

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

Lecture - 05
Introduction to Genomics - III
Transcriptome

Welcome to MOOC course on Introduction to Proteogenomics. In the last two lectures Dr. Kelly Ruggles have talked to you about some of the advancements of genomic technologies. In today's lecture Dr. Kelly is going to talk to you again about Transcriptome studies, especially how to utilize RNA sequencing with reference to the various type of data basis available and what are the challenges of using these technologies. She also talks about extended RNA sequencing workflow, which includes alignment of genome, count coverage per gene or transcript and differential expression study.

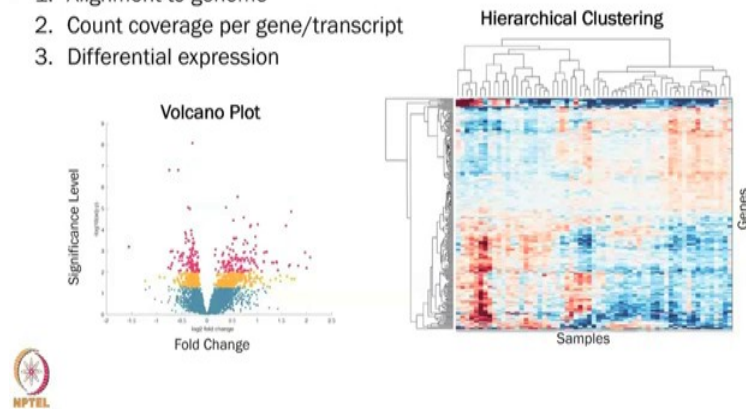
Today's lecture we will also talk about various software available for gene isoform quantification and for differential expression analysis. Dr. Kelly is going to talk about the effect of gene fusion where genes from different chromosomes come together, because of chromosomal translocation, inversion or interstitial deletion. The concepts of UCSC genome browser will be elaborated with all the data for each human gene is studied and mounted till now. Dr. Kelly is going to talk about the advancement in the field of transcriptomics where now one could also perform RNA sequencing using single cell. So, let us welcome Dr. Kelly Ruggles to give today's lecture.

So, standard RNA-seq workflow you again, you do your next gene sequencing, you have to align it to the genome, the most commonly used the liner for RNA-seq at this point a star. So, I would very much recommend you use that aligner if you are doing RNA-seq analysis.

(Refer Slide Time: 02:06)

Standard RNA-seq Workflow

1. Alignment to genome
2. Count coverage per gene/transcript
3. Differential expression



Then you use coverage, you count coverage per gene or transcript and then you do differential expression. So, some outputs to this are things like volcano plot. So, if you are familiar with those, so you are looking at, let us say you have disease and not unhealthy, you can look at the fold change of different transcripts versus the significance level of those transcripts and you are looking for things that are sort of either here or here in your data, you can do hierarchical clustering of your data. So, you can see if your disease versus healthy or different subtypes cluster together based on the expression of different genes.

(Refer Slide Time: 02:44)

RNA-seq Alignment Challenges

- Using RNA-seq for gene expression requires counting sequence reads per gene
- Must map reads to genes – but this is a difficult problem
 - Introns create gaps in alignment of mRNA to genome
 - What to do with reads that map to introns or completely outside exon boundaries?
 - What about overlapping genes?
- Use an aligner that supports splices (gaps) such as STAR



But there is some challenges to RNA-seq alignment, this includes the facts that you, there are introns, so with whole genome sequencing right you have chunks of DNA that you can just map back to your reference genome. Here, because you are looking at RNA, you have exons that have been spliced together. So, you have places where there will not be, there will be a gap and you have to account for that gap in your alignment.

And, so the aligners that work with RNA-seq have to be a little more sophisticated, because they have to deal with the fact that there are these gaps and they have to figure out where they are and where the boundaries are and then record the junctions of the of these boundaries. So, that is something that you have to keep in mind with these aligners.

And, then if the reads are in introns or in intergenic regions what does that mean, if like we do not expects that to be in RNA, is it real. And, I did want to mention some of the different ways that people do these counting the reads per gene, because there is a lot of different ways to do it.

(Refer Slide Time: 03:53)

Counting Reads per Gene

- Need a gene model for your reference genome (exons, introns)
- Various databases available – with differing levels of complexity
 - RefSeq
 - ENSEMBL
 - UCSC
- Expression Units: Normalizing reads for sequencing depth and gene length
 - RPKM: Reads Per Kilobase Million
 - Scaling factor = Total reads in sample/1,000,000
 - Normalized Reads = Read counts/Scaling Factor
 - RPKM = Normalized Reads/length of gene

Normalize Sequencing Depth First
 - FPKM: Fragments Per Kilobase Million
 - Same as RPKM but factors in paired reads
 - TPM: Transcripts Per Kilobase Million
 - Reads per kilobase (RPK) = Read counts/length of gene (kilobases)
 - Scaling factor = sum of RPKs in sample/1,000,000
 - TPM = RPK/scaling factor

Normalize Length First

Conesa et al., (2016) Genome Biology 17:13


So, there are, you need a gene model. So, in meaning you need a database that says these are where your exons are, this is where your start sites are. So, we have these gene models, we have these databases, some there are lots of databases available with different levels of complexity

So, for example, RefSeq or ENSEMBL these are all databases that exist that have files that say this is what we expect to see at the transcript level and you can kind of use those to measure how much, how many counts, how many reads you have of different transcripts or genes based on what we know about how those genes are structured in the genome. And there is lots of ways of methods of actually reporting, how much expression there is at the transcript or gene level. So, there is RPKM which is Reads Per Kilobase Million, there is FPKM which is Fragments Per Kilobase Millions. So, these are similar; the FPKM is typically used for these paired end reads versus the RPKM back to expression units.

So, as I mentioned there is, there are several ways that people will express the reads. So, you have to normalize the reads, you have reads, but certain genes are long right. So, if a gene is long you are going to have more reads that map to it, because it is just longer, that does not mean there is more of that gene that just means it's long right. So, we have to take that into account. We also have to take into account that maybe a certain sample just was had more reads in it, but that does not mean all of the genes in that sample are up that just means there is more reads in that we ended up getting in that sample.

So, these are two things that we have to normalize for. So, there is really two different ways of doing this; the first one normalizes by the depth. So, how many reads in the sample first and then normalizes by the length of the gene, so that is this RPKM and FPKM. The other one normalizes by the length of the gene first and then by the number of reads in the sample second and that is this TPM. I do not know which one is better, I do not know if anyone knows which one is better. There is a review here on the differences between both and they have again strengths and limitations.

So, whatever your problem is just spend some time thinking about this and know that there are different ways of doing this and they have different effects on your downstream analysis ok.

(Refer Slide Time: 06:28)

RNA-seq Software

- Gene/Isoform Quantification:
 - RSEM
 - Kallisto
 - Cufflinks
 - Salmon
- Differential Expression Analysis
 - EdgeR (R)
 - overdispersed Poisson distribution, Empirical Bayes method, Fisher's exact test
 - DESeq2 (R)
 - 75th percentile normalization, Negative binomial distribution, quantile distribution of variance (similar variance for genes expressed at similar levels)

Papers benchmarking these pipelines:

Teng et al., A benchmark for RNA-Seq quantification pipelines. *Genome Biology* 2016 17:74

Evaluation and comparison of computational tools for RNA-Seq isoform quantification. 2017 18:583

Schurch et al., How many biological replicates... and which differential expression tool should you use? *RNA* 2016 22:839

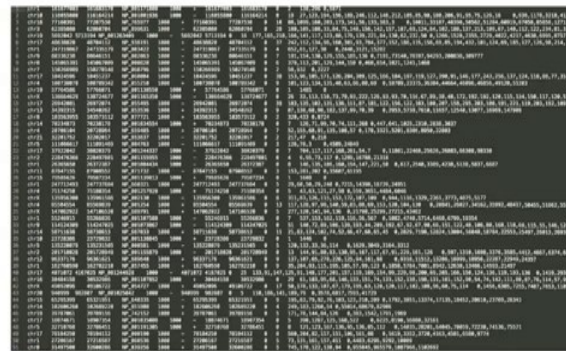


So, in terms of RNA-seq software that I just wanted to point out there is a lot of different ways of doing gene or isoform quantification. So, there is several I have listed here, there is one of these papers this one goes into lots and lots of details about them and more. So, you can look into this, if you are interested in learning more about how to quantify at the gene or isoform level. There is a several differential expression analysis packages DESeq 2 is a really commonly used one, there is also a EdgeR, this last paper actually compares the two and talks about one you should use one versus the other.

So, I have leaving these papers here for you guys in case this is something that you are going to do and you want to learn more about, but if you have specific questions about this just you can come find me and we could talk about it.

(Refer Slide Time: 07:24)

BED File Format



BED File Format

Columns:

- 1. Chromosome
- 2. Chromosome Start
- 3. Chromosome End
- 4. Name
- 5. Score
- 6. Strand (+or-)
- 7-9. Display info
- 10. # blocks (exons)
- 11. Size of blocks
- 12. Start of blocks



So, one of the things that you get from RNA-seq, if you have, I think for a lot of these packages you have to specifically ask for this. So, if you want this you should make sure you are near setting. So, you are asking for it, but you can get junction files which we are going to talk about that are in BED file format, also a lot of these like RefSeq in these annotation databases use the BED file to say this is where an exon is, this is where an intron is. So, these are the files that sort of tell you the structure of the actual genome, and so this is just an example BED file here I have included what the columns mean.

So, the first column is the chromosome, the second is the start and end of that of that. Sorry that is actually should be gene I can correct that and send it out again. So, it is gene start and gene ends a name, a score, a strand, and then there is this display info, because there is these browsers we will talk about, a little bit where you can change colors and you can have the display if you want to put it up on a specific browser you can have it look a certain way. So, there is columns for that. The number of exons or blocks, the size of the blocks or exons and the start of the blocks, and I am going to go through an example about what this looks like. So, for example, here we have one row from a BED file.

(Refer Slide Time: 08:50)

BED File

Chr	Start	End	Name	Score	str	Display info	# blocks	Block size	Block start
chr5	11106617	111091469	NP_004763	1000	-	11106617 111091469	0 3	126, 78, 3	0, 4509, 24849



So, we have chromosome 5, we know that this gene starts at this coordinate.

(Refer Slide Time: 08:59)

BED File

Chr	Start	End	Name	Score	str	Display info	# blocks	Block size	Block start
chr5	11106617	111091469	NP_004763	1000	-	11106617 111091469	0 3	126, 78, 3	0, 4509, 24849

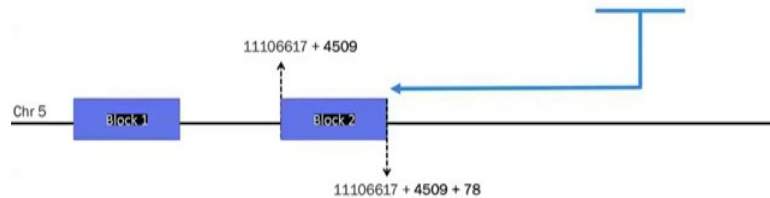


And, then we have block size and block starts, so these are the exons. So, we know that the exon is 126 base pairs is long and this is where it ends. So, it is the start is, the block start is this plus 0 and then the block end is this start of the gene plus the block size plus 126.

(Refer Slide Time: 09:22)

BED File

Chr	Start	End	Name	Score	str	Display info	# blocks	Block size	Block start
chr5	11106617	111091469	NP_004763	1000	-	11106617 111091469	0 3	126, 78, 3	0, 4509, 24849

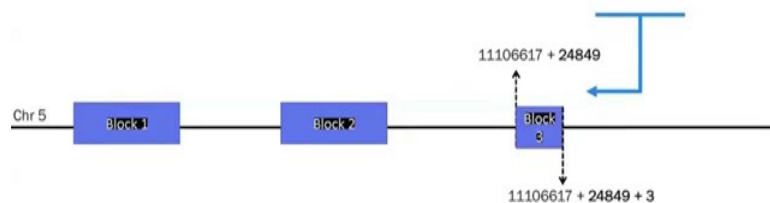


And then you have the second block or exons. So, you have again you start from the start of the gene and you add the block start number two which is 4509. So, this is where the second exon starts and you know that the second exon is 78 base pairs long. So, now you know where your block 2 ends.

(Refer Slide Time: 09:44)

BED File

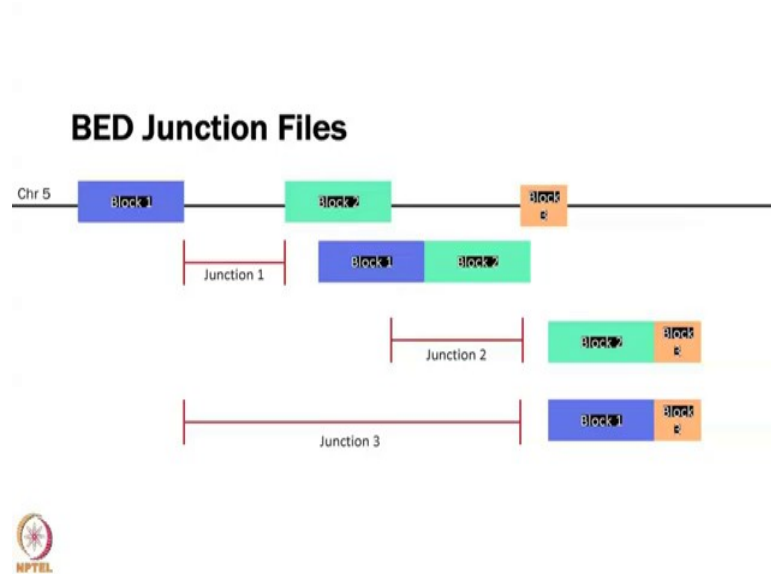
Chr	Start	End	Name	Score	str	Display info	# blocks	Block size	Block start
chr5	11106617	111091469	NP_004763	1000	-	11106617 111091469	0 3	126, 78, 3	0, 4509, 24849



And, then for block 3 you have again the start of the gene plus this block start 24849, you know that this starts here and it is only 3 long and ends there. So, then you know exactly where your exons are, based on what the BED file tells you about, about the

coordinates of these exons. So, what these junction files from the RNA-seq data will give you is not just reads that cover different exons, but also where the exons connect right.

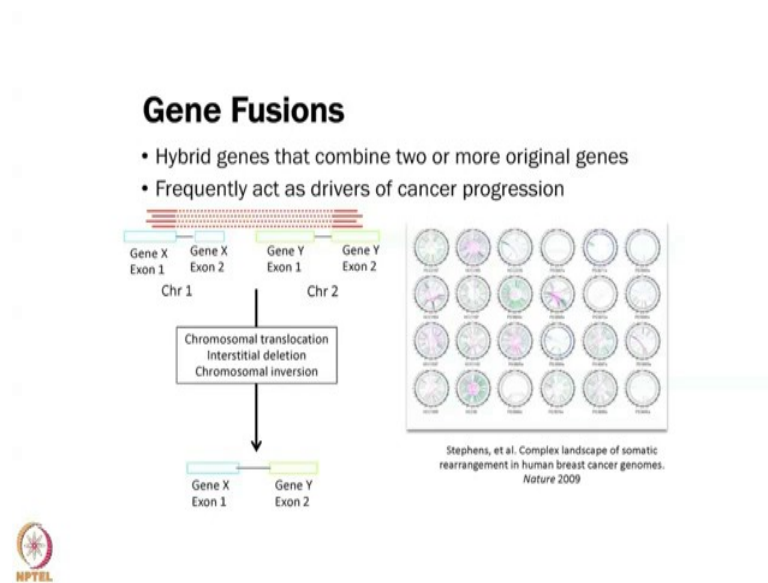
(Refer Slide Time: 10:11)



So, if you have this alternative splice, the splicing that is occurring where you have exon 1 spliced exon 2, there will be a read where there is a, it will show that these two are connected which is called a junction read, and it will be in this bed format. And so here like if you have exon 2 and exon 3 you would have another junction; junction 2 that would connect these so you would end up getting this or if you had a junction connecting exon 1 with exon 3 you get something that looks like this.

So, these junction files are something that comes out of RNA-seq and in addition to the expression analysis data and the last thing that can come out of this RNA-seq data is gene fusions.

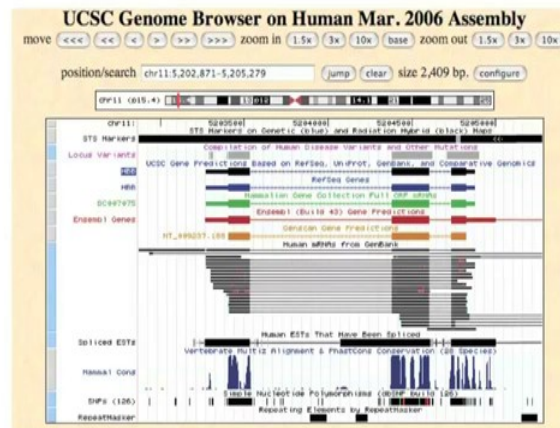
(Refer Slide Time: 10:58)



So, gene fusions are when a gene from one gene that is, can be from a different chromosome or from far away on the same chromosome, it is actually fused with another gene. So, here we have gene X on chromosome 1 and gene Y on chromosome 2, and you can see here that these two are connected, because of chromosomal rearrangements, which typically can occur and in cancer. So, this is a pretty cancer specific analysis. This is just a schematic showing each of these is a different breast tumor and each of these lines connects them based on how the genome has been rearranged.

So, you can see some of them have a lot of rearrangements, some of them do not. So, you will get fusion genes in certain samples and not others, but it is another thing to keep in mind when you have this RNA-seq data, this is another thing that you can look at as well.

(Refer Slide Time: 12:02)



And so, there are two browsers I am going to talk about that you can actually look at, you can take your data and upload it or you can look at the gene annotation for a specific gene, and so there is UCSC genome browser has anyone use this, it is pretty common and useful. So, if this is something you think you are going to be doing, I would spend some time exploring it. So, here you can see, you can look at a specific gene or part of the chromosome.

So, it has these really exons here and introns, you can have, there are so many tracks you could put hundreds of tracks. There is so much data on here you could spend days looking at it, exploring maybe your favorite gene, seeing what is available, they have lots of different publicly available datasets that have already been mounted. So, you can just click on them and see is this gene you know they have epigenetics data that is up there from encode, there is all sorts of things that you can look at.

(Refer Slide Time: 12:58)

Integrative Genomics Viewer (IGV)



<http://software.broadinstitute.org/software/igv/>



And in addition to that there is this integrative genomics viewer, where this is something it is a IGV from the broad that you actually download to your desktop and then you can upload your own data and also look they have different genomes and annotations there are already available in your, there are like mounted to it. So, for example, the human hg19 is up there and you can just already look at the annotation within that. And, so this is just I have uploaded a bunch of different dating.

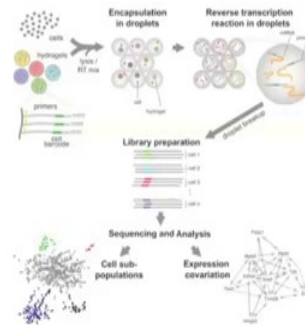
And then I did want to touch on single cell RNA-seq this is the hot field right now, people are really excited about it. And, so with this you can actually measure the RNA expression and a single cell versus what we normally do which is we just take a chunk of something. And, we look at the expression across many different cells, but you know there is heterogeneity especially in cancer.

So, you may be measuring normal tissue, you may be measuring a certain clone of a cancer one cancer you know you are, so you are kind of diluting out your results, but with single cell RNA-seq you do not get the same coverage right. So, you only get about a 1000 genes that you are measuring versus with RNA-seq where you are getting almost all, if not all of the genes, but it is still very cool; what you do with this there is a couple of ways of doing this.

(Refer Slide Time: 14:15)

Droplet Barcoding in Single Cell RNA-Seq

- Encapsulates cells into droplets with lysis buffer, reverse transcription (RT) reagents, and barcoded oligonucleotide primers
- mRNA released from each lysed cell remains trapped in the same droplet and is barcoded during synthesis of cDNA.
- After barcoding, material from all cells is combined by breaking the droplets, and the cDNA library is sequenced using established methods



So, this I am just going to talk about one way which is droplet bar coding. So, you have a cell and you encapsulate it into a droplet ok. So, you have all your cells and you put each one in a droplet, and within that droplet there is lysis buffer and there is everything you need essentially to do your library prep within the bubble, within the droplet. So, you lyse the cell, you release your RNA and you then can barcode it and do a cDNA synthesis and get it all ready to do sequencing essentially and you have bar coded it.

So, you know this is the one cell, everything in this droplet comes from one cell. And, then you just break the droplet and you do like the same kind of multiplexing you would do with lots of samples you just do them with lots of cells. So, then you just measure all of your RNA and then later after you have done your RNA-seq you can pull out each of the different cells based on the barcode.

So, that is currently have one of the ways that people are doing single cell RNA-seq and getting cell based data. The one thing that I have not seen, I know that some people are doing it, but I think it is a lot harder it is doing SNP calls from single cell data because it is just there is not enough coverage. So, that is something that is not currently happening, but I am sure eventually we will figure out how to do it and somebody will do it and it will be exciting yeah.

Student: What are these adaptors we are adding?

Say that again sorry.

Student: Bar coding.

It is similar to bar coding like the multiplex samples right, it is like adapter bar coding.

Student: PCR also involves lot of errors, how do you account for that?

Yeah.

Student: So, the PCR also.

Yeah.

Student: Involved a lot of error so.

Yes of course.

Student: So, how do you account for that?

I mean there is always error that that is we have to deal with in experiments, I mean I think we; so, how do we account for the PCR amplification error. So, what happens is especially with, so when you make those clusters in your library seq, when you are making your libraries right, you are making many many different copies. So, if only one of them has a certain SNP you are going to say that that was because of a PCR error.

So, it is usually just based on the, the genome aligner is kind of have all of that built in, I do not know the answer to that. The question was what if the errors is early enough that it is in every copy, I do not know that is a good question.

Yeah.

Student: We are having or some particular length of the gene whichever taking in the fragments and when we were bar coding that, but in this case.

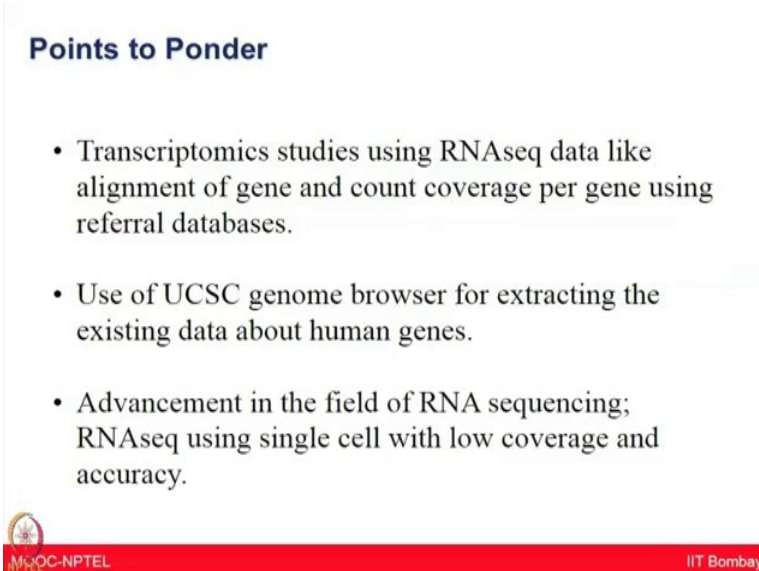
Yeah. So, the question, that is a good question. So, the question is with alumina there was a size filter essentially right, like you knew that your fragments were a certain size. I think actually they do I did not put that here, but I do think that is incorporated into this. So, you have kind of the similar size that you would have in like your bulk RNA seq, but

you are just have, it is just within every things kind of included in the droplet that you would do at the bulk level. So, you are, it is a very similar sequencing process, it is just there is less, there is less RNA. So, your coverage is lower.

Student: Coverage is lower.


Coverage is much lower with single cell RNA-seq than it is with a whole bulk, because with bulk you have lots of copies and you can you know with single cell you only have a certain number of copies, so you only measure up to like a 1000 genes versus 20,000 genes with bulk.

(Refer Slide Time: 18:04)



Points to Ponder

- Transcriptomics studies using RNAseq data like alignment of gene and count coverage per gene using referral databases.
- Use of UCSC genome browser for extracting the existing data about human genes.
- Advancement in the field of RNA sequencing; RNAseq using single cell with low coverage and accuracy.

 NPTEL IIT Bombay

So, in conclusion you have seen that how studying transcriptome can be very useful to provide the first level of functional information obtained from the genes. If you think about the central dogma from the DNA, the RNA being formed in the process of transcription and then from RNA the proteins are being formed in the process of translation; so, the first set of functional information comes in the form of transcripts. In today's lecture we have seen how introns become problem in the RNA sequencing data alignment as to the reference genome sequences, you also learnt about read and understand the BED file in sequence alignment for data analysis and representation.

I hope you got the concepts of RNA sequencing and how the droplet bar coding can be done for single cell RNA sequencing along with the pros and cons of this technique. In

the next lecture Dr. Ruggles will continue discussion about genomic and epigenomic technologies with more focus about epigenomic analysis. So, let us continue this discussion about genomic revolutions in the next lecture.

Thank you.