

**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Kelly Ruggles**  
**Department of Biosciences and Bioengineering**  
**Indian Institute of Technology, Bombay**

**Lecture – 44**  
**Variant Analysis**

Welcome to MOOC course on Introduction to Proteogenomics. After understanding the sequence centric Proteo-genomics, we will now listen Doctor Kelly Ruggles talk about Variant Analysis and its effect on other biomolecules leading to various clinical conditions. She will also talk about how to read dot VCF file, and also about informatics tools for creating the customized databases. She will discuss about how one could create variant peptide manually by using integrative genomics viewer or IGV to understand the basics.

However, once they have understood the basic mechanism, one can use different softwares to develop variant peptides. Let us now welcome Doctor Kelly Ruggles again, to tell us more in depth about creating and analyzing variant peptides.

I am going to walk you through doing this by hand, even though we now have tools to do this; but I think it is good to do these things by hand sometimes.

(Refer Slide Time: 01:33)

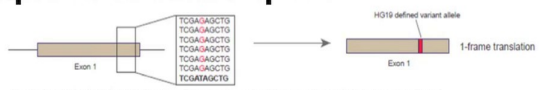
### Creating Variant Sequence DB



So, what you have is here your reference genome. So, this would be Hg38 or Hg19 and you have the sequence, you add your variants into the sequence within these exon boundaries, you do an in silico translation, and you throw these into your database and then you can actually search against your database to find them in your mass spec data.

(Refer Slide Time: 01:54)


### Example of Variant Peptide



Protein: NP\_001138550 zinc finger protein 805 isoform 2 [Homo sapiens]  
 Genome location: chr19:57764586+ 1485 0  
 DNA Variant: G183A  
 Protein Variant: V62I

```

MQGERLRPGLDSQKEKLP GKMSPKHDGLGTADSVCSRIIQDRVSLGDDVHDCDSHG
SGK NPTIQEENIEK NPTIQEENIEK CNECEKVFNKKRLRARHERIHSQVVPYECTECGKTFKSTY
LLQHHMVHTGKPKYKMECGKAFNRKSHLTQHQRIHSGEKPYKCECGKAFTHRST
FVLHNRSHTEGKPFVCKECGKAFRDRPGFIRHYIHSGENPYECFCGKVFKHRYSY
LMWHQQTHTGKPYECSECGKAFCEAA LIHHYVIHTGKPFCECGKAFNHRYSY
LKRHQRIHTGKPYVCSECGKAFTHCSTFILHKRAHTGKPFECGKAFSNRAD
LIRHFSIHTGKPYECMECGKAFNRSSGLTRHQRIHSGEKPYECIECGKTFCWSTN
LIRHSIHTGKPYECSECGKAFSRSSSLTQHRMHTGRNPISVTDVGRPFTSGQT
SVNIQELLLGKNFLNVTTEENLLQEEASYMASDRTYQRETPQVSSL
  
```




So, these are examples David actually just showed, but you know we might as well just look at them again, with the now understanding a little more about it. So, here you can see there is a SNP that is occurring at this location in this protein. And so, it is become goes from valine to an isoleucine, and then you can identify this within your data; or in some cases you can have a stop codon introduction.

So, instead of the being a full protein sequence, you will actually have part of it; as David mentioned this is a lot harder to validate right, because we do not always have coverage in proteomics. So, are we not seeing it because it is not, there are we not seeing it, because we are just not able to measure it.


(Refer Slide Time: 02:43)

### Example of Variant Peptide



Protein: NP\_899231 serine protease 48 precursor [Homo sapiens].  
Genome location: chr4:152198324+ 52,163,266,170,390 0,2623,4975,5944,13945  
DNA Variant: T984G  
Protein Variant: \*329E

```
MGPAGCAFTLLLLLGISVCGQPVYSSRVGGQDAAAGRWPQVSLHFDHNFYGGSLVSERLILTAACHIQPTWTFSTYVWLGSI TVGDSRKRKVKYVSKIVIHPKYQDITADVALLKSSQVTFSTAILPICLPSTKQLAIPFCWVTGWGKVKESDRDYHSALQEAEVPIIDRQACEQLYNPIGIFLPALEPVIKEDKICAGDTQNMKDSCKGDSGGPLSCHIDGVWIQTGVVSWGLECGKSLPGVYTNVIYYQKWINATISRANLDFSDFLFPIVLLSLALLCPSCAFGPNTIHRVGTVAEAVACIQGWEENAWRFSPRGREL TGEPLLTLGDFIYNLK
```




Or in some cases it can go from having a stop codon to having an amino acids. So, here we just continue on and you would get a normal protein, on that we continue translation after the original stop codon.

(Refer Slide Time: 03:01)

### Informatics Tools for Customized DB Creation

- **QUILTS**: python based tool to generate DB from genomic and RNA sequencing data (Ruggles et al., MCP 2015)
  - <http://quilts.fenyolab.org>
- **customProDB**: R package to generate DB from RNA-Seq data (Zhang B, et al. Bioinformatics 2013)
  - DOI: [10.18129/B9.bioc.customProDB](https://doi.org/10.18129/B9.bioc.customProDB)

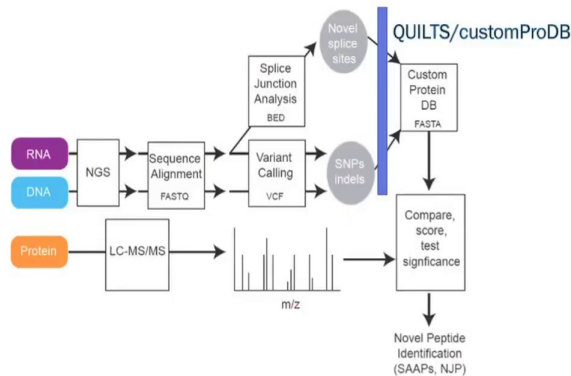


So, there are couple of different tools for this, and I mention two; because these are the two, that were created by two people who are in this room. He his group was responsible for customProDB and then David and I worked on quiltes; and so, you can use both of

these, I have link to them. So, you are able to put in either VCF files and or bed files and get back a database that has all of your information in it.

(Refer Slide Time: 03:29)

## Customized Protein Sequence DB Creation



So, it sort of sits here. So, after you do your next generation sequencing data, you get this splice junction, bed files and your variant, your VCF files and then quilts and custom Pro DB can then create databases for you, that you can then use to do your peptide identification.

(Refer Slide Time: 03:47)

## QUILTS web access



Welcome to QUILTS, a tool for creating sample specific protein sequence databases. It uses genomic and transcriptomic information to create comprehensive sample specific protein database that supports the identification of novel proteins, resulting from single nucleotide variants, splice variants and fusion genes.

To create your sample specific protein database, please choose a reference database:

Ensembl (GRCh37/hg19)  RefSeq (GRCh37/hg19)

And upload at least one sample-specific file:

**Germline variants (VCF)** (Example, only first 6 columns necessary): Choose File No file chosen

**Somatic variants (VCF)** (Example, only first 6 columns necessary): Choose File No file chosen

**Junction (BED)** (Example): Choose File No file chosen

**Fusion (TXT)** (Example, fifth column can be any quality measure): Choose File No file chosen

Variant quality threshold: 0.0

Threshold of supporting reads for novel splice junctions

Both boundaries annotated: 2

Left boundary annotated: 3

No boundaries annotated: 3

No missed cleavage

Once all data is uploaded click here: [QUILTS](#)

<http://quilts.fenyolab.org>



And here is quilts, if you want to go it is just a this is the web page. So, you are able to just upload your data here, and then you can get your databases out of it from there. This one right now it is you can just do human on the website; but if you are interested in other we can chat.

(Refer Slide Time: 04:10)

## Creating a Variant Peptide by Hand

Step 1. Open a genomics visualizer so you can see the gene annotation, reference genome sequence and variant in one window



Download IGV here: <http://software.broadinstitute.org/software/igv/>

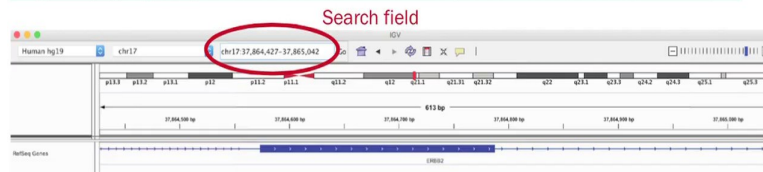


So this is where I wanted to talk about a kind of walk through what we are going to do on the hands on, before we actually do the hands on, let see if that helps with the actual hands on. So, you are welcome to try and follow one, but you know I have let us wait for questions until the actual hands on. So, I just kind of want to show you what we are going to be doing, and we are going to use a different example for the hands on itself. So, if you have not downloaded IGV please do so. So, we are going to do create a variant peptide by hands.

(Refer Slide Time: 04:46)

## Creating a Variant Peptide by Hand

Step 2. Move the RefSeq track to the top of the window and zoom into one exon on one gene by typing the gene name into the search field (ERBB2 used as example below)



The search field will change to chromosomal coordinates as you zoom in and out.



So, once you have IGV open you can zoom in and out on different genes pretty easily. And so, if you go to the search field and type in, I just picked a gene, so that you could just get sort of zoomed in. So, I picked ERBB 2 in the search field and then you are able to sort of zoom in on one gene and then you can change these coordinates either by picking a gene or just typing in different coordinates and it will move you along the genome.

(Refer Slide Time: 05:22)

## Creating a Variant Peptide by Hand

Step 3. Display the "Sequence Track" by dragging down just above the RefSeq track (red line below)



And so, the reason I wanted to show you this is because, there is this line this just you can drag up and down once you zoomed in enough, and well we if you do not have to do this right now, to we are going to do it during the hands on. But if you drag it down you actually get this sequence information of the genome and that is going to be something that is going to be really useful for actually doing the hands on itself.

So, you will end up getting something that looks like this and this is the. So, what is up here is the genome sequence and then this is the three through the three frame translation. So, this is amino acids and we will zoom in a little bit more. So, you can see, so it actually does the translation for you.

(Refer Slide Time: 05:57)

The screenshot shows a slide titled "Creating a Variant Peptide by Hand". Below the title is a green box with the text "Step 4. Zoom in further to see reference genome and proteome sequences (3 frame translation)". The main part of the slide is a screenshot of a genomic browser interface. The browser shows a genomic track for chromosome 17 (chr17) with a zoomed-in view of a 122 bp region. The sequence viewer displays the reference genome sequence in uppercase letters (A, C, G, T) and the proteome sequences in lowercase letters (a, c, g, t) for three different frames. The protein translation view shows the amino acid sequence for the ERBB2 gene, with the sequence "D L G L V Q G Y V L I A H N Q V K Q V F L Q S L R L R V I G C L Q L". The slide also includes the NPTEL logo in the bottom left corner.


So, you can zoom in and you see this is the genome sequence. And then here you can see that there is the amino acid translations from the genome sequence and then it has annotation for the actual ERBB 2 exons here as well.

(Refer Slide Time: 06:16)

## Creating a Variant Peptide by Hand

NOTE: You can flip the arrow to change the arrow to flip the strand in cases where genes are transcribed on the negative strand

\*\*Flip back if you're following along! We will be using the positive strand for this example\*\*



And then you can flip on this arrow to get if you are looking at a negative strand versus a positive strand. So, this is really important here and you can flip back and forth and it will flip the translation on the three, the three frame translation for you.

(Refer Slide Time: 06:32)


## Creating a Variant Peptide by Hand

Entry from your VCF file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
3	155623998	.	G	C	183	PASS	SOMATIC

Step 4. Upload your SNP VCF file to the browser so you can see the variants of interest

sequencePG-SNPs.vcf  
You will now have a track for the variants added to your display



So, when we are creating this variant peptide by hands, this is the first entry from the VCF file for the hands on, we are going to do the second entry; but you could upload files into, in your own files into this and then actually visualize them. So, here you would upload your the SNPs VCF file. So, it is the sequence PG dash SNPs dot VCF. So, you



would go to IGV there is a file and then you load the file and then you will see, if you go to this position chromosome 3 in the position within this VCF that VCF file. You actually are able to see it labels your variant. So, you have the variant labeled here.

(Refer Slide Time: 07:17)

### Creating a Variant Peptide by Hand

Entry from your VCF file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
3	155623998	.	G	C	183	PASS	SOMATIC

Step 5. Zoom in on your variant by entering the position (chr3:155623998) into the search field

Human Pyl19 chr3 chr3:155623998

Variant location

Sequence: T T A C A C A T G G A G A T A G T G T C A A A G T A G C T G A T G G A T T C

RefSeq Genes: GMP5

NPTTEL

So, then you can zoom in by entering into the search field exactly where the position is and you see the variant location of this gene here. And then what we will do is, we will just do the translation, in silico translation I am using the information that is provided here for the sequence. So, that we can actually sequence a peptide that would have this variant within it and then change that amino acid to the correct amino acid based on the SNPs that is there.

(Refer Slide Time: 07:51)

## Creating a Variant Peptide by Hand

Step 6. Determine the amino acid change occurring due to SNP

Entry from your VCF file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
3	155623998	.	G	C	183	PASS	SOMATIC

Variant:

Gene Sequence: `C T T A C C A C A T G G A G A T A G T G T A G A C A A A G T A G C T G A T G G A T T`

3 frame translation: `L T H G D S V D K V A D G I F`

Protein Sequence: `L T H G D S V D K V A D G I F`

*In silico translation*: L T H G D S V **H** K V A D .....

So, here we would start here. So, the variant again is here. So, it is a G to C. So, this is G. So, everything before the variant is the same right. So, we do not do anything, we can just kind of copy exactly what is there in the reference. So, it is L T H G D S V and then we get to the variant right and we got to figure out what it is. So, now, we know that.

(Refer Slide Time: 08:19)

## Creating a Variant Peptide by Hand

Step 6. Determine the amino acid change occurring due to SNP

Entry from your VCF file

ALT	QUAL	FILTER	INFO
C	183	PASS	SOMATIC

Ger:

3 frame Prote: `L T H G D S V D K V A D G I F`

tr: `L T H G D S V D K V A D G I F`

Reference Sequence: GAC → D (Asp)  
 SNP Sequence: CAC → H (His)

It was a GAC and now it is a CAC and then you can go up to your handy code on translations and figure out that it was a D and now it is an H. So, now you can add that in and then now you can have your In silico translation of your SNP at the protein level.


(Refer Slide Time: 08:41)

**Final Peptide (QUILTS output)**

>ENSP00000295920-D59H|guanine monphosphate synthetase  
[Source:HGNC  
Symbol;Acc:4378]|GN=GMPS|chr=3|type=S|SNP=G155623998C|qual=183.0  
00000|SAAP=D59H

MALCNGDSKMMNKVFGGTVHKKSVREDGVFNISVDNTCSLFRGLQK**EEVVLLT**  
**HGDSVHK**VADGFKVVARSGNIVAGIANESKKLYGAQFHPEVGLTENGKVLKNFL  
YDIAGCSGTFVQNRLECIREIKERVGTSKVLVLLSGGVDSTVCTALLNRALNQ  
EQVIAVHIDNGFMRKRESQSVEEALKKLGIQVKVINAHAHSFYNGTTLPISDEDR  
TPRKRIKTLNMTTSPEEKRKIIGDTFVKIANEVIGEMNLKPEEVFLAQGTLRPDL  
IESASLVASGKAELIKTHHNDTELIRKLREEGKVIPLKDFHKDEVRLGRELGLP  
EELVSRHPFPGPLAIRVICAEEPYICKDFPETNNILKIVADFSASVKKPHLLQR  
VKACTFEEDQEKLMQITSLHSLNAFLLLPIKTVGVQGDGRSRSYVCGISSKDEPD  
WESLIFLARLIPRMCHNVNRVVYIFGPPVKEPPTDVTPTFLTTLGVLSTLRQADFE  
AHNILRESGYAGKISQMPVILTPHFDRDPLQKQPSCQRSVVIRTFITSDFMTGIP  
ATPGNEIPVEVLLKMTVEIKKIPGISRIMYDLTSKPPGTTEWE

**Bold** = full tryptic peptide  
**Blue** = shown in demo  
**Red** = SNP



And then, so the output for quilts and custom Pro DB it will look something like this, well you will have a faster file, which has a header that sure has information about the SNP that was incorporated into this sequence.

And then it will have I bolded here the full tryptic peptide that would include that SNP, the blue is what I just showed in the demo I did not include the whole peptide, you could go through that one and scroll in and look at it and then the H is what they where the actual SNP is. So, we are going to do a similar example in demo a different SNP and actually going in the negative direction. Any questions on this that are not needed for ok, yeah go ahead.

Student: How is this tool different from ensemble, because this same thing we can do it here also?

Yeah there is a couple of different tools, I like this one the best that is why I use it ah; but there are I mean UCSC also you can use that, there is a couple of different genome browsers. So, like you can use your favorite, this is just the one that I decided was the most user friendly. So, I use it yeah, say that again.

Student: The variant file information, which the variant.

Yeah.

Student: The information only we can keep your in VCF file format or any other file format we also know like in this difficult tool which you are talking about.

So, you are saying can you use a different input format.

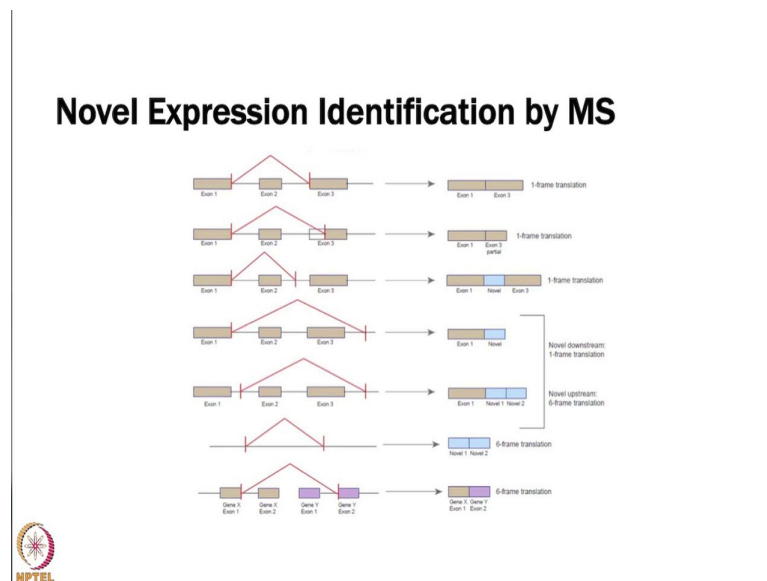
Student: Yeah.

You may be using MAP files, but I have not tried what kind of file format were you thinking.

Student: Like in file in which the header file like it is in fasta format and that header contain all type of useful information like at.

You will have to kind of parse it out and make it into a VCF file, is what I would recommend doing. It is definitely particular about it is file inputs, yeah. I think all of them are though, so other questions ok.

(Refer Slide Time: 11:00)



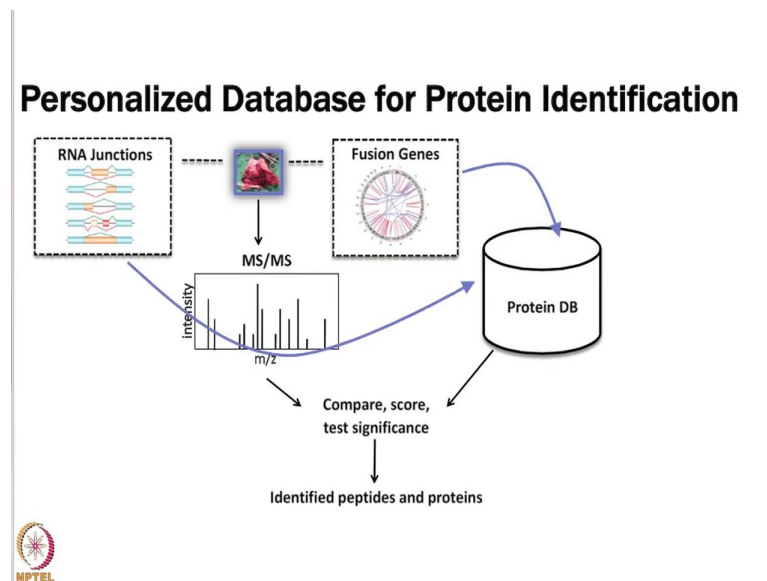
So, we can also look at novel expression; identification meaning novel alternative splicing or fusion genes. So, here are a couple of examples. So, for example, we are if we have two known exons, but let us say that they are combined in a way that is not annotated. So, that is a new alternative splicing event that does not have novel expression in terms of new exons; but it is just a new way of connecting exons. So, that is one example, you could have a example where one exon is connected to the middle of

another exon. So, it chops off the beginning of it, or maybe it is one exons connected to an intronic regions.

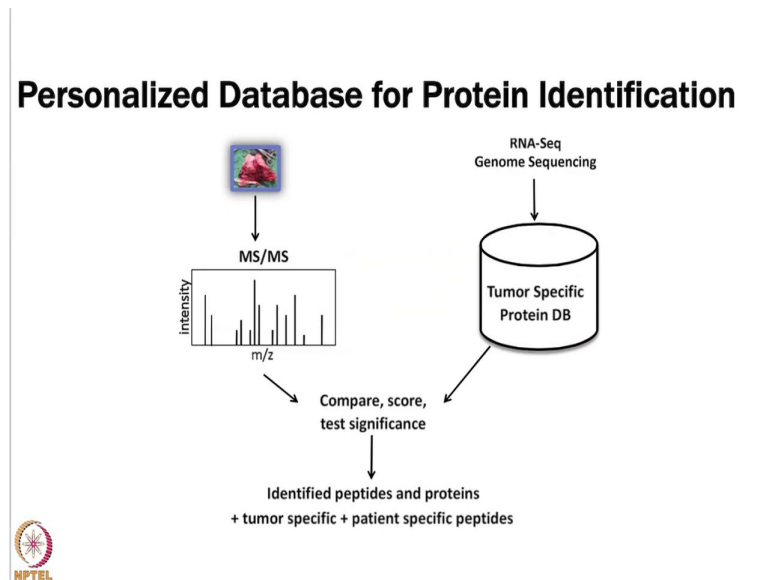
So, it is actually adding on some sequence before the annotated exon or maybe it is intergenic region, so it is a whole its past where we think the gene ends. Can occur or maybe it is just completely novel, maybe it is there is no boundaries, exon boundaries that exist that are annotated.

So, there is a lot of different ways that this can, these things can be combined and if you are doing genome annotation you have a similar this will be the same kind of problem that you will face. So, this can go for either of those questions. And also fusion genes, we talked to yesterday about fusion genes right; if you have gene x, one of the exons of gene x is connected to another exon and gene y, you are going to have a totally new gene or sequence that you would have to also add into your database in order to find it.

(Refer Slide Time: 12:28)

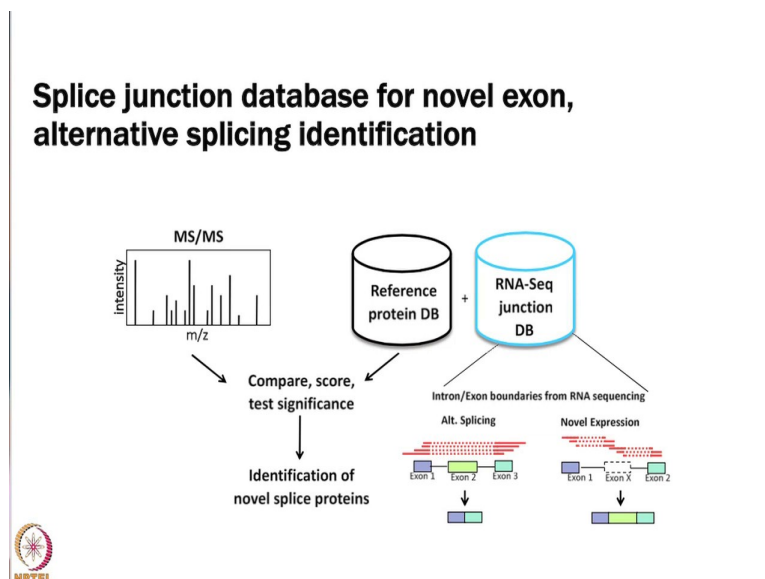


(Refer Slide Time: 12:35)



So, we can take our RNA seq data and the information we get from that including the junctions and the fusion genes throw those into our database and then we are able to find these tumor specific proteins as well.

(Refer Slide Time: 12:41)



So, yeah this is kind of the same. So, you can here have different new alternative splicing or it just completely novel expression ok.

(Refer Slide Time: 12:51)

## BED File Format

Chromosome	Start	End	Score	Strand	Name	Blocks
chr1	100000	200000	100	+	Gene A	100000-150000, 150000-200000
chr2	300000	400000	50	-	Gene B	300000-350000, 350000-400000
chr3	500000	600000	200	+	Gene C	500000-550000, 550000-600000
chr4	700000	800000	150	-	Gene D	700000-750000, 750000-800000
chr5	900000	1000000	80	+	Gene E	900000-950000, 950000-1000000

### BED File Format

Columns:

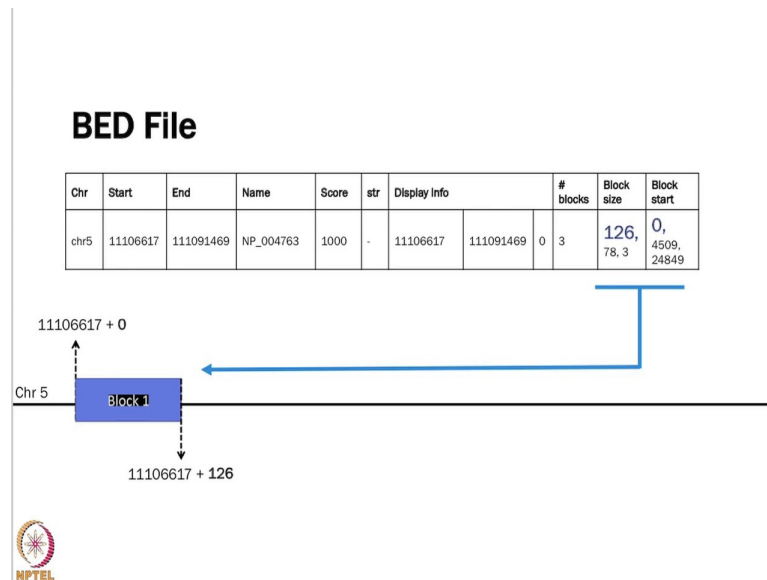
1. Chromosome
2. Gene Start
3. Gene End
4. Name
5. Score
6. Strand (+/-)
- 7-9. Display info
10. # blocks (exons)
11. Size of blocks
12. Start of blocks



So, that this really requires a bed file, I also include an example bed file in your zip folder, it has a very specific format; where you have the chromosome, your gene start and ends; in this case it would be a junction start and ends, but you can kind of treat it the same way.

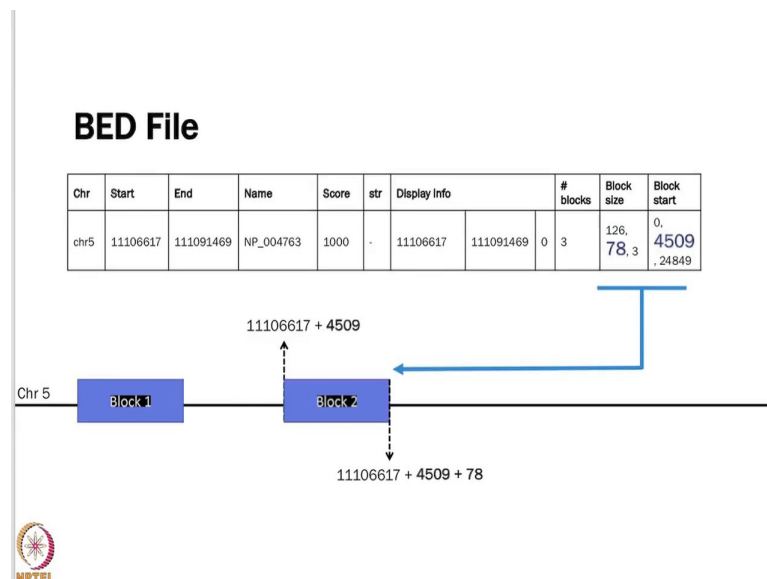
A name, a score, some information on strand, and just sort display info you will notice if you open the bed file I change some of the colors specifically, so that during the demo we can point things out. So, you can see there is a there is RGB numbers in there, that have been altered for that reason, the number of exons or blocks, the size of these exons at the start of these exons.

(Refer Slide Time: 13:36)



And I show this yesterday, I am just going to show it one more time because we are going to be thinking about this a lot. So, in our bed file, we have the start of the gene right and then we have the first block start it will be 0, we will start where the gene starts and then it will be the exon is 126 nucleotides long. So, it will make 126.

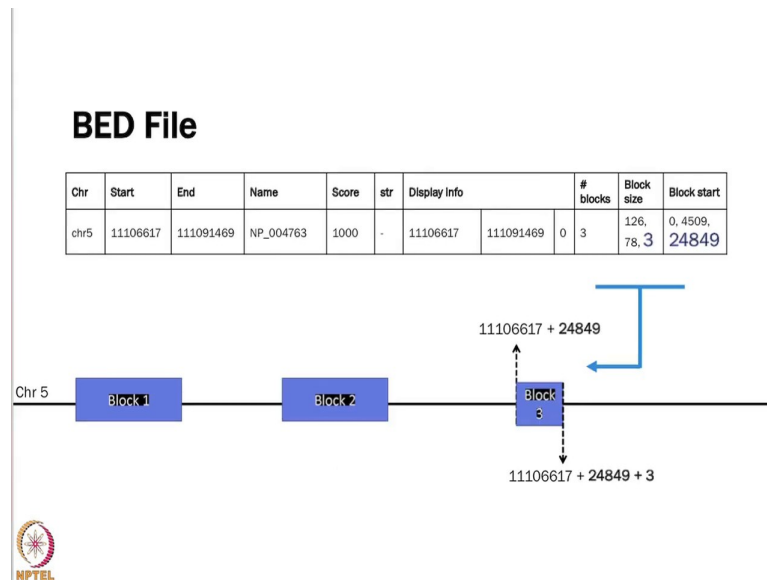
(Refer Slide Time: 13:58)



And then we will continue doing this. So, you always take the start and then you add the block start to get the start of the exon and then you add the block size to get the end of the exon.

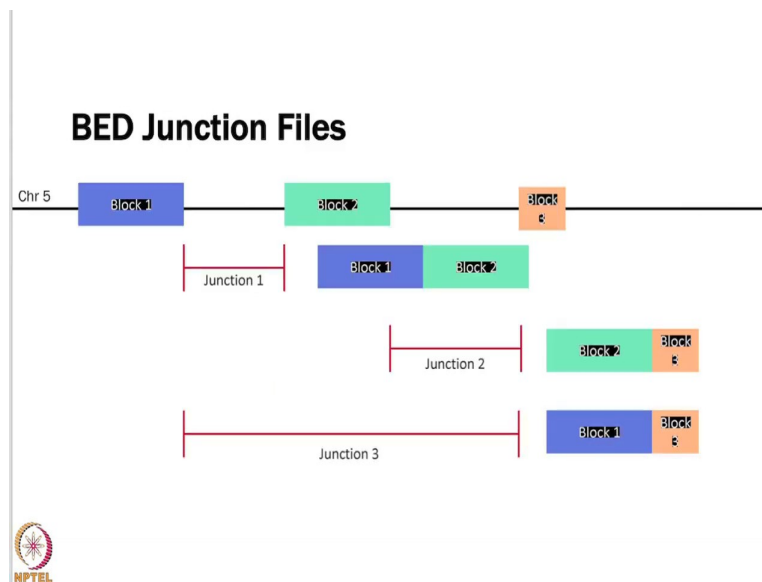


(Refer Slide Time: 14:09)



And you continue doing this for each of the different exons. So, that is how you surpass out the bed file to get the actual genome annotation from it.

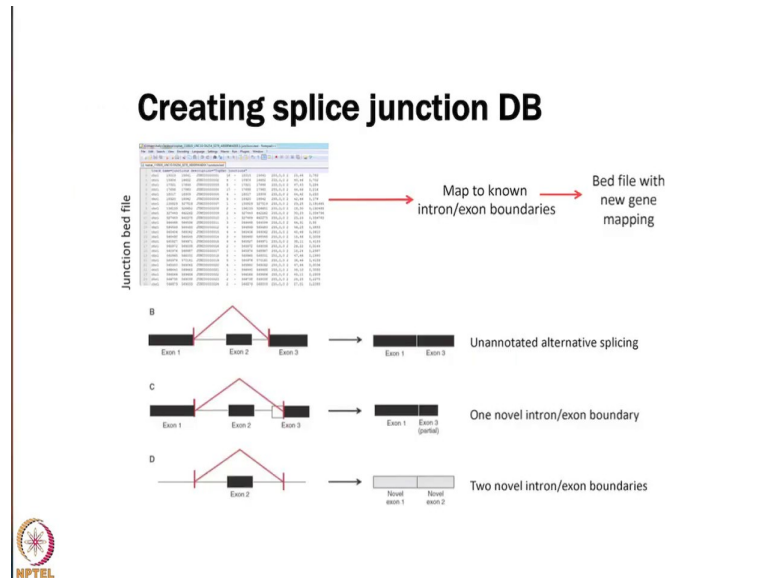
(Refer Slide Time: 14:19)



And the junction files if. So, if you run RNA seq on the sample; the junction files will just indicate where the boundaries between the exons that are spliced together are. So, if block 1 and block exon 1 and 2 are splice together, you will have a junction read here; if block 2 and 3 are spliced together, you will have a junction RNA here if block exon one

and three are spliced together so on and so forth. So, you will see how these junctions are you will be visualized in the IGV hopefully it will clear.

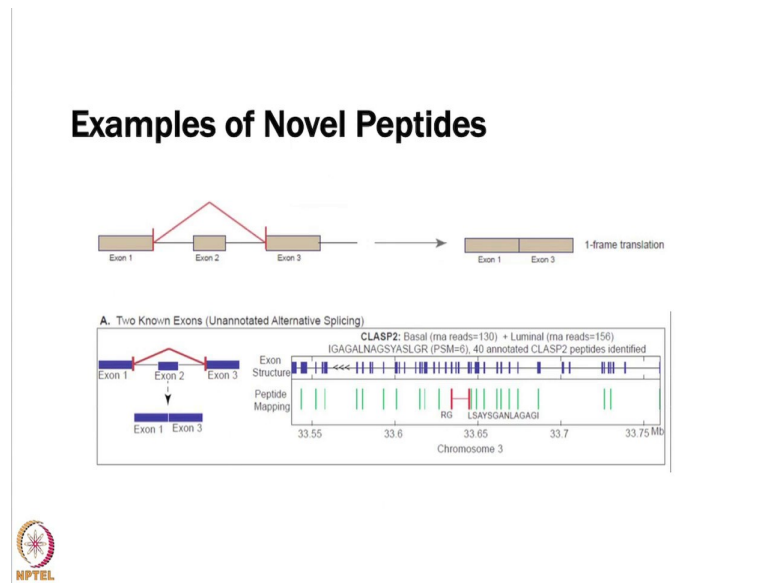
(Refer Slide Time: 14:54)



So, what we do to create this plus splice junction databases as we take this junction bed file, we compare it to the known annotation. So, we just take everything out that is known, because we know something we do not care about that we that is already included in our reference. So, we want to take those out and just get things that are new, and then so, we take the new file, bed file with the new gene mapping and then we figure out what kind of mapping it is.

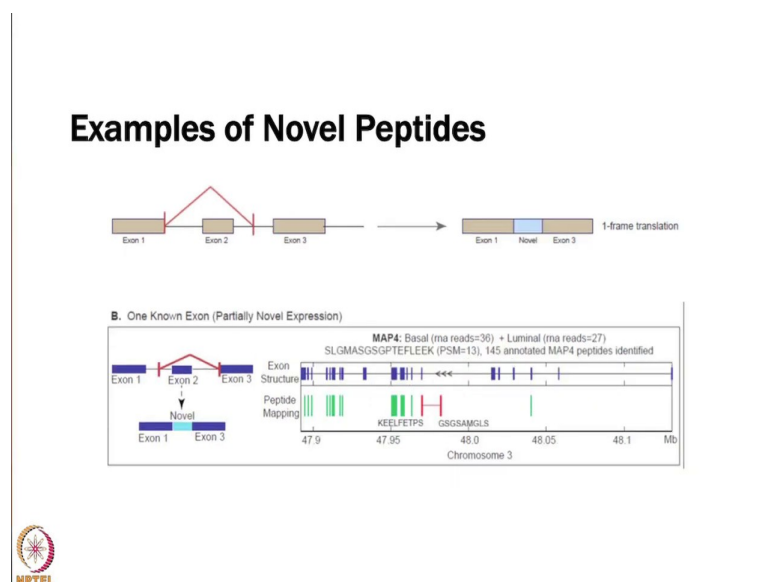
So, is it just a unannotated alternative splicing. So, we already know the exons, but it is sliced in a new way, is it what map it is mapped to one end of an exon; but not the other end is mapped to something new, or is it just completely new. And the way that we deal with this is changes based on what kind of novel splicing it is.

(Refer Slide Time: 15:48)



So, here would be an example of a alternative splicing with two exons that are known, right. So, and you can see here and this is the peptide data, and this is the exon structure you can see that, there is actually evidence for this in at the peptide level where you are connecting this exon and with a new with another exons that is known, but did not have annotation already in the database for that connection.

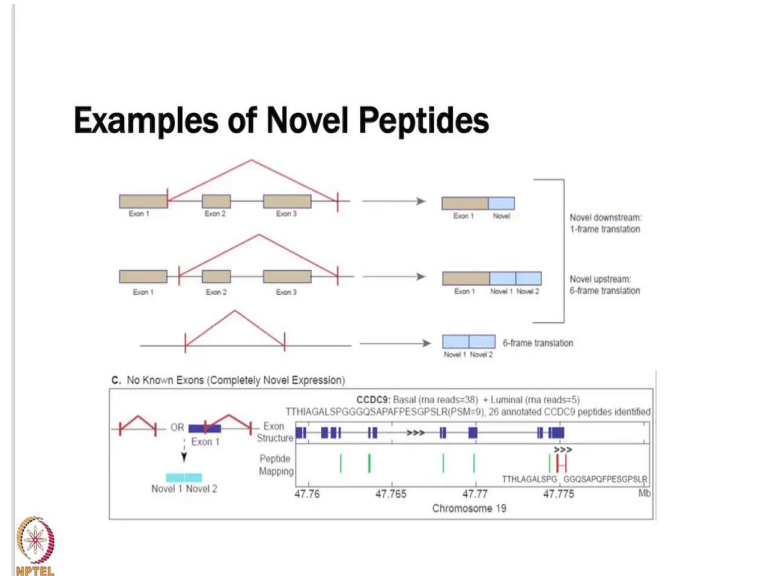
(Refer Slide Time: 16:15)



Or a connection between an exon and some intergenic regions, so here is an example where we have an exon that is annotated, connected to the this intronic region here and

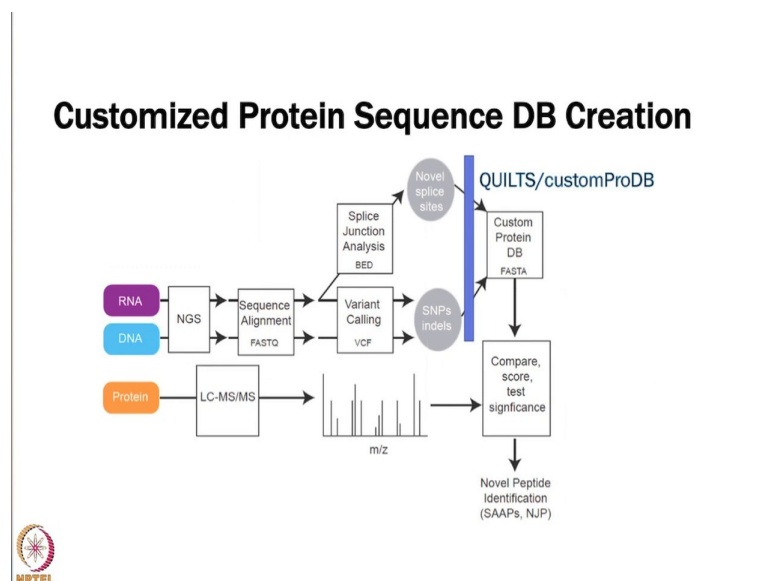
there is evidence at the peptide level for this connections; because we added this into the database.

(Refer Slide Time: 16:33)



Or also you can have as I mentioned these completely novel peptides, where it is just either in intronic or intergenic regions. And you can sort of see in this case, it was in the middle of an exon and then in the middle at the end of an exon. So, there is these are less likely especially in a really well annotated database, but you do find them.

(Refer Slide Time: 16:57)



So again you can put your bed file into these tools and just create your database using them; but we are going to do one by hand.

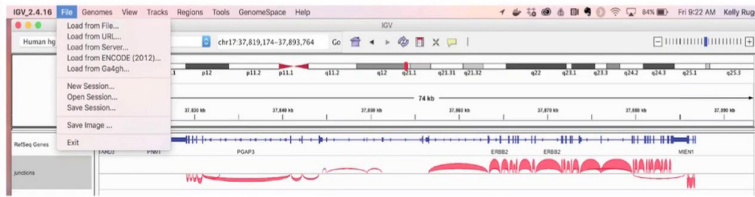
(Refer Slide Time: 17:10)

### Creating a Novel Splice Junction Peptide by Hand


Entry from your bed file

chr	start	end	name	score	strand	Display info	# blocks	size blocks	start of blocks

Step 1. Upload your bed file to the browser so you can see the junctions of interest



sequencePG-junctions.bed  
You will now have a track for the variants added to your display

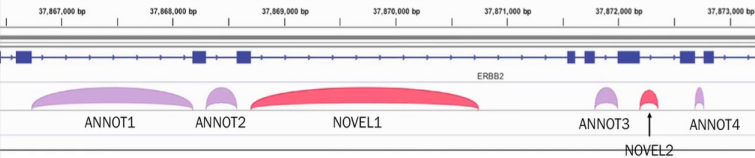


So, you upload your bed file, the same way you would upload your VCF file. So, you load here and I made you a very small bed file, but if you have the full bed file, you will have things that look like this, you will have like every single junction that connects all sorts of different exons.


(Refer Slide Time: 17:31)

### Creating a Novel Splice Junction Peptide by Hand

Step 2. Make sure you see 6 junctions in the ERBB2 gene  
(search for ERBB2 in the search field)

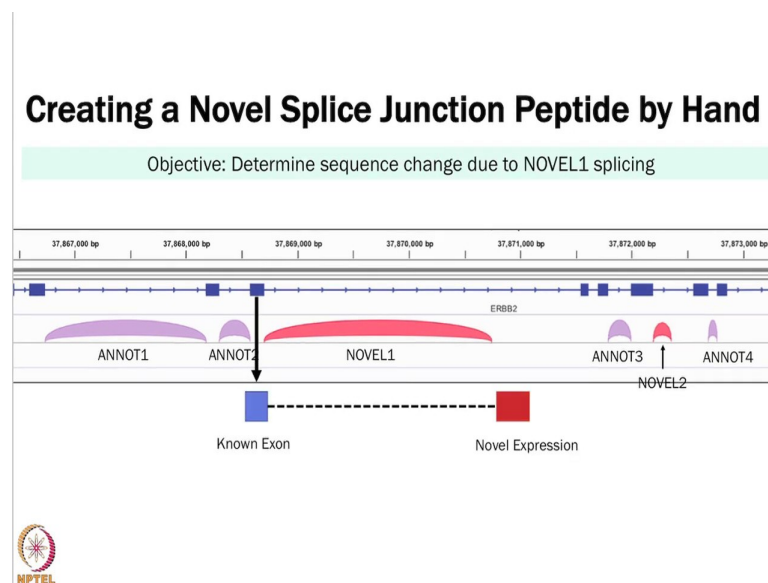


Purple = annotated alternative splicing  
Red = novel alternative splicing



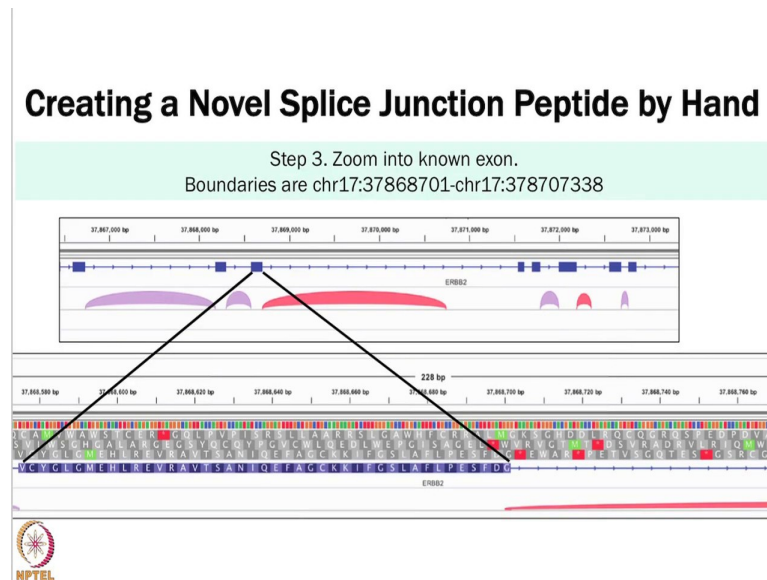
So, what I did give you a look should look like this. So, you will upload it and it will look like this there is 6 different junctions that are included. The purple ones are ones that are annotated and the red ones are novel; full disclosure I made up the novel junctions just for this purpose. So, and we may occur in reality, but I just made them for the demo. So, when we open this up and we will open it up and you can open it now, but we will open it up in the demo, you should see this.

(Refer Slide Time: 18:09)



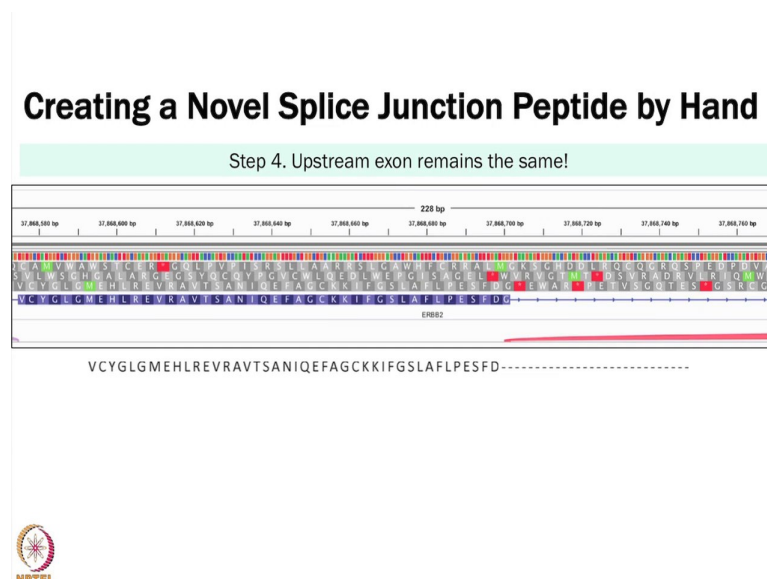
And what we will do right now, is just walk through how to by hand create this novel splice junction peptide. So, what you can see here is, these connects. So, the purple connects annotated exons. So, this will be, this exon connected to this one, this exon connects to this one and then there is this red junction that connects this exon to something in the middle of an intron. So, there is a novel expression here. So, we are going to do is figure out how to make the peptide that bridges the not the known exon with the novel expression

(Refer Slide Time: 18:43)



So, if we zoom in to this exon here, the known exon which I have included the boundaries for here; you will see that there is again this is the genome sequence and this is the three frame translation and then right here is the actual annotated gene. So, since this novel junction is actually downstream of a known exon, we can just keep this sequence as if right; because it is not changing frame and we know that this exon is still being transcribed in the same way.

(Refer Slide Time: 19:22)



So, now we have a translated sorry in the same way. So, now, we can just take this chunk of sequence and just sort of have it, and then we are going to add to the end of this the novel sequence; but what we have to pay attention to here right is that, we have this sequence that ends, but there is this G that is here, is actually not.

(Refer Slide Time: 19:52)

### Creating a Novel Splice Junction Peptide by Hand

Step 5. Check the boundary to see the last amino acid and the frame.

So 2 guanosines are left hanging and will attach to the other side of the boundary

Its assuming that the next thing that comes up is annotated, but it is actually just two G's that are hanging after the annotated D; and we are going to add something new to the end. So, we have to keep that in mind right.

So, we have these two quantities that are left hanging and they are going to attach to this new boundary. So, we keep the amino acid sequences, the peptide sequence from before this and we remember that the two G's are here.



(Refer Slide Time: 20:15)

### Creating a Novel Splice Junction Peptide by Hand

Step 1: Identify the splice junction boundary (chr17:378707338)


		Second letter			
		U	C	A	G
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCA } Ser UCG }	UAU } Tyr UAC } UAA } Stop UAG } Stop	UGU } Cys UGC } UGA } Stop UGG } Trp
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } CCA } CCG }	CAU } His CAC } CAA } CAG } Gln	CGU } Arg CGC } CGA } CGG }
	A	AUU } Ile AUC } AUA } Met AUG }	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } AAG } Lys	AGU } Ser AGC } AGA } AGG }
G	GUU } Val GUC } GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } Gly GGC } GGA } GGG }	

From previous exon | From novel expression

GG **A GAT GGT TAT ACC ACC ATG CCT...**

**C D C Y I I M R**

...GCKKIFGSLAFLPESFD**GD**GYTTP...

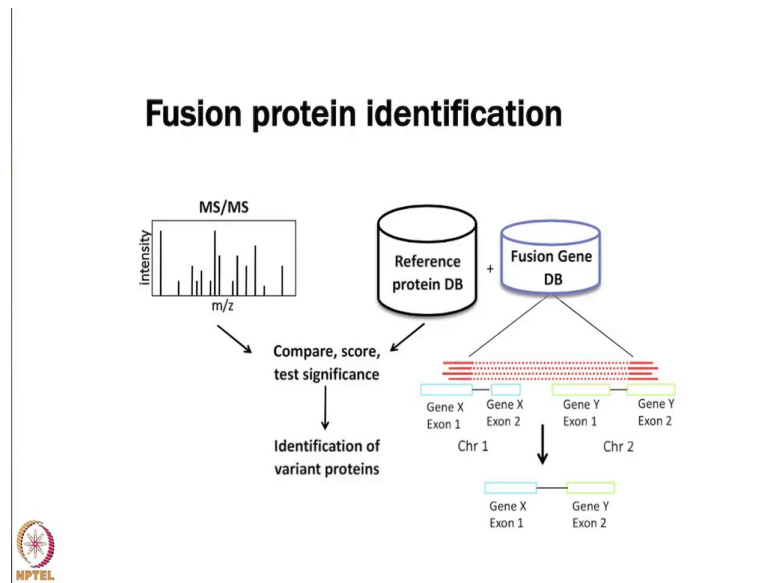


And then we look at the other side of the junction and we can see at the other side of the junction we take the two G's and then we add on the actual genome sequence from the other side of the junction here, from the novel expression. And then we are able to use this information right to actually encode, what this new novel expression would look like.

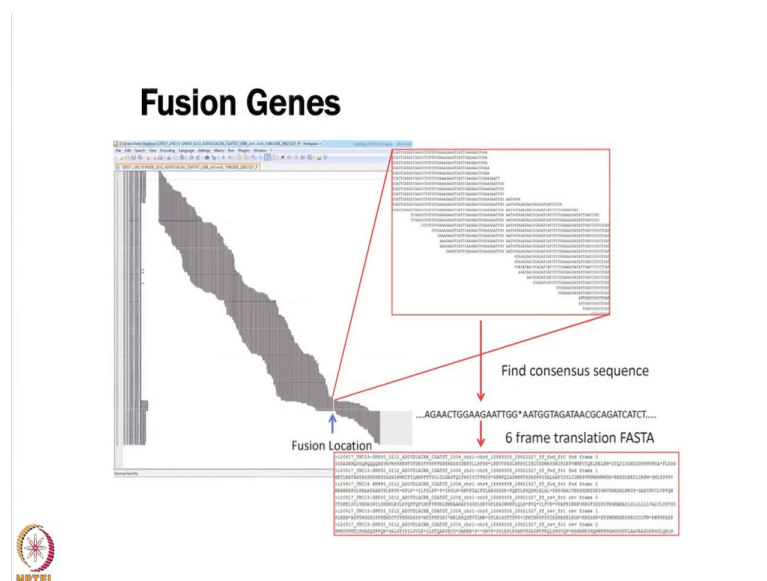
So, we can get our new amino acids and add those in. And you can then take these, so this G would be the barrier between the known and the unknown and throw these into our database as well. So, that we are able to find hypothetically this new boundary in our data. So, these are really hard to find right, because to prove to this is happening you really need to be able to identify this boundary between the known exon and the new exon.

And so, if you have to be able to find that one peptide that proves that is actually occurring, which is very it is not so likely. So, if you do not see, it does not mean it is not happening it just, if you do find it, it is exciting and then we will talk about the likelihood of finding these kinds of things in real data.

(Refer Slide Time: 21:30)



(Refer Slide Time: 21:36)



So, in addition to this as I mentioned, you can also include these fusion genes in your data. So, this is an example of what different outputs for fusion genes look like, I like this, because I think it looks, I do not know it looks like art.

So, if you zoom in on here these are actually sequences. So, everything up to this point is reads from RNA Seq for one gene and then reads from RNA Seq from another gene. It is just showing how the two are fused together and the this cancer data set. So, you would really want to take this boundary and add this into your database.

So, you would take the boundary here, find the consensus sequence of this boundary and then you do a 6 frame translations and add that into your database as well. Again very hard to identify this in real data, but worth trying if you have it ok.

(Refer Slide Time: 22:25)

## Sequence-Centric Cancer Proteogenomics

Cancer is characterized by altered expression of tumor drivers and suppressors

- Results from gene mutations causing changes in protein expression, activity
- Can influence diagnosis, prognosis and treatment

Cancer proteogenomics

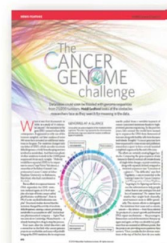
- Are genomic variants evident at the protein level?
- What is their effect on protein function?
- Can we classify tumors based on protein markers?



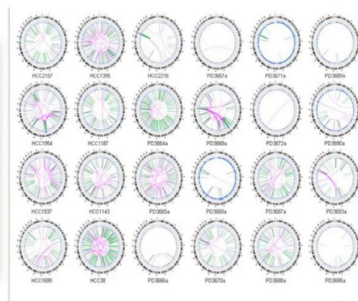
So, this is really specifically important in cancer, because as you have heard many times, there is a lot of altered expression either SNPs right and mutations causing changes in protein expression; and we really want to understand if they are these are found at the protein level, and if they are what their effect is on the protein function.

(Refer Slide Time: 22:45)

## Tumor Specific Proteomic Variation



Nature April 15, 2010



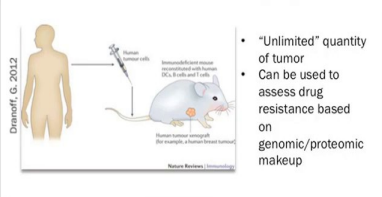
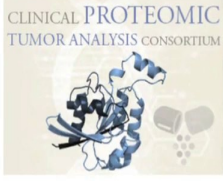
Stephens, et al. Complex landscape of somatic rearrangement in human breast cancer genomes. *Nature* 2009




This is showing that there is a lot of variability and how much variation occurs in different tumors, even if they are in the same type of tumor. So, here is human breast tumors, these circle plots just showing the rearrangement within each of the tumors and you can see that they are really different and highly variable. So, some of them do not have a lot, and some of them do.

(Refer Slide Time: 23:05)

## CPTAC Breast Cancer Projects

Patient Derived Xenograft (PDX) Tumors	CPTAC Retrospective Tumors
 <ul style="list-style-type: none"> <li>• "Unlimited" quantity of tumor</li> <li>• Can be used to assess drug resistance based on genomic/proteomic makeup</li> </ul>	
<p style="text-align: center;"><b>CompRef (2 PDX Tumors)</b></p> <ul style="list-style-type: none"> <li>• Two PDX tumors used as a quality control within experiments.</li> <li>• Measured repeatedly!</li> <li>• Use sample replicates to assess current depth of protein variant discovery</li> </ul>	<p style="text-align: center;"><b>Human Breast Cancer (77 Tumors)</b></p> <ul style="list-style-type: none"> <li>• 77 TCGA breast tumors from diverse breast cancer subtypes passed QC</li> <li>• Understand the functional effects of somatic mutations and chromosomal rearrangements</li> </ul>
<p>Ruggies KV et al., Mol Cell Proteomics 2016; 15(3):1060-71</p>	<p>Mertins P<sup>1</sup>, Mani DR<sup>4</sup>, Ruggies KV<sup>4</sup>, Gilette M<sup>4</sup> et al., Nature 2016; in press</p>



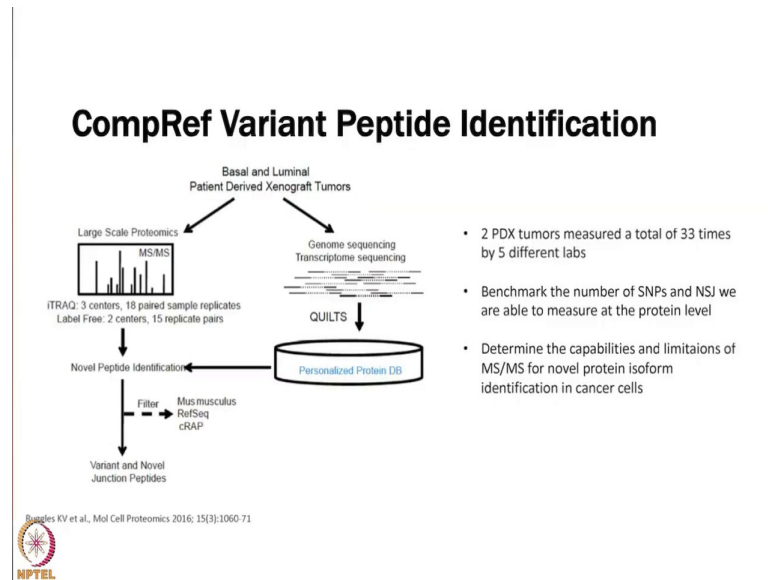
So, there is two studies that I wanted to talk about, one of them is a study you have already heard a lot about; it is the CPTAC retrospective breast study, where we looked at these 77 tumors to we looked at a lot of different things within these tumors. I am going to talk a little bit about how we looked at the effects of somatic mutations and also within the proteomics data; if we were able to, what we were able to find in terms of a mutations at the protein level.

And then also these patient derived Xenograft tumors. So, these are tumors that are injected into immunodeficient mice and they are able to grow on these mice. This is a very cool system, because you can have many many mice that have the same tumor; and you can treat them with different things or you can you can just grow lots and lots of the same tumor for all sorts of different QC experiments or just to better understand that tumor.

So, this was two different tumors that were really used this quality control within experiments, they are still being used actually for the studies that are ongoing. And what

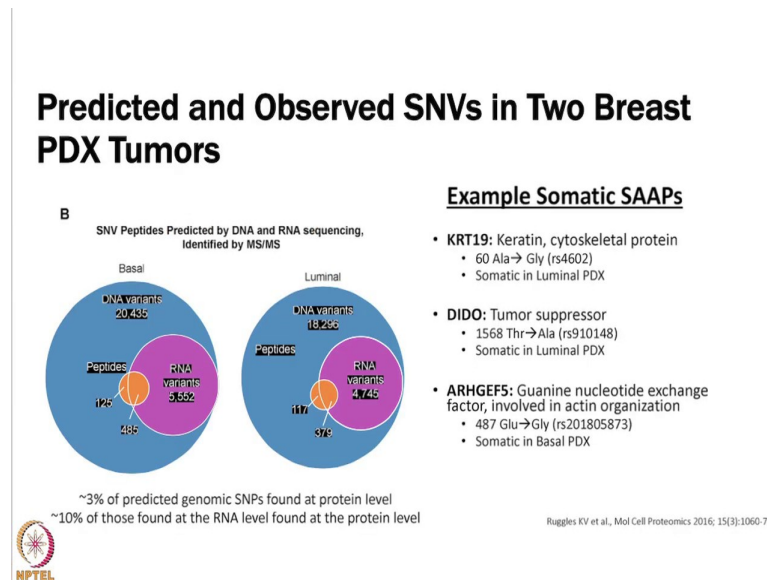
is cool about these is that they were measured over and over and over again. So, we used the fact that they were measured over again and over and over again to really understand where we are at in terms of our depth of discovery of these protein variants.

(Refer Slide Time: 24:30)



So, we will talk about that one first. So, we had these two tumors, they are a basal and luminal tumors. There was proteomics was completed as we have discussed with iTRAQ, and then there was genome and transcriptome sequencing that was done and these were incorporated into this protein database using quilts. And then we did novel peptide identification, filtered out the Mouse proteins and just the normal proteins we expect to see based on a reference database and then looked at things that were novel based, so either novel junctions or a somatic or SNPs.

(Refer Slide Time: 25:11)

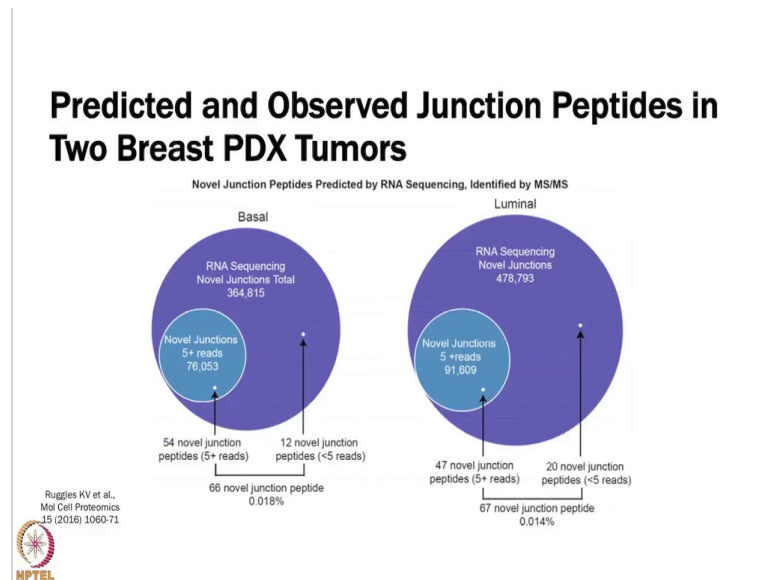


So, what we found was that. So, if you look these are the two different tumors. So, this is, the blue is all of the DNA variants that we identified, so in the genomics data. Then at the RNA level that is the purple, and then at the protein level it is orange; and this is for the basal tumor and the luminal tumor. So, what you can see from this is that, we were only able to find. So, about 3 percents of the predictive genomic SNPs were actually found at the protein level, and about 10 percent of those at the RNA level were found at the protein level.

So, there is a lot of reasons that this could happen right. So, maybe if SNP causes the protein to be degraded, so we are not going to find it; or maybe a SNP causes the protein, the peptide to be homologous to another part of the proteome. So, we are going to assume that it is normal, even if it is not; maybe we just do not have the coverage to find these. So, there is a lot of, just because we only find 3 percent does not mean that, those are the only 3 percent that make it to the protein level.

So, this was just kind of a way of assessing where we are at in terms of our ability to discover these. And this is just some examples of somatic SNPs that were identified that are cancer related. We also looked at the novel, these novel junctions, if we are able to identify novel junctions in our data.

(Refer Slide Time: 26:43)



And again the two different tumors, so the purple is all of the novel junctions that were identified by RNA Seq; the blue is ones that had at least 5 reads, so some of them if they are just like one read and probably just kind of garbage.

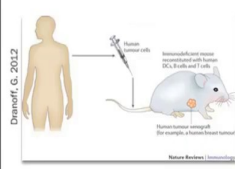
So, we wanted to make that clear as well, but if you could see these tiny little dots here, are the number of novel junctions that we were able to identify at the peptide level, so very very few. Again this may not be because they do not exist right, because in order to find to prove that these novel, splice site, splicing events are occurring, you have to find that peptide that exists right at that junction.

And maybe the peptide at that junction is too big or too small or homologous to something else or you know, so it is not that these are the only ones that exists, it is just that they are the only ones we were able to find using this method. I think this is also very, shows how low the human genome is annotated. I think that is the you know, if this was a less annotated, species we find a whole lot more new splice sites.

(Refer Slide Time: 28:02)

### CPTAC Breast Cancer Projects

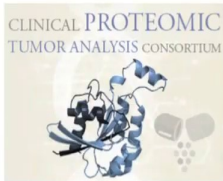
#### Patient Derived Xenograft (PDX) Tumors



- "Unlimited" quantity of tumor
- Can be used to assess drug resistance based on genomic/proteomic makeup

Dranoff, G. 2012

#### CPTAC Retrospective Tumors



**Human Breast Cancer (77 Tumors)**


- 77 TCGA breast tumors from diverse breast cancer subtypes passed QC
- Understand the functional effects of somatic mutations and chromosomal rearrangements

#### CompRef (2 PDX Tumors)

- Two PDX tumors used as a quality control within experiments.
- Measured repeatedly!
- Use sample replicates to assess current depth of protein variant discovery

Ruggles KV et al., Mol Cell Proteomics 2016; 15(3):1060-71

Mertins P\*, Mani DR\*, Ruggles KV\*, Gillette M\* et al., Nature 2016; In press

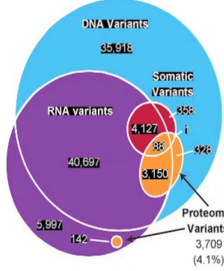


And then the last thing I will discuss is this just quickly is the, the CPTAC data, so that again the 77 breast tumors.

(Refer Slide Time: 28:16)

### CPTAC Breast Tumor SAAP Peptides

**a Non-synonymous Single Nucleotide Variants**  
Total: 90,806 (84,667 DNA; 54,201 RNA)




dbSNP 1930

COSMIC 1381

Novel 388

- **TP53:** Tumor suppressor
  - 273 Arg → Cys (rs121913343)
  - AAs 273-280 involved in DNA interaction
  - Somatic in 3 tumors
- **KRAS:** Cell proliferation regulating GTPase
  - 12 Gly → Val
  - Variant shown to cause constitutive activation
  - Somatic in 2 tumors
- **MYO1C:** Unconventional myosin IC
  - 826 Gln → Arg (rs9905106)
  - Somatic in 1 tumor, germline in 83

Mertins P\*, Mani DR\*, Ruggles KV\*, Gillette M\* et al., Nature 2016; In press



And so, here is actually an analysis where we combine all 77 and looked at them together; but it is a similar kind of Venn diagram, where we have all of our DNA variants, we have our RNA variants, and then we have only 4 percent of them were found at the protein level.




So, similar to what we showed in the last data set, and the red is just our somatic variants. And most of these were had been identified previously by, but just as existing the, so they either were in the dbSNP database, the COSMIC database and only a small percentage of them were completely novel. And then there are just a couple of examples here of, SNPs that were identified within this data ok.

(Refer Slide Time: 29:05)

### **Proteogenomic mapping**

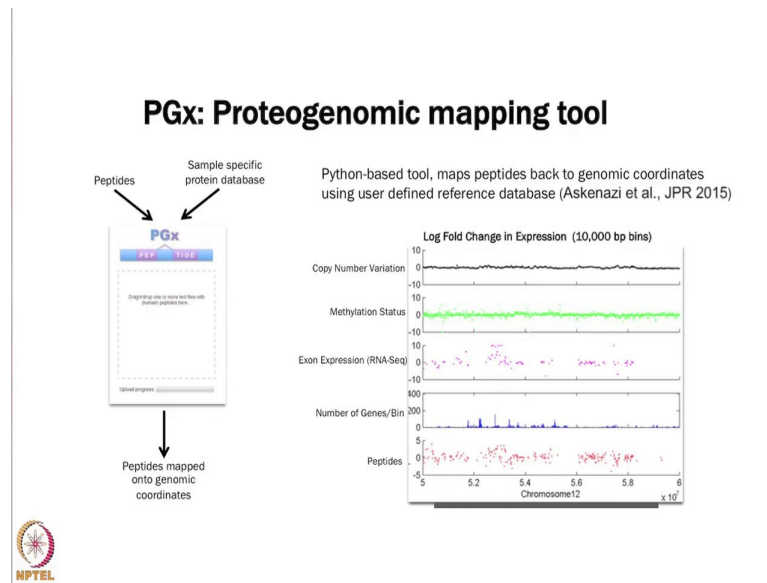
- Map back observed peptides to their genomic location.
- Requires tools to convert proteomic location to genomic coordinates
- Use to determine:
  - Exon location of peptides
  - Proteotypic
  - Novel coding region
  - Visualize in genome browsers (UCSC genome browser, Integrative Genomics Viewer (IGV))
  - Quantitative comparison based on genomic location



So, the last thing I just wanted to quickly talk about was map proteogenomic mapping. So, this is mapping. So, let us say you find new cool peptides, but you want to map them back onto the genome. So, this if you have a lot of them really requires automation. So, we have come up. So, this would be, the reason you would do this would be to try and visualize it alongside your genomics data.

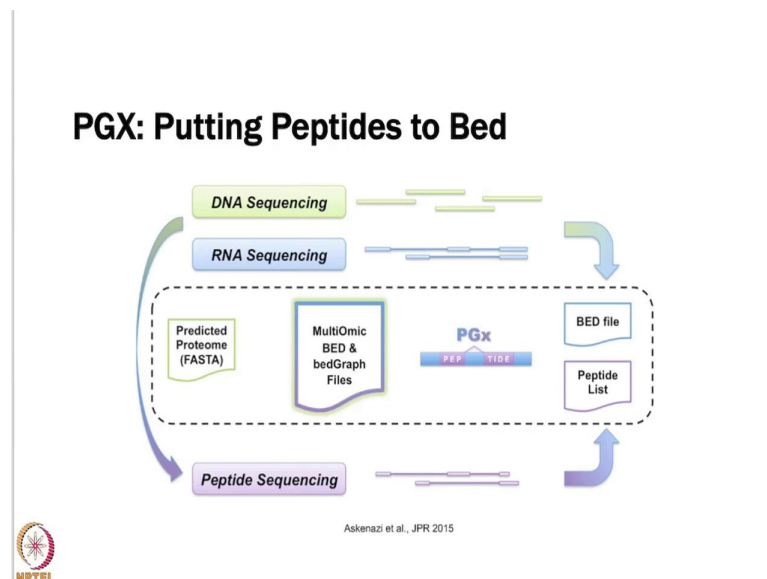
So, let us say you want to put it up in IGV and just see like we are alongside your junctions, make sure that it makes sense that your that is actually proving that your junction exists, and just kind of having it all on one in one browser.

(Refer Slide Time: 29:55)



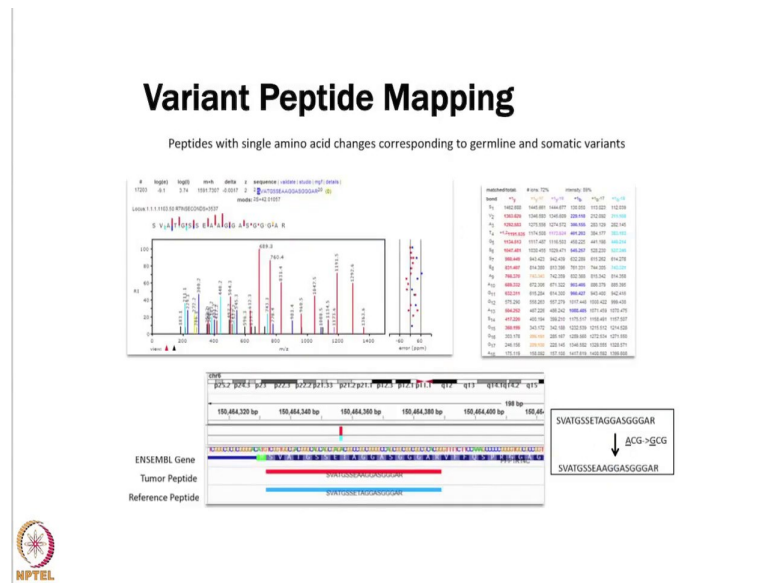
So, this tool PGx, where you can put your peptides and your sample specific database in and it will map on to genomic coordinates. So, here is just showing a schematic where you could have all your copy number, your methylation data, your RNA Seq data, and your peptides all maps to a chromosomal location and then you can look and see how things are quantitatively changing

(Refer Slide Time: 30:20)



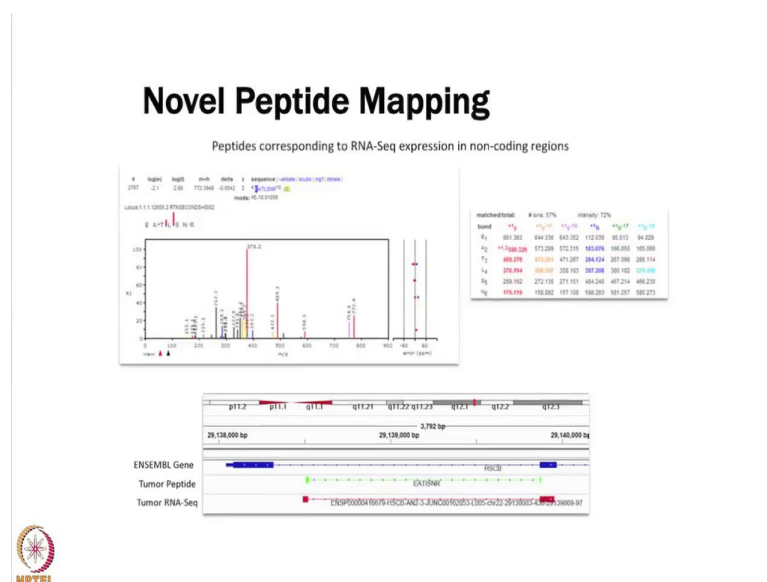
And so, it what happens is you can use all of this data to create bed files and the bed files again are what you then use an input into the IGV as we did with the junction data.

(Refer Slide Time: 30:37)



So, this is just an example. So, this is a spectra from this variant, that was identified in a tumor; I think this is probably from the compProdata. So, we have this, I can actually read about that is a T 2 A here and you can put this is a track showing where the variants are; and then you could actually just map your tumor peptide and just make sure it maps to the same place.

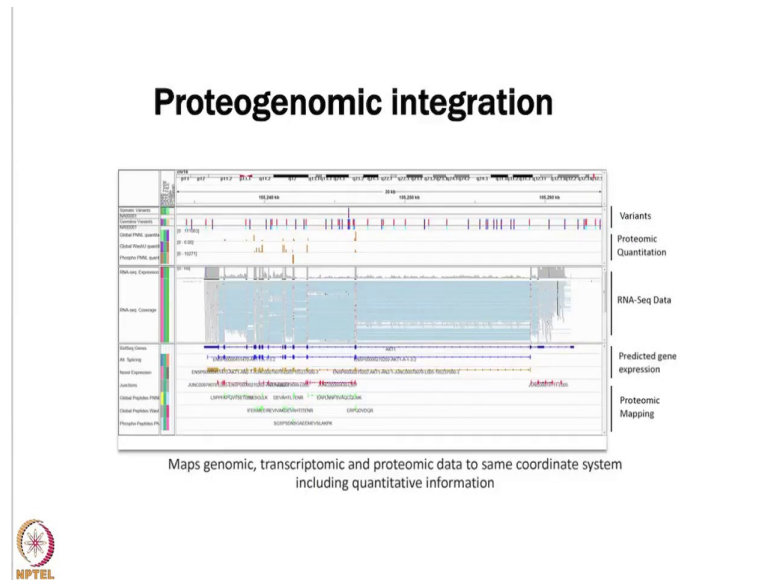
(Refer Slide Time: 31:03)



And then you could have your novel peptide data, where you have your junction. So, this is the RNA Seq data, and then you see in green this is peptide that spans the exact same

novel junction that we found at the RNA level. And so, you can actually visualize them together just to see that you are actually seeing the same boundary that was predicted by your RNA Seq.

(Refer Slide Time: 31:27)




And then you can throw everything up there and look at it all at once. So, this includes the reads from RNA Seq data. This is the annotation, this is proteomic mapping all of your variants. So, if you want to throw everything up there and look at it at once you can too.


(Refer Slide Time: 31:45)

## Thank you! Questions?

### Ruggles Lab


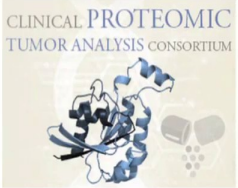



Dr. Hua Zhou   Dr. Amanda Erlund   Lili Blumenberg   Emily Kawaler   Angelina Volkova



Cooper Devlin   Tosh Cornwall   Shaleigh Smith   Lisa Katsnelson   Eii Niktab

### Fenyo Lab




So, I want to just thank everyone in my lab, the Fenyo lab and of course, CPTAC for almost of this work has been done.

(Refer Slide Time: 31:52)

### Points to Ponder

- Affect of variant and its analysis on other biomolecules leading to various clinical conditions
- Creation of customized database from genomics and RNAseq data
- Use of Integrative Genomics Viewer (IGV) to create variant peptide, manually using example of a cancer study.



MGOC-NPTEL IIT Bombay

I hope today you have learnt why is it important to know and understand the variant peptides. You also seen how integrative genomics viewer IGV can be operated and accessed to understand your data. You also heard about IGV helps us in finding what kind of mutations are present in a given gene by using VCF file, containing details of all the SNPs in the data.

Using the detail of mutated genes and type of mutation, one could create variant peptides as Doctor Kelly had just mentioned. IGV could also be used to visualize the novel expression due to the splicing. You also learnt about bed and bed junction files, which contain the information about various possible is splicing involved in a particular protein expression. The next lecture will be by a Master Spectrometry Scientist, Doctor Suman Thakur who will talk about proteomics in clinical studies.

Thank you.