An Introduction to Proteogenomics

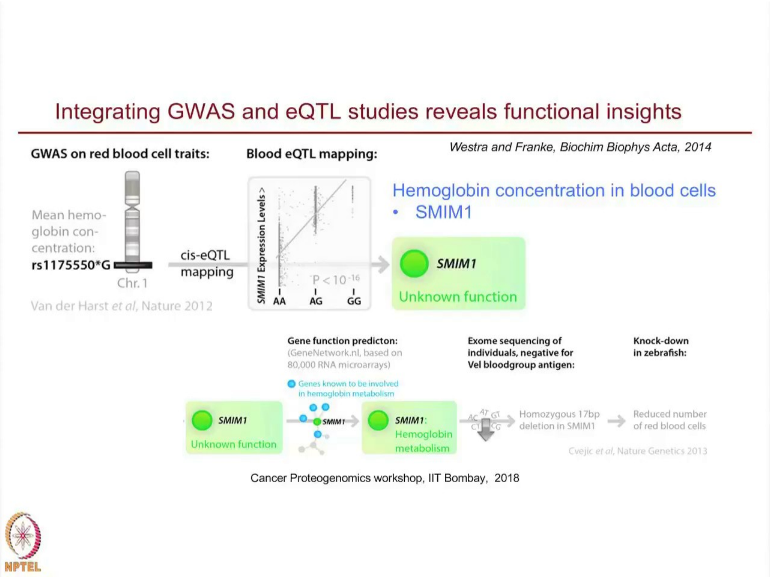
Dr. Sanjeeva Srivastava
Dr. Bing Zhang
Department of Biosciences and Bioengineering
Baylor College of Medicine
Indian Institute of Technology, Bombay

Lecture - 09
Genotype, Gene Expression and Phenotype - Part II

Welcome to MOOC course on Introduction to Proteogenomics. In the last lecture, Dr. Bing Zhang gave you very lucid elucidation of studying polymorphism. The QTL studies help to understand the effect of genetic variant on another gene located on the same or different chromosomes. GWAS studies help in identifying whether a particular gene variant from the entire DNA set could be correlated to a variation in the phenotype.

In this lecture, you are also introduced the single nucleotide polymorphism or SNPs and genome wide association studies or GWAS studies. Today's lecture from Professor Bing Zhang will be another effort to explain the power of integrating expression, quantitative trait loci study or eQTLs with Genome Wide Association Studies or GWAS. So, let us welcome Dr. Bing Zhang for today's lecture.

(Refer Slide Time: 01:31)



Integrating GWAS and eQTL studies reveals functional insights

Westra and Franke, Biochim Biophys Acta, 2014

Hemoglobin concentration in blood cells
- SMIM1

GWAS on red blood cell traits:
Blood eQTL mapping:
Mean hemo-globin con-centration:
rs1175550*G
Chr. 1
Van der Harst et al, Nature 2012
cis-eQTL mapping
SMIM1 Expression Levels >
P < 10-16
AA
AG
GG
SMIM1
Unknown function
Gene function predicton:
(GeneNetwork.nl, based on 80,000 RNA microarrays)
Genes known to be involved in hemoglobin metabolism
Exome sequencing of individuals, negative for Vel bloodgroup antigen:
Knock-down in zebrafish:
SMIM1
Unknown function
SMIM1
SMIM1: Hemoglobin metabolism
AC AT GT CT CG
Homozygous 17bp deletion in SMIM1
Reduced number of red blood cells
Cvejic et al, Nature Genetics 2013

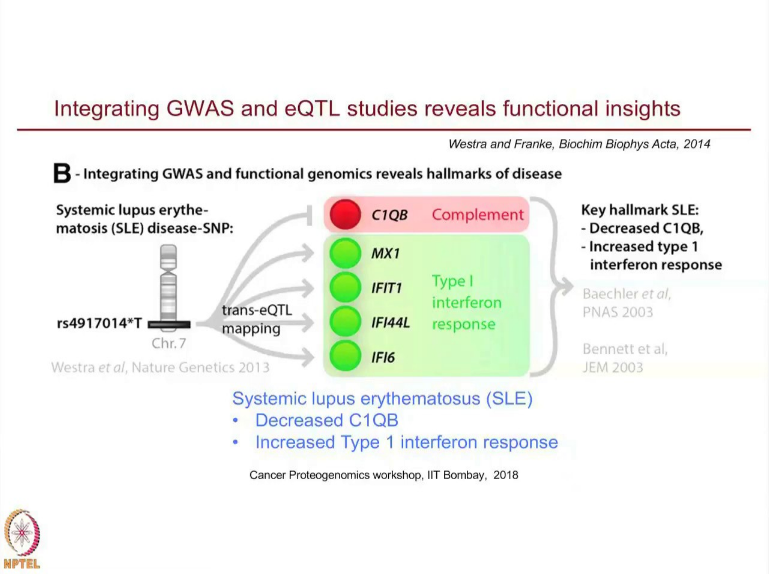
Cancer Proteogenomics workshop, IIT Bombay, 2018

So, let us look at some of the examples. The first example is; this group, so basically they know this SNP is associated with hemoglobin concentration in blood cells and then but they do not know how why this happened and of course, we are interesting interested in how this could happen right. So, they did eQTL analysis and through the cis-eQTL analysis they found this SNP this exact SNP has a good association with gene expression right next to the SNP the gene is called the SMIM 1.

So, and then you can generate a hypothesis that the SNP probably affects the this hemoglobin concentration through this gene expression right but that is just the hypothesis but at that time, there is nothing known about this gene but what they did was to through some gene network analysis, I am going to talk about in later in the next lecture and then they found this gene is particularly associated with hemoglobin metabolism genes in the network indicating it might have the similar function in that process and then through some functional experiments in human and also in model organisms they demonstrated indeed and this gene is involved of this protein deletion experiments.

They found the causal relationship between this gene expression and the phenotype they are interested in. So, this shows the power of doing the cis eQTL analysis.

(Refer Slide Time: 03:11)



Integrating GWAS and eQTL studies reveals functional insights

Westra and Franke, Biochim Biophys Acta, 2014

B - Integrating GWAS and functional genomics reveals hallmarks of disease

Systemic lupus erythematosus (SLE)

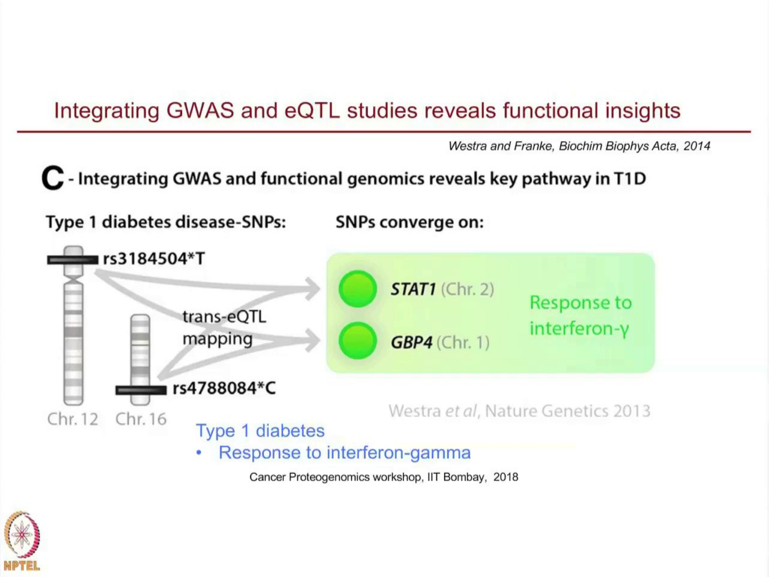
- Decreased C1QB
- Increased Type 1 interferon response

Cancer Proteogenomics workshop, IIT Bombay, 2018

And also we can think about the trans-eQTL analysis, let us say this SNP is associated with the Systemic Lupus Erythematosus or SLE this disease but at that time nobody knows how this SNP is associated with that disease and through eQTL mapping they found that this SNP is associated with multiple genes and one gene is C1QB which decrease the expression of C1QB and also increase the expression of multiple genes involving the type 1 interferon response pathway.

So, this helped some of them to think that maybe this SNP has a effect because it has through the alter the expression of these genes and the interesting sign is that and the decrease the C1QB and the increased type 1 interferon response has been a hallmark, I mean it is already known the disease has this phenotype. So, but now we know which genes are mediating this impact and similarly through the traditional GWAS analysis.

(Refer Slide Time: 04:31)



Integrating GWAS and eQTL studies reveals functional insights

Westra and Franke, Biochim Biophys Acta, 2014

C - Integrating GWAS and functional genomics reveals key pathway in T1D

Type 1 diabetes
- Response to interferon-gamma

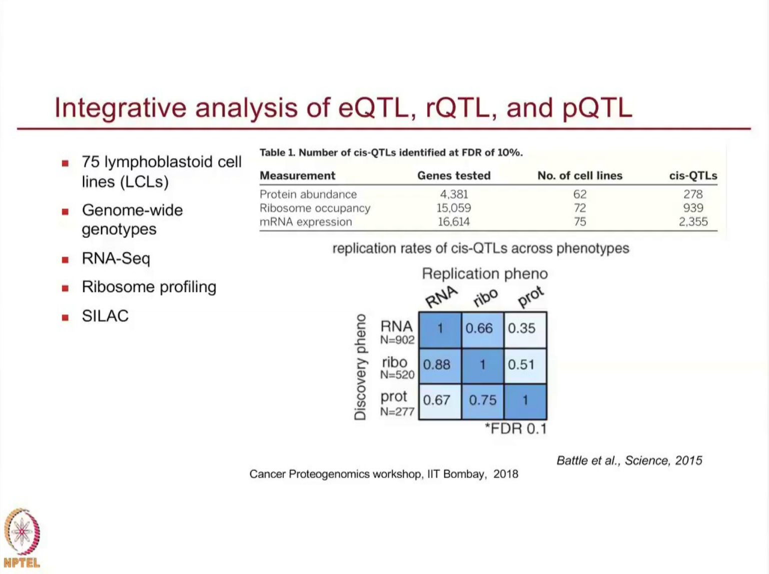
Cancer Proteogenomics workshop, IIT Bombay, 2018

There multiple SNPs that has been associated with type 1 diabetes but again we do not know how these SNPs execute the effect but through eQTL mapping this group found that although they these SNPs are located at different locations of the genome but they converge. So, they alter or they control the expression of the same genes that means, different alterations genotypes converge through the same gene expression alterations and eventually change the phenotype.

So, this shows I mean examples basically show how you canintegrate the GWAS analysis through getting the DNA sequence and also the for example, RNA-Seq to get the gene expression and the combining them and of course the phenotype and then you will be able to not only associate the SNPs with disease phenotype but you also know how what are the gene expressions that are involved in this.

You would think the protein expression is also important and we now know that the protein does not necessarily or perfectly correlate with gene expression. I mean if you do this at the protein level you probably will also get additional new insights right but because of the technology is less matured or lags behind the RNA-Seq those technologies the relatively fewer pQTL type of analysis but this audience I think we should think about this approach in order to integrate the GWAS study for example, is pQTL studies.

(Refer Slide Time: 06:08)



Integrative analysis of eQTL, rQTL, and pQTL

- 75 lymphoblastoid cell lines (LCLs)
- Genome-wide genotypes
- RNA-Seq
- Ribosome profiling
- SILAC

Table 1. Number of cis-QTLs identified at FDR of 10%.

Measurement	Genes tested	No. of cell lines	cis-QTLs
Protein abundance	4,381	62	278
Ribosome occupancy	15,059	72	939
mRNA expression	16,614	75	2,355

replication rates of cis-QTLs across phenotypes

Discovery pheno \ Replication pheno	RNA	ribo	prot
RNA N=902	1	0.66	0.35
ribo N=520	0.88	1	0.51
prot N=277	0.67	0.75	1

*FDR 0.1

Battle et al., Science, 2015

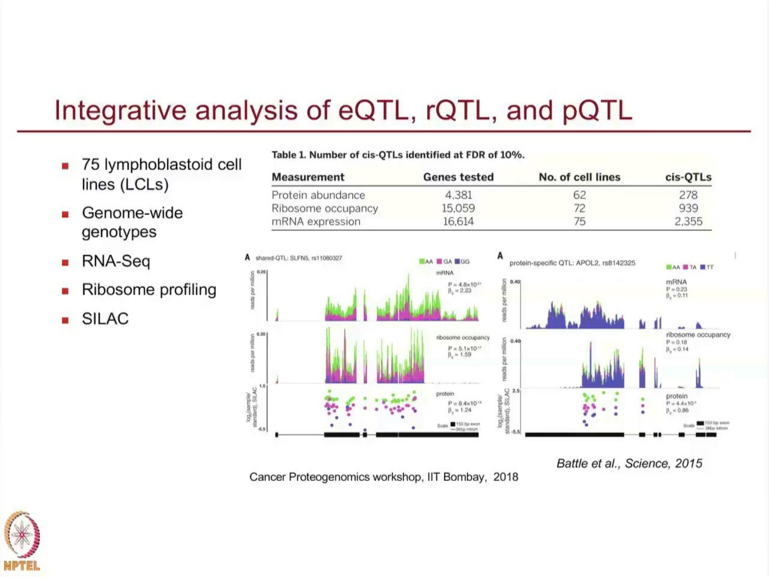
Cancer Proteogenomics workshop, IIT Bombay, 2018

And the actually there are some groups started to do this and the in this study so basically they look at the 75 lymphoblastoid cell lines and then they did the genome-wide genotype study, and the RNA-Seq, ribosome profiling through Ribo-Seq and the SILAC proteomics experiments. So, now you have the genotype, you have the mRNA expression, and the ribosome occupancy and the protein expression.

So, and they were able to identify, the focussed only the cis eQTL, rQTL and the pQTLs, and they found hundreds of and sometimes thousands of this cis-QTLs and then they asked the I mean if I found cis-eQTL what is a likelihood I am going to find the same like rQTL and pQTL like meaning the SNP controls the gene expression also the mRNA expression, also the ribosome occupancy and the protein expression that means, consistent right.

But they found for example, if you look at the RNA of course, RNA-RNA is one, but only 88 percent of the RNA the eQTLs can be replicated in the ribo-seq experiment or the rQTL level and only 67 percent can be replicated at the protein level the pQTLs. So, this basically indicates not all the eQTL effects are reproducible as a pQTL level. It is the same manner if you have a protein QTL, only 35 percent of them can be observed at the eQTL that is kind of consistent with our observations the mRNA and the protein expression are not perfectly coordinated, but this also indicate that for example, SNP might have a effect only at the protein level but not necessarily at RNA level. This could happen for example, if the SNP is in a region that is controlling translation of the protein rather than the transcription of the gene.

(Refer Slide Time: 08:29)



Integrative analysis of eQTL, rQTL, and pQTL

- 75 lymphoblastoid cell lines (LCLs)
- Genome-wide genotypes
- RNA-Seq
- Ribosome profiling
- SILAC

Table 1. Number of cis-QTLs identified at FDR of 10%.

Measurement	Genes tested	No. of cell lines	cis-QTLs
Protein abundance	4,381	62	278
Ribosome occupancy	15,059	72	939
mRNA expression	16,614	75	2,355

Battle et al., Science, 2015

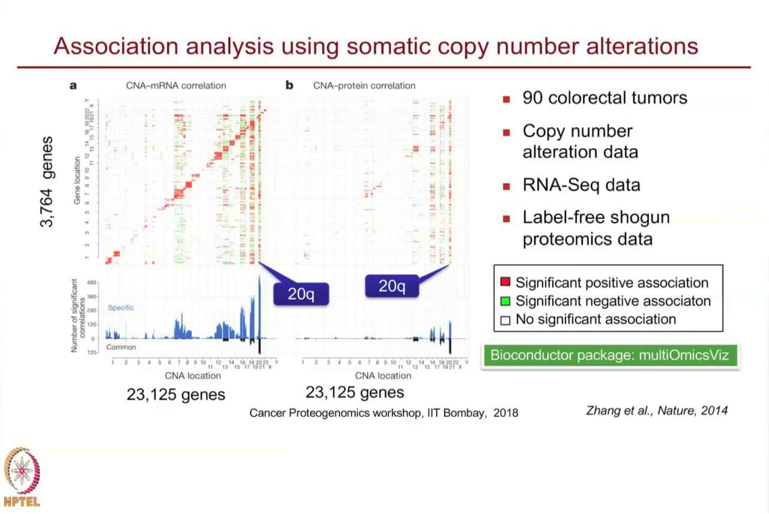
Cancer Proteogenomics workshop, IIT Bombay, 2018

So, but we can look at these examples; for example, in this case, we can see the effect is consistent at the protein level or the ribosome occupancy level and RNA level, because the three genotypes the effect is a kind of colored by three different colors. We can see the difference can be observed at the RNA level, the ribosome occupancy level and the protein level. But in this case, we can see the effect is only observable at the protein level but not at ribosome and the RNA level that means this SNP is affecting the translation of this protein or the maybe the stability of this protein without affecting its RNA and ribosome translation ribosome occupancy.

So, this indicates if we do both eQTL analysis and pQTL analysis, it will give us more information than just doing the eQTL analysis. But for this study particularly I mean they only look at the cis-eQTL and they did not look at the trans-QTLs because as I said I mean when your sample size is small and the trans eQTL is less difficult to observe because the effects is relatively smaller.

And finally, I want to show one example and what we have talked about so far or the about SNPs which are the germline alterations right because this audience is interested in cancer. So, we can borrow the same idea and apply to the cancers studies and in this case we did not look at the SNPs we look at the somatic copy number alterations because in cancers they are lot of chromosome regions that are getting amplified or deleted right and then we can consider that as the genotype or change right and then we want to ask whether the genotype change will affect the gene expression, mRNA expression and the protein expression of that gene or maybe it will affect other genes in the genome.

(Refer Slide Time: 10:38)



Association analysis using somatic copy number alterations

- 90 colorectal tumors
- Copy number alteration data
- RNA-Seq data
- Label-free shogun proteomics data

🟥 Significant positive association
🟩 Significant negative associaton
□ No significant association

Bioconductor package: multiOmicsViz

Cancer Proteogenomics workshop, IIT Bombay, 2018

Zhang et al., Nature, 2014

So, and for this we did analysis in 90 colorectal tumors and then we calculated the copy number alteration for individual genes based on just SNP array data and then we get RNA-Seq data. So, basically we have the mRNA abundance for each gene and the we did the label-free shogun proteomics in this study and got the protein abundance for the genes and we have only focused on the genes have both mRNA and the protein measurement. This correspond to close to 4000 genes at that time and but for copy number we have the data for 23,000 genes.

And now for each copy number data and all the mRNAs we can calculate but in this case the genotype is also continuous right because it is a copy number measurement and then we can calculate the Pearson correlation between the genotype and the mRNA expression and the if there is a significant positive correlation we put a red dot here, that means the copy number change of this gene will affect the mRNA abundance of that gene. And then if there is a significant negative association, we put a green dot here and if there is no significant, then we leave it blank. And this is a plot we get from this analysis, and the interesting sign is we observed two types of interesting patterns.

One is at the time we see a very strong diagonal pattern and the other type of pattern is these stripes I mean in the vertical lines. Let us say if we use this gene as an example, this locus as a example. If it this copy number amplification will increase the mRNA expression of the gene in that locus, so you are going to have a significant copy number change and the mRNA expression change and then you are going to get a red dot because here all the genes are ordered based on the chromosome location and the here it is also based on the same chromosome location.

So, if it is a cis effect we are going to see a red dot here right, but let us see this gene is a transcription factor and the DNA amplification of the transcription factor not only cause higher abundance of that gene, but this transcription factor will in turn activate or deactivate a lot of other genes. So, then you also see a genome wide effect of that DNA copy number change.

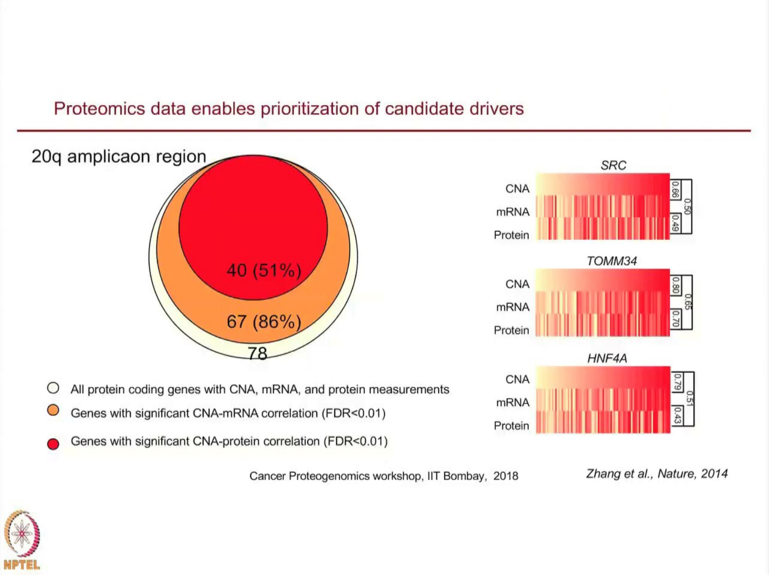
So, we call this the cis band on the diagonal because it is the copy number change that will alter the mRNA abundance of the same gene but we also see the genome wide these vertical bands, these are the trans band meaning copy number change at this position may affect a lot of other genes in the genome.

And then we also look at the protein data. So, here is a correlation between copy number and the protein. We can see both those cis and the trans bands getting weaker I mean we get kind of similar patterns but it is weaker. This indicate not all the impact at the mRNA level can be carried over to the protein level. This is called phenotype dampening meaning there is a reduce of the effect.

For example, if the copy number amplification give you more mRNAs of a certain gene, but this gene does not give you the close advantage of a cancer cell. The protein is not needed, then we do not need to make that extra protein, so that means, there is additional regulation, it is not never that we will remove those effects that is not necessary.

So, and we also sometimes see the copy number of protein correlation only observed at the protein level but not at the corresponding mRNA level. This indicate that copy number change might only affect the protein especially for the trans-effect. So, this is very helpful because now we can look at this plot and we can say ok, there are certain chromosome regions that have particularly strong impact at the global level. For example, this 20q region seems to have a big impact genome wide for mRNA and the protein, and there might be some interesting genes in that chromosome.

(Refer Slide Time: 15:15)



Proteomics data enables prioritization of candidate drivers

○ All protein coding genes with CNA, mRNA, and protein measurements

● Genes with significant CNA-mRNA correlation (FDR<0.01)

● Genes with significant CNA-protein correlation (FDR<0.01)

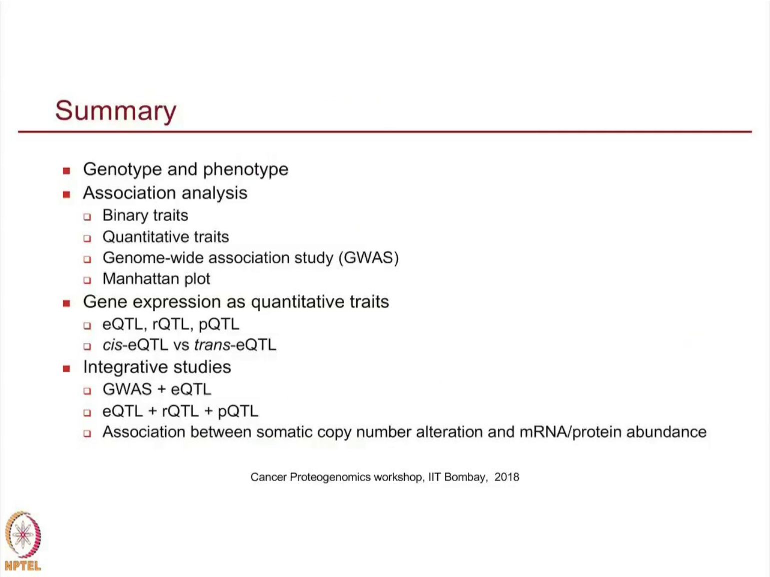
Cancer Proteogenomics workshop, IIT Bombay, 2018

Zhang et al., Nature, 2014

And we also can look at the trans-effect or cis-effect and we can see and this indicate the large circle indicate all the protein-coding genes in this 20q amplification region that we have both mRNA and the protein measurements and this indicates the genes with the good copy number, and the mRNA correlation, and this indicates the genes with good copy number and the protein correlation. So, this help us to narrow down to the genes that the copy number will not only increase the protein or mRNA level, but also increase the protein level, and these are the very likely cancer drivers genes in the region.

And for example through this approach we were able to identify SRC which is a very well established oncogene in colon cancer, and then TOMM34 has also been reported, and we were able to identify and a new candidate driver HNF4A which could be a new discovery to be tested in the future.

(Refer Slide Time: 16:21)



Summary

- Genotype and phenotype
- Association analysis
    - Binary traits
    - Quantitative traits
    - Genome-wide association study (GWAS)
    - Manhattan plot
- Gene expression as quantitative traits
    - eQTL, rQTL, pQTL
    - *cis*-eQTL vs *trans*-eQTL
- Integrative studies
    - GWAS + eQTL
    - eQTL + rQTL + pQTL
    - Association between somatic copy number alteration and mRNA/protein abundance

Cancer Proteogenomics workshop, IIT Bombay, 2018

So, just to give a quick summarized summary of the talk. So, we this talked about the genotype and phenotype, and the using association study to understand the relationships or test the relationships and depending on whether it is a binary trait or quantitative trait, we have to use different types of statistical tests and if we do this at the genome-wide scale, we it is called the GWAS study and we can use Manhattan plot to visualize results and then if we trait the gene expression as a quantitative traits, and then we can do the eQTL, rQTL, pQTL type of analysis.

And this QTLs can be divided into the cis-eQTL and the trans eQTL based on their in positional relationship between the SNP and the genes and we showed a examples that you can indicate GWAS and the eQTL or eQTL, rQTL, pQTL or copy number and the mRNA and the protein expression to understand the relationship between the genotype and gene expression and the phenotype but maybe we can have one or two questions or maybe we can discuss during the lunch.

Student: Copy number analysis slide.

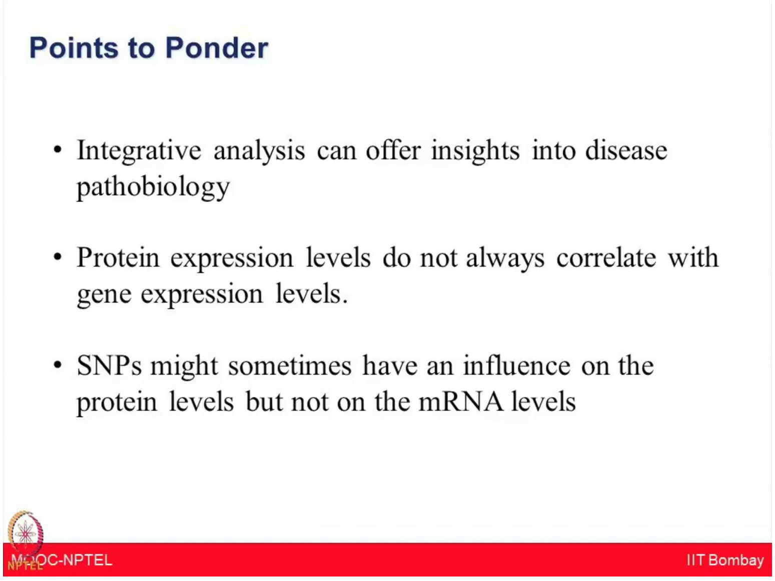
Yes.

Student: Can you please go to that slide? In that the last colum, so can it be like the mRNA degradation, it is not been degraded and which is one of the reason most of the proteins are stable, I mean the transcripts are stable .

So, I think it yeah it could be well, but if you think about the RNA-Seq right it is actually measuring the steady state mRNA abundance. So, I think the degraded RNA will not be made at the RNA-Seq I mean yeah if you consider that measurement is at the steady state and that has already been taken care of but of course, I mean it is kind of dynamic.

So, it might partially reflects that effect but I would think the most of the mRNA measurement we have is for steady state, so that it is already incorporated both RNA generation and the degradation has both being measured in that measurement. But also I mean this is not the abundance of the mRNA or protein right. It is association between the copy number and mRNA or protein. But I think the discrepancy is more likely to be caused by the altered translation or like the protein degradation, the half life of the proteins, yeah.

(Refer Slide Time: 19:18)



Points to Ponder

- Integrative analysis can offer insights into disease pathobiology
- Protein expression levels do not always correlate with gene expression levels.
- SNPs might sometimes have an influence on the protein levels but not on the mRNA levels

MOOC-NPTEL
IIT Bombay

Today’s lecture broadly explained the use of various integrative analysis to understand disease pathobiology. Using examples from published literature cis-eQTL and trans-eQTL mapping integrated with GWAS study were shown to correlate the gene expression with phenotype. Additionally, the integrative analysis of somatic copy number variations and mRNA abundance were seen to be directly correlated. In the next lecture, you will be introduced to the next generation sequencing technology and its application by a industry application scientist.

Thank you.