

Introduction to Proteogenomics

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

Lecture – S16 **Integrative Genomics Viewer (IGV)**

Welcome to MOOC course on Introduction to Proteogenomics. After understanding the sequence centric proteogenomics, we will now listen Doctor Kelly Ruggles and she will talk about how to use IGV using one of the examples of a human gene. The prerequisites for the hands on is to download IGV on your system. You can see this link on your screen. If you have not downloaded the IGV yet, please pause this video and download the software to move forward.

(Refer Slide Time: 00:50)

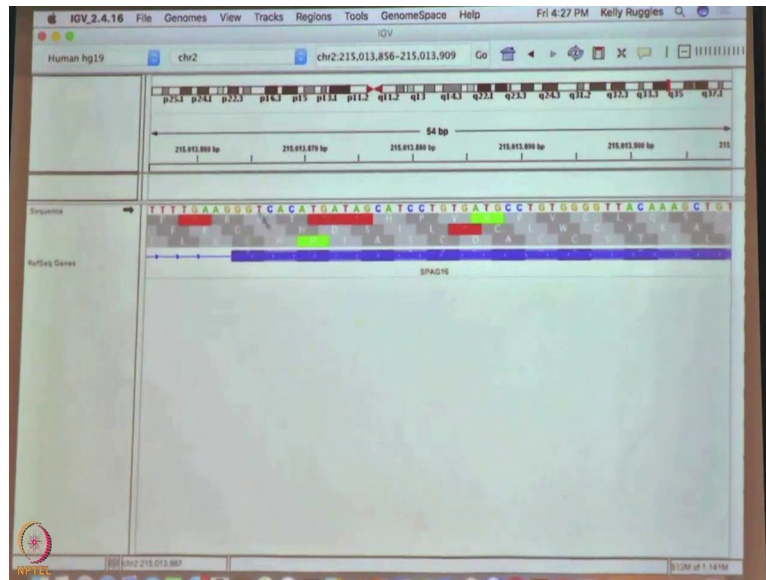
Download IGV on your system –

Link:

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

Keep a note you need to have Java on your system before you are start using the software and it is hands on session, else IGV will not work. You should also have the VCF file and dot ved files shared with you this week. Please download those files because those will be the input files for the software which we are going to use. Files name will be sequence pgs nps dot vcf and sequence pg underscore junctions dot bed. So, now you have java and igv installed in the system. So, let us welcome Doctor Kelly Ruggles for today's hands on session.

(Refer Slide Time: 01:54)



Ok. So, move the refseq up click on the specific area and then you should be able to like you will hover over this area here and you should be able to drag it down. Sometimes it takes a little. You have to zoom in quite a bit and you should be able to get the sequence information which of course I am not going to do, all right. I am doing a demo. There we go.

So, you should be able to zoom in enough that you are able to then drag it down and you will get the sequence information and what this is, is you zoom in further is the nucleotide sequence and then the actual amino acid three frame translation. So, everybody at least trying it to the point where you can see the sequence show translation.

Student: Ok.

If you are only getting the nucleotide sequence and not the amino acid sequence, if you left click and you will it will give you some options or right click and it will say show translation. So, click on show translation and then it will show you the three frame translation.

Student: Got this.

Yes same thing. So, if you right click.

Student: Ok.

Show translation.

Student: Right.

Yeah.

Student: I think you forgot to mention to use Java.

Oh.

Student: And windows.

If you do not have Java, it will not work. Thank you. Java.

Student: 8 0.

Is that the she you guys are having oh sorry take for granted these things.

Student: Sequence of these coming, but.

Yeah. So, again if you only see the nucleotide sequence, right click on the sequence and I will and then it will give you the option of opening the translation.

Student: But this one how you rating these. I am just getting to the point.

Yeah. So, left right click on the sequence.

Student: Ok.

Show translation

Student: Ok.

I am going to give you two more questions.

Student: I am just and this thing how to open these sequences.

Oh yeah. If you cannot open the sequences, you just have to zoom in really far and you will see them.

Student: Ok, but you know getting something.

Yeah then you have to right click on the sequence track right there. I will show you and then it should open translation. You are good.

Student: Yeah good thanks.

Right zoom in very very far. So, here I will show you, you can do like this. Oh yeah you can just select the window.

Student: Ok selecting the window ok.

Keep going; keep going and then it should pop up and if it does not, you drag this down, you play with this until it.

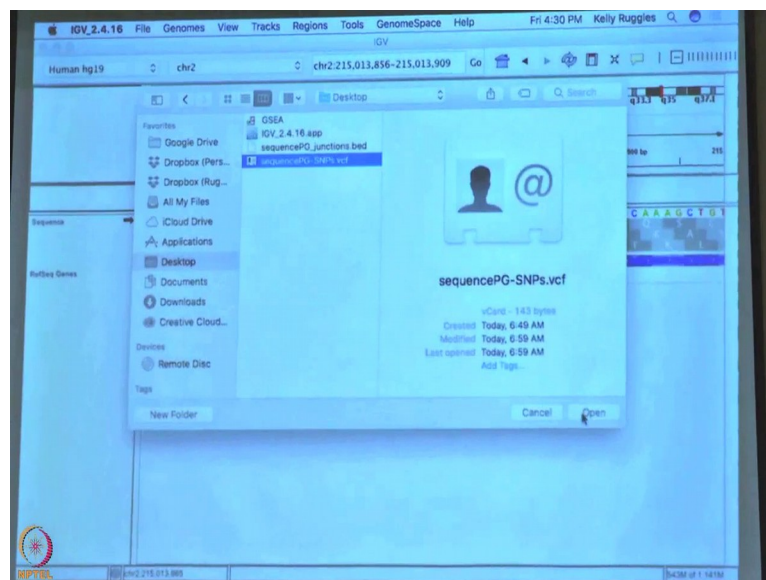
Student: Ok.

There you go.

Student: Yeah thank you.

Yeah ok. You should have your vcf and your bed files ready to go. So, go to load from file.

(Refer Slide Time: 04:55)



And then pick your vcf file. So, it should be sequence pg dash SNPs dot vcf. So, you will want to put in the example location of the second SNP which is chromosome 1 and then a very long number that you should copy I guess I could say it, but it will be easier if you copy it is 155, 646 348 and you will click on that and it should bring you to that location. So, you will see here that I have a gray box at that exact location since indicating that based on that vcf file, we uploaded the SNP is in the same location as that vcf file.

(Refer Slide Time: 05:44)

1. Creating a Variant Peptide by Hand

Entry from your VCF file

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	155646348		T	G	50	PASS	AC=CC

Step 5. Zoom in on your variant by entering the position (chr1:155646348) into the search field

Variant location

So, if you have gotten to this point what you will notice with this gene that is different from the last one. We looked at is that it is on the negative strands. So, you have to flip your sequence or else everything is going to be wrong. So, hit the arrow here and it will flip the nucleotide sequence and it will also do a it will also flip the translation. So, it will be in the correct frame and it will be no correct direction. So, we are moving in the negative direction right. So, we know that our variant is right here.

(Refer Slide Time: 06:19)

1. 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
1	155646348		T	G	50	PASS	AC=CC

Move in negative direction!

Variant
Gene Sequence
3 frame translation
Protein Sequence

? Q Q L R K R Q A P D In silico translation

Reference Sequence: CAG → Q (Glutamine)
SNP Sequence: GAG → H (Histidine)

And we know from our vcf file. So, this it goes from C to G. In this case it is G to C because we flipped it right. The objective here is to go through the sequence and just you can go from wherever you want to start and just by hands, just figure out how to create this peptide that has this SNP in it.

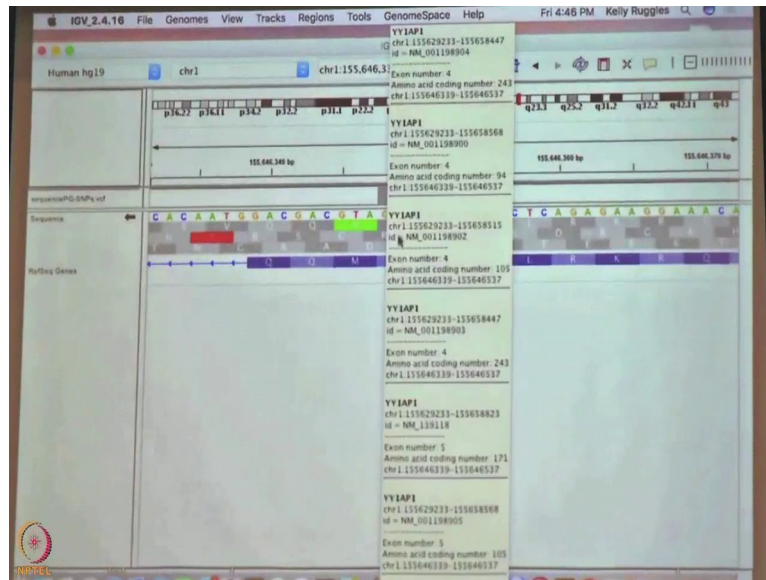
So, just take some time and what you will want to do is kind of what I did in the last one where you go from amino acid to amino acid until you hit the SNP and then you just make sure that you encode it correctly and then, you keep going just to get the final tryptic peptide. Does that make sense to everybody? So, you are going to do an in silico translation. I chose to start at this D. You can choose to start wherever you want to and you are essentially just going to keep moving oops until you hit the.

Student: Hit the?

variant. So, you are just going to move in this direction till you hit the variant and then you will figure out and I already gave you the answer what. what this is going to look like. So, just kind of think through it and then the junction one is a lot harder. So, I would rather us work through this now and then and so you are just going to creative tryptic peptide from right to left and then when you hit the SNP, just make sure you change it accordingly. So, there because you know that this G is changed to C, all I wanted you to do is to take a look at the to go.

So, this is what you should essentially have in front of you in some level. So, if we wanted to put this SNP into our database, we would be moving from right to left and we would be creating a peptide sequence. So, we know that trypsin is going to cut at R. So, I guess we could start here.

(Refer Slide Time: 08:18)



And we can go l q q and then at this q we are going to have instead of CAG, we are going to have CAC. So, when I was showing you guys this during the talk. Ok this guy.

(Refer Slide Time: 08:34)

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

	U	C	G	
U	UUU Phe	UUC Tyr	UGU Cys	U
U	UUC Leu	UUA Stop	UGC Cys	C
U	UUA Leu	UAG Stop	UGA Stop	A
U	UUG Leu	UAC Tyr	UGG Trp	G
C	CUU Leu	CAU His	CCU Pro	U
C	CUC Leu	CAC His	CCC Pro	C
C	CUA Leu	CAA His	CCA Pro	A
C	CUG Leu	CAG His	CCG Pro	G
A	AUU Ile	AUA Stop	AUG Met	U
A	AUC Ile	AUA Stop	AUG Met	C
A	AUA Ile	AUA Stop	AUG Met	A
A	AUG Met	AUA Stop	AUG Met	G
G	GUU Val	GAU Asp	GGU Gly	U
G	GUC Val	GAC Asp	GGC Gly	C
G	GUA Val	GAA Asp	GGG Gly	A
G	GUG Val	GAG Asp	GGG Gly	G

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

So, you should be able to figure out and so, what that SNP is going to encode instead of the one that it is currently.

(Refer Slide Time: 08:51)

1. 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
1	155646348	ex2	C	G	151	PASS	SOMATIC

Move in negative direction!

Variant Gene Sequence: CAGGACCTTACGAGAGAAAGAAACAGCCCA

3 frame translation

Protein Sequence: ... Q Q L R K R Q A P D

In silico translation

Reference Sequence: CAG
SNP Sequence: GAG

So, right now we have a CAG, but really it is changing here. So, figuring out what that what that looks like is this the right; this is the right example right yes.

(Refer Slide Time: 09:23)

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  
MALCNGDSKMMNKVFGGTVHKKSVREDGVFNISVDNTCSLFRGLQKEEVLLT  
HGDSVHKVADGFKVVARSGNIVAGIANESKKLYGAQFHPEVGLTENGKVLKNFL  
YDIAGCSGTFVTQNRLEECIREIKERVGTSKVLVLLSGGVDSTVCTALLNRALNQ  
EQVIAVHIDNGFMRKRESQSVVEEALKKLGIVKVINAAHSFYNGTTLPISDEDR  
TPRKRIKTLNMTTSPEEKRKIIGDTFVKIANEVIGEMNLKPEEVFLAQTLRPDL  
IESASLVASGKAELIKTHHNDTELIRKLREEGKVIPLKDFHKDEVIRILGRELGLP  
EELVSRHPFPGPLAIRVICAEEPYICKDFPETNNILKIVADFSASVKKPHTLLQR  
VKACTTEEDQEKLMTSLHSLNAFLLPKTVGVQGDGCRSYYVCGISSKDEPD  
WESLIFLARLIPRMCHNVNRVYIFGPPVKEPPTDVTPTFLTTGVLSTLRQADFE  
AHNILRESGYAGKISQMPVILTPLHFDRLPLQKQPSCQRSSVIRTFFITSDFMTGIP  
ATPGNEIPVEVLLKMYTEIKKIPGISRIMYDLTSKPPGTTWE
```

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

So, this is all I wanted you to think about what it was, how replacing this nucleotide would end up with a new amino acid that then we would then change the faster file that we would use in our database. So, here is what the QUILTS, you can put this into QUILTS as a vcf files. You do not have to actually do this by hand. So, if you were to put this one line of the vcf file into QUILTS, it would give you a new protein sequence that would include that

change in the amino acid based on the SNP that was given in the vcf file. So, that is just sort of what I wanted everyone to be able to like get a feel for using the IGV itself, so that sound does everyone kind of get where the next one is going to be a little more difficult.

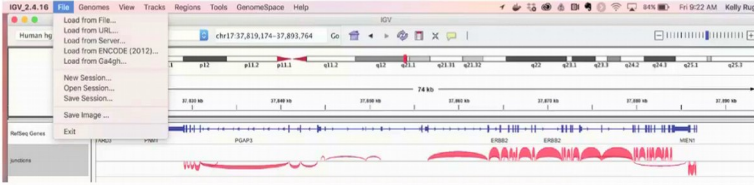
(Refer Slide Time: 10:02)

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

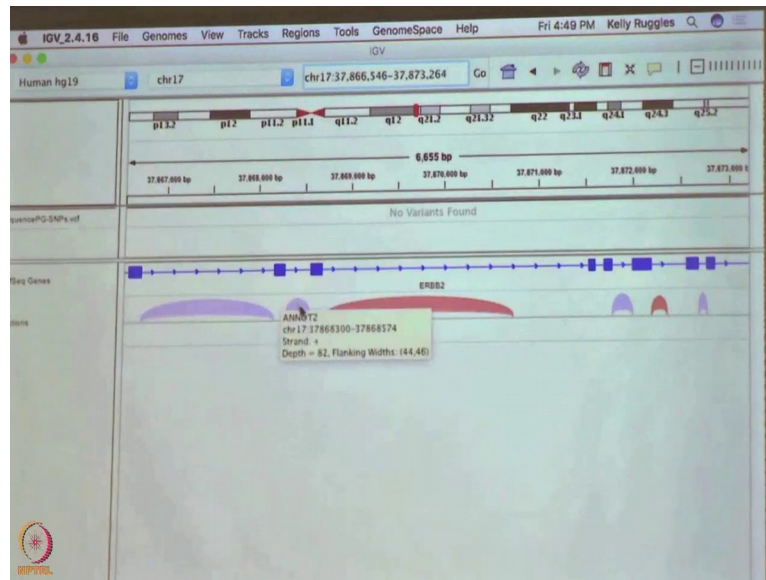


sequencePG-junctions.bed
You will now have a track for the variants added to your display



So, the other thing we can do is looking at these novel splice sites. So, I made this junction file. So, if you upload the junction file very similarly to what you did before. So, if we go to file load from file junctions that bed and you open this. So, this one I only if you I just have junctions for ERBB 2, I made it easy. So, in the field here go ERBB 2 and it will bring you to the ERBB 2 annotation. Once you are there, you will see that there are a couple of junctions that should be in your junction file.

(Refer Slide Time: 10:49)



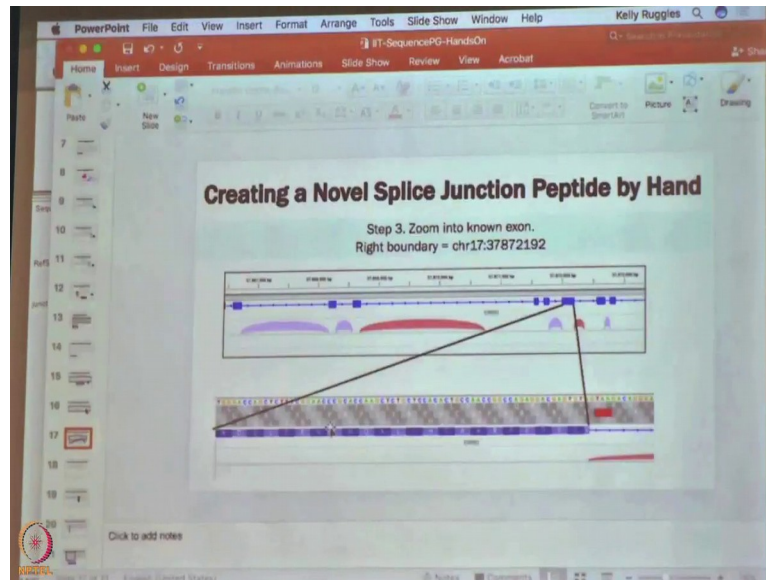
There I changed the color. So, that the purple ones are ones that are annotated. So, you can see that they connect known exons here and the red ones are novels. So, the one I did we looked at one of these novel ones during the lecture, the other novel one is what I was.

So once you have the file open, it looks like this we are going to look at what this novel junction 2. What the translation of that novel junction 2 would look like if we were throw it into the sequence database? So, what you want to do is zoom in on this second red junction. So, what you do how you zoom is you can just come up here and you actually just create the window around what you want to see. So, it might take a second to load, there we go.

So, you should see the end of this exon and then the beginning of this junction and so, the junction is just showing like essentially the RNA seek indicated that there was a connection between this exon, the end of this exon and this some area within this intron; this is what it is telling us.

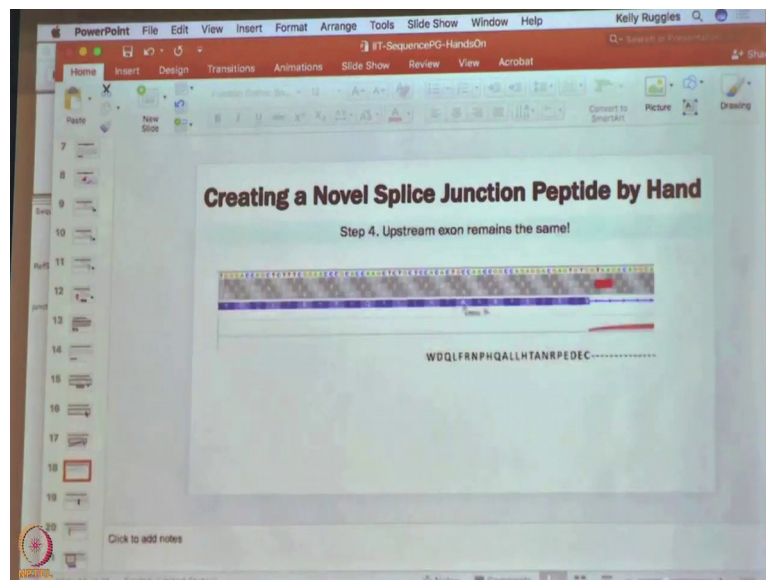
So, if you go to the end of the exon you will see that you can get oh by the way we are looking at a gene that is going it is on the forward strand. So, change your sequence back, arrow back or else it is not going to work. It will be very confusing. Does that make sense to everyone, ok. So, what you will do here is, we are going to make another tryptic peptide. So, again we will start after the arginine. So, it will be PEDEC and then there is this extra g hanging over like we saw on the original we had two g's in the last one and now we have one g.

(Refer Slide Time: 12:46)



So, I am going to show you in my PowerPoint because I think that might be easier.

(Refer Slide Time: 12:50)



So, you zoom in here and we can get the sequence because we know when there is a boundary. So, there is a junction between a known exon and something new, right. So, the thing that is known we can keep as is we could just take the sequence from that and then we just have to figure out with the boundary next to it what that extra sequence is going to look like. So, we take this sequence, we have this g that is hanging and then we look at the other side of the boundary. So, zoom in to the other side of this boundary here and this is where

you will get the nucleotide sequence that will continue on from the original exon and then, you can figure out you can do an in silico translation to figure out what that full peptide will look like. So, what these are showing are that are the boundaries. So, it is just showing that these two connect by splicing.

Student: Ok

So, it is just showing that this exon connects to this exon which we know.

Student: This one is connecting these two exons. This we know now.

Yeah we already know that we that is already annotated in like the genome. We know that these two exons place together.

Student: Well you know now ok.

We knew that that is like somebody else figure that out.

Student: Who somebody else?

Is in the database.

Student: So, why that is my question of why it has not come as an exon here then?

So, it is the junctions. Just show the boundaries between exons.

Student: The this will be it.

So, just showing how they connect.

Student: Ok.

Yeah.

Student: You mean to say this is only the junction part.

Yes.

Student: Ok.

Yes.

Student: Ok maam.

Student: Excuse me maam where these red, red and green?

The, so the green are Methionine and the red are stop codons.

Student: That is ok.

It is just showing starts and stops essentially.

(Refer Slide Time: 14:37)

Creating a Novel Splice Junction Peptide by Hand

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

junction boundary (chr17:378707338)

From previous exon: GG
From novel expression: A GAT GGT TAT ACC ACC ATG CCT...

...GCKKIFGSLAFLPESFDGDGYTTPM...

So, you should be able to get from the sequence data, you should be able to manually figure out what the nucleotide sequence would look like at that boundary. So, here in the file I sent you it actually has it in there and then you can do an in silico translation to figure out what that amino acid sequence would look like as well.

So, this sequence here you would throw into your database to see if this boundary actually came up at the with the protein level essentially. So, we have software to do this. You do not have to do this by hand, but I think by doing it by hand, you better understand what that what these databases actually are and if you may at some point have to do something like this by hand. So, yeah.

Student: So, it here it is same that these are the different junction's right 6 exon?

Yes.

Student: Per ERBB 2 what are junctions exactly that splicing?

Yeah they are kind of showing how things place it together. So, they are showing the connections between exons.

Student: Ok these are the connections. So, if I am zooming it in.

Yes.

Student: So, here.

That is just because you are too far and.

Student: Ok. So, those two ends are going to join. So, here these are the stop codons. So, these are.

So, those are only. So, if you have since you have three frames that is why one of the.

Student: Ok.

Frames in your in an intron here. So, like none of that is what matters.

Student: Ok, so this is the intron right.

Yeah.

Student: Ok.

So, if you zoom out you would not see in a exon.

Student: Using that ok.

Yeah.

Student: So, these two are going to join in because this is an intron.

Yes.

Student: So, these two will join.

Yes.

Student: And here there is no joining, but here it is joining here.

In the middle of nowhere.

Student: So?

Student: You can detect a intuitive splicings.

Student: What slicing?

So, that is so that's i what your sleeve what you do want to do is find a sequence here and then the nucleotide sequence here and then figure out what that would look like if it actually joins.

Student: Ok. So, if it actually joins what the amino acid will.

Exactly.

Student: Ok.

Yes

Student: First we are going to the genome the file we loaded is it loading from the file, then can you just interpret what exactly we are doing and what we will get it.

Yeah so.

Student: Please conclude then it will be.

So, I just wanted to show two examples of if you have a SNP and you how do we think about how that SNP would be encoded into the proteome. So, that we could throw it in the database since then find it. So, if you have that one SNP how does it impact the peptide?

Student: Ok.

For the junctions it is if you have some expression.

Student: Because of that change in a SNP.

Exactly.

Student: Whether a new peptide will be there?

Exactly.

Student: Or not?

It will be yeah.

Student: That is what you will have trying to see.

Yes

Student: But here from.

And then this one. So, for the junctions the junctions are showing.

Student: How to see that?

Where the two exons are connecting.

Student: Ok.

So, that say that again. So, like this is showing these two exons are connecting.

Student: Ok.

And this one is showing the red one showing there is an exon that is connecting to the middle of an intron which makes no sense right like we would not expect to see that because we expect to see two exons joining, but not an exon in an expression in an intron.

Student: Ok.

So, if it is like cancer we want to see is that intronic expression real is that like some new isoform that we have never seen a normal tissue that is existing in cancer and if we want to see that, we have to be able to encode the sequence in the intron as a protein because it would never be in a normal database because it is a new expression in an area. That is not normally expressed.

Student: So, this is just me, what you are doing it is based on the genomics data. So, whatever.

Yeah.

Student: It may have to SNP genotyping data. So, suppose for the same you have MS/MS data also.

Yeah.

Student: So, how to integrate these?

So, yeah that is a good question if you have. So, that is the whole the reason we would do this is we would we would get the new peptides. And then we would search the MS data within each peptide.

Student: That that new peptide?

And see if it comes up in our data.

Student: We will look at it a SNP and in a now what are we suppose to do like you have to just compare

Student: These this should be s right or not?

So, you just want to look at what certainly I think before what does nucleotide have change to you.

Student: Ok.

And then change the amino acid.

Student: All right.

To fix that and then I just that would be the new cup time, you would put in your database.

(Refer Slide Time: 19:03)

Points to Ponder

- Visualization of SNPs and its affect on all the possible frames of translation of the gene.
- Mapping SNPs of a gene to the reference genome and preparation of variant peptides.
- Novel variants can be found using IGV where introns may be the contributor to the novelty.



MOOC-NPTEL

IIT Bombay

I hope today you have learned how to see SNPs in a gene with respect to the reference genome of human and one could also look into their data using the reference genome of their target organisms. SNPs in the genomic viewer enables us to look in all three frames of translation and possible effect of the SNPs on the translation. We also saw how one can find the truncated proteins and splice junctions, we also learnt how to look at the junctions which may include exons and part of introns to form the variant peptides due to their SNPs. I hope this hands on session was useful and now you will start using this for different applications.

Thank you.