

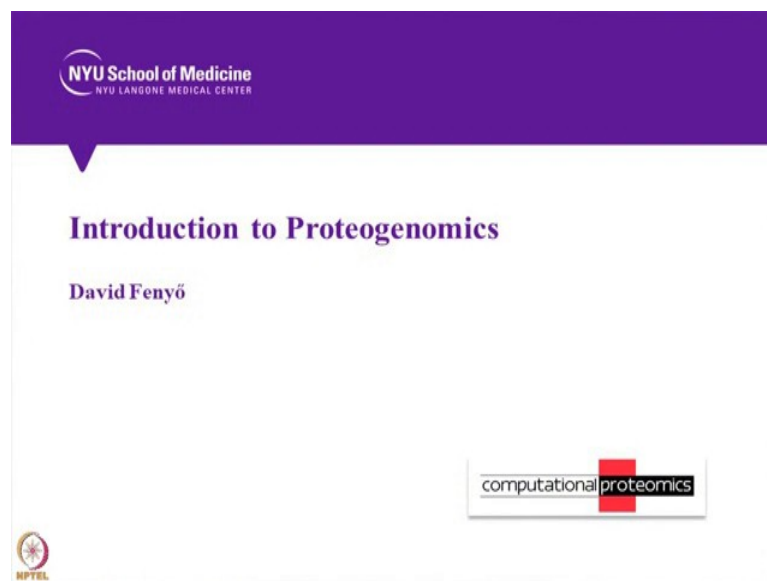
Introduction to Proteogenomics
Prof. Sanjeeva Srivastava
Dr. David Fenyo
Department of Biosciences and Bioengineering
Department of Biochemistry and Molecular Pharmacology
Indian Institute of Technology, Bombay
Institute for Systems Genetics

Lecture - 41
Introduction to Proteogenomics - I

Welcome to MOOC course on Introduction to Proteogenomics. In the course so far, the emphasis was to give you the good understanding of genomics and proteomics, different basic concepts, and tools available to do data analysis. But, now we will shift gear and move on how to integrate data from both genomics and proteomics in the form of proteogenomics, because neither the genomics or proteomics platforms could provide you the complete picture.

Proteogenomic technologies, proteogenomic tools have power to combine various genomic approaches along with proteomic data sets to provide a comprehensive, very broad overview at transcriptional and transitional level. In today's lecture Professor David Fenyo will introduce you to the basic concepts of proteogenomics. So, let us welcome Dr. Fenyo for his lecture.

(Refer Slide Time: 01:29)

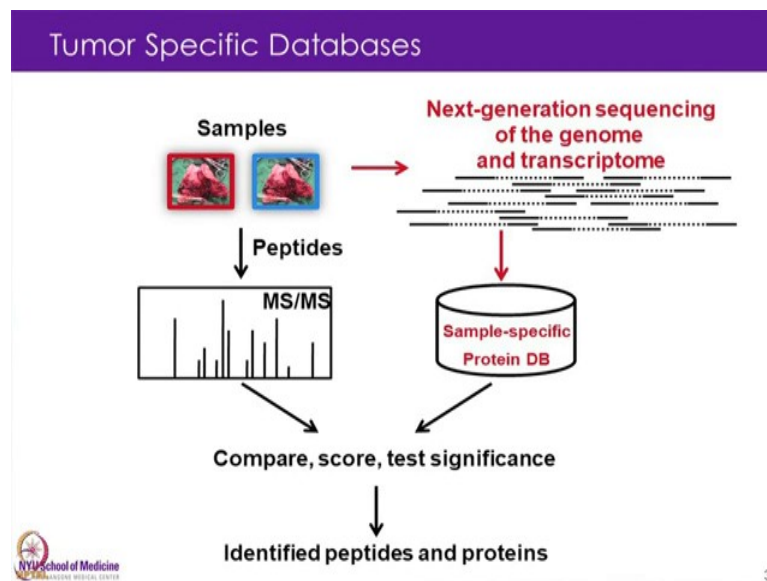


The slide thumbnail features a purple header with the NYU School of Medicine logo and text. Below the header, the title 'Introduction to Proteogenomics' and the name 'David Fenyo' are displayed in purple. At the bottom right, there is a logo for 'computational proteomics' with a red cross symbol. A small NPTL logo is visible in the bottom left corner.

So, far you have had first an introduction to genomics, then to proteomics, and the also to machine learning; and now I will try to give an introduction to proteogenomics, where we combine the genomics and proteomics and see what additional things we can do, when we have both kinds of data. So, as you heard from Karl, when we often, when we want to the first step in proteomics is to identify peptides and the proteins.

And so, one very important thing with this database search is that, if the protein, the exact protein sequence that we are looking for is not in the database we would not be able to find it. So, what one can do then, and of course, we know that in cancer there are a lot of changes in the genome; and then these could lead to changes in protein sequence. But if you just use the reference protein sequence database, we would not be able to identify this.

(Refer Slide Time: 03:02)



And, but now since we have both genomics and proteomics data, we can use the genomics data to modify your database, to add in all the effects of the genomic changes. So, that is and then we make a more comprehensive database that then we should be able to see the specific changes that happen in this tumor.

(Refer Slide Time: 03:28)

Example of variant peptide

Protein: NP_001138550 zinc finger protein 805 isoform 2 [Homo sapiens]
Genome location: chr19:57764586+ 1485 0
DNA Variant: G183A
Protein Variant: V62I

MQGERLRPGLDSQKEKLP GKMSPKHDGLGTADSVCSRIIQDRVSLGDDVHDCDSHG
SGKNPVIOEENIEKCNECEKVFNKKRLLARHERIHSGVKPYECTECGKTFISKSTY
LLQHHMVHTGEKPKYKMECGKAFNRKSHLTQHQRHSGEKPYKCECGKAFTHRST
FVLHNRSHTEGKPFVCKEKGKAFDRPGFIRHYIIHSGENPYECFECGKVFKHRSY
LMWHQQTHTGEKPYECSECGKAFCEAAALIHYYVIHTGEKPFCELECGKAFNHRSY
LKRHQRHTEGKPYVCSECGKAFTHCSTFILHKRAHTGEKPFCECKEKGKAFSNRAD
LIRHFSIHTGEKPYECMECGKAFNRRSGLTRHQRHSGEKPYECIECGKTFWCSTN
LIRHSIIHTGEKPYECSECGKAFSRSSSLTQHQRMHTGRNPISVTDVGRPFPTSGQT
LLLGKNFLNVTTEENLLQEEASYMASDRTYQRETQVSSSL

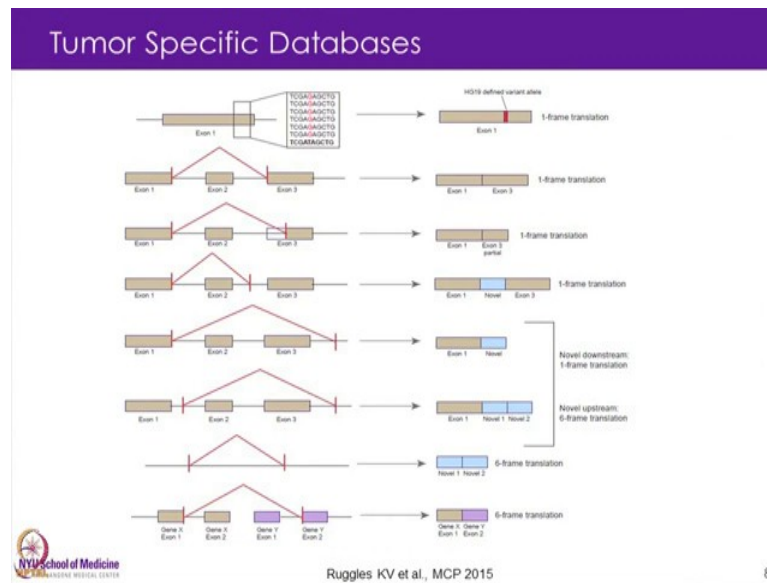
5

And so, for example, just to look at a few examples, if we have a single nucleotide change, like in this case here we have one base that is changing that then can lead to a change in an amino acid. So, this is just one example of, in this protein we have a wonder in the DNA we observe that the G at position 183 changes to an A; which then leads to that this valine at position 62 in the reference database changes to an isoleucine.

So, the tryptic peptide is underlined where we have the valine. So, the change we see is that, we then instead get modified peptide that then we will have a different mass and a different fragmentation spectrum. So, what we, but if we go through and take all the single nucleotide variants that we see in the genome; and see which extra tryptic peptides we get, we can then make a larger database. And so, this is probably the simplest change on the genomic level that then propagates to the proteome, we can have more complex changes.

So, for example, in this case we have a more dramatic change where this on the genome, the DNA this C at position 155 changes to A; and that means, that the tyrosine changes to a stop codon. So, what will happen then is that, most of the protein will not be able to be produced in our sample in the tumor. So, we see we can only expect to see peptides from the first part. And so, of course, this is much more difficult case in the sense that, the in proteomics our coverage is limited. So, if we do not see something that is not proof that it is not there; so but, that is one example of things that can change.

(Refer Slide Time: 05:54)



And so, what the people do then is to, sort of evaluate all the possible effects that we can have and Kelly is going to talk more about what kind of changes or; but I will just quickly say that most of the different ones are related to that splice variants that we can get on the from the RNA sequencing. So, this one up here is just simply that we have three exons and according to reference database exon 1 is connected to 2 which is connected to 3; but then in the RNA seq we also see that, exon 1 is connected directly to exon 3.

So, that is, but we also see a lot of other cases where they get connections from the middle of an exon from an intron and so on. And so, Kelly's going to talk more in detail about these things; but what you can expect and how to create these tumor specific databases using combination of usually exome sequencing where we get two variants and insertions deletions and their RNA seq where we get the different splice variants. So, what are the effects then that we see from these genomic changes?

(Refer Slide Time: 07:28)

Effects of Sequence Variation on the Proteome

- Protein sequence changes
- A modification site is changed
- Protein sequence does not change but the protein level increases or decreases

NYU School of Medicine
9

So, it can be the protein sequence; can change as we looked at these two examples. We can also have that the modification site is changing and, but we can also in a lot of cases have mutation that does not change the protein sequence, but changes its level.

So, it is either from a mutation you get more of the protein or less; and the same for modifications, that mutations can lead to both increasing and decreasing of the mutations. So, I have a quiz. So, why do we care, why what do we use these. So, now, we make a big catalogue of modifications that we see on the genomic level and that we also see on the proteomic level. So, anyone have any suggestions for what, why, what do we do after that, after we made this catalog, no yeah.

Student: Like if you have that catalogue where we can test out sample data set and we can find their net variants and we can cross check with the catalogue, whether this variance is specific to cancer or not.

Yeah. So, we can look downstream and see, if this it I mean both that it is correlated with cancer or that maybe it happens in a tumor suppressor gene or an oncogene that changes things yes. And an another thing that we can of course look at is, if these peptides are only present in cancer and not in normal samples; then we can try to target them and see if that we can for example, I try to activate the immune system to attack cells that display these peptides or in other ways do targeting and there are examples of treatments like that already.

Student: Or we can also use this catalog to create a database with which we can use for proteomics mass spec search.

Yeah, no, so that is what we do, yeah. So, yeah, but I was thinking of after that, after it done, yeah.

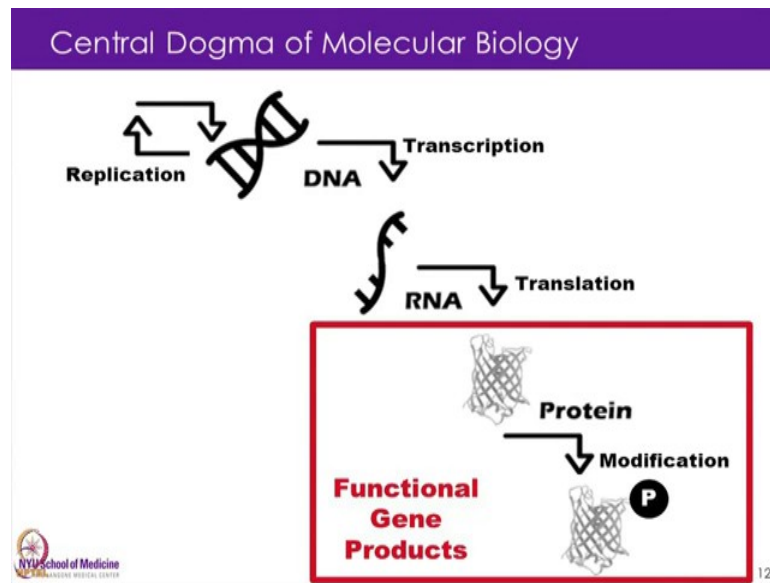
Student: My question is are you saying, what do you mean by peptide specific to cancer.

Yes that is what I was thinking. So, I to right, so either peptides specific to the tumor or that the protein is activated that is usually not active that is maybe not expressed at all in most cells, but gets because of mutations gets expressed. So, that is another thing that we can. So, for example, in breast cancer HER 2 would be an example of that; in most cells they have very low amounts, but in a subtype of breast tumors, we have it is very highly expressed on the surface of the cancer cells. So, we can then target, we have targeted treatment for that.

Student: But we are doing the mass spec analysis. So, we are getting list of peptides. So, like if we are saying that a protein is cancer specific for that protein is also available with a normal sample one. So, I would like, I do not know; but how we can say any peptide protein, we can say this is cancer specific or not.

I mean yeah. So, that is depends on from your experimental design. So, if you for example, you will have to then analyze a lot of tumors and also normal samples to see that, in most of the normal samples you do not see it. And I mean prefer I mean what one can do is for example, look at studies like GTEx that looks at a lot of normal samples because in for good treatment you want that there is no expression in any of the organs pretty much.

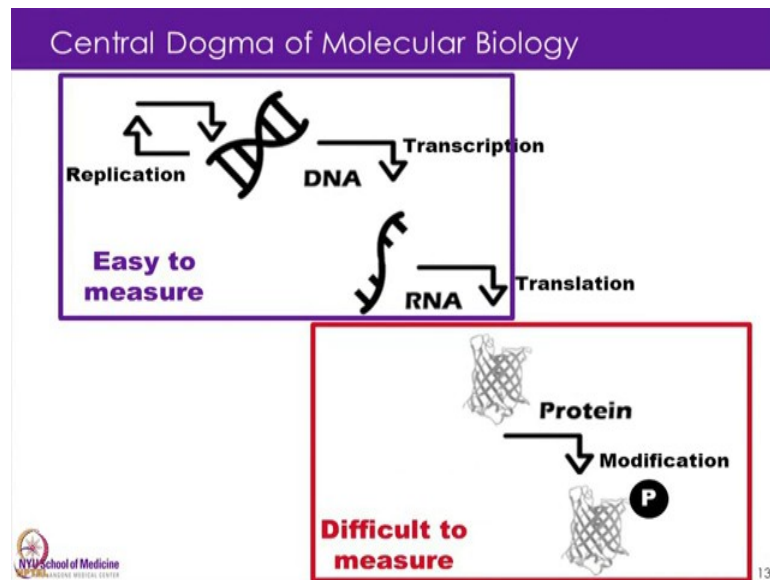
(Refer Slide Time: 12:14)



And you can, so you can use these public datasets to compare; another thing that we can do is to go back to the central dogma of the molecular biology. Since we now are measuring these different kinds of molecules. So, we are doing whole genome sequencing, we have RNA seq data, we measure how the protein levels and we measure how much modifications we have.

And so, that is in CPTAC that is what we usually do in some time, for some project we have other types of measurements, but these at least basic measurements that we have. So, we can look a little bit at what, how these relationships are. So, for one as I was already mentioned several times that, the I mean the proteins and the modifications are the functional gene products. So, that is what makes the phenotype.

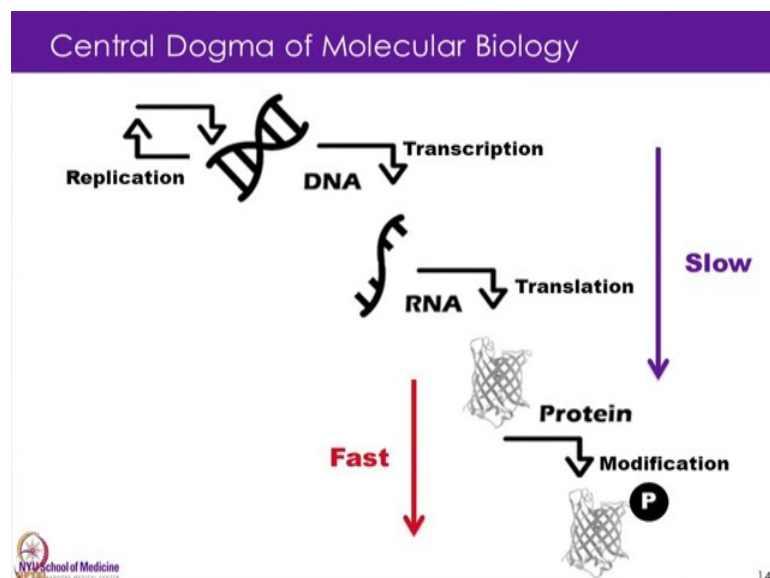
(Refer Slide Time: 13:17)



And, but then of course, it is much easier to measure DNA and RNA there are much more automated methods, and that still even though this been a lot of improvement, it still rather much more difficult to measure proteins and modified proteins.

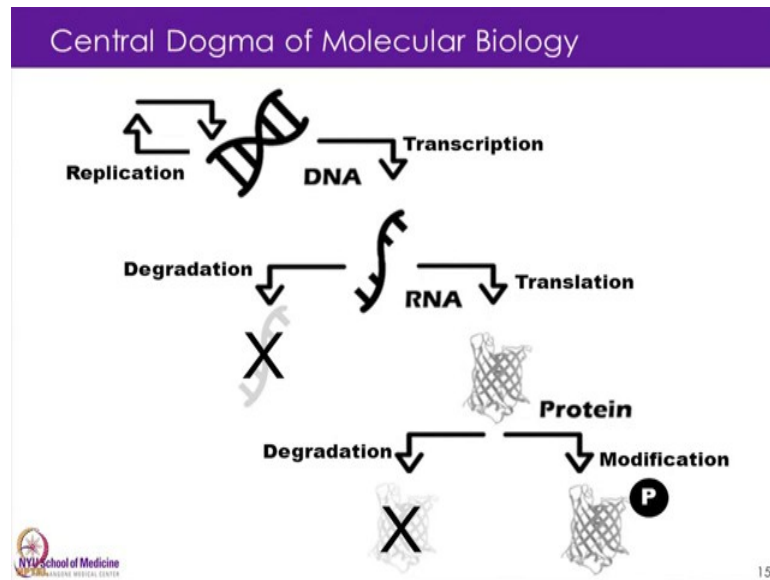
And so, it is often one can probably have in most studies, where one only looks at for example, RNA seq of tumors, there are these are many more samples are analyzed. And, but the, so then the question is ok. So, it is cheaper and faster to measure, RNA it can we just measure RNA instead and not worry about the proteins and their modifications.

(Refer Slide Time: 14:17)



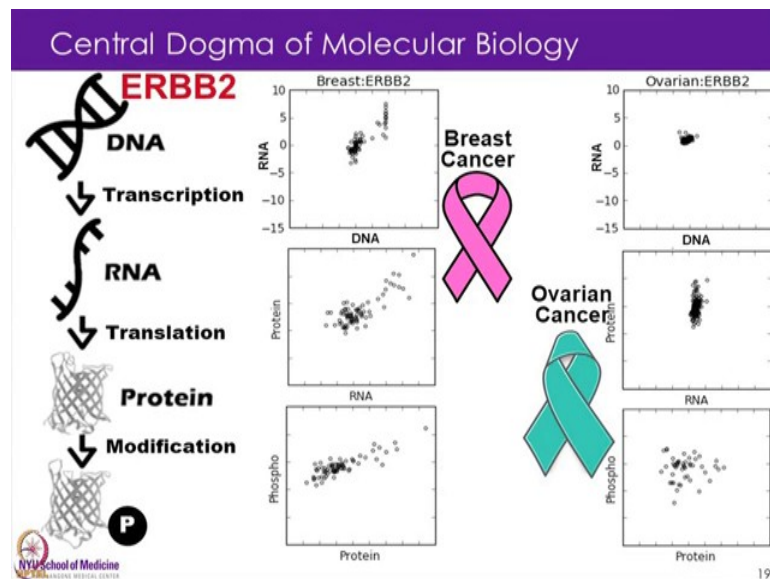
So, there are many arguments for why that is not a good idea, and I mean first of all there is a lot of additional regulation then this; the other thing is that this process, the transcription and translation is rather slow, but the modifications can be done, both added and removed very fast.

(Refer Slide Time: 14:38)



And then we have different degradation rates, usually for RNA and proteins. So, we should not be surprised that there are substantial differences between RNA and protein measurements.

(Refer Slide Time: 14:56)

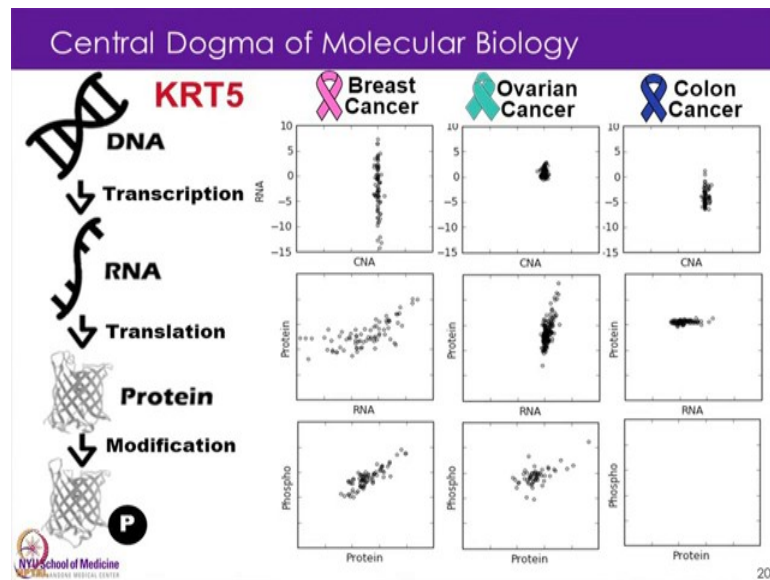


So, if we just look at one example. So, this is in breast cancer it is ERBB 2. So, if we look here, we have copy number and RNA levels. So, we see that, there are is a group of samples that have a higher copy number of this gene or there will be two. And of course, this is the HER 2 sub type and we see that that is correlated with transcript levels. So, we have more copies of the gene, we have, we will have more RNA. So, we have a nice correlation there. So, if we go down and look at the translation, we see that also RNA is correlated with the protein here. So, then we have many transcripts, we also have more protein.

And finally, if we look at one of the phosphorylation sites, it is also very highly correlated with the protein levels. And so, that is what, we at if there is no additional regulation that is what we would expect. So, we have more copies of the gene, we get more transcripts and more proteins; but if we look in the same gene in this example in another tumor type. So, in ovarian cancer, we first of all we do not have really any copy number changes, all the samples have the same copy number, they pretty much have the same transcript levels also.

And, but even though there is the same transcript level, we see quite a range of variation of the protein. And almost it is a little bit smaller than the range of proteins in breast cancer; but it is definitely very comparable. And finally, we do not see any correlation between protein and the same phosphopeptide as we see; so very strong correlation here. So, these are observations that we see and we can look at different genes.

(Refer Slide Time: 17:19)

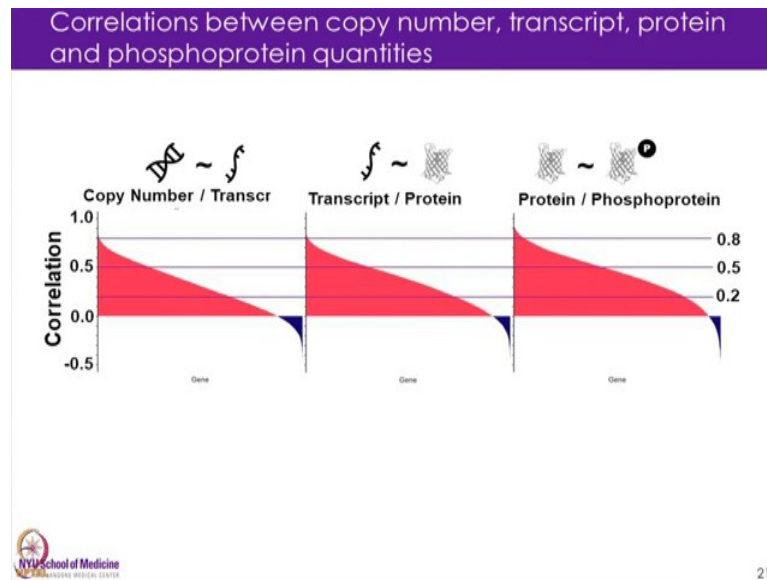


And we see large differences in between different tumor types, between different subtypes. Just to look at one more example, this is one of the keratins. So, here we do not have any. So, we are looking at Breast cancer, Ovarian cancer, and Colon cancer and we do not see any changes in copy number in any of the tumors; we see, but see very large range, difference in range of the RNA levels.

So, in breast cancer we have a very large range, in ovarian cancer almost no change in transcript levels; and in the colon cancer somewhere in between. And then if we look at comparing looking at the translation, so between RNA and protein; breast cancer again we see correlation, and in we see some correlation, but it is the range of RNA is very small for ovarian cancer. And even though we have quite a bit of variation of the transcript levels, there is no protein levels are constant.

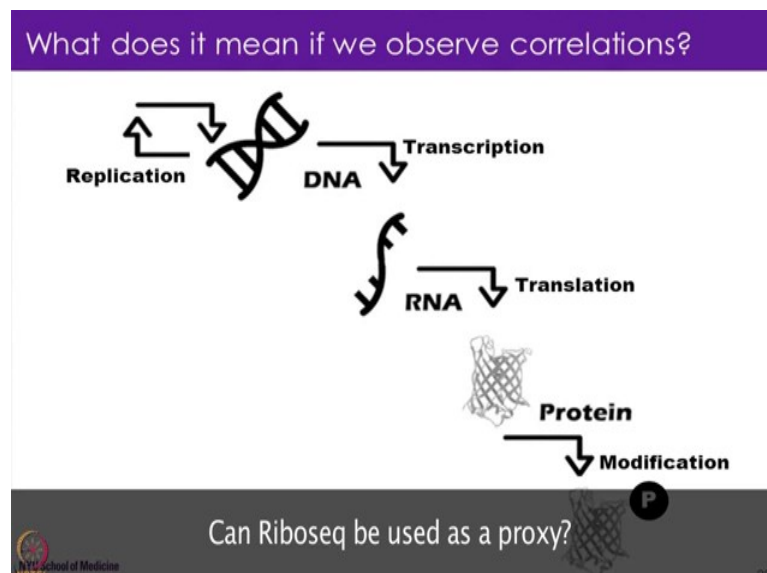
And so, we see that, in this case we have very different regulation of both the transcript and the protein.

(Refer Slide Time: 18:35)



And so, if we look, so these are just two examples. So, now, if we look more globally, we see that there is a wide range of, so this is copy number transcript; we see that there are some that are highly correlated others are, there is a even anti correlated and we see this wide range, and for every comparison both transcript protein and protein phosphoprotein.

(Refer Slide Time: 19:06)



So, on, so then, so what does it mean now? So, we see, if we see a correlation in one case, we do not see a correlation what do we, how do we start thinking about this; any suggestions?

Maybe we can start, let us say what to we say if we, see a color what any yeah, any among.

Student: So, the reason could be the DNA is not transcribed into RNA or the some part of the DNA is only transcribed into RNA. And the RNA is not, might be the RNA is translated into protein, some of the protein has been degraded and some of them is used in downstream processing.

Yeah.

Yes. So, and they can also have a case where the proteins are produced somewhere else in the body, like there will also be see a lot of blood proteins that will be in the tumor; but there is no RNA there, because it is being produced somewhere else. So, there are of all these things and we have the degradation. So, we have a lot of different things going on.

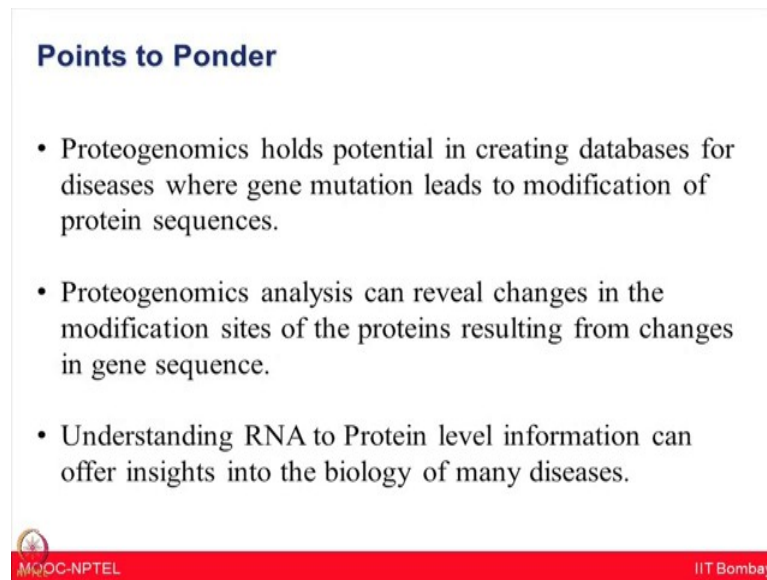
Student: Ribo Seq be used as a proxy for translation .

Yes we can that is.

Student: as it uses.

Yeah. So, Ribo Seq would be somewhere in between here, it would measure the actual translation. It would not measure the amount of protein that is there, but it would measure how much is actually translated at the moment. So, that yeah, that is definitely that would be going one in the step in between the RNA and the protein.

(Refer Slide Time: 20:56)



Points to Ponder

- Proteogenomics holds potential in creating databases for diseases where gene mutation leads to modification of protein sequences.
- Proteogenomics analysis can reveal changes in the modification sites of the proteins resulting from changes in gene sequence.
- Understanding RNA to Protein level information can offer insights into the biology of many diseases.

MOOC-NPTEL IIT Bombay

In today's lecture you got a very broad detailed glimpse of how proteogenomics could help to reduce changes in protein sequences. Due to the mutations in the gene, how to identify the changes in modification sites of proteins; also in which way proteogenomics could help to understand the changes in the level of protein expression, due to change in the gene sequences.

In diseases which are characterized by the changes in the protein sequence due to mutations at the genetic level, the proteogenomic analysis could help to provide us the development of disease a specific databases; where the modified protein sequence information could be made available. Taking the example of ERBB 2 and using proteogenomics, Dr. Fenyo showed that there exists a clear correlation between RNA, protein and phospho protein expression in breast cancer.

However, same did not hold true for ovarian cancer cases, as there was only correlation between RNA and protein, but not the phospho protein levels. So, you can see that you know there is no clear pattern, depending on each disease, the specific context, the correlations could actually vary; and therefore, these analysis on individual data set by looking at a specific questions are very relevant. Though proteogenomic studies may not always show a direct correlation at all the levels, they still offer and provide information, which could be used to answer questions which are very relevant to disease pathobiology.

In the next lecture you will be introduced to a few more concepts of proteogenomics in clinical studies by Dr. David Fenyo.

Thank you.