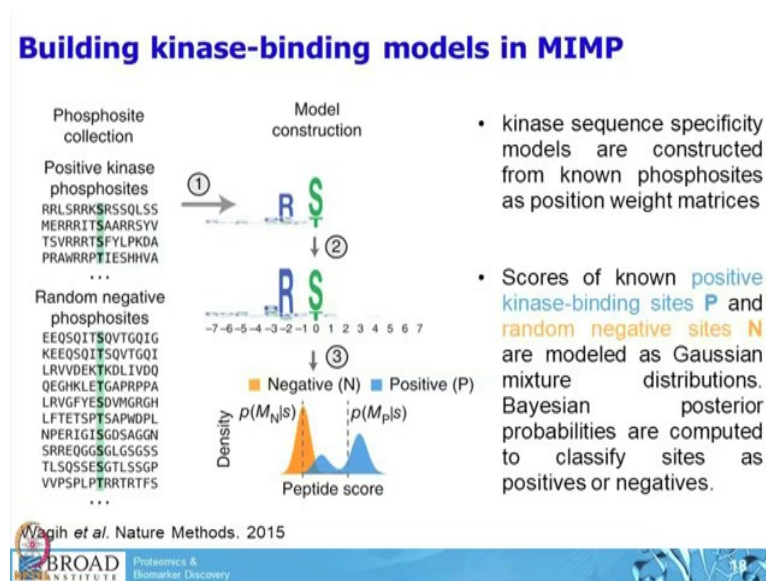


**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivatsava**  
**Dr. Karsten Krug**  
**Department of Biosciences and Bioengineering**  
**Broad Institute of MIT and Harvard**  
**Indian Institute of Technology, Bombay**

**Lecture – 54**  
**Mutations and Signaling - II**

Welcome to MOOC course on Introduction to Proteogenomics. In the last lecture, you were introduced to the effect of mutations on gene expression and how they could alter the signaling pathways. You were also taught about how these mutations can lead to the modifications of phosphosites and play role in downstream signaling processes. Today's lecture by Dr. Karsten Krug is a continuation of the last lecture, and he will cover the use of two important tools active DB and MIMP.

(Refer Slide Time: 01:03)



So, we are going to go quickly through the different steps here, I mean here the idea is to get the principle how it works. So, if you do not understand like all the details that is that is not cool shortly, you know to follow a hands on session. But I mean this is again very similar to what we have just looked at. So, you have two sets of phosphosite collection. So, one is ok, so this is actually how we built the model here. So, one is all positive kinase phosphosites. So, these are known, substrates and then to get a like a background data set or like a negative distribution it choose randomnegative phosphosites. And then you construct your model and

then you use a Bayesian approach to you know to determine whether you are phosphosite that you that you are interested it is comes from a negative distribution or from the positive distribution.

There is you know money was talking about mixture, modeling yesterday and this is kind of a related approach here.

(Refer Slide Time: 02:05)

### Amino acid frequencies calculated by Position Weighted Matrices (PWM)

- Construct one PWM for each kinase using known binding sites

*S* ... set of *n* known binding sites of length *m* for kinase

	1	...	<i>j</i>	...	<i>m</i>										
1	K	R	K	A	A	V	L	S	D	S	E	D	E	E	K
...	E	A	D	E	E	D	V	S	E	E	E	A	E	S	K
...	L	R	S	R	G	R	A	S	P	G	G	V	S	T	S
...	P	P	E	E	E	N	E	S	E	P	E	E	P	S	G
<i>k</i>	S	L	L	G	P	G	P	S	P	P	S	A	L	T	P
...	S	K	I	L	L	V	D	S	P	G	M	G	N	A	D
<i>n</i>	P	S	S	R	A	E	S	P	G	P	G	S	R		

$$f_{ij} = \frac{1}{n} \sum_{k=1}^n \lambda_i(s_{kj}) + \epsilon \quad \lambda_i(q) = \begin{cases} 1, & \text{if } i = q \\ 0, & \text{otherwise} \end{cases}$$

$\epsilon$  = background probability of the amino acid multiplied by 0.01

Wagih *et al.* Nature Methods. 2015

So, in this case amino acid frequencies are calculated by what they call a position weighted matrices. Again, it is very similar to what we have looked at, but. So, it is the relative frequency of each amino acid at each position. So, again we have the positions on the x axis and amino acids on the y axis and they add some error term here. So, that is basically the only difference. So, from and from that you can already calculate these sequence windows here.

(Refer Slide Time: 02:41)

### Measuring similarity between phospho sequence and PWM

- Given a phosphosite sequence  $q$  of length  $m$ ,  $q = q_1, \dots, q_m$ , calculate a **Matrix Similarity Score (MSS)** between  $q$  and the **PWM** of kinase  $K$
- MSS measures **information content** of each sequence position and normalizes against the highest and lowest relative frequencies per position in the PWM:

$$MSS = \frac{current - min}{max - min}$$

$$current = \sum_{j=1}^m I(j) f_{q_j, j}$$

$$min = \sum_{j=1}^m I(j) f_j^{min}$$

$$max = \sum_{j=1}^m I(j) f_j^{max}$$

$$I(j) = - \sum_i f_{i, j} \log \left( \frac{f_{i, j}}{f_b} \right)$$

$f_b$  ... amino acid frequency in human proteome

phosphosite sequence  $q$

$q = \begin{matrix} 1 & \dots & j & \dots & m \\ q_1 & q_2 & q_3 & & q_m \end{matrix}$

PWM<sub>K</sub>

		Offset position														
		-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
Amino acid	1	A														
		C														
		D														
		...														
		V														
		W														
	Y															
	20															

$f_{ij}$

Walsh et al. Nature Methods. 2015

BROAD INSTITUTE Proteomics & Biomarker Discovery

So, in order to, yeah in order to calculate the score for yours for a specific phosphosite they came up with these with this matrix similarity score which is actually a concept that has been introduced, but for proteogenomics sequence analysis. So, again like the principle is you want to calculate you want to calculate a score given a phospho sequence how likely it is or how similar is it to my position weight matrix here.

So, this is, this you have for each kinase and you calculate a score for each phosphosite how likely or how similar is it to that motif and you know it is based on information content. Again, we do not have to go into too much detail here, but what you do you calculate the phosphosite take the score for your sequence and you subtract the minimum given your matrix and you scale it by the range.

(Refer Slide Time: 03:45)

### Bayesian approach to calculate kinase binding scores

- Calculate MSS for all **known binding sites P** and a set of **random negative phosphosites N**
- Resulting distributions  $M_P$  and  $M_N$  are used to train a Gaussian Mixture Model (GMM)
  - $M_P$  ... MSS scores of true positive kinase-bound sequences
  - $M_N$  ... MSS scores of randomly sampled nonphosphorylated sequences
- Using the Bayesian theorem, two probabilities can be calculated for any peptide score MSS score S:
  - $p(M_N|S)$  ... Prob that S belongs to  $M_N$
  - $p(M_P|S)$  ... Prob that S belongs to  $M_P$

Phosphosite collection

Positive kinase phosphosites

RRLSRK\$KSSQLSS  
MERRRIT\$AARSYV  
TSVRRIT\$FYLKDA  
PRARRP\$EESHVA  
...

Random negative phosphosites

EEGSIIT\$QVVGIG  
KEEGSIIT\$QVVGIG  
LRVDEP\$KQLLYQG  
QEGHLE\$TGAPPPA  
LRVGFY\$QVVGRRH  
LFTETS\$APDPL  
NPERIG\$GDSAGN  
SRREGG\$GLGSSS  
TLSSSE\$GTLSSGP  
VVPSLP\$RRTRIFS  
...

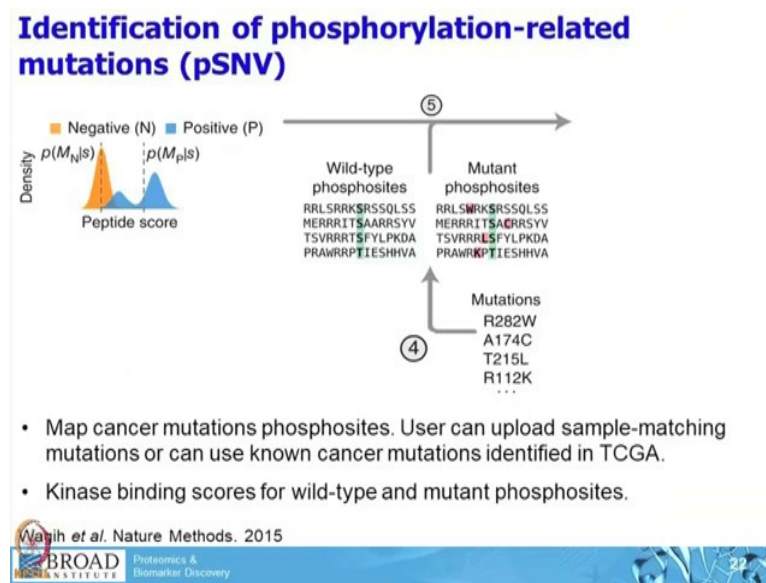
Wahj et al. Nature Methods. 2015

BROAD INSTITUTE Proteomics & Biomarker Discovery

So, in order to build up these Bayesian models what the software does it calculates exactly this score for all non-binding sites, this is one distribution. This is the blue distribution here and then in order to get a negative distribution it takes randomly sampled phosphosites from the human proteome, all right. And then it can calculate again all of these scores and then you end up having a negative distribution.

Then for a given this is what I just said, and then for a given peptide let us say, now, you are looking at your peptide that you measured you get a score which is maybe here and then using Bayesian theory we can calculate the probability whether this score is more likely to be derived from that distribution or from that distribution that is the entire concept.

(Refer Slide Time: 04:33)



So, now with this approach we can actually calculate kinase substrate specificity, given a phospho-peptide how likely it is that it has been phosphorylated by this particular kinase or maybe by another one. So now, how do we connect that to mutations?

So, you basically calculate these specificities for both versions of one is wild type phosphosite and then if there is mutation that happened in this sequence window here, we calculate the same specificity you know using the mutated sequence, right. And these again these are can be known cancer mutations from TCGA or you can upload your own mutations that is both possible.

(Refer Slide Time: 05:19)

### Prediction of mutation-induced phosphorylation loss and gain

Predict phosphorylation loss/gain by calculating joint probabilities:

$P_{\text{loss}} > 0.5$

- wild-type: positive kinase-binding site
- Mutant: negative binding site

$P_{\text{gain}}$

- wild-type: negative binding site
- Mutant: positive kinase-binding site

$$P_{\text{loss}} = P(M_{\text{P}}|S_{\text{w}}) \times P(M_{\text{N}}|S_{\text{m}})$$

$$P_{\text{gain}} = P(M_{\text{N}}|S_{\text{w}}) \times P(M_{\text{P}}|S_{\text{m}})$$

Wang et al. Nature Methods. 2015

BROAD INSTITUTE Proteomics & Biomarker Discovery

And so, then you basically look for differences in these probabilities. So, one is your wild type was a very likely that you know kinase AUR Kinase B phosphorylated this wild type sequence and now after the mutation how likely is it that still AUR Kinase B can phosphorylate that. So, this is how MIMP you know calculates the effect of, predicts the effect of mutations on these kinase binding motifs. So, this is the entire concept of membrane. I hope that we will make that work during the hands on.

(Refer Slide Time: 05:52)

### MIMP example

Gene*	Psite	Mut.	Psite Seq.	WT score	MT score	Prob.	Log ratio	Effect	PWM
ARID1B	1555	R1552K	NHISRAPS <sup>S</sup> PAS <sup>F</sup> QRS NHISKAPS <sup>S</sup> PAS <sup>F</sup> QRS	0.669	0.0431	0.975	-3.96	Loss	AKT1 
ARID1B	1555	R1552K	NHISRAPS <sup>S</sup> PAS <sup>F</sup> QRS NHISKAPS <sup>S</sup> PAS <sup>F</sup> QRS	0.776	0.153	0.921	-2.35	Loss	CAMK2A 
ARID1B	1555	R1552K	NHISRAPS <sup>S</sup> PAS <sup>F</sup> QRS NHISKAPS <sup>S</sup> PAS <sup>F</sup> QRS	0.585	0.144	0.882	-2.02	Loss	PRKACA 

BROAD INSTITUTE Proteomics & Biomarker Discovery

So, this is one example output here that you get from form in. So, basically this is the mutation that I have fed in. So, it is an arginine at 1555 which has been mutated into a lysine, effect in this particular gene here and here you actually see you know the sequence window so this is the wild type sequence window, it has this arginine at position minus 2 and after the mutation his arginine has been replaced by a lysine.

So, and here in this case we see that the predicted effect is actually a loss of phosphorylation exactly. So, the motif is gone. So, I do not see, yeah. So, the wild type sequence has a probability of 0.97 to be phosphorylated by AKT 1, which apparently recognizes the arginine at minus 3. After mutation AKT 1 is no longer able to phosphorylate it, to phosphorylate this site. So, this is how we read these kind of plots here. And then for the same site you have other kinase motifs. I mean what I have probably forgot to mention is these kinase motifs are very loose and not very specific, so many kinases share the same motif or at least parts of the motif right as you can see here you know this CAMK2A also is able to recognize an arginine it.

Student: What do you mean by phosphorylation is lost?

Loss means AKT cannot phosphorylate this phosphosite any more. So, that is just an example. I mean we will go through more examples during hands on I hope. So, here you see, so these are the scores that we have calculated a couple of slides back for the wild type sequence its 0.6 um. For the wild type sequence and this is for a mutated sequence almost 0, right and you know this difference kind of determines the probability. Further questions? Before I move on to.

Student: Is this the real data or are you showing the example data

This is actual data from a patient. So, the question was whether this is real data or like some example data, but this is a TCGA sample that I have used here.

Can you use mic please?

Student: Because the R and K are mostly of the same nature, so it is highly unlikely that this replacement will lead to loss of function, phosphorylation. Both are positively charged and the protein needs one positive charge to balance the negative charge on phosphate, so it is highly unlikely that this R to K replacement will lead to loss of phosphorylation

So, what you are saying is it is very unlikely that this is a loss because it is too unspecific.

Student: No, because it is R to K. Suppose if R to E happens or R to A or P anything, then that may lead to the loss but in this case it is unlikely to happen. That is why I am asking if is real data or just an example data

I mean the data that goes into this analysis is real data, but I mean you know keep in mind these are all predicted events, right. So, we do not know.

Student: So this is the predictive data?

Of course, yeah. So, I mean using this computational framework which I just you know tried to explain to you we calculate giving this data we calculate the probability you know that this might happen given our statistical framework. So, that is not an experimental proof that this phosphorylation site is actually loss, right. And also you must you know that we do not know these kind of sequence motifs for all kinases.

So, this model has been done on, 120 kinases or so, right for which we actually have very highly curated high quality substrate sites. So, it might very likely be that this phosphorylation you know site after it has been muted it can be recognized by any other kinase that we might not know the motif of, all right. Again, this is all computational it is all predicted, but you know it is a way to kind of approach these kind of relationships between mutations and signaling.

Student: What is the basis? So, if we some validation on this prediction it would be more better.

Sure, but you know I am not the one who is going in the wet lab. So, that is your part.

Sure. So, I mean what goes in here on phosphosites that have met would that it has been measured like in the lab, you know that could have been measured in the sample and in a patient sample in this case and also we have genomics data it comes from the same patient right, mutation somatic mutations in a patient. So, these are the data that goes into this analysis or there is already mentioned.



(Refer Slide Time: 11:27)

### MIMP implementation

- Web server:  
<http://mimp.baderlab.org/>
- R-package:  
<https://github.com/omarwagih/rmimp>

```
install.packages("devtools")  
library("devtools")  
install_github("omarwagih/rmimp")  
library(rmimp)
```




So, you can access them on our web server. So, I encountered a couple of difficulties while trying to run analysis on a web server. So, that is why actually I decided to use the R-package, ok.

(Refer Slide Time: 11:44)

### ActiveDriverDB – database resource to study mutations and PTMs

- Proteogenomic database that annotates disease mutations and population variants through the lens of PTMs.
- Integrates two major types of omics data:
  - >385,000 PTM sites
  - ~3.6 million substitutions (nsSNVs)
- Prediction of network-rewiring impact of mutations by analyzing gains and losses of kinase-bound sequence motifs using MIMP

Grassowski et al. 2018 *Nucleic Acids Res.*



Let us move on. So, the second tool I briefly want to talk about in the next 10 minutes or so is it is more like a database its it is not really like a tool, but this database also enables you to upload your own mutations and see what you know could possibly be the effect of these mutations on phospho signaling events. Again, so these are all predictions right. So, we there

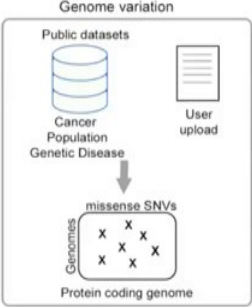
is no validation whatsoever. So, active driver DB is very very well developed and very complete data database proteogenomic, database that annotates disease mutations, but also population variants and relates them to PTMs.

So, we have like two major types of omics data which has been integrated here. So, one are like post translational modification sites like more than 385,000 and also like more than 3.5 or 3.6 million, SNPs. And yeah, so basically it also predicts network rewiring impact of mutations and also it uses actually the MIMP software we just talked about, so it basically comes from the same lab.

(Refer Slide Time: 13:03)

### ActiveDriverDB – three types of human genome variation datasets

- 1) Somatic cancer mutations of ~9000 tumor samples of 34 types from exome sequencing by the TCGA
- 2) Inherited disease mutations from the ClinVar database
- 3) Interindividual genome variation of the human population includes the 1,000 Genomes Project with >2500 genomes and the ESP6500 with >6500 exomes.



The diagram, titled 'Genome variation', shows the flow of data into the database. It starts with 'Public datasets' which include 'Cancer Population Genetic Disease' (represented by a database icon) and 'User upload' (represented by a document icon). An arrow points from these sources to a box labeled 'missense SNVs'. This box is situated between 'Genomes' (on the left) and 'Protein coding genome' (on the right). Inside the 'missense SNVs' box, there are several 'X' marks representing mutations.

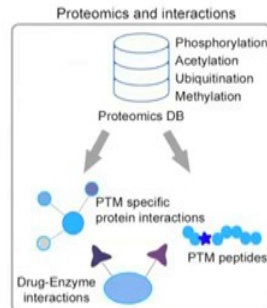
Kraśkowski et al. 2018 *Nucleic Acids Res.*  
BROAD INSTITUTE Proteomics & Biomarker Discovery

So, there are 3 types of human genome variation data sets that goes in here. So, one is TCGA, the other are disease mutations that come from the Clinvar database and also mutations from the one thousand genomes project. So, it again it uses these publicly available datasets, but also enables a user to upload your own mutation files. And again, this is something that we are trying to do to in hands on.

(Refer Slide Time: 13:34)

## ActiveDriverDB – proteomics and interaction data

- 1) Experimentally determined human PTM sites retrieved from:
  - PhosphositePlus
  - Phospho.ELM
  - Human Protein Reference Database (HPRD)
- 2) Site-specific protein-protein interactions of substrate proteins and upstream enzymes (primarily kinases) are also included from these databases
- 3) DrugBank database to annotate drugs that target upstream enzymes of PTMs



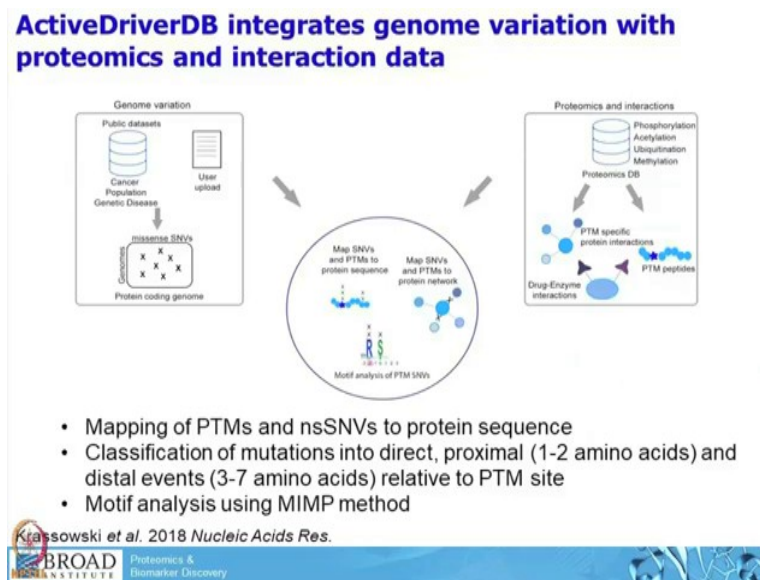
Krassowski *et al.* 2018 *Nucleic Acids Res.*

BROAD INSTITUTE  
Proteomics & Biomarker Discovery

So, in terms of proteomics and interaction data it uses you know data that is available in the several different databases like PhosphositePlus, Phospho EML and Human Protein Reference Database. And in terms of PTMs severe we are not only looking at phosphorylation here we have also looking at acetylation, ubiquitination and methylation and also pulls in information of you know drug enzyme interactions and also PTM site specific, protein interactions.

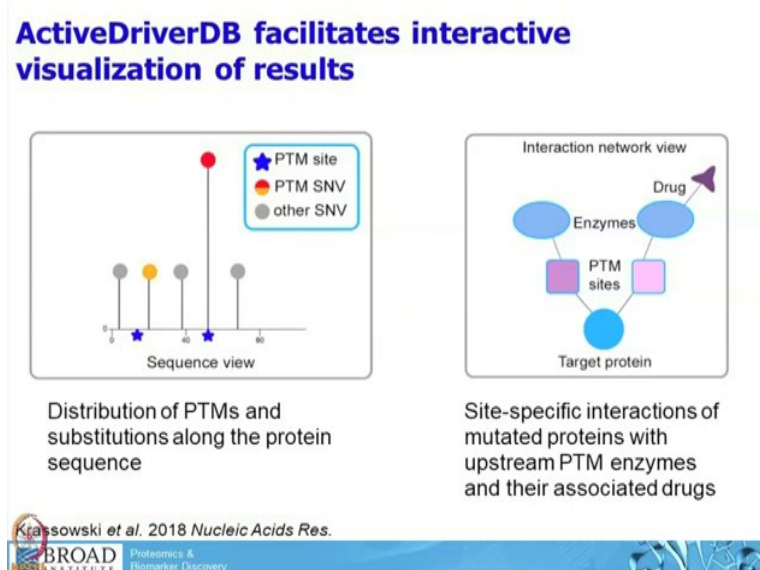
So, these are you know, these information all passed from publicly available databases and now are gathered in one place and it is a very nice very intuitive user interface which enables you know everybody who does not have the computational skills to do these kinds of analysis tool to do those.

(Refer Slide Time: 14:28)



So, what you do is very similar to what we did you know what we have learned in the previous part of the talks we look at. So, we combine genome variation with PTM data and to basically map SNVs and PTMs to protein sequence and look what are the non-synonymous events in your protein and you calculate motifs or using you know the effect of mutations on the motif along that. But again, so here we are not only looking at phosphorylation, but also other modifications.

(Refer Slide Time: 15:07)



So, it is very interactive this entire web webpage there is many different ways how to view your data. So, one is the sequence view we have they can look at your protein sequence you know along the x axis and then you have like all mutation events or PTM events you know highlighted in this in this protein sequence you can actually look at distribution of PTMs along the sequence. And it also has a way to build up these interactive network views between proteins, kinase and drugs, and PTMs and so on and so forth.

(Refer Slide Time: 15:50)



**ActiveDriverDB**

- Web server:  
<https://www.activedriverdb.org/>
- GitHub:  
<https://github.com/reimandlab/ActiveDriverDB>

 **BROAD** INSTITUTE Proteomics & Biomarker Discovery

So, this is something again we are going to explore during the hands on session. So, there is a web server available which works pretty well and also there is a GitHub page where you can access the tool you can download it, you can fork it, you can modify it and run it locally on your computer.

(Refer Slide Time: 16:12)

## Summary

- Mutations affect signaling networks mediated by PTMs
- Millions of mutations have been identified to be associated with cancer
- Integrative analysis of mutations and PTMs has the potential to provide insights into the molecular consequences of associated mutations
- Among other tools, MIMP and ActiveDriverDB are promising approaches for integrative analyses of PTMs and mutations



So, to sum up mutations effect signaling networks, I do not think that there is any doubt that this is not going to happen, but that this is not happening. So, there is millions of mutations that have been identified that are associated with cancer, but we still do not know the exact molecular mechanism. So, how? So, what is the association between phenotype and genotype or genotype and phenotype?

So, when I think that integrative analysis of these mutations and PTMs has a great potential to insights into these kind of relationships. And here, I just pointed out two of these tools as many out and there is also many you know ongoing efforts I know that Bing's group is also working a very nice tool that uses the neural networks who look into these kind of relationships. So, I just picked these two because there is they are kind of easy to use way into intuitive to use and the output is very easy to interpret and so on and so forth. This does not mean that these are the best tools, right.

(Refer Slide Time: 17:18)

## References


Schwartz, D., & Gygi, S. P. (2005). **An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets**. *NATURE BIOTECHNOLOGY VOLUME*, 23. <https://doi.org/10.1038/nbt1146>

Wagih, O., Reimand, J., & Bader, G. D. (2015). **MIMP: predicting the impact of mutations on kinase-substrate phosphorylation**. *Nature Methods*, 12(6), 531–533. <https://doi.org/10.1038/nmeth.3396>

Reimand, J., & Bader, G. D. (2013). **Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers**. *Molecular Systems Biology*, 9(637), 637. <https://doi.org/10.1038/msb.2012.68>

Krassowski, M., Paczkowska, M., Cullion, K., Huang, T., Dzieladze, I., Ouellette, B. F. F., ... Reimand, J. (2018). **ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins**. *Nucleic Acids Research*, 46(D1), D901–D910. <https://doi.org/10.1093/nar/gkx973>

Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., ... Mani, D. R. (2017). **Methods, tools and current perspectives in proteogenomics**. *Molecular & Cellular Proteomics : MCP*, mcp.000024.2017. <https://doi.org/10.1074/mcp.MR117.000024>




(Refer Slide Time: 17:22)

## References

Ryu, G.-M., Song, P., Kim, K.-W., Oh, K.-S., Park, K.-J., & Kim, J. H. (2009). **Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases**. *Nucleic Acids Research*, 37(4), 1297–1307. <https://doi.org/10.1093/nar/gkn1008>

Hu, J., Rho, H.-S., Newman, R. H., Zhang, J., Zhu, H., & Qian, J. (2014). **PhosphoNetworks: a database for human phosphorylation networks**. *Bioinformatics*, 30(1), 141–142. <https://doi.org/10.1093/bioinformatics/btt627>

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). **PhosphoSitePlus, 2014: mutations, PTMs and recalibrations**. *Nucleic Acids Research*, 43(D1), D512–D520. <https://doi.org/10.1093/nar/gku1267>





And here I just put in a lot of references for all of the papers and tools that I have used here in these my talk.

(Refer Slide Time: 17:29)

**Points to Ponder**

- Active DB annotates disease mutations and population variants considering PTMs.
- Active DB can predict network-rewiring impact of mutations analyzing gain and loss of kinase-bound sequence from MIMP.
- It can be used to integrate genome variations with data from proteomics and interaction analysis.



Today, you were introduced to two tools namely MIMP and active driver DB, which have been used to explore the mutations and their effects on a few cancers. You are also shown how MIMP predicts the gain or loss of a mutation by calculating a probabilistic score at different locations in a given sequence. Active DB is another proteogenomic database that annotates disease mutations and population variants as PTMs.

It also provides information about various molecular interactions from publicly available databases. There many more tools freely available and you are encouraged to try out using them and then you will explore many features, which could really help you for much better and deeper understanding. In the next lecture, we will look at pathway enrichment using a few widely used tools and software.

Thank you.