

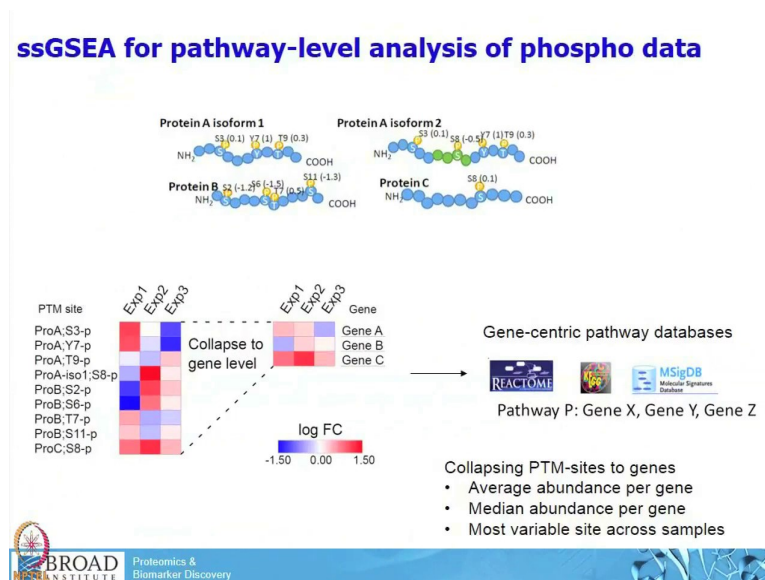
**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Karsten Krug**  
**Department of Biosciences and Bioengineering**  
**Broad Institute of MIT and Harvard**  
**Indian Institute of Technology, Bombay**

**Lecture – 56**  
**Pathway Enrichment-II**

Welcome to MOOC course on Introduction to Proteogenomics. After understanding the two approaches for pathway enrichment and basic differences between both the approaches, we will now listen Dr. Karsten Krug who will explain the use of GSEA at the pathway level analysis for different PTMs by taking an example of phospho data.

Dr. Krug will talk about the way one could analyze the available data at gene level to take it to the PTM level. He will also talk about the recent work which is an initiative to make a PTM site curated database at Broad Institute. This involves a scoring of each P-site by mapping it against the database. So, let us now welcome Dr. Karsten to talk more about how GSEA can be used to map PTM pathways for the analysis.

(Refer Slide Time: 01:28)



So, if you wanna apply that kind of approach for phospho PTM data. now I am going to talk about phospho data, but this is actually true for all kinds of PTMs. So, there we measure

different phosphorylation sites on proteins, right. And it might happen that we have multiple phosphosites in the same protein, it might happen that we have different isoforms of the same gene or protein you know which we measure different phosphosites on that and so on so forth. But in order to do pathway analysis all of these pathway, databases are gene-centric, as I said pathways is the list of genes.


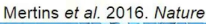
So, right now, all of the database that we have are curated gene level. So, meaning if you want to do any pathway-level analysis of our phospho data, you first have to collapse all of these measurements of phosphosites into gene level. So, it basically your throwing way a lot of information, but it is this is something we have to deal with but now because it has now there are no databases that are curated at PTM site level or at protein or isoform level.

So, you would do that by for example, taking the average protein or median or looking across the most variable site you know across your samples. So, variants means informations, that is why you would pick that one.

(Refer Slide Time: 02:53)

**Example: ssGSEA for phospho-pathway clustering**

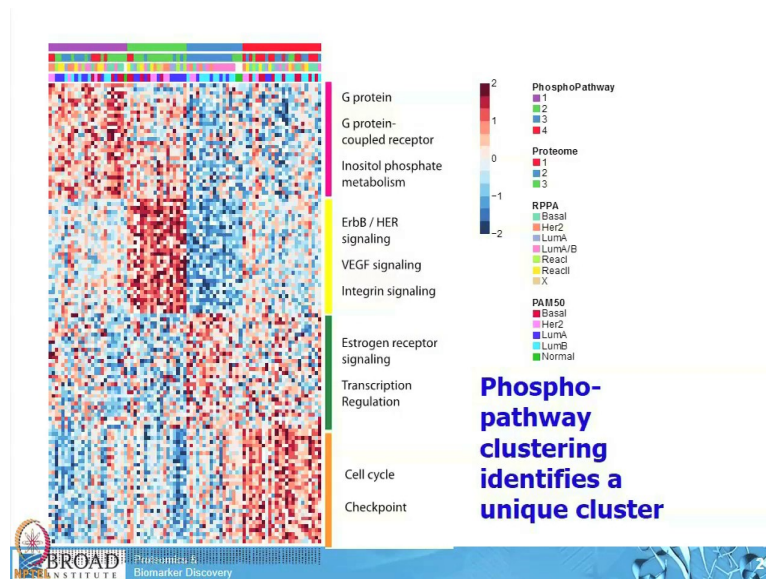
- Dataset: 5,914 phosphoproteins
  - Filtered Phosphoproteome data
    - Phosphosites with <81 missing values
    - Standard deviation > 0.5 across all samples
  - Phosphosites rolled-up to proteins using median ratio
  - Map phosphoproteins to genes
- Map samples to MSigDB pathways using ssGSEA
  - 908 curated pathways
- Consensus *k*-means clustering in pathway space
- Assess cluster coherence

  Proteomics & Biomarker Discovery

Now, this is one example which where we have you know boarders approach. This is again from the breast cancer study, what is the Nature 2 years back where you know, we started with about this 6000 phosphoproteins and performed the same kind type of analysis I just said. So, we using the sample GSEA to map this phosphoproteins which we of course first collapsed to genes using median ways show and so on and so forth. The projected those into

this space of pathways and in order to pathways, and then we performed consensus clustering on this data matrix.

(Refer Slide Time: 03:41)



And interestingly we saw like a clustering like a unique cluster which we only saw on pathway space and now we would not have seen that if we would look at phosphosite or phosphoprotein level. So this one example where these kind of analysis, we just you know projected data onto a higher level of annotation and perform some analysis, can give you new insides that you probably would have missed if you would have not done that.

Student: Can you please repeat what you mean to say according to the phosphorylated site?

Can you please use the mike that everybody can understand, you know.

Student: So, just coming to at this kind you are saying according to each gene and all according to our biological function, it will arrange the phosphosites.

No, here we are not looking at phosphosites anymore. So.

Student: No, the word clustering that phosphosites in each signaling pathway, is not it?

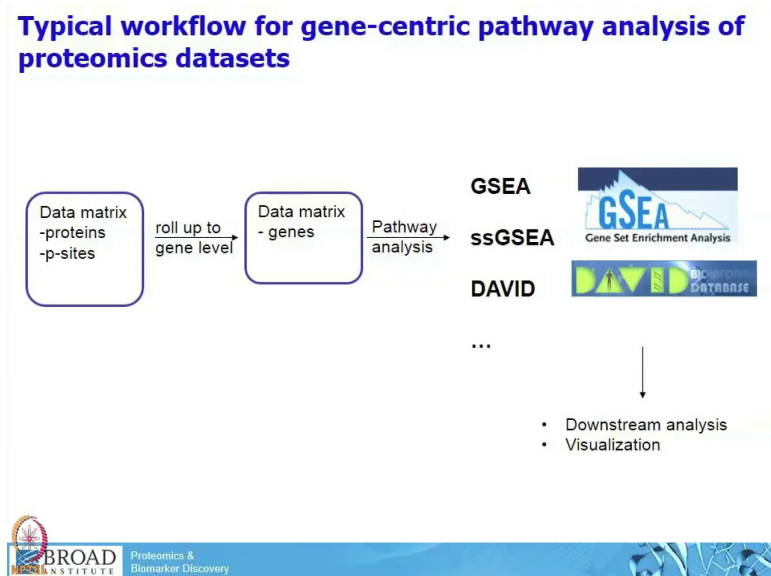
No, we calculate for each I mean we, now we are looking we are not looking at phosphosites, but each row is a pathway, and the data is actually there which means go I was this quite being a couple of slides back.

Student: Yeah.

So, we have already combined our phosphosites to phosphoproteins to genes, and then we performed our signature projection method. So, meaning in each sample, so these are the 77 breast cancer samples that we are looking at here. And in each sample we have a score and which means score for this particular pathway, right. So, red means, this pathways more active; blue mean this pathways less active in this particular sample. Then we can do in this case we have done an unsupervised cluster analysis on that.

So, very convenient way to you know combine multiple measurements of phosphosites, it map to the same gene through the same like through a gene-centric levels actually morpheus which is a very versatile matrix visualization to it. We are going to use it again and so on. It is very powerful, but we all, we are going to use it for this particular purpose here.

(Refer Slide Time: 05:53)



So, just like to repeat what I have just said. So, we start with data matrix which can be either proteins, phosphosites or genes transcripts whatsoever. For the first step that we always have

to do is we have to roll up our expression values to the level of genes. Because, all of these databases that we are using for our pathway analysis or gene-centric.

And after we have done that then we can continue with our pathway analysis drawing GSEA, single sample GSEA you know DAVID what else, so many different tools and approaches out there. It is you know number of tastes what you like more, but we highly recommend we highly prefer doing some sort of GSEA, Genes Enrichment type of Analysis, because we do not have to throw away we do not have to filter you list before hand.

(Refer Slide Time: 06:50)

**Site-centric analysis requires pathways curated at PTM site level**

- Pathway analysis of phosphoproteomics datasets is typically done in gene-centric space, due to lack of pathway database curated at the site level.
- PTM signatures database (PTMsigDB) provides a curated resource for phosphosite-specific pathway analysis

PTM site

PTM site	Exp1	Exp2	Exp3
ProA:S3-p	Red	Blue	Blue
ProA:Y7-p	Blue	Blue	Blue
ProA:T9-p	Blue	Blue	Blue
ProA-iso1:S8-p	Blue	Blue	Blue
ProB:S2-p	Blue	Blue	Blue
ProB:S6-p	Blue	Blue	Blue
ProB:T7-p	Blue	Blue	Blue
ProB:S11-p	Blue	Blue	Blue
ProC:S8-p	Blue	Blue	Blue

log FC


-1.50 0.00 1.50

Site-centric pathway database

**PTMsigDB**

Pathway P: Site X0, Site Y0, Site Z0

- Multiple PTM sites measured on the same protein contribute to the enrichment score

 Proteomics & Biomarker Discovery

So, as I have just told you all of these databases that are available now are curated at the gene-centric level. I just want to introduce you tool like a projected we are doing at a broad together, many of the other collaborators where we actually tried to come up with the pathway database, it is curated at the site level at phosphosite level. So, this involves many people many resources ok, let us say for moment. So, we call that the PTM signatures database which in theory what that is actually to go to be able to is go each and every single phosphorylation site directly you consider database.

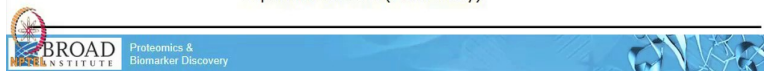
Thats a large curation effort as you can imagine. It is very difficult to curated pathways at the gene level, but if you want a even you know go deeper I want to curate every single phosphorylation site, what does it do in this pathway does it go up or down and is it involve that all. So, it is lot of curation effort.

(Refer Slide Time: 07:57)

## PTMsigDB – curated resource for phosphosite-specific pathway analysis

- PTM signatures database – signatures curated at site level
- Developed in collaboration with curators of molecular pathway and PTM databases

Signature Source	Signature Type	URL
PhosphoSitePlus	• Perturbation-specific signatures with annotated directionality of change	www.phosphosite.org
NetPath	• Pathway-specific phosphosites with annotated phosphorylation / dephosphorylation status	www.netpath.org
WikiPathways	• Pathway-specific phosphosites with annotated phosphorylation / dephosphorylation status	www.wikipathways.org
LINCS	• Perturbation-specific signatures from experimental data (P100 assay)	



Student: Point 2, this is the human data here?

In most of the signatures are human. We started to do at for mouse and rat too, but it is mostly human. So, we teamed up with other database curators from phosphosite plus, from netpath or so wikipathways as I mentioned earlier.

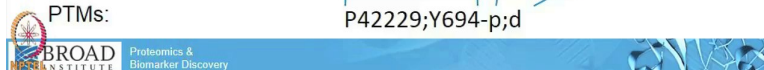
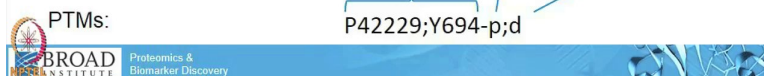
(Refer Slide Time: 08:26)

## PTM sites are robustly represented in PTMsigDB

- Unambiguous representation of PTM sites is crucial:
  - Different database accession numbers: UniProt, RefSeq, Ensemble, ...
  - Residue numbers vary between isoforms and database versions
- PTMsigDB supports three types of PTM site identifier:

Database format	Site accession	Example
Uniprot-centric	Uniprot_acc;site-type;direction	Q06609;Y315-p;u
Flanking sequence	+/-7aa flanking seq-type;direction	ETRICKIYDSPCLPE-p;u
PSP site group id	site_grp_id-type;direction	448324-p;u

- Format generalizable to other



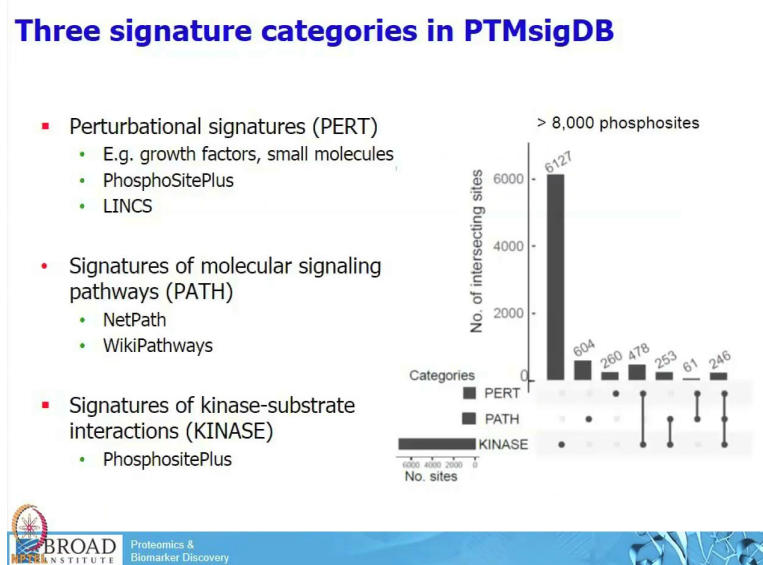
And nobody have all of these people involved here, we you know started to curate this database. And other very important aspect when you want to do something like this, how do I represent the PTM site robustly which might some preview, but it is actually like a big problem like gene symbols have been standardized in a way right, this is study you go gene symbols, which try to harmonize and standardize human gene symbols. If you look into protein databases, you know you looked the same protein might have different accessional number in one database. So, this is now uniprot database, if you look after same protein at RefSeq you would have completely different unrelated ID, so it is very difficult to cross reference those, right.

And this is even more like a even more severe problem if you look at PTM sites. So, how do you robust you to send the PTM site? So, there is different ways how we try to approach these problems. So, one is uniprot-centric. So, uniprot is well highly curated protein database. So, you know we picked uniprot and we represent the site as uniprot id modified residue and the PTM type. And in this database you also have information whether it goes up or down in the specific pathway or for the patients.

So, another way to represent phosphosite or flanking sequences. So, this is what we just looked at in a morning right. So, we look at like plus minus 6 or 7 amino acid or whatsoever you know around the phosphorylation site. This is a pretty unique identified already if you compare that in the human proteome.

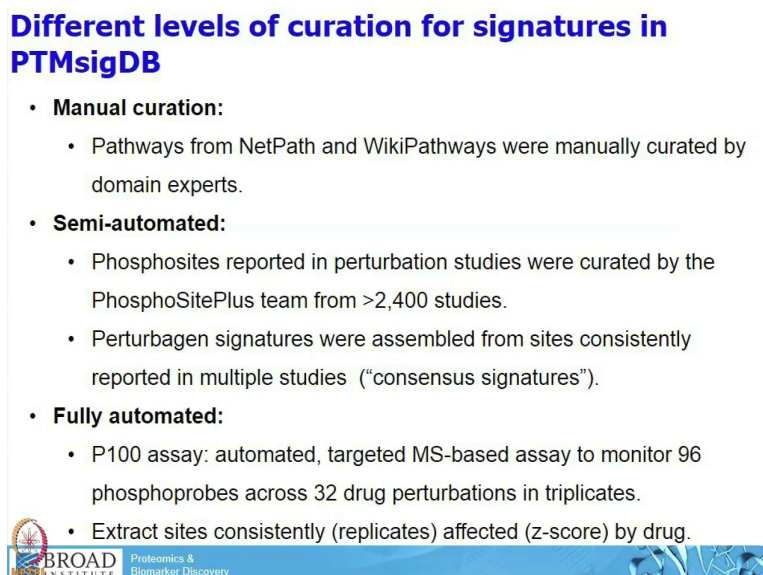
Now, also phosphosite plus they are actually trying to come up with a unambiguous way to group, sites across or within protein, families just give me a mind you know this residue number might change if you look at isoform A or isoform B, right. Its might be the same site, but residue number might be completely different. So, in the site group id tries to harmonize those kind of events, ok.

(Refer Slide Time: 10:48)



So, I have quickly go through here, so right now we have like pathways we have a kind of substrate signatures used to be talked about and we have lot of perturbations by growth factors or so small molecules in this database.

(Refer Slide Time: 11:01)



And so most of them like this pathways have been completely manually curated by curators from NetPath, WikiPathways. And we also extract that signatures semi-automatically and fully automatically. So, this is we try to come up with this standardized way how to extract



automatically, automatically do I have these kind of signatures from known literature. Let me quickly go to these slides here.

(Refer Slide Time: 11:34)

### "Consensus-signatures" reduce experimental noise

- Functional annotations of PTM sites from the literature are often not consistent across studies.
  - Different cell types, experimental conditions, protocols, ...
- **Consensus-signature:** PTM sites that have been consistently reported across multiple independent studies:

PhosphoSitePlus	Perturbation P						phosphorylation acetylation ubiquitination
	PTM site	1	2	3	...	n-1	
Study 1	↑	↑	↓				↑
Study 2	↓	↑	↓				↑
Study 3		↑	↓			↑	↑
...							
Study N	↑	↑	↓			↓	↑

Signature S: 1 (red), 2 (blue), ..., m (red)

PTMsigDB

BROAD INSTITUTE Proteom Biomarker Discovery

So, and actually in order to extract signatures we came up with the you know method to extract consensus-signatures. So, let us say you have one perturbation or one pathway that has been studied by different studies by different labs you know there is different papers that might be put you know this site goes up on this perturbation. And other study we both the same sides, but it goes down. So, this kind of inconsistency you find all over the place, right. This might be due to you know the labs may used different cell types, experimental conditions, different protocols whatsoever. So, what we try to do, we try to come up with a consensus between at least two independently published signature like papers you know.

(Refer Slide Time: 12:33)

### PTM Signature Enrichment Analysis (PTM-SEA)

- PTM-SEA extends Gene Set Enrichment Analysis (GSEA) to directional PTM signatures.

Bi-directional signature S:

S = site 1 site 2 site 3 ... site m

- Score reflects degree of agreement between signature and data

**BROAD INSTITUTE** Proteomics & Biomarker Discovery

In order to include these signatures now database. I am uses very similar approached compared to GSEA we actually standard at scoring scheme in order to look for these signatures.

(Refer Slide Time: 12:44)

### Benchmark dataset: EGF stimulated HeLa cells

- Phosphoproteome profiling of HeLa cells in three experimental conditions:
  - Control (DMSO)
  - Mitotically arrested (nocodazole) and released cells
  - Cells stimulated with epidermal growth factor (EGF)
- ~36K localized and quantified phosphosites
- Both signatures (EGF and nocodazole treatment) are part of PTMsigDB and were used to assess PTM-SEA.

**Cell Reports**  
Volume 5, Issue 5, 11 September 2014, Pages 1553–1594  
Open Access

Reprints  
Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling  
Kirk Sharma<sup>1</sup>, Rachelle C. J. O'Souza<sup>1</sup>, Stefa Tjebkova<sup>1</sup>, Christoph Schaab<sup>1</sup>, Jacek R. Wroniewski<sup>1</sup>, Jürgen Cox<sup>1</sup>, Matthias Mann<sup>1</sup>

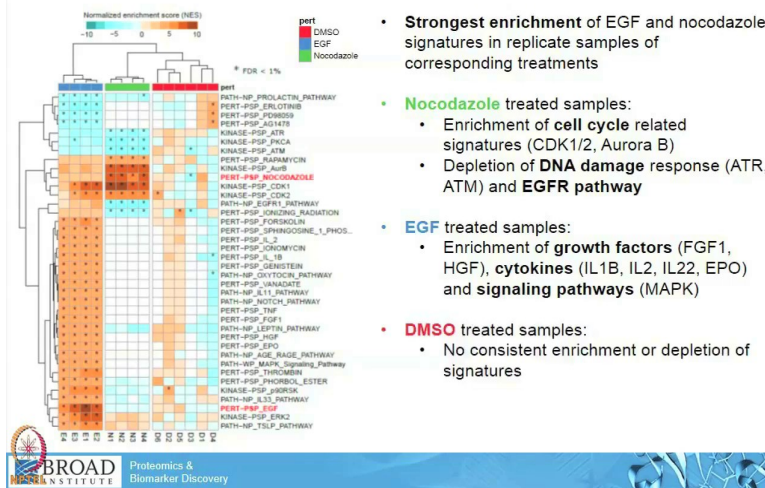
**BROAD INSTITUTE** Proteomics & Biomarker Discovery

So, we testes that against a very well studied dataset or EGF stimulated HeLa cells has been published in couple of years back now in 2014. It is a very good like system biology data set tool test your computational tools. So, we pick that one because the author used you know in

nocodazole to mitotically arrested the HeLa cells and also EGF to stimulate you know phosphorylation lower life in general signaling.

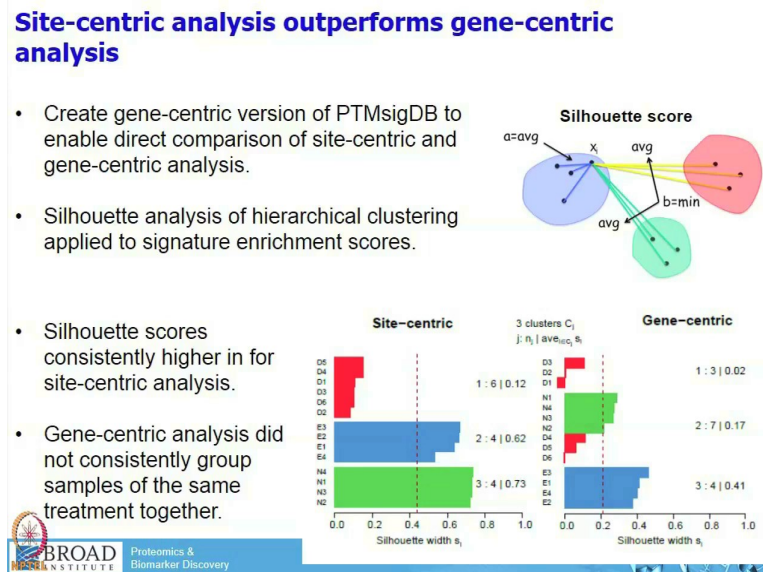
(Refer Slide Time: 13:17)

### PTM-SEA captured underlying signaling cascades in EGF- and nocodazole-treated HeLa cells



And both of these signatures are in our database. So, this was our benchmark data sets you know. If you can pick up these signatures, we I think we are in a good track that what we are doing is the right way. And what you are looking here is the heat map of enrichment scores. So, each row here is enrichment score of a signature and here I liked different experimental conditions. So, where this DMSO, blue is EGF, green is nocodazole measured in different number of replicates right. So, here is four replicates, four replicates here, six replicates. And luckily, we see that in nocodazole here in green the highest enrichment actually we observe in nocodazole treated samples. So, that is a good, but thing in the same as tool for EGF, right.

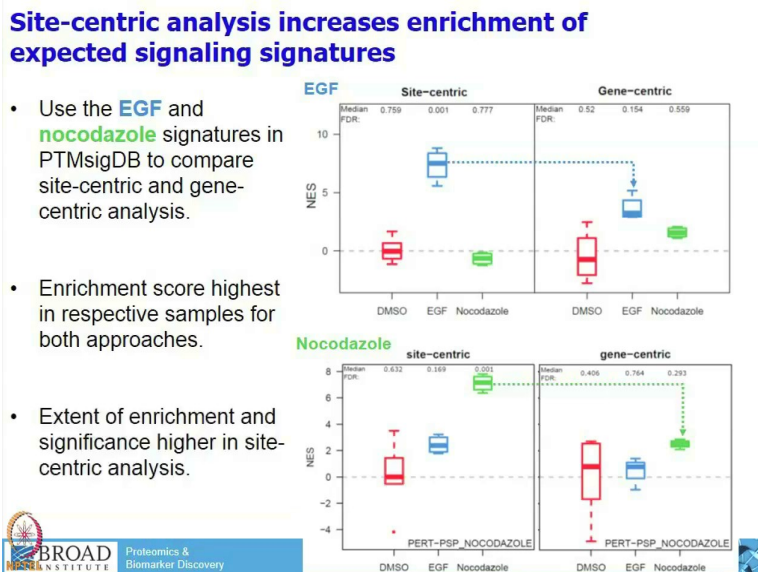
(Refer Slide Time: 14:09)



Now, so in the controlled sample, we do not see any consistent enrichment. So, this was our kind of you know global approach to prove that you know we are in a good track, and then we specifically focused on the clustering matrix. So, how well do these additional score cluster our data. And we compare that to a gene-centric approach. So, here in the left you look at this is how clean our clustering is if you do a site-centric approach, and this is how clean our clustering is if we first projected to genes, then do the same type of analysis. And we see that you know in the site-centric type of analysis we see very clean clusters. So, the higher these bars are you know look better represented if is my clustering.

So, now, so what we can see in the gene-centric space is that samples from you know from the control in a DMSO are clustering together which is not would be, what we would expect.

(Refer Slide Time: 15:06)



Now, see if you just look at these EGF and nocodazole signatures alone, so I am to compare the signatures scores across these different treatments. And when compare site-centric and gene-centric, we definitely see that you know we see a higher enrichment compared to gene-centric approach in this site-centric approach for EGF and also for nocodazole. So, we can pick up what we have a better signal of our of the biological pathways if we do with site-centric compared to gene-centric.

(Refer Slide Time: 15:41)

### PTMsigDB and PTM-SEA / ssGSEA are available on GitHub and GenePattern

<https://github.com/broadinstitute/ssGSEA2.0>

<https://tinyurl.com/PTM-SEA-GP>

The figure shows two screenshots. The top one is a GitHub repository page for 'ssGSEA2.0/PTM-SEA' with a README file. The bottom one is a screenshot of the GenePattern web interface, showing the 'PTM-SEA' analysis module with various input fields and file upload buttons.

(Refer Slide Time: 15:46)

## Pathway Enrichment Analysis: From Gene lists to Pathways

TABLE III  
Computational resources for pathway and gene ontology enrichment

Name	URL	Reference	Remarks
DAVID	<a href="http://david.abcc.ncifcrf.gov/">http://david.abcc.ncifcrf.gov/</a>	(163)	GO/Pathway annotation and enrichment
GoMiner	<a href="http://discover.nci.nih.gov/gominer/">http://discover.nci.nih.gov/gominer/</a>	(204)	GO analysis
GSEA	<a href="http://software.broadinstitute.org/gsea/">http://software.broadinstitute.org/gsea/</a>	(123)	Identifies pathways/GO terms with gene enrichment based on gene/protein ranking
InnateDB	<a href="http://www.innatedb.com/">http://www.innatedb.com/</a>	(205)	GO/Pathway annotation and enrichment, visualization
KEGG Atlas	<a href="http://www.kegg.jp/kegg/atlas/">http://www.kegg.jp/kegg/atlas/</a>	(206)	Pathway enrichment, visualization
LINCS	<a href="http://www.lincsproject.org/">http://www.lincsproject.org/</a>	(207)	Identifies common perturbation networks based on drug treated human cell lines
PHOXTRACK	<a href="http://phoxtrack.molgen.mpg.de/">http://phoxtrack.molgen.mpg.de/</a>	(167)	Modified GSEA approach focused on phosphosite-level profiling
SPIA	<a href="http://vortex.cs.wayne.edu/ontoexpress/">http://vortex.cs.wayne.edu/ontoexpress/</a>	(166)	GO/pathway enrichment modified to take into account pathway topology
WebGestalt	<a href="http://www.webgestalt.org/">http://www.webgestalt.org/</a>	(164)	GO/Pathway annotation and enrichment

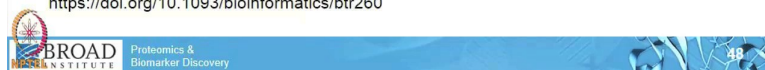


I think now I am going to where up here. So, these pores, so these tools are available on github and gene pattern. There is many other tools that do similar kind of things like phoxtracks that specifically look at kinase substrate and the actions. And you know, there is no other database that can do pathway or perturbation analysis, but there is other databases that do kind of substrate analysis as we have thought a lot that in this morning. And tomorrow I think you are going through here where to start and so on and so forth.

(Refer Slide Time: 16:17)

## References

- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature Protocols*, 4(1), 44–57. <https://doi.org/10.1038/nprot.2008.211>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a., ... Mesirov, J. P. (2005). **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Susan, E., Dunn, I. F., ... Hahn, W. C. (2010). **Systematic RNA interference reveals that oncogenic KRAS- driven cancers require TBK1**, 462(7269), 108–112. <https://doi.org/10.1038/nature08460>. Systematic
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). **The Molecular Signatures Database Hallmark Gene Set Collection**. *Cell Systems*, 1(6), 417–425. <https://doi.org/10.1016/J.CELS.2015.12.004>
- Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. (2011). **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics*, 27(12), 1739–1740. <https://doi.org/10.1093/bioinformatics/btr260>



So, and with this I want to end my thought. Again here I have put some references, I am open for one or two questions before Bing is going to talk about. Any questions?

Student: In PTMs sig database so, how do you are taking care of isoforms, in uniform number which you have shown here.

Yes.

Student: So, if we are some phospho data if we have 12 isoforms. So, there will be hyphen 1, 2. So, have like a can we give out data in that format.

So, we recommend to use the sequence when those which I know that you do not have any database and your data set, but the sequence you know just represent a phosphosite by its sequence window, but franking sequence is much more robust identifier.

Student: In this is tool also be applicable.

Yes.

Student: In this tool also we have to give the sequence?



We do not have to, but it is you are on safer side if you do so.

Student: Ok, thank you.

(Refer Slide Time: 17:23)

**Points to Ponder**

- Gene Set Enrichment Analysis (GSEA) can be used for pathway analysis for all the different kinds of PTMs.
- Three signature categories of PTMSigDB are perturbational signatures, signatures of molecular signaling pathways and signatures of kinase-substrate interactions.
- PTM-SEA (modified version of GSEA) contain curated of PTM sites using GSEA



So, today in conclusions we learnt about GSEA and how it can be used for pathway analysis for all the different kinds of PTMs. We also learnt that all the pathways are gene-centric, all the databases are gene curated which may dilute your efforts. Hence we need a PTM curated database for proper analysis of PTM data studies.

We also heard that why different ids and different isoforms curation is important, but very challenging. We also heard about the three signature categories of PTMSig database for example, perturbational signatures, signatures of molecular signaling pathways, and signature of kinase substrate interactions. Curation of PTM sites using GSEA can be done manually by semi-automated or fully automated function. Based on this, they have also made PTM SEA or the modified version of GSEA to look at the signatures of PTMs.

We also saw with an example that site-centric data grouping is more efficient and properly grouped as compared to the gene-centric. The next lecture is going to be hands on exercise by Dr. Karsten to show us how one can use GSEA for pathway enrichment.

Thank you.