Supplement: Phospho-signature predicts dasatinib response in non-small cell lung cancer

Martin Klammer Marc Kaminski Alexandra Zedler
Felix Oppermann Stefan Müller Andreas Tebbe Klaus Godl
Christoph Schaab

1 Cost matrix example

The example in Fig. S1A shows how the introduction of cost matrices influences the support vector classification. The figure shows a classification example that aims at separating red stars from blue crosses. Each class contains 10 samples with two features. The values of both features were sampled from normal distributions (N(1,1) and N(-1,1) for crosses and stars, respectively). The black line represents the separating hyperplane of the SVM classification with linear kernel (parameter C=1), when no explicit cost matrix is applied (i.e. the cost of misclassifying a star is the same as the cost for misclassifying a cross). One can clearly see that the data is not linearly separable, which leads to one misclassified cross and one misclassified star. The red line shows the hyperplane when the cost for the false classification of stars is twice as high as the cost for star misclassification. As a result, the separating hyperplane is shifted towards the cloud of red stars, but the classification result is still the same. By increasing the cost factor of cross misclassification to ten times the cost of star misclassification, the hyperplane (blue line) is shifted further and all crosses are classified correctly. However, instead of one falsely predicted star there are now four. Finally, when using a cost factor of 200 (see purple line), all samples would be classified as crosses leading to ten wrongly predicted stars.

This shifting of the hyperplane can be used to calculate the receiver operating characteristic (ROC) curve and the area under it. A ROC curve based on the four different cost matrices above would look like Fig. S1B (assuming that the crosses are the positives and the stars the negatives in the ROC statistics). The point at (1.0|1.0) corresponds to the purple hyperplane, where all crosses are classified correctly and all stars wrongly; the point at (0.4|1.0) to the blue discrimination line, where all crosses are classified correctly and 4 stars are falsely predicted as positives; the point at (0.1|0.9) to both the red and black hyperplane, where 9 crosses are classified correctly and one star wrongly as positive; and finally one more point at (0|0) that is not depicted in Fig S1A but represents the extreme when all samples are assumed to be negatives (stars), which can be considered the opposite of the purple discrimination line. Finally, the area under the curve can be computed, which is 0.93 in this example.

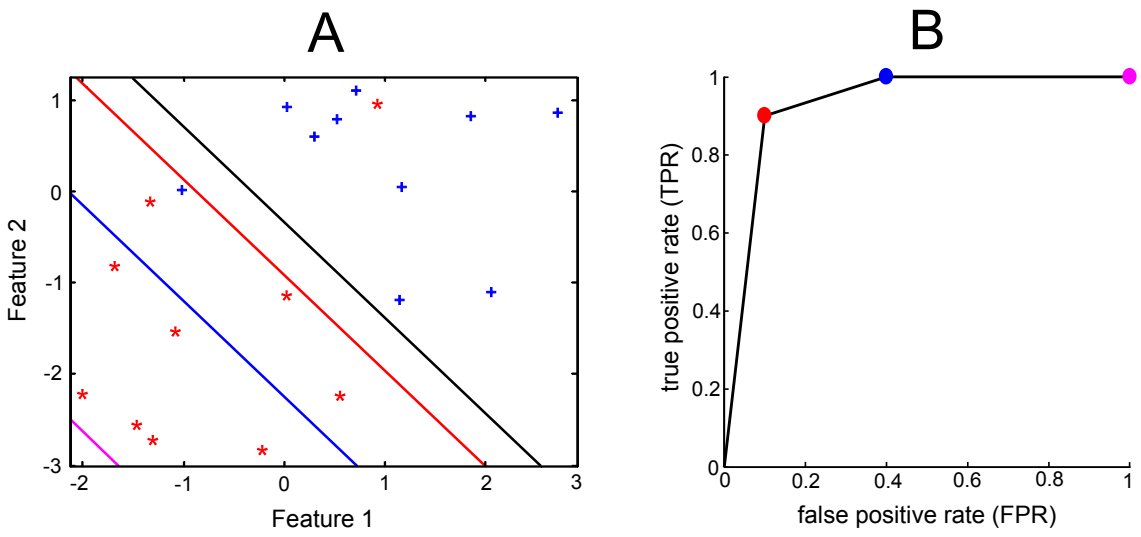

Fig. S 1: Classification example using a linear SVM with different cost matrices (A), and the corresponding ROC curve (B).

2 Details on SVM prediction

The decision function of the SVM classification is given by

$$f(\vec{x}) = \mathrm{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k\left(\vec{x_i}, \vec{x}\right) + b\right), \tag{1}$$

where m is the number of training samples (cell lines), y_i the class label of the i^{th} training sample (-1 or 1 for sensitive and resistant cell lines, respectively), α_i the respective Lagrange multiplier, $\vec{x_i}$ a vector of length f (f being the number of selected features) holding the ratios of the i^{th} training sample, $\vec{x}$ a vector of length f holding the ratios of the test sample, and b the bias (i.e. the translation of the hyperplane with respect to the origin). $k(\vec{x_i}, \vec{x})$ is called a kernel, i.e. a function that characterizes the similarity of two vectors. Equation [1] can be rewritten as

$$f(\vec{x}) = \mathrm{sgn}\Big(k\left(\vec{w}, \vec{x}\right) + b\Big), \tag{2}$$

with the weight vector $\vec{w}$, whose elements represent the importance (influence) of the corresponding features, defined as $\vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{x_i}$. In the case of the linear SVM, the kernel function is defined as the dot product of the two vectors, which leads to the linear decision function

$$f(\vec{x}) = \mathrm{sgn}\left(\sum_{j=1}^{f} w_j x_j + b\right). \tag{3}$$

So far, changes in the phosphorylation level were represented by ratios, which can be expressed as $x = S - S_{ref}$, where S is the signal of the phosphosite in the corresponding cell line and S_{ref} the signal of the site in the reference cell line pool. Here, the signal is defined as log intensity of the corresponding phosphosite. For data produced by other methods such as multiple reaction monitoring or ELISA, where the quantitative data are represented by intensities, one can still

make predictions with the proposed phospho-signature, but the decision function (Equation [2]) has to be modified to

$$f(\vec{S}) = \mathrm{sgn}\Big(k\big(\vec{w}, \vec{S}\big) + \underbrace{b - k\big(\vec{w}, \vec{S}_{ref}\big)}_{\tilde{b}}\Big). \tag{4}$$

Note, that only the bias term has to be modified while the weight vector $\vec{w}$ stays the same. In geometrical terms, the orientation of the hyperplane does not change, but is translated to the new position. In the case of the linear SVM the decision function thus changes to

$$f(\vec{S}) = \mathrm{sgn}\left(\sum_{j=1}^{f} w_j S_j + \tilde{b}\right). \tag{5}$$

3 Supplemental figures mentioned in the main article

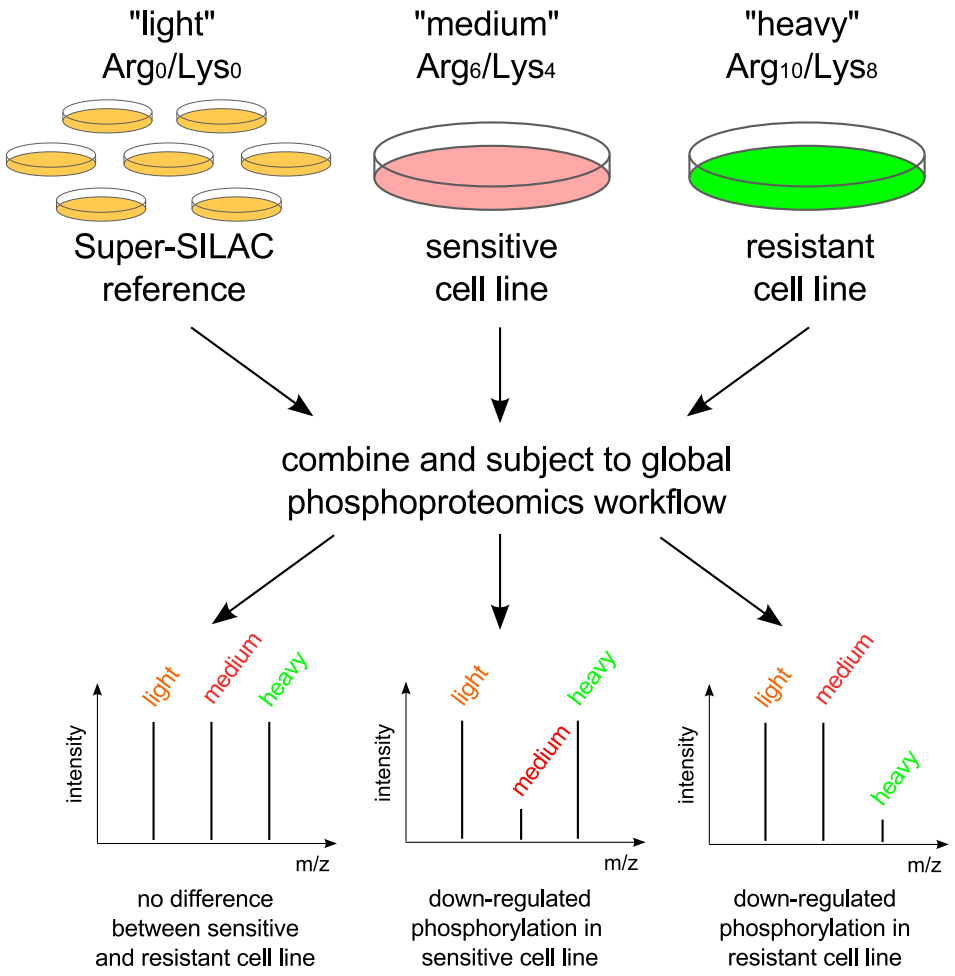

Fig. S 2: SILAC labelling diagram. The scheme illustrates how isotopic labelling enables relative quantification of phosphorylation amounts via a spike-in reference (SuperSILAC).

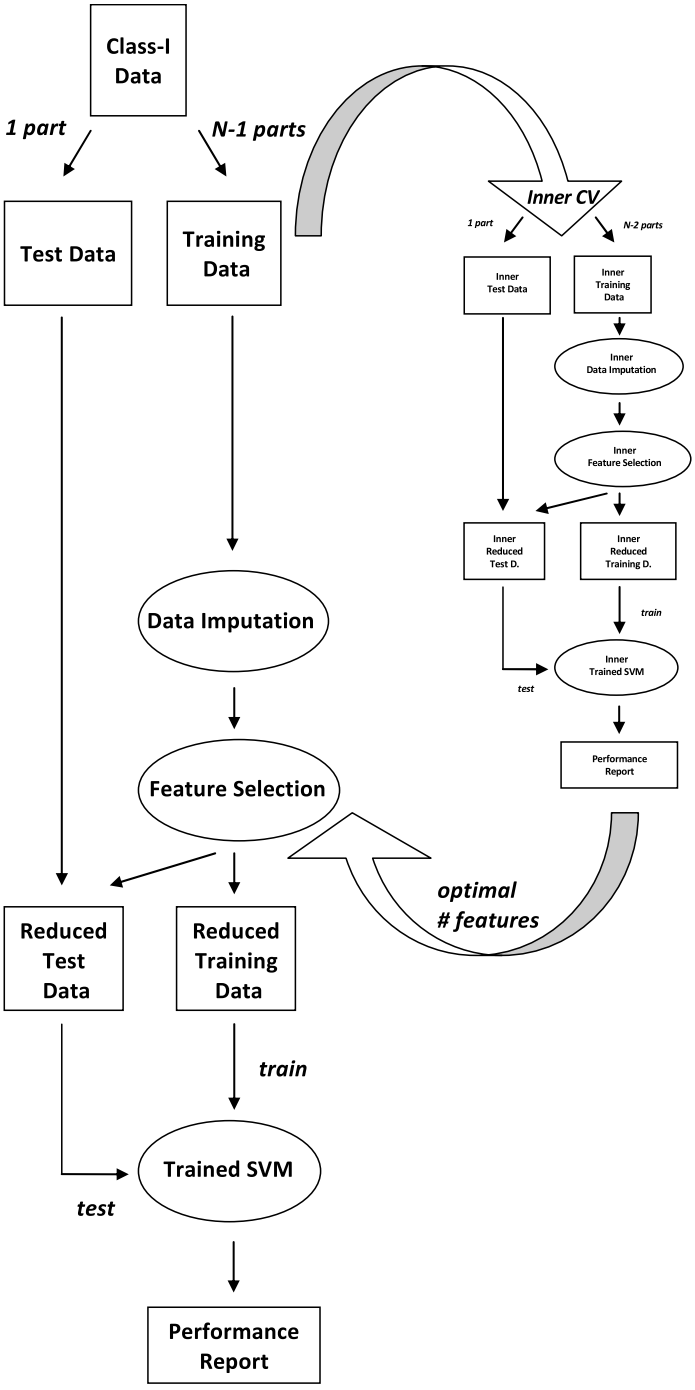

Fig. S 3: Workflow diagram for prediction quality assessment. Two cross validation loops are applied to estimate the prediction accuracy: in the inner CV loop the optimal number of features is determined. This number is then used in the feature selection process in the outer CV loop. Subsequently, an SVM is trained and tested with the respective data sets. The prediction results in each outer CV loop are combined and the prediction accuracy is calculated.

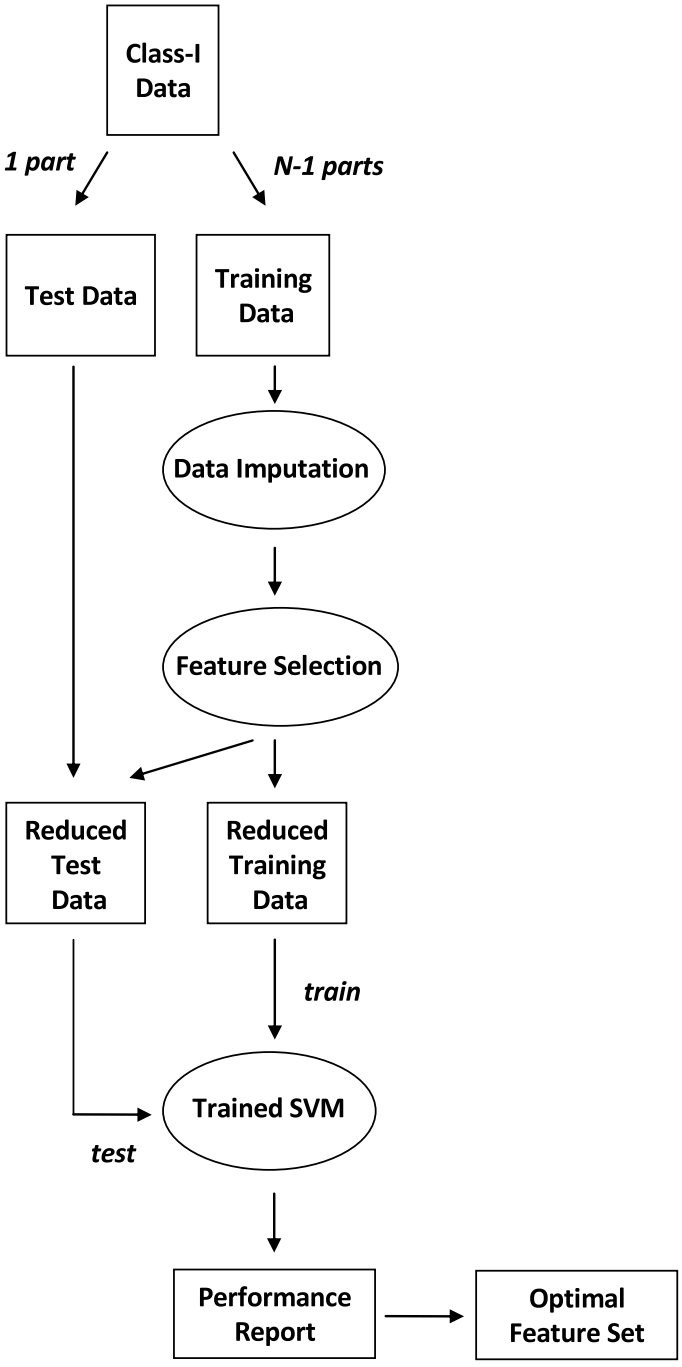

Fig. S 4: Workflow diagram for finding the final phospho-signature. The workflow corresponds to one inner CV loop in Fig. S2 resulting in the optimal set of features, which is then used to train the final SVM predictor.

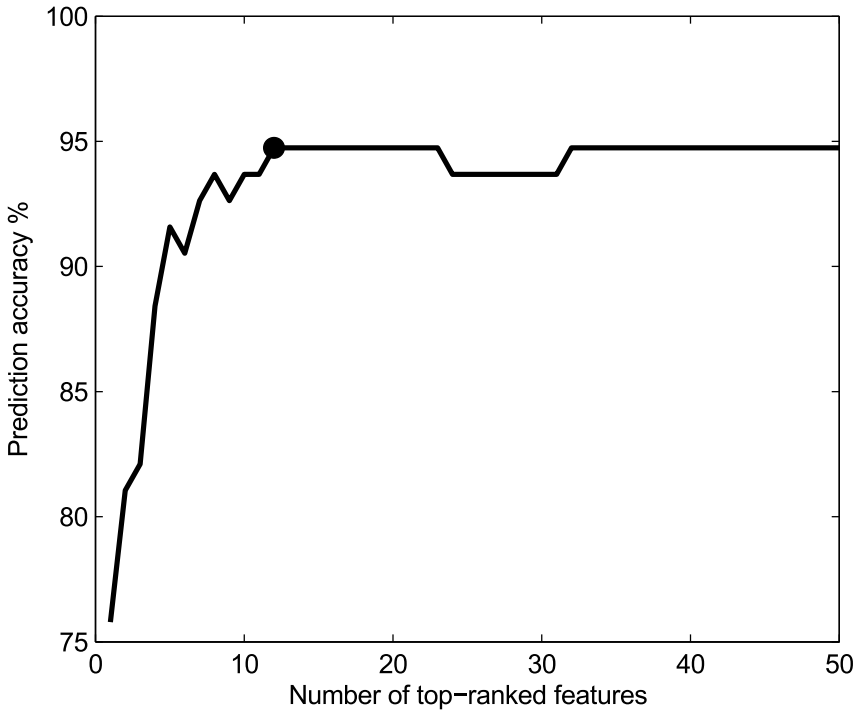

Fig. S 5: The prediction accuracy depending on the number of top-ranked features incorporated into the phospho-signature. While the accuracy increased with the first few features, it reached its maximum at 12 features (circle), where it saturated.

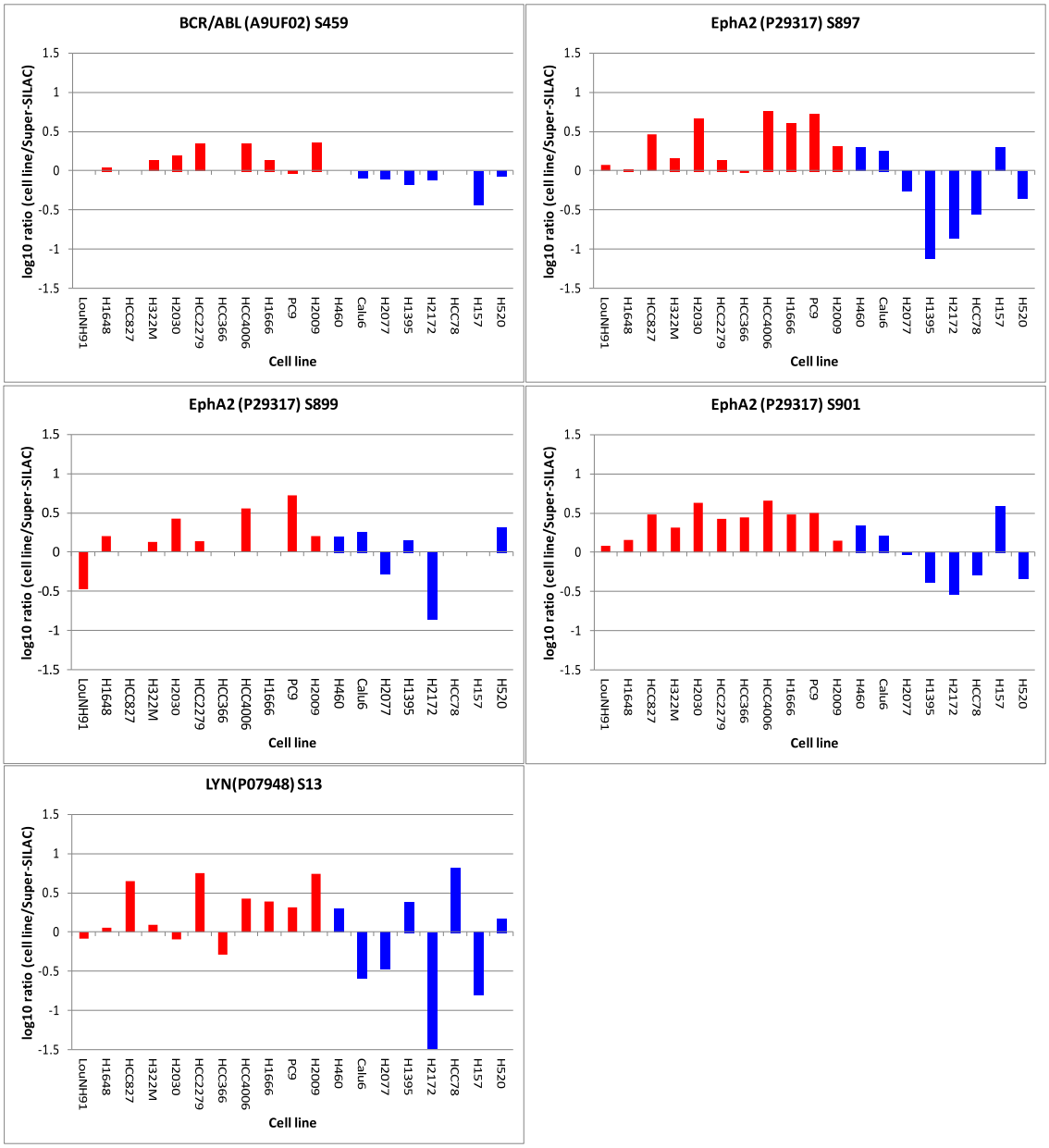

Fig. S 6: Bar charts of log10 ratios (cell line/Super-SILAC) of phosphorylation sites on tyrosine kinases quantified in at least two thirds of the experiments.

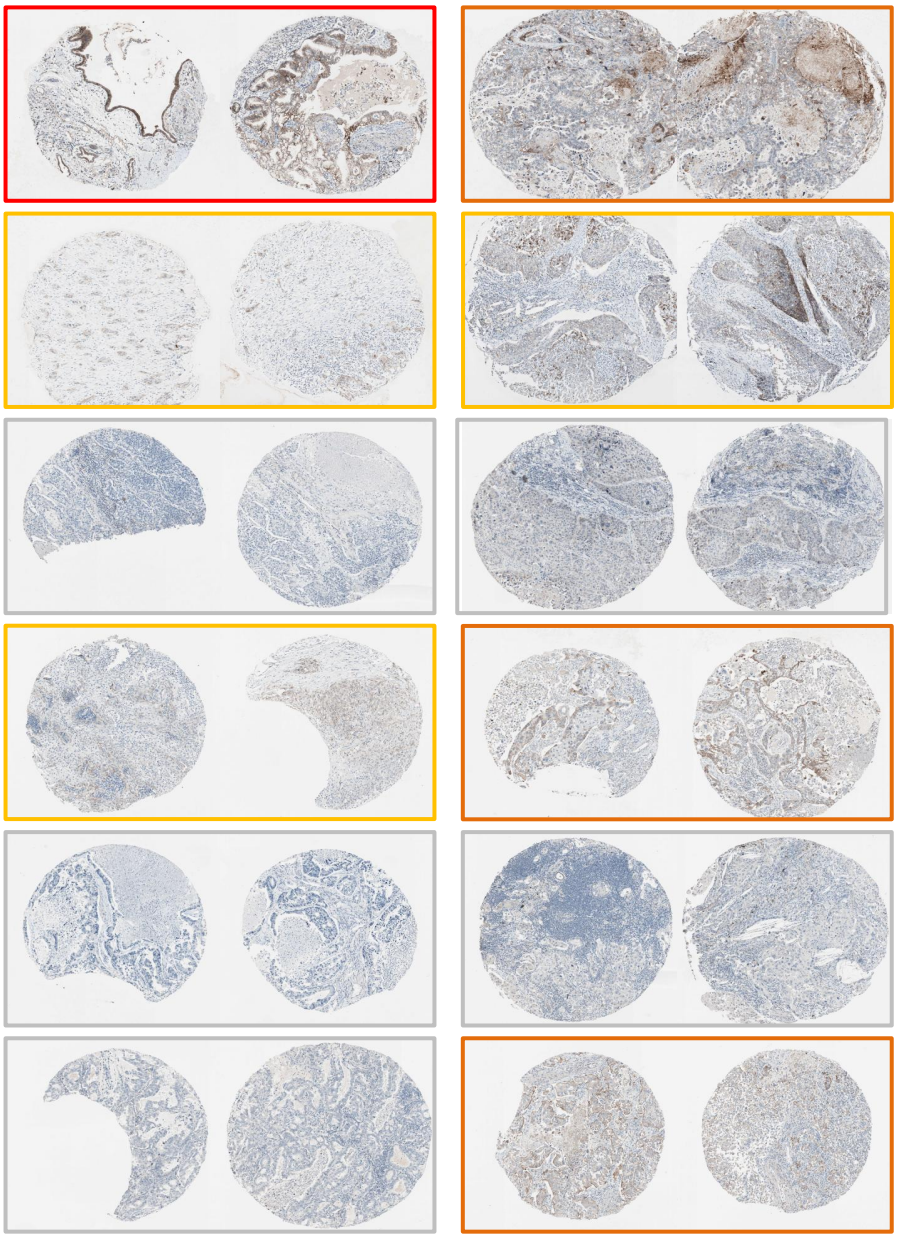

Fig. S7A: Immunohistochemical staining of lung cancer tissues from the Human Protein Atlas. A red border indicates heavy staining, orange moderate, yellow weak and grey no staining.

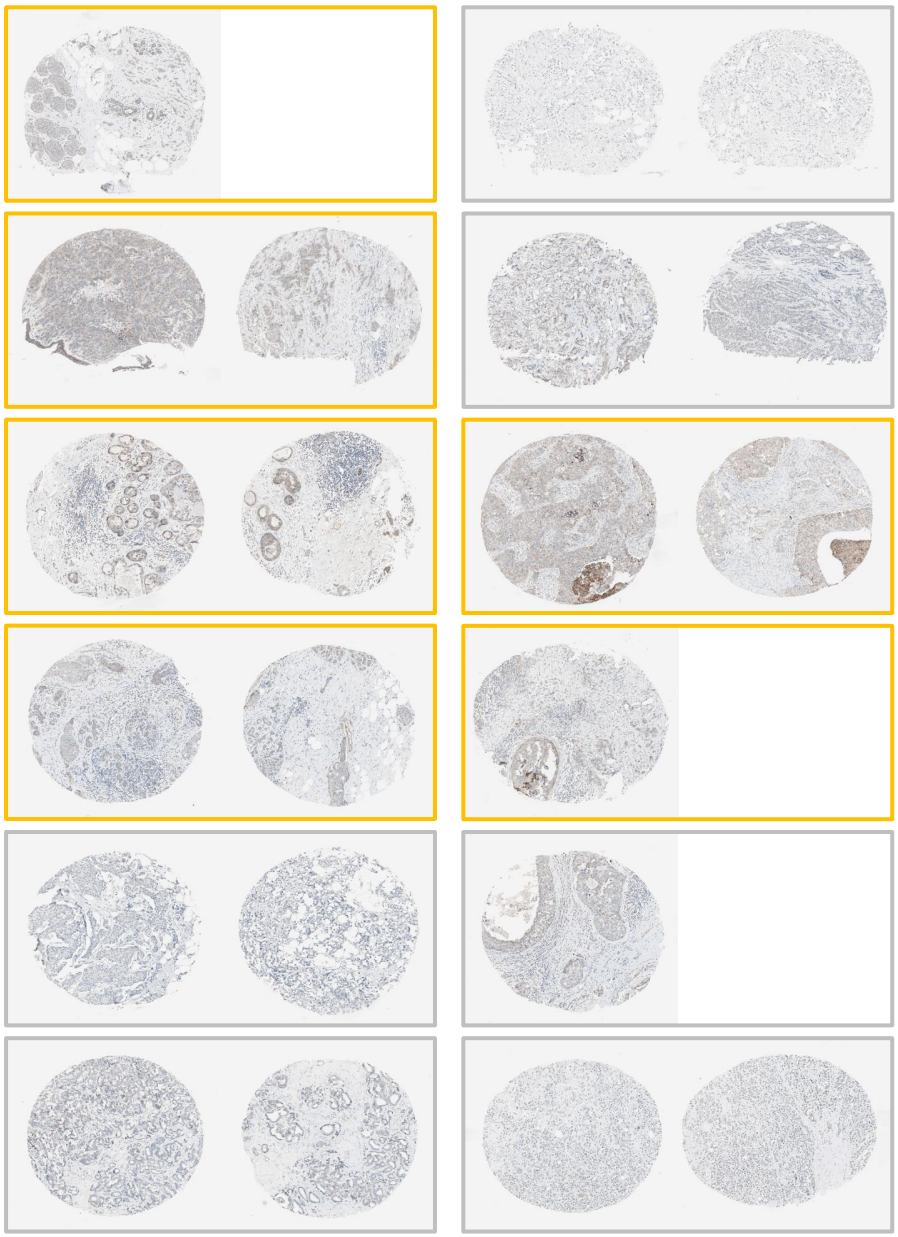

Fig. S 7B: Immunohistochemical staining of breast cancer tissues from the Human Protein Atlas. A yellow border indicates weak staining and grey no staining.

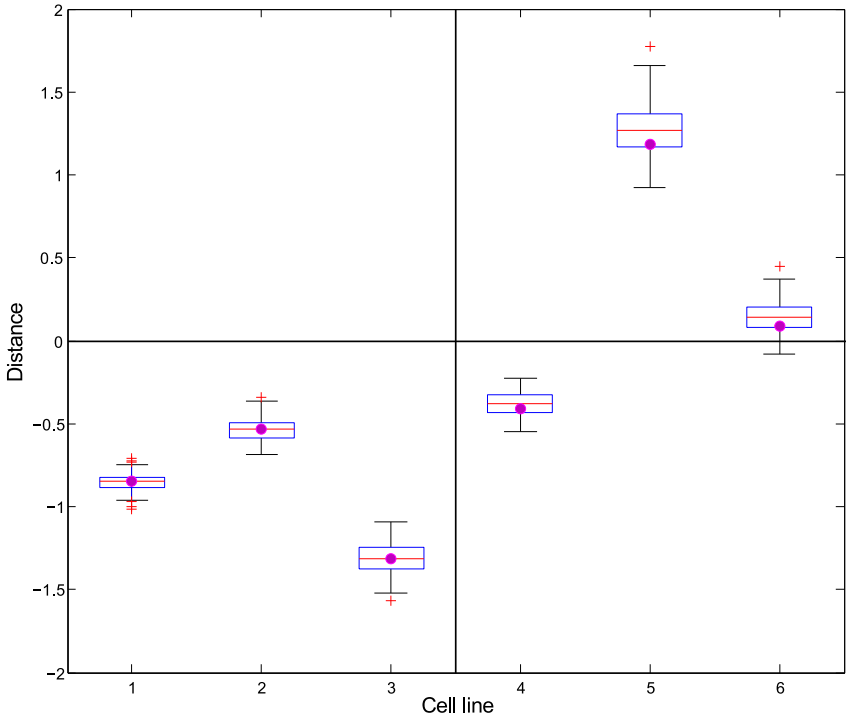

Fig. S 8: Effect of the imputation method for the final predictor when applied to the breast cancer samples. Purple dots indicate the classification result with the predictor trained on the mean-imputed NSCLC data, the box plots show the results for the iputation based on 100 samplings from the respective normal distribution of each feature and class.

4 Supplemental tables mentioned in the main article

Table S 1. Cell line information

Cell line	Indication	Origin	Supplier number	$GI_{50}(\mu M)$ dasatinib literature*	$GI_{50}(\mu M)$ dasatinib this paper	Class	Valid†	TP53 mutation‡	TP53 status	Doubling time (h)	$GI_{50}(\mu M)$ sorafenib *literature
Calu6	NSCLC	ATCC	HTB-56	22.54	2.8	-	YES	c.586C>T; Arg->STOP	MUT	25	30
H1395	NSCLC	ATCC	CRL-5868	31.12	4.7	-	YES	WT	WT	50	7.24
H1568	NSCLC	ATCC	CRL-5876	0.8975	5.44	+	no	–	–	59	6.46
H157	NSCLC	MPI§	–	10.54	2.63	-	YES	c.892G>T; Glu->STOP	MUT	25	6.61
H1648	NSCLC	ATCC	CRL-5882	0.0593	0.079	+	YES	c.102_103ins1; Leu->?	MUT	50	6.03
H1666	NSCLC	ATCC	CRL-5885	0.175	0.076	+	YES	WT	WT	30	30
H2009	NSCLC	ATCC	CRL-5911	0.7465	0.085	+	YES	c.818G>T; Arg->Leu	MUT	50	11.09
H2030	NSCLC	ATCC	CRL-5914	0.1183	0.022	+	YES	c.785G>T; Gly->Val	MUT	25	7.76
H2077	NSCLC	MPI	–	10.07	4.75	-	YES	–	–	50	5.37
H2172	NSCLC	ATCC	CRL-5930	16.71	5.85	-	YES	–	–	50	–
H2887	NSCLC	MPI	–	11.3	0.176	-	no	–	–	40	13.65
H322	NSCLC	MPI	–	0.2588	2.1	+	no	c.743G>T; Arg->Leu	MUT	–	5.43
H460	NSCLC	ATCC	HTB-177	24.16	3.9	-	YES	WT	WT	25	30
HCC827	NSCLC	ATCC	CRL-2868	0.1456	0.033	+	YES	–	–	43	5.25
H520	NSCLC	ATCC	HTB-182	11.56	1.43	-	YES	c.438G>A; Trp->STOP	MUT	42	4.84
H647	NSCLC	ATCC	CRL-5834	12.39	0.016	-	no	c.782+1G>T; [intron!]	MUT	–	12.45
HCC1359	NSCLC	MPI	–	11.3	0.52	-	no	c.388C>T; Leu->Phe	MUT	30	11.89
LCLC103H	NSCLC	DSMZ	ACC 384	13.9	0.08	-	no	c.646G>T; Val->Leu	MUT	–	9.66
LouNH91	NSCLC	DSMZ	ACC 393	0.113	0.068	+	YES	–	–	55	4.68
HCC366	NSCLC	DSMZ	ACC 492	0.482	0.017	+	YES	–	–	53	6.03
HCC4006	NSCLC	ATCC	CRL-2871	0.8376	0.95	+	YES	–	–	–	6.46
HCC78	NSCLC	DSMZ	ACC 563	13.9	17.05	-	YES	–	–	–	11.09
H322M	NSCLC	MPI	–	0.0819	0.311	+	YES	c.743G>T; Arg->Leu	MUT	–	14.13
HOP62	NSCLC	MPI	–	12.76	0.014	-	no	c.633_634ins36; Phe->?	MUT	–	9.44
HCC2279	NSCLC	MPI	–	0.139	0.045	+	YES	c.701A>G; Tyr->Cys	MUT	–	12.45
PC9	NSCLC	MPI	–	0.4603	0.02	+	YES	c.743G>A; Arg->Gln	MUT	25	15.85
BT-20	Breast c.	ATCC	HTB-19	0.1652	0.497	+	YES	c.394A>C; Lys->Gln	MUT	–	–
BT-549	Breast c.	ATCC	HTB-122	9.0576	1.71	-	YES	c.747G>C; Arg->Ser	MUT	–	–
MDA-MB-468	Breast c.	ATCC	HTB-132	7.1258	2.8	-	YES	c.818G>A; Arg->His	MUT	–	–
MDA-MB-231	Breast c.	ATCC	HTB-26	0.0095	0.036	+	YES	c.839G>A; Arg->Lys	MUT	–	–
MCF7	Breast c.	ATCC	HTB-22	>9.524	3.27	-	YES	WT	WT	–	–
HCC1937	Breast c.	ATCC	CRL-2336	0.07	0.082	+	YES	c.916C>T; Arg->STOP	MUT	–	–

*NSCLC data from Sos, *et al.* (2009), breast cancer data from Huang, *et al.* (2007)

†whether the GI_{50} values from the literature and this paper agree

‡according to the IARC TP53 database [Petitjean, *et al.* (2007)] version R15

§Max Planck Institute for Neurological Research (Cologne, Germany)

Table S 2 . Mass spectrometric pairing scheme

Exp. number	Group medium	Group heavy	Cell line light	Cell line medium	Cell line heavy
1	+	-	CELLMIX	LouNH91	H460
2	+	-	CELLMIX	H1648	Calu6
3	+	-	CELLMIX	HCC827	LCLC103H*
4	+	-	CELLMIX	H322M	H2077
5	+	-	CELLMIX	H2030	H1395
6	+	-	CELLMIX	HCC2279	H2172
7	+	-	CELLMIX	H1568*	H647*
8	+	-	CELLMIX	H322*	HOP62*
9	+	-	CELLMIX	HCC366	HCC78
10	+	-	CELLMIX	HCC4006	HCC1359*
11	+	-	CELLMIX	H1666	H157
12	+	-	CELLMIX	PC9	H520
13	+	-	CELLMIX	H2009	H2887*
14†	-	+	CELLMIX	H2077	H322M
15‡	-	+	CELLMIX	H2887*	H2009
16	+	-	CELLMIX	BT-20	MDA-MB-468
17	-	+	CELLMIX	BT-549	MDA-MB231
18	+	-	CELLMIX	HCC1937	MCF7

*this cell line's GI50 value turned out to be inconsistent with the one reported in Sos, *et al.* (2009); thus, the cell line was not used in the analysis

†label switch of experiment 4

‡label switch of experiment 13

Table S 3 . Significantly enriched GO terms

GO term	GO id	Category*	q-value†
cell communication	GO:0007154	GOBP	2.5E-06
signal transduction	GO:0007165	GOBP	2.5E-06
signal transducer activity	GO:0004871	GOMF	5.1E-04
cytoskeletal protein binding	GO:0008092	GOMF	5.1E-04
small GTPase regulator activity	GO:0005083	GOMF	5.5E-04
regulation of cellular process	GO:0050794	GOBP	7.2E-04
GTPase regulator activity	GO:0030695	GOMF	1.1E-03
protein kinase activity	GO:0004672	GOMF	2.0E-03
protein serine/threonine kinase activity	GO:0004674	GOMF	2.9E-03
protein tyrosine kinase activity	GO:0004713	GOMF	2.9E-03
receptor activity	GO:0004872	GOMF	2.9E-03
phosphotransferase activity, alcohol group as acceptor	GO:0016773	GOMF	3.6E-03
lipid binding	GO:0008289	GOMF	3.6E-03
kinase activity	GO:0016301	GOMF	3.7E-03
actin binding	GO:0003779	GOMF	3.7E-03
zinc ion binding	GO:0008270	GOMF	1.2E-02
Ras protein signal transduction	GO:0007265	GOBP	1.2E-02
protein amino acid phosphorylation	GO:0006468	GOBP	1.2E-02
regulation of cell communication	GO:0010646	GOBP	1.3E-02
transferase activity, transferring phosphorus-containing groups	GO:0016772	GOMF	1.3E-02
regulation of signal transduction	GO:0009966	GOBP	1.7E-02
GTPase activator activity	GO:0005096	GOMF	2.0E-02
cell surface receptor linked signal transduction	GO:0007166	GOBP	2.1E-02
intracellular signaling cascade	GO:0007242	GOBP	2.2E-02
enzyme activator activity	GO:0008047	GOMF	2.2E-02
vesicle-mediated transport	GO:0016192	GOBP	2.4E-02
Rho protein signal transduction	GO:0007266	GOBP	2.5E-02
transport	GO:0006810	GOBP	2.5E-02
establishment of localization	GO:0051234	GOBP	2.5E-02
amine binding	GO:0043176	GOMF	2.7E-02
epidermal cell differentiation	GO:0009913	GOBP	2.9E-02
phosphorylation	GO:0016310	GOBP	2.9E-02
transmembrane receptor activity	GO:0004888	GOMF	3.1E-02
cytoskeleton organization	GO:0007010	GOBP	3.5E-02
locomotory behavior	GO:0007626	GOBP	3.5E-02
transition metal ion binding	GO:0046914	GOMF	3.8E-02
phosphate metabolic process	GO:0006796	GOBP	4.0E-02
Rho guanyl-nucleotide exchange factor activity	GO:0005089	GOMF	4.3E-02
actin filament binding	GO:0051015	GOMF	4.6E-02
post-translational protein modification	GO:0043687	GOBP	4.7E-02

*GOBP: biological process, GOMF: melecular function

†adjusted p-value

Table S 4 . Additional phosphorylation site information

Gene Name	Site*	Canonical Uniprot id†	Canonical site‡	All Uniprot ids§	Known ¶site
ITGB4	S1448	P16144	S1518	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2;P16144;P16144-4	YES
BAIAP2	S509	Q9UQB8-5	S509	Q9UQB8-5;B3KPV9	no
ITGB4	S1387	P16144	S1457	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2;P16144;P16144-3;P16144-4	YES
ITGB4	T1385	P16144	T1455	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2;P16144;P16144-3;P16144-4	no
ITGB4	S1069	P16144	S1069	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2;P16144;P16144-3;P16144-4	‖YES
GPCR5A	S345	Q8NFJ5	S345	A8K556;Q8NFJ5	YES
ITPR3	S916	Q14573	S916	Q14573;Q59ES2;A6H8K3	YES
TNKS1BP1	S429	Q9C0C2	S429	Q9C0C2;B3KXS7	YES
ARHGEF18	S1101	Q6ZSZ5	S1101	Q6ZSZ5;B5ME81;D6W646;Q6ZSZ5-2;A8MV62;Q6ZSZ5-3	YES
IASPP	S102	Q8WUF5	S102	Q8WUF5;Q6ZNZ8	YES
APG16L	S269	Q676U5	S269	Q676U5;Q676U5-3;Q676U5-4;Q17RG0;Q53SV2	YES
TPD52L2	S141	O43399	S161	O43399;Q6FGS1;Q53GA0;B4DDV4;O43399-2;Q68E05;B4DPJ6	YES

*as reported throughout the paper

†the main Uniprot entry of the corresponding protein

‡the position in the canonical Uniprot entry

§all Uniprot accession numbers from which the corresponding phosphopeptide could originate

¶according to PhosphoSitePlus (www.phosphosite.org) accessed on 6thAugust 2011

‖detected in mouse only

Table S 5 . Log10 ratios of cell lines versus SuperSILAC mix

Cell line	Indication	Class	ITGB4 S1448	BAIAP2 S509	ITGB4 S1387	ITGB4 T1385	ITGB4 S1069	GPCR5A S345	ITPR3 S916	TNKS1BP1 S429	ARHGEF18 S1101	IASPP S102	APG16L S269	TPD52L2 S141
LouNH91	NSCLC	+	0.265	0.239				0.192	0.560	0.840	-0.042	0.312		
H1648	NSCLC	+		0.735	0.643	0.507	0.412	0.644	-0.103	0.693	0.074	0.402	0.345	0.393
HCC827	NSCLC	+							0.033	0.032	0.558		-0.194	0.734
H322M	NSCLC	+	0.926	0.645	0.819	0.909	0.852	0.588	-0.070	-0.479	0.085	0.456	-0.118	0.399
H2030	NSCLC	+	0.463		0.377		0.383	0.567	-0.305	0.070	0.439	0.421	0.089	0.746
HCC2279	NSCLC	+	1.012	0.442	0.758	0.656	0.943	0.484	0.194	0.396	0.422		0.124	
HCC366	NSCLC	+	0.896		0.746	0.655	0.818	0.562		0.011	-0.008			0.259
HCC4006	NSCLC	+	0.890	0.261	0.900	0.903		0.603	-0.034	0.044	0.121	0.799	-0.235	0.461
H1666	NSCLC	+	0.717	0.865	0.913	0.690	0.865	0.032		0.130	0.386	0.529	-0.008	0.810
PC9	NSCLC	+		0.296	0.644	0.644	1.101	0.173	0.160	-0.132	0.123	0.399	-0.021	0.580
H2009	NSCLC	+	0.962	0.685	0.996	0.820	1.466	0.172	-0.456	0.138	0.047	0.605	-0.082	0.279
H460	NSCLC	-	-0.142	-0.866				-0.073	-1.025	-0.736	-0.484	-0.429		
Calu6	NSCLC	-		-0.477	-0.421	-0.554	-0.544	-0.223	-0.479	-0.998	-0.467	-0.188	-0.716	-0.597
H2077	NSCLC	-	-0.597	-0.757	-0.609	-0.410	-0.892	-0.349	-0.579	-0.692	-0.411	-0.139	-1.069	-0.367
H1395	NSCLC	-	-0.765		-0.787		-0.857	-0.353	-0.792	-1.086	-0.211	0.058	-1.077	-0.077
H2172	NSCLC	-	-0.705	-0.549	0.042	-0.174		-0.350	-0.839	-0.998	-0.263		-0.381	
HCC78	NSCLC	-	-0.936		-0.049	-0.218	0.257	-0.071		-0.239	-0.334			0.192
H157	NSCLC	-	-0.109	-0.797	-0.233	-0.310	-0.211	-0.971		-0.990	-0.226	-0.040	-0.478	0.017
H520	NSCLC	-		-0.348	-0.189	-0.189	0.029	-0.552	-0.986	-0.807	-0.129	0.051	-0.776	-0.127
BT-20	Breast c.	+		0.585	0.457	0.575	0.478	0.135	0.083	0.295		0.668	-0.008	0.911
MDA-MB-231	Breast c.	+	0.580		0.403	0.432	0.738	0.243	0.547	-0.114			-0.188	-0.431
HCC1937	Breast c.	+	0.495	0.555	0.723	0.685	0.834	-0.487	0.648	0.807	0.092	0.098	0.569	1.252
MDA-MB-468	Breast c.	-		-0.163	0.160	0.290	-0.055	-0.181	-0.147	0.747		0.316	0.327	0.634
BT-549	Breast c.	-	-0.934		-1.239	-1.428	-0.622	-0.296	-0.642	0.153			-0.009	-0.494
MCF7	Breast c.	-	-0.471	0.305	-0.181	-0.114	-0.795	0.127	-0.059	0.231	-0.531	0.177	0.049	0.586

Table S 6 . Log10 ratios (cell line versus SuperSILAC) of the eight non-modified ribosomal peptides used for the alternative normalisation

Peptide Seq.	FNADEFEDMVAEK	FTPGTFTNQIQAAFREPR	HGSLGFLPR	HMYHSLYLK	ILDSVGIEADDDRLNK	NIEDVIAQGIGK	TIAECLADELINAAK	VCTLAIIDPGDSDIIR
Name	RPL10	RPSA	RPL3	RPL19	RPLP2	RPLP2	RPS5	RPL30
Uniprot Id	P27635	P08865	P39023	P84098	P05387	P05387	P46782	P62888
LouNH1	0.247	0.243	0.273		0.080	0.157	0.255	0.161
H1648	0.282	0.238	0.220		0.320	0.257	0.255	0.287
HCC827	0.182	0.146			0.056	0.147	0.180	0.130
H322M			0.277	0.306	0.177			
H2030		0.196	0.264	0.308	0.181	0.250		
HCC2279	0.270	0.063	0.219	0.154	0.307	0.232	0.132	0.218
HCC366	0.238	0.158	0.140	0.148	0.295	0.221	0.220	0.259
HCC4006					0.301	0.161		
H1666	0.095	0.121	0.138	0.149	0.151	0.125	0.112	0.147
PC9	0.208	0.143	0.228	0.257	0.218	0.272	0.276	0.282
H2009	0.087	0.174	0.103	0.096	0.223	0.021		
H460	0.426	0.417	0.471		0.252	0.353	0.461	0.361
Calu6	0.340	0.278	0.303		0.397	0.383	0.330	0.339
H2077			0.445	0.456	0.409			
H1395		0.068	0.241	0.344	0.264	0.249		
H2172	0.214	0.198	0.173	0.182	0.302	0.238	0.194	0.149
HCC7	0.203	0.167	0.243	0.280	0.240	0.228	0.211	0.193
H157	0.200	0.189	0.145	0.211	0.210	0.100	0.129	0.111
H520	0.149	0.198	0.205	0.250	0.221	0.217	0.225	0.192

5 Supplemental files

Supplement 1: This document. Contains supplementary information, Figures S1-S8 and Tables S1-S6.

Supplement 2: Data of the 25,020 class-I sites from the NSCLC experiments.

Supplement 3: Data of the 13,730 class-I sites from the breast cancer experiments.

Supplement 4: MS/MS spectra of the 25,020 class-I sites from the NSCLC experiments. The spectra are named xxx_yyy.svg, where xxx is the id of the phosphosite in Supplement 2 and yyy is the scan number.

Supplement 5: MS/MS spectra of the 13,730 class-I sites from the breast cancer experiments. The spectra are named xxx_yyy.svg, where xxx is the id of the phosphosite in Supplement 3 and yyy is the scan number.