Phosphosignature Predicts Dasatinib Response in Non-small Cell Lung Cancer*[S]

Martin Klammer‡, Marc Kaminski‡, Alexandra Zedler‡, Felix Oppermann‡, Stephanie Blencke‡, Sandra Marx‡, Stefan Müller‡, Andreas Tebbe‡, Klaus Godl‡, and Christoph Schaab‡§¶

Targeted drugs are less toxic than traditional chemotherapeutic therapies; however, the proportion of patients that benefit from these drugs is often smaller. A marker that confidently predicts patient response to a specific therapy would allow an individual therapy selection most likely to benefit the patient. Here, we used quantitative mass spectrometry to globally profile the basal phosphoproteome of a panel of non-small cell lung cancer cell lines. The effect of the kinase inhibitor dasatinib on cellular growth was tested against the same panel. From the phosphoproteome profiles, we identified 58 phosphorylation sites, which consistently differ between sensitive and resistant cell lines. Many of the corresponding proteins are involved in cell adhesion and cytoskeleton organization. We showed that a signature of only 12 phosphorylation sites is sufficient to accurately predict dasatinib sensitivity. Four of the phosphorylation sites belong to integrin β4, a protein that mediates cell-matrix or cell-cell adhesion. The signature was validated in cross-validation and label switch experiments and in six independently profiled breast cancer cell lines. The study supports that the phosphorylation of integrin β4, as well as eight further proteins comprising the signature, are candidate biomarkers for predicting response to dasatinib in solid tumors. Furthermore, our results show that identifying predictive phosphorylation signatures from global, quantitative phosphoproteomic data is possible and can open a new path to discovering molecular markers for response prediction. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.016410, 651–668, 2012.***

The introduction of targeted drugs for treating cancer is a major biomedical achievement of the past decade (1, 2). Because these drugs selectively block molecular pathways that are typically overactivated in tumor cells, they are more precise and less toxic than traditional chemotherapeutics. However, although many cancer patients benefit from a specific targeted therapy, many others do not. Therefore, predictive molecular markers are needed to confidently predict patient response to a specific therapy. Such markers would facilitate therapy personalization, where the selected therapy is based on the molecular profile of the patient.

Predictive tests currently used in the clinic are frequently based on one particular marker that is often linked to the drug target. A well known example for a predictive test is assessing HER2/neu overexpression using immunohistochemistry or fluorescent *in situ* hybridization to predict the response to therapy with trastuzumab (Herceptin®; Roche) (3, 4). However, in some cases the expression or mutational status of the target or other singleton markers might not be sufficient to predict a therapeutic response. Recently, several studies tried to identify molecular signatures comprising multiple markers for response predictions, usually based on gene expression profiling (5, 6). To our knowledge, no study successfully identified a signature from global phosphoproteomic profiles so far.

Recent advances in mass spectrometry, methods for enriching phosphorylated proteins or peptides, and computer algorithms for analyzing proteomics data have enabled the application of mass spectrometry-based proteomics to monitor phosphorylation events in a global and unbiased manner. These methods have become sufficiently sensitive and robust to localize and quantify the phosphorylation sites within a peptide sequence (7–9). Phosphorylation events are important in signal transduction, where signals caused by external stimuli are transmitted from the cell membrane to the nucleus. Aberrations in these signal transduction pathways are particularly important for understanding the mechanisms of certain diseases, such as cancer, inflammation, and diabetes (10, 11).

Approximately 391,000 incidences and 342,000 deaths from lung cancer were estimated in Europe in 2008 (12), accounting for nearly 20% of all cancer deaths in Europe. Approximately 85% of all lung cancer incidences are non-small cell lung cancer (NSCLC)[1] (13). Dasatinib (Sprycel®;

From the ‡Evotec München GmbH and the §Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

Received December 14, 2011, and in revised form, April 16, 2012
Published, MCP Papers in Press, May 22, 2012, DOI 10.1074/mcp.M111.016410

[1] The abbreviations used are: AUROC, area under the receiver operating characteristic curve; CV, cross-validation; FDR, false discovery rate; NSCLC, non-small cell lung cancer; SILAC, stable isotope labeling by amino acid in cell culture; SVM, support vector machine; GO, gene ontology; KEGG, kyoto encyclopedia of genes and genomes.

Bristol-Myers Squibb) is a multikinase inhibitor targeting BCR-ABL, the Src kinase family, c-Kit, ephrin receptors, and platelet-derived growth factor receptor β (14, 15). It is currently approved for chronic myelogenous leukemia and Philadelphia chromosome-positive acute lymphoblastic leukemia. Recently, dasatinib was clinically evaluated in patients with advanced NSCLC. Dasatinib had modest clinical activity, with only one partial response and 12 stable diseases among 30 patients. Neither Src family kinase activation nor EGFR and K-ras mutations could predict the response to dasatinib (16).

In this study we wanted to identify a signature of protein phosphorylation that predicts the response to dasatinib in NSCLC cell lines. In total, 26 NSCLC cell lines were tested for their response to dasatinib. The identical cell lines were profiled in a global, unbiased, phosphoproteomics study, and the obtained phosphoproteome profiles were used to assemble a biomarker signature of 12 phosphorylation sites. We evaluated the performance of this signature in a cross-validation setup and investigated the robustness of the selected predictive features. Finally, we confirmed the predictive power of the signature in an independent set of breast cancer cell lines.

In a recent study, Andersen *et al.* (17) identified phosphorylation sites predicting response to phosphatidylinositol 3-kinase inhibitors. Their study differs in two aspects from the study presented here. First, the authors focused on the phosphatidylinositol 3-kinase and MAPK pathways by immunoprecipitating phosphorylated peptides with antibodies directed against corresponding phosphomotifs. In contrast, we followed an unbiased approach, where no hypothesis about the involved signaling pathways has to be made. Second, the authors first investigated the regulation of phosphorylation sites upon drug treatment in one sensitive cell line and subsequently confirmed the applicability of one site to response prediction by evaluating its basal phosphorylation in a panel of cell lines. Here, we started directly by investigating the basal phosphoproteome of a panel of sensitive and resistant cell lines.

EXPERIMENTAL PROCEDURES

Cell Culture

Based on the half-maximum growth inhibitory concentrations (GI_{50}) of dasatinib on a panel of 84 NSCLC cell lines reported in supplemental Table 5 of Sos *et al.* (18), 13 cell lines with low and 13 with high GI_{50} values were selected (cf. supplemental Table S1). These 26 cell lines were obtained from LGC Standards (Wesel, Germany), from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (Braunschweig, Germany), and Roman Thomas's group at the Max Planck Institute for Neurological Research (Cologne, Germany). The six breast cancer cell lines were obtained from LGC Standards (see supplemental Table S1).

All of the cell lines were cultivated in RPMI 1640, 10% fetal bovine serum, 2 mM glutamine, 1 mM sodium pyruvate, and penicillin/streptomycin (PAA Laboratories, Cölbe, Germany). The cells were routinely monitored for mycoplasma infection using the MycoAlert reagents (Lonza, Cologne, Germany).

Metabolic labeling of the cell lines was performed using stable isotope labeling with amino acids in cell culture (SILAC) (19). The cells were cultivated in media containing SILAC-RPMI (PAA) and dialyzed fetal bovine serum (Invitrogen). L-Lysine and L-arginine were replaced by normal L-lysine (Lys^0) and L-arginine (Arg^0), or medium isotope-labeled L-$D_4{}^{14}N_2$-lysine (Lys^4) and L-$^{13}C_6{}^{14}N_4$-arginine (Arg^6), or heavy isotope-labeled L-$^{13}C_6{}^{15}N_2$-lysine (Lys^8) and L-$^{13}C_6{}^{15}N_4$-arginine (Arg^{10}). Isotope-labeled amino acids were purchased from Cambridge Isotope Laboratories (Andover, MA). The cells were cultivated for a minimum of six doubling times to obtain an incorporation efficiency for the labeled amino acids of at least 95%.

16 NSCLC cell lines were selected as a reference pool: A549, Calu6, H1395, H1437, H1755, H2030, H2052, H2172, H28, H460, HCC827 (obtained from LGC Standards), LCLC103H, LouNH91 (obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen), H322M, HCC2279, and HCC2429 (obtained from MPI for Neurological Research). The selected cell lines were grown in SILAC medium supplemented with the natural "light" forms of arginine and lysine. The labeled cells of each cell line were lysed, pooled, aliquoted, and stored at −80 °C. In total, 40 aliquots with 12 mg of protein each were generated.

Determination of Cellular Growth Inhibition

Sensitivity of the cell lines for dasatinib was determined by measuring the cellular ATP content after 96 h of treatment using the CellTiter Glo chemiluminescent viability assay (Promega, Mannheim, Germany). The cells were cultivated in 96-well plates (Greiner, Frickenhausen, Germany) in the presence of dasatinib (LC Laboratories, Woburn, MA) within a concentration range between 3 nM and 30 μM.

The raw data from the chemiluminometer (FLUOstar OPTIMA; BMG Labtech, Offenburg, Germany) was used to determine the GI_{50} value. First, the background was determined by calculating the median value of the plate's border wells, which contained only growth media. This value was then subtracted from each inner well. Because two experiments were conducted on one 96-well plate with 10 compound concentrations each (0 nM (DMSO control), 3 nM, 10 nM, 30 nM, 100 nM, 300 nM, 1 μM, 3 μM, 10 μM, and 30 μM), three data points per concentration and experiment were available. Ratios representing the percentage of growth inhibition were calculated by dividing each data point coming from a concentration >0 nM by the median of the DMSO values. A logistic regression was performed to fit a curve to those ratios and compute the GI_{50} value.

Classification into Sensitive/Resistant

The calculated GI_{50} values of the 26 selected cell lines were compared with the values reported in Ref. 18. Although the correlation between the two sets was strong (Pearson correlation = 0.50, p = 0.009 on logged GI_{50} values), a few cell lines showed inconsistent behavior. By setting the threshold to discriminate between sensitive and resistant cells to a GI_{50} value of 1 μM, seven cell lines were classified inconsistently (five were resistant in Ref. 18 but sensitive in this study, and two were sensitive in Ref. 18 but resistant in this study). Consequently, these cell lines were excluded from the workflow that aims at finding a predictive phosphosignature.

Phosphoproteomics Workflow

Responsive and nonresponsive cell lines were grown in medium or heavy SILAC medium, and after washing twice with ice-cold PBS, the cells were lysed directly on the plates by the addition of ice-cold lysis buffer (8 M urea, 50 mM Tris, pH 8.2, 5 mM EDTA, 5 mM EGTA, Sigma HALT phosphatase inhibitor mix, Roche Applied Science complete protease inhibitor mix). After sonication cell debris was sedimented by centrifugation, and the protein concentration was determined by Bradford assays. Equal protein amounts of the reference cell culture mix and a medium and heavy labeled cell line (7 mg protein each)

were mixed as depicted in supplemental Fig. S2 and subsequently subjected to reduction (20 mM DTT, 30 min 37 °C) and alkylation (50 mM iodoacetamide for 30 min at room temperature) prior to proteolytic cleavage. Then 80 μg of LysC (Wako) was added for 4 h followed by a 4-fold dilution with 50 mM Tris, pH 8.2. Proteolytic cleavage was continued by the addition of 120 μg of trypsin (Promega) overnight. The peptide mixtures were acidified by the addition of TFA to a final concentration of 0.5% and subsequently desalted via C18 SephPack columns (Waters). The peptides were eluted with 50% ACN and dried under vacuum. For a first separation of phosphorylated and nonphosphorylated peptides, the dried peptide powder was reconstituted in 1 ml of strong cation exchange chromatography buffer A (5 mM K_2HPO_4, pH 2.7, 30% ACN) and loaded onto a polysulfoethyl column (9.4 × 250 mm; PolyLC) using an ÄKTA purifier chromatography system equipped with a fraction collector. The peptides were separated by a linear gradient to 25% strong cation exchange chromatography buffer B (buffer A supplemented with 500 mM KCl) over 40 min at flow rate of 3 ml/min. Twenty fractions (12 ml each) were collected across the gradient.

Prior to immobilized metal ion affinity chromatography enrichment, the solvent of the strong cation exchange chromatography fractions was removed by lyophilization. Dried peptides were reconstituted in 1 ml of 0.1% TFA and desalted by using C18 reversed phase cartridges (Waters). The bound peptides were eluted with 50% ACN, 0.5% HOAc, and the peptides were lyophilized again. Dried peptides were reconstituted in 40% ACN, 25 mM formic acid, and phosphopeptides were captured using PhosSelect (Sigma) according to the manufacturer's instructions. Eluted phosphopeptides were subjected to mass spectrometric analysis.

LC-MS/MS Analysis

Mass spectrometric analysis was carried out by on-line nano LC-MS/MS. The sample was loaded directly by an Agilent 1200 nanoflow system (Agilent Technologies) on a 15-cm fused silica emitter (New Objective) packed in-house with reversed phase material (Reprusil-Pur C18-AQ, 3 μm; Dr. Maisch GmbH) at a flow of 500 nl/min. The bound peptides were eluted by a gradient from 2 to 40% of solvent B (80% ACN, 0.5% HOAc) at a flow of 200 nl/min and sprayed directly into a LTQ-Orbitrap XL or LTQ-Orbitrap Discovery mass spectrometer (Thermo Fischer Scientific) at a spray voltage of 2 kV applying a nanoelectrospray ion source (ProxeonBiosystems). The mass spectrometer was operated in the positive ion mode and with a data-dependent switch between MS and MS/MS acquisition. To improve mass accuracy in the MS mode, the lock mass option was enabled. Full scans were acquired in the orbitrap at a resolution $r = 60{,}000$ (Orbitrap XL) or 30,000 (Orbitrap Discovery) and a target value of 1,000,000 ions. The five most intense ions detected in the MS were selected for collision-induced dissociation in the LTQ at a target value of 5000. The resulting fragmentation spectra were also recorded in the linear ion trap. To improve complete dissociation of phosphopeptides, the multistage activation option was enabled, applying additional dissociation energy on potential neutral loss fragments (precursor minus 98, 49, and 32.7 Thompson). Ions that were once selected for data-dependent acquisition were dynamically excluded for 90 s for further fragmentation.

MaxQuant Analysis

The raw mass spectral data were processed using the MaxQuant software (version 1.1.1.25) (20) applying the Andromeda search engine for peptide and protein identification. The human UNIPROT database (version: 57.12) was used comprising 110,595 database entries including the UNIPROT splice variants database. The minimal peptide length was set to 6 amino acids, trypsin was selected as proteolytic enzyme, and maximally three missed cleavage sites were allowed. Carbamidomethylation of cysteines was selected as a fixed modification, whereas methionine oxidation, N-terminal protein acetylation, and phosphorylation of serine, threonine, and tyrosine residues were considered as variable modifications. Because MaxQuant automatically extracts isotopic SILAC peptide triplets, the corresponding isotopic forms of lysine and arginine were automatically selected. The maximal mass deviation of precursor and fragment masses was set to 20 ppm and 0.5 Da before internal mass recalibration by MaxQuant. A false discovery rate (FDR) of 0.01 was selected for proteins and peptides and a posterior error probability below or equal to 0.1 for each MS/MS spectrum was required. The MaxQuant results were uploaded to the MaxQB database (21) for further analysis.

Data Preprocessing

Data from the MaxQuant PhosphoSTY table were the data source for identifying a predictive phosphosignature. Each entry in this table describes one specific phosphosite along with information about its localization, confidence, and regulation. The regulation of a phosphosite is provided as ratio of the site's abundance between each cell line and the super-SILAC standard. MaxQuant already provides normalized ratios, which were used in this study. There are two coefficients that account for the reliability of identification and localization of a phosphosites, *i.e.* localization probability and score difference. Sites that satisfy the constraints localization probability ≥ 0.75 and score difference ≥ 5 were considered to be sufficiently reliable (class I sites). Furthermore, sites that are flagged as reverse or contaminant hits were also excluded. All phosphosites that fulfill both requirements (class I, no contaminant/reverse) were subjected to further analysis. The identification and quantification data on the class I sites, as well as the fragment spectra of the best localization evidence, are accessible in supplemental Files 2–5.

Analysis of Differential Phosphorylation Sites

Significance Analysis—After preprocessing the data, a Wilcoxon rank sum test was applied to find differentially abundant phosphorylation sites between sensitive and resistant cell lines. For this analysis only phosphosites with values in at least two-thirds of the experiments in each group were considered (*i.e.*, at least 8 of 11 sensitive and 6 of 8 resistant data points had to be present). Subsequently, the *p* values reported by the Wilcoxon rank sum test were corrected for multiple hypotheses testing by applying Benjamini-Hochberg FDR correction (22).

Enrichment Analysis—To analyze whether proteins harboring differentially abundant phosphorylation sites are enriched in certain GO terms (23) or KEGG pathways (24), FatiScan enrichment analysis (25) was applied. In brief, FatiScan performs a segmentation test, which checks for asymmetrical distribution of biological labels (*e.g.*, GO terms, KEGG pathways) associated with proteins in a ranked list. For this purpose, the phosphorylation sites were sorted according to their *q* values, and the algorithm was set up to search for a possible enrichment in the low *q* value area of this ranked list. The analysis was performed via the Babelomics web interface (http://babelomics.bioinfo.cipf.es/, version 4.2).

Detection of Significantly Different Subnetworks

To visualize and interpret the data in a network context, the SubExtractor algorithm was applied (26). In brief, SubExtractor combines phosphoproteomic data with protein-protein interaction data via a Bayesian probabilistic model. Regulated subnetworks are found with a genetic algorithm and subsequent significance evaluation based on

the global rank test (27). The STRING database version 8.3 (28) was used as source for protein-protein interactions. It was preprocessed to contain only human interactions with a confidence score larger than 0.9 without considering text mining evidences. The algorithm's parameters were set to $\alpha = 0.5$ and $\sigma = 5.0$, and subnetworks with an FDR smaller than 0.1 were reported.

To calculate *z* scores required as input for the algorithm, pair-wise phosphorylation abundance differences between sensitive and resistant cell lines had to be computed first. Because the number of experiments in the two groups are not balanced (11 and 8, respectively), sampling with replacement was applied to the smaller group (*i.e.*, it was sampled 11 times from eight experiments while ensuring that each experiment was chosen at least once). Subsequently, the pair-wise differences could be computed along with the estimated global standard deviation as suggested in Ref. 26, and finally the *z* scores were calculated.

Identification and Evaluation of Phosphosignature

Cross-validation—The data set containing $N = 19$ objects was split into two parts, one containing data of one cell line, and the other containing the data of the remaining $N - 1$ cell lines. The larger part was then used for training a predictor (training set) and the smaller one for testing this predictor (test set). By alternating the cell lines that made up the training set, each cell line was used once for testing. Each of the *N* cross-validation steps included missing data imputation, feature selection, predictor training, and predictor testing.

A phosphosite was only considered as a potential feature if it had training data values in at least two-thirds of the experiments in each class (*e.g.*, if the training set contained data from 10 sensitive and 8 insensitive cell lines, at least 7 and 6 training data points had to be present, respectively). Because this criterion uses the class labels, the features have to be filtered within the CV loop. It further means that the filtered features may be different in each CV step.

Data Imputation—For each phosphosite and class, the mean and standard deviation was computed, and the missing values were filled by sampling from the resulting normal distribution. This procedure was only applied to the training data, because the test data should be handled as if the class association was unknown. Nevertheless, test data can also contain missing values. If so, the mean of the corresponding two group means was imputed, which is an unbiased way of replacing the missing value that does not involve information about the test sample's class association. Geometrically speaking, the imputed test sample value is located exactly halfway between the two class means, which should minimize its influence on the prediction process.

Feature Selection—In this study, a simple Wilcoxon rank sum test in combination with the ensemble feature selection method (29) was used. As the Wilcoxon test often delivers identical *p* values because of its rank-based nature, ties were broken by preferring features that have a larger difference in their two classes' medians. The core idea of the ensemble method is that robust features should still rank among the best if the data set is slightly modified. For this purpose, different samplings of the training data were generated by drawing (with replacement) 50 different bootstrap samples (*i.e.* if the training set consists of 10 sensitive and 8 resistant cell lines, one randomly draws 10 and 8 times with replacement from the respective set to get one bootstrap sample). The Wilcoxon rank sum test is applied to each sample, and thus a diverse set of feature rankings is generated. The ranks of each feature were then averaged across all bootstrap runs and sorted in ascending order according to this metaranking. Subsequently, the *k* best features were used to train and test the predictor. By varying K and assessing the prediction accuracy and area under the receiver operator curve (AUROC), one can find the optimal number of features.

Support Vector Machine Training—Once a set of features has been selected, and the training and test data have been modified to include only those features (*i.e.*, "reduced" sets), a SVM with linear kernel (30) can be trained. In addition to the kernel function, an SVM has a parameter *C* that controls the trade-off between margin maximization and training error minimization, if the hyper plane cannot perfectly separate the two classes. The default value of $C = 1$ was used throughout the analysis. First, the SVM was trained with the training data. Subsequently, the class association of the test data was predicted with the trained SVM. The result of this prediction is the probability of the test sample belonging to either of the two classes (the closer the test data is to the decision boundary, the less confident the prediction is). The class prediction with the larger probability was then taken and compared with the actual class association. In this way, correct predictions were counted across all cross-validation steps.

Area under the Receiver Operating Characteristic Curve—To calculate the AUROC, the separating hyperplane of a trained SVM was shifted by introducing cost matrices. For example, by shifting the hyperplane toward the group of sensitive training samples, it becomes more likely for a test sample to be classified as resistant. Ultimately, this shifting leads to the extreme that every test sample is classified as resistant, which means that all resistant test samples have been classified correctly (true negative rate = 1 and false positive rate = 0, given that the resistant ones are the negatives) and all sensitive test samples have been classified wrongly (true positive rate = 0). The exact opposite is true if the separating hyperplane is shifted toward the resistant group. Thus, by applying different cost values, one can control the degree of shifting, calculate the respective true positive rates and false positive rates, and compute the resulting area under the curve by means of the trapezoidal rule (see supplemental materials for an example).

Random Seeds—For the imputation of missing values, a random number generator is needed to sample values from a normal distribution. Different seeds of the random generator will produce different imputation data. To avoid a bias of the data toward the seeding, the entire cross-validation procedure was repeated five times using different random number generator seeds. The prediction accuracies, AUROC values and global feature rankings for different numbers of selected features (*k*) were averaged over the five CV runs and used for the final selection of the phosphosignature.

Data Normalization—Among the fraction of nonphosphorylated peptides, 15 peptides had values in at least two-thirds of the experiments and a standard deviation of <0.1 (log 10 scale). Eight of them were from ribosomal proteins, which are expected to be constantly expressed. Thus, for each experiment the median of the corresponding eight ratios was computed and used as an alternative normalization approach (by subtracting the median from each phosphosite's non-MaxQuant-normalized logarithmic ratio).

Final Predictor Construction—When selecting the final set of phosphosites (phosphosignature) to be used for the prediction of future samples, the optimal number of features was determined in a CV loop. This is essentially the same as the inner loop in the quality assessment process (see also supplemental Fig. S4).

Therefore, after running the cross-validation process five times with different random number generator seeds, we obtained the following results: a 200×5 prediction result matrix (200 being the rows, 5 being the columns) containing the number of correct CV predictions for $k = 1 \ldots 200$ selected features (*i.e.*, *k* best ranking in each CV step) across the five random seeds; a 200×5 AUROC matrix containing the corresponding area under the ROC curve values; and a $25{,}020 \times 19 \times 5$ rank matrix holding the rank of each feature in each CV step across the five random seed runs (features that were not subjected to imputation/feature selection because of too many missing values

received the rank maxRank+1, where maxRank is the number of features that were subjected to imputation/feature selection).

The primary criterion for selecting the best subset of features was the number of correct predictions. For this purpose the values in the prediction matrix and AUROC matrix were row-averaged, leading to a vector of 200 average correct predictions and area under the curve values. Within this vector the indices (numbers of features) that lead to the best number of correct predictions were determined. Among those, the one index that had the highest AUROC value was selected as best performing feature number, which was 12.

Next, the final feature rank was determined by averaging first over the third and subsequently over the second dimension of the rank matrix. The resulting vector of length 25,020 containing the average rank of each feature was sorted in ascending order, and the 12 top-ranked were selected. These were the phosphosites described in Table II.

The 12 selected final features were then used to train the final predictor. However, because these features also contained missing values, imputation had to be performed first. The original sampling should reflect the variance within each feature and class, which is crucial for the quality of a feature. Because the best features had already been selected at this stage, sampling can influence the feature weights in the final predictor only. We used the mean of each feature and class for replacing missing values in the data set for the final predictor. Alternatively, we could use the same sampling approach as above and then aggregate the resulting predictors by, for example, averaging the classification score. The differences in these two alternatives are only marginal (supplemental Fig. S8). Finally, a SVM based on the predictive 12-site phosphosignature (again with linear kernel and $C = 1$) was trained and can now be applied to the classification of new samples.

Quantitative Western Blot Analysis

For protein detection in human lung cancer cell lines, exponentially growing cells from 15-cm dishes were used. After cell lysis 80 μg of total protein was separated on 4–12% Bis-Tris NuPAGE gels (Invitrogen) for the detection of integrin β4 or on 7.5% Tris-glycine gels (Bio-Rad Mini PROTEAN) for the detection of tankyrase 1-binding protein (TNKS1BP1). The proteins were transferred overnight to 0.2-μm nitrocellulose membranes and probed with the appropriate antibodies in LI-COR Odyssey blocking buffer. All of the primary antibodies were used in 1:1000 dilutions: anti-integrin β4 antibody [M126] (ab29042; Abcam); anti-TNKS1BP1 (SAB4503414; Sigma-Aldrich); and anti-actin (I-19) (sc-1616-R; Santa Cruz Biotechnology). Actin served as a loading control. Following primary antibody incubation, membranes were probed with IRDye 800CW-conjugated goat anti-mouse IgG (H+L9 (LI-COR number 926-32210), dilution 1:15,000 for the detection of integrin β4; or IRDye 800 conjugated affinity-purified anti-rabbit IgG, (611-732-127; Rockland), dilution 1:20,000, for the detection of TNKS1BP1 and actin; or DyLight 800-conjugated affinity-purified anti-rabbit IgG (H+L) (611-145-122; Rockland), dilution 1:50,000 for the detection of actin. The signals were detected at 800 nm using the LI-COR Odyssey infrared system.

RESULTS

Dasatinib Sensitivity Was Confirmed—Based on the half-maximum growth inhibitory concentration (GI_{50}) of dasatinib reported previously (18), 13 sensitive and 13 resistant NSCLC cell lines were preselected. For these 26 cell lines, we repeated viability assays to verify the reported GI_{50} values. We chose the median GI_{50} as classification threshold, so that depending on the GI_{50}, the cell lines were assigned to sensitive ($GI_{50} < 1$ μM) and resistant ($GI_{50} > 1$ μM) classes. For 19 of 26 cell lines, the assignment was consistent. For 7 cell lines, the assignment based on the sensitivity determined here differed from that reported previously (18). By using only the cell lines for which the sensitivity could be reproduced in two different labs, we maximize the reproducibility of the cell line assignment and therewith the robustness of the predictive signature. The other cell lines were therefore excluded from the training set (see supplemental Table S1 for GI_{50} values). The remaining 19 cell lines (11 sensitive and 8 resistant) were used to identify a predictive phosphosignature. The peak dasatinib plasma concentration (C_{max}) obtained in a phase II trial in patients with advanced NSCLC was 124 $\pm$ 59 ng/ml (16). The corresponding molarity is below the classification threshold chosen above. However, only the GI50 values of two cell lines, HCC4006 and H322M, are marginally higher than the average peak plasma concentration.

Phosphoproteomic Profiling Reveals Differentially Phosphorylated Proteins—To quantitatively compare the cell lines to be analyzed, we isotopically labeled sensitive and resistant NSCLC cell lines using stable isotope labeling by amino acid in cell culture (SILAC) (19). The sensitive cell lines were grown in SILAC medium supplemented with the medium forms of arginine and lysine (Arg^6/Lys^4), whereas the resistant cell lines were grown in heavy medium (Arg^{10}/Lys^8; see supplemental Table S2 for experimental pairing scheme). A Super-SILAC reference (31) was generated by mixing protein lysates of 16 randomly selected cell lines in unlabeled (light, Arg^0/Lys^0) medium. The Super-SILAC reference serves as a spike-in standard, enabling accurate cross-sample comparison (see supplemental Fig. S2). Equal protein amounts of the Super-SILAC reference, a sensitive cell line, and a resistant cell line were mixed and subsequently subjected to a global, quantitative phosphoproteomics workflow using strong cation exchange chromatography and immobilized metal ion affinity chromatography followed by LC-MS/MS analysis (see "Experimental Procedures" for details). In total, 37,747 phosphosites were identified in the 26 profiled cell lines. 88% of all quantified phosphorylation sites had a cell line to Super-SILAC ratio <4-fold, which allowed for accurate quantification of phosphorylation changes between the analyzed cell lines. From the 37,747 identified phosphorylation sites, 25,020 were rated as class I sites, *i.e.*, sites that could be identified with high localization confidence (7). Only these sites were used in the following analyses. The frequency distribution of the phosphorylated residues (serine, 83.2%; threonine, 15.3%; and tyrosine, 1.5%) is similar to the frequency distribution observed by Olsen *et al.* (7).

We first tried to identify proteins that are differentially phosphorylated between the sensitive and resistant cell lines. To this end, the Wilcoxon rank sum test was applied to the set of phosphosites with data values in at least two-thirds of the experiments (leading to 4457 valid sites with ~11% missing values on average). Indeed, 58 phosphosites were signifi-

cantly regulated between the group of 11 sensitive and 8 resistant cell lines at a FDR of 10% (Table I). The regulated sites reside on 41 unique proteins. Most of the regulated sites (53 or 91%) are phosphorylated more strongly in sensitive cell lines. Only 5 (9%) sites are more strongly phosphorylated in resistant cell lines. For three known dasatinib targets, Bcr-Abl, EphA2, and Lyn (14, 15), we could detect phosphosites that were quantified in two-thirds of the experiments. The phosphorylations of EphA2 and Lyn cannot differentiate between the sensitive and resistant groups (supplemental Fig. S6). Only the site Ser^{459} on the breakpoint cluster region protein (Bcr) is differentially phosphorylated (see supplemental Table S1 and Fig. S6).

We next investigated whether any KEGG pathway or Gene Ontology term is enriched in the set of proteins with differential phosphosites. The list of proteins ordered by the Wilcoxon rank sum test statistic of their most significant phosphosite were analyzed with FatiScan (25). Only the KEGG pathway "regulation of actin cytoskeleton" (hsa04810) is significantly enriched at an FDR of 5%. Many of the significantly regulated phosphosites are located on proteins involved in this pathway. A similar analysis revealed that 40 terms of the biological process and the molecular function gene ontologies are significantly enriched (see supplemental Table S3). Many of them relate to very generic and not surprising terms, like "kinase activity" (GO:0016301) or "signal transduction" (GO:0007165). However, a few of them are more specific, like "Ras protein signal transduction" (GO:0007265) and "Rho protein signal transduction" (GO:0007266) in the biological process ontology and "cytoskeletal protein binding" (GO:0008092) and "actin binding" (GO:0003779) in the molecular function ontology.

As a next step, we applied the SubExtractor algorithm (26) to the phosphoproteomic data. SubExtractor detects significantly regulated subnetworks in the STRING protein-protein interaction network (28). The tool combines local as well as topological information, *i.e.*, information about the regulation of a certain node (represented by the protein's strongest regulated phosphorylation site) and information about the connectivity with its neighbors. The largest subnetwork that has been identified by SubExtractor (Fig. 1) clustered around the EGF receptor, with most of the proteins again being more strongly phosphorylated in the sensitive cells. The largest subnetwork comprises many proteins involved in cell adhesion and actin cytoskeleton organization, such as ajuba (JUB), catenin α1 (CTNNA1) and δ1 (CTNND1), ephrin type-A receptor 2 (EPHA2), brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAP2), integrin β4 (ITGB4), and plectin (PLEC1).

A Predictive Phosphosignature Was Identified—Following the general workflow for detecting phosphosignatures (Fig. 2), a predictive phosphosignature was identified, and its accuracy was estimated by cross-validation (CV) based on the cell line data set (19 valid cell lines). Feature selection was applied within each CV loop to reduce dimensionality of the data and thus avoid overfitting the resulting predictor. We used a Wilcoxon rank sum test combined with the ensemble method (29) for selecting the phosphosites used for the signatures. The number of phosphosites is optimized in an inner leave-one-out cross-validation loop. The phosphosites were used to train a support vector machine (SVM) with linear kernel, which was chosen as the predictor because it offers state-of-the-art prediction quality and has been successfully applied several times to biological data (32–34). SVMs separate two classes by a hyper plane, such that the margin between the classes becomes as wide as possible (30).

The final phosphosignature comprises 12 phosphosites (Table II) located on nine different proteins. The phosphorylation degrees of the 12 identified sites strongly separate the class of sensitive and resistant cell lines (Fig. 3). All of them are more strongly phosphorylated in the sensitive cell lines. The five highest ranked phosphosites show ~10-fold differences in their medians. The differences between the 25th and 75th percentiles are still ~5-fold. Interestingly, four of the highest ranked phosphosites are located on the same protein integrin β4 (ITGB4 or CD104). The second highest ranked phosphosite is located on the brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAP2). Further, we identified phosphosites that are located on the G-protein-coupled receptor family C group 5 member A (GPCRC5A), the inositol 1,4,5-triphosphate receptor type 3 (ITPR3), the 192-kDa tankyrase-1-binding protein (TNKS1BP1), the Rho guanine nucleotide exchange factor 18 (ARHGEF8), the RelA-associated inhibitor (IASPP), the autophagy-related protein 16–1 (APG16L), and the tumor protein D54 (TPD52L2).

Signature Is Sensitive and Specific—To determine the prediction performance, leave-one-out cross-validation was applied. It has been shown that CV, including leave-one-out cross-validation, estimates the true prediction performance accurately and shows a low bias (35). Because not all phosphosites discriminate well between sensitive and resistant cell lines, feature selection is applied in each CV step, which selects a defined subset of predictive phosphosites. First the features are ranked according to their discriminative power, and then the optimal number of top-ranking features is determined by an inner parameter optimization cross-validation. In this inner CV procedure, different numbers of top ranking features ($k = 1 \ldots 200$) are used, and their respective performances are assessed. The smallest number of features leading to the best prediction quality in the inner CV loop is then applied to the feature selection in the outer cross-validation loop (see also supplemental Fig. S3). Subsequently, a SVM predictor is trained on the reduced training data (reduced in the sense of containing only features that passed the feature selection criteria) and tested with the reduced test data. It is important to note that the test sample is used neither for optimizing the number of features nor for selecting the features within cross-validation. Furthermore, the preprocessing steps and classification workflow were fixed before acquiring

TABLE I

Significantly different phosphorylation sites

Shown are sites that are differentially phosphorylated between sensitive and resistant cell lines.

Uniprot identification code	Gene name	Protein name	Site	Modified sequence[a]	Median difference[b]	*q* value[c]
A8K556	GPCR5A	Retinoic acid-induced protein 3	Ser^{345}	AHAWPpSPYKDYEVK	0.872	0.047
Q6ZSZ5	ARHGEF18	Rho guanine nucleotide exchange factor 18	Ser^{1101}	pSLSPILPGR	0.419	0.047
Q13177	PAK2	Serine/threonine protein kinase PAK 2	Ser^{141}	YLpSFTPPEKDGFPSGTPALNAK	0.315	0.047
Q15149-2	PLEC1	Plectin	Ser^{42}	pSGGGAGSNGSVLDPAER	0.334	0.047
Q9C0C2	TNKS1BP1	182-kDa tankyrase-1-binding protein	Ser^{429}	RFpSEGVLQSPSQDQEK	0.968	0.047
P16144-2	ITGB4	Integrin β4	Ser^{1424}	DYNpSLTR	1.406	0.055
P16144-2	ITGB4	Integrin β4	Ser^{1387}	MDFAFPGSTNpSLHR	0.992	0.055
Q6ZSZ5	ARHGEF18	Rho guanine nucleotide exchange factor 18	Ser^{1103}	RSLpSPILPGR	0.345	0.055
Q3KQU3	MAP7D1	MAP7 domain-containing protein 1	Ser^{116}	RSSQPpSPTAVPASDSPPTK	0.514	0.055
Q86SQ0	LL5B	Pleckstrin homology-like domain family B member 2	Ser^{212}	KMpSIQDSLALQPK	0.721	0.055
Q8IVF2	AHNAK2	Protein AHNAK2	Ser^{2657}	FKMPpSFR	0.909	0.055
Q92614	KIAA0216	Myosin-XVIIIa	Ser^{1970}	LEGDpSDVDSELEDRVDGVK	0.657	0.055
Q9Y2U5	MAP3K2	Mitogen-activated protein kinase kinase kinase 2	Ser^{153}	RLpSIIGPTSR	0.494	0.055
P49792	RGP3	E3 SUMO-protein ligase RanBP2	Thr^{799}	pTPPRWAEDQNSLLK	-0.261	0.055
B2R5W6	MAPRE3	Microtubule-associated protein RP/EB family member 3	Thr^{164}	LIGTAVPQRTSPpTGPK	0.447	0.055
B8QGS6	PKP2	Plakophilin-2	Ser^{151}	LEIpSPDSSPER	0.742	0.079
B4DIK2	NUP153	Nuclear pore complex protein Nup153	Ser^{338}	RIPSIVSSPLNpSPLDR	-0.317	0.079
Q13177	PAK2	Serine/threonine protein kinase PAK 2	Ser^{2}	pSDNGELEDKPPAPPVR	0.234	0.079
O15231-3	ZNF185	Zinc finger protein 185	Ser^{469}	RESCGpSSVLTDFEGK	1.662	0.080
O43399-2	TPD52L2	Tumor protein D54	Ser^{141}	KLGDMRNpSATFK	0.563	0.080
Q14573	ITPR3	Inositol 1,4,5-trisphosphate receptor type 3	Ser^{916}	pSIQGVGHMMSTMVLSR	0.782	0.080
Q676U5	APG16L	Autophagy-related protein 16-1	Ser^{269}	RLpSQPAGGLLDSITNIFGR	0.725	0.080
Q86SQ0	LL5B	Pleckstrin homology-like domain family B member 2	Ser^{513}	KDpSLPDADLASCGSLSQSSASFFTPR	0.606	0.080
B8QGS6	PKP2	Plakophilin-2	Ser^{154}	RLEISPDpSpSPER	0.688	0.082
P16144-2	ITGB4	Integrin β4	Ser^{1445}	DYSTLTpSVSSHDSR	1.473	0.082
P16144-2	ITGB4	Integrin β4	Ser^{1448}	DYSTLTSVSpSHDSR	1.544	0.082
P16144-2	ITGB4	Integrin β4	Ser^{1069}	LLELQEVDpSLLRGR	1.236	0.082
A6NDI6	FNBP1L	Formin-binding protein 1-like	Ser^{490}	RHSpSDINHLVTQGR	0.239	0.082
A8K1D2	LASP1	LIM and SH3 domain protein 1	Ser^{146}	MGPSGGEGMEPERRDpSQDGSSYR	0.366	0.082
A8K7M3	Sep 10	Septin-10	Ser^{451}	KNpSNFL	1.015	0.082
A9UF02	BCR/ABL	BCR/ABL fusion protein	Ser^{459}	HQDGLPYIDDpSPSSSPHLSSK	0.270	0.082
B3KSZ4	GATAD2B	Transcriptional repressor p66-beta	Ser^{129}	LTPSPDIIVLpSDNEASSPR	-0.181	0.082
D6W4Y8	ASAP2	Arf-GAP (SH3, ANK, PH) protein 2[d]	Ser^{701}	LLHEDLDEpSDDDMDEKLQPSPNR	0.430	0.082
O60303	KIAA0556	Uncharacterized protein KIAA0556	Ser^{691}	KDpSLSQLEEYLR	0.618	0.082
Q52LW3	ARHGAP29	Rho GTPase-activating protein 29	Ser^{1019}	IRPVpSLPVDR	1.340	0.082
Q8WUF5	IASPP	RelA-associated inhibitor	Ser^{102}	SEpSAPTLHPpYSPLSPK	0.528	0.082
Q9UQB8-5	BAIAP2	Brain-specific angiogenesis inhibitor 1-associated protein 2	Ser^{509}	pSMSSADVEVARF	1.197	0.082
P16144-2	ITGB4	Integrin β4	Thr^{1385}	MDFAFPGSpTNSLHR	0.937	0.082
B8QGS6	PKP2	Plakophilin-2	Ser^{155}	RLEISPDSpSPER	0.854	0.083
O15231-3	ZNF185	Zinc finger protein 185	Ser^{466}	REpSCGSSVLTDFEGK	1.560	0.083
P23528	CFL	Cofilin-1	Ser^{156}	LGGpSAVISLEGKPL	0.445	0.083
Q13439	GOLGA4	Golgin subfamily A member 4	Ser^{78}	VPpSVESLFRpSPIK	0.468	0.083
Q8N4C8	MINK	Misshapen-like kinase 1	Ser^{699}	pSNSAWQIYLQR	0.486	0.083
Q14573	ITPR3	Inositol 1,4,5-trisphosphate receptor type 3	Ser^{934}	KQpSVFSAPSLSAGASAAEPLDR	0.788	0.086
Q9BY89	KIAA1671	Uncharacterized protein KIAA1671	Ser^{1800}	KRQpSLYENQV	0.422	0.086
B8QGS6	PKP2	Plakophilin-2	Ser^{251}	pSMGNLLEK	0.655	0.096
B2RBM8	ADNP	Activity-dependent neuro-protector homeobox protein	Ser^{769}	KpSFLpTKYFNK	0.793	0.096
Q8NEY8	HSPC206	Periphilin-1	Ser^{133}	DNTFFREpSPVGR	-0.213	0.096
D3DXE9	BAZ1B	Tyrosine-protein kinase BAZ1B	Ser^{1468}	LAEDEGDpSEPEAVGQSR	-0.217	0.096
P28066	PSMA5	Proteasome subunit alpha type-5	Ser^{16}	GVNTFpSPEGR	0.430	0.096
Q53EP0	FAD104	Fibronectin type III domain-containing protein 3B	Ser^{208}	LNpSPPSSIYK	0.391	0.096
Q6ZRV2	FAM83H	Protein FAM83H	Ser^{870}	GpSPTSAYPER	1.049	0.096
Q6ZRV2	FAM83H	Protein FAM83H	Ser^{936}	GpSLTLTISGESPK	1.026	0.096
Q6ZRV2	FAM83H	Protein FAM83H	Ser^{785}	pSLESCLLDLR	0.795	0.096
Q86SQ0	LL5B	Pleckstrin homology-like domain family B member 2	Ser^{415}	KSpSISSISGR	0.631	0.096
Q86YV5	SGK223	Tyrosine-protein kinase SgK223	Ser^{696}	SApSFAFEFPK	0.716	0.096
Q8TDM6	DLG5	Disks large homolog 5	Ser^{264}	NLLQQpSWEDMKR	0.518	0.096
Q3KQU3	MAP7D1	MAP7 domain-containing protein 1	Thr^{118}	RSSQPpSPpTAVPASDSPPTK	0.396	0.096

[a] Sequence of the peptide on which the phosphosite was detected; p indicates that the subsequent amino acid was phosphorylated.

[b] Median difference of log10 ratios between sensitive and resistant classes.

[c] FDR-corrected Wilcoxon rank sum *p* value.

[d] Full name: Arf-GAP with SH3 domain, ANK repeat, and PH domain-containing protein 2.

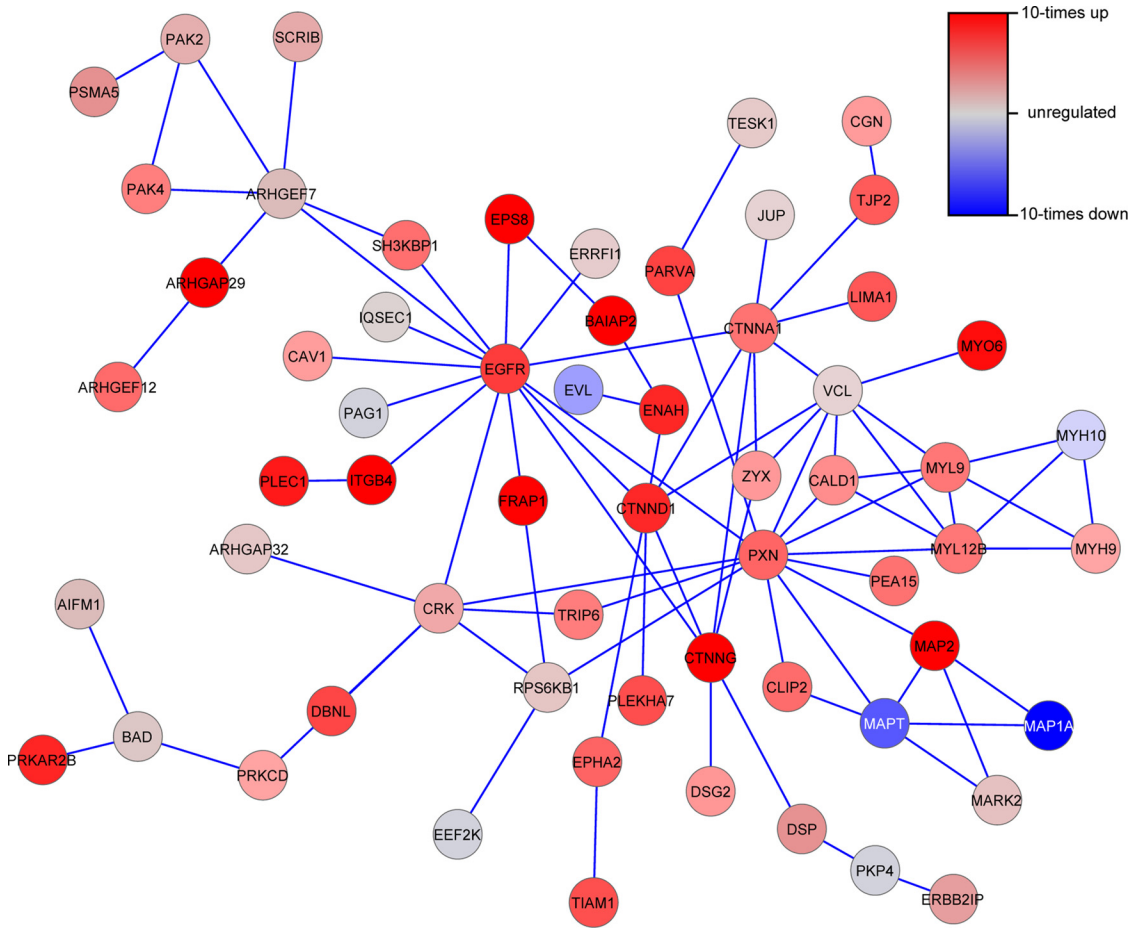

FIG. 1. **Protein-protein interaction subnetwork showing differential phosphorylation in sensitive and resistant cells.** The subnetworks were identified using the SubExtractor algorithm. Only the largest network is shown. *Red nodes* are more strongly phosphorylated, and *blue nodes* are more weakly phosphorylated in sensitive than in resistant cells.

the NSCLC data. Otherwise, the prediction accuracy would be overestimated.

Missing data are a common phenomenon in shotgun proteomics. Although the quantitative information (*i.e.*, SILAC peaks) of a peptide may be present in the MS spectrum, at least one of the SILAC peaks has to be selected for fragmentation. In this case, the resulting fragment spectrum is used to identify the corresponding peptide. Because the selection of peptides for fragmentation is data-dependent, a certain peptide may be selected in some MS runs but not in others. Therefore, a missing value does not necessarily mean that the corresponding phosphopeptide was not present. This is particularly true when applying the Super-SILAC approach like in this study.

Because many machine learning techniques (SVMs among them) cannot handle missing values, they were replaced by estimated values that were randomly sampled from the respective empirical distribution. As a consequence, the entire assessment was carried out five times with different seeds for the random number generator used for imputation, leading to five distinct prediction results. The five results were strikingly similar as can be expected from a robust set of features, *i.e.*, four times only one cell line was misclassified (HCC78), and once two were falsely classified (HCC78 and HCC827), which leads to a prediction accuracy of 94% and an area under the receiver operating characteristic curve (AUROC) of 0.92 (Fig. 4*A*). Each *circle* in Fig. 4*A* shows the averaged predicted outcome of this cell line when all other cell lines were used as training data. A sensitive cell line is predicted correctly if the SVM predictor assigns a negative value, and a resistant cell line is predicted correctly if the SVM predictor assigns a positive value. The larger the distance to the separating hyperplane (*i.e.*, the distance from 0 in the plot), the more confident the prediction is. It can be clearly seen that 18 of 19 cell lines were predicted correctly by cross-validation.

For the final predictor, the workflow was carried out with only one CV loop, corresponding to the inner loop during the prediction quality assessment (see supplemental Fig. S4). This resulted in identifying a predictive phosphosignature containing the 12 phosphosites. Interestingly, the average number of selected features within the inner parameter optimization loop during the prediction quality assessment was

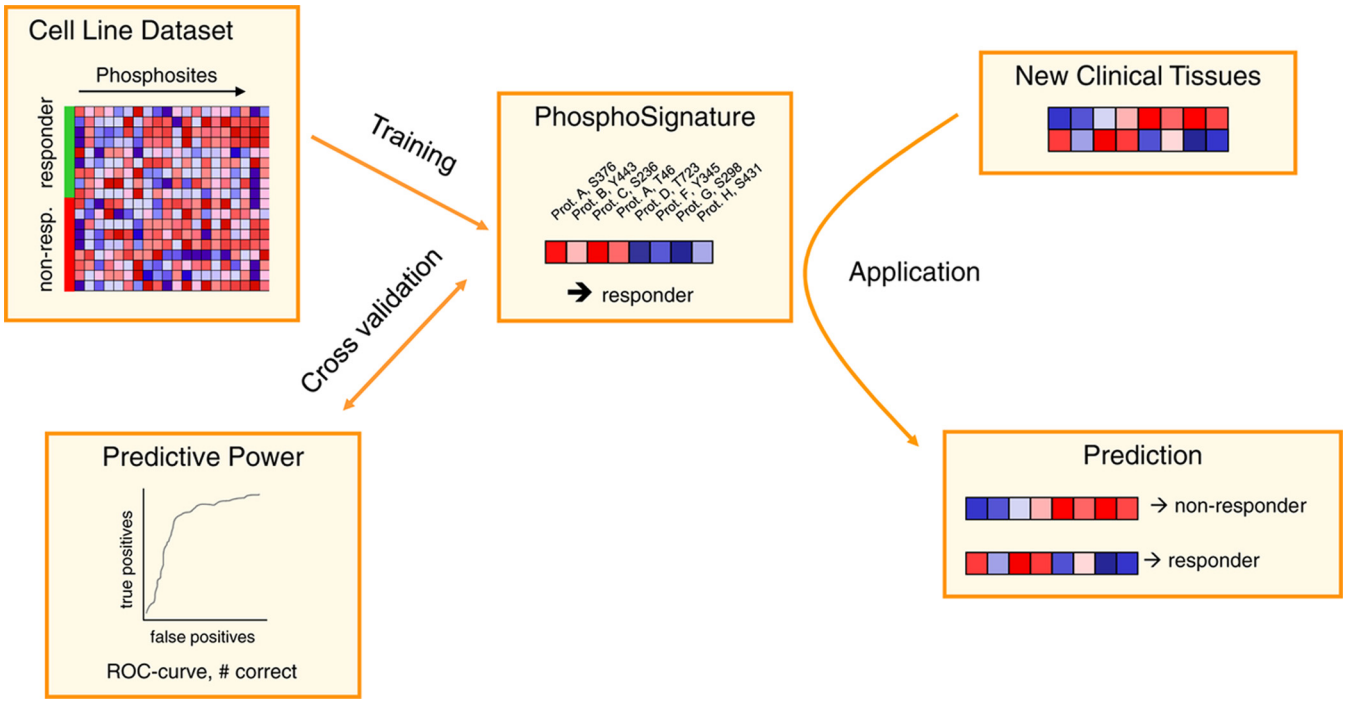

FIG. 2. **The general workflow of phosphobiomarker classification.** First, a predictive phosphosignature is identified based on phosphoprofiles of sensitive and resistant cell lines using the cross-validation approach (described in detail in the text). Once this signature has been identified, it can be applied to new samples to predict the response of the donor to the respective drug.

also ~12, which further supported the robustness of the selected set of phosphosites. The sites are listed in Table II sorted by their global feature ranks and depicted as a heat map in Fig. 5 (see also supplemental Table S4 for more details and supplemental Table S5 for observed ratios). With an increasing number of features the prediction accuracy also increased, until it saturated at 12 features (see supplemental Fig. S5). Additional features did not improve the prediction accuracy.

These results show that a predictive phosphosignature can be identified from phosphoproteomics data. However, the question remains regarding whether the identified signature is specific to dasatinib or whether it also works for other substances not related to dasatinib. As a first step to answer this question, we applied the prediction quality assessment workflow to randomized class labels. Strikingly, the prediction accuracy was only 51% (AUROC = 0.53), which is almost exactly what one would expect if predicting the classes by chance. Thus, a predictive signature cannot be found for arbitrary class associations. As a next step, we investigated whether the classification scores of the final predictor correlate with the cell doubling times of untreated cell lines. The classification score corresponds to the distance from the SVM classification hyperplane and can be interpreted as the confidence in correct classification. In particular, the score is negative (positive) if the sample is predicted as being sensitive (resistant). The cell doubling times range from 25 to 55 h (supplemental Table S1). A Pearson correlation coefficient of −0.08 (p value 0.79) indicates that the doubling times are not associated with the classification. In contrast, the correlation between classification scores and GI_{50} values of dasatinib is significant (0.81, p = 2.6E-6). Finally, we sought to show whether the dasatinib signature is predictive for other substances. The small molecule sorafenib (Nexavar®; Bayer) is a multikinase inhibitor targeting the Raf/Mek/Erk and the vascular endothelial growth factor receptor pathway. The correlation between the doubling times and GI_{50} values of sorafenib (18) is −0.05 (p value 0.83). Taking these results together, we could demonstrate that the identified phosphosignature is specific for predicting response to treatment with dasatinib.

The Phosphosignature Is Robust—A good feature and consequently a good set of features should be robust to small variations in the data. Only when slight changes in the data composition still lead to correct predictions is the biomarker reliably applicable to samples not used for training. Therefore, robustness already plays a crucial role in the process of feature selection. First, a robust feature is chosen frequently by the feature selection method across all of the cross-validation steps. Second, within each cross-validation step, slight variations in the training data should also result in the constant selection of robust features.

To identify such robust phosphosites, we applied the Wilcoxon rank sum test in combination with the ensemble feature selection method (29) to get a feature ranking in each CV step. The average of these ranks across all CV iterations for the

TABLE II
Phosphorylation sites of the final phosphosignature

Uniprot identification code	Gene/protein name	Site	Modified sequence[a]	Average rank[b]	Median difference[c]	Rank ≤12[d]	SV weight[e]
P16144-2	ITGB4 Integrin β4	Ser[1448]	DYSTLTSVSpSHDSR	2.716	1.544	18	−0.386
Q9UQB8-5	BAIAP2 Brain-specific angiogenesis inhibitor 1-associated protein 2	Ser[509]	pSMSSADVEVARF	3.611	1.197	18	−0.311
P16144-2	ITGB4 Integrin β4	Ser[1387]	MDFAFPGSTNpSLHR	4.337	0.992	19	−0.155
P16144-2	ITGB4 Integrin β4	Thr[1385]	MDFAFPGSpTNSLHR	5.716	0.937	18	−0.275
P16144-2	ITGB4 Integrin β4	Ser[1069]	LLELQEVDpSLLRGR	7.937	1.236	13	−0.076
A8K556	GPCR5A Retinoic acid-induced protein 3	Ser[345]	AHAWPpSPYKDYEVK	9.632	0.872	16	−0.174
Q14573	ITPR3 Inositol 1,4,5-trisphosphate receptor type 3	Ser[916]	pSIQGVGHMMSTMVLSR	14.168	0.782	8	−0.205
Q9C0C2	TNKS1BP1 182-kDa tankyrase-1-binding protein	Ser[429]	RFpSEGVLQSPSQDQEK	15.032	0.968	1	−0.159
Q6ZSZ5	ARHGEF18 Rho guanine nucleotide	Ser[1101]	pSLSPILPGR	16.874	0.419	0	−0.188
Q8WUF5	IASPP RelA-associated inhibitor exchange factor 18	Ser[102]	SEpSAPTLHPYSPLSPK	17.516	0.528	7	−0.145
Q676U5	APG16L Autophagy-related protein 16–1	Ser[269]	RLpSQPAGGLLDSITNIFGR	18.19	0.725	13	−0.24
O43399-2	TPD52L2 Tumor protein D54	Ser[141]	KLGDMRNpSATFK	18.274	0.563	8	−0.155

[a] Sequence of the peptide on which the phosphosite was detected; p indicates that the subsequent amino acid was phosphorylated.
[b] The average rank of the feature across all cross-validation steps.
[c] Median difference of log10 ratios between sensitive and resistant classes.
[d] The number of times the feature was among the 12 best across all CV steps.
[e] The importance of the feature in the SVM predictor (the larger the absolute weight, the more important).

signature's 12 features along with the number of times each of them was ranked under the first 12 positions are listed in Table II. The best features turned out to be very stable, *e.g.*, the top four have an average rank smaller than 6 and were among the 12 best more than 90% of all iterations. The importance of these features is also indicated by their high weight in the SVM. Overall, 7 features are among the 12 best in more than two-thirds of the iterations, and only 2 in less than one-third.

To ensure that the SILAC labeling procedure of cell lines has no effect on the results, label switch experiments were performed, where originally medium-labeled cell lines were now labeled with heavy amino acids and vice versa. The classification results of the final predictor applied to these experiments are depicted in Fig. 4*B*. For two of the three label-switched samples, the prediction is virtually identical to the original data (Fig. 4*B*, *circles* and *crosses* at positions 11 and 14). In the case of the position 4 (H322M), the difference is somewhat larger, but the corresponding label switch experiment is still classified correctly.

Because phosphosites in this study are detected in a global and unbiased way, we applied global normalization strategy during the discovery phase. However, when the phosphosignature is applied in the clinic, a method that specifically measures the phosphosites of the signature in a robust and cheap way is more likely to be used (see the supplemental materials for how SVM predictor can be adapted to use data from other methods). Such targeted methods could be either based on phospho-specific antibodies (*e.g.*, immunohistochemistry or ELISA based assays) or targeted mass spectrometry methods such as multiple reaction monitoring (36, 37). Because a global normalization strategy is not applicable to targeted methods, it is necessary to develop an alternative. We focused on nonphosphorylated peptides that showed a very low variance across the regulation data of all cell lines regardless of whether the cell line was sensitive or resistant. Although the phosphoproteomic workflow is designed to specifically enrich for phosphorylated peptides, a significant fraction of nonphosphorylated peptides is still present. In this study, a normalization factor based on a set of nonphosphorylated ribosomal proteins exhibiting low variance across all cell lines proved useful (see supplemental Table S6 for normalization data). The classification results of the ribosomal protein normalized data are depicted in Fig. 4*C*, which shows that the prediction quality is essentially as good as for the globally normalized data the predictor was trained on.

The Phosphosignature Was Validated in Breast Cancer Cells—To test whether the phosphosignature is also applicable to other cancer types, we selected three sensitive and three resistant breast cancer cell lines. Again, the GI_{50} values

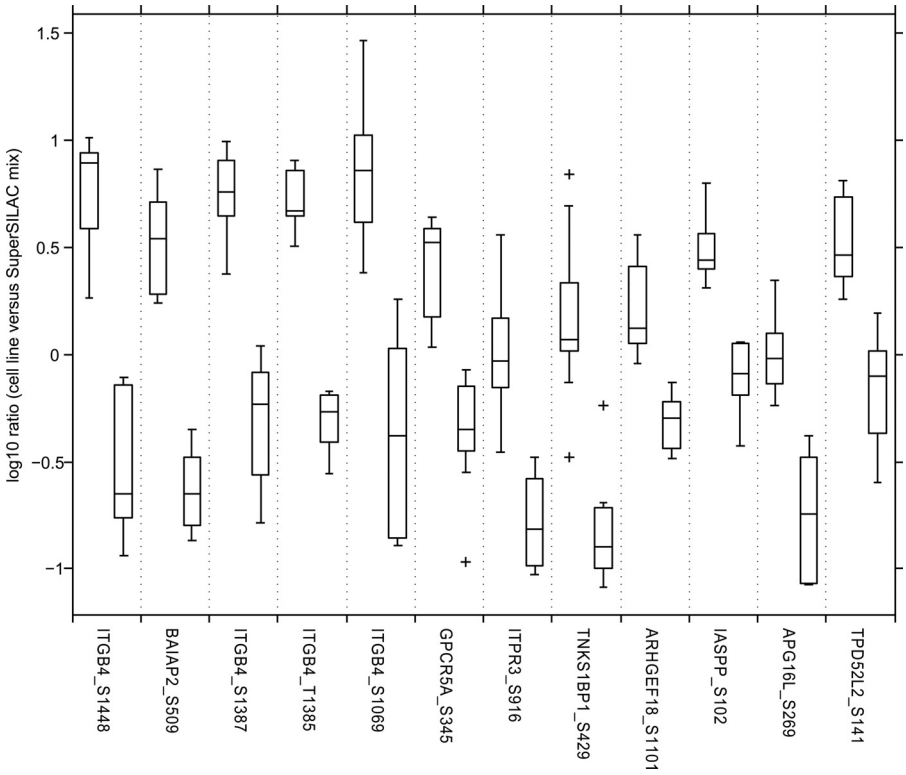

FIG. 3. **Final phosphosignature consisting of 12 phosphosites.** Each pair of boxes corresponds to one phosphosite. The *left box* represents the sensitive cell lines, and the *right box* represents the resistant cell lines. On each box, the *central mark* is the median, the edges of the box are the 25th and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and outliers are marked individually with *crosses*.

were also determined in-house and compared with the previously reported values (5). This time, all of the data were consistent (supplemental Table S1), and the six breast cancer cell lines were subjected to our global phosphoproteomics workflow (see supplemental Table S5 for data).

Subsequently, the cell lines were classified with the SVM predictor trained on the set of NCSLC cell lines. Strikingly, five of the six breast cancer cell lines could be classified correctly (Fig. 4*D*); only one resistant sample was wrongly predicted to be sensitive (MDA-MB-468). These findings indicate that the proposed phosphosignature is also predictive for dasatinib sensitivity in other cancer types.

Integrin β4 Expression Can Be Used as Surrogate Marker—Four of the highest ranked predictive phosphosites reside on the protein Integrin β4 (ITGB4; supplemental Table S2). Because we did enrich for phosphorylated peptides and did not measure the abundance of the nonphosphorylated peptides or the total protein, it is principally impossible to distinguish between differences in the phosphorylation degree and differences in the expression of the corresponding protein. However, in case of ITGB4, it is likely that the differences in the phosphorylation of the four sites are caused by differences in the abundance of the protein itself. To prove that the expression of this protein is indeed different in the two classes of the NSCLC cell lines, we performed quantitative Western blots using antibodies against the total protein of ITGB4 and 182-kDa tankyrase-1-binding protein (TNKS1BP1). We selected TNKS1BP1 as one of the eight proteins for which only one phosphosite was identified as predictive feature. While TNKS1BP1 is present in almost all cell lines and its expression shows no correlation with the sensitivity of the cell line to dasatinib, ITGB4 can be detected in eight sensitive cell lines but in only two resistant cell lines (Fig. 6*A*). This is confirmed by quantitative analysis of three replicate experiments (Fig. 6, *B* and *C*). The background-corrected signals of ITGB4 correlate with the phosphorylation degree measured by mass spectrometry (Pearson correlation 0.88, $p = 2 \times 10^{-6}$). The signals of most resistant cell lines are low, whereas strong signals can be determined in the sensitive cell lines. This clearly shows that expression of ITGB4 is also predictive and that it can be used as surrogate marker instead of its phosphorylation. Indeed, if choosing the average of the median signals in each group as classification threshold, all resistant and eight sensitive cell lines would be correctly classified, whereas three sensitive cell lines would be falsely classified as resistant. Nevertheless, the prediction accuracy of ITGB4 expression (84%) is not as high as the accuracy of the full phosphosignature (94%). In contrast, the signals for total TNKS1BP1 do not correlate with sensitivity, although its phosphorylation is predictive.

Integrin β4 Is Expressed in Subpopulation of Lung and Breast Cancer Tissues—We demonstrated that the signature consisting of 12 phosphorylation sites and the expression of ITGB4 is predictive in NSCLC and breast cancer cell lines. To explore whether ITGB4 is also expressed in cancer tissues, we examined immunohistochemistry images of several cancer tissue slices. The Human Protein Atlas (38) systematically analyses the human proteome in cell lines, normal tissues,

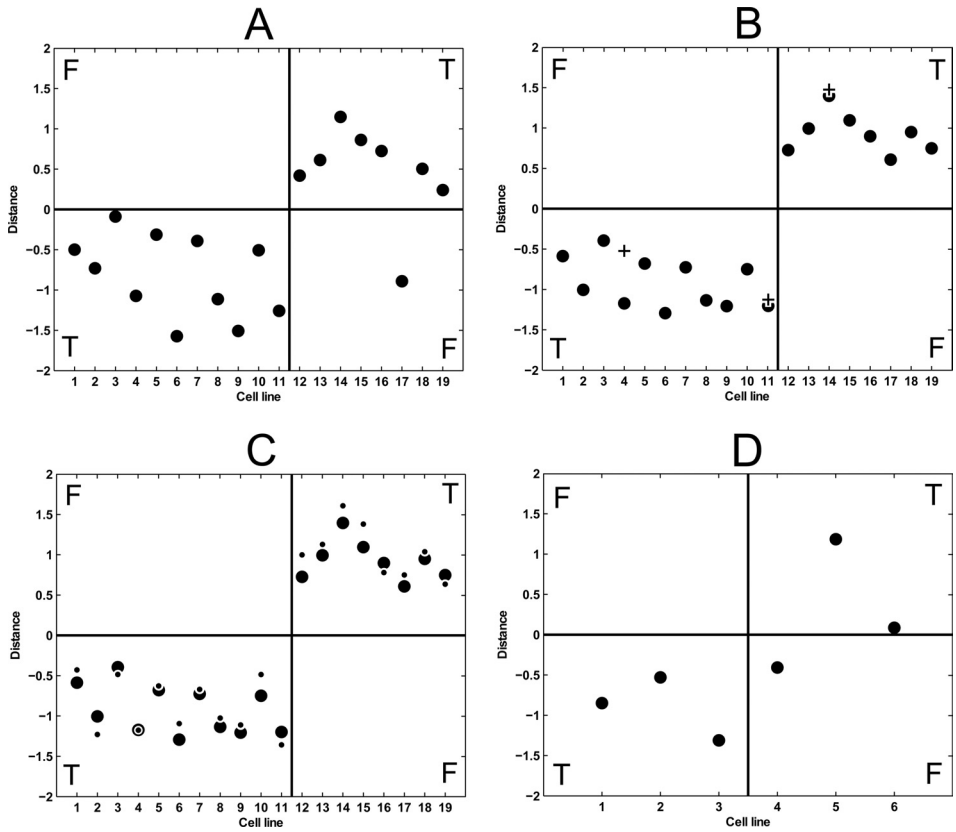

FIG. 4. **Classification results represented by distance to the separating hyperplanes of the respective SVMs.** The cell lines in *A*, *B*, and *C* are: *1*, LouNH91; *2*, H1648; *3*, HCC827; *4*, H322M; *5*, H2030; *6*, HCC2279; *7*, HCC366; *8*, HCC4006; *9*, H1666; *10*, PC9; *11*, H2009; *12*, H460; *13*, Calu6; *14*, H2077; *15*, H1395; *16*, H2172; *17*, HCC78; *18*, H157; and *19*, H520. The cell lines in *D* are: *1*, BT-20; *2*, MDA-MB-231; *3*, HCC1937; *4*, MDA-MB-468; *5*, BT-549; and *6*, MCF7. Sensitive cell lines (*left half*) are predicted correctly if they get assigned a negative value; resistant ones (*right half*) are correct if they are assigned a positive value. *A*, the results of the prediction quality assessment. *B*, prediction results of the final predictor when applied to the same date as used for training (*circles*) along with the results for the label switch experiments (*crosses*). *C*, prediction results of the final predictor when applied to the same data as used for training (*circles*), along with the results for the same data when normalized by the selected set of ribosomal proteins (*dots*). *D*, prediction results of the final predictor when applied to the breast cancer samples.

and cancer tissues using antibodies. In particular, it contains a number of immunohistochemistry images of cancer tissues stained with an antibody (CAB005258) against total protein of ITGB4. Five lung cancer samples (42%) are negative, whereas seven samples show weak to strong expression of ITGB4 (supplemental Fig. S7*A*). Similarly, six breast cancer samples (50%) are negative, whereas six samples show weak expression (supplemental Fig. S7*B*). In summary, we could show that the expression of ITGB4 can be used as surrogate marker for its phosphorylation. The marker is measurable by immunohistochemistry in clinical tissue samples, and it is present in a subpopulation of ~50% of the investigated cancer tissues.

DISCUSSION

This study shows that the identification of response prediction markers from global and unbiased quantitative phosphoproteomics experiments in a preclinical setting is possible. Detection of a few ten thousands of phosphorylation sites across a panel of cancer cell lines is feasible. The use of a pool of cell lines as a common reference enabled the accurate quantification of the detected sites. The accuracy and reproducibility of the phosphoproteomic workflow was demonstrated in label switch experiments. Measuring protein phosphorylation levels allowed us to monitor overactivation and repression of disease-specific signaling pathways. Because kinase inhibitors, such as small molecules and monoclonal antibodies interfere with signal transduction pathways, we hypothesized that determining the basal activity of these pathways will allow predicting a response to therapy with such an inhibitor.

We identified 58 phosphosites that are differentially abundant between sensitive and resistant cell lines. Enrichment analysis of gene ontology terms and KEGG pathways as well as subnetwork analysis shows that many of the differentially phosphorylated proteins are involved in cell adhesion and cytoskeleton organization, where most phosphorylations are higher in the sensitive group. Interestingly, it has been shown that dasatinib inhibits migration and invasion of various solid

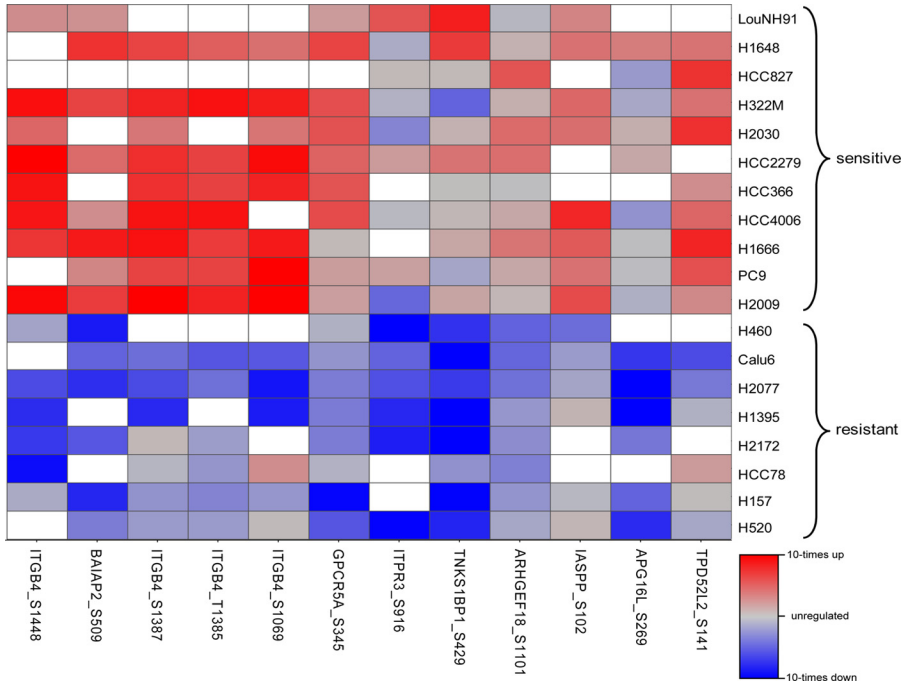

FIG. 5. **Heat map of the final 12 selected phosphorylation sites.** *Rows* are the 19 cell lines that were used to identify the phosphosignature (the upper 11 are sensitive, and the lower 8 are resistant), and *columns* are the phosphosites ordered by their importance ranks (*left* is the best). *Red* indicates up-regulation, *blue* indicates down-regulation, and *gray* indicates no regulation. Missing values are colored *white*.

tumors through inhibition of the Src kinase (39–41), which is one of the main targets of dasatinib (14, 15). We thus hypothesize that cells, in which pathways related to cell adhesion and cytoskeleton organization are overactivated, respond to a treatment with dasatinib. Src is a nonreceptor tyrosine-protein kinase. That none of the differentially phosphorylated residues is a tyrosine, does not contradict the hypothesis, because we studied the basal phosphoproteome of untreated cells. Proteins that are causal for resistance to Src inhibition may be located downstream or upstream of the direct Src kinase substrates in the signaling cascades.

We showed that a phosphosignature consisting of only 12 phosphorylation sites is sufficient to predict the response from the basal phosphoproteome of a cultured cell. The predictor model was based on a support vector machine with linear kernel. We validated the accuracy of the prediction in a leave-one-out cross-validation procedure. 18 of 19 cell lines could be classified correctly. The obtained prediction accuracy was 94%, and the area under the curve was 92%.

The 12 phosphorylation sites were located on 9 different proteins (see Table II and Fig. 5). Four of the phosphorylation sites are located on integrin $\beta 4$ (ITGB4 or CD104). In general, integrins mediate cell-matrix or cell-cell adhesion and are involved in transducing signals to regulate transcription and cell growth. The subunit $\beta 4$ associates with $\alpha 6$, and the resulting integrin $\alpha 6\beta 4$ is a receptor for the laminin family of extracellular matrix proteins. Integrin $\beta 4$ is linked to various signaling pathways such as the MAPK, phosphatidylinositol 3-kinase-Akt, and Src-Fak pathways (42–44). Furthermore, expression of $\alpha 6\beta 4$ is associated with poor patient prognosis in various cancers (45–47). According to the PhosphoSite database (48) the sites Ser^{1457} and Ser^{1518} were detected in previous mass spectrometry-based proteomics experiments, but to our knowledge the functions for none of the four sites have been described so far. All four sites are phosphorylated more strongly in sensitive cells than in resistant cells.

In addition to the integrin $\beta 4$ phosphorylations, the signature comprised eight additional phosphosites on eight other proteins. Like integrin $\beta 4$, the brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAP2) and the Rho guanine nucleotide exchange factor 18 (ARHGEF18) are involved in regulating the actin cytoskeleton. BAIAP2 (also called insulin receptor substrate p53, IRSp53) serves as an adaptor linking a Ras-related protein Rac1 with a Wiskott-Aldrich syndrome protein family member 2 (WAVE2). The recruitment of WAVE2 induces Cdc42 and the formation of filopodia (49, 50). ARGHEF18 acts as guanine nucleotide exchange factor for the GTPases RhoA and Rac1 (51, 52). Activation of RhoA induces actin stress fibers and cell rounding.

The RelA-associated inhibitor (PPP1R13L, also called inhibitor of ASPP protein, IASPP) and the G-protein-coupled receptor family C group 5 member A (GPRC5A, also called retinoic acid-induced protein 3, RAI3) are functionally connected to the tumor suppressor p53. PPP1R13L binds to p53 and inhibits its activation by ASPP1 and ASPP2 (53). On the other hand, p53 was demonstrated to bind to the promoter of GPRC5A and thereby negatively regulates its expression (54).

The tumor suppressor p53 is associated with at least two signature proteins. At the same time, p53 is inactivated by mutations in a large proportion of tumor cell lines. We therefore investigated whether p53 status alone is predictive of a response to dasatinib. According to the IARC TP53 database (55), six of seven sensitive and three of five nonsensitive cell lines have a mutation in the p53 protein (seven cell lines were

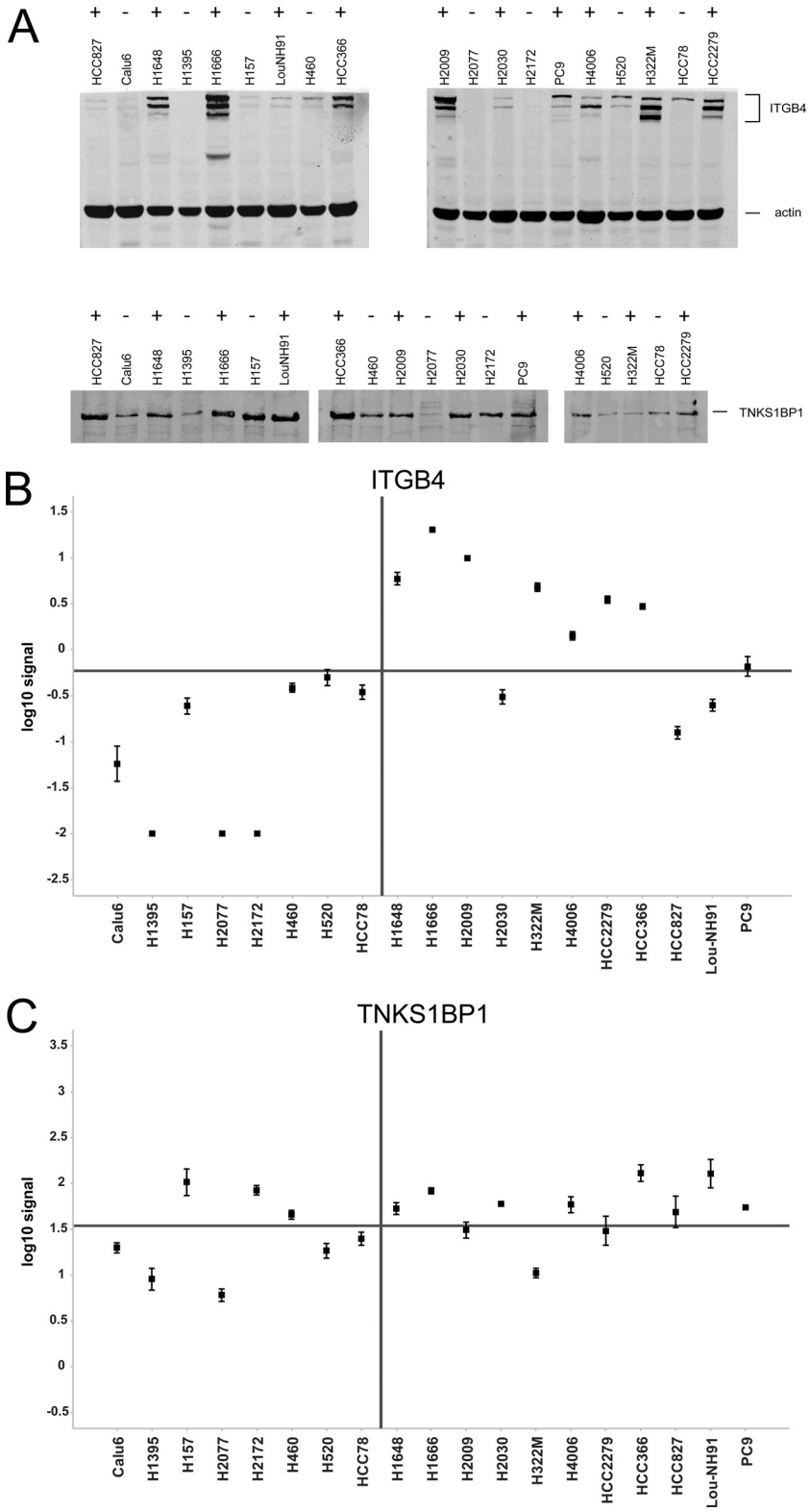

FIG. 6. **Western blots of ITGB 4 and TNKS1BP1 in NSCLC cell lines.** *A*, Western blot images for one replicate. The *top panel* shows Western blots for ITGB4, and the *bottom panel* shows Western blots for TNKS1BP1. The sensitivity to dasatinib treatment is noted by ± above the cell line labels. *B*, quantitative readout for ITGB4 in resistant (*left*) and sensitive (*right*) cell lines. The *error bars* represent the standard error across three replicates. The *horizontal line* represents the average of the class medians. *C*, quantitative readout for TNKS1BP1.

not listed; see also supplemental Table S1). Because the functional effect is not known for all mutations, we assumed that any mutation, apart from neutral or silent mutations, is functionally relevant. The null hypothesis that sensitivity to treatment with dasatinib does not differ between p53-mutated and p53 wild type cell lines cannot be rejected (Fisher's exact test p value is 0.52). Therefore, the mutation status of p53 is not a good predictor of dasatinib sensitivity.

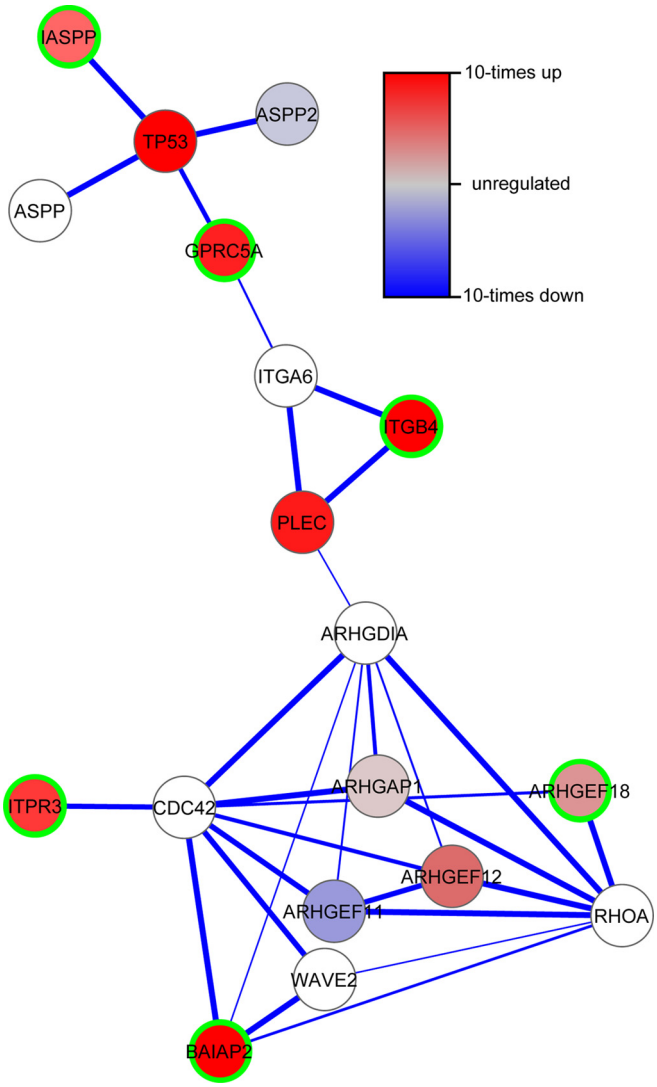

FIG. 7. **Protein-protein interaction network that shows the relationship between six of nine predictive signature proteins (marked with green border).** The network was obtained using STRING.

Although, based on the current literature, a direct link cannot be made between the other four proteins inositol 1,4,5-triphosphate receptor type 3 (ITPR3), 182-kDa tankyrase-1-binding protein (TNKS1BP1), autophagy-related protein 16–1 (APG16L), tumor protein D54 (TPD52L2), and the main dasatinib targets, the fact that their phosphorylation correlates with the treatment response supports their use in the predictive model.

From the discussion above it is clear that many of the signature proteins are related to each other. Indeed, when mapping the nine proteins to the STRING protein-protein interaction network (28), we revealed one network involving six signature proteins and few additional proteins (Fig. 7). Phosphorylation sites for most of the proteins in this network are less abundant in the resistant cell lines than in the sensitive cell lines.

The difference in phosphorylation of a specific site between two cell lines may be due to a difference in either expression of the corresponding protein, the degree of phosphorylation of this site, or a combination of both. The phosphoproteomic data do not allow distinguishing between the three possibilities. However, as long as the abundance of a certain phosphorylated peptide consistently differs between sensitive and resistant cell lines, the cause for its difference is not important for its use as a predictive biomarker. In case of ITGB4, we could indeed show that its protein expression is also predictive. Contrary, the protein expression of TNKS1BP1 does not differentiate between sensitive and resistant cell lines.

The study also showed that the predictor identified from a panel of NSCLC cell lines can be used in other cancer cell lines. Five of six breast cancer cell lines were correctly predicted (prediction accuracy 83%). Only one resistant cell line (MDA-MB-468) was predicted to be sensitive.

A few markers for dasatinib have been suggested in the literature or are already applied in the clinic. For example, Huang *et al.* (5) identified a predictive six-gene model from gene expression profiles. Obviously, the phosphorylation grade may be largely independent of the mRNA expression level. Nevertheless, we investigated whether the phosphorylation sites on the corresponding proteins are also predictive. We detected phosphorylation sites on five of the six proteins: EPHA2, CAV1, CAV2, ANXA1, and PTRF. Although the phosphorylation tends to be high in sensitive cell lines and low in resistant cell lines, the relationship is not as sound as for the markers identified in this study. All of the sites are not significantly different between the two classes. As an example, supplemental Fig. S6 shows three sites on the Ephrin type-A receptor (EPHA2). Additionally the tyrosine phosphorylations p-Src(Y418), p-BCR-ABL(Y412), p-Crkl(Y207), p-Pax(Y31), and p-Fak(Y576) have been described as pharmacodynamic markers for dasatinib in mouse experiments and in clinical trials (57–59). These markers are modulated after treatment with dasatinib, and their basal levels do not necessarily differentiate between sensitive and resistant subjects. Nevertheless, we were interested in their behavior across the untreated cell lines. We could detect the phosphorylation site Tyr^{418} of Src in five cell lines but could not identify any relationship to the sensitivity of these cell lines. The site ABL(Y412) on the fusion protein BCR-ABL was not detected. However, a different site BCR(S459) was detected in almost all cell lines and is significantly modulated between the sensitive and resistant group (supplemental Table S1 and Fig. S6).

We demonstrated our method for the identification of a predictive phosphosignature in a set of NSCLC and breast cancer cell lines. The application to cultured cells has a number of advantages: the cell population is very homogenous; sample amounts from cell lines are not limited; experiments are easily reproducible; and the drug's efficacy can be experimentally determined. However, whether the signature or parts of the signature are also predictive in clinical samples

has to be shown in future studies with clinical samples. Instead of applying shotgun phosphoproteomics, it is possible to apply targeted detection methods, such as immunological methods or the mass spectrometry-based multiple reaction monitoring method (60). These methods allow the quantification of marker phosphosites of relatively low sample amounts and can be applied to large number of samples. Because fresh frozen tissues are rare, the translation of our results to the clinic requires the analysis of formalin-fixed and paraffin-embedded tissues. It has been assumed that the cross-linking of proteins prevents a proteomic analysis. Recently, it could be shown that proteins can be effectively extracted from formalin-fixed and paraffin-embedded samples and that the proteins and phosphorylations are quantitatively preserved compared with fresh frozen tissues (56, 61, 62).

As an alternative, we demonstrated that the expression of ITGB4 can be used as surrogate marker for its phosphorylation. The marker is measurable by immunohistochemistry in clinical tissue samples, and it is present in a subpopulation of ~50% of the investigated cancer tissues.

In this study, the phosphorylation data were globally normalized, assuming that the overall phosphoproteome is fairly well conserved between the different cell lines. However, this strategy is no longer applicable to targeted detection of the selected phosphosites, because all measured phosphosites will be regulated. We proposed an alternative normalization strategy using the expression of eight nonregulated ribosomal proteins. It could be demonstrated that the prediction of sensitivity using the phosphosignature is stable for the application of the alternative normalization strategy.

In summary, the identified phosphosignature consisting of 12 phosphorylation sites is highly predictive for the sensitivity to treatment with dasatinib in NSCLC cell lines as well as breast cancer cell lines. The results suggest that the phosphorylations of integrin β4 as well as eight further proteins are candidate biomarkers for predicting response in solid tumors to dasatinib and potentially to other Src family kinase inhibitors. That many of the signature proteins have related function and are connected in a protein-protein interaction network further supports the generalizability of the predictive signature.

In this study we proposed a general method for identifying response prediction biomarkers based on a phosphorylation signature. The method is hypothesis-free insofar as the investigated phosphorylation sites do not have to be preselected, and no assumptions about the mechanism of action of the therapeutic drug have to be made. The basis of the method is the global quantitative phosphoproteomic analysis of base-line samples. Although we demonstrated that the method permits identifying a highly predictive phosphorylation signature for response to dasatinib treatment in NSCLC cell lines, it can be assumed that the method can also be applied to other drugs, particularly other kinase inhibitors, and to other tumor types.

Acknowledgments—We are grateful to colleagues at Evotec for useful discussions and excellent technical assistance. We thank Nicole Jordan for tremendous support in the cell culture, as well as Roman Thomas and his group at the Max Planck Institute for Neurological Research for providing cell lines.

* This work is based on Project 0315011 supported by the Federal German Ministry of Education and Research.

S This article contains supplemental material.

¶ To whom correspondence should be addressed: Evotec München GmbH, Am Klopferspitz 19a, 82152 Martinsried, Germany. Tel.: 49-89-45-244-65-23; Fax: 49-89-45-244-65-20; E-mail: christoph.schaab@evotec.com.

REFERENCES

1. Katzel, J. A., Fanucchi, M. P., and Li, Z. (2009) Recent advances of novel targeted therapy in non-small cell lung cancer. *J. Hematol. Oncol.* **2,** 2
2. Reichert, J. M., and Valge-Archer, V. E. (2007) Development trends for monoclonal antibody cancer therapeutics. *Nat. Rev. Drug Discov.* **6,** 349–356
3. Ross, J. S., and Fletcher, J. A. (1999) HER-2/neu (c-erb-B2) gene and protein in breast cancer. *Am. J. Clin. Pathol.* **112,** S53–S67
4. Cobleigh, M. A., Vogel, C. L., Tripathy, D., Robert, N. J., Scholl, S., Fehrenbacher, L., Wolter, J. M., Paton, V., Shak, S., Lieberman, G., and Slamon, D. J. (1999) Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. *J. Clin. Oncol.* **17,** 2639–2648
5. Huang, F., Reeves, K., Han, X., Fairchild, C., Platero, S., Wong, T. W., Lee, F., Shaw, P., and Clark, E. (2007) Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: Rationale for patient selection. *Cancer Res.* **67,** 2226–2238
6. Dressman, H. K., Berchuck, A., Chan, G., Zhai, J., Bild, A., Sayer, R., Cragun, J., Clarke, J., Whitaker, R. S., Li, L., Gray, J., Marks, J., Ginsburg, G. S., Potti, A., West, M., Nevins, J. R., and Lancaster, J. M. (2007) An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J. Clin. Oncol.* **25,** 517–525
7. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648
8. Macek, B., Mann, M., and Olsen, J. V. (2009) Global and site-specific quantitative phosphoproteomics: Principles and applications. *Annu. Rev. Pharmacol. Toxicol.* **49,** 199–221
9. Schaab, C. (2011) Analysis of phosphoproteomics data. *Methods Mol. Biol.* **696,** 41–57
10. Blume-Jensen, P., and Hunter, T. (2001) Oncogenic kinase signalling. *Nature* **411,** 355–365
11. Kaminska, B. (2005) MAPK signalling pathways as molecular targets for anti-inflammatory therapy: From molecular mechanisms to therapeutic benefits. *Biochim. Biophys. Acta* **1754,** 253–262
12. Ferlay, J., Parkin, D. M., and Steliarova-Foucher, E. (2010) Estimates of cancer incidence and mortality in Europe in 2008. *Eur. J. Cancer* **46,** 765–781
13. Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., and Thun, M. J. (2008) Cancer statistics, 2008. *CA Cancer J. Clin.* **58,** 71–96
14. Sharma, K., Weber, C., Bairlein, M., Greff, Z., Kéri, G., Cox, J., Olsen, J. V., and Daub, H. (2009) Proteomics strategy for quantitative protein interaction profiling in cell extracts. *Nat. Methods* **6,** 741–744
15. Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C., Reader, V., Sweetman, G., Bauer, A., Bouwmeester, T., Hopf, C., Kruse, U., Neubauer, G., Ramsden, N., Rick, J., Kuster, B., and Drewes, G. (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **25,** 1035–1044
16. Johnson, F. M., Bekele, B. N., Feng, L., Wistuba, I., Tang, X. M., Tran, H. T., Erasmus, J. J., Hwang, L. L., Takebe, N., Blumenschein, G. R., Lippman, S. M., and Stewart, D. J. (2010) Phase II study of dasatinib in patients with advanced non-small-cell lung cancer. *J. Clin. Oncol.* **28,** 4609–4615

17. Andersen, J. N., Sathyanarayanan, S., Di Bacco, A., Chi, A., Zhang, T., Chen, A. H., Dolinski, B., Kraus, M., Roberts, B., Arthur, W., Klinghoffer, R. A., Gargano, D., Li, L., Feldman, I., Lynch, B., Rush, J., Hendrickson, R. C., Blume-Jensen, P., and Paweletz, C. P. (2010) Pathway-based identification of biomarkers for targeted therapeutics: Personalized oncology with PI3K pathway inhibitors. *Sci. Transl. Med.* **2,** 43ra55
18. Sos, M. L., Michel, K., Zander, T., Weiss, J., Frommolt, P., Peifer, M., Li, D., Ullrich, R., Koker, M., Fischer, F., Shimamura, T., Rauh, D., Mermel, C., Fischer, S., Stückrath, I., Heynck, S., Beroukhim, R., Lin, W., Winckler, W., Shah, K., LaFramboise, T., Moriarty, W. F., Hanna, M., Tolosi, L., Rahnenführer, J., Verhaak, R., Chiang, D., Getz, G., Hellmich, M., Wolf, J., Girard, L., Peyton, M., Weir, B. A., Chen, T. H., Greulich, H., Barretina, J., Shapiro, G. I., Garraway, L. A., Gazdar, A. F., Minna, J. D., Meyerson, M., Wong, K. K., and Thomas, R. K. (2009) Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *J. Clin. Invest.* **119,** 1727–1740
19. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
20. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
21. Schaab, C., Geiger, T., Stoehr, G., Cox, J., and Mann, M. (2012) Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell. Proteomics* **11**, 10.1074/mcp.M111.014068
22. Benjamini Y., H., Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* **B 57,** 289–300
23. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., and Consortium, G. O. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32,** D258–D261
24. Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30
25. Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., and Dopazo, J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics* **8,** 114
26. Klammer, M., Godl, K., Tebbe, A., and Schaab, C. Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC Bioinformatics* **11,** 351
27. Zhou, Y., Cras-Méneur, C., Ohsugi, M., Stormo, G. D., and Permutt, M. A. (2007) A global approach to identify differentially expressed genes in cDNA (two-color) microarray experiments. *Bioinformatics* **23,** 2073–2079
28. Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009) STRING 8: A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37,** D412–D416
29. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **26,** 392–398
30. Schölkopf, B., and Smola, A. J. (2002) *Learning with Kernels*, The MIT Press, Cambridge, MA
31. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7,** 383–385
32. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 15149–15154
33. Hutter, B., Schaab, C., Albrecht, S., Borgmann, M., Brunner, N. A., Freiberg, C., Ziegelbauer, K., Rock, C. O., Ivanov, I., and Loferer, H. (2004) Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. *Antimicrob. Agents Chemother.* **48,** 2838–2844
34. Thuerigen, O., Schneeweiss, A., Toedt, G., Warnat, P., Hahn, M., Kramer, H., Brors, B., Rudlowski, C., Benner, A., Schuetz, F., Tews, B., Eils, R., Sinn, H. P., Sohn, C., and Lichter, P. (2006) Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer. *J. Clin. Oncol.* **24,** 1839–1845
35. Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005) Prediction error estimation: A comparison of resampling methods. *Bioinformatics* **21,** 3301–3307
36. Lange, V., Picotti, P., Domon, B., and Aebersold, R. (2008) Selected reaction monitoring for quantitative proteomics: A tutorial. *Mol. Syst. Biol.* **4,** 222
37. Hüttenhain, R., Malmström, J., Picotti, P., and Aebersold, R. (2009) Perspectives of targeted mass spectrometry for protein biomarker verification. *Curr. Opin. Chem. Biol.* **13,** 518–525
38. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28,** 1248–1250
39. Buettner, R., Mesa, T., Vultur, A., Lee, F., and Jove, R. (2008) Inhibition of Src family kinases with dasatinib blocks migration and invasion of human melanoma cells. *Mol. Cancer Res.* **6,** 1766–1774
40. Shor, A. C., Keschman, E. A., Lee, F. Y., Muro-Cacho, C., Letson, G. D., Trent, J. C., Pledger, W. J., and Jove, R. (2007) Dasatinib inhibits migration and invasion in diverse human sarcoma cell lines and induces apoptosis in bone sarcoma cells dependent on SRC kinase for survival. *Cancer Res.* **67,** 2800–2808
41. Johnson, F. M., Saigal, B., Talpaz, M., and Donato, N. J. (2005) Dasatinib (BMS-354825) tyrosine kinase inhibitor suppresses invasion and induces cell cycle arrest and apoptosis of head and neck squamous cell carcinoma and non-small cell lung cancer cells. *Clin. Cancer Res.* **11,** 6924–6932
42. Dans, M., Gagnoux-Palacios, L., Blaikie, P., Klein, S., Mariotti, A., and Giancotti, F. G. (2001) Tyrosine phosphorylation of the $\beta 4$ integrin cytoplasmic domain mediates Shc signaling to extracellular signal-regulated kinase and antagonizes formation of hemidesmosomes. *J. Biol. Chem.* **276,** 1494–1502
43. Chung, J., Bachelder, R. E., Lipscomb, E. A., Shaw, L. M., and Mercurio, A. M. (2002) Integrin (alpha 6 beta 4) regulation of eIF-4E activity and VEGF translation: A survival mechanism for carcinoma cells. *J. Cell Biol.* **158,** 165–174
44. Dutta, U., and Shaw, L. M. (2008) A key tyrosine (Y1494) in the $\beta 4$ integrin regulates multiple signaling pathways important for tumor development and progression. *Cancer Res.* **68,** 8779–8787
45. Tagliabue, E., Ghirelli, C., Squicciarini, P., Aiello, P., Colnaghi, M. I., and Ménard, S. (1998) Prognostic value of $\alpha 6\beta 4$ integrin expression in breast carcinomas is affected by laminin production from tumor cells. *Clin. Cancer Res.* **4,** 407–410
46. Lu, S., Simin, K., Khan, A., and Mercurio, A. M. (2008) Analysis of integrin $\beta 4$ expression in human breast cancer: Association with basal-like tumors and prognostic significance. *Clin. Cancer Res.* **14,** 1050–1058
47. Van Waes, C., Kozarsky, K. F., Warren, A. B., Kidd, L., Paugh, D., Liebert, M., and Carey, T. E. (1991) The A9 antigen associated with aggressive human squamous carcinoma is structurally and functionally similar to the newly defined integrin $\alpha 6\beta 4$. *Cancer Res.* **51,** 2395–2402
48. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4,** 1551–1561
49. Miki, H., Yamaguchi, H., Suetsugu, S., and Takenawa, T. (2000) IRSp53 is an essential intermediate between Rac and WAVE in the regulation of membrane ruffling. *Nature* **408,** 732–735
50. Yamagishi, A., Masuda, M., Ohki, T., Onishi, H., and Mochizuki, N. (2004) A novel actin bundling/filopodium-forming domain conserved in insulin receptor tyrosine kinase substrate p53 and missing in metastasis protein. *J. Biol. Chem.* **279,** 14929–14936
51. Blomquist, A., Schwörer, G., Schablowski, H., Psoma, A., Lehnen, M., Jakobs, K. H., and Rümenapp, U. (2000) Identification and characterization of a novel Rho-specific guanine nucleotide exchange factor.

Biochem. J. **352,** 319–325
52. Niu, J., Profirovic, J., Pan, H., Vaiskunaite, R., and Voyno-Yasenetskaya, T. (2003) G Protein betagamma subunits stimulate p114RhoGEF, a guanine nucleotide exchange factor for RhoA and Rac1: Regulation of cell shape and reactive oxygen species production. *Circ. Res.* **93,** 848–856
53. Bergamaschi, D., Samuels, Y., Sullivan, A., Zvelebil, M., Breyssens, H., Bisso, A., Del Sal, G., Syed, N., Smith, P., Gasco, M., Crook, T., and Lu, X. (2006) iASPP preferentially binds p53 proline-rich region and modulates apoptotic function of codon 72-polymorphic p53. *Nat. Genet.* **38,** 1133–1141
54. Wu, Q., Ding, W., Mirza, A., Van Arsdale, T., Wei, I., Bishop, W. R., Basso, A., McClanahan, T., Luo, L., Kirschmeier, P., Gustafson, E., Hernandez, M., and Liu, S. (2005) Integrative genomics revealed RAI3 is a cell growth-promoting gene and a novel P53 transcriptional target. *J. Biol. Chem.* **280,** 12935–12943
55. Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S. V., Hainaut, P., and Olivier, M. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. *Hum. Mutat.* **28,** 622–629
56. Berg, D., Malinowsky, K., Reischauer, B., Wolff, C., and Becker, K. F. (2011) Use of formalin-fixed and paraffin-embedded tissues for diagnosis and therapy in routine clinical settings. *Methods Mol. Biol.* **785,** 109–122
57. Luo, F. R., Yang, Z., Camuso, A., Smykla, R., McGlinchey, K., Fager, K., Flefleh, C., Castaneda, S., Inigo, I., Kan, D., Wen, M. L., Kramer, R., Blackwood-Chirchir, A., and Lee, F. Y. (2006) Dasatinib (BMS-354825) pharmacokinetics and pharmacodynamic biomarkers in animal models predict optimal clinical exposure. *Clin. Cancer Res.* **12,** 7180–7186
58. Luo, F. R., Barrett, Y. C., Yang, Z., Camuso, A., McGlinchey, K., Wen, M. L., Smykla, R., Fager, K., Wild, R., Palme, H., Galbraith, S., Blackwood-Chirchir, A., and Lee, F. Y. (2008) Identification and validation of phospho-SRC, a novel and potential pharmacodynamic biomarker for dasatinib (SPRYCEL), a multi-targeted kinase inhibitor. *Cancer Chemother. Pharmacol.* **62,** 1065–1074
59. Herold, C. I., Chadaram, V., Peterson, B. L., Marcom, P. K., Hopkins, J., Kimmick, G. G., Favaro, J., Hamilton, E., Welch, R. A., Bacus, S., and Blackwell, K. L. (2011) Phase II trial of dasatinib in patients with metastatic breast cancer using real-time pharmacodynamic tissue biomarkers of Src inhibition to escalate dosing. *Clin. Cancer Res.* **17,** 6061–6070
60. Kitteringham, N. R., Jenkins, R. E., Lane, C. S., Elliott, V. L., and Park, B. K. (2009) Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **877,** 1229–1239
61. Ostasiewicz, P., Zielinska, D. F., Mann, M., and Wiśniewski, J. R. (2010) Proteome, phosphoproteome, and *N*-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. *J. Proteome Res.* **9,** 3688–3700
62. Gámez-Pozo, A., Sánchez-Navarro, I., Calvo, E., Díaz, E., Miguel-Martín, M., López, R., Agulló, T., Camafeita, E., Espinosa, E., López, J. A., Nistal, M., and Vara, J. Á. (2011) Protein phosphorylation analysis in archival clinical cancer samples by shotgun and targeted proteomics approaches. *Mol. Biosyst.* **7,** 2368–2374