Introduction to Proteogenomics

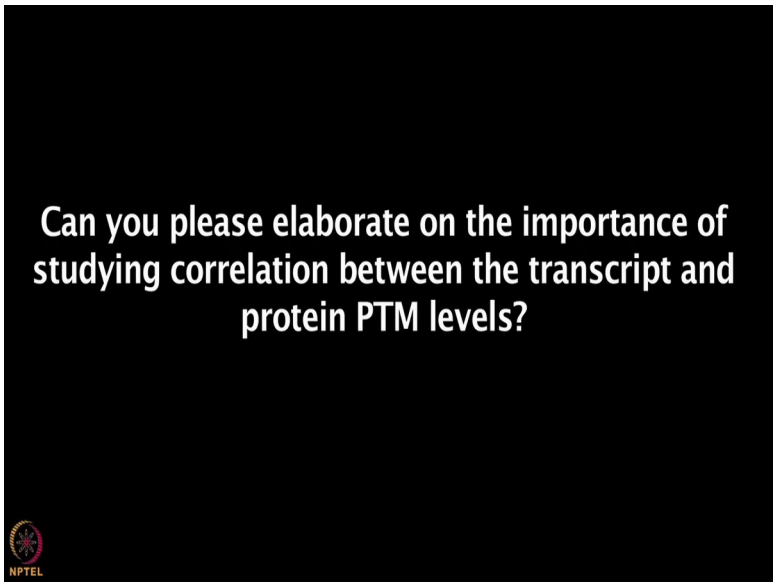
Dr. Sanjeeva Srivastava
Dr. David Fenyo
Dr. Karl Clauser
Dr. Kelly Ruggles
Department of Biosciences and Bioengineering
New York University
Broad Institute
Indian Institute of Technology, Bombay

Supplementary lecture - 19

A perspective on Proteogenomics – IV

So, my name is David Fenyo. And I am a professor at New York University, and my group works on integrating different types of a biomedical data and with a focus on integrating proteomics and genomics.

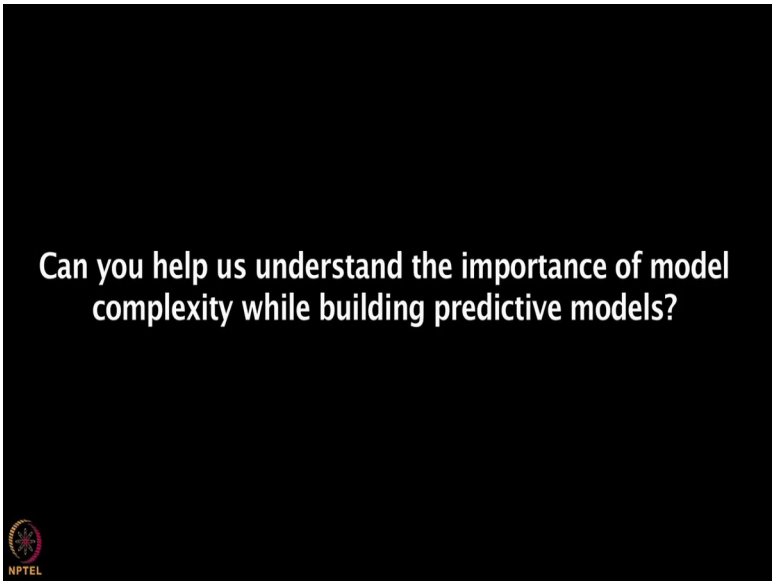
(Refer Slide Time: 00:46)



So, by looking at the correlations between different these different data types. We can better understand underlying biology. And for example, find modules that are proteins that are working together in complexes, we can find the transcriptional regulation, and also signalling on the phosphorylation level. And we can by looking at how these different measurements are correlated we can then understand how the biological function of the cell and try to

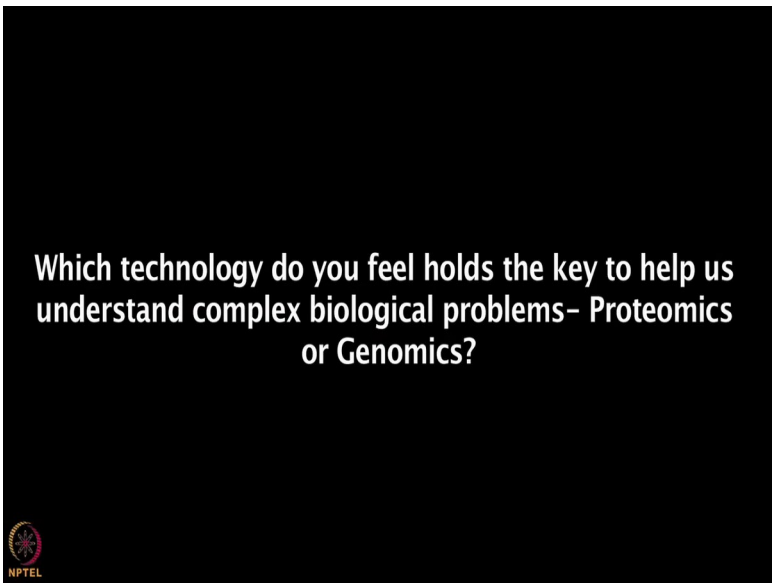
elucidate the complexity and of these functions. And also in cancer, we can then see how these cellular networks are dysregulated and can lead to cancer.

(Refer Slide Time: 01:57)



So, while we building these predictive models, we always have to make a trade off between having a complex model, and or a more simple model. And what we when we make it the model too complex then we risk over fitting our data, and then it the model would not generalize very well. But on the other hand, when we make it too simple, it would not have very good predictive power. So, we have to find this balance between simplicity and complexity. And this is something that will depend very much on our data that depending on also this is both the size and the quality of the data and how complex we can make our models.

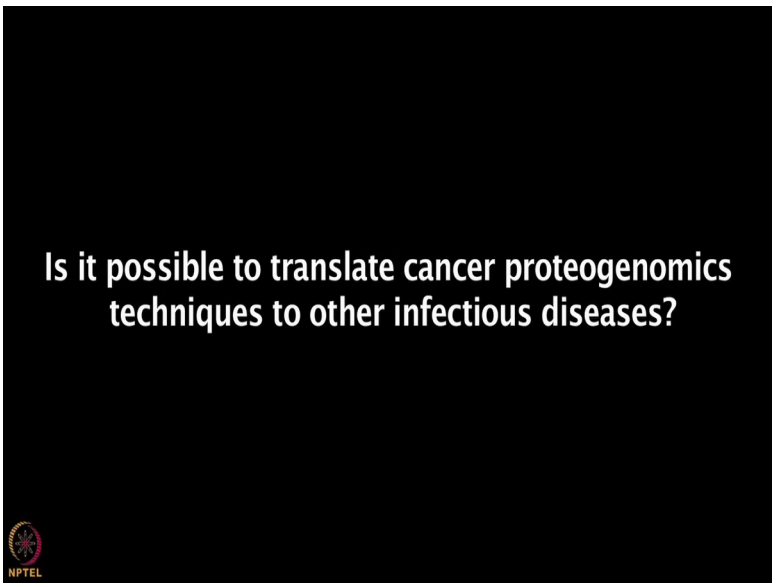
(Refer Slide Time: 02:58)



So, I think this is a question that we get a lot, but the mainly it is that they are very complementary the measurements. When we measure do genomics and transcriptomic measurements in tumours, we see a lot of things that change and it is very difficult to prioritize which of these changes are important, which of the changes drive the cancer.

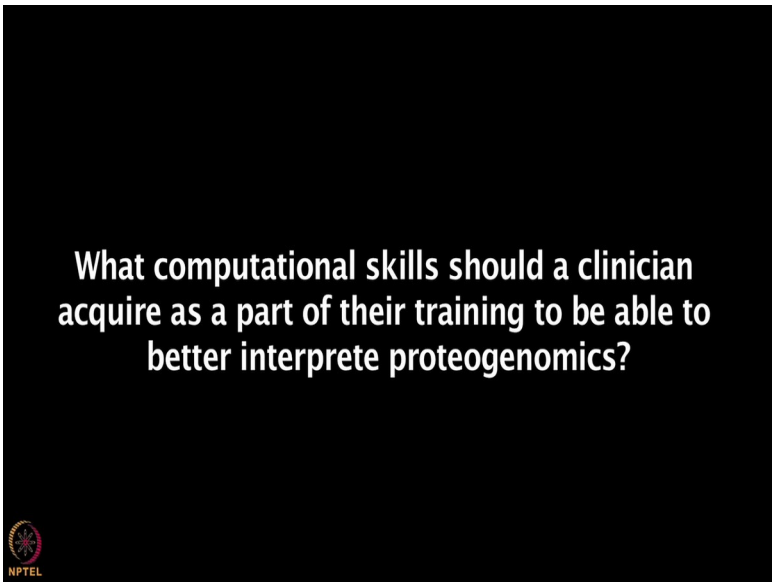
And so by adding in the proteomics the what we can do is use the proteomics to prioritize which of the genomic changes are actually important. And the main reason for that is the proteomics I mean the proteins or the functional gene products, so those are the ones that are closer to phenotype. And so if for example, we have a genomic changes that do not result in any changes in the proteins, then probably they are less important than the ones that lead to dramatic protein changes.

(Refer Slide Time: 04:21)



Yes. So, people have done that other groups including my group have worked on applying proteogenomics techniques to different infectious diseases including HIV and malaria. And the angle that we and our collaborators did was to look at the immune response to these infections, so there, the proteogenomics approach was to first do targeted sequencing of the variable regions of antibodies. And then to get a survey of the immune response, and then a targeted mass spectrometry approach to find which of those antibodies have high affinity to the infectious agent.

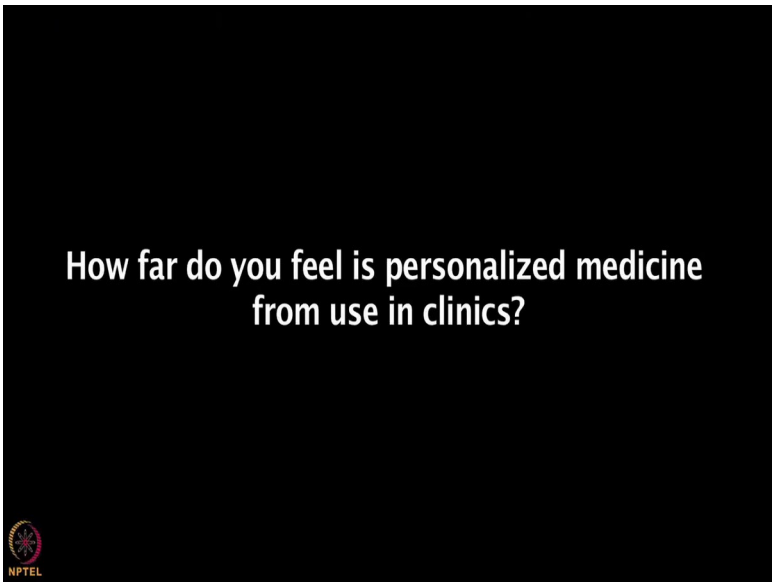
(Refer Slide Time: 05:34)



Yeah. So, we are very interested in actually expanding bioinformatics and computational education into the many different aspects of medical education. But in during medical school that doctor should already done learn to be able to handle and analyze both clinical and molecular data. And as we have seen there is more and more data that more and more measurements that are done on patients that results in data. And it is very important that medical doctors understand, how what are the possibilities of computational methods that can help analyze these datasets, because they are the ones that are best placed they see what is needed in the clinic and the data is available.

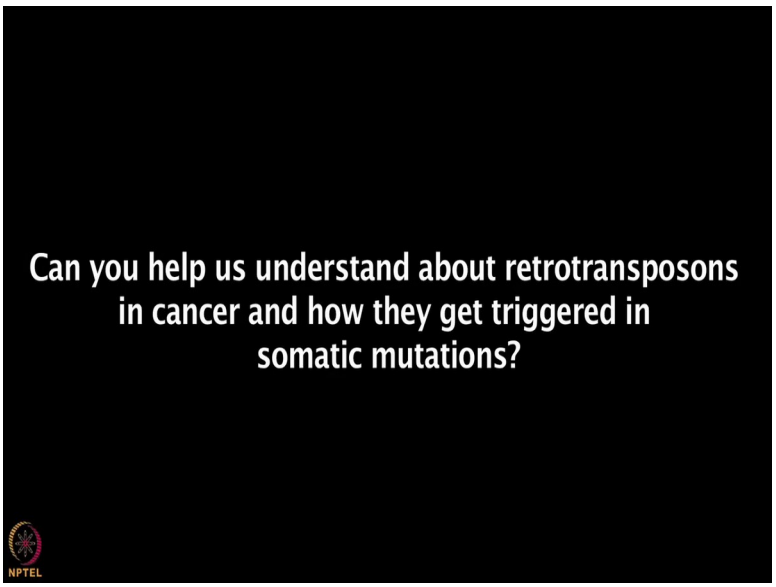
So, if they also understand what is possible with computational methods I think then we can really much faster change and improve the healthcare based and may and really make it sort of predictive and based on data.

(Refer Slide Time: 07:08)



So, I think in some aspects we are already achieving it, but it is a very special areas and like one thing there is be in the treating diabetes, for example, we do the personalized practice personalized medicine today because we based on measurements adjust the medication on maybe even a daily basis. But it is still these are very limited cases where we can do it. And to really do it on a larger scale, I think we therefore that we still needs a lot more research. But, we are slowly getting there and at least today we can imagine how we would get there even though it might be quite a few years away.

(Refer Slide Time: 08:15)



So, about 50 percent of our genome is made out of remnants of retrotransposons sequence. So, this is a very large part of our genome, and most of these are truncated. So, they are not active anymore, but they at some point they were active. And, but there are about a hundred positions in the genome where there is full length of the one retrotransposon called LINE-1, and that have the potential of being active. And the activity in this case means that they are transcribed, proteins two proteins are made, one protein binds to its own RNA, and actually both bind to its own RNA, and one of them is also and the nuclease and the reverse transcriptase.

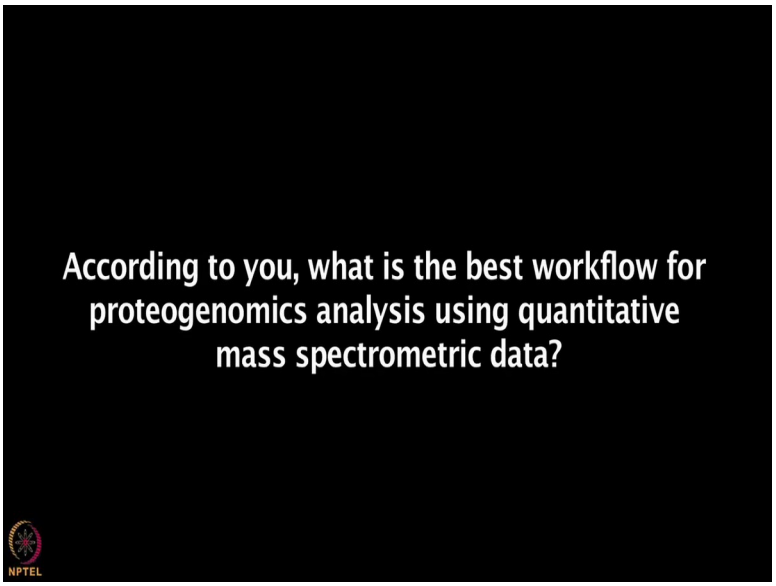
So, after this ribinucleoparticle is imported, it can the one protein can cut, then the nucleus can cut the genome and in reverse transcriptase and insert itself in new position. And this of course, can if it is inserted into a gene or a promoter region, this can cause all sorts of problems, because it can disrupt the gene or so it really the host has developed really a very efficient suppression mechanism. So, the retrotransposons are not or in most somatic cells not active. And but what happens in a lot of tumors what people have observed is that we get transcription. And so we approach this with the proteogenomics approach where we look at both the transcription and the proteins.

So, what we have seen is that we can see where a lot of transcription factor binding to one retrotransposon , and we can see that it has lot of transcription in certain tumors and it also has the proteins are produced. And finally, we can only I am also developed a matter to look

at novel insertions in the genome. So, taking this together, we are now looking also at what host proteins are needed for this process, what proteins do they interact with and we are looking at what regulates this both on the transcription factor level, but also on the translation level. So, and the nice thing is that the proteogenomics data that is coming out of several labs nowadays on tumors, we can actually do these studies with existing public data.

So, my name is Karl Clauser; I am a Principal Scientist at the Broad Institute of MIT and Harvard in Boston. And I been for quite a few years now doing research in proteogenomics mainly oriented around cancer.

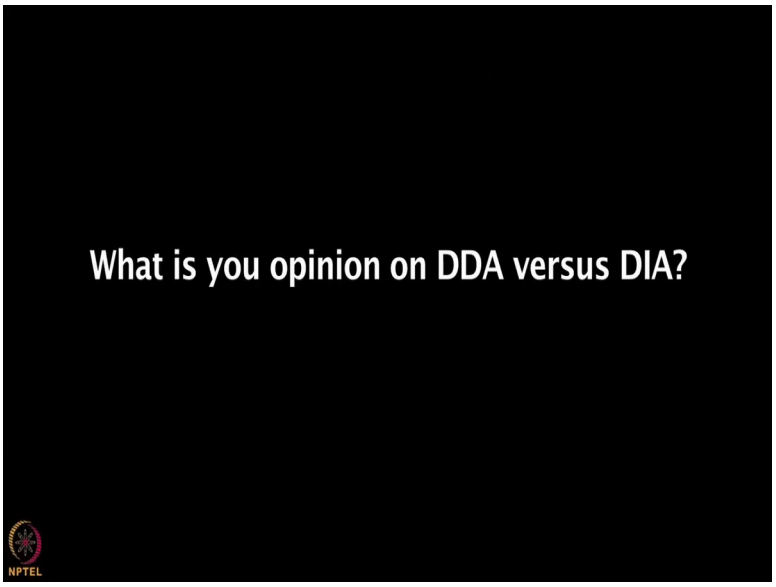
(Refer Slide Time: 12:06)



The research we do is mostly trying to analyze tumors that come directly from patients where we are trying to get an overview of the proteomics and the genomics of cancer. And so what we seek to do is have tumors from over a 100 patients and we want to have depth of coverage that in the proteomic side is say 10,000 proteins or more. So, in order to make effective use of instrument time, we use multiplexing strategy that involves TMT labelling, peptide fractionation, and then automated mass spectrometry that does LC-MS/MS and doing that we get both proteomic information and phosphoproteomic information.

And the phosphoproteomic information comes from a step that of isolating and enriching for phospho peptides used using immobilized metal affinity chromatography that gives us one set that we do proteomic work, and one set that we do phosphate proteomic work. Then we like collects millions of mass spectra that software is used to interpret the spectra I am responsible for building some of that software. And that creates large amounts of information that we then seek to integrate with genomic information and learn things about the cancer process and processes, and how to put different types of cancer into better classifications and help ultimately to get better treatments in diagnostics for patients.

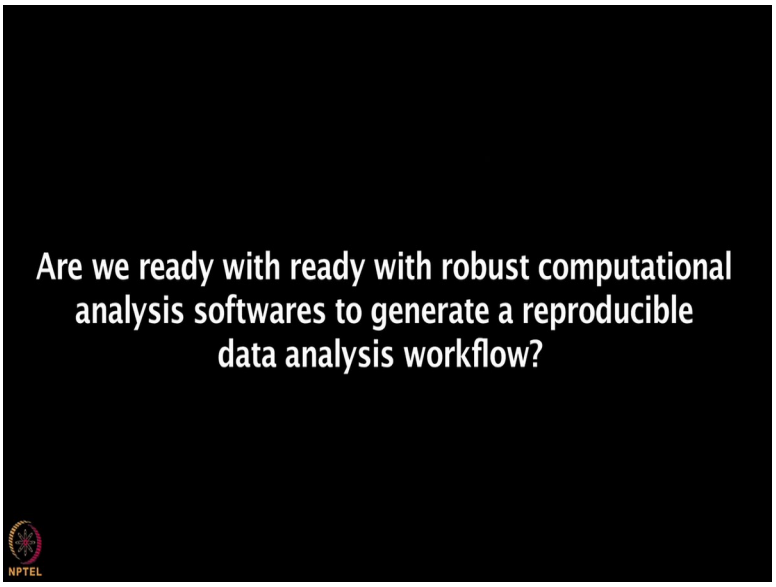
(Refer Slide Time: 13:51)



Well. So, DIA is data independent acquisition, and DDA is data dependent acquisition. And I think amongst the practitioners in the field, there is a bit of controversy at this point right. So, the DDA is a bit of older technique, and the engineers that build and design instruments have been for many years working to do that very well ok. And DIA has as emerged in the last few years as people with those instruments basically trying to run them in a way that gets what they like to think of as more comprehensive data by collecting many things at a time. And some people would say that it makes the data a bit more of a mess ok. Simply put, I think it is like people who are fans of DIA are a bit like New York, Yankees fans that live in Boston ok. I myself I am Boston Red Sox fan, and as far as where I stand let us just say the Red Sox win the world series this year ok.

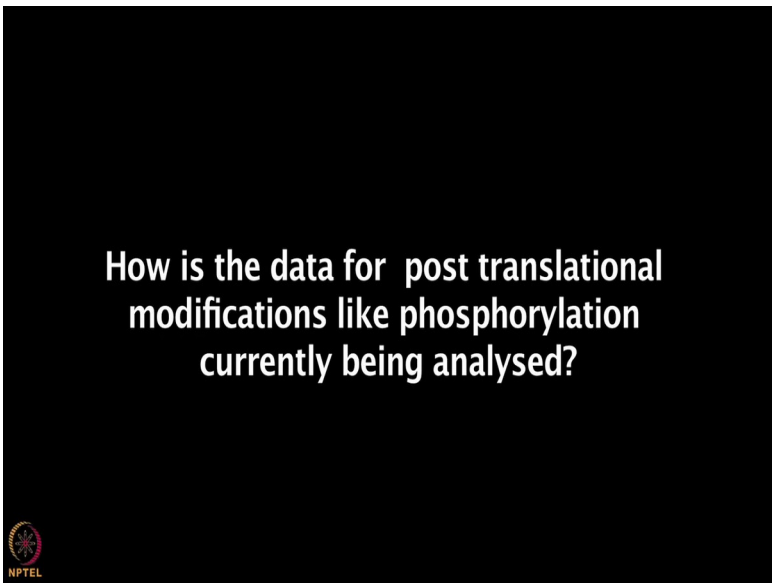
Now, if I take put back on my scientist hat, I would say that were probably headed for an area say with the next generation of instruments or so, where the two techniques are going to become more merged right. So, the instrumentation is already becoming faster. And I think DIA makes not enough use of that information in order to trigger acquisition, and some of the compromises that are currently made to collect data in a DIA fashion will no longer be limiting it with say one more generation of instruments. And you will you will have the instruments being fast and sensitive as well as being specific and I think that combination will look a bit more like what traditional DDA is.

(Refer Slide Time: 16:01)



Well, I think my lab in is already producing data that is of high quality and gets published in top notch journals, but at the same time it is a bit like being a homeowner. You got to live in the home and your family grows and you want to make the home a better place all right. So, things are constantly being improved, but I think were already to the point where we can claim to be robust and reproducible, but that is not to say that were satisfied right. We would like to be able to do things more efficiently, we would like them to be more robust and more reproducible, so that we can get even higher quality data.

(Refer Slide Time: 16:47)



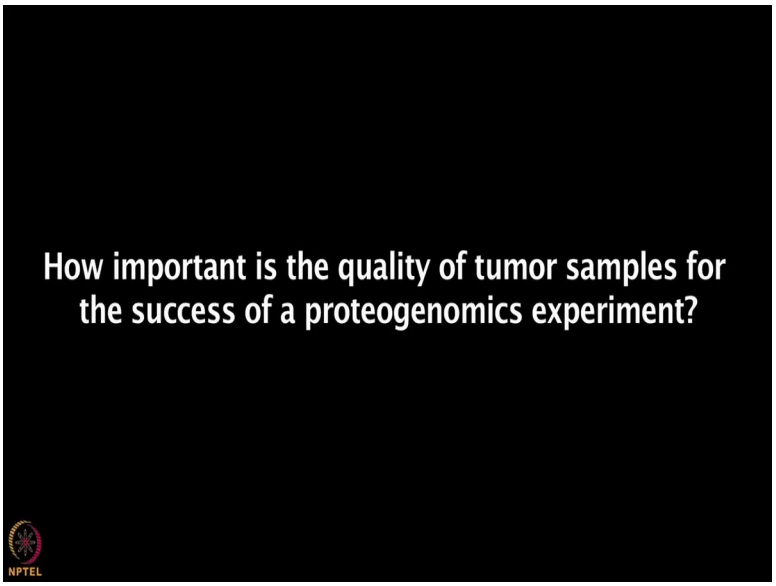
Well. So, it is already possible to do in a high throughput manner to do phospho proteome analysis. We are now doing a acetylome analysis where we enriched for acetylated peptides, peptides that are acetylated on lysine residues. And we do that using antibody enrichment where than anti acetyl lysine antibody. The phosphopeptide enrichment today is doing done by using an immobilized metal affinity chromatography approached.

When we have a complete data set, those data sets often have significant numbers of missing values that make and drawing conclusions from those data sets harder. If we can improve it, I think the enrichment process is probably one of the most limiting things at this point in the most room for improvement. Sensitivity anyway you can get it always helps these things and I guess that is that is one of the major features that. That inevitably I think right now we also have a certain amount of uncertainty with regard to localizing the sites of modifications when you have multiple residues that are possible to be modified in the same peptide.

And the limitations in doing that are not really software there are more the underlying data. So, MS fragmentation tends not to give complete sequence and so you often end up with uncertainty. So, today when we do phospho proteomic analysis maybe 70 percent of the phosphopeptides, we identify we can confidently localize the site to a particular serine threonine or tyrosine residue and if we had more complete fragmentation that would improve ok. For acetyl lysine containing peptides that is much less of a problem because there is not as much potential for there to be multiple lysines in a peptide ok. And if there is it is going to

be one lysine at the c terminus because of tryptic peptide, and another lysine that is somewhere else that is probably the acetylated one. So, it is a less of a problem with localization.

(Refer Slide Time: 19:06)



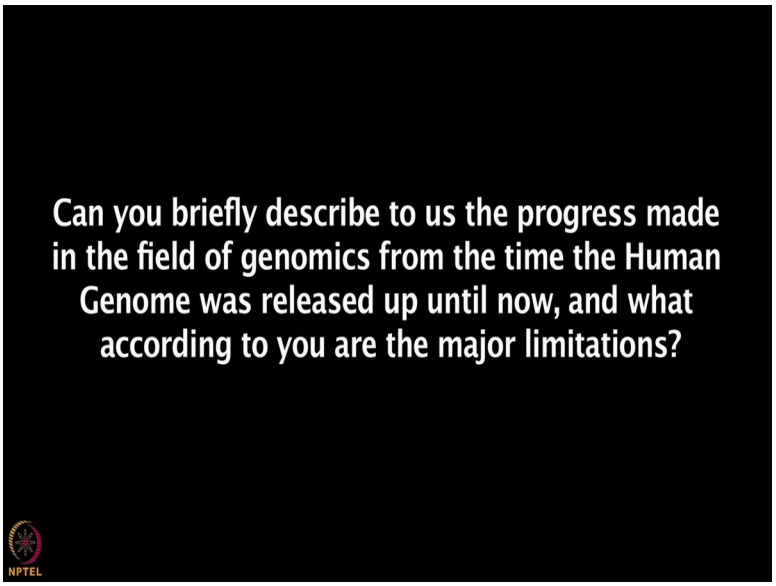
I think in order to do good science you always need to have good samples to start with right. Most of the work that I am involved in these days is related to the CPTAC program that is in the United States run by the National Institutes of Health. And at this point that the tumors are collected under a protocol that is been optimized to make sure that we can get as good proteomic data quality as possible. Those samples come from different places in the world. And it is it is critical, I think to have an effective program to have partnerships with hospitals and cancer centres that can provide those materials.

Now, if we were to improve the technical aspects of our work particularly by being able to work with less and less material and still get the as information the data quality out that we want we could do even better. So, right now we tend to require a bit larger tumors that are often easy for surgeons to obtain from patients. And what were actively trying to do is reduce the amount of material that it takes for us to generate the data, so that we can work effectively with just biopsies of tumors. And then I think that is going to open up areas to larger studies that hopefully can produce even better data.

Hi. So, I am Kelly Ruggles; I am an assistant professor at NYU School of Medicine. I am also the director of academic programs for the Sackler Institute also at the NYU School of

Medicine. So, I am I am involved in both research and education at NYU. And my lab really focuses on multi-omics integration proteogenomics and microbiome and cancer lots of different areas, but really interested specifically in looking at how we can integrate these diverse data types to understand human disease.

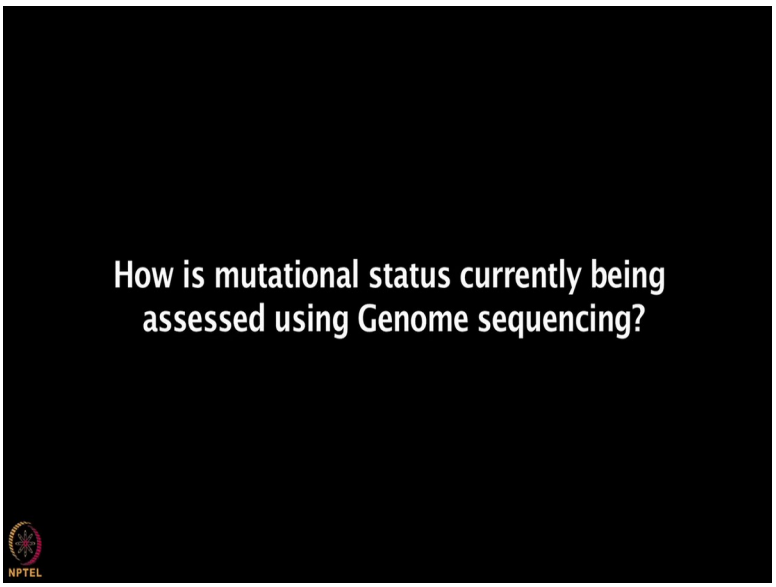
(Refer Slide Time: 21:22)



So, a lot has happened since the human genome project and mostly because the technologies become so much better. So, we are able to assess genomics at a much higher depth, we have much higher coverage to become much cheaper to do sequencing. So, we have many more organisms that have been sequenced and we can look at things like epigenetics and the transcriptome and all sorts of levels of omics data.

So, a tremendous amount that is been done in terms of the limitations you know we still have only sequenced a small percentage of the total organisms that are on the world. So, there is a lot more we could do, but we are limited with that. And also we to get really good depth with like whole genome sequencing is still very expensive. So, I think with time we may see even more improvements as the technology improves.

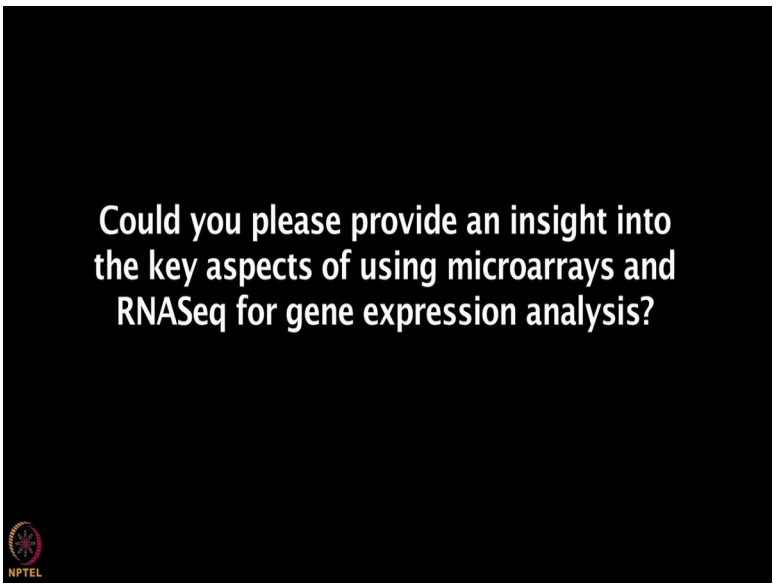
(Refer Slide Time: 22:24)



Yes. So, mutation status that we deal with a lot specifically with our cancer data because it is something that were really interested in and little is understanding how somatic and germline mutations are identified, and how they affects on tumors. And so the actual identification of variants occurs through several different pipelines the TCGA. So, the cancer genome analysis has been instrumental in coming up with these informatics pipelines that allow for variant calling from either whole genome or whole exome sequencing and there is also SNP arrays that are available.

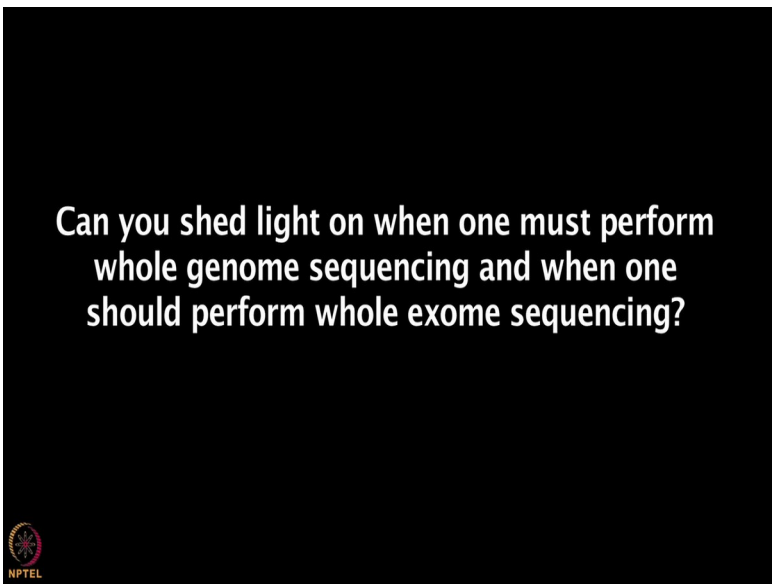
So, there is a couple of different ways that you can do this. And it is been really well developed because of all the work that is been put into this. And like I said a lot of it is been done by the cancer genome atlas as well as other big consortiums and smaller groups as well. So, we have we have made a lot of progress with that and I think it is become a really interesting way to understand cancer.

(Refer Slide Time: 23:33)



So, micro arrays have sort of gone out of style as RNA seq has become the primary method for measuring transcriptomics. And the main reason for this is because with micro arrays, you really you need to choose your genes before you measure them. So, there are specific probes that you pick. So, you pick a certain number of genes essentially that you are able to measure. And then you only can measure those, but with RNA seek you are able to measure everything that is in your sample. So, it is a, it is an unbiased approach and it is something that has really pushed the feel forward having the ability to not choose beforehand what you are measuring, and really being able to measure whatever you want in your sample.

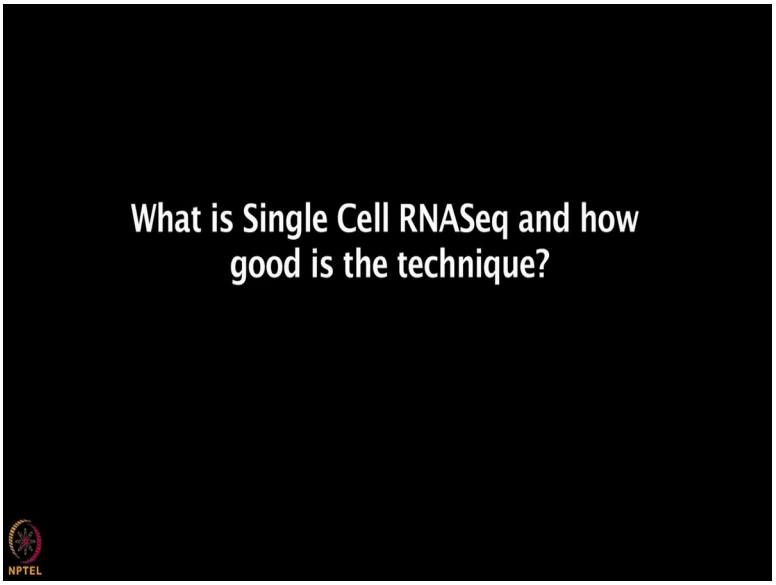
(Refer Slide Time: 24:22)



It is a good question. You know they are both they are both really important and they are good methods, I think that the main reason to use one or the other is usually cost and the question that you are asking. So, whole exome sequencing with exomes especially, so for if we are talking about the human genome for example, it is about 2 percent of the genome as an is exome. So, if you want to have really high depth and really high coverage of what you are sequencing, you would choose exome sequencing if you do not have you know an unlimited number of funds.

So, if you are lucky enough to have a ton of money, then a whole genome sequencing is a great way to go because you can get a lot of information about the non coding regions which were learning more and more are extremely important to understanding the cellular process. So, it really depends on your question, and it also depends on how much how much money you have to spend.

(Refer Slide Time: 25:24)



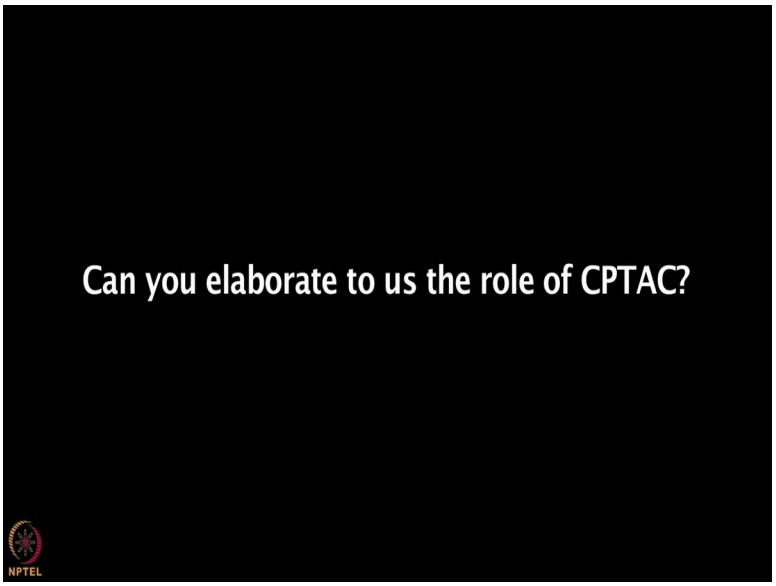
Single cell RNA seq is a really hot field and technique right now. It is something that i am not doing specifically, but I work with and collaborate with a lot of people who are doing single cell RNA seq. And it is, it is a method where you are able to measure to separate cells out one by one and measure specifically the gene expression within that cell.

So, there are a couple of methods that you can use to do this one of which is to use droplets. And you are able to put cell each cell in a different droplet and actually do the whole library prep within that droplet and barcode the RNA for each cell within the droplet and then

sequence all of them together and pull out afterwards the specific cell specific RNA expression. And it is a really interesting and great method for understanding heterogeneity of samples, and it is something that the coverage right now is not very high I think it is about a thousand genes depending on how people do it.

So, I think as again as our technology improves, I think this is going to become an even more exciting fields.

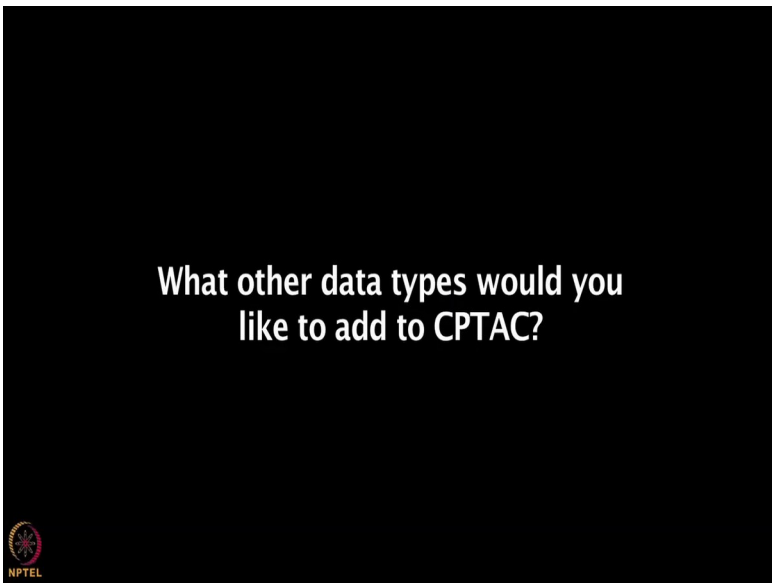
(Refer Slide Time: 26:41)



Sure. So, CPTAC - the Clinical Proteomic Tumor Analysis Consortium is as is funded by the NCI and within the NIH. And it is it is a great consortium because we are working, it is a large group of us who are working together to try and understand cancer using proteogenomics. So, we were using proteomics, phosphoproteomics, genomics, transcriptomics to really try and understand if we integrate this data can we identify biomarkers can we find signatures can we understand drug toxicity and predict how people will respond to different drugs.

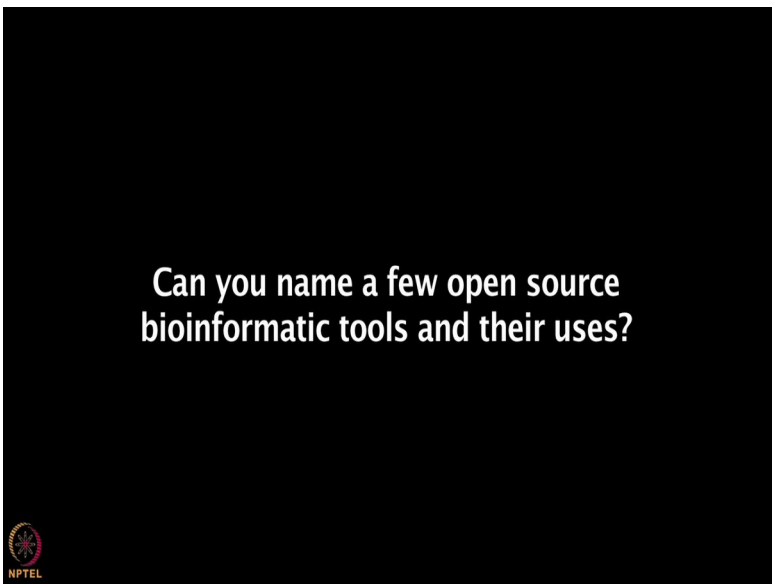
I am so trying to harness all of this data to really understand the clinical aspects of cancer and come up with new ways of treating and diagnosing people. So, we are working on a lot of different tumor types right now, and it is something that is going to continue to go on I am part of one of the data analysis sorry the data analysis teams. So, we are really working were and we are in the data and we are trying to figure out how best to analyze this data and how best you understand cancer using even more levels of data than we have used before.

(Refer Slide Time: 28:02)



That is a good question yeah. So, I would love if we also had metabolomics data. So, metabolomics data you know really complements proteomics and genomics data in that. You can see exactly what sort of enzymatic reactions are occurring, and you can try and figure out there are certain things that are building up in the cell or if there certain parts pathways that are up or down that are causing different metabolites to change in terms of their concentrations. So, that is something that I think you know is another data type that we can really benefit from, and it is something that is becoming more and more popular in these multi omics analysis that I personally would be really excited to work with as well.

(Refer Slide Time: 28:51)



My lab works on a lot of creating a lot of open source tools and we work with a lot of people who create open source tools, and you know I think it is it is a really important thing for us as the scientific community to contribute. So, open source meaning you know we create these pipelines that we can make public something that my lab and something I am particularly interested is making things interactive.

So, having it be on a web server and having it be available for people who are not computational who can upload their data, and it and really explore it in an interactive way. So, that they are able to ask their own questions and not relying on someone who is a bioinformatics expert to always be taking their data and doing something with it and giving it back and having this iterative process.

I think having this sort of having it available to the scientists themselves to ask their own questions and play around with the data is something that I think is really important. And something that we should all work towards especially the computational field is allowing other scientists to really who do not have the same skills to really be able to look at their data themselves.

(Refer Slide Time: 30:12)

Though I am very involved in our computational biology program at NYU, I am part of the masters. I help lead masters program and I am very involved in the Ph.D. program. And you know training our scientists at this point to really understand how to also do some programming or to at least understand how the programming works, they do not have to

become experts you know, we cannot all be experts in these fields. And I think also really teaching people how to be collaborative. I think we are at a point in science where we all rely on each other and we it is hard to run a lab and just be insular and do everything yourself.

So, having you know people who do the wet lab and people who do the informatics, and people who sort of translate between those two and training our next generation of scientists to understand this and be able to work it better in groups I think is something that is really important. And also just training them to you have the statistical and computational backgrounds that is required to drive the field forward the entire scientific field forward I think is something that we all need to think about and invest in.