take it places and sequence I guess whatever you want to. So, there are a lot of different file formats, that we will talk about just, because I you know I sit in the data analysis side of most of this.

So, I do not typically do I do work with people who do sequencing and I work with the genomics technology center at NYU a lot and so, I am familiar what they do. But, I typically just get the data after, that is what I prefer; so, I wanted to talk a little bit about file formats. So, the raw data is typically, in what is called a FASTQ format, there is a GenBank format, which is an SRA, there is alignment formats, which are SAM's and BAM's. There are these genome browser formats and then there is genomic variant formats and we will talk about each of these in some level of detail.

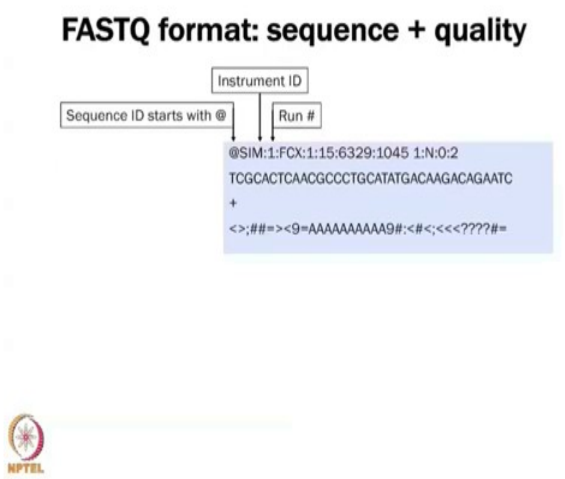
(Refer Slide Time: 28:55)



NGS File formats

- Raw data (FASTQ)
- GenBank formats (SRA)
- Alignments (SAM/BAM)
- Genome Browser formats (wig, bed, gff, etc)
- Genomic variants (VCF, MAF, bed)

(Refer Slide Time: 29:05)

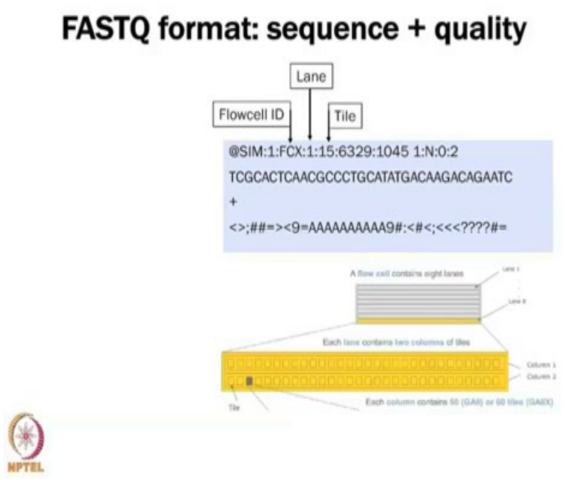


FASTQ format: sequence + quality

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

So, the FASTQ has anyone seen a FASTQ, has everyone seen a FASTQ? Who is seen a FASTQ? Ok, some people ok. So, this is what this is an example of what it looks like. So, it just starts with information on the actual instrument and then run number and then it has information on the flow cell.

(Refer Slide Time: 29:27)



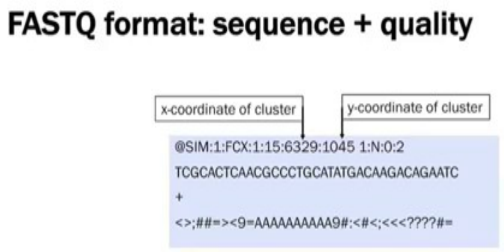
FASTQ format: sequence + quality

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

A flow cell contains eight lanes
Lane 1
Lane 8
Each lane contains two columns of tiles
Column 1
Column 2
Tile
Each column contains 50 (GAII) or 60 tiles (GAIIX)

So, this is an Illumina, FASTQ it depends on the instrument, just showing you the Illumina; Illumina FASTQ because I think that is again the most common for cancer genomics. So, the flow cell ID, the lane and the tile. So, you can see here if there is a flow cell contains 8 lanes and then each lane has 2 columns of tiles. So, you really get exactly where this sequence was in terms of the flow cell coordinates.

(Refer Slide Time: 29:55)

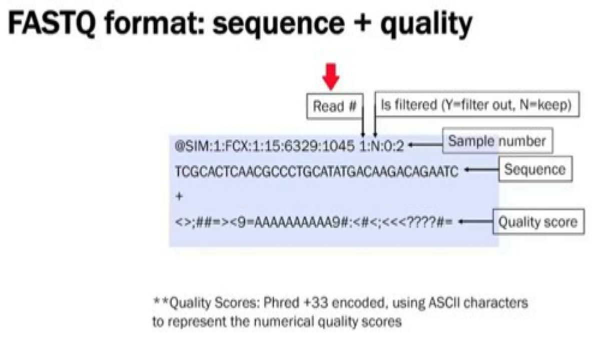


FASTQ format: sequence + quality

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

And, then you also get an x and a y coordinate of the cluster on the tile.

(Refer Slide Time: 30:03)



FASTQ format: sequence + quality

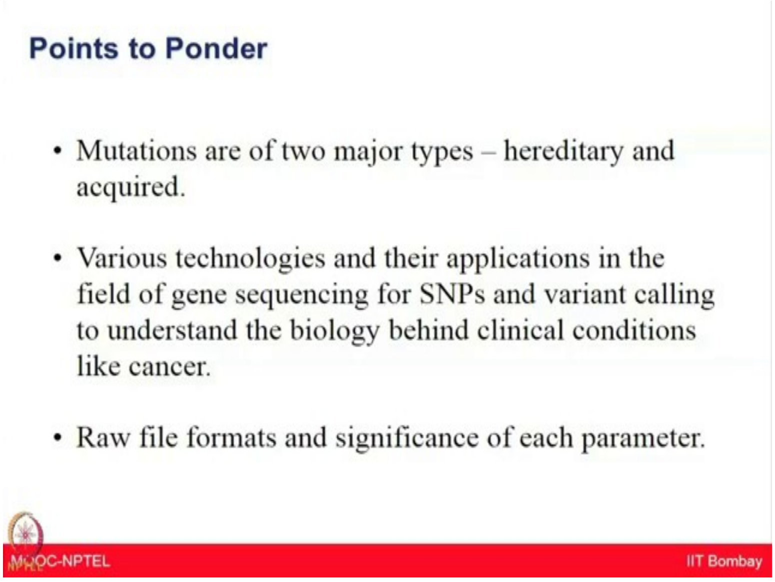
```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

**Quality Scores: Phred +33 encoded, using ASCII characters to represent the numerical quality scores

You get a read number. So, if it is single reads, you will only have a 1, if it is paired end you will have a 1 and a 2. And, then if it is filtered out for quality reasons you get a yes here, if it is kept you get a no, and then there is a sample number. And, then there is the actual sequence and then you get this quality score, which is a it is an encoded score using ASCII characters to represent the quality score of each of the bases.

So, it ranges from 0 to 40 in terms of quality and you can use there, its just a key that goes along with Illumina that just shows exactly what each of these corresponds to so.

(Refer Slide Time: 30:46)



Points to Ponder

- Mutations are of two major types – hereditary and acquired.
- Various technologies and their applications in the field of gene sequencing for SNPs and variant calling to understand the biology behind clinical conditions like cancer.
- Raw file formats and significance of each parameter.

MOOC-NPTEL
IIT Bombay

In today’s lecture you got an overview of genomic technologies and especially, its relevance in context of cancer and how it could actually provide a broader overview to understand any disease or clinical conditions. You also learnt, why in genomic sequencing two side reads are important as compared to the single reads and how it can affect the accuracy of sequencing.

You also learned about gene sequencing, how it has evolved over the years and now, we have not only achieved much higher accuracy of sequencing, in a much shorter time frame and in a very-very cost effective manner, but also the instruments have become much more miniaturized. And, one could see the examples like Oxford, Nanopore technology MinION, which is very easy to carry and transport anywhere as per the requirements.

Dr. Kelly has also introduced and made you familiarized with availability of datasets and how to obtain different type of raw files and format them for various parameters, for further analysis. Let us continue the next lecture in the same flow of genomics, where Dr. Kelly is going to talk to you about the sequencing alignments and various factors which affects that.

Thank you.