An Introduction to Proteogenomics

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

Lecture - 04
Introduction to Genomics - II

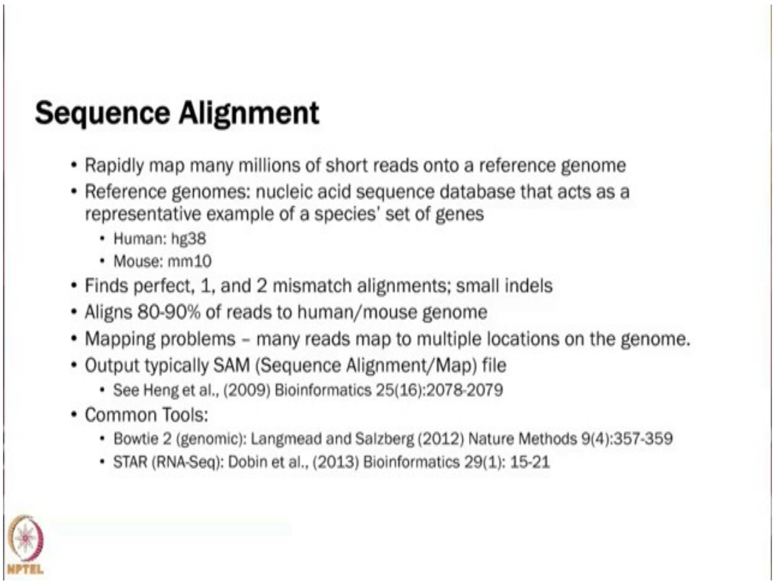
Welcome to MOOC course Introduction to Proteogenomics. From the last lecture, Dr. Kelly Ruggles provided you an overview of genomics and genomic technologies, how they are making revolutions for various diseases; especially in context of cancer. Today is going to be second lecture by Dr. Kelly Ruggles and she will talk to you about sequence alignments with respect to the reference genome. Also, what are the terminologies like coverage and depth refers in terms of alignment of genes to the referral genome.

She will talk to you about exome and whole genome sequencing, how it helps to understand the genome and different variations like; copy number variants and mutational status which may lead to various clinical conditions. The lecture will also describe about various type of SNP's, SNP arrays and applications of GWAS or Genome Wide Associational Studies in a population with reference to disease.

She will then cover about transcriptomic fields which is going to look beyond the genome, how the transcript of form, how you can study RNA expression. And, by using the RNA sequencing data or NGS technologies and various applications how one could get some functional information just looking beyond the genome. So, let us welcome Dr. Kelly Ruggles for her second lecture ok.

So we are going to just continue where we left off so, at this point I have sort of walked us through getting to the fast cube raw data files and now we are going to talk about what we do with it, once we have those data files for all different kinds of omics analysis. So, the first thing that we do is we have to align these sequences to a reference genome, if a reference genome exists. I am going to assume for the purpose of this that we have a reference genome and so, what that means, is you take the short sequences and you match it against a genome that represents whatever species that we are looking at.

(Refer Slide Time: 02:49)



Sequence Alignment

- Rapidly map many millions of short reads onto a reference genome
- Reference genomes: nucleic acid sequence database that acts as a representative example of a species' set of genes
  - Human: hg38
  - Mouse: mm10
- Finds perfect, 1, and 2 mismatch alignments; small indels
- Aligns 80-90% of reads to human/mouse genome
- Mapping problems – many reads map to multiple locations on the genome.
- Output typically SAM (Sequence Alignment/Map) file
  - See Heng et al., (2009) Bioinformatics 25(16):2078-2079
- Common Tools:
  - Bowtie 2 (genomic): Langmead and Salzberg (2012) Nature Methods 9(4):357-359
  - STAR (RNA-Seq): Dobin et al., (2013) Bioinformatics 29(1): 15-21

So, for example, for humans there is a reference genome, the current updated reference genome is hg38. The version before this was hg19; lot of things still are in hg19, some things are hg38. This is something that if you are doing an alignment or you are using data that is aligned to a genome, you should definitely always check the reference genome, because it will completely mess you up if you use the wrong reference genome and you assume it is one and it is actually aligned to the other.

So, reference genome is just a sequence database that acts as a representative sample of a species and so, as I mentioned for human, the current version is hg38. There is a mouse version mm10, there is a whole every species that has been sequenced, has gene a reference genome you can look up. If these are not including your favorites species to work with and what the alignment does is it finds perfects matches. So, anything where that 150 or 200 base pairs perfectly aligns to the genome or it can allow for a certain amount of mismatches and depending on the aligner and the settings, you can sort of put in how much mismatch you will want to allow.

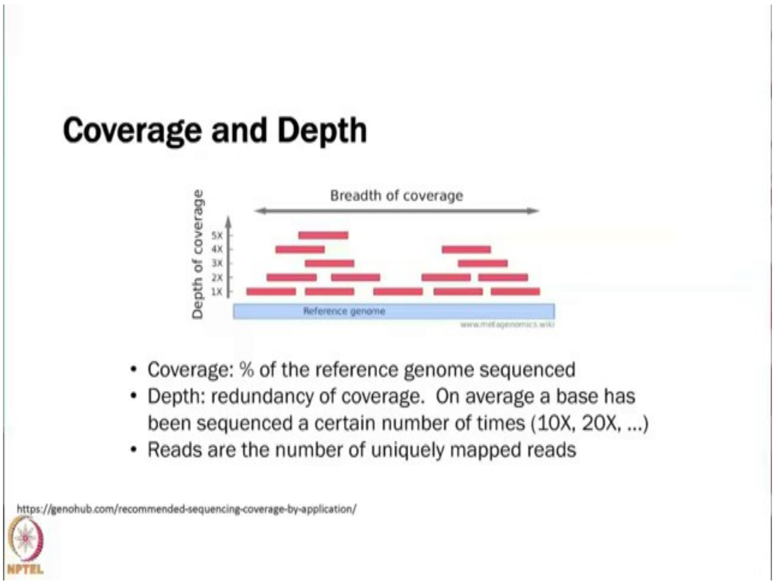
And then, if you use let us say an alumina you can get about 80 and 90 percent of the reads that the map to either human or a mouse genome. There are a lot of problems that occur you know if you have a chunk that maps many different parts of the genome, you do not know which one it came from. I mean there are you can read about depending on which aligner you are using sort of the limitations and the strengths of each of the different aligners. And typically the output is what is called a SAM file, which is a sequence alignment map file. I am not going to go into the details about this. We could have a whole day on SAM files, but if you want to learn more I did include reference here that is pretty thorough in terms of the SAM file format.

And, then in there are some common tools I mentioned you here so Bowtie which is typically now used for genomic alignment and STAR which is used for RNA seq. I also included here some references if you are interested in learning more about either of those, we do not have time to go through all of the details about them today, but I did want to mention them. Has anyone used Bowtie?

STAR? No. Neither, great all new.

So, this is that is good. So, and so, something to keep in mind here to is you hear lot about coverage and depth, when you hear about next gen sequencing and what that really means is the coverage is the percent of the reference genome that you were able to sequence and then the depth is the redundancy of coverage.

(Refer Slide Time: 05:35)



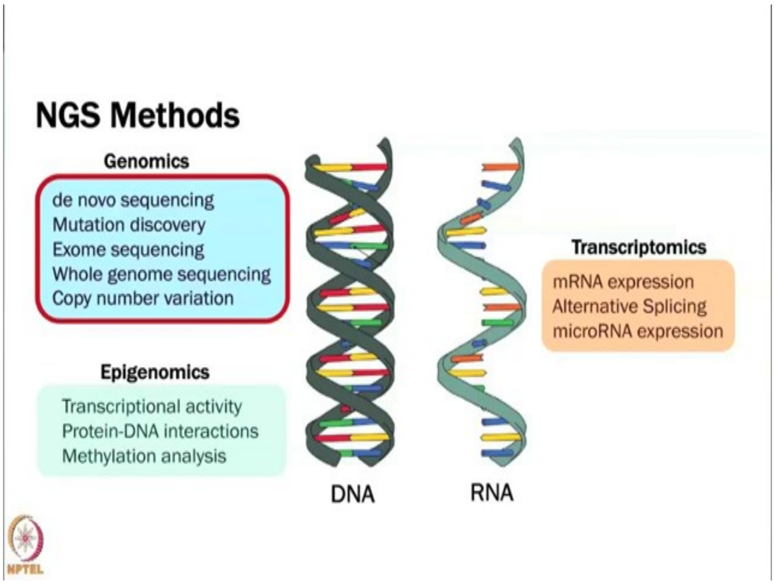
Coverage and Depth

- Coverage: % of the reference genome sequenced
- Depth: redundancy of coverage. On average a base has been sequenced a certain number of times (10X, 20X, ...)
- Reads are the number of uniquely mapped reads

https://genohub.com/recommended-sequencing-coverage-by-application/

So, how many reads you were able to get out of this so, on average at a certain point within the reference genome. So, for example, 10X would mean 10 reads on average that you are able to cover across the reference genome. And, the reads are the number of uniquely mapped reads here ok.

(Refer Slide Time: 05:58)



NGS Methods

Genomics

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

Epigenomics

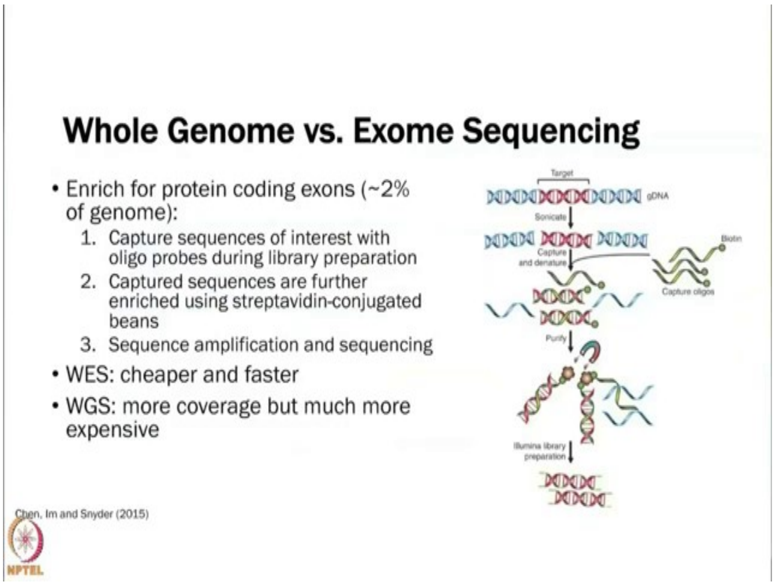
Transcriptional activity
Protein-DNA interactions
Methylation analysis

Transcriptomics

mRNA expression
Alternative Splicing
microRNA expression

So, I wanted to go through some examples next gen sequencing methods and how they are used. So, we will start with genomics, then we will move into transcriptomics and then we will end on epigenomics. So, for genomics the common two commonly used methods are whole exome and whole genome sequencing. So, whole genome sequencing just means that you are taking everything in the genome and you are sequencing all of it. Depending on the species you are working with, if you are working with humans, that is a lot that is an enormous amount to sequence.

(Refer Slide Time: 06:15)



Whole Genome vs. Exome Sequencing

- Enrich for protein coding exons (~2% of genome):
  1. Capture sequences of interest with oligo probes during library preparation
  2. Captured sequences are further enriched using streptavidin-conjugated beans
  3. Sequence amplification and sequencing
- WES: cheaper and faster
- WGS: more coverage but much more expensive

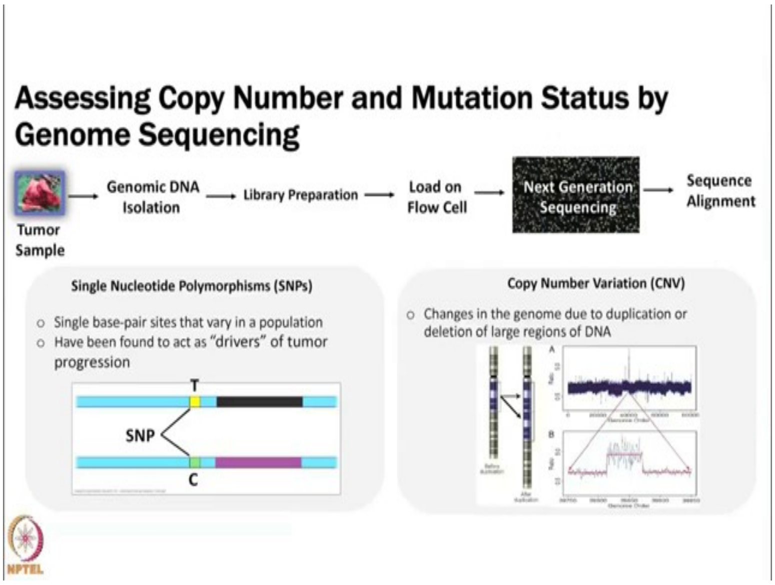
Chen, Im and Snyder (2015)

So, if you do not care about the things that are intergenic or intronic regions and you just care about protein coding genes or exons, then you just want to look at the exome. And so, what you can do is you can actually capture the exome sequences. So, you use these oligo probes that are match to exon sequences that are able to pull out and enrich for these an exon sequences during the library preps.

You kind of get rid of everything that is intergenic or so, what you can do is you can enrich for these exons before you do your while you are doing your library prep. And, then you do all of the sequence amplification and sequencing following that so, that you are only really looking at the exons. So, you get rid of everything else and this is 2 percent of the whole genome which really cuts back on my costs, quite a bit if you do not care about the other stuff.

So, it is another method to keep that the a lot of people use. So, it is cheaper and faster. With whole genome sequencing its more coverage, but it is much more expensive. So, depending on what you care about you decide which one you want to actually do ok.

(Refer Slide Time: 07:44)



Assessing Copy Number and Mutation Status by Genome Sequencing

Single Nucleotide Polymorphisms (SNPs)

- Single base-pair sites that vary in a population
- Have been found to act as "drivers" of tumor progression

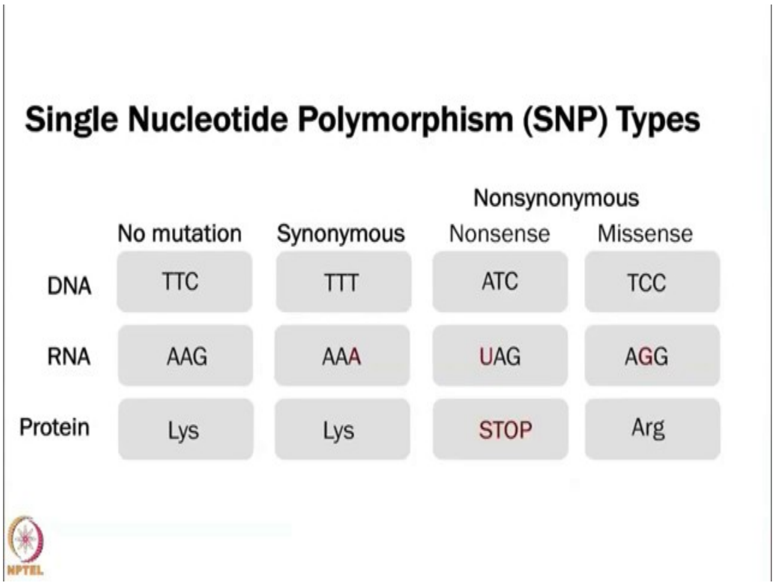
Copy Number Variation (CNV)

- Changes in the genome due to duplication or deletion of large regions of DNA

And then so, two of the main things that you can do with the genomics is to identify single nucleotide polymorphisms or SNPs, which are these single based parasites that vary. So, for examples of in your reference genome there is a T and in your sample there is a C, then you know that there was some sort of mutation that occurred and some of as I mentioned in the beginning some of these have been shown to be drivers of tumor progression. So, in cancer these are particularly interesting. And, you can also look at copy number variation which is just changes in the genome, because there is a large duplication or deletion of DNA.

So, instead of so, you can see here if this is along the chromosome you see that there is this big chunk that is been duplicated here, and then if you look at the copy number level you can actually see that there is double the number like approximately double the copies of this area of the genome that you see in your actual reads. So, you are able to actually get information on these duplications and deletions using these sequencing methods.

(Refer Slide Time: 08:53)



Single Nucleotide Polymorphism (SNP) Types

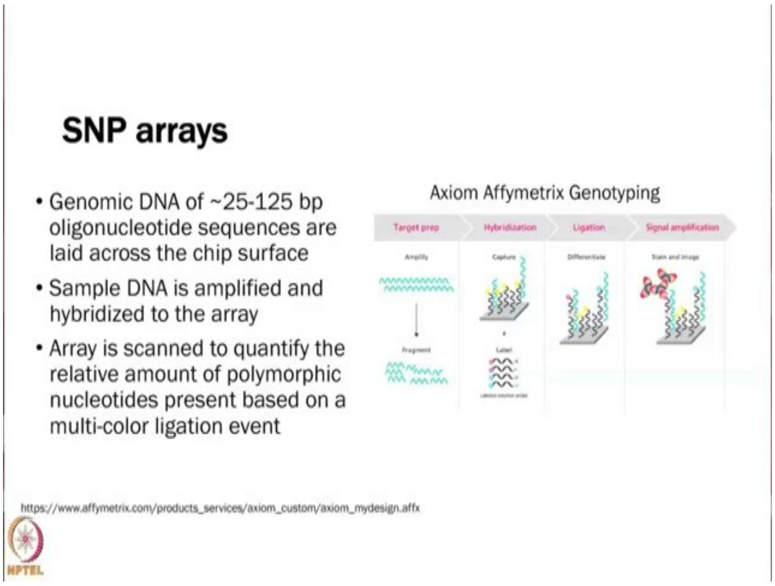
			Nonsynonymous	
	No mutation	Synonymous	Nonsense	Missense
DNA	TTC	TTT	ATC	TCC
RNA	AAG	AAA	UAG	AGG
Protein	Lys	Lys	STOP	Arg

There are couple of different kinds of SNPs. So, let us say this your DNA as I mentioned so, there was codons that encode for different protein amino acids. So, in this example there this is your reference genome. So, you have no mutation and then if this encodes for this RNA sequence, which then encodes for a lysine. You could have a synonymous mutation where this C is turned to a T, which causes the RNA to change to 3 A’s, in during translation. And, then our transcription sorry and then the protein though is still becomes a lysine, because there is some overlapping RNA in codons that encode for the same protein.

So, this does not actually cause the change at the protein level. But then you can have some nonsynonymous SNPs. So, for example, if you change the A the T to a A here, you get UAG at the RNA level which encodes for a STOP codon. So, now you have instead of your protein going on and continuing to grow, you actually have a truncated protein, or you could have a missense SNP where you have the middle T becomes the C at the RNA level is AGG and then you have an arginine. So, its changed the protein.

So, these were the mutations that people typically focus on, because they have an impact at the protein level. But when we do SNP calling you find you actually identify all of the mutations regardless of whether or not they are synonymous or non-synonymous. So, they are in addition to next gen sort of the standard next gen sequencing we were talking about there is also SNP arrays and these are still pretty commonly used. So, I did want to talk about them a little bit and these are just they are actual arrays that have specific SNP that they measure.

(Refer Slide Time: 10:44)



SNP arrays

- Genomic DNA of ~25-125 bp oligonucleotide sequences are laid across the chip surface
- Sample DNA is amplified and hybridized to the array
- Array is scanned to quantify the relative amount of polymorphic nucleotides present based on a multi-color ligation event

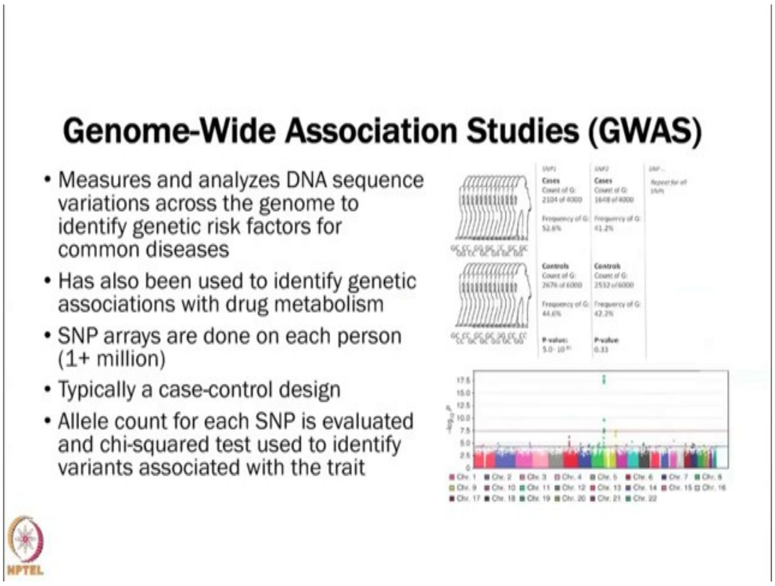
Axiom Affymetrix Genotyping

https://www.affymetrix.com/products_services/axiom_custom/axiom_mydesign.affx

So, you are only going to measure the SNP that are on the array. When you do whole genome sequencing, you can measure whatever, it is whatever is whatever you find you find. In a array you are actually asking do I find these SNPs and how can I measure them, what is the connotation of those SNPs are different populations. So, here you have used genomic DNA and you fragment it and then you lay across the chip surface.

So, this is there is a couple of different kinds of chips I just chose, one of the newer ones; and then the DNA is amplified and hybridized to whatever your array is which is here. And, then its scanned and you are able to quantify how much of which of the SNPs is occurring. In this case it is because of they have these four the fluorescent labels that specifically hybridize different SNPs. So, you are able to actually measure which SNP is present in which sample. There is a couple of different ways you can do this, but that is essentially sort of the overview of how this works.

(Refer Slide Time: 11:55)



Genome-Wide Association Studies (GWAS)

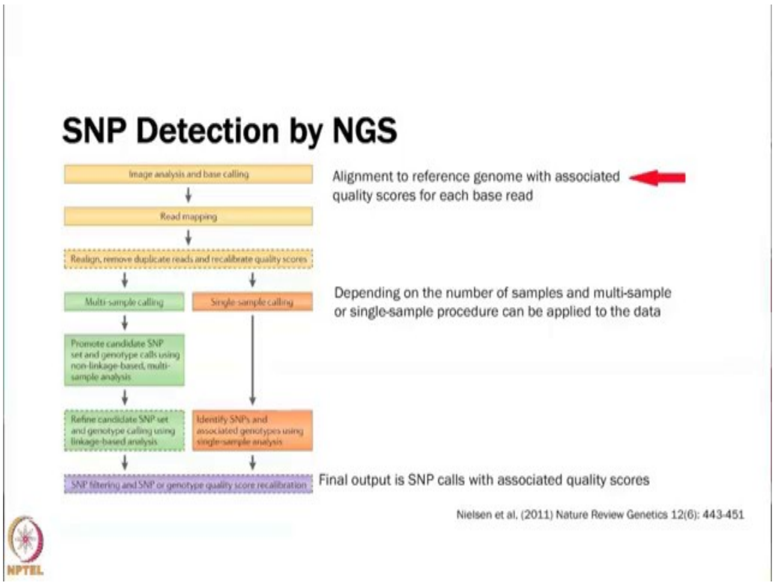
- Measures and analyzes DNA sequence variations across the genome to identify genetic risk factors for common diseases
- Has also been used to identify genetic associations with drug metabolism
- SNP arrays are done on each person (1+ million)
- Typically a case-control design
- Allele count for each SNP is evaluated and chi-squared test used to identify variants associated with the trait

And these are commonly used in these genome-wide association studies. So, these GWAS studies if anyone's done like 23 and me I do not know any of these like sequence your own SNPs, they are done with SNP arrays. So, where GWAS studies, they just measure and analyze these SNPs across different populations. So, they are typically trying to understand, it is a case control study.

So, if you have a population of people with disease x and a population of people without disease x, can you find a SNP that occurs more often and it is at a statistically significant level in disease x versus the control. So, and these were I think like they were super popular maybe 10 years ago. People still do them, but they definitely were huge deal for a little while and there are certain cases where they are still really useful. And, see you can just see here this is just looking across all the chromosomes and it is showing that at this points in chromosome I can tell which one is since based on the color, but that there is a significance association.

So, this is a log a negative log 10 p value with the disease versus the control, and you can also use SNP arrays if you are doing like a cancer study and you just want to look its specific SNPs in your population, its one way of doing. It is cheaper than doing the next gen sequencing ok. So, another way that you can do SNP detection is just using either the whole genome or whole exome sequencing that we talked about before.

(Refer Slide Time: 13:36)



SNP Detection by NGS

Image analysis and base calling

Read mapping

Realign, remove duplicate reads and recalibrate quality scores

Multi-sample calling

Single-sample calling

Promote candidate SNP set and genotype calls using non-linkage-based, multi-sample analysis

Refine candidate SNP set and genotype calling using linkage-based analysis

Identify SNPs and associated genotypes using single-sample analysis

SNP filtering and SNP or genotype quality score recalibration

Alignment to reference genome with associated quality scores for each base read

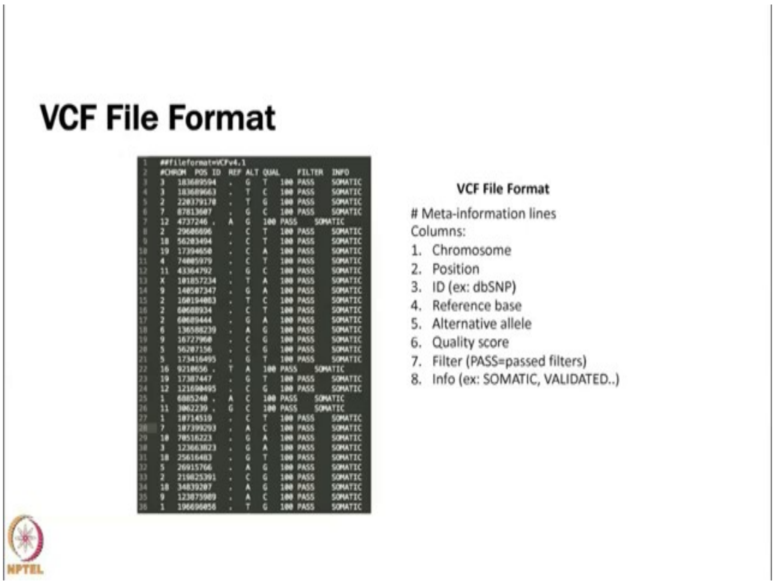
Depending on the number of samples and multi-sample or single-sample procedure can be applied to the data

Final output is SNP calls with associated quality scores

Nielsen et al. (2011) Nature Review Genetics 12(6): 443-451

So, you have your whole your sequencing data and you align it is the reference genome again as we discussed and then you have those quality scores, those fed scores for each of the different reads. So, you know how confident you are that the base that you are calling at that specific location is true or not. And, then you can remove some of these reads or you know you do a QC step and then depending on the number of samples, you can either do this multi sample calling or a single sample calling and then you there is many algorithms that you this, I will talk a little bit about which ones are available.

(Refer Slide Time: 14:28)



VCF File Format

```
##fileformat=VCFv4.1
#CHROM  POS ID  REF ALT QUAL     FILTER  INFO
3   183609594   .   G   T   100 PASS    SOMATIC
3   183609663   .   T   C   100 PASS    SOMATIC
2   220379170   .   T   G   100 PASS    SOMATIC
7   87813607    .   G   C   100 PASS    SOMATIC
12  4737246 .   A   G   100 PASS    SOMATIC
2   29606696    .   C   T   100 PASS    SOMATIC
10  56203494    .   C   T   100 PASS    SOMATIC
19  17394650    .   C   A   100 PASS    SOMATIC
4   74005979    .   C   T   100 PASS    SOMATIC
11  43364792    .   G   C   100 PASS    SOMATIC
X   101857234   .   T   A   100 PASS    SOMATIC
9   140507347   .   G   A   100 PASS    SOMATIC
2   160194083   .   T   C   100 PASS    SOMATIC
2   60680934    .   C   T   100 PASS    SOMATIC
2   60689444    .   G   A   100 PASS    SOMATIC
6   136588239   .   A   G   100 PASS    SOMATIC
9   16727960    .   C   G   100 PASS    SOMATIC
5   56207156    .   C   G   100 PASS    SOMATIC
5   173416495   .   G   T   100 PASS    SOMATIC
16  9210656 .   T   A   100 PASS    SOMATIC
19  17307447    .   G   T   100 PASS    SOMATIC
12  121690495   .   C   G   100 PASS    SOMATIC
1   6005240 .   A   C   100 PASS    SOMATIC
11  3062239 .   G   C   100 PASS    SOMATIC
1   10714519    .   C   T   100 PASS    SOMATIC
7   107390293   .   A   C   100 PASS    SOMATIC
10  70516223    .   G   A   100 PASS    SOMATIC
3   123663023   .   G   A   100 PASS    SOMATIC
10  25616483    .   G   T   100 PASS    SOMATIC
5   26915766    .   A   G   100 PASS    SOMATIC
2   219025391   .   C   G   100 PASS    SOMATIC
10  34839207    .   A   G   100 PASS    SOMATIC
9   123075909   .   A   C   100 PASS    SOMATIC
1   196690056   .   T   G   100 PASS    SOMATIC
```

VCF File Format

Meta-information lines

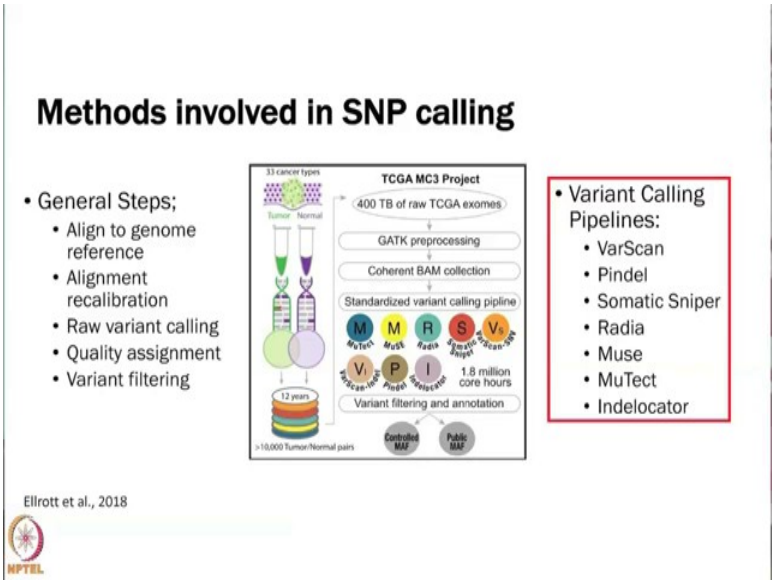
Columns:

1. Chromosome
2. Position
3. ID (ex: dbSNP)
4. Reference base
5. Alternative allele
6. Quality score
7. Filter (PASS=passed filters)
8. Info (ex: SOMATIC, VALIDATED..)

So, the algorithms will call different SNPs and then it outputs SNP calls which in VCF format which typically in VCF format which I will talk about what that format looks like. Here we go so, this is a VCF file and what it has is information on these SNPs and where they are located. So, you can see the first column is chromosome and the second column is the position of the SNP and the third column is an ID. So, in this case is that is just left blank. But, sometimes it has information like from a different database that exists like cosmic or dbSNP which we will talk about some of these have been annotated so, we will just put in whatever that SNP is.

It has the reference base so, what it is in like ref seq or whatever your sorry in the hg19 database and then it has whatever the real that is different in your sample is. And, then it has a quality score and you can have all sorts of columns that go on and on I talked about what it is. So, it depends on the data, but the 6 the first 6 are always there and they are the most important.

(Refer Slide Time: 15:38)



Methods involved in SNP calling

- General Steps;
  - Align to genome reference
  - Alignment recalibration
  - Raw variant calling
  - Quality assignment
  - Variant filtering

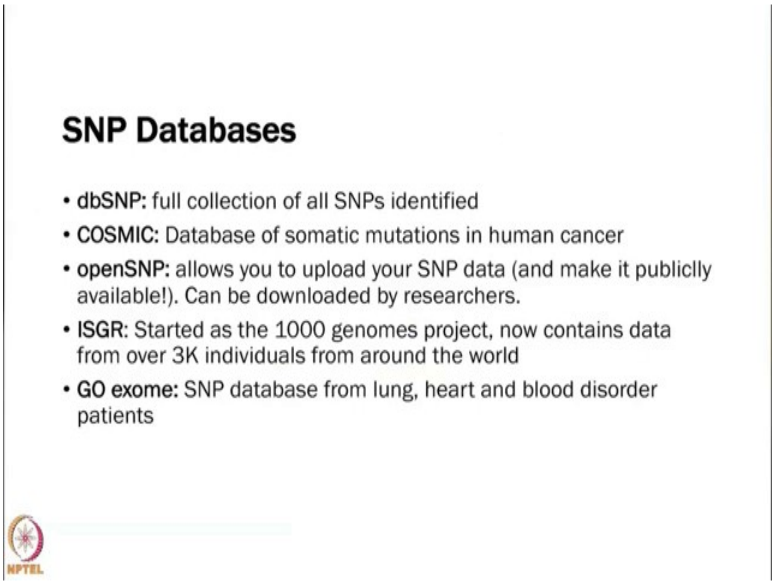
- Variant Calling Pipelines:
  - VarScan
  - Pindel
  - Somatic Sniper
  - Radia
  - Muse
  - MuTect
  - Indelocator

Ellrott et al., 2018

So, there was a paper there was a paper that came out 2018, I reviewed on some of these variant calling pipelines I have included it here. So, again the general steps for all of these algorithms or to align it to the genome reference do this recalibration in QC step, then do the variant calling and look at the quality of those and then filter out the variants based on the quality that they come out in your variant caller.

So, there is a whole bunch of pipelines, everyone has a favorite it is usually the one that they created or that they know the person who created it, that is how these things work right. But, a lot of people what they do is they use several of them and then they look for overlaps and that seems to be the best way of doing this, because you know that if many of them all of them have different strengths, limitations and then you know if it was called by several then the overlap is probably the best way to go.

(Refer Slide Time: 16:30)



SNP Databases

- **dbSNP:** full collection of all SNPs identified
- **COSMIC:** Database of somatic mutations in human cancer
- **openSNP:** allows you to upload your SNP data (and make it publiclly available!). Can be downloaded by researchers.
- **ISGR**: Started as the 1000 genomes project, now contains data from over 3K individuals from around the world
- **GO exome:** SNP database from lung, heart and blood disorder patients

And then there are several SNP databases that are really useful, if you are working with these SNP. So, for example, there is dbSNP which is just the collection of every SNP essentially that is been identified. Yeah, explained variant calling.

Student: Yeah.

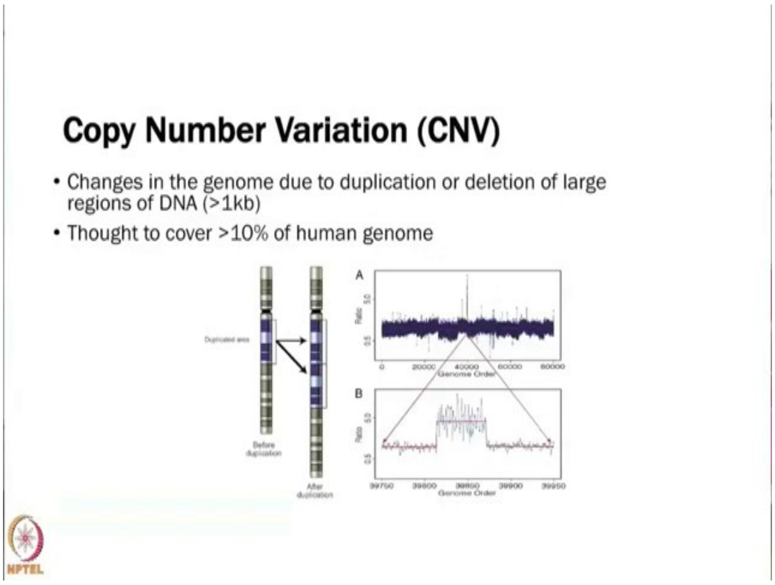
That you said yeah so, what it is this is essentially it is taking your chunk of sequence and it is comparing it to the reference and then its saying this nucleotides always comes up different; in a whole bunch of reads, this one nucleotide. And, it will pull out the fact that that nucleotide is different and a whole bunch of reads and then it gives it a quality score. So, they will output that information ok.

So, there is a whole bunch of SNP databases; dbSNP which I mentioned is just the collection of all a bunch of SNP that have been identified. COSMIC, which is specific for somatic mutations in cancer, so, if we are working with cancer this is a really a good one to look at; openSNP where you can actually upload your SNP data, I can believe people do this, people do this they get their SNP data from these companies and then they upload it. So, that other people can use it, they are very trusting, this ISGR which was started as the 1000 genomes projects, is anyone heard about the 1000 genomes projects?

Student: (Refer Time: 17:54).

Great. It is pretty interesting, and I think it is they keep adding more and more data. So, it is really useful if you are looking at SNPs in different populations. They are just trying to get DNA from people all over the world's to try and see what SNPs occur in different populations. And, then there is GO exome which is a SNP database from a long heart and blood disorder projects ok. And, then copy number variation, as I mentioned is just looking at these changes in the genome due to these duplication or deletions of large regions of DNA and this is also really often occurs in cancer. So, it is something that we pay a lot of attention to and use whole genome or whole exome sequencing to get information on.

(Refer Slide Time: 18:23)



Copy Number Variation (CNV)

- Changes in the genome due to duplication or deletion of large regions of DNA (>1kb)
- Thought to cover >10% of human genome

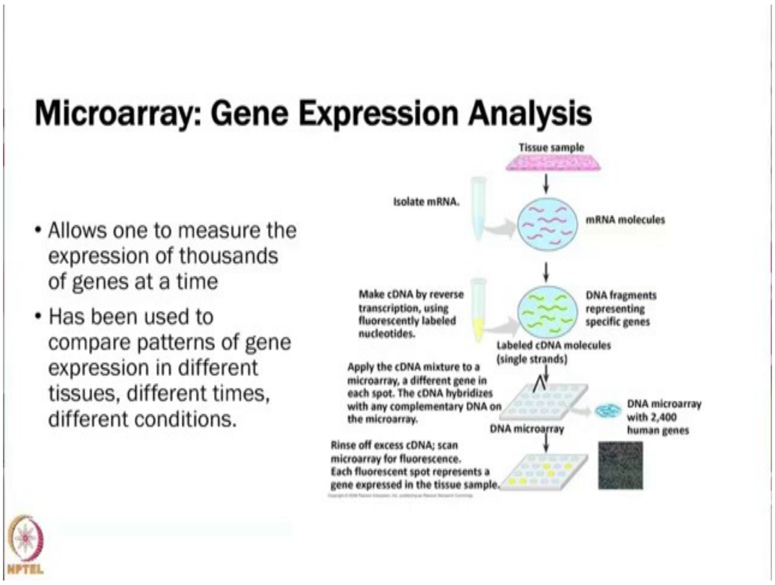
Student: We were doing you know that CNV, we were looking for CNV, but after that we changed and we are not looking CNV.

So, you think that CNV is like less popular now.

Student: Yeah, I am asking you would like to.

I mean I all of the projects I have worked on, we still do it and we still do included in our data analysis. I think also if you are doing you know if you are doing the whole genome or the whole exome sequencing already to get your SNPs, then why not do copy number, but so I think it is also like if you are already have the data you are going to do it ok. So, I am going to move on to transcriptomics.

(Refer Slide Time: 19:28)



Microarray: Gene Expression Analysis

- Allows one to measure the expression of thousands of genes at a time
- Has been used to compare patterns of gene expression in different tissues, different times, different conditions.

Isolate mRNA.

Make cDNA by reverse transcription, using fluorescently labeled nucleotides.

Apply the cDNA mixture to a microarray, a different gene in each spot. The cDNA hybridizes with any complementary DNA on the microarray.

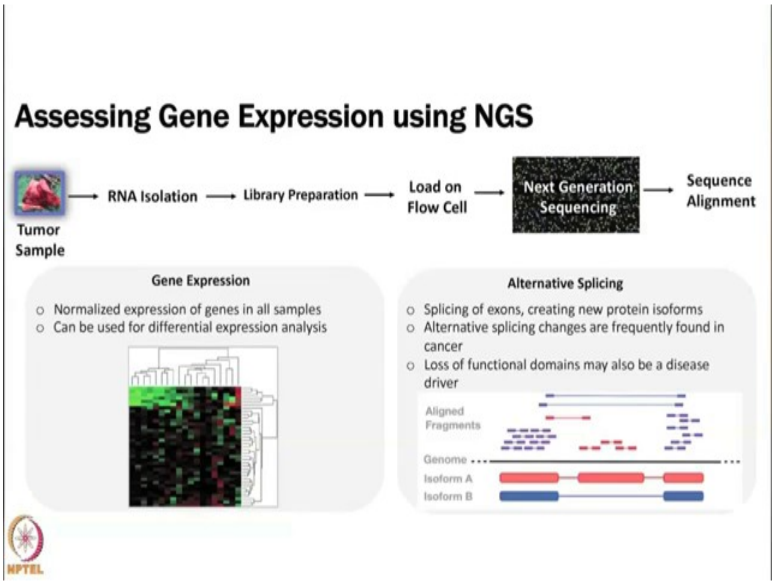
Rinse off excess cDNA; scan microarray for fluorescence. Each fluorescent spot represents a gene expressed in the tissue sample.

Tissue sample
mRNA molecules
DNA fragments representing specific genes
Labeled cDNA molecules (single strands)
DNA microarray
DNA microarray with 2,400 human genes
NPTEL

So, the old way of doing high throughput transcript genomics was using a micro arrays. So, these are gene expression arrays, where similar to the snip arrays where you have the set number of genes that or now in this case genes that you want to measure; and you can then you have this chip where you make this cDNA of your so, you take your RNA and you make it cDNA using reverse transcriptase and then so, you are just taking your single strand and you are making it a double strand.

And, then you fragment it and you put it on you have you label it with this fluorescence and then you add it to this microarray and it hybridizes. So, you have probes for specific genes or transcripts in your microarray that you then measure. So, you just if you have a lot of a certain transcript there will be a lot of things that sticks to that probe and then you can measure based on how much fluorescence there is in each of the cells, how much transcription you have in that gene. So, this is the sort of the old way I think there are some people who are still using this, but for the most part people have moved right into RNA seq. So, I am going to spend some time on RNA seq.

(Refer Slide Time: 20:36)



Assessing Gene Expression using NGS

Tumor Sample → RNA Isolation → Library Preparation → Load on Flow Cell → Next Generation Sequencing → Sequence Alignment

Gene Expression

- Normalized expression of genes in all samples
- Can be used for differential expression analysis

Alternative Splicing

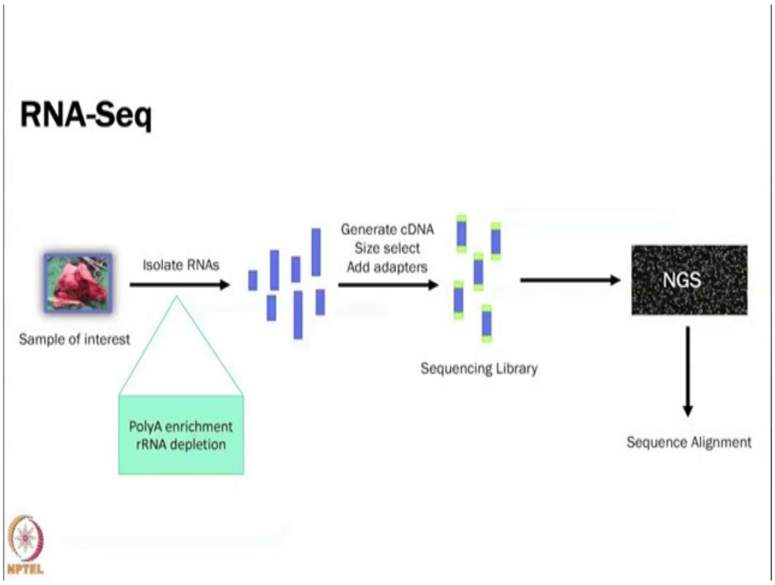
- Splicing of exons, creating new protein isoforms
- Alternative splicing changes are frequently found in cancer
- Loss of functional domains may also be a disease driver

So, in RNA seq it is similar to the whole genome and whole exome sequencing where you have your sample, you isolate your RNA and you do a library prep like we discussed before. You load on your flow cell and you do this next gen sequencing. In this case the only difference is that you are measuring RNA instead of DNA, and what this can be used for is gene expressions. So, you can look for the expression of genes in transcripts in all of your samples, you can do differential expression analysis and you can also look at alternative splicing.

So, with RNA right you are going to get anything the you are going to get splicing of different exons that you are going to be able to see, because at the genome level the exons are separated by entrance. And, then once they are transcribed you can get them in you can see what they look like at the once they have alternative splicing has occurred.

So, that is so, that is a benefit to RNA seq that you get more information than you would get from your genome sequencing. Some people also to do SNP calling from RNA seq; from I have talked to a lot of people about this and it is there is a higher error rate. So, there is some worries about using RNA seq to do SNP calling, but it is something that people to do.

(Refer Slide Time: 22:01)



RNA-Seq

Sample of interest

Isolate RNAs

PolyA enrichment
rRNA depletion

Generate cDNA
Size select
Add adapters

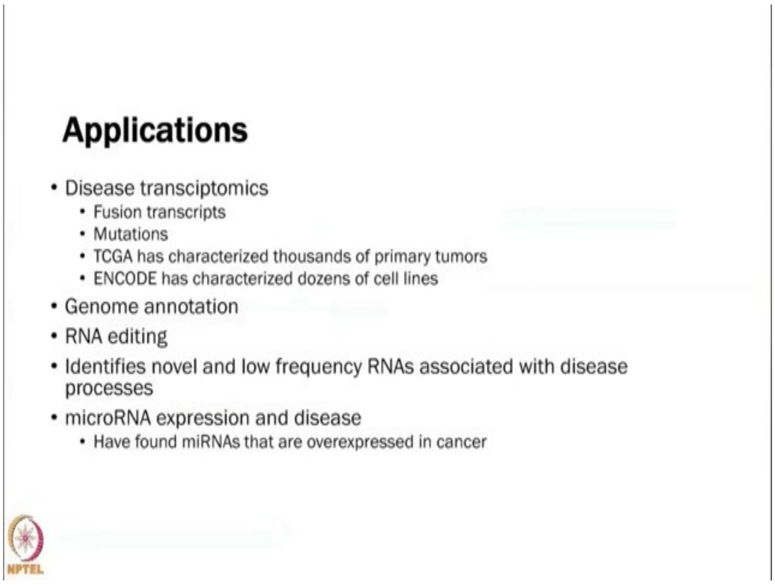
Sequencing Library

NGS

Sequence Alignment

So, how does this works? So, you actually have to enrich your RNA’s when you do this, so you have your sample of interest you isolate your RNA as since either using a polyA enrichment. So, you pull them out of your sample or you can deplete ribosomal RNA, those were the two different methods you can use. So, you get this enrichment of mRNA’s; then you select for specific sizes from this, from the RNA’s that you have enriched for and you add adapters; similar to what we did I showed previously with genomic sequencing and then you just do the next gen sequencing as we sort of discussed. And, this is again a lot of times it is done using alumina or similar instruments.

(Refer Slide Time: 22:46)



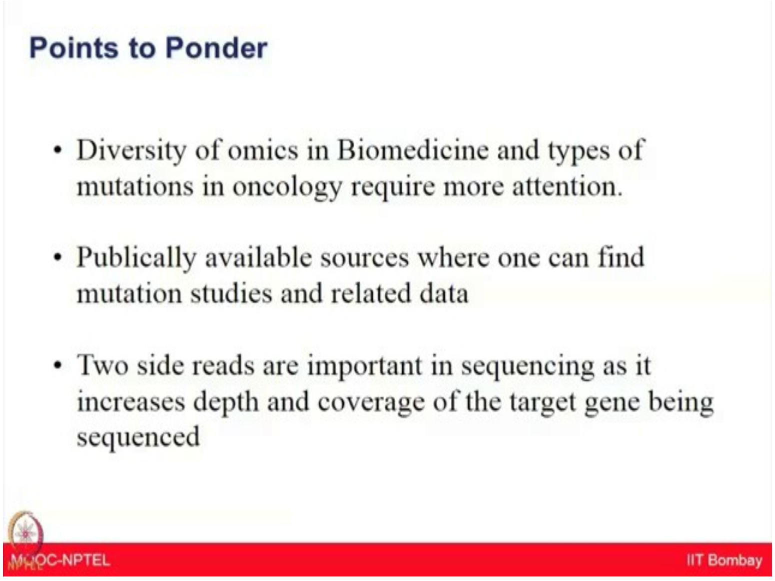
Applications

- Disease transciptomics
  - Fusion transcripts
  - Mutations
  - TCGA has characterized thousands of primary tumors
  - ENCODE has characterized dozens of cell lines
- Genome annotation
- RNA editing
- Identifies novel and low frequency RNAs associated with disease processes
- microRNA expression and disease
  - Have found miRNAs that are overexpressed in cancer

And there is a lot of applications for RNA seq, I think we have if you are in the fields at all you have read a lot of papers where people use RNA seq. It is very popular right now. You can look for fusion transcripts which we will talk a little bit about mutations. The TCGA which I will talk about more later has used RNA seq to characterize thousands of tumors. ENCODE which I will also talk about has also characterized dozens of cell lines.

You can look at annotation of genomes; so how the genomes are actually structured and then you can identify RNA's that are associated with disease. You can also look at micro RNA's that takes a totally different process of sample prep, but it is the same once you kind of isolate that micro RNA's you can sequence them in and look similarly at how they are expressed in different diseases.

(Refer Slide Time: 23:41)



Points to Ponder

- Diversity of omics in Biomedicine and types of mutations in oncology require more attention.
- Publically available sources where one can find mutation studies and related data
- Two side reads are important in sequencing as it increases depth and coverage of the target gene being sequenced

MOOC-NPTEL IIT Bombay

So, today's lecture you have learned how sequence alignment could be done and factors which help in increasing the efficiency of your analysis for the big datasets obtained from genome data sets. You also learned about GWAS which contains experimental data related to SNP's in various genes leading to different clinical conditions. Dr. Kelly has also helped us to understand how to use the raw files of NGS in SNP data analysis and I hope you have also learned about various SNP databases like COSMIC which contains somatic mutations in human cancer, GO exome which is SNP database for lung, heart and blood disorders and many more.

So, I hope you know by understanding, by listening these lectures not only you are getting refreshed about the genomic and transcriptomic and basic of these technologies, but also some of the databases and resources which are available, from where you can obtain lot of new information from the publicly available datasets. In next lecture Dr. Kelly will talk to you about how one could use RNA sequencing for transcriptomic studies and interpretation of data for much more meaningful insights of a given disease.

Thank you.