

Introduction to Proteogenomics

Dr. Sanjeeva Srivastava

Ms. Shalini Agarwal

**Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay**

Supplementary - 08

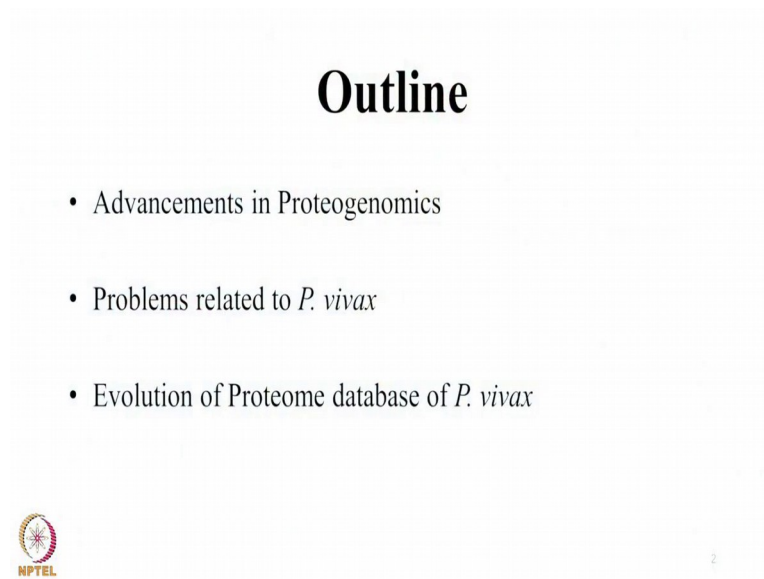
Topics in Proteogenomics: Malaria Case Study

Welcome to MOOC course on Introduction to Proteogenomics. You have studied about genomics proteogenomics and various tools involved in proteogenomics. We thought to give you some case studies and examples; how proteogenomics can be used for some applications. In this light today in the proteogenomics case study, we will have a TA of this course Ms. Shalini Aggarwal, who is a research scholar at IIT, Bombay. She will talk to you about how proteogenomics analysis have been applied for some clinical applications.

She will talk to you about malaria and which way the proteogenomics research have made contribution to understand the parasite Plasmodium vivax which is less studied less known parasite. But it is now causing much more issues in India and many parts of the world. She will talk about how proteogenomics can help in better understanding of the parasite, as proteogenomics is now emerging as a broad tool to solve various problems and especially in the clinical field, it is started showing its applications. So, let us welcome Shalini to give her lecture about proteogenomics case studies on malaria.


Hi all. So, my name is Shalini Aggarwal and I am a research scholar at IIT, Bombay. In today's lecture, I am going to tell you about proteogenomics in infectious diseases such as malaria.

(Refer Slide Time: 02:07)



Outline

- Advancements in Proteogenomics
- Problems related to *P. vivax*
- Evolution of Proteome database of *P. vivax*

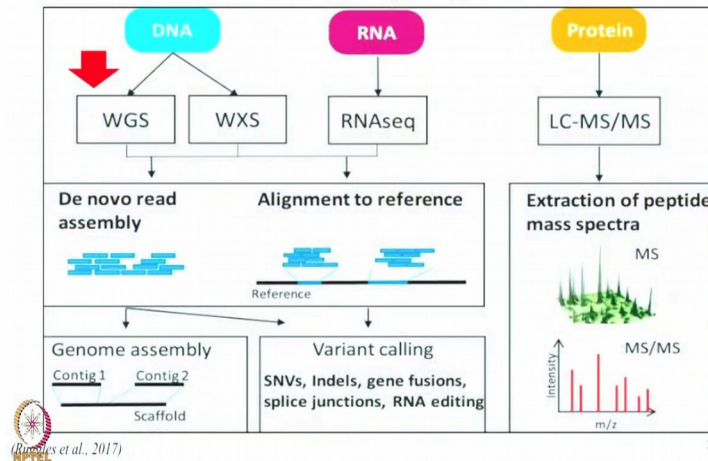
 2

So, in this lecture, we are going to cover three things; one is advancements in proteogenomics, problems related to Plasmodium vivax and then, evolution of proteome database of Plasmodium vivax using proteogenomics approach.

So, for the reference first of all, I would like to explain you about what do you mean by proteogenomics and how one can use it for understanding the biology of an organism better. So, in a paper by Ruggles et al. in 2017, with the title “Advancements in Proteogenomics Methods, Tools and Current Perspectives”, they have explained about the three major part of central dogma that is DNA, RNA and protein.

(Refer Slide Time: 02:49)

Advancements in Proteogenomics – Methods, tools and current perspectives



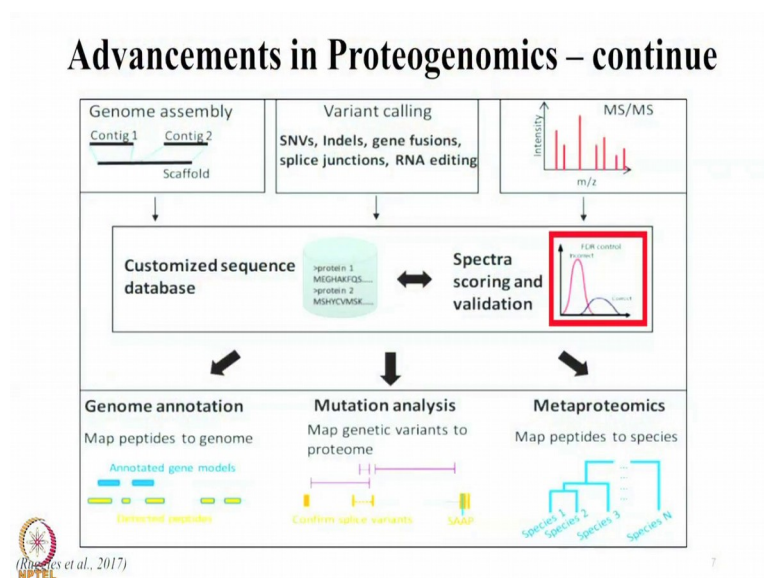
How these three components can be used for further sequencing and analysis of an organism to understand the organism better. So, as you can see in the slide for for the first case DNA. DNA can be analyzed by two different methods; one is WGS and other is WXS. So, WGS stands for Whole Genome Sequencing, where this whole genome sequencing is done when you do not know anything about the organism and you are trying to study that organism for the very first time.

So, you extract the genome of the organism and you go for the whole genome sequencing. So, this helps us in de novo read assembly and it will give a reference genome for all the further studies which one can do in future. Then, comes whole exome sequencing. Here, one can sequence all the exomes specifically by leaving the introns, but this will only be applicable once you have a reference genome sequence which is done by whole genome sequencing. So, whole exome sequence helps in alignment of genes on the reference. Sorry, whole whole exome sequencing will help us in alignment of the genes on the reference genome. Similarly, in the case of RNAseq, one extracts the RNA and go for the RNA sequencing. RNA sequencing adds on the value by telling us the information regarding the introns which are also involved in the production of a protein.

There are many protein, which are not only formed by exomes, but also include some part of introns. So, if we do whole genome sequencing or whole exome sequencing we will miss out those introns which are contributing to the production of a protein. But in the case of

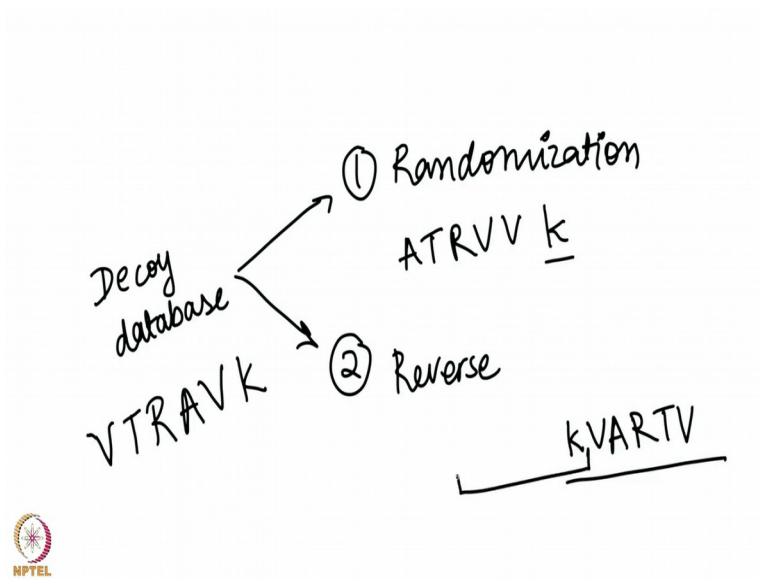
RNAseq, we will be able to identify those introns which will miss out otherwise. So, after aligning the sequence genes, we align them on the reference genome and then we can understand two things; one is genome assembly that how the gene is getting assembled on the reference genome and second thing is variant calling which includes SNVs that is Single Nucleotide Variants, Indels that is insertions and deletions, gene fusions, splice junctions and RNA editing. We will be able to understand all these things if we use RNAseq, WXS or WGS. But after that one more component is very important that is protein; one can analyze the proteome database of an organism and then, subject it to mass spectrometry to extract the possible spectra for all the proteins which are present and processed.

(Refer Slide Time: 05:33)



And then, this protein which we have obtained from mass spectrometry data, we can analyze it using the reference database which we can make by using the information which we have obtained from whole genome, whole exome and RNAseq data. So, here one more component plays a very important role that is FDR control. FDR stands for False Discovery Rate and it helps in removing all the contaminants which can possibly come, which can possibly come from the very first step of sample collection to the last step which is subjection to mass spectrometry. So, for this we need to make a decoy database. Let me show you how we can make decoy database. There are two methods by which one can make decoy database.

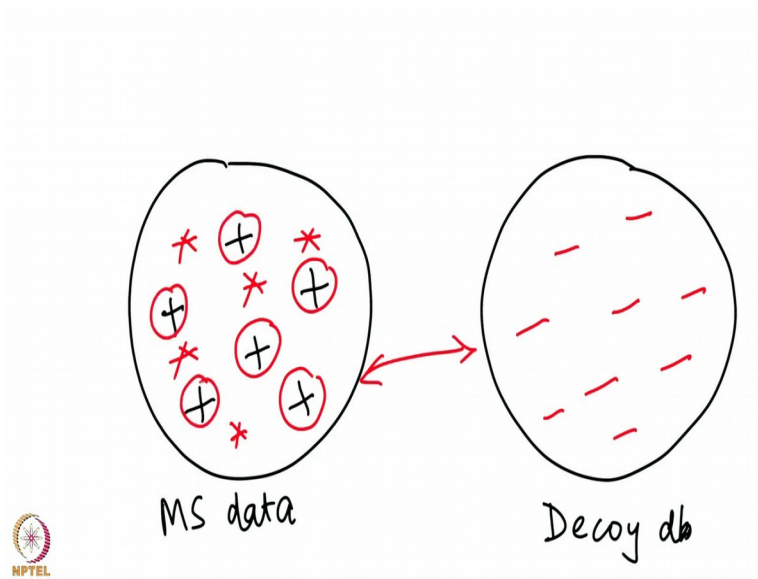
(Refer Slide Time: 06:23)



So, one is randomization and second is reverse. So, let us take an example of a peptide sequence; let us say this is a sequence of a peptide we have and here if we go for randomization, it will randomize all the amino acids except k because mostly we do trypsin digestion that is why we are keeping the k constant. Now, it will randomize all the amino acids by keeping 1 amino acid constant and the next one is reverse. In this one, it reverses the whole sequence of the peptide. So, if this is the peptide, we will not be getting this peptide if we cut it by using trypsin digestion. It will give some peptide which is this way because trypsin is C terminal digesting enzyme.

So, now let us see how this decoy database can be useful in understanding the FDR control.

(Refer Slide Time: 07:51)

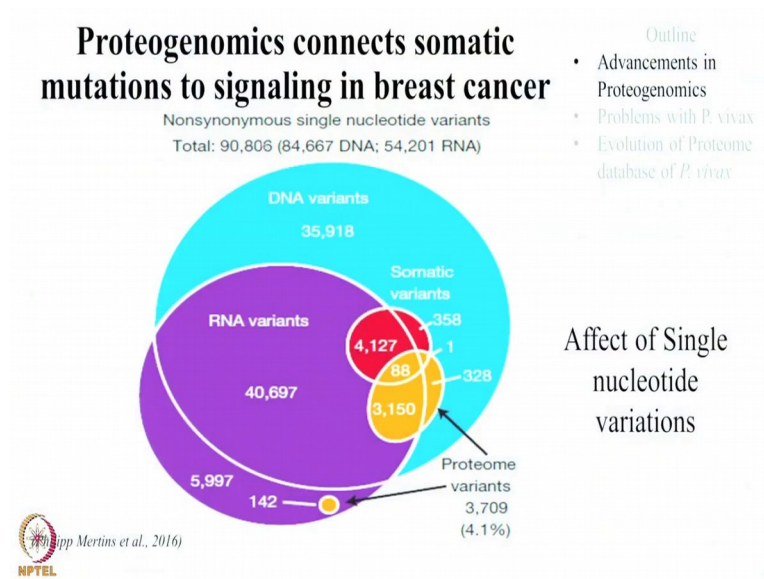


So, let us assume that we have these two circles; where, one represents the mass spec data and the other one represents decoy database and in a mass spec data, let us assume positive stands for the peptides which belong to the sample; whereas, the red ones are the ones which are coming from the contamination while processing the sample or any other source. So, now, what we do we overlap these two circles and then we remove all the negatives which are present in the mass spec data and after removing it, we will only have our peptides which are useful for us in the understanding of the proteome of a particular sample.

So, this is how one can find the FDR control and remove all the contaminants from the sample and then, use the proper sample peptides to analyze the results. So, after this. Now, we know about the gene sequence, RNA sequence and proteome data and then, we can use all these information for various purposes like genome annotation, where we can map the peptides on the genome. Then, for mutation analysis, where it will tell us about all the possible mutation which are leading to a particular clinical condition like malaria, cancer or any other disease and also for meta proteomics. Proteogenomics is vastly studied in the case of cancers and the many people have already taken it to a very high level.

So, I would like to take an example of a study which is done by Philip Martin et al in 2016, but the title of the paper is proteogenomics connects somatic mutation to signaling in breast cancers, where they have disapproved the conventional thinking that DNA mutation leads to RNA mutant and then, the RNA mutant leads to a protein mutant which causes a disease.

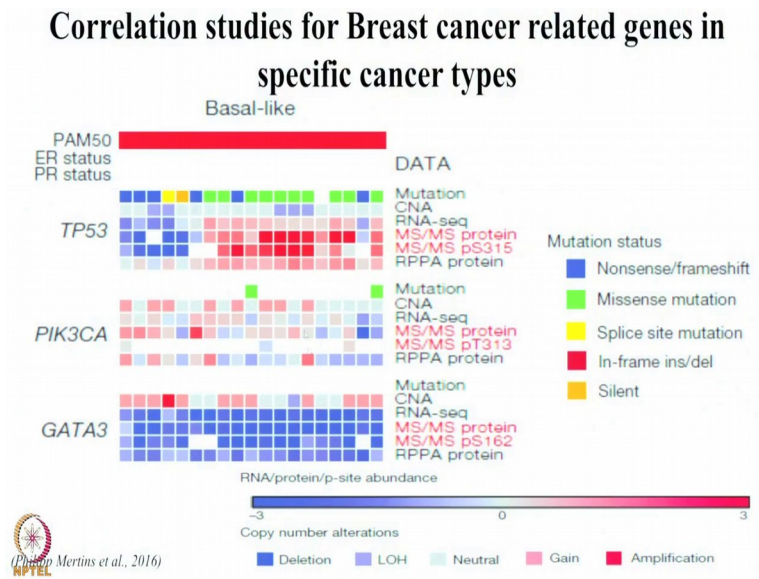
(Refer Slide Time: 10:11)



But in this paper as you can see here that they have taken approximately 90,806 mutations out of which 84,667 were DNA mutations and out of those mutations only 40,697 of gene led to the RNA variants. Whereas, others were only present in DNA variant and they were not further replicated on RNA or protein level. But out of these RNA variants also very few have translated into onto proteome level.

So, this tells us about the variations on different level which one can face that is why, studying only one platform RNA, DNA or proteome will never give you a perfect answer for understanding the biology behind any disease.

(Refer Slide Time: 11:05)



So, here in this paper I have taken one part of their study, where I have taken example of Basal-like cancer. They have taken, they have confirmed the cancer by using PAM50 test and then, they had also checked for ER and PR status, where each box represents a patient and these different level shows different level of analysis, they had done. One is mutation; mutation they had checked for this particular gene that is TP53 and here on the right hand side, you can see the different color which represent different type of mutations.

So, just like green represents mis-sense mutation; blue represents Nonsense mutation or frameshift mutation. CNA stands for Copy Number Aberrations, which in the bottom is represented by different colors that is dark blue represents deletion and dark red represents amplification. So, similarly, RNAseq and for other mass spec data, they have given the scale from minus 3 to 3 which shows further up regulation or down regulation of the protein. So, as you can see here in these 3 patients are showing, in these three patients as you can see that these 3 are having mis-sense mutation which is leading to heterozygosity loss. But still the amount of protein which is produced is up-regulated. So, one cannot tell directly or for sure, by looking at a mutation of on gene or RNA level that this particular protein will up regulate or down regulate. So, hence the study at all the three levels is very important.

So, now I would like to take you towards Plasmodium vivax and why this particular parasite is very harmful or very important to study this parasite. So, Plasmodium vivax is a parasite which causes malaria.


(Refer Slide Time: 13:03)

Malaria

- Caused by Plasmodium species viz. *falciparum*, *vivax*, *ovale*, *knowlesi*, and *malariae*
- *P. falciparum* and *P. vivax* contributes ~90%
- WHO 2017 malaria reports, ~3.5 million cases.
- Vector - cross borders without any inhibitions

(Research gate)

WHO 2017, Reports



And it is having a life cycle, where it includes two hosts; one is mosquito and other one is human. So, malaria can be caused because of five major parasites which are falciparum, vivax, ovale, knowlesi and malariae. Out of these five Plasmodium falciparum and Plasmodium vivax contribute to 90 percent of the malaria cases across the world. And according to WHO 2017 malaria reports, it was reported that approximately 3.5 million cases were found. So, the major issue with the parasite is not the vector that is mosquito can cross the borders without any inhibitions handled or taken care properly. So, problem related to Plasmodium vivax are many; one of those is diagnosis and differentiation.

(Refer Slide Time: 13:57)

Problems related to *P. vivax*


- Diagnosis and Differentiation
 - Selective infection of Reticulocytes
 - Low parasitemia

J Beeson et. Al, 2013

Microscopy

- No distinguishable marker for *P. vivax* (Pan plasmodium)

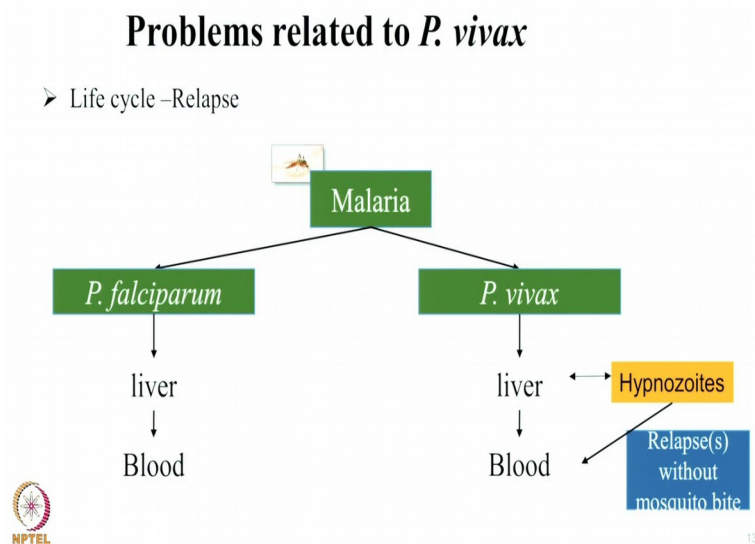
Rapid Diagnostic Test (RDT)



Where, Plasmodium vivax infects only reticulocytes that are immature RBCs. Reticulocyte number in whole blood is very less and in that also if parasite is infecting only reticulocytes, the chances of finding an infected reticulocyte becomes very less. Plus because of which the parasitemia level is almost every time very low in the case of Plasmodium vivax, which makes it very difficult for being diagnosed by using microscopy. Whereas, the other method is RDT that is Rapid Diagnostic Test. So, we have two types of protein and RDT one is PFHRP 2 which is specific for plasmodium falciparum; whereas, the other ones are aldolase and lactose dehydrogenase which tells us about the remaining all parasites except falciparum.

So, RDTs which are existing can only tell us whether a person is suffering from malaria and the parasite which is causing it is falciparum or not. It will never tell us about whether the other, whether the patient is infected with vivax or not.

(Refer Slide Time: 15:17)



So, apart from that it also has a major issue with the life cycle where when the mosquito bites an individual and causes malaria, if it is falciparum; the parasite goes into the liver and then it infects blood. Whereas, in the case of vivax it infects liver and it stays there in the form of hypnozoites. Those are the dormant condition, they are in the dormant condition and it can relapse whenever the conditions are suitable.


So, it can relapse up to 6 months after the infection also without any further mosquito bite because of this, the malaria because of plasmodium vivax is becoming annual rather than being seasonal which was the case earlier. It also has another problem which is because of

discontinuous culturing capability of the parasite because in the case of *Plasmodium falciparum*, one can continuously grow culture of *Plasmodium falciparum* in vitro; whereas, in the case of *vivax*, one cannot take the culture beyond 48 to 72 hours.

(Refer Slide Time: 16:23)

Problems related to *P. vivax*


➤ Lack of understanding about the strain



No continuous culture

Hence, we lack –

1. Proper proteome database
2. Poorly understood dormancy
3. No rapid diagnostic method availability
4. Hemolysis due to G6PD deficiency - primaquine



14

So, we because of this we lack proper proteome database and we poorly understand the dormancy of the parasite because one cannot diagnose the dormant condition parasite in the patient. And we also do not have any rapid diagnostic methodology available for diagnosing the infection, with also G6PD deficiency that is glucose 6 phosphate dehydrogenase deficiency in an individual may lead to severe hemolysis if primaquine is given. Primaquine is the only drug which helps us in reducing or removal of dormant states of *Plasmodium vivax* that is hypnozoites, but if a person G6PD deficient and if primaquine is given to that person, it may lead to severe hemolysis and then a patient may die.

(Refer Slide Time: 17:13)

Problems related to *P. vivax*

➤ Understanding of Severity

WHO criteria – same as *P. falciparum* except correlation with parasitemia

```
graph LR; Malaria[Malaria] --> Host[Host (Human)]; Malaria --> Parasite[Parasite]; Host --> Proteogenomics[Proteogenomics to understand contributing organism]; Parasite --> Proteogenomics;
```

Outline

- Advancements in Proteogenomics
- Problems with *P. vivax*
- Evolution of Proteome database of *P. vivax*

NPTEL

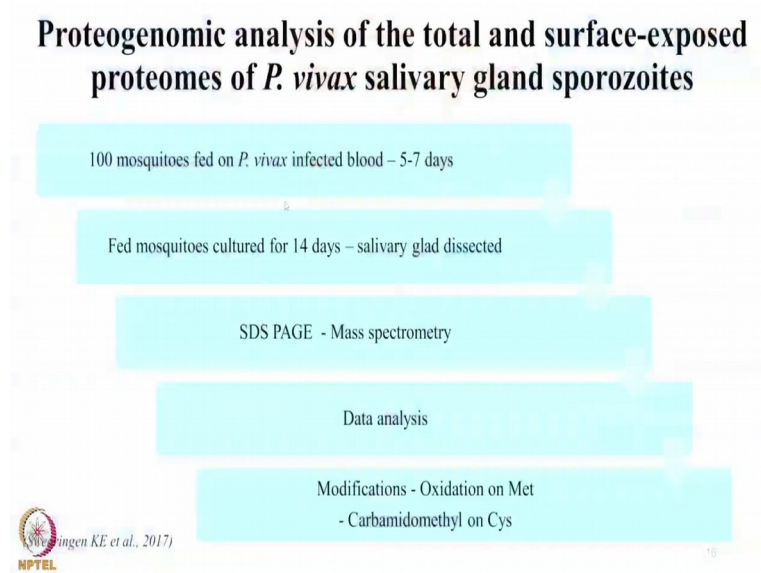
15

So, understanding the parasite is very important and to have a diagnostic kit is also very important.

Apart from that we also do not know what is the criteria for vivax infection to be severe or non-severe. So, according to WHO, they have provided a criteria for differentiating severe, severe infection of vivax from non severe vivax condition and this is exactly same as falciparum because vivax is still not well studied to differentiate it properly on the basis of vivax specifically. Except one condition where parasitemia level is not directly correlated in the case of plasmodium vivax; whereas, in the case of plasmodium falciparum, if the parasitemia level is high one can say that the person, the patient will go towards the severity.

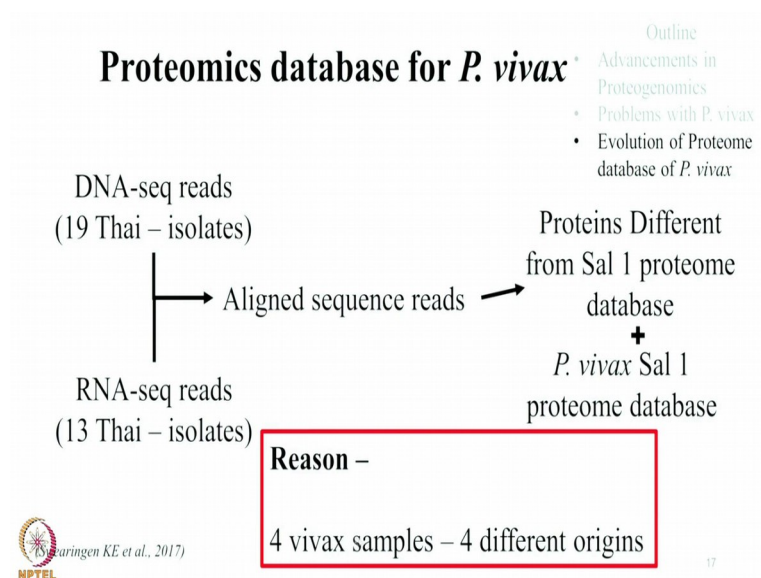
So, at present we do not know whether the severity in the patient is because of the patient in whom the infection is being caused or because of the parasite level. So, in this case proteogenomics can help us in an understanding what is leading to the severity. So, for this I would like to take an example of proteogenomic analysis of the total and surface exposed proteome of plasmodium vivax salivary glands sporozoites, where they have tried to incorporate proteogenomics in the understanding of sporozoites of plasmodium vivax in saliva glands from the mosquito.

(Refer Slide Time: 18:43)



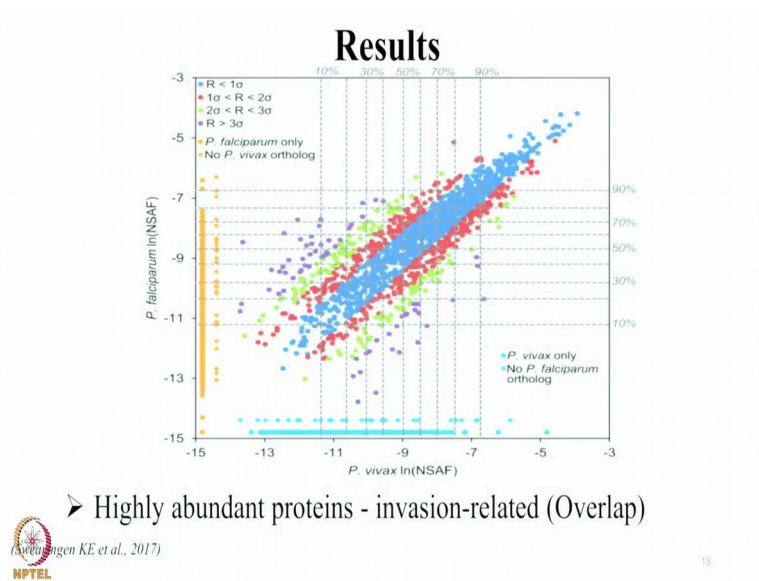
So, what they had taken they had collected 4 different Plasmodium vivax infected samples blood samples and then, they had used 100 mosquitoes to feed them on the plasmodium infected blood for 5 to 7 days and the fed mosquitoes were taken further for culturing for say 14 days and the salivary glands were dissected out of those mosquitoes. Then, the lysate of the salivary gland was taken and run on the SDS PAGE and further processed it for subjecting it on the mass spectrometry. Then, the data was analyzed and; they had data was analyzed and they had considered the oxidation on methionine and carbamethylation as a modification which is contributed because of the sample preparation.

(Refer Slide Time: 19:29)



In this, how they have taken the help of proteogenomics in understanding the parasite is they have taken 19 Thai-isolates and extracted the DNA sequence reads and then, they had taken 13 Thai-isolate, RNA sequence read, then they have aligned these sequence reads and then, they have aligned this customized database of proteome. On the existing proteome database on plasmid db and then, they have removed the duplicates and made a customized database for Thai population. Why they have done this because in the very starting I told you that they have taken 4 vivax samples from 4 different origins to cover the maximum variations, they had taken the different DNA and RNAseq isolates. So, when they try to correlate between Plasmodium falciparum and Plasmodium vivax data, they were able to see that most of the proteins are correlating.

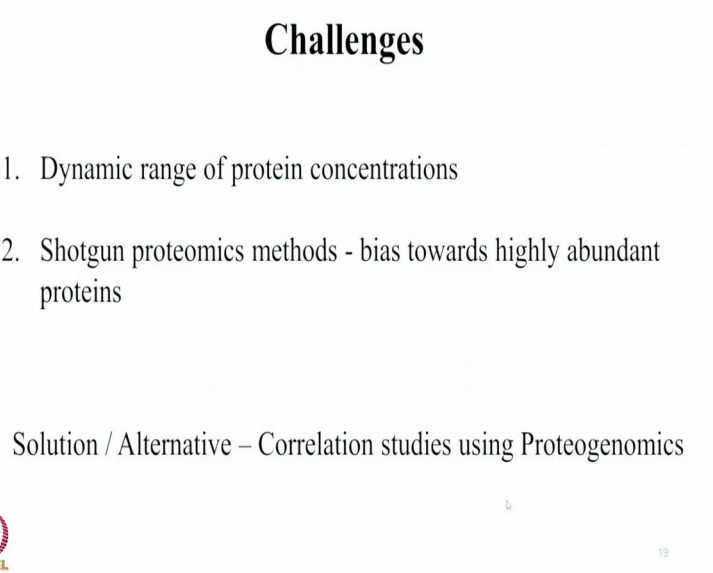
(Refer Slide Time: 20:21)



➤ Highly abundant proteins - invasion-related (Overlap)

This blue proteins in the diagonal are showing the maximum correlation and the ones which are towards the x axis are more related to the plasmodium vivax; whereas, the one which are more towards the y axis are more related to plasmodium falciparum.


(Refer Slide Time: 20:49)



Challenges

1. Dynamic range of protein concentrations
2. Shotgun proteomics methods - bias towards highly abundant proteins

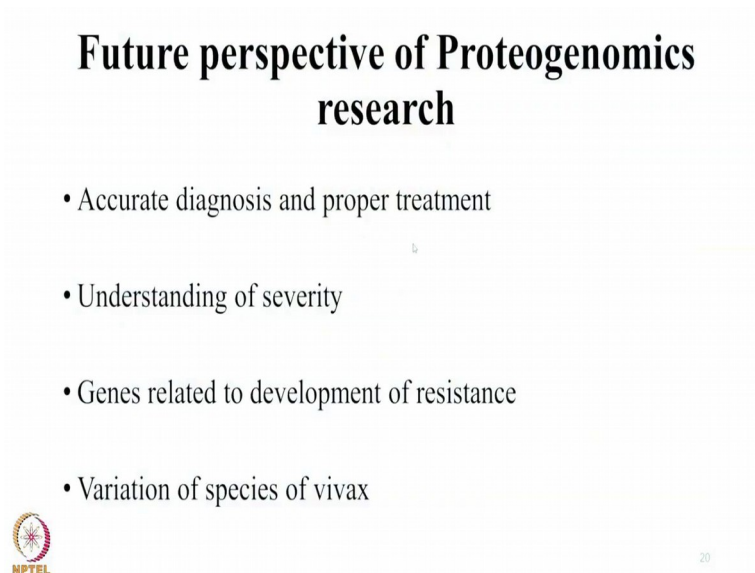
Solution / Alternative – Correlation studies using Proteogenomics



19


So, the challenges with the plasmodium vivax in the study here is the dynamic range of protein concentrations were present because of which the amount of protein which are in abundant quantity are mostly taken up and then, the protein which are in lesser quantity are mostly neglected because this is a shortcoming of the shotgun proteomics. So, it is biased towards the highly abundant proteins and because of which one can miss out on the low abundant proteins. Alternate thought is that one can go for correlation studies using proteogenomics because of which one can also include the low abundant proteins and the genes which are having variations related to the particular protein.

(Refer Slide Time: 21:33)



Future perspective of Proteogenomics research

- Accurate diagnosis and proper treatment
- Understanding of severity
- Genes related to development of resistance
- Variation of species of vivax



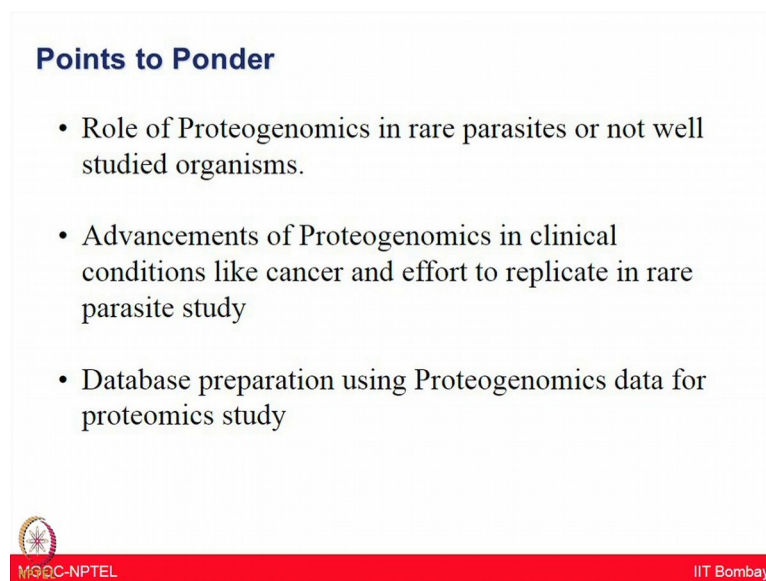
20

So, future perspective of proteogenomics research can be accurate diagnosis and proper treatment, understanding of severity of the parasite Plasmodium vivax, genes related to development of resistance can also be studied by using proteogenomic analysis and also variations of species of vivax. As India itself contained maximum variations of Plasmodium vivax as compared to the remaining, rest of the world. So, we need to understand the variations and vivax as much as possible to treat the patients in India effectively.

So, with this I would like to end the lecture.

Thank you.

(Refer Slide Time: 22:09)



Points to Ponder

- Role of Proteogenomics in rare parasites or not well studied organisms.
- Advancements of Proteogenomics in clinical conditions like cancer and effort to replicate in rare parasite study
- Database preparation using Proteogenomics data for proteomics study

MGC-NPTEL IIT Bombay

I hope in this case study, you learnt about how proteogenomics is emerging rapidly to solve the intricacies of various diseases and this approach can also help to study the parasites or organisms which are not very well understood like Plasmodium vivax. Major issues like the lack of proteome database can also be complemented by transcriptomics and exome sequencing data as we have seen in today's case study. We will also provide you few more case studies in context of cancer and that will probably give you much better idea that how proteogenomics have it started making its impact in the actual clinical field.

Thank you.