Introduction to Proteogenomics

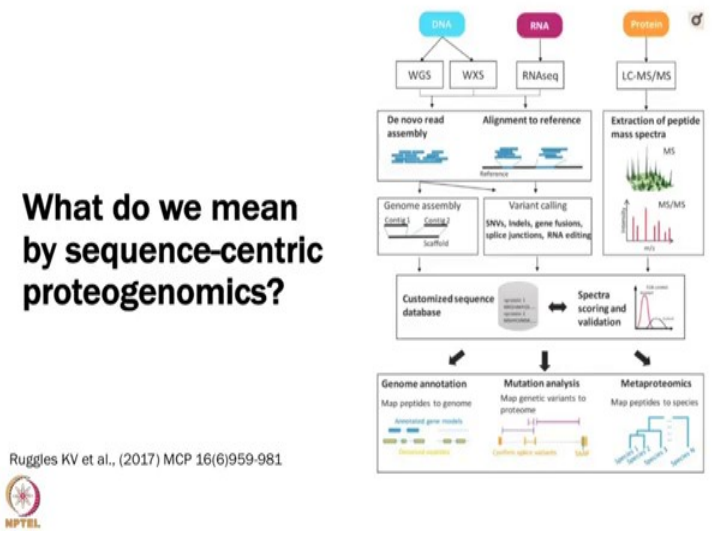
Dr. Sanjeeva Srivastava
Dr. Kelly Ruggles
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay
New York University

Lecture – 43
Sequence centric in proteogenomics

Welcome to MOOC course on Introduction to Proteogenomics. After getting a glimpse of proteogenomic concepts we looked at David Fenyo, we will now move on to the next step of understanding the sequence centric proteogenomics by Dr. Kelly Ruggles. Dr. Kelly will talk about the basic workflow and requirements of Sequence centric proteogenomics.

She will also talk about the reference databases, like RefSeq, UniProt and ensemble. She will also talk about the gene annotation, and if genomic data capable of facilitating the search of novel peptide or identification of functional proteins. So, let us welcome Dr. Kelly Ruggles to know the answers of some of these interesting important questions, and also to understand the sequence centric proteogenomics approaches.

(Refer Slide Time: 01:29)



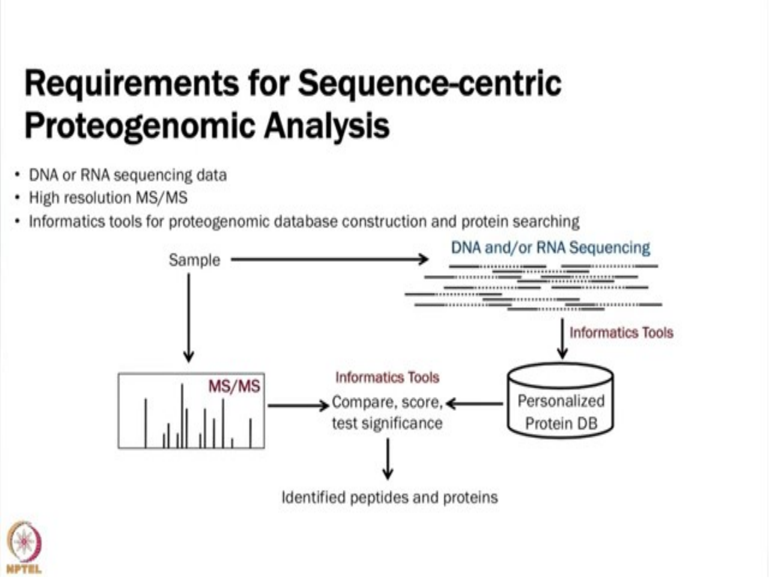
What do we mean by sequence-centric proteogenomics?

Ruggles KV et al., (2017) MCP 16(6)959-981

We are going to talk now about sequence centric proteogenomics and so what, but what do we mean by this right. So, what does this mean? This just means we are really focusing on the seq the sequencing data in terms of what information we get from it like the SNVs, the indels, the fusions, the splice junctions which is what you know I discussed yesterday and trying to combine that use this information you get more information out of our proteomics data.

So, whether that means, genome annotation which I will talk about in detail or looking at the actual mutation analysis specifically in tumor samples. And you can also use this for Meta proteomics, but this is an outside of the scope. So, this is something that we do work on, but it is in the micro biome.

(Refer Slide Time: 02:08)



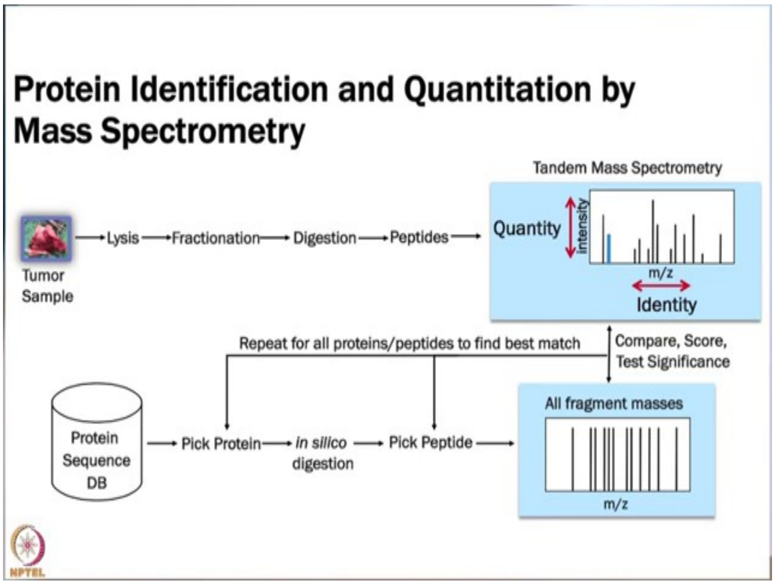
Requirements for Sequence-centric Proteogenomic Analysis

- DNA or RNA sequencing data
- High resolution MS/MS
- Informatics tools for proteogenomic database construction and protein searching

Sample
DNA and/or RNA Sequencing
Informatics Tools
MS/MS
Informatics Tools
Compare, score, test significance
Personalized Protein DB
Identified peptides and proteins

So, there are a couple of requirements to this you of course, need DNA or RNA sequencing if you have both, you can use both in different ways. Some sort of high resolution mass spec data, and then the actual tools to combine these and we will talk about some of those later on in the session.

(Refer Slide Time: 02:27)

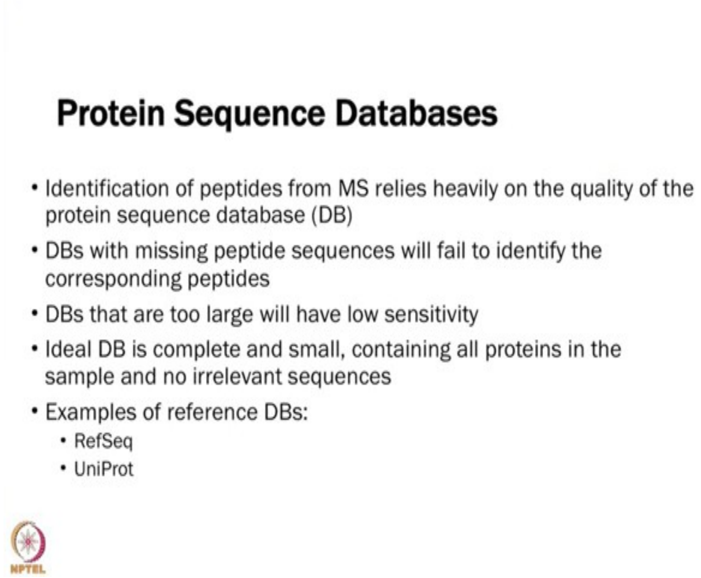


Protein Identification and Quantitation by Mass Spectrometry

And just as a review I know you have heard this a bunch of times, but I want to really focus in on the importance of the protein sequence database. So, I want to touch on it a little couple of slides of about that. So, when we are doing for a protein identification and quantification by mass spec right, we have our sample fractionation digestion you have peptides, you have run them on the mass spec. And then in order to actually identify them you need a protein sequence database or you could be like Karl and do it by hand, but let us assume that we do not we want to have the protein sequence database.

So, from our database, so this is something like RefSeq or UniProt, you are the algorithm will pick a protein, do an in silico digestion, pick a peptide and then have the fragment masses here and do a comparison test and test for the significance and we will continue to do this, over and over again. But of course, if your peptide or your proteins not in this database, you are never going to find it.

(Refer Slide Time: 03:27)



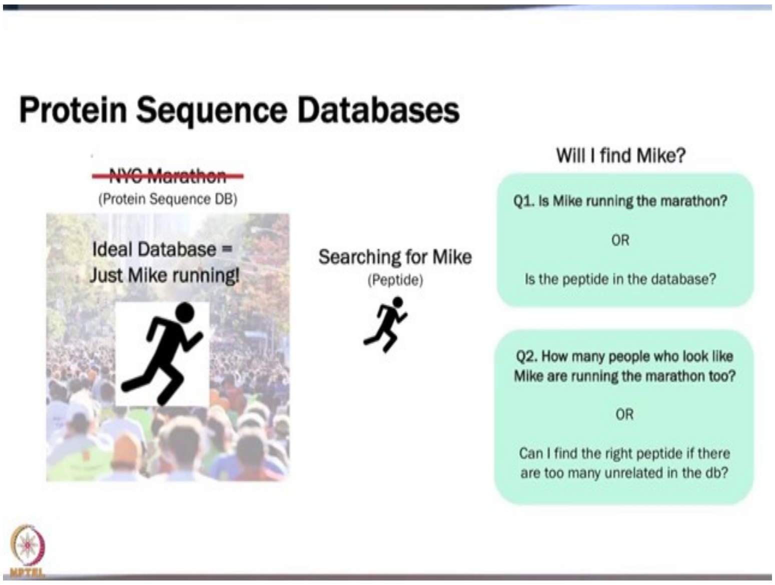
Protein Sequence Databases

- Identification of peptides from MS relies heavily on the quality of the protein sequence database (DB)
- DBs with missing peptide sequences will fail to identify the corresponding peptides
- DBs that are too large will have low sensitivity
- Ideal DB is complete and small, containing all proteins in the sample and no irrelevant sequences
- Examples of reference DBs:
  - RefSeq
  - UniProt

So, there this is a very important thing that sequencing can help us really make sure we have the right sequences in our database. And so databases with these missing, peptide sequences will fail to be identified and if we make our database too big, you know sometimes people will say well why do not just put everything you could possibly put in it, then we are going to lose sensitivity. So, we do not want to do that either.

So, really we want to make sure the database is small, but complete. So, ideally it would contain all of the proteins that you expect to see in the sample, but nothing else obviously that is not; but usual you are not going to happen, but you want to get as close as you can to that right. So, the example of reference databases I already mentioned or RefSeq and UniProt Ensembl and there are more.

(Refer Slide Time: 04:16)



Protein Sequence Databases

~~NYC Marathon~~

(Protein Sequence DB)

Ideal Database = Just Mike running!

Searching for Mike

(Peptide)

Will I find Mike?

Q1. Is Mike running the marathon?

OR

Is the peptide in the database?

Q2. How many people who look like Mike are running the marathon too?

OR

Can I find the right peptide if there are too many unrelated in the db?

But so I do not know how many of you know about the New York City marathon, I was just watching it recently, I am so I wanted to make a comparison between this in the marathon. So, this is what the marathon looks like, there is 50000 people running in the streets and I was looking for one person, a friend of mine who really wanted me to see him, this is very hard to do.

So, I am searching for Mike. So, he is the peptide the marathons or database, so will I find Mike. So, one is he running the marathon that is important if he is not that I am wasting my time right, so is he in the database is the peptide in the database at all. To having people who work exactly like Mike or very close to Mike are also running in the marathon, the answer to that was a lot; it was very hard for me to find.

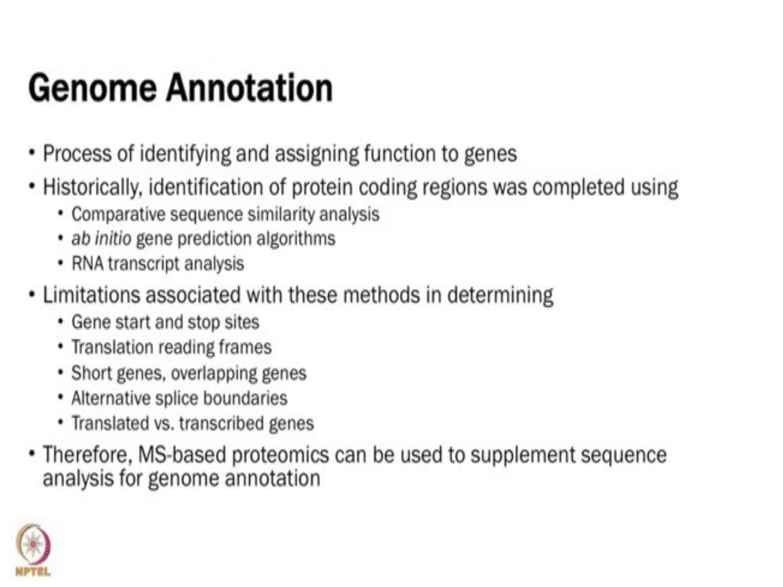
But so if you add more and more people right, can I find the right peptide if there are too many unrelated ones in the database. So, the perfect ideal database here would be just mike running through the streets of New York, I would obviously find him. So, I this just happened to me. So, what I thought it was a nice example of why we have to make our databases really the best that we can to find that the peptides of interest

(Refer Slide Time: 05:33)

Genome Annotation

So, the first example I am going to talk about is genome annotation. So, using sequencing data for genome annotation this is not cancer specific, but this is just another use that I think a lot of people in the room may end up using in their work, so I wanted to mention it.

(Refer Slide Time: 05:51)



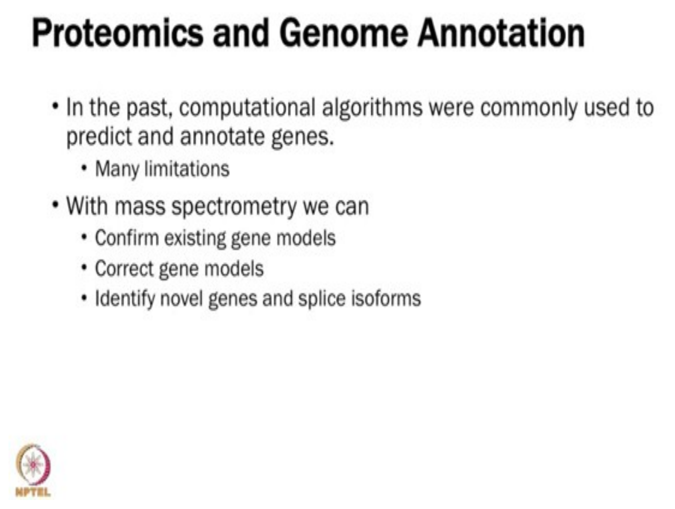
Genome Annotation

- Process of identifying and assigning function to genes
- Historically, identification of protein coding regions was completed using
  - Comparative sequence similarity analysis
  - *ab initio* gene prediction algorithms
  - RNA transcript analysis
- Limitations associated with these methods in determining
  - Gene start and stop sites
  - Translation reading frames
  - Short genes, overlapping genes
  - Alternative splice boundaries
  - Translated vs. transcribed genes
- Therefore, MS-based proteomics can be used to supplement sequence analysis for genome annotation

So, what is genome annotation? It is the process of identifying and assigning functions to genes. So, humans the genome has been fairly well annotated the mouse genome, just awful all of our favorite model organisms, but there is tons of organisms out there right that have not been annotated. So, if you are working on one of those this may be a really good way of trying to help further annotate your genome. So, historically there is been some there is been software and models that have been developed to try and predict genes from genomes.

And those are ok, but they are not perfect and we will talk about how some of them can fail and a little bit of detail I am not going to go into too much detail. And also RNA transcripts, transcription analysis is of course really important here, but to really understand what is making it to the protein level you have to do proteomics. So, for really good genome annotation proteomics and Proteo-genomics is really important. So, really proteomics has been used in several instances to supplement our sequence analysis to really get the best genome annotation we can.

(Refer Slide Time: 07:03)

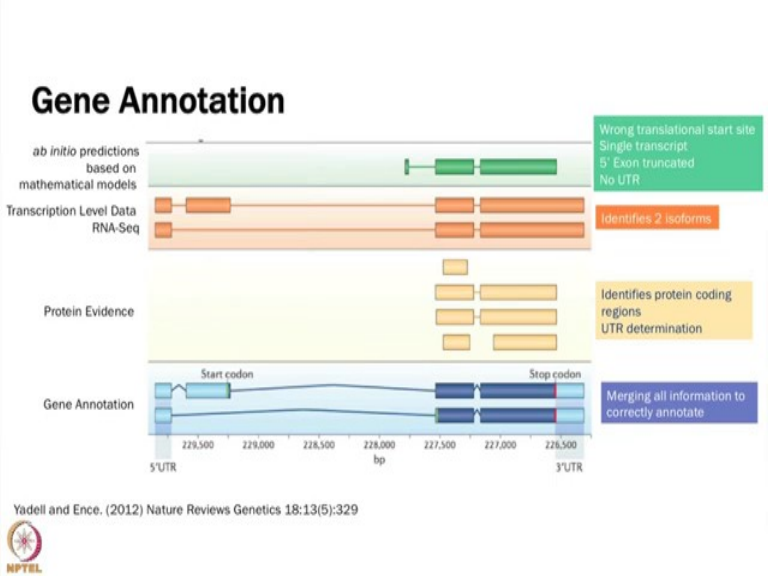


Proteomics and Genome Annotation

- In the past, computational algorithms were commonly used to predict and annotate genes.
  - Many limitations
- With mass spectrometry we can
  - Confirm existing gene models
  - Correct gene models
  - Identify novel genes and splice isoforms

So, we can use mass spec data to confirm gene models to correct gene models, and to also identify novel genes as like isoforms.

(Refer Slide Time: 07:14)



Gene Annotation

Yadell and Ence. (2012) Nature Reviews Genetics 18:13(5):329

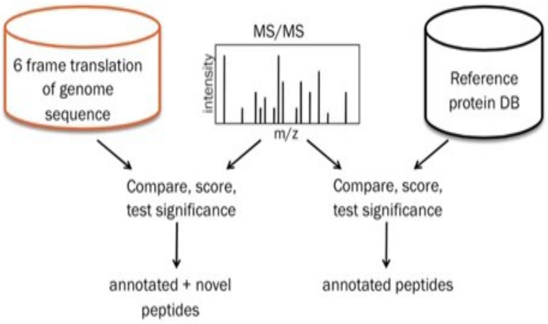
So, here is an example and I am sorry you cannot read this, I will or you can hear ok, ok. So, so the green here would be our predictive models, where they takes the sequence and it predicts what actually is an exon essentially. You can see so here is the actual annotation, so you can see that it is missing a whole lot of things right. So, its does not have the right transcriptional start site, it only has one transcript; it does not have this the 5 prime exon, it does not have the UTR, so its missing a lot of information.

Now, if we add an RNA-seq data, we end up finding at least that there are two isoforms here. But then when we add in proteomics we figure out what is actually making it to the protein, right. So, then we can find out where the start codon is we can better understand, where the UTRs are, so merging all this information together is really necessary to correctly annotate genomes.

(Refer Slide Time: 08:20)

Genome Sequence-based DB for Genome Annotation

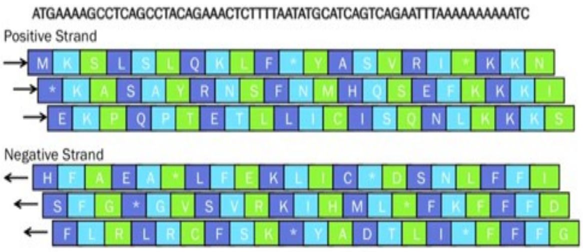
Commonly used method is to search MS against 6 frame translation, removing bias based on established annotation


6 frame translation of genome sequence
MS/MS
intensity
m/z
Reference protein DB
Compare, score, test significance
Compare, score, test significance
annotated + novel peptides
annotated peptides

So, and how do we do this. So, we have our reference database or whatever we have available. So, if there is a lot of databases for understudied organisms that kind of exists, but they are incomplete. So, you can use that and then if you do a sequence, if you could do a whole genome sequence of whatever organism you want, and then you just do a 6 frame translations we talked about what that is and add that in. And then you confirm that find new peptides in that will supplement what we already know about the organism annotation, right.

(Refer Slide Time: 08:56)

Creating 6-frame translation database



ATGAAAAGCCTCAGCCTACAGAAACTCTTTTAATATGCATCAGTCAGAATTTAAAAAAAAAAATC

Positive Strand

→ M K S L S L Q K L F * Y A S V R I * K K N

→ * K A S A Y R N S F N M H Q S E F K K K I

→ E K P Q P T E T L L I C I S Q N L K K K S

Negative Strand

← H F A E A * L F E K L I C * D S N L F F I

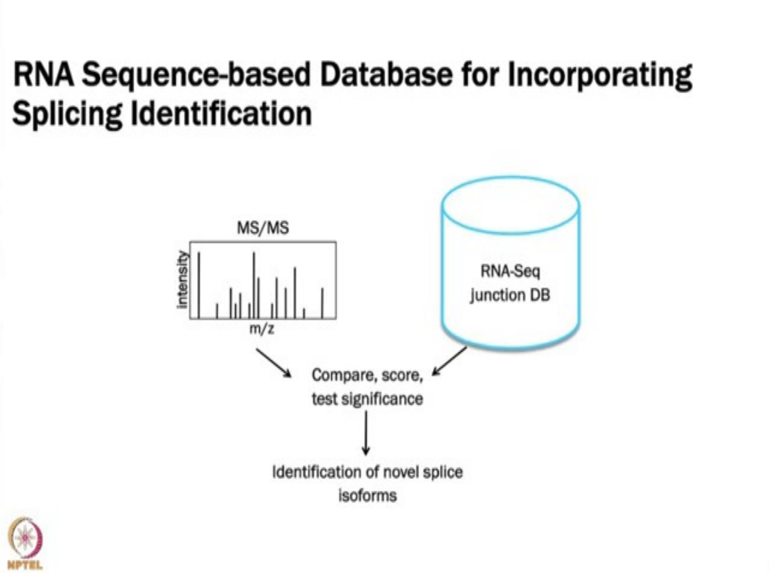
← S F G * G V S V R K I H M L * F K F F F D

← F L R L R C F S K * Y A D T L I * F F F G

So, here is what so, we have our sequence. So, you do a positive strands three frame translations. So, you know you start at ATG and you go from there GTGA and you go from there. So, you do every, every frame, and then you go in the negative direction and you get the other three frames. So, then you have six frames, and then you can use this to supplement your reference database.

This is if the only the bet the best way to go if you do not have RNA seq data, and you know because you are really you are blowing up your database, you are making it enormous if you do this. So, it is not necessarily the best thing to do if you have other data available that we will talk about. But it is one option especially when you are working with like an understudied organism and you only have whole genome sequencing.

(Refer Slide Time: 09:47)



RNA Sequence-based Database for Incorporating Splicing Identification

MS/MS

Intensity

m/z

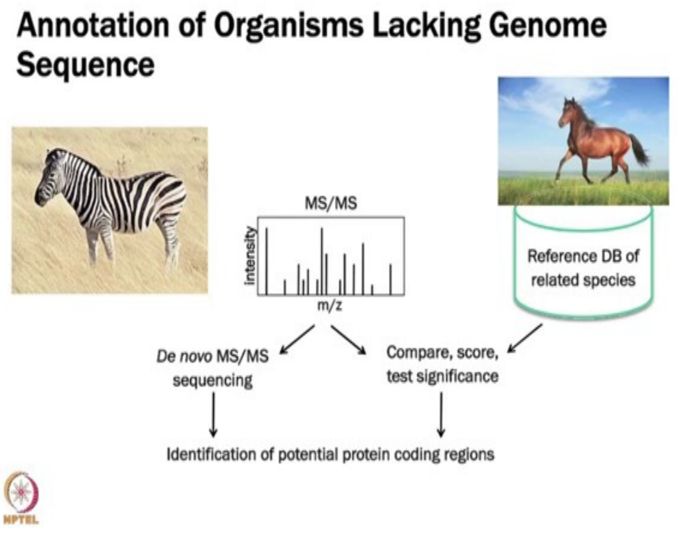
RNA-Seq junction DB

Compare, score, test significance

Identification of novel splice isoforms

So, if you have RNA seq data as I mentioned then you can add in splicing information which will be even more it will help you annotate even further.

(Refer Slide Time: 09:56)



Annotation of Organisms Lacking Genome Sequence

MS/MS

intensity

m/z

Reference DB of related species

De novo MS/MS sequencing

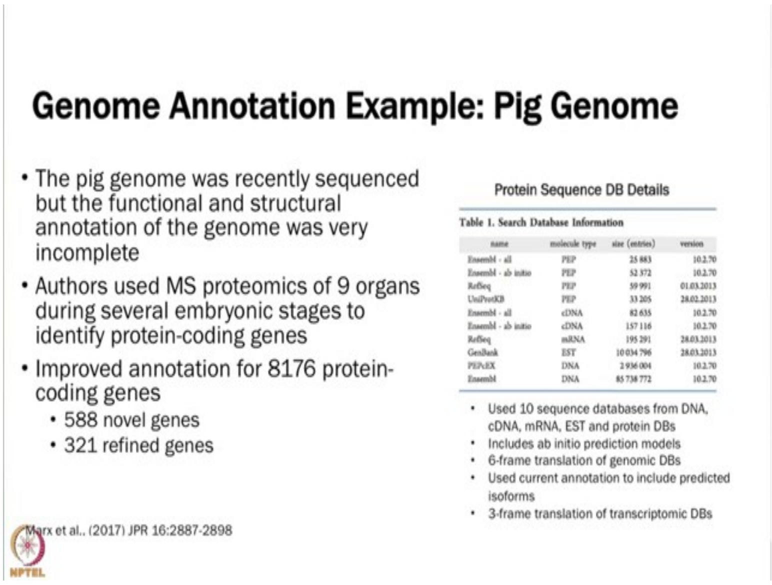
Compare, score, test significance

Identification of potential protein coding regions

And if you let us say you want to study an organism that has no genome sequence. So, zebras do not have there is their sequence genome, genome sequenced I have checked last night. There is a whole bunch of organisms that do not, but let us say you want to study zebras and you are interested in this.

But you do not have anything about it well you could you could use a horse its close enough, and you can try and see what you find using the horse sequence to see if you are able to find interesting related proteins, and then you can do some de novo sequencing as well to try and supplement this. So, this is an option is anyone studying zebras.

(Refer Slide Time: 10:41)



Genome Annotation Example: Pig Genome

- The pig genome was recently sequenced but the functional and structural annotation of the genome was very incomplete
- Authors used MS proteomics of 9 organs during several embryonic stages to identify protein-coding genes
- Improved annotation for 8176 protein-coding genes
  - 588 novel genes
  - 321 refined genes

Protein Sequence DB Details

Table 1. Search Database Information

name	molecule type	size (entries)	version
Ensembl - all	PEP	25 883	10.2.70
Ensembl - ab initio	PEP	52 372	10.2.70
RefSeq	PEP	59 991	01.03.2013
UniProtKB	PEP	33 205	28.02.2013
Ensembl - all	cDNA	82 635	10.2.70
Ensembl - ab initio	cDNA	157 116	10.2.70
RefSeq	mRNA	195 291	28.03.2013
GenBank	EST	10 034 796	28.03.2013
PEPxEX	DNA	2 936 004	10.2.70
Ensembl	DNA	85 738 772	10.2.70

- Used 10 sequence databases from DNA, cDNA, mRNA, EST and protein DBs
- Includes ab initio prediction models
- 6-frame translation of genomic DBs
- Used current annotation to include predicted isoforms
- 3-frame translation of transcriptomic DBs

Marx et al., (2017) JPR 16:2887-2898

So, one example a recent example of this, this type of method was in the pig genome. So, in 2017, there is the paper that came out you can find it I mean if you are interested, you can go look at it where the pig genome had been recently sequenced, but the actual annotation of the genome had was not complete. So, they use mass spec, they did mass spec on 9 organs during different stages of development. And then they were able to improve the annotation for over 8000 protein-coding genes. So, they think this is like the perfect example of how you can use these two things together to really better annotate genomes.

And there is a list of their databases. So, they used all sorts of sequence databases, they used prediction models, they use 6 frame translations of the genome, they use transcriptome data. So, they kind of did all of the things we already talked about. So, I think it is a really nice example of how you can use this data for annotation. Any questions about annotation, I am going to move into, go ahead yeah.

Student: So, I am talking on model organism other than man.

Yeah.

Student: So, for which I do not have the all genome sequence, but I have sequence called relative spaces. So, I am identifying protein. So, how this will be useful that whatever genome informational with the from relative spaces and then using for my target species.

See you have similar situation to zebra

Student: Yeah.

And you are so

Student: I am working on fish which is closely related to Zibra fish, Danio rerio species, but Danio whole genome sequence is available, but where in case of this species which is closer related whole direct genome sequence is not available. So, I am trying to identify the protein using relative database, then again I am going for suppose proteogenomics.

Yeah.

Student: How this will be heading?

So, I mean can you do sequencing on your species that would be the best thing to do right. If you cannot yeah I have you tried to use a related species database.

Student: Yeah that is what I am doing, I am doing that and how it would be valid how much validity it will be there.

Depends on how related they are. So, we could talk about it offline. This is interesting because we could chat about the this exact question, but it really depends right I know David Fenyo was working on rat and mouse work that is was similar to this. So, he may also be a good person to talk to you about it.

Student: Excuse me Ma'am.

Yeah.

Student: We are actually he is talking about my project.

Student: Yes and I am working on *Labeo rohita* fish and the database that we are using for this is from the another closely related speices Danio rerio,

Yeah.

Student: that is most commonly known as Zebra fish.

Yeah.

Student: So, they are how closely they are related is that they belong to same family that is why we chosen we have chosen this, and most of the information available is for this related fish only. So, he is asking that.

Ok.

Student: How much means validate, validate, how it is validated or not if we use proteogenomics using the same genomic data using the other like zebrafish like still now you are identifying protein using its relative genomic data.

It is not perfect right like you are going to be missing a lot of information, but I think it is worth trying and I do not know how similar your species are right. So, it is hard for me to say, but I think it is worth it sounds like you are already going to do it. So, you might as well try, you can let me know how it works and then we can talk about it.

Student: There, but still the like appropriate database like this, this is the early option that we can study this using the related organism

Yeah, yeah, I think it is vary case by case right. So, yeah.

Student: Conservation scores on the protein coding region.

Conservation scores on the protein coding region that we said what about it?

Student: Yes

What about it?

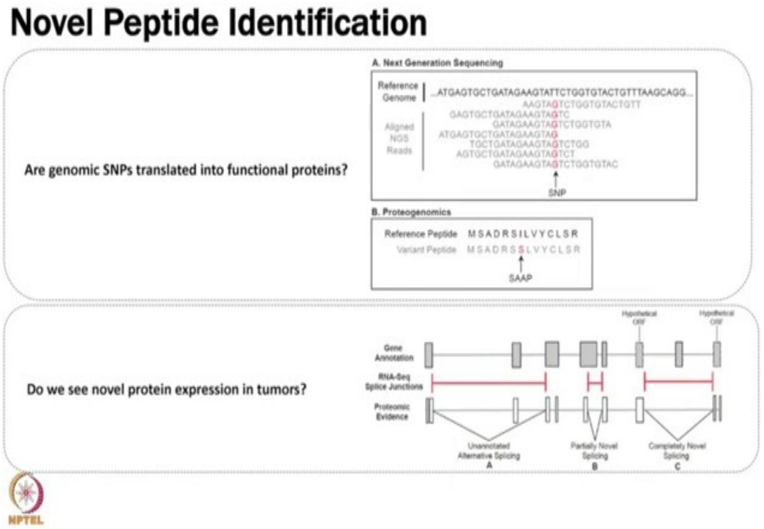
Student: Reduce the reject the like protein coding if the.

Yeah. So, you can you can predict what would be protein coding.

Student: Isnt it right?

Yes correct, yeah, yes, correct is that I do not know what the question was, but I agree with what you said ok. Other questions move on ok. So, I am going to spend most of the time talking about variant identification so variant peptide identification.

(Refer Slide Time: 15:31)



Novel Peptide Identification

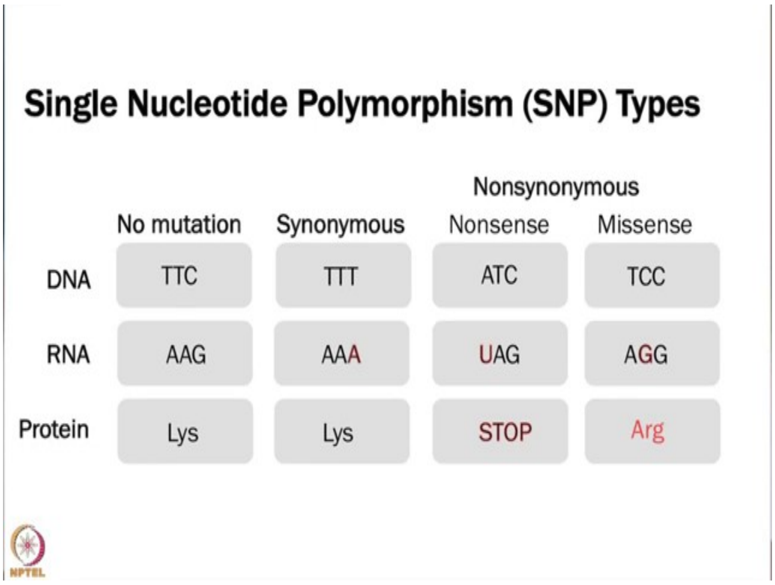
Are genomic SNPs translated into functional proteins?

Do we see novel protein expression in tumors?

So, what do we mean by this the novel peptide identification? So, what the questions we want to answer here are whether or not genomics SNPs are translated into functional proteins. So, when we have a single nucleotide polymorphism does it make it to the protein, and if it does do we see it at the protein level. And also do we see novel protein expression. I am going to be specifically really looking at tumors in this case, but you could apply this to whatever you are you know whatever you are looking at.

It is just typically in tumors there is there you see more of this novel expression. So, it is something that we pay a lot of attention to it. So, here this would be looking at different RNA seq splice junctions which we talked about yesterday so things, and we will talk about all of the different sort of combinations of splice junctions that are novel, and how we deal with them in terms of proteomics.

(Refer Slide Time: 16:24)

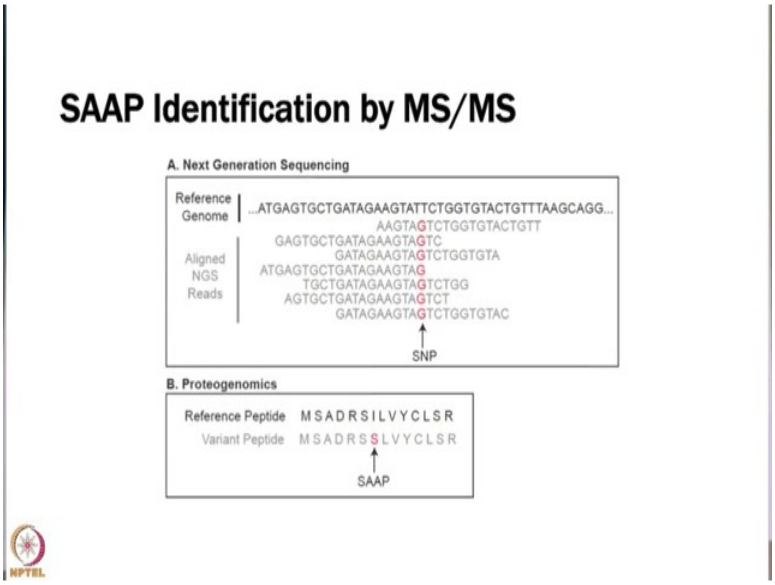


Single Nucleotide Polymorphism (SNP) Types

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

There are a couple of different kinds of SNPs. So, if we have these are our codons, if we have no mutation we get a lysine; if we have a synonymous SNP, where there is g to a, but it does not actually change the amino acid. We can have non-synonymous SNPs, where it turns into a stop codon. So, it will just stop the protein synthesis early or we can have a missense mutation where we actually get a complete change in the in the amino acid. And we will talk about these in more detail.

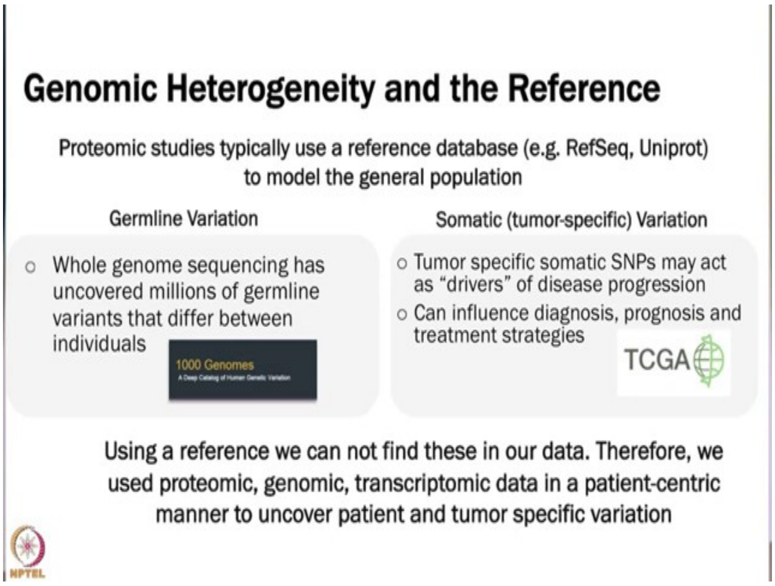
(Refer Slide Time: 16:59)



SAAP Identification by MS/MS

So, we are really going to focus on how do we how do we put these peptides, how do we get them into our database so we can even find them.

(Refer Slide Time: 17:06)



Genomic Heterogeneity and the Reference

Proteomic studies typically use a reference database (e.g. RefSeq, Uniprot) to model the general population

Germline Variation

- Whole genome sequencing has uncovered millions of germline variants that differ between individuals

1000 Genomes
A Deep Catalog of Human Genetic Variation

Somatic (tumor-specific) Variation

- Tumor specific somatic SNPs may act as "drivers" of disease progression
- Can influence diagnosis, prognosis and treatment strategies

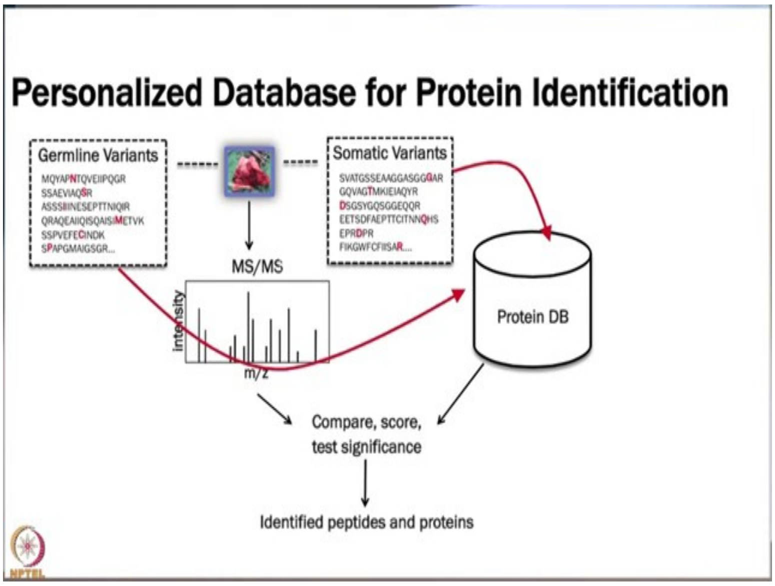
Using a reference we can not find these in our data. Therefore, we used proteomic, genomic, transcriptomic data in a patient-centric manner to uncover patient and tumor specific variation

So, and one of the reasons this is really important is because there have been several studies that well. So, most proteomic studies that had been done especially previously we have gotten better about this is that they usually use a reference database. So, either RefSeq or Uniprot to model whomever, but as we talked about yesterday a reference database is just trying to represent the population, but it does not have all of the different in variation that occurs in a population.

So, and there was this thousand genomes project which we also talked about yesterday that really uncovered how much variation there is person to person, and they are not necessarily disease causing SNPs they are just SNPs we just exist. So, if we model everyone using a reference database, we are going to miss a lot of information. And also in cancer there are somatic mutations that occur, so mutations that are just occurring in the tumor.

So, if you are trying to measure proteins in a tumor and you do not include these somatic SNPs, you may actually miss those peptides. And, these are really-really interesting because many of them act as they are very much involved in disease progression. So, if we use a reference we cannot find these SNPs in our data. So, we really need to figure out how to make sure we include them in our data, so that we can find these and to uncover both patient if it is if we are looking at cancer in tumor specific variation.

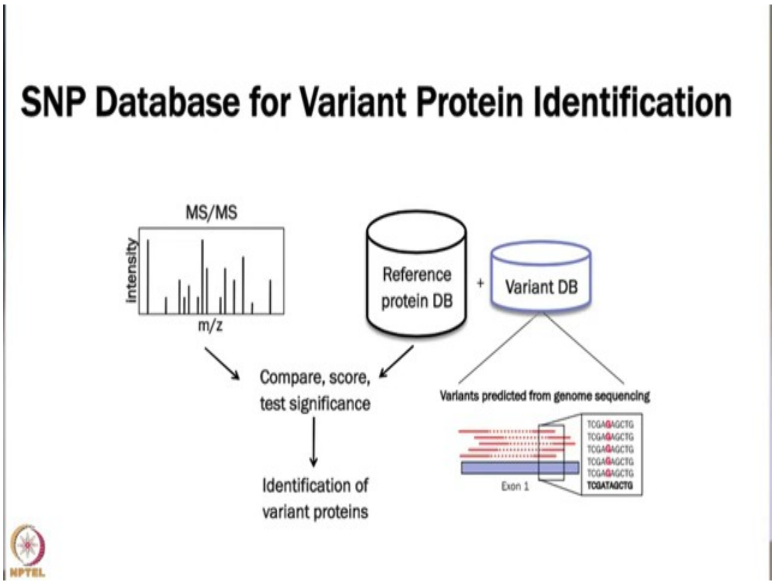
(Refer Slide Time: 18:47)



Personalized Database for Protein Identification

So, for example, if we have our mass spec flow, if we have germline mutations, so these are just mutations occurring in people; and somatic mutations occurring in the tumor itself. We have to figure out how to get these into the database, so that we can actually find them.

(Refer Slide Time: 19:04)

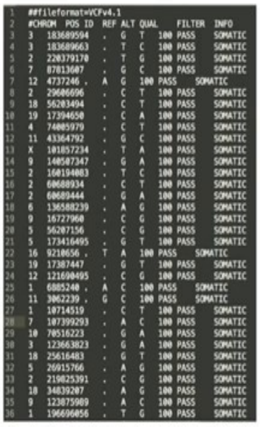


SNP Database for Variant Protein Identification

So, this is just a representation of the same thing.

(Refer Slide Time: 19:10)

VCF File Format



```
##fileformat=VCFv4.1
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
3   183689594   .   G   T   100 PASS    SOMATIC
3   183689663   .   T   C   100 PASS    SOMATIC
2   220379170   .   T   G   100 PASS    SOMATIC
7   87013607    .   G   C   100 PASS    SOMATIC
12  4737246 .   A   G   100 PASS    SOMATIC
2   29606696    .   C   T   100 PASS    SOMATIC
18  56203494    .   C   T   100 PASS    SOMATIC
19  17394650    .   C   A   100 PASS    SOMATIC
4   74005979    .   C   T   100 PASS    SOMATIC
11  43364792    .   G   C   100 PASS    SOMATIC
X   101857234   .   T   A   100 PASS    SOMATIC
9   140507347   .   G   A   100 PASS    SOMATIC
2   160194083   .   T   C   100 PASS    SOMATIC
2   60688934    .   C   T   100 PASS    SOMATIC
2   60689444    .   G   A   100 PASS    SOMATIC
6   136588239   .   A   G   100 PASS    SOMATIC
9   16727960    .   C   G   100 PASS    SOMATIC
5   56207156    .   C   G   100 PASS    SOMATIC
5   173416495   .   G   T   100 PASS    SOMATIC
16  9210656 .   T   A   100 PASS    SOMATIC
19  17387447    .   G   T   100 PASS    SOMATIC
12  121690495   .   C   G   100 PASS    SOMATIC
1   6885240 .   A   C   100 PASS    SOMATIC
11  3062239 .   G   C   100 PASS    SOMATIC
1   10714519    .   C   T   100 PASS    SOMATIC
7   107399293   .   A   C   100 PASS    SOMATIC
10  70516223    .   G   A   100 PASS    SOMATIC
3   123663823   .   G   A   100 PASS    SOMATIC
18  25616483    .   G   T   100 PASS    SOMATIC
5   26915766    .   A   G   100 PASS    SOMATIC
2   219025391   .   C   G   100 PASS    SOMATIC
10  34839207    .   A   G   100 PASS    SOMATIC
9   123075909   .   A   C   100 PASS    SOMATIC
1   196696056   .   T   G   100 PASS    SOMATIC
```

VCF File Format

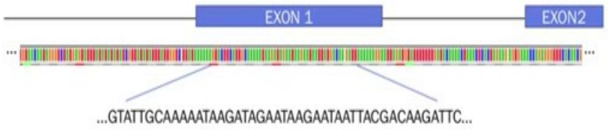
# Meta-information lines

Columns:

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

So, then there is the VCF file format is the most commonly used format for looking at these variants in I uploaded these are the columns again, we have the chromosome the position of the SNP an identifier. So, sometimes there is nothing not really anything there it is just the dot. The reference I mean a nucleotide and then the new nucleotide, and then a quality score and then some other information about the SNP itself.

Creating Variant Sequence DB



(Refer Slide Time: 19:47)

Student: Many times when you look at variant calling

Yeah.

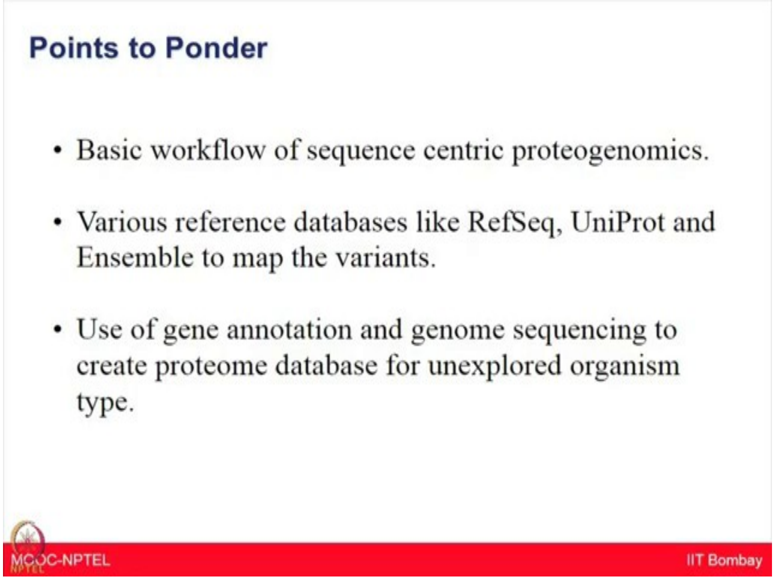
Student: Mentioned as true variant etcetera, but when you actually do PCR verification you will find that actually it is a false.

Yeah.

Student: So, what is the criteria that which we can actually pick out true variants for sequencer.

I mean the best way is to do PCR verification. So, there are a lot of ways of validating it, one of them is that way. So, it just depends on the study, how much work a person’s willing to do to validate. If you have hundreds and hundreds of them, you cannot do that, so that is why I mentioned yesterday that there are several SNP collars. So, right now a lot what a lot of people do is they will use a whole bunch of them, and then look at the overlap, and then trust the overlap versus just using one, because you are going to have a lot of false positives, and that seems to work fairly well.

(Refer Slide Time: 20:43)



Points to Ponder

- Basic workflow of sequence centric proteogenomics.
- Various reference databases like RefSeq, UniProt and Ensemble to map the variants.
- Use of gene annotation and genome sequencing to create proteome database for unexplored organism type.

MOOC-NPTEL IIT Bombay

In conclusion, I hope today you have learned how one can use gene annotation and genome sequencing to create the proteome databases for unexplored organism type. I would like to emphasize, it is very crucial to learn this information because many time you are working on the unknown organisms for which databases are not available. And therefore, your searches are going to revert back with unknown or hypothetical proteins.

So, refining databases is very crucial especially if you are not working on the human and other model systems. So, you may have to first try to establish good databases for doing the search for proteomics data. You also learned why it is better to know your targets while searching for a SNPs - Single Nucleotide Polymorphism as you do not want to sequence non-pathogenic SNPs in this process. We also heard about how one could make the personalized protein databases for specific studies.

The next lecture is about variant analysis and their effect on RNA and protein expression in clinical conditions, lecture will be continued by Dr. Kelly Ruggles.

Thank you.