

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

Lecture - 08
Genotype, Gene Expression and Phenotype - Part I

Welcome to MOOC course on Introduction to Proteogenomics. In the last lectures, Dr. Kelly Ruggles have given you very detailed overview of genomics, transcriptomics and epigenomics. Continuing in the same theme, let us talk about SNPs. SNPs are most common type of gene polymorphism, and they located in the promoter regions of genes thus the bringing the changes in gene expression.

Today's lecture will be given by Dr. Bing Zhang who is a Professor of Molecular and Human Genetics at Baylor's College of Medicine in USA. Prof. Zhang will introduce you to the concepts of DNA polymorphism, and how they bring about the variability in a given population. The lecture also aims to provide an understanding on how genotype could influence the trait of a phenotype. So, let us welcome Dr. Bing Zhang for today's lecture.

(Refer Slide Time: 01:32)

Genotype

- Genotype
 - The unique genetic makeup of an individual organism
 - Encoded in DNA
- DNA polymorphism
 - Difference in the nucleotide sequence between individuals
 - Single nucleotide polymorphism (SNP)
 - A single nucleotide (A, T, C, or G) differs between individuals
 - Most frequent type of DNA polymorphism
 - Deletions
 - Insertions
 - Copy number changes of a given DNA sequence

Reference genome
TCGAGGTATTAAC
TCGAGGTATTAAC
TCGAGGTATTAGC
TCGAGGTATTAAC
TCGAGGTATTAGC
TCTAGGTATTAAC

Biallelic
Major allele: G, 80%
Minor allele: T, 20%

Biallelic
Major allele: A, 60%
Minor allele: G, 40%

Cancer Proteogenomics workshop, IIT Bombay, 2018

So, I will start with the simple definition of genotype. We know that genotype refers to the unique genetic makeup of individual organisms and its encoded in the DNA sequence we know that right. So, the difference in the DNA sequence between individuals is called DNA polymorphism. And the if we look at this example and let us say this sequence, you can probably download lets say human genome sequence and the it is just the short fragment right. We can see some individuals next this guy has exactly the same sequence as a reference, but this guy for example has a different nucleotide at this position. And then if we look at this position and there are two possible alleles one is the G, the other is the T right.

So, and the so this is the biallelic locus meaning there are two different types of alleles. And then the major allele is the G, because it occupies 80 percent of I mean this very small population right, and the minor allele is the T which is only 20 percent one out of 5. And there is another biallelic locus here, where we found two individuals have the G where all others have A; and then this is also a biallelic position and the with the major allelic 60 percent and is the minor allelic 40 percent.

So, this is a very simple type of DNA polymorphism and the because difference only occurs in a single nucleotide it is called single nucleotide polymorphism or a SNP. And this is a very common type of a DNA polymorphism in the genome and the it is very frequent for example, in the human genome probably around 10 million SNPs in the genome. And the, there are other types of DNA polymorphism like deletions, insertions or even copy number alterations; for large fragments.


So, and the phenotype is obviously, the observable traits or characteristics of individuals and the depending on the nature of the phenotype, it could be binary phenotypes meaning the only two possible selections like any disease, you are either diabetes or non-diabetes.

(Refer Slide Time: 04:06)

Phenotype

- Phenotype
 - The observable properties or traits of an organism
 - Binary traits (*e.g.*, diabetes vs non-diabetes)
 - Quantitative traits (continuous, *e.g.*, height)
- Produced by the interaction of the genotype and the environment

Cancer Proteogenomics workshop, IIT Bombay, 2018




Or it could be continuous in nature, like the quantitative traits like our height or weight and these are continuous or quantitative traits. And the phenotype especially if we were interested in this for example, we are interested in the disease phenotypes right and the it is usually the interaction between the genotype and the environment and the interaction. And the genotype plays a very important role in determining one's phenotype.


(Refer Slide Time: 04:37)

Association analysis

- Consider a genetic marker consisting of a single biallelic locus with alleles *a* and *A* (*i.e.*, a SNP), there are three possible genotypes
 - *a/a*
 - *a/A*
 - *A/A*
- Association analysis tests the association between the locus and a trait of interest
 - Binary
 - Quantitative



Cancer Proteogenomics workshop, IIT Bombay, 2018



So that is why a major goal in biomedical research is to understand how the genotype determines the phenotype. And in order to do this a very simple, but powerful way is

through the association analysis. So, in the association analysis what we want to do is if we are interested in a phenotype we want to test, whether there is an association between the genotype and the phenotype. And depending on whether the phenotype is the binary phenotype or a quantitative phenotype, we need different types of statistical tests in order to establish that relationship.

So, let's first look at the binary traits and this is usually referred to as case control studies. I mean for example, what is the case hence the lung diseases control right.

(Refer Slide Time: 05:31)

Association analysis: binary traits (case/control)

Contingency table, genotype count					Test	Contingency table description	Degrees of freedom (d.f.)
Genotype	a/a	A/a	A/A	Total	Genotypic association	2x3 table (a/a, a/A, A/A)	2
Case	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$	Dominant model	2x2 table (a/a, not a/a)	1
Control	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$	Recessive model	2x2 table (A/A, not A/A)	1
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n			

Balding, Nat Rev Genet. 2006
Clark et al., Nat Protoc. 2011

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$$

where $E[n_{ij}] = \frac{n_{i\cdot} n_{\cdot j}}{n}$

Chi-square test

Cancer Proteogenomics workshop, IIT Bombay, 2018

And the if you have a population or a sample size of n and then some of them are disease samples or case samples some of them are control samples right, and then some of them have this genotype and this genotype and this genotype. So, basically you can look at your individuals and the put some numbers into this 2 by 3 tables, and so basically and these are the cases with this small a , small a which is genotype right. You can fill this table very easily, after you if you know all the genotype and the phenotype for that all individuals.

And then we can use the very simple chi-square test in order to test the association between the binary phenotype and the genotype and this is a formula showing how chi-square test can be done right. But sometimes I wonder for example, the major allele might have a dominant effect that means, whether you have a this heterozygous locus or this homozygous locus, you are going to have the same phenotype. So, and then in that

case we can combine these two into the same column in the table and the 2 by 3 table can become a 2 by 2 table right by doing this we actually can power, because we reduced the degree of freedom from 2 to 1. And then if the phenotype is actually the major allele has a dominant effect, we actually can power by doing that.

And on the other side and if under a recessive model and the way would expect and this to a only the a this one will have the same effect, but this two will have the same effect; and then again we can combine those 2 columns into 1. And then we can do the chi-square test. And but sometimes maybe the a there is a additive effect, let us say if whether you are in this genotype or this genotype.

So, let us see the count the number of the minor alleles is kind of linearly associated with the proportion of the case or the disease proportion in the population like this. And then if we simply do the chi-square test that we would not be able to capture that relationship right. So, what can we do? So people have come up with this; it is called the Cochran-Armitage trend test.

(Refer Slide Time: 08:19)

Association analysis: binary traits (case/control)

Contingency table, genotype count					Test	Contingency table description	Degrees of freedom (d.f.)
Genotype	a/a	A/a	A/A	Total	Genotypic association	2x3 table (a/a, a/A, A/A)	2
Case	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$	Dominant model	2x2 table (a/a, not a/a)	1
Control	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$	Recessive model	2x2 table (A/A, not A/A)	1
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n	Cochran-Armitage trend test	2x3 table (a/a, a/A, A/A)	1

Balding, Nat Rev Genet. 2006
Clark et al., Nat Protoc. 2011

$$\chi^2 = \frac{\left[\sum_{i=1}^3 w_i (n_{1i} n_{2i} - n_{2i} n_{1i}) \right]^2}{n_{1\cdot} n_{2\cdot} \left[\sum_{i=1}^3 w_i^2 n_{i\cdot} (n - n_{i\cdot}) - 2 \sum_{j=1}^2 \sum_{i=1}^3 w_i w_j n_{ij} n_{i\cdot} n_{j\cdot} \right]}$$

where $w = (w_1, w_2, w_3)$ are weights chosen to detect particular types of association.

$w=(0,1,2)$: Additive effect
 $w=(0,1,1)$: Dominant model
 $w=(0,0,1)$: Recessive model

Cancer Proteogenomics workshop, IIT Bombay, 2018

So with this test by playing with weights the w is in the formula. So, you will be able to for example, if you set the w as 0, 1, 2; you will be able to test this additive effect and with the formula. And the interesting thing about this, test is that you can change the weight setting to 0, 1, 1 or 0, 0, 1 and then you will be able to also test the dominant model and the recessive model.

So, usually we do not really know with which model SNP or phenotype is determined right and the people one way you can do is to test all the possible models, and then you get the most significant one and the use that for your report.

(Refer Slide Time: 09:10)

Association analysis: binary traits (case/control)

Contingency table, genotype count

Genotype	a/a	A/a	A/A	Total
Case	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Control	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n


Contingency table, allele count

Allele	a	A	Total
Case	m_{11}	m_{12}	$m_{1\cdot}$
Control	m_{21}	m_{22}	$m_{2\cdot}$
Total	$m_{\cdot 1}$	$m_{\cdot 2}$	$m (=2n)$

Test	Contingency table description	Degrees of freedom (d.f.)
Genotypic association	2x3 table (a/a, a/A, A/A)	2
Dominant model	2x2 table (a/a, not a/a)	1
Recessive model	2x2 table (A/A, not A/A)	1
Cochran-Armitage trend test	2x3 table (a/a, a/A, A/A)	1
Allelic association	2x2 table (a,A)	1

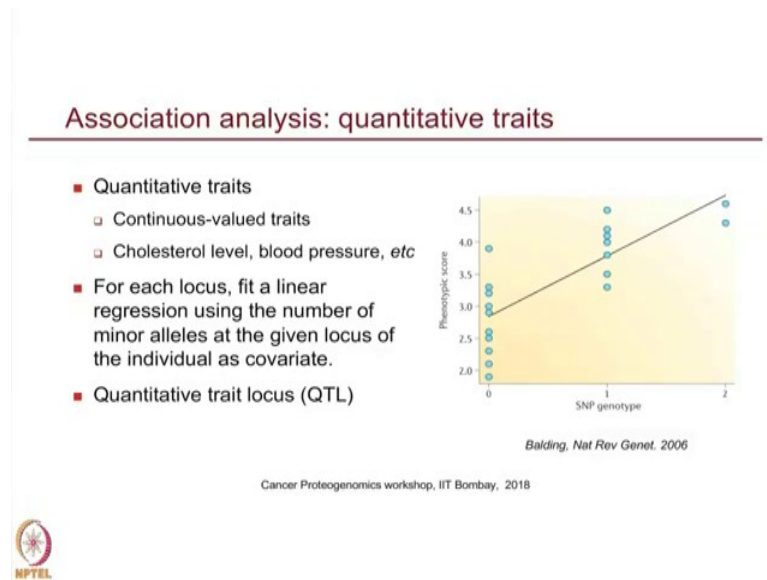
$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$ where $E[n_{ij}] = \frac{n_{i\cdot} n_{\cdot j}}{n}$ Chi-square test

Clark et al., Nat Protoc, 2011 Cancer Proteogenomics workshop, IIT Bombay, 2018



And another way to and deal with this additive effect is you can consider converting the genotype count table into a allele count table. So, basically each person will contribute two alleles to your count table. So, your sample size actually increase from m to two times of n right and then you count if one have heterozygous, then allele locus and then it contribute to both of the rows in this table right, so that way you can still do the chi-square test and the degree of freedom is still one for this, but this also assumes the additive effect. So, those are for the binary traits, but for the continuous or the quantitative traits and for example, the blood pressure or cholesterol levels this type of things.

(Refer Slide Time: 10:03)



We can what we can do is to fit a linear regression model against the data. So, basically is the covariates will be the number of minor alleles individuals and then you correlate with that the continuous measurement of the phenotype for example, blood pressure and then you can test the goodness of fit of that linear regression. And the if you find a locus is actually associated with the quantitative trait, because that locus a quantitative trait locus or QTL and probably you have heard about QTL many times.


And so far we have been talking about just you are interested in one SNP and the one of phenotype, and then you try to reestablish the relationship, but often times I mean you start with the disease of interest or phenotype of interest. And then you do not really know which SNP or which position of the DNA sequence is associated with that phenotype right.

(Refer Slide Time: 11:12)

Genome-wide association study (GWAS)

- Scanning genetic variants across the complete set of DNA, or genomes, of many individuals to see if any variant is associated with a particular trait.
- Requirements
 - Low cost, accurate method for genotyping
 - Array-based method (requires a catalog of human genetic variants)
 - Next-generation sequencing (truly unbiased analysis)
 - Large number of informative samples
 - Statistical analysis
 - Association test
 - Multiple test adjustment

Cancer Proteogenomics workshop, IIT Bombay, 2018



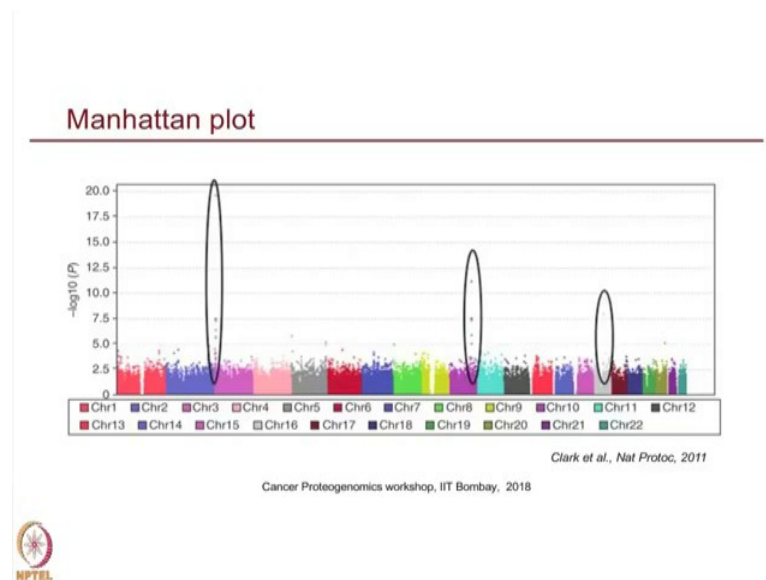
In order to find out, you have to scan the whole genome the whole DNA sequence and all the SNPs in order to find which one is actually associated with that phenotype right. In order to do that we have to do the genome wide association study, which is a GWAS; I think again probably you have heard about this term already. And in order to do the Genome Wide Association Study, we first the need the very kind of low cost assay on our platform in order to do the whole genome genotyping and this can be done by array based platforms, and the nowadays and the arrays can go up to one million SNPs, so basically you can scan 1 million SNPs at a time.

And you can also go with the next-generation sequencing based approach, because array-based approach we are talking about the 10 million SNPs in the human genome, but the array can only cover one million right. But if you do a next-generation sequencing with very good depths, and then you this is truly unbiased assay, but it is still more expensive than the array based technology. And then you also need a notch population with both case and the controls always different variation for quantitative trait.

And then of course, we need the statistical analysis in order to analyze the data, but the basics statistics is simple is the one we just talked about the and either the chi-square table test, the trend test or the and linear regression test for the quantitative trait. And one thing we want to remember is, because now you are testing a million SNP for example, against one phenotype and then you are doing a lot of tests so this is the multiple testing

problem. I think Dr. Mani has talked about that yesterday, so we need to be careful to adjust for the multiple test because you did so many tests and by random chance you are getting going to get some positive hit right; we want to make sure that what we get is not because of that reason.

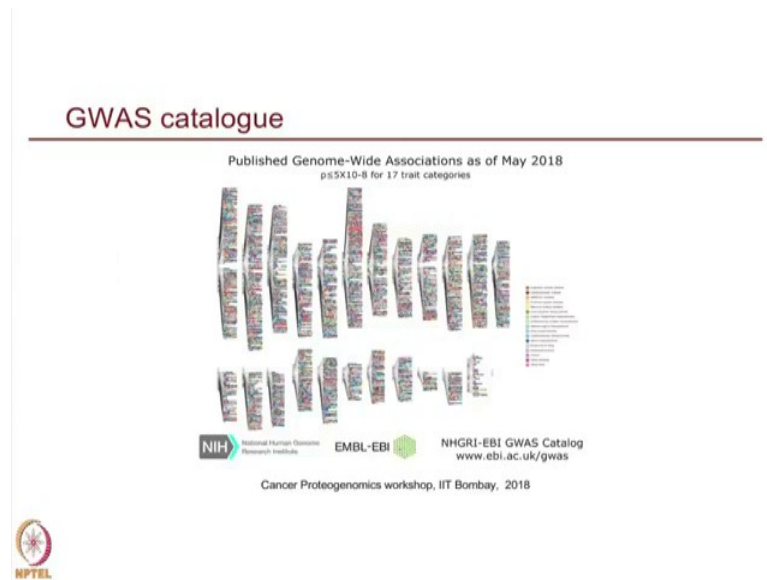
(Refer Slide Time: 13:23)



So, and when you get the result from the GWAS analysis, it can be visualized in this Manhattan plot. So, in this plot so on the x-axis, these are the genome locations of the locus and the each chromosome is indicated by a different color. So, here is chromosome 1 and the here is chromosome 22 usually we don't include the sex chromosomes here. And another y-axis, this is a p binaries in the minus log scale, because if you take minus log and there is a smaller p, when you have a bigger a value right and then it will appear at the top of the plot.

And then immediately this is a very effective way for validation and immediately you can see at this chromosome 3 and the this chromosome and this chromosome, the 3 chromosomes in at these positions, there is that kind of the SNPs are associated the phenotypes you are interested in. So that is very effective way to visualize and the understand your data.

(Refer Slide Time: 14:37)



And well so the GWAS catalogue is a resource that captures all the and the SNPs associated with a different kinds of diseases of phenotypes categorized into I think 17 categories and then you have all the associations in that resource. And the then they actually used the p value of 5 multiplied by 10 to the minus 8 in order to determine the significance.

So, can anyone tell me why they pick this cutoff rather than just a 0.01 or other values? The 0.05 it is yeah, 5 times 10 to the minus 8, why that is used as a cutoff. So, if you are going to select the significance from a test, what p value are you going to use?

Student: 0.05.

0.05 right, but why they used such a stringent, because there is a 0.05 divided by 1 million right.

Student: To reduce the numbers, so we can get the specific traits and not so many findings.

Because there are too many findings, you want to get the most significant ones.

Student: Correct.

Any other answers?

Student: Also here when we are going for a whole genome sequencing then we are going multiple reads because of which we want to make sure that a very less and very like there are no variation in between, so variation is there we need to remove it that is where we are going for a most stringent criteria.

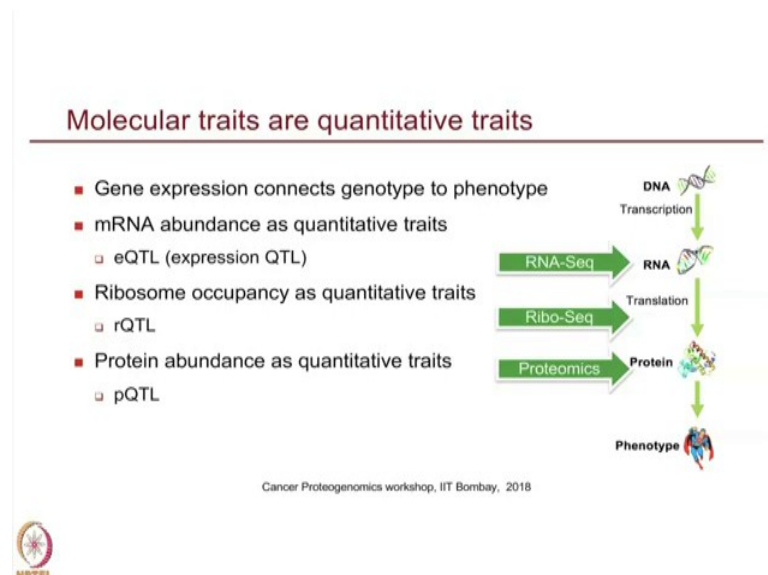
Ok, it is because of the variation you think.

Student: Yes, because multiple result being taken care.

Yeah, yeah, exactly. So, it is a multiple testing we are talking about and this is the reason they p takes this value is, because most of this data were generated on the SNP arrays which measure a million SNPs that means, you have a million times of tests. So, they just do the very simple Bonferroni correction as Dr. Mani mentioned yesterday. So, basically you have to use the cut off, so basically the 0.05 as you would typically do if you only have one SNP; you have to divide that by 10 one million, so that is why you get this number. Student: Yeah.

So, I think so far and we talk about very basic genetics and the GWAS studies or association studies, I think most of you probably have already been familiar with those.

(Refer Slide Time: 17:25)



So, in the next part I am going to talk about some molecular traits as the quantitative traits that is probably more relevant to this audience. So, we know although we are interested in the association between DNA and the phenotype, but we also know that

there is a lot of other things going on between the genotype and the phenotype, that means, the DNA has to be transcribed into RNA, and the RNA has to be translated into proteins and there are lot of regulations going on right. When you have both DNA, you would not end up with the same protein, you would not necessarily end up with the same phenotype.

So, in order to understand this whole process, we can consider the gene expression measurements as also as quantitative traits. If you think about, if you do a RNA-Seq experiment; for example, in 400 patients, and then you are going to get a gene expression measurement for example, TP 53 abundance of a 100 times. And then that is actually quantitative you can think that as the quantitative trait right. And then if you also have the genotype data from the same cohort and then you can actually correlate those. And the if you do that this is called the mRNA abundance as quantitative traits or its expression QTL error or eQTL error.

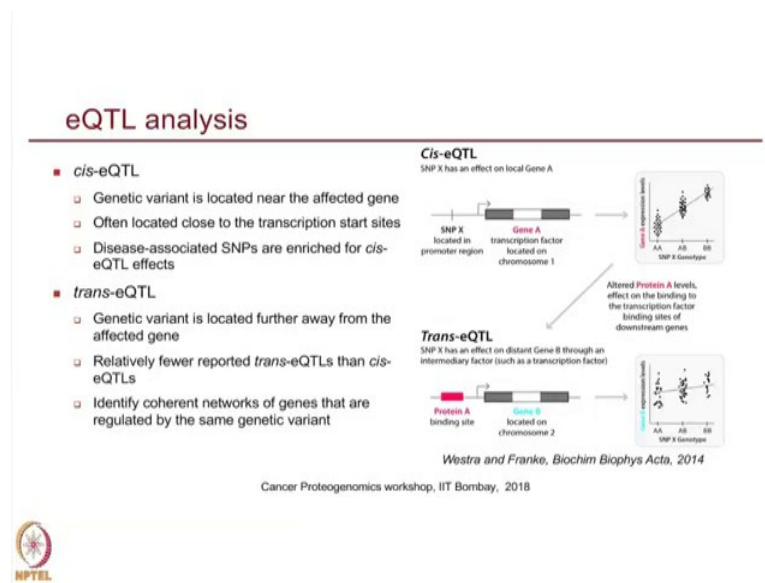
So, basically you are trying to establish the relationship between a genotype, but against the gene expression not the final disease phenotype. And the, I think yesterday somebody asked about the chip-seq right, you can also do chip-seq in order to get the actively translated the region of the RNA. And the then so this indicates the ribosome occupancy and then if you associate those measurements with the genotype, and then if you find QTL, it is called the rQTL or ribosome occupancy QTL. And similarly we can do protein abundance measurements and then we can associate the genotype with a protein expression. And then we if we get a association, it is called pQTL.

So, in these studies, if you only when you do a disease study you have one phenotype of interest and then you scan for many SNPs. But this time you have tens maybe 10000 genes with all the measurements and then so basically it further increases the test, you can do it very powerful because you can test the so many hypothesis at the same time, but just be careful about your multiple test adjustment. But the beauty is that you can understand the regulator regulatory mechanism for all the SNPs and all the gene of protein expression that is pretty cool.

But the basics it is very simple I mean because you treat each gene expression or protein expression as a phenotype or quantitative traits, and then you can do the same linear regression, and to find the relationship. So, I will show you some before we go to some

examples, I want to mention there is a difference between the typical disease phenotype and the gene expression or protein expression or the rQTL ribosome occupancy, this type of phenotype. Because both of them like the genes and the SNPs, they are or they have the location relationship on the genome like.

(Refer Slide Time: 20:54)



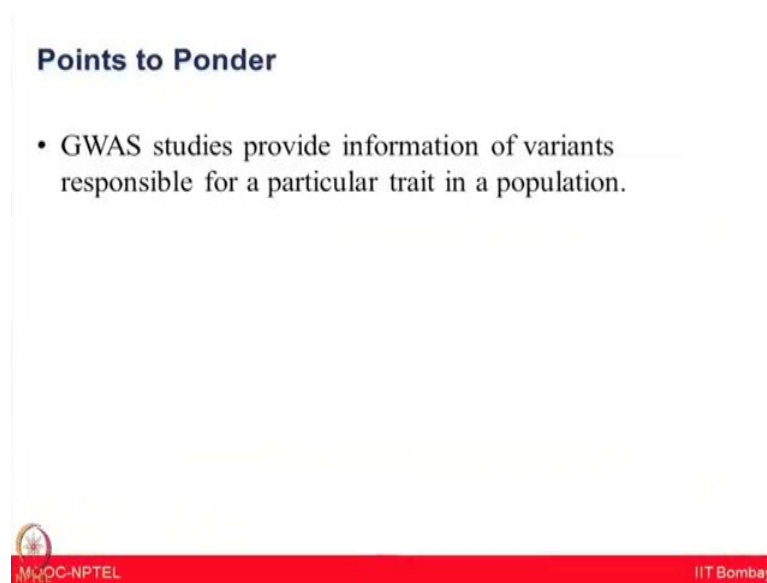
And then sometimes a SNP can be very close to the gene. And there are lot of times you can find SNPs in the promoter region, because there are lot of regulatory sequences in that region, and then you can find and the relationship between genotype in that region and the gene expression right next to that SNP. And these are called the cis-eQTL error, or similarly I mean if its protein expression is called cis-pQTL. And the as you can see in this example, if you have a SNP here, it may in especially in the promoter region and that may if effects the expression of this gene that is very understandable right and it has been shown that the cis-eQTL is very important because the disease associated SNPs are enriched in this cis-eQTLs.

And lets assuming this gene actually encode a protein which is a transcription factor, and then it could have additional effect on the other proteins not in the same region, but maybe a further away from the SNP, or even another chromosome you can still find the relationship. And in this case, it is called the trans eQTLs, that means, the gene expression or protein expression, you are interested in is further away from the SNP you

identified, even is this effect is because it is relatively indirectly effect it is relatively weaker .

So, usually you need larger sample size to identify those that is why they are fewer also like trans eQTLs not have been reported so far. But if you do find some of them, it is very nice because and it will show you maybe how one SNP may alter a pathway or network, because there are multiple genes might be controlled by the same SNP that can tell you how the network regulatory network works.

(Refer Slide Time: 23:02)

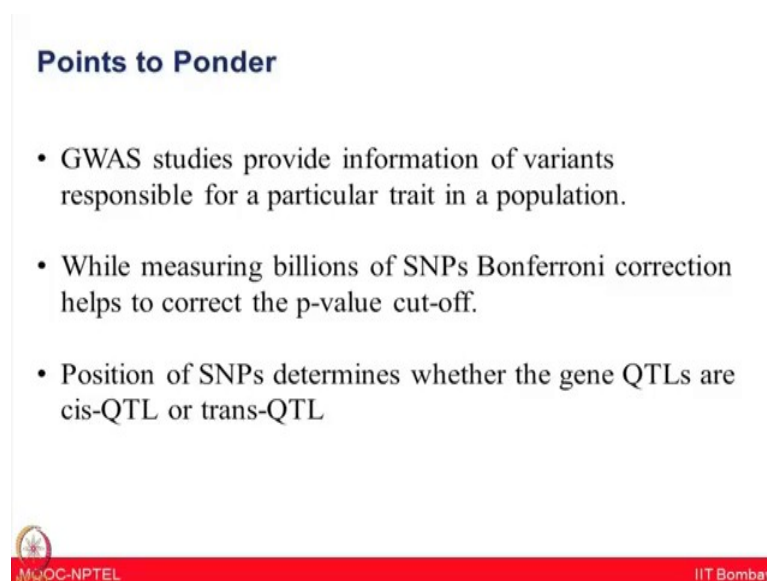


Points to Ponder

- GWAS studies provide information of variants responsible for a particular trait in a population.

MOOC-NPTEL IIT Bombay

(Refer Slide Time: 23:12)



Points to Ponder

- GWAS studies provide information of variants responsible for a particular trait in a population.
- While measuring billions of SNPs Bonferroni correction helps to correct the p-value cut-off.
- Position of SNPs determines whether the gene QTLs are cis-QTL or trans-QTL

MOOC-NPTEL IIT Bombay

In today's lecture, you were introduced to the association analysis, which is used to understand the relationship between genotype and phenotype various statistical test could be used to understand the relationship depending on whether the traits are quantitative or binary. Manhattan plots could be used to understand the results when GWAS or genome wide association studies are performed. Depending on the position of single nucleotide polymorphism SNPs and the gene QTLs, it could be either cis-QTL or trans QTL; those could be analyzed using these tools.

In next lecture, you will be given concepts of the power of integrative GWAS and eQTL analysis using various examples from literature.

Thank you.