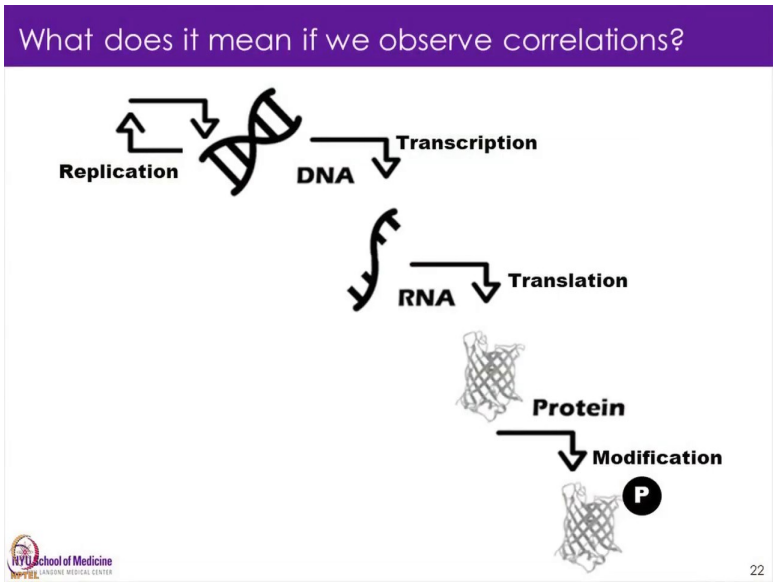
**Introduction to Proteogenomics**

**Dr. Sanjeeva Srivastava**
**Dr. David Fenyo**
**Department of Bioscience and Bioengineering**
**Department of Biochemistry and Molecular Pharmacology**
**Indian Institute of Technology, Bombay**
**Institute for Systems Genetics**

**Lecture - 42**
**Introduction to Proteogenomics - II**

Welcome to MOOC course on Introduction to Protogenomics. In the last lecture by Dr. David Fenyo you were introduced to the concept of protogenomics and its ability to provide expression level information at multiple levels. In today's lecture Dr. Fenyo will introduce you to few more capabilities of protogenomics and its applications in various clinical problems. So, let us welcome Prof. David Fenyo for his today's lecture.

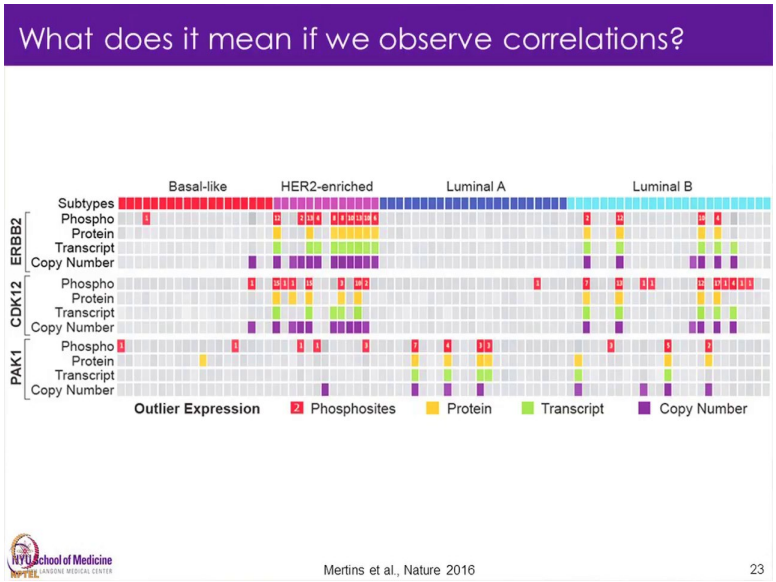
(Refer Slide Time: 00:55)




So, one thing that people observed a lot in the cancer genome atlas and in many cancer genomic studies is that there are a lot of changes when you look in the either whole exome sequencing or RNA seq. And there it is a lot of changes and it is difficult to say which changes are interesting, which changes are important. So, one thing that one can use the proteomics for is to focus in on the changes that actually have an effect on the proteome because I mean it is really in general often does not matter if we have a copy

number change, that does not lead to any changes in the protein that is probably not so interesting than if we have something that actually changes the proteome.

(Refer Slide Time: 01:57)

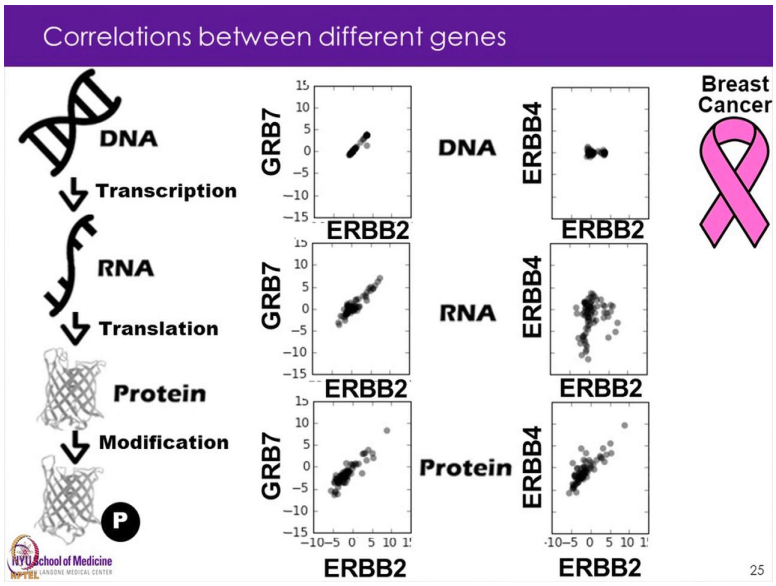



So, one thing that we can then look at is so this is from the CPTAC breast paper. We can look at when we have consistent, we see consistency between the different measurements. So, this shows copy number transcript protein and phospho protein, and we see that this is again for ERBB2 here in on top. So, we see that usually when we have an amplification, so a copy number change, we then also see that the transcript levels are high, the protein levels are high, and a lot of the phosphorylation levels are high.

So, when we see this consistency between the different data types, we can then we do see that the ERBB 2 which is well-known to be an important driver in subset of breast tumors, so that comes out. And then we see this for a few other kinases were in other samples like PAK-1 for example, also has this consistent and there are a few others.

But that says so we can definitely focus on things where all the different data are correlated, and tells us the same thing. But we will if we only do that, that will be very limiting. So, we but that is one way to get started to see what are the contestant things between different data levels. So, on another thing is we can look at then correlations between different genes.

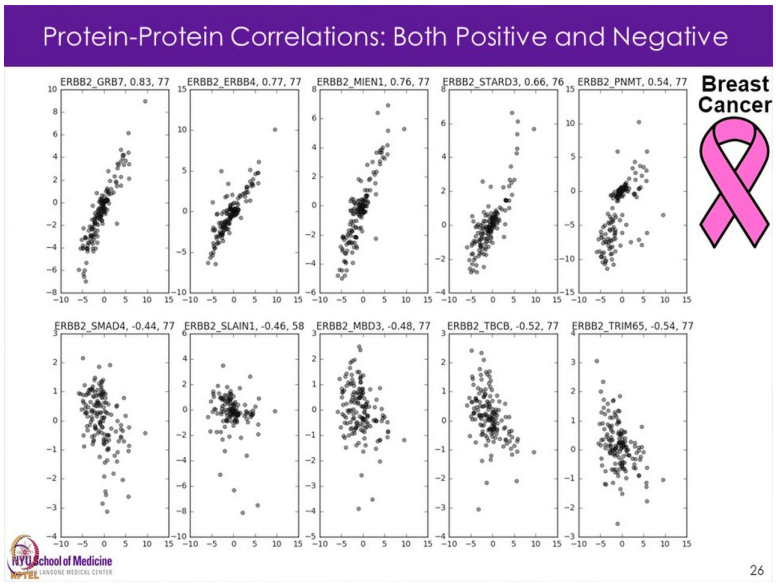
(Refer Slide Time: 03:54)




So, this is one example again ERBB2, now comparing it to GRB7. And we do this both on the DNA, RNA and protein level, in this case we have very consistent result, they are highly correlated on all the measurements. And so the reason why we have the copy number change, so highly correlated is that they are very close to each other on the same chromosome. So, this is one example.

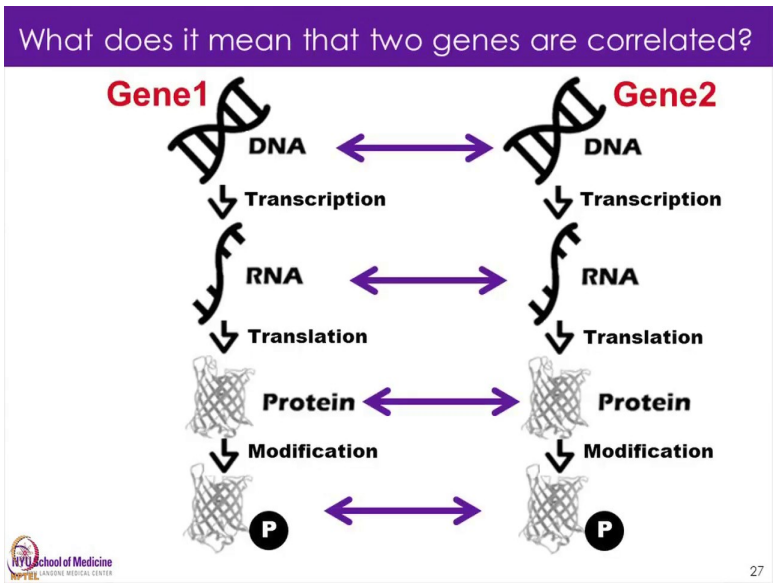
But if we compare ERBB 2 to ERBB 4, we see that we have ERBB 4 does not have any copy number changes. And the transcript levels are not correlated, but we have rather high correlation on the protein level. And this is quite common for proteins that work together. So, if we look at ERBB 2 in general, and just rank how in breast cancer, and how which are the genes that are highly correlated with it. Now, it is only on the protein level we see that we have some high

(Refer Slide Time: 04:57)




So, these are the two that we looked at GRB 7 and ERBB 4. So, those are the two highest correlation, but then there are others that are highly have high positive correlations also. And we have some that have rather high negative correlations at least. So, this one is minus 0.54 correlation coefficient. And so then we can start by looking at this we can start sort of seeing which what are the sort of other proteins that each protein works with.

(Refer Slide Time: 05:44)

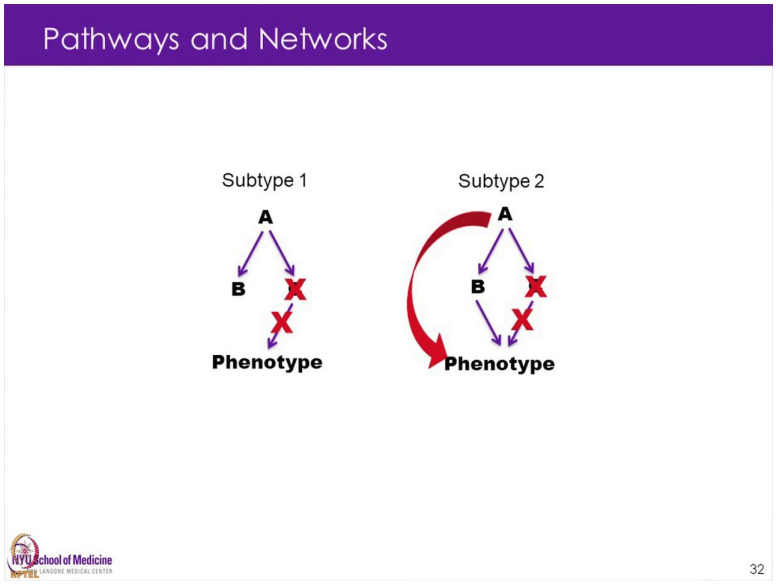



So, then so we looked, so the question is what does this then mean? So, we saw that for the simple thing was that on if we see correlation on their copy number level, it is usually

just that we have that they are close to each other on the genome also if the copy number one changes, it is usually the copy number changes in a larger region. So, then and if they are close to each other, they will both change. And but then it gets more involved what we can and we saw then on the protein level, we could have that if they were working together in a complex, they were regulated in the protein level maybe by that the complexes are formed.

And if they are one of the components is on its own, and left over, and it does not have anyone's pair with it gets degraded at a high rate for example. But they could be other and but we see also correlations at RNA level. So, then we can start thinking about how do we, then use this information. And one way to do it is if we have large enough experimental data set, we can start to see look at networks, and pathways to see how if we see any differences.

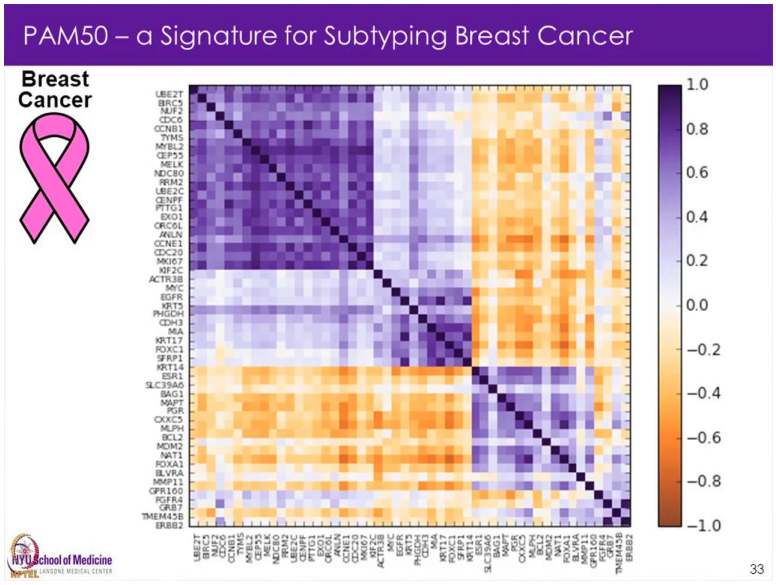
(Refer Slide Time: 07:24)




So, so, for example, this is just one example, where we have 3 genes, A will affect B, A will affect C also, and C will affect the phenotype. So, in this case, we see that A will affect the phenotype through C. So, if we then inhibit C, we will also inhibit the affect so A now cant affect the phenotype. So, if there if we have one subtype of our tumors that have this structure, then we see that inhibiting C would be a good way to change the phenotype in this case treats the tumor.

But, if we have another subtype where we also have a connection from B to the phenotype, that will then change things. So, in this case if we have a drug against C, it will stop this pathway, but we will still have the other pathway to so that is one thing that one can start thinking about when one has these correlation data between different genes on different levels.

(Refer Slide Time: 09:18)

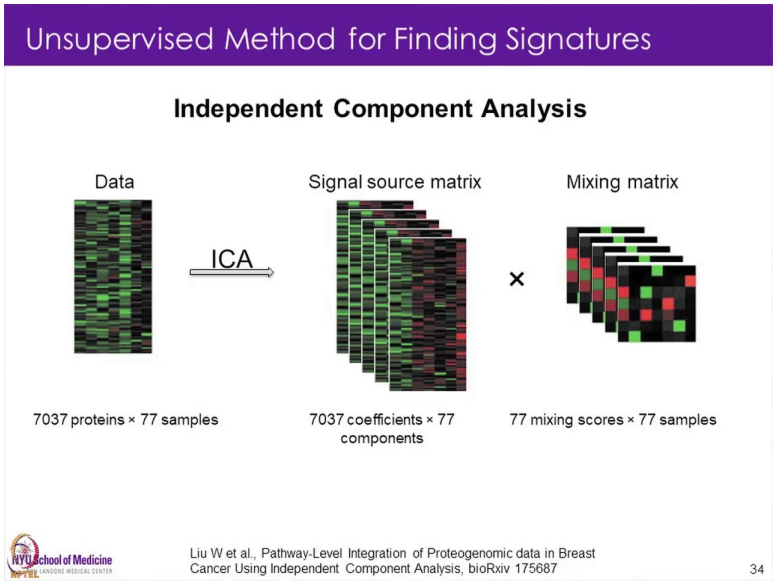



And so then we are going to talk a lot about signatures and already now we have talked about different signatures also. For breast cancer, we have the PAM 50 signature which is a set of 50 genes that are used to subtype the to determine the four major subtypes of breast cancer. So, here we just listed the different the 50 different genes. And we are looking just at the correlation between them. And then we see that and this is very common with signatures that we select genes that are quite highly correlated, but not perfectly. So, we see these regions of high correlation.

And then we see another group that be the quite highly correlated within the group, but anti-correlated with the other group. So, the some of the genes go up and down, and that is how they then build a signature. So, the question is then how do we get a signature? So, Mani talked about unsupervised analysis and that is one way that we can try to extract signatures. And there as he mentioned there are many different ways to do the unsupervised analysis. One example of it is independent component analysis. And independent component analysis was initially developed for if you have a room with

several people talking and then you have a microphone you hear an overlay of the different voices and so the whole point is to then separate out the different sources.

(Refer Slide Time: 11:17)




And the way that it is done is that we have the composite of all the sources and then we separate them out, so that we have into two matrices. One that would be the signal source matrix and the other one is the mixing matrix on how to mix them. And so we can do the same thing for tumors. And there the idea is that there are several biological processes that both basic biological processes that does not have anything to do with the cancer that we measure, but some of these biological processes are related to cancer. But what we are measuring is a sum of these different processes going on. So, if we separate them out, so the first step is completely unsupervised. So, that the yeah.

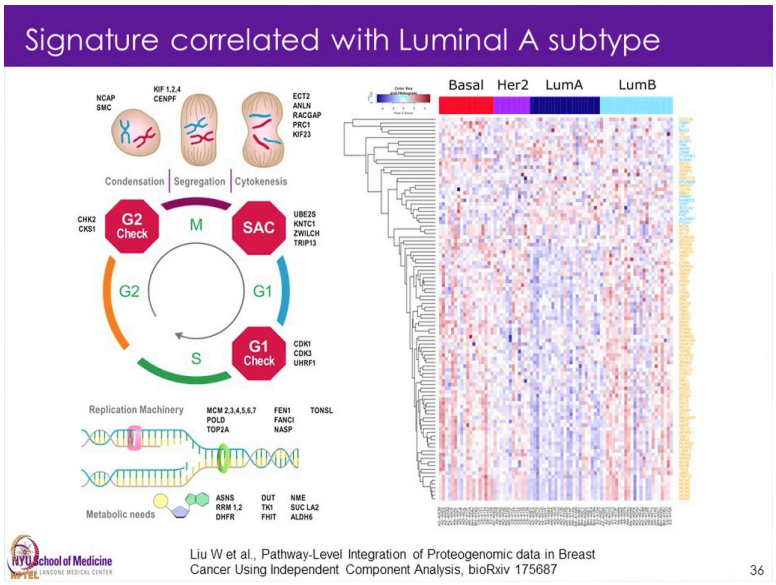
Student: So, sir in the case of this mixing score that we are seeing is it because of differentially expressed proteins and these components of the populations were are taking here

Yeah. So, these could be either the proteins levels that we measure or RNA we could do it for either or we could even combine it and yeah. So, we gonna actually talk about that how in an analysis like this, how one could potentially combine both proteins and the transcriptomic measurements.

Student: Sir I was having a doubt the mixing score exactly stands for?

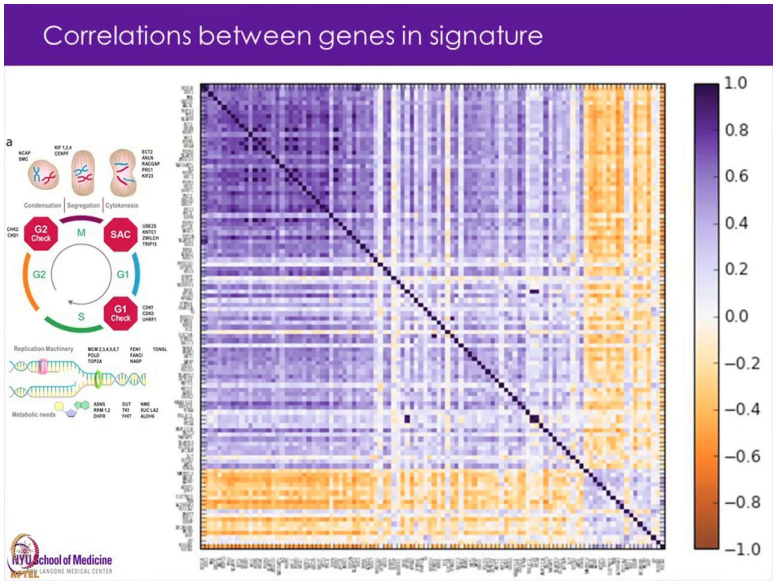
Yeah. So, the signal source matrix are the sort of signatures of potential different biological processes. And, then the mixing scores are how large effect is of each of those signatures on the overall measurement. So, yeah, so as I said the first step is then completely unsupervised, but then of course, since we both get signatures from basic cellular processes and or processes and they are related to cancer, we will then go back and look at which of these signatures are correlated to different clinical data types.

(Refer Slide Time: 14:02)



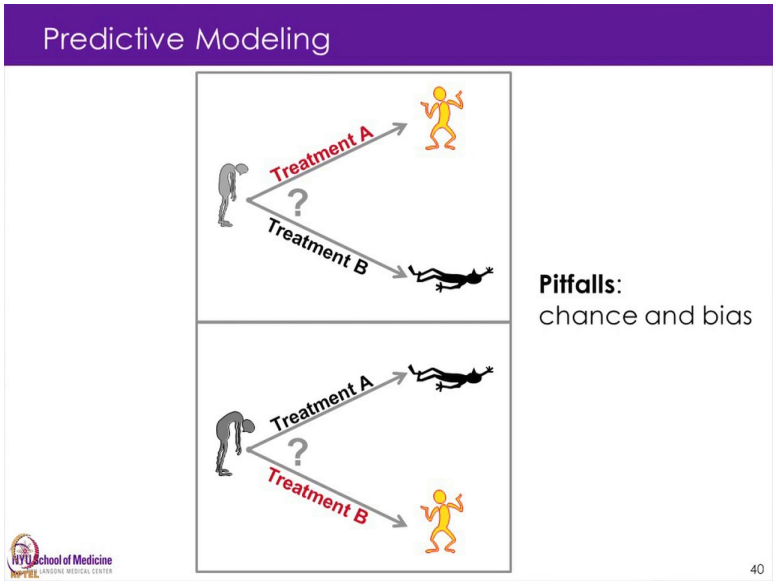

So, then one example would be here the one of the signatures are highly correlated to the luminal A subtype, and then what we see in that signature is that there are a lot of genes associated with the cell cycle. And most of them are down. So, this is the luminal A (Refer Time: 14:26) subtypes to see that most of the genes are blue in this case. So, they are down. So, these cell cycle genes are much lower in this signature. And this is actually well-known signature of luminal A. But we got it through an unsupervised analysis and then we are able to recover it without making any initial assumptions.

(Refer Slide Time: 14:55)




And, if you then look at this signature again, we just look at the correlation between the different genes in the signature. We again see the same thing that we have most of the genes are in this case positively correlated and, but there is quite a bit of variation, and that is one thing that we have still variation between the genes, but they are quite well correlated and then we have another group that is internally positively correlated, but then anti correlated to the rest.

(Refer Slide Time: 15:40)

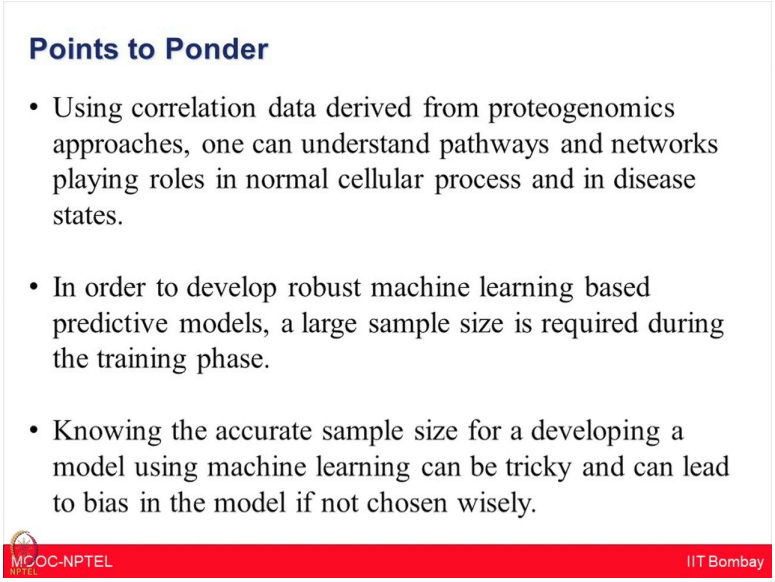


So, now, we going another thing that we are going to talk about this how do we use this for predictive modeling. So, this is just a cartoon showing that patient here who has cancer can either be treated with treatment A or treatment B. And treatment A is preferable, but at the time when we start the treatment we do not know. So, we definitely would love to have a be able to do a measurement at this point and predict which of these two treatments are would be the best.

So, that is an for another person, of course it could be the opposite the treatment B is what would be preferable. So, this is one thing that one example, but there of course, a lot of other things we want to predict. And we going to talk more on Saturday about how to do this prediction Mani talked and introduced it, but we will talk in more depth about it. And the biggest problems with this kind of analysis is the that we can have bad luck, If you do not have enough samples especially which we almost never do, they can just have random matching and that will give us them predictive models that only work for the patients that we trained it on and not generalize. And the other serious thing is bias. So, on the for to treat the random variation, there we have quite well established method that is already Mani started introducing and, but for bias is actually a very problematic thing because one and one really has to be very careful because all these when we build these predictive models we use machine learning algorithms that we train.

So, and they will only give us back the answer to what the kind of data we used to train them. So, if we choose the control samples, for example, in a wrong way then that is then it will not give us something that generalizes. And this is the bias especially is really problematic. And one should spend a lot of time on trying to think about that.

(Refer Slide Time: 18:25)




I hope today you learned that how proteogenome analysis could provide you very useful and novel insights. In addition to providing RNA to protein correlation, the proteogenomics can also provide information on the association between two or more genes or proteins. The network and pathway analysis can provide information on how the presence of a protein or gene influences the expression of other protein or gene.

A comprehensive understanding on the concepts such as predictive analysis, pathway enrichment, mutation and signaling as well as marker selection will help you appreciate the role of proteogenomics in accelerating clinical research.

Thank you.