Supplementary material and methods

The peptide samples of *C. elegans* (15 mg) and *S. cerevisiae* (20 mg) were prepared as described recently[1]. The dried down peptides were resolubilized to a final concentration of 1 mg/ml in off-gel electrophoresis buffer containing 6.25% glycerol and 1.25% IPG buffer (GE Healthcare). The peptides were separated on pH 3-10 IPG strips (GE Healthcare) with a 3100 OFFGEL fractionator (Agilent) as previously described[2] using a protocol of 1 hour rehydration at maximum 500 V, 50 mA and 200 mW followed by the separation at maximum 8000V, 100 mA and 300 mW until 50 kVh were reached. After iso-electric focusing, the fractions were concentrated and cleaned up by C18 reversed-phase spin columns according to the manufacturer's instructions (Harvard Apparatus).

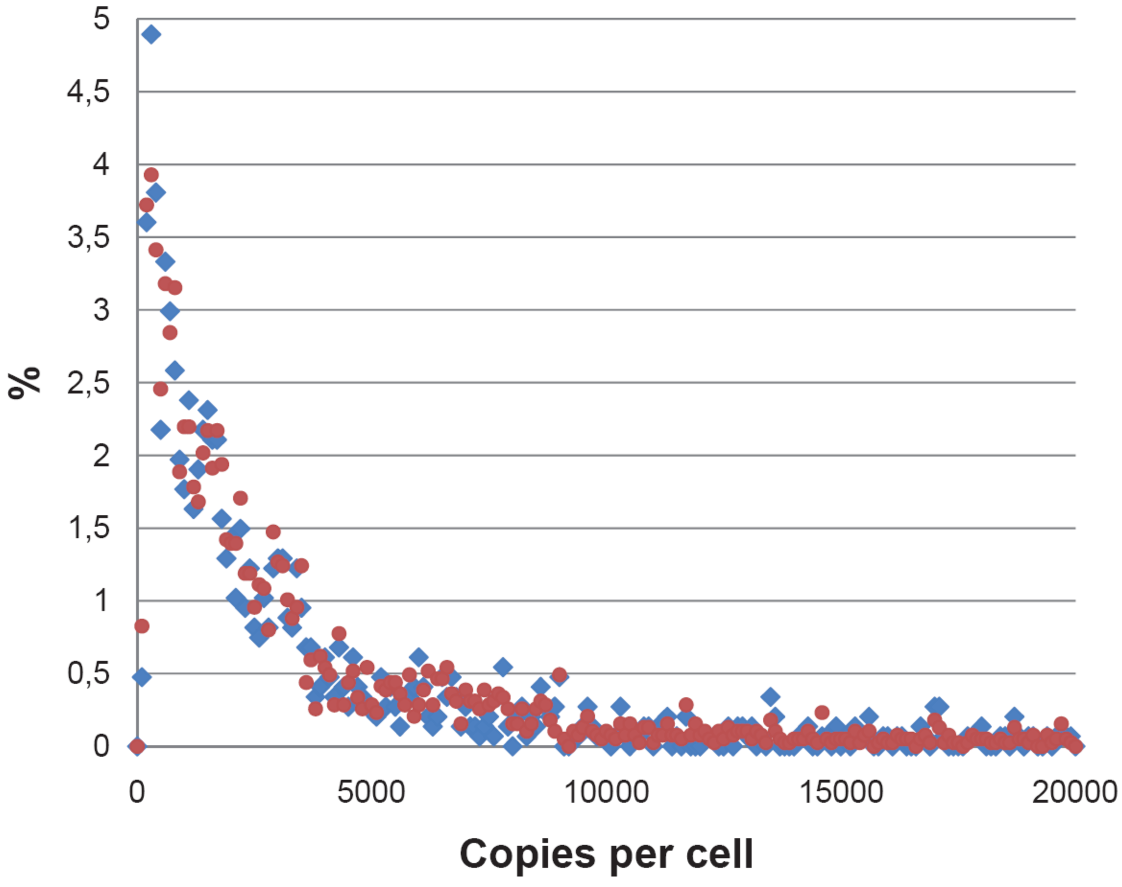
Phosphopeptides in each fraction were isolated as described in[3] and [4], analyzed using LC-MS/MS and database searches (based on the SGD database (20. Oct 2007) for yeast[5], the wormpep183 database for worm[6] and the IPI database v3.23 for human[7]) were done as described in[1]. We made all the data in PhosphoPep searchable by spectral matching through SpectraST[8]. (http://www.phosphopep.org/spectrast/index.php). Specifically, for each distinct phosphopeptide identified in this study, all corresponding MS2 spectra were collapsed into a single consensus spectrum. Unknown query spectra can then be identified by spectral searching against the library of phosphopeptide consensus spectra. SpectraST can be used both as a web interface in PhosphoPep, and as a stand-alone application released as part of the TPP suite of software[1, 9]. The identified (phospho)peptides were mapped to all possible proteins/gene products present in the corresponding database.

*For **Table 1*** the "total phosphorylation sites" includes all sites of phosphopeptides with a dCn >= 0.0 as computed by Sequest[10]. A phosphorylation sites was considered to have an assigned site if a dCn (between the first and second Sequest output entry) threshold was exceeded[1, 11]. In case of the *D. melanogaster* dataset a dCn >= 0.1 as computed by Sequest[10] and for the *S. cerevisiae*, *C. elegans* as well as the human dataset a dCn >= 0.125 as computed by Sequest[10] was used to define a phosphorylation site as assigned.

For Supplementary Figure 1A the protein copies per cell were taken from the publication by Ghaemmagami *et al*[12].

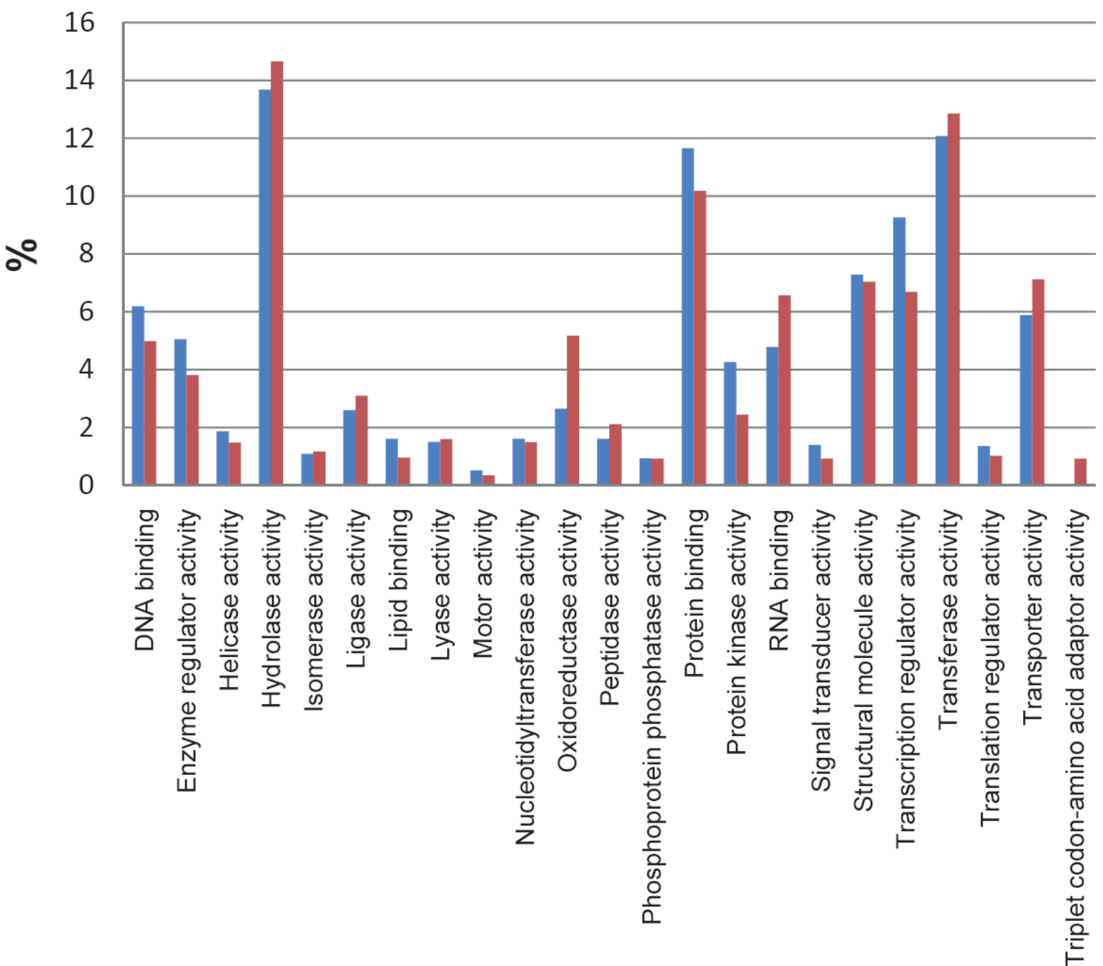
For **Supplementary Figure 1A and 1B** both the “all proteins” as well as the GO annotations were taken/retrieved from the yeast SGD database[5].

For the ***C. elegans*** dataset we omitted the GO analysis as for the 2,959 proteins identified only for 348 a GO annotation “molecular function” and for 373 a GO annotation “biological process” is given. Also so far no dataset is published which accurately predicts or describes protein abundances in *C. elegans* and therefore an analogous analysis as for Supplementary Figure 1A was omitted as well.



Supplementary Figure 1A

A comparison of the yeast phosphoprotein abundance (blue) with the abundance of most proteins of the yeast proteome (red) as determined by Ghaemmaghami *et al*[12]. Proteins with more than 20,000 copies per cell are not displayed. The distribution of proteins with more than 20,000 copies per cell is nearly identical between the identified phosphoproteome and yeast proteome. The X-axis displays the protein copies per cell, the Y-axis the percentage of protein counts per copies per cell bin (bin size 100) divided by all proteins from the phosphoprotein or Ghaemmaghami *et al*[12] data set.



Supplementary Figure 1B

Fraction of identified phosphoproteins (blue) assigned to a given biological function according to gene ontology. As comparison all yeast proteins (red) assigned to a given biological function are shown.

Supplementary References

1. Bodenmiller, B. et al. PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* **3**, 139 (2007).
2. Heller, M. et al. Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.* **4**, 2273-2282 (2005).
3. Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B. & Aebersold, R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat. Methods* **4**, 231-237 (2007).
4. Bodenmiller, B. et al. An integrated chemical, mass spectrometric and computational strategy for (quantitative) phosphoproteomics: application to Drosophila melanogaster Kc167 cells. *Mol. BioSys.* **3**, 275-286 (2007).
5. Cherry, J.M. et al. Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387**, 67-73 (1997).
6. Rogers, A. et al. WormBase 2007. *Nucleic Acids Res.* **36**, D612-617 (2008).
7. Kersey, P.J. et al. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985-1988 (2004).
8. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667 (2007).
9. Keller, A., Eng, J., Zhang, N., Li, X.J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 2005 0017 (2005).
10. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989 (1994).
11. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285-1292 (2006).
12. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).
13. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383-5392 (2002).
14. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207-214 (2007).

New PhosphoPep help page and tutorial for biologists

Buttons used in PhosphoPep

	View KEGG pathways for this protein
	Start cytoscape network with this protein
	View orthologs/homolog information
	Search for protein interaction networks in String
	Look up protein information in Peptide Atlas
	Search protein sequence at Scansite

Importantly, as the current knowledge about cellular pathways is far from complete, only a portion of the phosphoproteins can be placed into their pathways. This partial knowledge also applies for the orthologous protein information as well as to the prediction of a kinase for a given phosphorylation sites.

Scores and numbers used in PhosphoPep

PeptideProphet	When interpreting tandem mass spectrometry data, it is crucial to determine if an identification is correct. The PeptideProphet computes a probability of a given fragment ion spectrum to be correctly assigned to a peptide sequence by a given database search algorithm and assigns a score accordingly. The range of the score is from 0 (worst) to 1(best). Depending on the dataset or database the probabilities can slightly vary at a given threshold/Score.
Tryptic Ends	As we analyze peptides in our tandem mass spectrometry experiments we have to digest the proteins using a protease. This is normally done by using trypsin. Trypsin cleaves after arginine and lysine but exhibits also some unspecific cleavage. 2 tryptic ends means that both ends were specifically cut by trypsin.
Peptide Mass	Molecular mass of the phosphopeptide
DeltaCN	The deltaCn score (dCn) is a score computed by the Sequest algorithm which we use to interpret tandem mass spectra. Oversimplified, the dCn tells you how big the distance between the best database search hit is from the second (best) for a given tandem mass spectrum. In the case of phosphopeptides the dCn also correlates to the correctness of

	the phosphorylation site assignment within the phosphopeptide sequence.
# Obs	Number of times the phosphopeptide was identified in our experiments
# Mappings	Maps to # of gene models / maps to # of transcripts

How to assess the quality of a phosphopeptide identified using tandem mass spectrometry

In order to understand the basic methods of peptide identification using tandem mass spectrometry we strongly recommend studying the presentation which you can find under the link

http://www.proteomesoftware.com/Proteome_software_pro_interpreting.html

The presentation is very easy to understand and is one of the best introductions to proteomics we saw so far.

When phosphopeptides are analyzed using liquid chromatography – tandem mass spectrometry and the resulting spectra are assigned to phosphopeptide sequences using database search algorithms two types of error can occur. The first type of error is the miss-assignment of the fragment ion spectrum to a peptide sequence. The second type of error is the miss-assignment of the site of phosphorylation in an otherwise correctly identified phosphopeptide.

Here we explain how each of the errors was assed and how the users of PhosphoPep can use the computed scores and some simple rules to judge if a phosphopeptide was correctly identified and the site correctly assigned.

Is the phosphopeptide correctly identified?

As mentioned above, one type of error in the automatic interpretation of tandem mass spectra is the miss-assignment of the fragment ion spectrum to a peptide sequence. This type of error can be estimated by applying a statistical mixture model, PeptideProphet[13] and/or by using a decoy sequence databases[14].

All data loaded into PhosphoPep were assessed using both statistical tools and we already applied a stringent cut off on all data. Therefore the false positive content in the case of the fly data is about 2.6 % (for yeast, worm and human this number is similar). This means that if you don't apply any further filter criteria 1 out of 38 phosphopeptide entries is wrong. For bioinformatic large scale analyses this false positive rate is in most cases very acceptable, but for a biologist who wants to perform follow up experiments this can already be too high and therefore it is desirable to choose your own false positive rate. So how do you choose your own false positive rate?

One of the statistical tools to compute the false positive rate, the Peptide Prophet, computes a score (ranging from 0 (worst) to 1(best)). This score is displayed for every peptide in PhosphoPep. As mentioned above, we have already prefiltered the data, therefore the lowest PeptideProphet score you will find is 0.8 (~ 2.6 % false positive rate). The closer the score is to 1.0 the lower is the chance that you pick a wrongly identified phosphopeptide. For example, at a Peptide Prophet cut off of 0.99 approximately 0.2 % of all entries (equal or above this score) are false positive assignments (1 out of 500 phosphopeptide entries).

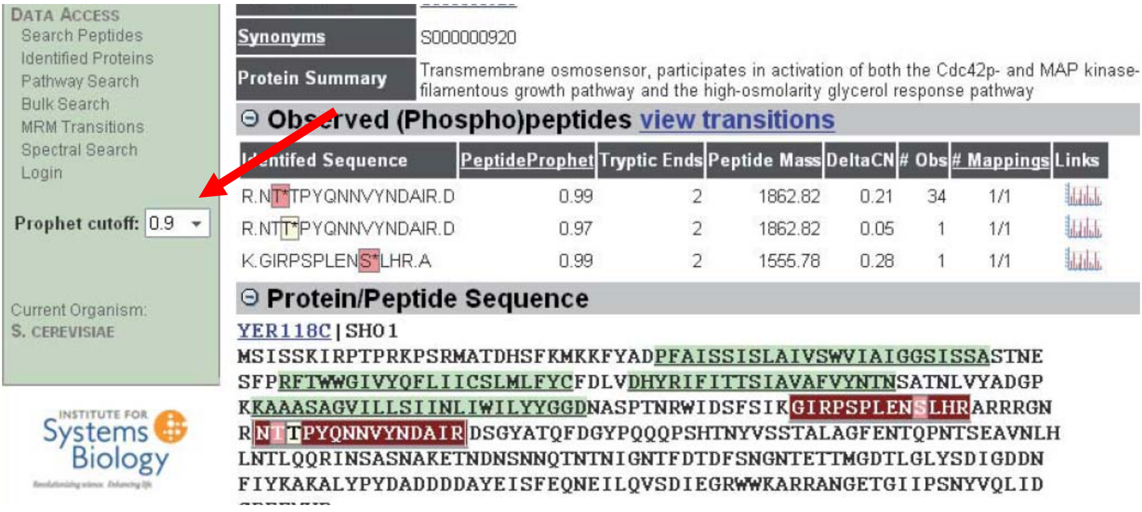
⊖ **Observed Phosphopeptides** view transitions

Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.TTS*SSFESEIK.S	0.92	2	1295.52	0.23	4	1/1	
K.SNGANRDS*SDLAPTLR.S	0.96	2	1753.78	0.16	19	1/1	
K.SNGANRDSS*DLAPTLR.S	0.84	2	1753.78	0.02	1	1/1	
K.RVS*DVLPK.R	0.95	2	993.52	0.21	6	1/1	

⊖ **Protein/Peptide Sequence**

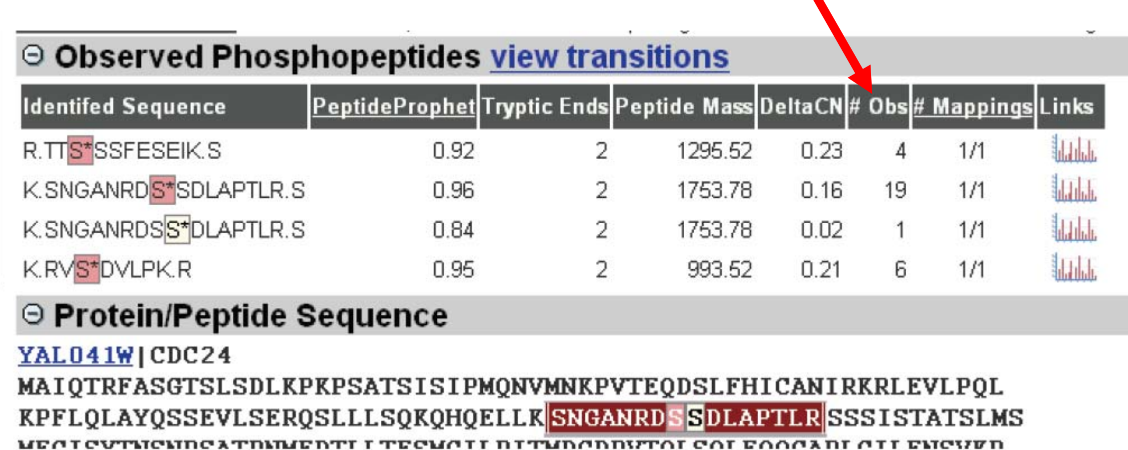
YAL041W|CDC24
MAIQTRFASGTSLSDLKPKPSATSISIPMQNVMNKPVTEQDSLFHICANIRKRLEVLPQL
KPFLQLAYQSSEVLSERQSLLLSQKQHQELLKSNGANRDSSDLAPTLRSSSISTATSLMS

With the button to the left you can choose the Prophet Score cut off on your own.



Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.NT*TPYQNNVYNDAIR.D	0.99	2	1862.82	0.21	34	1/1	
R.NTT*PYQNNVYNDAIR.D	0.97	2	1862.82	0.05	1	1/1	
K.GIRPSPLENS*LHR.A	0.99	2	1555.78	0.28	1	1/1	

One further criterium which increases the certainty that a phosphopeptide was identified correctly is the "# Obs" which tells you how often a phosphopeptide was identified in our experiments.



Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.TTS*SSFESEIK.S	0.92	2	1295.52	0.23	4	1/1	
K.SNGANRDS*SDLAPTLR.S	0.96	2	1753.78	0.16	19	1/1	
K.SNGANRDSS*DLAPTLR.S	0.84	2	1753.78	0.02	1	1/1	
K.RVS*DVLPK.R	0.95	2	993.52	0.21	6	1/1	

The chance that a phosphopeptide which was identified multiple times is wrong is lower than that of a phosphopeptides that was just identified once (but keep in mind that this is only a rule of thumb).

So taken together, if you choose a phosphorylation site for follow up experiments make sure that it has a high PeptideProphet score and was observed several times.

Is the site of phosphorylation correctly assigned?

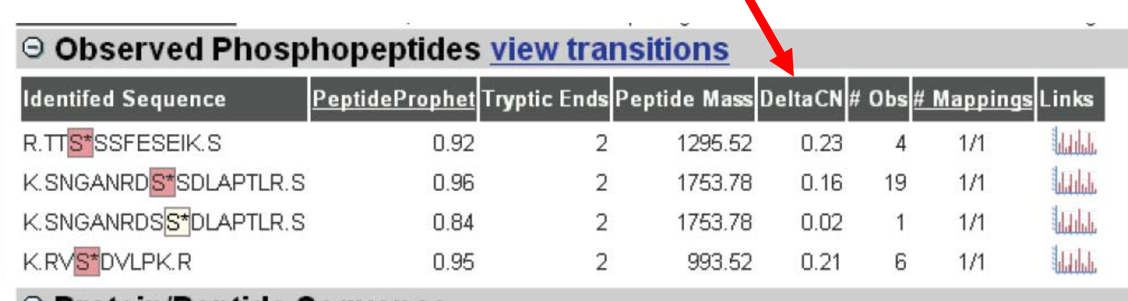
Often phosphopeptides are rich in serines and threonins which can sometimes puzzle the algorithm for the automatic interpretation of tandem mass spectra in regards to which serine/threonine was phosphorylated. Therefore another type of error connected to phosphopeptides identified using tandem mass spectrometry is the miss-assignment of the site of phosphorylation in an otherwise correctly identified phosphopeptide.

This error was estimated by comparing the search engine output scores for the potential phosphorylated forms of a peptide, assuming that any hydroxy-amino acid in a phosphopeptide could be phosphorylated. Based on this estimation we highlighted the phosphopeptides either red (high probability of correct assignment) or yellow (low probability of correct assignment).

As one typical approach to study protein phosphorylation is to mutate the site of phosphorylation either to an alanine or an aspartate it is advisable to ascertain that you choose the right amino acid. There are several steps you can take in order to assure that the site of phosphorylation was correctly assigned.

1) *Take a look at the dCn value.*

The first step to determine the certainty in the phosphorylation site assignment is to look at the dCn score (In short, the dCn score describes how much the number one hit in a database search differs from the second hit for a given tandem mass spectrum. Now if the first and second hits are the same phosphopeptide, but the algorithm has problems to unequivocally assign the site, the score will be very low).



⊖ Observed Phosphopeptides view transitions

Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.TTS*SSFESEIK.S	0.92	2	1295.52	0.23	4	1/1	
K.SNGANRDS*SDLAPTLR.S	0.96	2	1753.78	0.16	19	1/1	
K.SNGANRDSS*DLAPTLR.S	0.84	2	1753.78	0.02	1	1/1	
K.RVS*DVLPK.R	0.95	2	993.52	0.21	6	1/1	

⊖ Protein/Peptide Sequence

YAL041W|CDC24
MAIQTRFASGTSLSDLKPKPSATSISIPMQNVMNKPVTEQDSLFHICANIRKRLEVLPQL
KPFLQLAYQSSEVLSERQSLLLSQKQHQELLKSNGANRDSSDLAPTLRSSSISTATSLMS

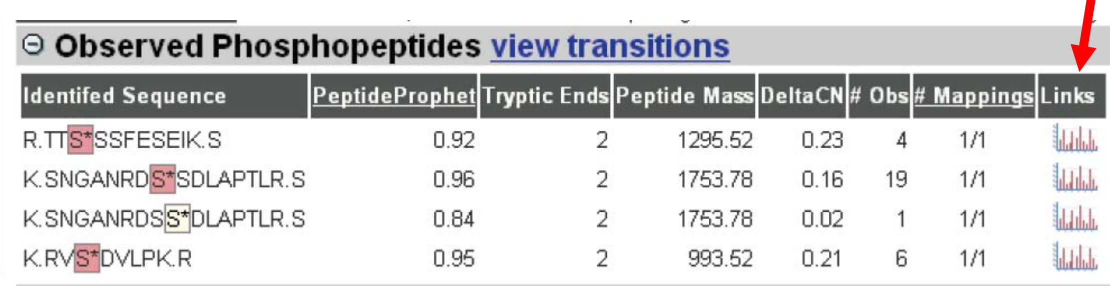
Again as a rule of thumb: The higher the dCn score the more certain is the phosphorylation site assignment. In addition, a score of dCn > 0.125 corresponds to a very high certainty that the site is correctly assigned.

Below a phosphopeptide is shown which was identified several times but the site of phosphorylation could never be assigned with high certainty. As a result the same phosphopeptide exists in several versions in PhosphoPep. Such agglomerations of the same peptide with many different phosphorylation sites is a hint that the site is not well assigned (but keep in mind, some places in proteins can serve as phosphorylation platforms and therefore the same peptide will exist in different phosphorylation forms).

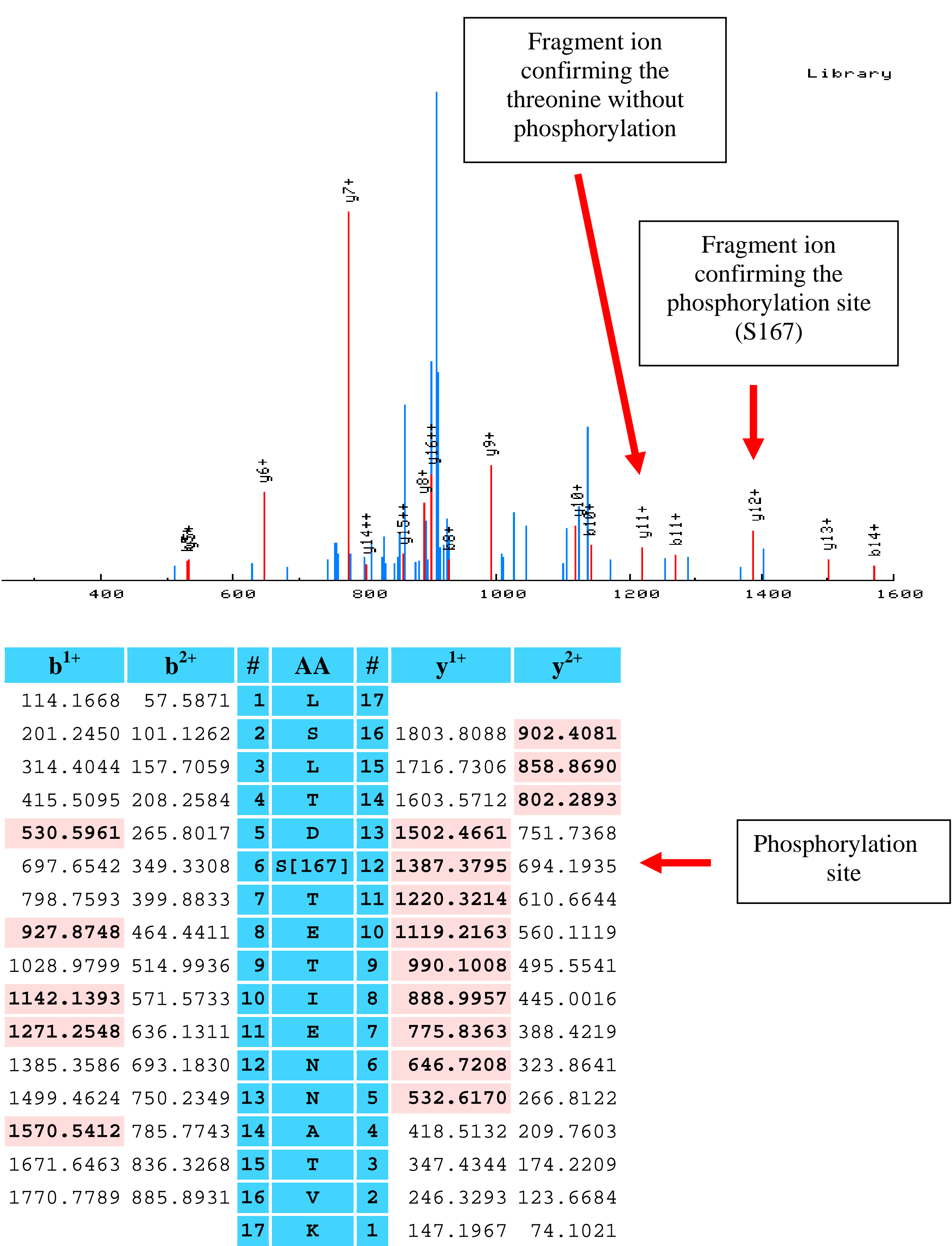
R.APSSVSMVS*PPPLHK.N	0.95	2	1613.76	0.12	1	1/1	
R.APS*SVS*MVSPPPLHK.N	0.96	2	1693.72	0.03	1	1/1	
R.APS*S*VSMVSPPPLHK.N	0.92	2	1693.72	0.08	1	1/1	

2) *Take a look at the tandem mass spectrum*

After having assessed the dCn value it is always advisable to take a look at the tandem mass spectrum of the phosphopeptide. You can open it by clicking on the symbol .



This will open a new window in which the tandem mass spectrum is displayed (See next page). In the upper window you see the tandem mass spectrum in which the fragment ion peaks are assigned with y-ion or b-ion together with a number (ion assignment nomenclature) as well as below the spectrum the amino acid sequence of the phosphopeptide (the phosphorylation site is assigned with a number behind the serine/threonine/tyrosine and is 167/181/243) is shown. Here you have to look for the following: Left and right of the amino acid sequence the fragment ion signals which were found and could be assigned in the tandem mass spectrum are highlighted in red. In our example the question is, if really the serine (at position 6) is phosphorylated LSLTDS$_{167}$TETIENNATVK or the adjacent threonine LSLTDST$_{167}$ETIENNATVK at position 7.

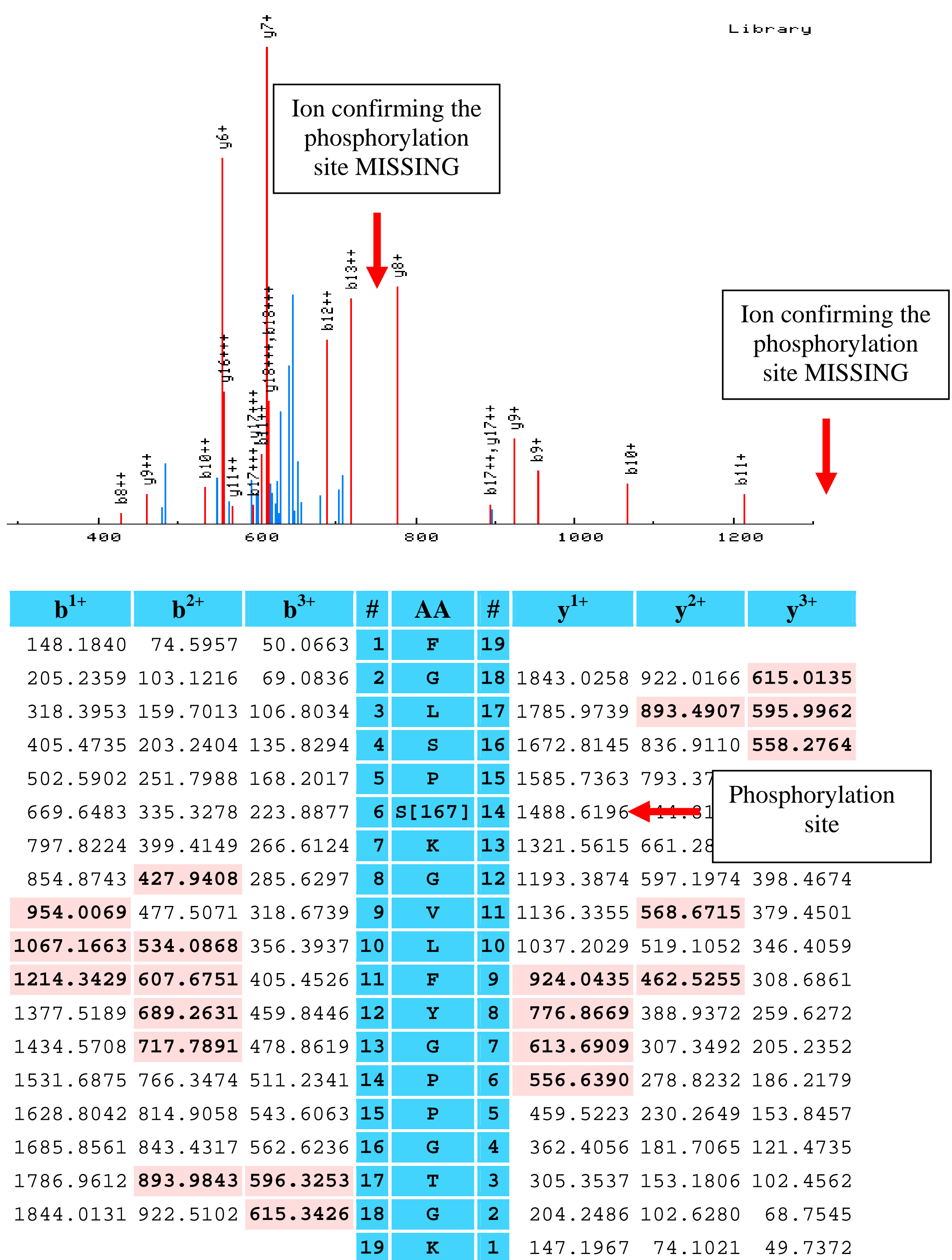


b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
114.1668	57.5871	1	L	17		
201.2450	101.1262	2	S	16	1803.8088	**902.4081**
314.4044	157.7059	3	L	15	1716.7306	**858.8690**
415.5095	208.2584	4	T	14	1603.5712	**802.2893**
530.5961	265.8017	5	D	13	**1502.4661**	751.7368
697.6542	349.3308	6	S[167]	12	**1387.3795**	694.1935
798.7593	399.8833	7	T	11	**1220.3214**	610.6644
927.8748	464.4411	8	E	10	**1119.2163**	560.1119
1028.9799	514.9936	9	T	9	**990.1008**	495.5541
1142.1393	571.5733	10	I	8	**888.9957**	445.0016
1271.2548	636.1311	11	E	7	**775.8363**	388.4219
1385.3586	693.1830	12	N	6	**646.7208**	323.8641
1499.4624	750.2349	13	N	5	**532.6170**	266.8122
1570.5412	785.7743	14	A	4	418.5132	209.7603
1671.6463	836.3268	15	T	3	347.4344	174.2209
1770.7789	885.8931	16	V	2	246.3293	123.6684
		17	K	1	147.1967	74.1021

The inspection of the highlighted ions shows that indeed all ions, including the phosphorylated serine as well as the non-phosphorylated threonine were identified and assigned, strengthening that the shown serine phosphorylation is correct. In addition, take a look at the spectrum. Here you can see that both assigned ions are rather intense and that the y1+ fragment ion at m/z 1300 (or y2+ fragment ion at m/z 650) which would indicate that

threonine 7 is phosphorylated is missing. Both findings finally confirm the correct assignment of the phosphorylation site.

Here is an example of a non-assigned phosphorylation site, as you can see the ion of the phosphorylation site was not found.



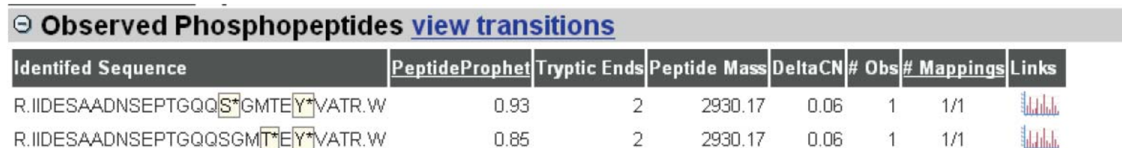
b^{1+}	b^{2+}	b^{3+}	#	AA	#	y^{1+}	y^{2+}	y^{3+}
148.1840	74.5957	50.0663	1	F	19			
205.2359	103.1216	69.0836	2	G	18	1843.0258	922.0166	**615.0135**
318.3953	159.7013	106.8034	3	L	17	1785.9739	**893.4907**	**595.9962**
405.4735	203.2404	135.8294	4	S	16	1672.8145	836.9110	**558.2764**
502.5902	251.7988	168.2017	5	P	15	1585.7363	793.37	
669.6483	335.3278	223.8877	6	S[167]	14	1488.6196	[illegible]	
797.8224	399.4149	266.6124	7	K	13	1321.5615	661.28	
854.8743	**427.9408**	285.6297	8	G	12	1193.3874	597.1974	398.4674
954.0069	477.5071	318.6739	9	V	11	1136.3355	**568.6715**	379.4501
1067.1663	**534.0868**	356.3937	10	L	10	1037.2029	519.1052	346.4059
1214.3429	**607.6751**	405.4526	11	F	9	**924.0435**	**462.5255**	308.6861
1377.5189	**689.2631**	459.8446	12	Y	8	**776.8669**	388.9372	259.6272
1434.5708	**717.7891**	478.8619	13	G	7	**613.6909**	307.3492	205.2352
1531.6875	766.3474	511.2341	14	P	6	**556.6390**	278.8232	186.2179
1628.8042	814.9058	543.6063	15	P	5	459.5223	230.2649	153.8457
1685.8561	843.4317	562.6236	16	G	4	362.4056	181.7065	121.4735
1786.9612	**893.9843**	**596.3253**	17	T	3	305.3537	153.1806	102.4562
1844.0131	922.5102	**615.3426**	18	G	2	204.2486	102.6280	68.7545
			19	K	1	147.1967	74.1021	49.7372

3) Take a look at the kinase phosphorylation motif

An additional step to take in order to confirm a site of phosphorylation is to look at the possible kinase motif surrounding the phosphorylation site. In the example below phosphorylation sites on the protein FUS3, a MAPK, are shown. Here is it not clear whether

R.IIDESAADNSEPTGQQ**S***GMTE**Y***VATR.W
or
R.IIDESAADNSEPTGQQSGM**T***E**Y***VATR.W

is correct. Knowing that the MAP kinases are activated by the phosphorylation in the T*XY* motif we can assume that the R.IIDESAADNSEPTGQQSGMT*EY*VATR.W is correct.



Observed Phosphopeptides view transitions

Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.IIDESAADNSEPTGQQS*GMTEY*VATR.W	0.93	2	2930.17	0.06	1	1/1	
R.IIDESAADNSEPTGQQSGMT*EY*VATR.W	0.85	2	2930.17	0.06	1	1/1	

4) Predict the motif using Scansite

In case you do not have all kinase motifs memorized you can use the ScanSite algorithm to search the protein sequence for possible kinase motifs. For this simply click on the button

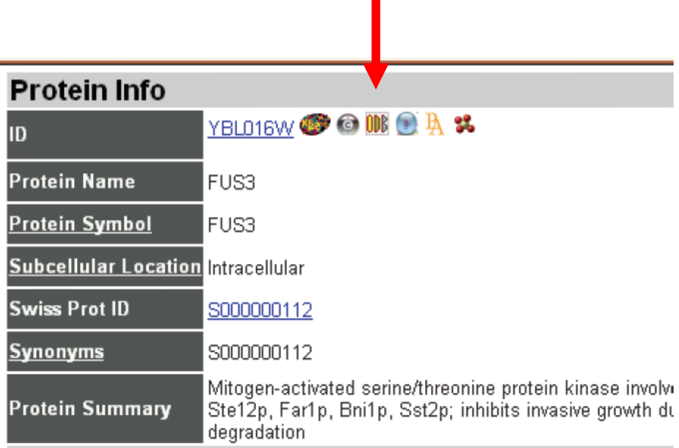
“Search protein sequence at Scansite” in the Protein Info section.

Protein Info

ID	YBL016W
Protein Name	FUS3
Protein Symbol	FUS3
Subcellular Location	Intracellular
Swiss Prot ID	S000000112
Synonyms	S000000112
Protein Summary	Mitogen-activated serine/threonine protein kinase involv Ste12p, Far1p, Bni1p, Sst2p; inhibits invasive growth du degradation

5) *Check the evolutionary conservation of the site*

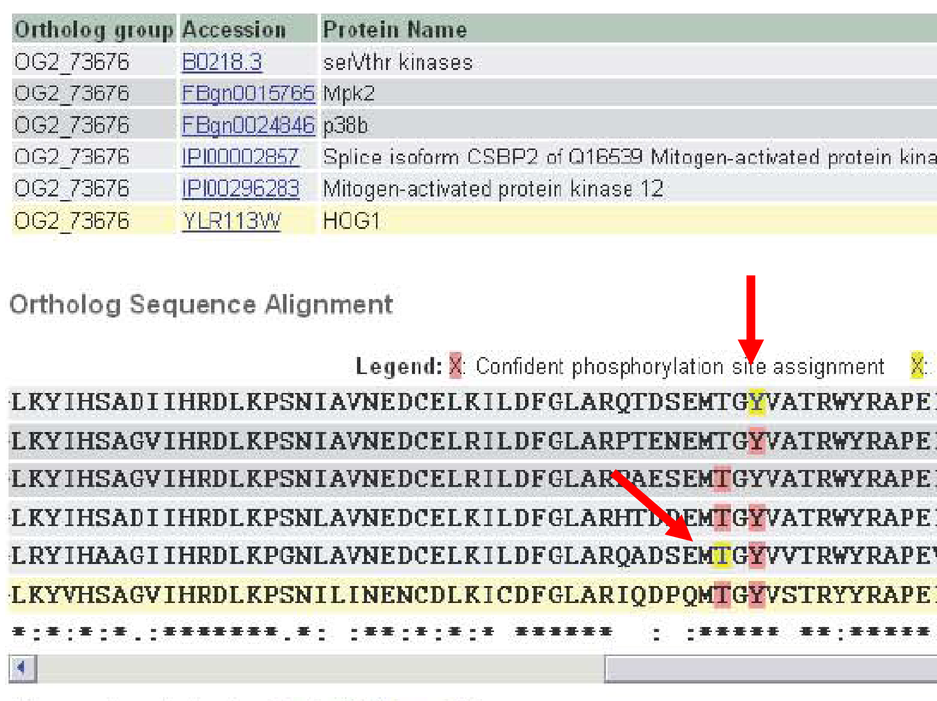
Finally, you can also check whether your phosphorylation site of interest is evolutionary conserved which is an additional indication for the correct assignment of a phosphorylation site. For this click on the button "View orthologs/homolog information"



Protein Info

ID	YBL016W
Protein Name	FUS3
Protein Symbol	FUS3
Subcellular Location	Intracellular
Swiss Prot ID	S000000112
Synonyms	S000000112
Protein Summary	Mitogen-activated serine/threonine protein kinase involv Ste12p, Far1p, Bni1p, Sst2p; inhibits invasive growth du degradation

and a new window will be opened showing the alignment of the phosphorylation site between yeast, worm fly and human. In the examples below we can conclude that the unassigned threonine (highlighted in yellow) is correctly assigned and that in first amino acid sequence either the tyrosine or threonine in the TXY motif must be phosphorylated.



Ortholog group	Accession	Protein Name
OG2_73676	B0218.3	ser/thr kinases
OG2_73676	FBgn0015765	Mpk2
OG2_73676	FBgn0024846	p38b
OG2_73676	IPI00002857	Splice isoform CSBP2 of Q16539 Mitogen-activated protein kina
OG2_73676	IPI00296283	Mitogen-activated protein kinase 12
OG2_73676	YLR113W	HOG1