

**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Karsten Krug**  
**Department of Biosciences and Bioengineering**  
**Broad Institute of MIT and Harvard**  
**Indian Institute of Technology, Bombay**

**Lecture – 55**  
**Pathway Enrichment – I**

Welcome to MOOC course on Introduction to Proteogenomics. After understanding how mutations in a given gene are specifically on P-sites can alter expression of the gene and its effects on the signaling pathways. We will now learn about how gene lists can be transformed into the pathways by Doctor Karsten Krug. He will talk about how one could interpret the role of differentially expressed genes in clinical conditions as compared to the healthy individuals by analyzing the pathways which are affecting them.

He will also talk about various pathway and databases which can be used to analyze pathways such as MSigDB, WikiPathways, KEGG and others. He will also talk about two basic ways to perform pathway enrichment. So, let us now welcome Doctor Karsten Krug to talk in more detail about how one can use various tools and transform the gene list to the pathways and make sense out of the data which one is obtained using various OMICs technologies.

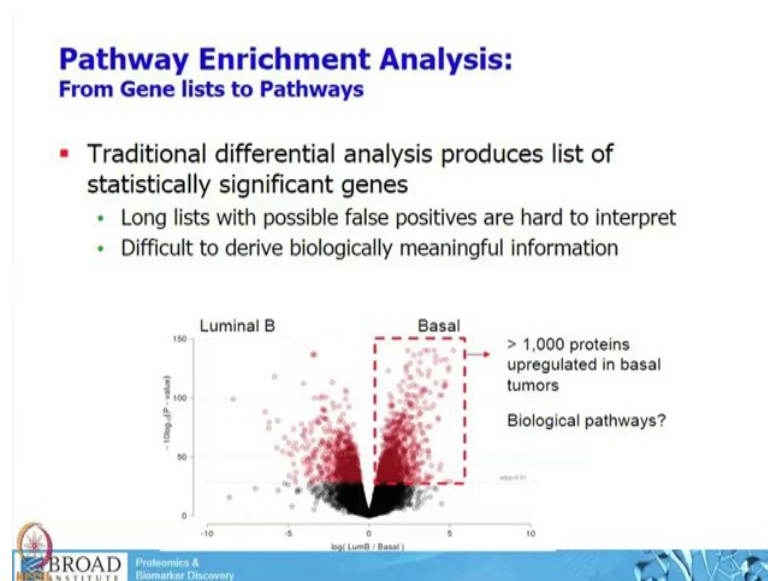
(Refer Slide Time: 01:58)

**Pathway Enrichment Analysis:**  
**From Gene lists to Pathways**

- Traditional differential analysis produces list of statistically significant genes
  - Long lists with possible false positives are hard to interpret
  - Difficult to derive biologically meaningful information

So, in this lecture we want to talk about how we can come or how we can transform gene lists into pathways. You know if you perform the experiment to compare your wild type and knockout, you perform your statistical test what Mani was talking about you end up with long lists of differentially expressed proteins phosphosites of genes and possibly ah you know a high weight or like a higher proportion of those might be false positives meaning them actually not differentially expressed in your in your sample. And, this makes it very hard to interpret these results biologically.

(Refer Slide Time: 02:27)



So, this is just an example here. So, if you look at luminal and basal breast cancer subtype that we have looked at yesterday in a hands on session. We see there is more one 1000 proteins upregulated in basal or differentially regulated between basal and luminal you know, but actually what we want to know what are the biological pathways to try off these kind of separations between luminal and basal.

(Refer Slide Time: 02:49)

## Pathway databases are curated at gene level

- Pathway = Group of genes that are members of a pathway

$$P = [Gene X, Gene Y, \dots, Gene Z]$$

- Pathway databases:

Database	Description	URL
MSigDB	Molecular Signatures Database	<a href="#">Link</a>
WikiPathways	Community driven pathway curation	<a href="#">Link</a>
NetPath	Manually curated resource of human pathways	<a href="#">Link</a>
KEGG	Kyoto Encyclopedia of Genes and Genomes	<a href="#">Link</a>
Reactome	Manually curated and peer-reviewed pathway database	<a href="#">Link</a>



So, in order to do that there is many different ways how to perform pathway analysis and it all starts with a pathway database. And, you know how do we represent the pathway in a computer? So, pathway in the most simplest case is a crew of genes that are members of a pathway or they share some common biological process all right.

So, it is basically a pathway is a list of genes gene symbols. And, there is many different resources for pathway databases like MSigDB developed at the broad WikiPathways, NetPath which has been developed here and by Alex.Pandey's group. it has the Kyoto Encyclopedia of Genes and Genomes, Reactome and there is even more. So, here these if you click on these links they will directly ah forward you to the actual website to these databases

(Refer Slide Time: 03:49)




So, I just want to briefly point out WikiPathways which is a very promising research resource for curated pathways which has been you know developed over the last couple of years, but now it really starts to take off. So, this is like a Wikipedia for pathways. So, everybody who is you know studying a specific pathway. So, maybe you in your particular lab you are interested in one specific pathway.

So, actually you are the expert to do this kind of curation of a pathway right and this website or this entire resource you know should enable you as researcher to help the community to provide well and highly accurate curated pathways Somebody also using that resource a lot and you know they also have like a curator of the week here and things like that. So, if you are really contributing a lot to that resource you might end up on their webpage

(Refer Slide Time: 04:43)

**Pathway Enrichment Analysis:**  
From Gene lists to Pathways

- **Hypergeometrical test (Fisher's exact test)**
  - Test for over/underrepresentation of pathways in a list of genes **differentially expressed** (DE) in two conditions (e.g. tumor, normal)
  - List of DE has to be determined beforehand, e.g. DE genes at 1% FDR
- **Gene Set Enrichment Analysis (GSEA)**
  - Combines evidence from many small but coordinated changes to achieve statistical significance when individual features do not
  - **All measurements** are taken into account

 **BROAD** INSTITUTE Proteomics & Biomarker Discovery

So, there is pathway databases. So, now, we want to do pathway enrichment and there is basically two different ways how to perform that or how to approach that problem. So, one is you know based on some sort of test. So, here just pointed out the Fishers Fisher's test where you test for over or under representation of pathways in a list of genes in a list of differentially expressed genes you know in two conditions. Let us say tumor and normal.

So, meaning, what that means, you have to define this list of differentially expressed proteins beforehand before you do your pathway analysis. Let us say you compare basal and luminal and then you will do your statistical test two sample T test and you look at everything that is differentially expressed at 1 percent FDR. So, that is the input to a pathway analysis


The other approach is so called gene set enrichment analysis which has been introduced in 2005 it is you know highly cited it is you know very convenient way to look for you know small, but coordinated changes you know that you observe in your sample. Let us say, like a protein does not maybe pass threshold for being statistically significant in this pathway, but if you observe many different members of the same pathway that might not change a lot, but they are all changed into the same direction right which increases the evidence that your pathway might be enriched.


So, and the main difference compared to the first approach is that here we are looking at all measurements. We do not filter anything beforehand, but we look at all measurements you know advance

(Refer Slide Time: 06:37)

**Fisher's Exact Test**  
Hypergeometric Test

- Statistical Significance test used in the analysis of contingency tables
- Valid for all sample sizes—especially small samples
  - Equivalent to Chi-Square test for large samples
- Originally devised by R. A. Fisher to address claims by "lady tasting tea" (Dr. Muriel Bristol)
  - Can you tell if milk or tea was poured in first?



 **BROAD** INSTITUTE Proteomics & Biomarker Discovery

So, briefly about Fisher's exact test; I am sure that many of you guys are aware of that it is a test to you know to test significance of contingency tables. So, table said. So, you know that is a little story that Mani used to tell here, so has been developed by Fisher and to address claims by a good friend that he had she was called Miss Bristol and she kind of insisted that she is able to tell whether the milk or tea was poured first in a cup.

And so, in order to prove her wrong he conducted a little experiment you know by you know conducting this experiment 8 times and he would pour 3 times the tea first into the cup and 3 times the milk first into the cup and then he just you know fill out this continuity and he counted the number of successes of the lady you know and filled out this contingency table and then you can basically calculate Fisher's exact P-value, but just enumerate about overall possibilities in calculator to P-value So, in this case he proved her wrong.

(Refer Slide Time: 07:56)

### Fisher's Exact Test


#### Application to Pathway Enrichment

Contingency (Confusion) Matrix

		In gene list (G)		
		yes	no	row TOTAL
In Pathway (P)	yes	$N_{pG}$	$N_{p\bar{G}}$	$N_p$
	no	$N_{\bar{p}G}$	$N_{\bar{p}\bar{G}}$	$N_{\bar{p}}$
	column TOTAL	$N_G$	$N_{\bar{G}}$	$N$

- Gene List:
  - Sorted list of input genes
    - Ranked by correlation, differential signal, etc.
- Pathway (Gene Set):
  - Set of genes from biological pathway or functional experiments

$N = 18988$   
(total number of protein-coding gene symbols)



And, you can do the same with pathways. So, let us say you want to compare your pathway and your differentially expressed list of genes. So, here you have your gene list on the x axis and you ask the question whether the gene is in your list or not and you know why you look for the pathways meaning if the gene is in your pathways or not all right. And, then you can fill out this matrix it should you always compare against a background.

So, the total ends should sum up for example, to all genes new human genome which is this case depending on which database you are using this number might be slightly different in this case we have roughly 19000 genes.

(Refer Slide Time: 08:40)

### Fisher's Exact Test Example


- Pathway: ATP1A1, LARP1, PCDH9, CNTN5, CHD4, THBD, EGRI, EIF2B4, MAPK1
- Gene List: UGT1A1, CRYL1, LARP1, PTPN11, PCDH9, CNTN5, CHD4, EGRI, NR1, PGF, MAPK1
- Background: UGT1A1, CRYL1, LARP1, PTPN11, PCDH9, CNTN5, CHD4, EGRI, NR1, PGF, MAPK1, ATP1A1, THBD, EIF2B4, ACOX1, FADS2, SPON2, TNS4, EB13, THEM60

In Pathway (P)	In gene list (G)			row TOTAL
	yes	no		
Yes	6	3	9	
No	5	6	11	
column TOTAL	11	9	20	

```
> x <- matrix(c(6,3,5,6), nrow=2, byrow = T)
> x
     [,1] [,2]
[1,]  6   3
[2,]  5   6
> fisher.test(x, alternative = 'greater')

Fisher's Exact Test for Count Data

data:  x
p-value = 0.311
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3752773      Inf
sample estimates:
odds ratio
 2.294738
> |
```



And, here is a little example. So, let us say this is my pathway this is not like an arbitrary a theoretical example. This is no you know does not have any biological meaning, but you have your pathway you know and everything that is highlighted in bold here does overlap with your gene list. So, meaning 6 of your gene list members on this pathway and so on and so forth and your background is in this case is the universe of 20 genes.

Again, just as an example to how to fill up this contingency table and then you can basically again using R you can just calculate Fisher's P-value In this case it is not significant whatsoever all right. So, this is something you do for every pathway that you want to test


One way you know convenient tool to you is a very powerful tool is it is called David which has been popular several years back, but is very powerful and very easy to use. So, you can just as I just described you can just paste in your list of different of genes. You can tell the software whether this is my genes of interest or whether this is whether this is my background list then you can perform these sort of tests that I just described and you will get for all on every pathway you will get the P-value and enrichment scores and things like that.

So, it is very convenient to use because to use because you just go into excel to your go into the result of your statistical test in excel you filter your significant genes and then just paste them in here



(Refer Slide Time: 10:17)



## Gene Set Enrichment Analysis (GSEA)



- **GSEA** takes all genes into account and does not require to define a gene list of interest *a priori*
- GSEA can combine evidence from **many small but coordinated changes** to achieve statistical significance when individual features do not
- **Gene set:** group of genes that share a common biological function, chromosomal location, or regulation.

INSULIN\_SIGNALING\_PATHWAY      GCK CALM2 PPARGC1A ELK1 CALM1 EIF4 ...

<http://software.broadinstitute.org/gsea/msigdb>




BROAD INSTITUTE Proteomics & Biomarker Discovery

So, now we want to talk about Gene Set Enrichment Analysis or GSEA and this is also something we want to try during hands on sessions. So, I hope that we make that work. So, as I already mentioned here you take into account all genes you do not have to filter before you do the analysis and so, it is basically what I just said. So, gene sets; so pathways are also called gene sets or basically with the introduction of GSEA the board also came up with a collection of gene sets which again it is just a collection of genes which might refer to a pathway or you know they share biological process and so on and so forth, but in general gene set is nothing else in a pathway.


(Refer Slide Time: 11:16)

## Gene set collections in MSigDB



<b>H</b> <b>hallmark gene sets</b> are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.	Set of 50 highly curated gene sets of cancer hallmark pathways
<b>C1</b> <b>positional gene sets</b> for each human chromosome and cytogenetic band.	
<b>C2</b> <b>curated gene sets</b> from online pathway databases, publications in PubMed, and knowledge of domain experts.	C2 canonical pathways (CP) contains curated pathways from KEGG, Reactome and Biocarta
<b>C3</b> <b>motif gene sets</b> based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.	
<b>C4</b> <b>computational gene sets</b> defined by mining large collections of cancer-oriented microarray data.	
<b>C5</b> <b>GO gene sets</b> consist of genes annotated by the same GO terms.	
<b>C6</b> <b>oncogenic gene sets</b> defined directly from microarray gene expression data from cancer gene perturbations.	
<b>C7</b> <b>immunologic gene sets</b> defined directly from microarray gene expression data from immunologic studies.	

<http://software.broadinstitute.org/gsea/msigdb>



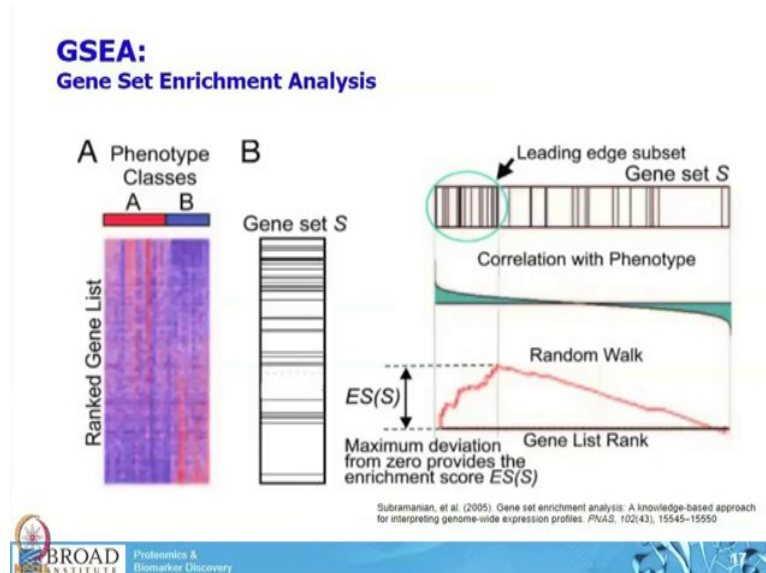
BROAD INSTITUTE Proteomics & Biomarker Discovery

And, there is a large collection of those you can find in in the molecular signatures database or MSigDB. So, if you go to this website you will find this kind of overview. So, these are different categories of gene sets which you can find in MSigDB. And, yeah, I just highlighted I guess which are the most commonly used gene sets in MSigDB; so, one is the so called hallmark gene set data dataset or gene set which is if is a very small databases are just 50 signatures or 50 pathways, but they are highly curated and represents very important and common cancer hallmark pathways.

And, the other one is the. So, called C2, CP; CP stands for canonical pathways. So, that is a collection of gene sets that have been derived from other you know pathway databases like KEGG or Reactome and so on and so forth. Canonical means you know these are, so, what we believe is the actual pathway and there is others that might or might not be of interest.

For example, if you are used to geo like gene ontology terms you can also look into the category C5 and so on and so forth. So, the general principle of GSEA is shown on this slide. Here it is actually a figure one I believe from the Original Publication and back in 2005.

(Refer Slide Time: 13:00)



So, you again you start with the data matrix. So, we have seen this kind of data format a couple of times in our workshop. So, you measure features in this case it is genes on your works and you measure these features across a set of samples. And, you know you always want to compare at least two phenotypes two conditions right. In this case it is condition a phenotype A and phenotype B; let us say one is tumor, one is normal.

So, and you somehow rank these gene lists and you would rank it in a way that you rank that you would rank differentially genes accordingly. For example, you do your two sample t-test then you rank it according to your P-value right. So, the most significantly differential genes are on the top and then the further down you go the less significant it becomes. So, you have a list of ranked gene sets genes.

And, then you take your pathway of interest. So, this is now just one pathway again and to look where in my ranked list of genes do the members of this pathway fall. So, all of these like horizontal bars are locations of members of this gene set in my excel data right. And, here you see that here in this upper part of this plot you see there is many more horizontal bars and down here. So, there is an enrichment of this pathway just visually you can see that, but so, there is enrichment of this pathway among you know among genes that are differentially between A and B that is the principal.

So, how you calculate that? So, they calculate a so called enrichment score ES, by basically so, now, we are we have transposed this matrix. So, we are now looking at the ranks on the horizontal axis here. So, from high to low and now we are basically calculating enrichment score by working down this list here.


And, whenever we see a member we observe member our data set we increase our running some statistic here and if you do not observe a member we decrease it right. And, this builds up this kind of mountain plot here and at some point you do not observe enough members of this pathway anymore that that so, meaning this running some statistics starts stops to increase you constantly decrease.

And, then there is different ways how to quantify this enrichment scores one is just taking the you know the maximum deviation here or you can also calculate the area under the curve and things like that. So, there is you know different nuances through that type of analysis, but that is the general principle.


So, just to mention, so, this is one enrichment score in order to calculate the P value for that enrichment score you would you know repeat this analysis 1000 times you know you would do 1000 permutations and you would do you would permute your class labels. You would shovel your class labels just to get a background distribution and you would repeat this analysis you get a distribution and then from that distribution you can calculate an empirical P-value for your observed enrichment score all right.

(Refer Slide Time: 16:16)

**single sample GSEA (ssGSEA)**  
**Signature projection method**

- Single sample GSEA is a derivative of GSEA that works on single class datasets, e.g.
  - Expression values in a single condition
  - mRNA/Protein correlation coefficients
  - logFC between two conditions
  - ...
- ssGSEA enables the projection of gene/protein expression matrix into gene set (pathway) space:

```
graph LR; A[Gene centric] -- Signature projection --> B[Pathway centric]
```
- Pathway centric matrix can then be subjected to cluster analysis, marker selection, etc.



So, that is the general principle of GSEA you compare two phenotypes and you have measured a sufficient number of replicates like biological replicates in your phenotype A and B. So, another approach to this kind of gene set enrichment analysis is also it is the so called signature projection method or seeing the sample GSEA which basically works in the single class data set.

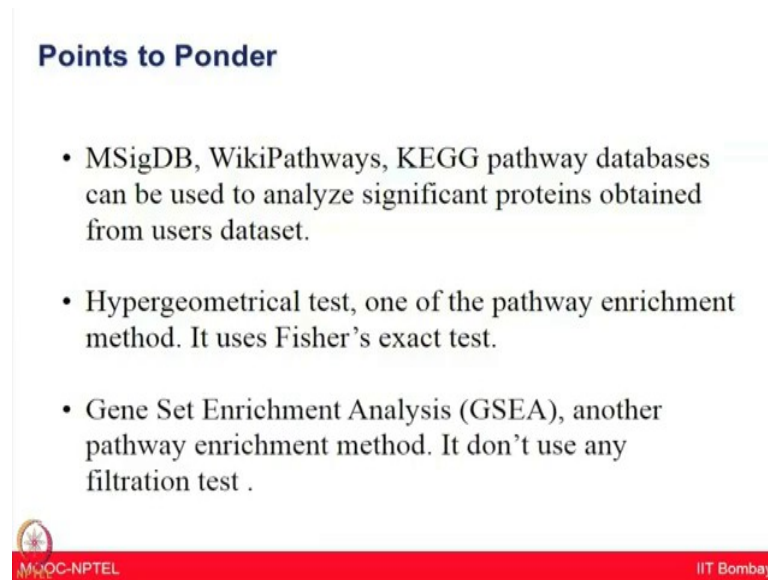
So, they you do not necessarily want to compare you know two conditions or you might not have any replicate. So, you just have a single data vector let us say this data vector can be anything. It can be you know expression values in a single condition. So, you just have one experiment measure the proteome in one time point, let us say and you want to look at high abundant and low abundant proteins in general. Or you can do you can you know look at on a protein correlation coefficients you can your ranking would be you know highly correlated protein or genes in proteome and on a space you know compared to low correlate or negatively correlated genes and so on and so forth. So, that is the principle.

So, basically what you do if you have a data matrix you would project your let us say these are gene centric matrices and you would project each column into pathway space and we will do that during hands on and I hope that this will become a bit clearer, but basically that is that is why it is called a signature projection method.

So, you start with your genes or proteins whatsoever and then you apply this method and in the end you have the same kind of matrix, but instead of looking at genes you are looking at



pathways. And, with this matrix you can again do like all kinds of statistical analysis like you know supervised or unsupervised marker selection type of analysis and so on so forth

(Refer Slide Time: 18:20)



**Points to Ponder**

- MSigDB, WikiPathways, KEGG pathway databases can be used to analyze significant proteins obtained from users dataset.
- Hypergeometrical test, one of the pathway enrichment method. It uses Fisher's exact test.
- Gene Set Enrichment Analysis (GSEA), another pathway enrichment method. It don't use any filtration test .

 MOOC-NPTEL 

I hope you have learned how the pathways provide information for the combination of correlated genes making a network to make a system functional. You also learned that WikiPathways is like Wikipedia for pathways. We also learned about hypergeometrical test which is based on Fisher's exact test where one can compare clinical conditions with healthy individuals and make a pathway based on differentially expressed genes.

Other pathway enrichment approaches like Gene Set Enrichment Analysis or GSEA, it includes all the proteins in this study without filtration and mapping a pathway. The next lecture is going to be the continuation of the pathway enrichment by Doctor Krug.

Thank you.