Cell Reports
Resource

Cell PRESS
Open ACCESS

Global Proteome Analysis of the NCI-60 Cell Line Panel

Amin Moghaddas Gholami,[1,4,*] Hannes Hahne,[1,4] Zhixiang Wu,[1,4] Florian Johann Auer,[1] Chen Meng,[1] Mathias Wilhelm,[1] and Bernhard Kuster[1,2,3,*]
[1]Proteomics and Bioanalytics, Technische Universität München, Emil-Erlenmeyer-Forum 5, 85354 Freising, Germany
[2]Center for Integrated Protein Science Munich, Department of Chemistry and Biochemistry, Butenandtstr. 5–13, 81377 Munich, Germany
[3]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[4]These authors contributed equally to this work
*Correspondence: amin@tum.de (A.M.G.), kuster@tum.de (B.K.)
http://dx.doi.org/10.1016/j.celrep.2013.07.018
This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

SUMMARY

The NCI-60 cell line collection is a very widely used panel for the study of cellular mechanisms of cancer in general and in vitro drug action in particular. It is a model system for the tissue types and genetic diversity of human cancers and has been extensively molecularly characterized. Here, we present a quantitative proteome and kinome profile of the NCI-60 panel covering, in total, 10,350 proteins (including 375 protein kinases) and including a core cancer proteome of 5,578 proteins that were consistently quantified across all tissue types. Bioinformatic analysis revealed strong cell line clusters according to tissue type and disclosed hundreds of differentially regulated proteins representing potential biomarkers for numerous tumor properties. Integration with public transcriptome data showed considerable similarity between mRNA and protein expression. Modeling of proteome and drug-response profiles for 108 FDA-approved drugs identified known and potential protein markers for drug sensitivity and resistance. To enable community access to this unique resource, we incorporated it into a public database for comparative and integrative analysis (http://wzw.tum.de/proteomics/nci60).

INTRODUCTION

Cell lines derived from human tumors are very widely used model systems for the study of cancer biology and drug discovery. Particularly in combination with system-level explorative profiling technologies, the comparative analysis of cancer cell lines can reveal distinct similarities and differences in biological processes between cancer cells that can be exploited in many different ways. For instance, the 59 cancer cell lines (NCI-60 panel) of the National Cancer Institute's (NCI's) Developmental Therapeutics Program (DTP; http://dtp.nci.nih.gov) are an established tool for in vitro drug screening. The collection represents, at least to some extent, the tissue type and genetic diversity of human cancers (Shoemaker, 2006). Since its inception, the NCI-60 panel has led to many important discoveries, including a general advance in the understanding of cancer mechanisms (Boyd and Paull, 1995; Weinstein, 2006), the identification of mechanisms of action of drugs, and the approval of new chemotherapeutic agents (e.g., bortezomib). Hundreds of thousands of potential anticancer agents have by now been screened using the NCI-60 panel (Holbeck et al., 2010; Shoemaker, 2006), and multiple technology platforms have been used to characterize the cells on the molecular level including, but not limited to, array comparative genomic hybridization (Bussey et al., 2006), karyotype analysis (Roschke et al., 2003), DNA mutational analysis (Abaan et al., 2013; Ikediobi et al., 2006), DNA fingerprinting (Lorenzi et al., 2009), microarrays for transcript expression (Scherf et al., 2000; Shankavaram et al., 2007), microarrays for microRNA expression (Blower et al., 2008; Liu et al., 2010), single-nucleotide polymorphism arrays to identify DNA copy number alterations (Garraway et al., 2005), and DNA methylation (Ehrich et al., 2008). Although proteins carry out virtually all cellular processes and represent the vast majority of anticancer drug targets, very few studies have focused on the analysis of protein expression across the NCI-60 panel (Nishizuka et al., 2003; Park et al., 2010; Shankavaram et al., 2007). In particular, reverse-phase protein microarrays from cellular lysates have been employed in this context, and although these studies focused on a rather confined number of proteins, their results highlight the potential of systematic protein expression analyses for cancer research in general and drug discovery in particular. Mass spectrometry (MS)-based proteomics has undergone rapid progress in past years (Aebersold and Mann, 2003; Mallick and Kuster, 2010), and systematic analyses can now be carried out to identify and quantify the majority of proteins expressed in a human cell line (Beck et al., 2011; Burkard et al., 2011; Geiger et al., 2012; Lundberg et al., 2010; Nagaraj et al., 2011). In addition, important oncogene classes such as kinases, which are often of low cellular abundance, can now be systematically queried using MS-based proteomics (Bantscheff et al., 2007; Wu et al., 2011). Protein kinases are key players of intracellular signal transduction, and dysregulation of protein kinases can be cause or consequence of cancer and therefore are among the most important anticancer drug targets today (Cohen,

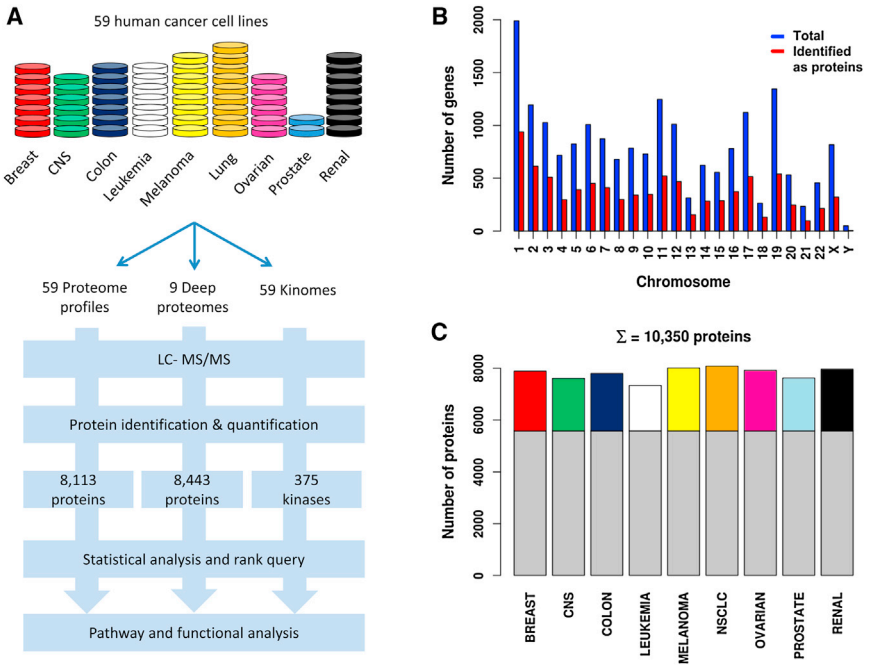

Figure 1. Proteomic Analysis of the NCI-60 Cell Line Panel

(A) Experimental strategy.

(B) Coverage of the human genome by chromosomes. Distributions of the identified genes are shown in red versus total genes (blue) for each chromosome.

(C) Core cancer proteome and contributions of different tissue groups.

See also Figure S1.

2002; Knapp et al., 2013). More generally speaking, proteomic analyses are now valuable tools for molecular and clinical cancer research (Hanash and Taguchi, 2010) and drug discovery (Schirle et al., 2012) and will be more and more used in concert with DNA/RNA-level investigations in the future.

In the present study, we employed a comprehensive analysis using MS-based proteomics to obtain quantitative proteome profiles of all 59 cell lines of the NCI-60 panel. Collectively, 10,350 proteins, including 375 protein kinases, were identified and 6,003 proteins were consistently quantified in at least 5 out of 59 cell lines. Unsupervised bioinformatic analysis revealed strong cell line clusters according to tissue type, and further analyses disclosed 522 differentially regulated proteins and several pathways predominant in certain tissue types. Integration with public transcriptome data showed a significant degree of correlation between messenger RNA (mRNA) and protein expression, and the integration with profiles of 108 drugs approved by the US Food and Drug Administration (FDA) revealed known and potentially novel protein markers involved in mediating drug resistance and sensitivity. To enable access to this unique resource, we incorporated the proteomics data including the peptide fragment spectra into a database for comparative and integrative analysis. These data can, for example, be used to obtain reference expression profiles for proteins of interest both within and across experiments and cell lines.

RESULTS AND DISCUSSION

Proteomic Analysis of the NCI-60 Panel Identifies the Core Cancer Proteome

The NCI-60 panel comprises 59 individual cancer cell lines derived from nine different tissues (brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate, and skin), which we analyzed using three proteomic approaches (Figures 1A and S1A–S1C). The proteome profiles of all individual cell lines followed a conventional one-dimensional PAGE followed by in-gel digestion and liquid chromatography-tandem mass spectrometry (GeLC-MS/MS) approach (Schirle et al., 2003) and yielded 8,113 proteins (Figure S1D). To increase tissue-specific proteome coverage, one representative cell line from each of the nine tissue groups was analyzed in more depth and resulted in the identification of 8,443 proteins (deep proteomes; Figures S1A and S1E). Kinase profiles were obtained for the complete NCI-60 panel using immobilized nonselective kinase inhibitors (kinobeads) followed by MS-based protein identification (Bantscheff et al., 2007; Figures S1B, S1D, and S2A). The kinobead approach resulted in the identification of 220 protein kinases of which 106 were not identified in the other two approaches, thus accessing a part of the proteome of too low abundance for conventional shotgun proteomics. In addition, 155 protein kinases, which have low or no affinity for kinobeads (Bantscheff et al., 2007; Lemeer et al., 2013), were exclusively identified in the proteome profiling experiments.

Collectively, the three data sets comprise 10,350 distinct proteins corresponding to 8,739 unique genes and representing 46% of the protein-coding human genome (Figure 1B). Protein identifications are spread evenly across autosomes with an average coverage of 44%. Interestingly, while a similar coverage was also obtained for the X chromosome, we found only five proteins encoded by Y-chromosomal genes (12% coverage) in cell lines of male origin, which is consistent with previous findings (Geiger et al., 2012).

The large number of cell lines investigated allowed us to analyze expression profiles of proteins systematically across cell lines. To identify proteins ubiquitously expressed in cancer cells of different origins, we reconstructed the common proteome for all nine tissue groups. A protein was considered as part of the core proteome if it was identified in at least one cell line of every single tissue group. The resulting 5,578 proteins can thus be regarded as the core cancer proteome. The remaining ~5,000 proteins show a more distinct expression pattern between tissues and each tissue contributing ~2,000 proteins not contained in the core proteome (Figure 1C).

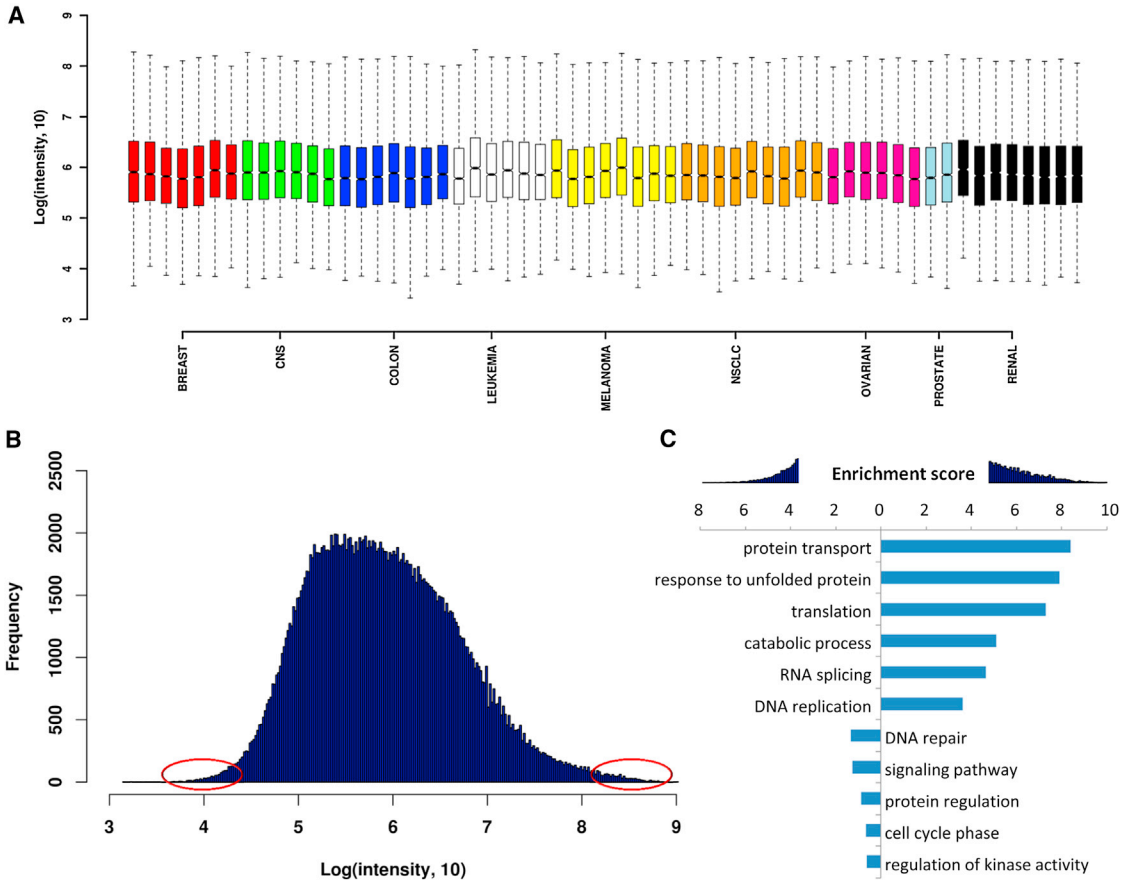

Figure 2. Proteome-wide Label-free Quantification of the NCI-60 Cell Line Panel

(A) Distribution of protein intensity from proteome profiling experiments of all 59 cell lines. Whiskers represent the most extreme data point.

(B) Distribution of the median logarithmic protein intensity of all proteins identified in the proteome profiling experiments.

(C) Gene ontology enrichment analysis of the most and least abundant proteins. Enrichment score of biological functions and cellular components was measured by modified Fisher's exact test (Hosack et al., 2003).

See also Figures S2, S3, and S4.

Proteomics Detects Proteins between 100 and 10 Million Copies per Cell

We used intensity-based label-free quantification for the relative quantification of proteins across cell lines in which the summed protein intensity from proteome profiling experiments (not deep proteomes or kinomes) served as a proxy for protein abundance within a cell line (Luber et al., 2010). Protein abundance distributions of all 59 cell lines are generally very similar and span at least five orders of magnitude (Figures 2A and S2B), which is consistent with recent estimates of protein copy numbers in mammalian cell lines (Beck et al., 2011; Geiger et al., 2012; Nagaraj et al., 2011; Schwanhäusser et al., 2011).

To identify proteins that appear uniformly among the most and least abundant proteins, we calculated median abundance values across all cell lines (Figures 2B, S2B, and S2C) and estimated protein copy numbers. Assuming a median number of ~10,000 copies per protein (Beck et al., 2011) and a linear correlation between measured protein abundance and copy number (Beck et al., 2011; Malmström et al., 2009), we estimate that the identified proteins have copy numbers between 100 and 10,000,000 copies per cell. The 10% most highly expressed proteins contain mostly structural proteins and proteins involved in basic cellular machineries that are known to be much more abundant than regulatory proteins (Figures 2C and S3A–S3D). For instance, proteins involved in transport processes, as classified by Gene Ontology (GO) annotation (Ashburner et al., 2000), formed a tight cluster at the top end of the distribution of protein expression levels. Similarly, proteins with roles in protein folding, molecular transport, and translation are significantly enriched among the most abundant proteins. Conversely, among the 10% least abundant proteins are mainly regulatory proteins, membrane proteins, and a large proportion of as-yet-uncharacterized proteins. The correlation of protein abundance of all NCI-60 cell lines was relatively high (average Pearson correlation of R = 0.81; Figure S3E). Despite this strong correlation, there are clearly also proteins that differ by orders of magnitude in expression between cell lines (Figure S3F). Taken together, these observations are consistent with previous comparative studies of multiple cell lines (Geiger et al., 2012) as well as with studies of single cell lines (Beck et al., 2011; Nagaraj et al., 2011).

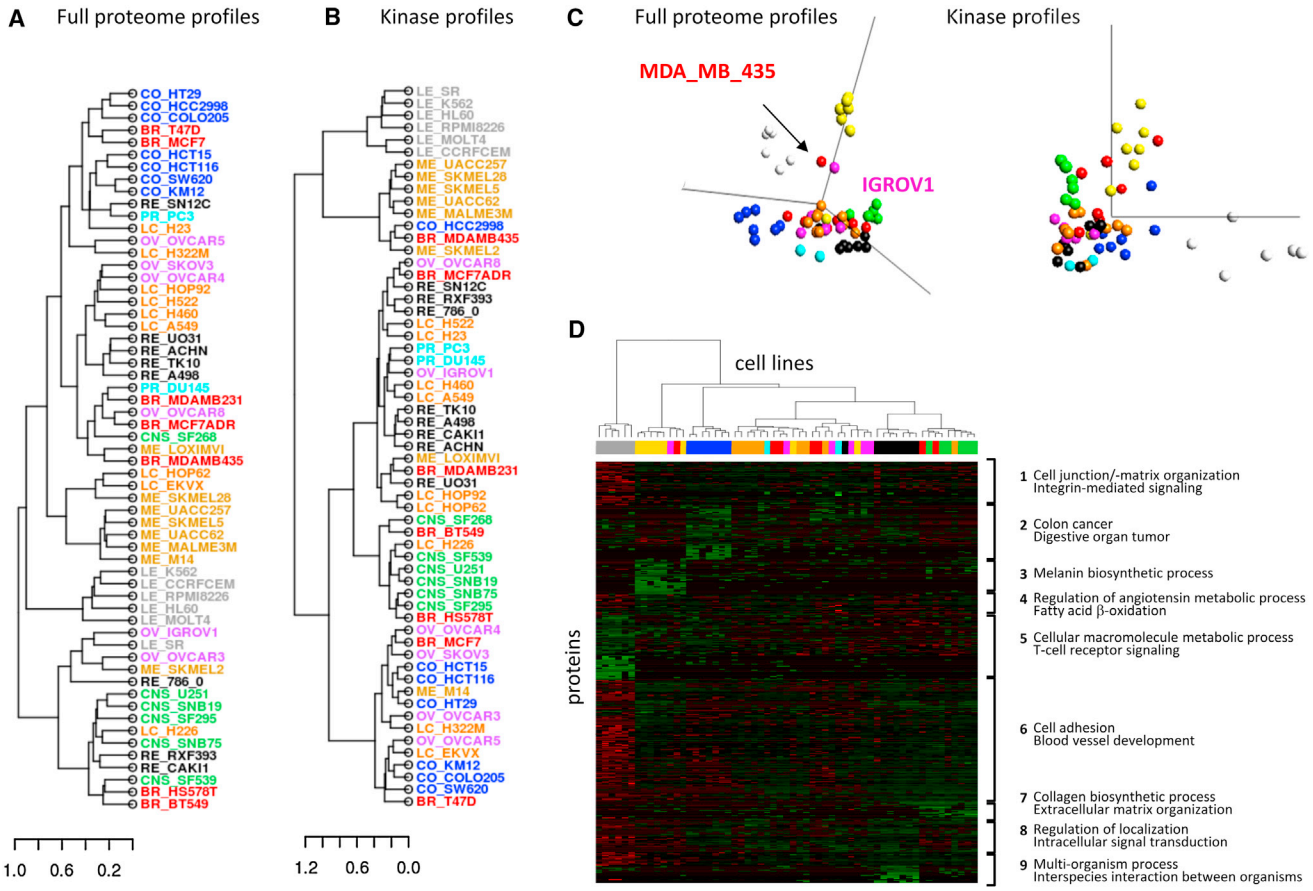

Figure 3. Hierarchical Clustering and PCA Analyses of the Proteome and Kinase Profiling Experiments

(A and B) Unsupervised hierarchical clustering of cell lines based on proteome profiles and kinase profiles. Dendrograms show average linkage hierarchical clustering using Spearman rank correlation with Ward metric.

(C) Cell line PCA based on 473 and 49 differentially expressed proteins and protein kinases, respectively.

(D) Hierarchical clustering of all cell lines and differentially expressed proteins including kinases. Top biological functions and pathways enriched in defined clusters are indicated. Cell lines are colored as in Figure 1.

See also Figure S4.

Clustering of Proteomes Highlights Common and Distinct Molecular Signatures of Cancer Cells

We next examined the similarity of individual cell line proteomes using unsupervised hierarchical clustering using 6,003 proteins that were quantified in at least 5 out of the 59 cell lines (Figure 3A). In general, cell lines originating from the same tissue converged into the same or closely related clusters. Hierarchical clustering revealed subclusters of colon (seven out of seven), leukemia (five out of six), CNS (five out of six), and melanoma (six out of eight) cell lines in which the samples segregated largely according to their tissue of origin. Interestingly, one melanoma line that did not cluster (LOX IMVI) has been reported to lack melanin production (Stinson et al., 1992), which is a strong determinant for the melanoma cluster. Cell lines from breast and ovarian cancers show a more pronounced distribution across multiple clusters, indicating that their protein expression patterns are quite heterogeneous. We observed, for instance, that the estrogen receptor (ER)-negative breast cancer cell line Hs578T clustered with the stromal/mesenchymal cluster of glioblastoma and renal tumor cell lines. In contrast, the ER-positive breast cancer lines MCF-7 and T47D clustered to colon cancer lines displaying an epithelial phenotype, which is, for instance, characterized by the expression of proteins involved in cell-junction signaling.

We separately performed hierarchical clustering on kinase profiles of all NCI-60 cell lines (220 protein kinases; Figure 3B), which revealed five distinct clusters at an average intercluster correlation coefficient of $\geq$0.6. While CNS (six out of six), leukemia (six out of six), prostate (two out of two), and melanoma (six out of eight) cell lines clustered on single branches, ovarian, and non-small-cell lung cancer lines diverged into multiple clusters, likely reflecting a more heterogeneous molecular phenotype than the aforementioned tumor entities. Interestingly, the clustering analysis shows that the similarity of cell lines according to their tissue of origin is less pronounced on the kinase level compared to the proteome level. On the one hand, this suggests that cancer cell lines preserve, at least to a significant extent, the basic molecular makeup and biological functions of their tissue of origin and thus may be valuable model systems for studying

tissue-specific functions and processes. On the other hand, the striking kinase expression heterogeneity observed for some cell lines may reflect molecularly diverged signal-transduction mechanisms among and between cancer cell types.

Given that genomic instability is a hallmark of cancer cells, it is of note that particularly highly abundant proteins, such as those involved in cellular maintenance (e.g., energy metabolism and protein synthesis), evolve more slowly than proteins of lower abundance, which include most regulatory proteins (Beck et al., 2011; Pál et al., 2001; Subramanian and Kumar, 2004). Highly expressed proteins are apparently under strong selective pressure, likely because of energy constraints (Lane and Martin, 2010) and/or because of a requirement for translational robustness (i.e., minimizing the risk for protein aggregation and toxicity; Drummond et al., 2005). Consistent with the common notion that aberrant signal transduction is an important driver of cancer development and progression, the lower selective pressure of low-abundance proteins, in turn, might represent a general mechanism of how regulatory functions of cancer cells evolve and diverge.

To identify proteins differentially expressed between different tissue categories, we grouped cell lines accordingly. Overall, the correlation of protein expression between tissue groups is strong ($R > 0.8$; Figure S4A). Proteins significantly changing between groups were detected by fitting a linear model to the normalized data and calculating empirical Bayes and moderated F and t statistics. In total, this test determined 522 such proteins ($q < 0.05$) in at least one of the groups across the whole NCI-60 panel. The majority of these proteins are of high abundance (Figures S4B and S4C), which is in accordance with the observation that abundant proteins can be consistently quantified across multiple cell lines.

Principal component analysis (PCA) of these proteins also revealed convergence of cell lines according to their tissue of origin (Figure 3C). Again, the separation of cell lines was less pronounced in the kinase profiles than in the proteome profiles, which is consistent with the hierarchical clustering results. Two cell lines are particularly notable in this context. The MDA-MB-435 cell line, derived from the pleural effusion of a patient with breast cancer, coclustered with melanoma cell lines. This cell line was originally reported as a breast carcinoma cell line, but more recent evidence indicates that it is a derivative of the M14 melanoma cell line (Rae et al., 2007). This is supported by the proteomic data, which suggests that MDA-MB-435 is indeed a melanoma line. We also observed high similarity between MDA-MB-435 and the ovarian line IGROV1, raising the possibility that the latter may also originate from an (occult) melanoma.

Hierarchical clustering of differentially expressed proteins and cell lines disclosed that significantly enriched biological functions and biochemical pathways of selected protein clusters are consistent with actual biological functions of the respective tissue type (Figure 3D). For instance, cluster 3 comprises proteins highly expressed in melanoma cell lines, and the only biological function associated with proteins in this cluster is melanin biosynthesis and pigmentation. In contrast, proteins in cluster 5 are highly expressed in leukemic cell lines and participate, for instance, in intracellular signaling pathways of immune cells (e.g., the tyrosine-protein kinase BTK or the proto-oncogene Vav). We also note that the leukemia lines all show markedly reduced levels of proteins involved in processes more relevant for solid tumors (e.g., cell junction, cell adhesion, blood vessel development, and others).

Protein kinases are the major constituents of cellular signal transduction. Among the 220 kinases identified from kinobead purifications, 49 are differentially expressed between tissue groups. Kinases were identified across all branches of the phylogenetic tree and, in particular, tyrosine kinases (TKs) display lineage-specific differential expression (Figure 4). The TK family comprises multiple targets of anticancer drugs, and many of them show significant differences between tissue types, such as EGFR, EPHA2 and EPHB2, SRC, or MET.

Moreover, the profiles of differentially expressed kinases underscore known drug/target combinations and suggest yet-unexplored potential targets or applications of known drugs. For instance, BTK represents an attractive drug target as it has been found solely in leukemia cell lines. This is consistent with recent results that indicate that the BTK inhibitor ibrutinib has significant activity and is well tolerated in patients with relapsed or refractory B cell malignancies (Advani et al., 2013). Our results also highlight platelet-derived growth factor receptor alpha (PDGFRα), which has been exclusively identified in CNS cell lines, indicating that the inhibition of angiogenesis via PDGFRα might represent a potential target in malignant glioblastomas. A yet-unexplored drug target, for instance, might be the ribosomal protein S6 kinase alpha-4 (RPS6KA4), which has been found to be significantly overexpressed in prostate cancer cell lines. It is downstream of important signaling pathways and involved in the phosphorylation and regulation of various transcription factors including activation of the proto-oncogenes c-Fos and c-Jun (FOS and JUN; Pierrat et al., 1998; Soloaga et al., 2003).

Comparative Analysis of Proteome and Transcriptome Profiles

Transcriptome data are frequently used for global gene expression analysis of cancer cells, and the approach has also been applied to the NCI-60 panel (Pfister et al., 2009). Given that proteins are translated from mRNA templates, it is logical to compare mRNA and proteome profiling data as molecular descriptors of gene expression. The normalized transcriptome data (Figures S5A–S5C) mapped to 13,741 genes above the detection limit, and of all the expressed transcripts, 8,065 (59%) were detected as proteins with a clear bias toward higher-abundance transcripts (Figures 5A and 5B). While 81% of the most abundant mRNAs (first quartile) could be identified on the protein level, this was only the case for 21% of the least expressed transcripts (last quartile). Proteins and transcripts were equally well identified across subcellular localizations (Figure 5C). This is of note as it is a common (albeit probably inaccurate) notion that membrane proteins are underrepresented in proteomic data.

To identify trends and relationship between mRNA and protein abundances across the NCI-60 cell lines, we performed multivariate analyses. Coinertia analysis (CIA) was used to visualize relationships between transcriptome and proteome profiles (Figure 5D). In CIA plots, each arrow represents one of the NCI-60

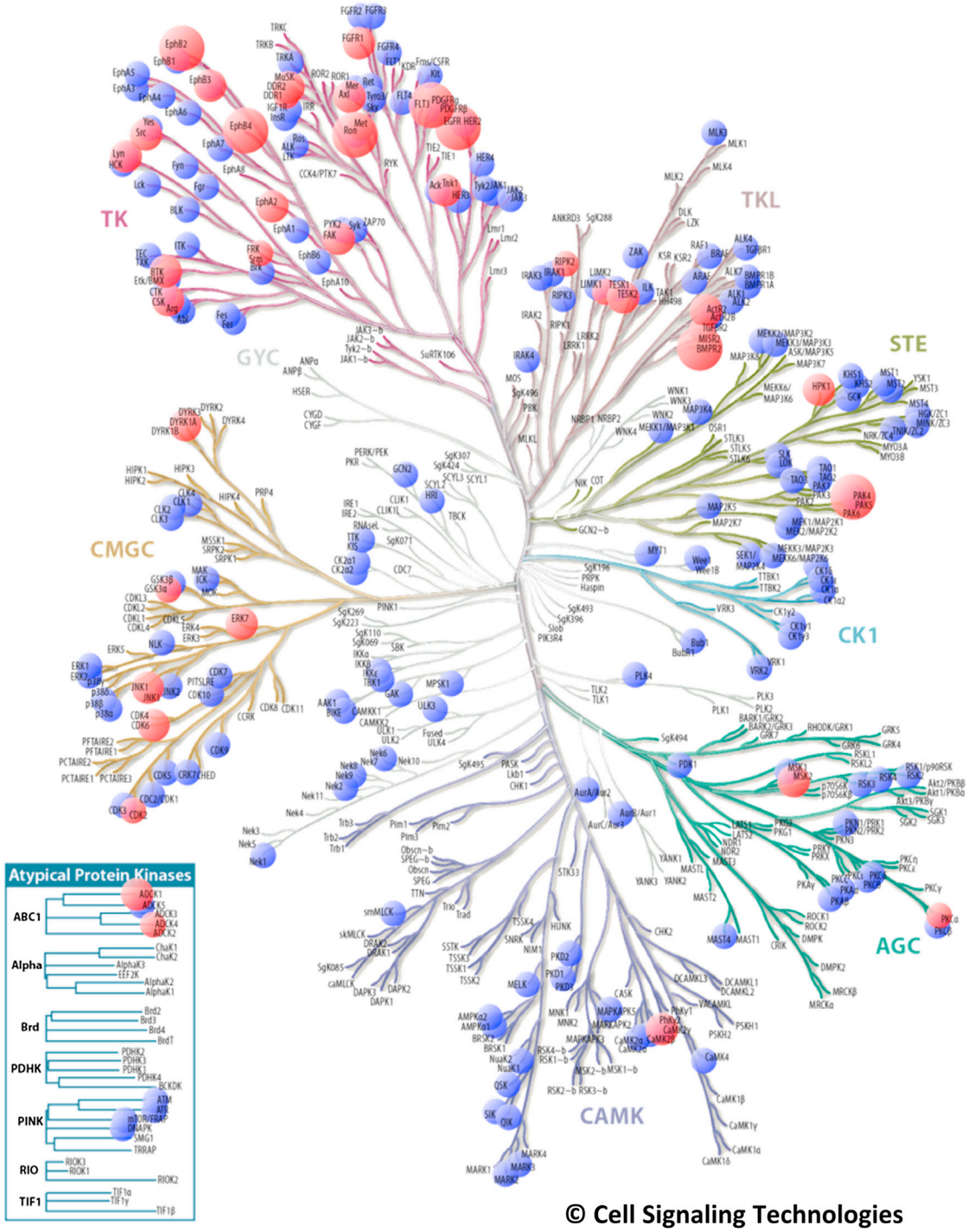

Figure 4. Kinome Tree of Identified and Differentially Expressed Kinases

Mass spectrometric analysis of kinobead purifications from 59 NCI-60 cell lines identified 220 protein kinases across all branches of the phylogenetic tree. Identified and differentially expressed ($q < 0.05$, ANOVA) protein kinases are indicated in blue and red, respectively. Red dots are sized according to the negative logarithmic q value.

Illustration reproduced courtesy of Cell Signaling Technology (http://www.cellsignal.com).

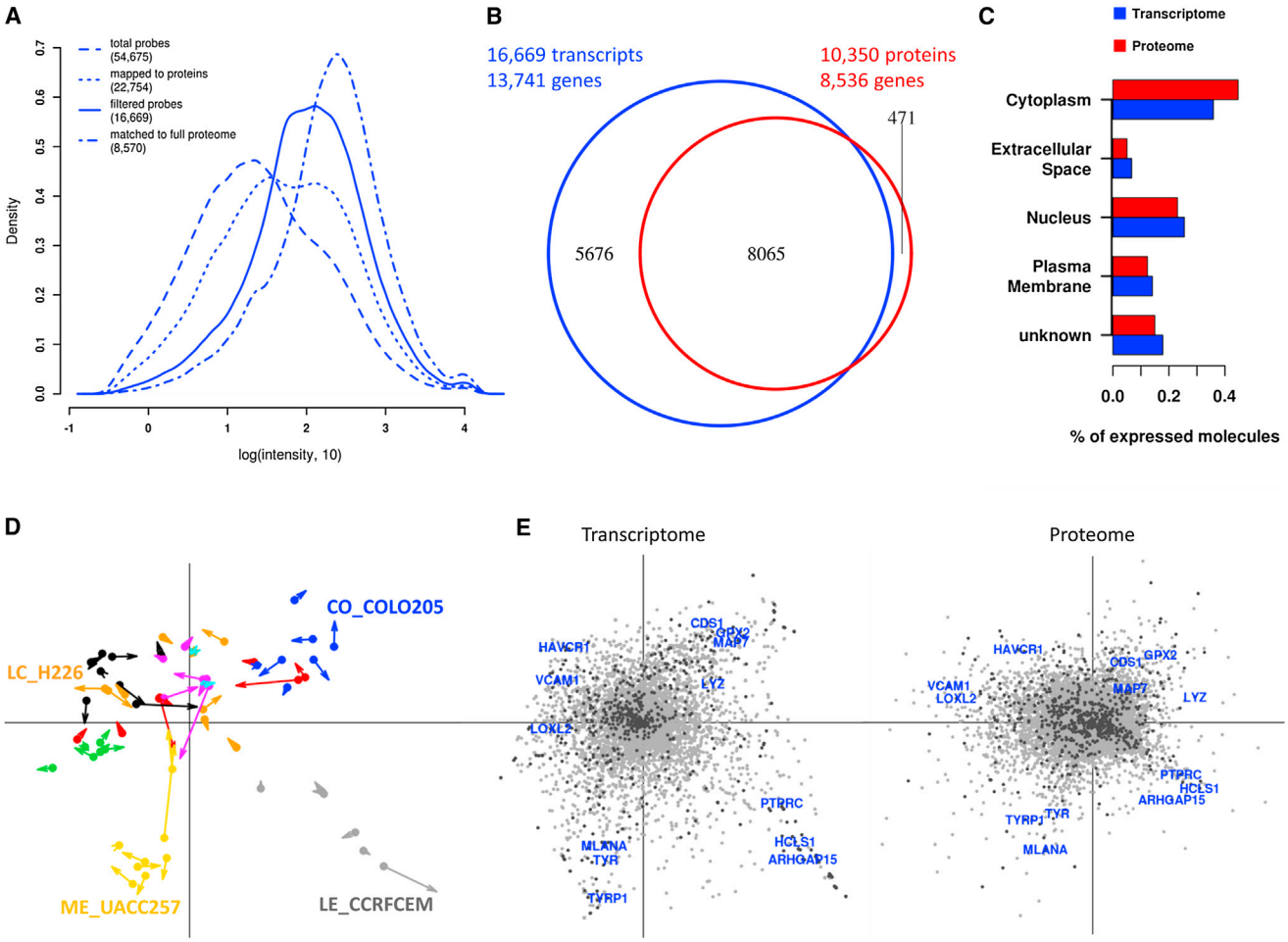

Figure 5. Comparison of the NCI-60 Proteome and Transcriptome

(A) Distribution of mRNA intensities.

(B) Venn diagram of genes detected on mRNA and protein level.

(C) Subcellular localization of mRNA and proteins.

(D) Coinertia analysis of mRNA and protein expression across the complete NCI-60 cell line panel. Arrows represent the projected coordinates of transcriptome (arrow base) and proteome (arrow tip) of the respective cell lines. The length of the arrow is proportional to the divergence between the data sets. The global correlation between the data matrices was 0.76, indicating a high costructure between transcriptome and proteome of the NCI-60 cell lines. Colors represent the nine NCI-60 cell line classes as in Figure 1.

(E) PCA of mRNAs and proteins, respectively, plotted in the same orientation as the CIA.

See also Figures S5 and S6.

cell lines. The base of each arrow represents the position of that cell line in transcriptome space, and the tip of each arrow gives the position in proteome space. The length of an arrow represents the divergence between transcriptome and proteome of a given cell line (the shorter the arrow, the higher the level of concordance between the mRNAs and proteins for a particular cell line). Generally, the previously described cell line clusters are also observed in the CIA plot (for transcriptome and proteome hierarchical clustering clusters, see Figures S5D and 3A, respectively) and represent the same or similar biological functions. This indicates that the high-level biological information content of transcriptomes and proteomes is similar and equally well suited to cluster cell lines into the correct tissue space. We also note that the more heterogeneous cell line groups (e.g., breast cancer cell lines) do not exhibit a considerably higher degree of divergence between their transcriptome and proteome profiles, indicating that the observed spread reflects the heterogeneity on cellular level.

The position of cell lines along the axes of the CIA plot in Figure 5D can be explained by distinct cellular properties. The horizontal axis separates cell lines according to their proliferation rate (Spearman rank correlation of $p = 4.7 \times 10^{-5}$ and $p = 6.4 \times 10^{-6}$ for transcriptome and proteome, respectively; Figure S6A) and the vertical axis separates according to the tissue type. While the majority of cell lines isolated from carcinomas and sarcomas do not exhibit strong divergence, leukemia and melanoma cell lines form distinct clusters toward one end of the vertical axis. We note that many cell lines are projected

further in protein space than in gene space, indicating that the protein expression profiles of these cell lines contain more information than the corresponding mRNA profiles and thus contribute more to the trends on the axes. To disclose mRNAs and proteins responsible for the separation of tissue clusters, we selected highly correlated mRNAs and proteins of the same PCA direction (Figure 5E). While the majority of genes exhibits considerable variation between transcriptome and proteome expression, 629 genes are highly correlated ($R > 0.7$; Figure 5E, black dots) and are strongly expressed on both the transcriptome (Figure 5E, left panel) and proteome levels (Figure 5E, right panel) of colon, leukemia, melanoma, CNS, or renal cell lines. For instance, the integrated analysis of mRNA and protein expression disclosed several markers for leukemia cancer cell lines, including protein tyrosine phosphatase PTPRC (CD45), the hematopoietic lineage cell-specific protein HCLS1, and the RacGTPase-activating protein ArhGAP15, all of which fulfill well-studied functions in immune cells (Costa et al., 2011; Rhee and Veillette, 2012; Yamanashi et al., 1993).

Taken together, despite the moderate depth of proteome coverage per single cell line (again, 6,003 proteins were quantified in at least 5 out of 59 cell lines), proteome and transcriptome data both appear to be powerful molecular descriptors of cancer cells, and their integrative analysis enables a more comprehensive view on the multiple layers of cellular regulation than any technique alone.

Identification of Protein Signatures for Drug Sensitivity and Resistance

In many cases, sensitivity or resistance of cell lines to anticancer therapeutics cannot be simply attributed to single genes or proteins (Garnett et al., 2012). To uncover protein signatures significantly associated with drug sensitivity and resistance, we employed elastic net regression analysis (Zou and Hastie, 2005), which disclosed cooperative interactions between protein expression, their mutational status (Reinhold et al., 2012) and the response signature of anticancer therapeutics (DTP; Rubinstein et al., 1990) across the NCI-60 panel, thereby revealing complex molecular signatures, which might be used as ''panel biomarkers'' (or ''multifeature biomarkers'') to predict drug sensitivity or resistance. Overall, elastic net modeling identified 20,743 protein-drug associations from which 1,801 associations corresponding to 97 different drugs were defined as highly significant (effect size > 90% percentile, frequency > mean frequency × 2 SDs). In most instances, the identified signatures are complex and involve both, mutation and differential protein expression. Interestingly, a small number of proteins were recurrently associated with increased sensitivity to drugs from different classes (Figure 6A), most notably the antiapoptotic regulator Bcl-2 (11/32, significant/detected protein-drug associations) and the helicase CHD4, the latter representing a major component of the nucleosome remodeling and histone deacetylase (NuRD) repressor complex (11/15). Somewhat counterintuitively, Bcl-2 has recently been shown to facilitate initiation of apoptosis after binding to paclitaxel (Ferlini et al., 2009). However, the sensitizing effect of Bcl-2 expression in our data was not observed for tubulin-targeting drugs such as paclitaxel, but it has been significantly and recurrently associated with other classes of chemotherapeutics (e.g., topoisomerase inhibitors or alkylating agents). Although we cannot provide a clear explanation for this observation, the overexpression of Bcl-2 in numerous sensitive cell lines suggests a general role of Bcl-2 in mediating drug sensitivity, potentially by facilitating cells to undergo apoptosis upon treatment.

Proteins often correlated with drug resistance exhibit a variety of cellular functions. The most frequently observed protein, the 14-3-3 protein zeta/delta (YWHAZ; 17/19), has recently been implicated in drug resistance of breast cancer patients (Li et al., 2010) and likely affects drug sensitivity through inhibition of apoptosis via sequestration of the proapoptotic Bcl-2-associated death promoter protein (BAD; Hermeking, 2003; Neal et al., 2009; Niemantsverdriet et al., 2008). Thus, 14-3-3 zeta/delta represents an attractive target to overcome resistance to a broad range of anticancer agents. A notable group of proteins recurrently associated with drug resistance is involved in membrane trafficking and regulation and includes four members of the Rab family (Rab5B, 11/11; Rab1B, 10/13; Rab11, 10/12; and Rab14, 7/10) as well as Sec22B (9/11), indicating that these proteins mediate resistance by facilitating degradation and extrusion of drug molecules, thus keeping intracellular levels of the active drug low. We note that protein kinases were only rarely associated with sensitivity to more than one drug, underscoring the notion that aberrantly expressed and regulated protein kinases represent selective drug targets for selected cancer subtypes. Drug resistance, however, might be more frequently mediated by protein kinases, such as PAK-4 (8/9), which may, again, modulate drug sensitivity through the inhibition of apoptosis (Gnesutta and Minden, 2003; Gnesutta et al., 2001).

Elastic net analysis enabled the identification of drug sensitivity and resistance features for both targeted drugs as well as chemotherapeutics. For instance, sensitivity to paclitaxel, a mitotic drug targeting tubulin, is strongly associated with high expression of aryl hydrocarbon receptor interacting protein (AIP; Figures 6B and S6B), for which several lines of evidence suggest a role as a tumor suppressor (Georgitsi et al., 2007; Nord et al., 2010) and proapoptotic protein (Kang and Altieri, 2006). Despite numerous associated mutations, the strongest correlates for paclitaxel resistance were actually the expression of glutamate dehydrogenase 1 (GluD1) and Rab5B (Figure 6B). Dasatinib sensitivity, in contrast, was notably associated with expression of Src kinase, one of the primary targets of dasatinib, and accompanied by expression of numerous Src substrates, such as STAT1 or the small subunit 1 of the calcium-dependent protease calpain (CAPNS1) as well as integrin beta-1 (ITGB1), a Src activator (Figure 6C). Consistent with this, although STAT1 activation is frequently increased in tumors and cell lines, its functional role has previously been associated with growth suppression and, hence, can be considered as a potential tumor suppressor (Bromberg et al., 1996). Resistance to dasatinib is significantly associated with proteins involved in RNA processing as well as cellular compromise, such as the apoptosis-inducing factor (AIF; Figure 6C). AIF plays a central role in caspase-independent cell death; it has not yet been ascribed a potential role in resistance to targeted cancer drugs, but it might represent valuable therapeutic potential in resistant tumor cells (Lorenzo and Susin, 2007).

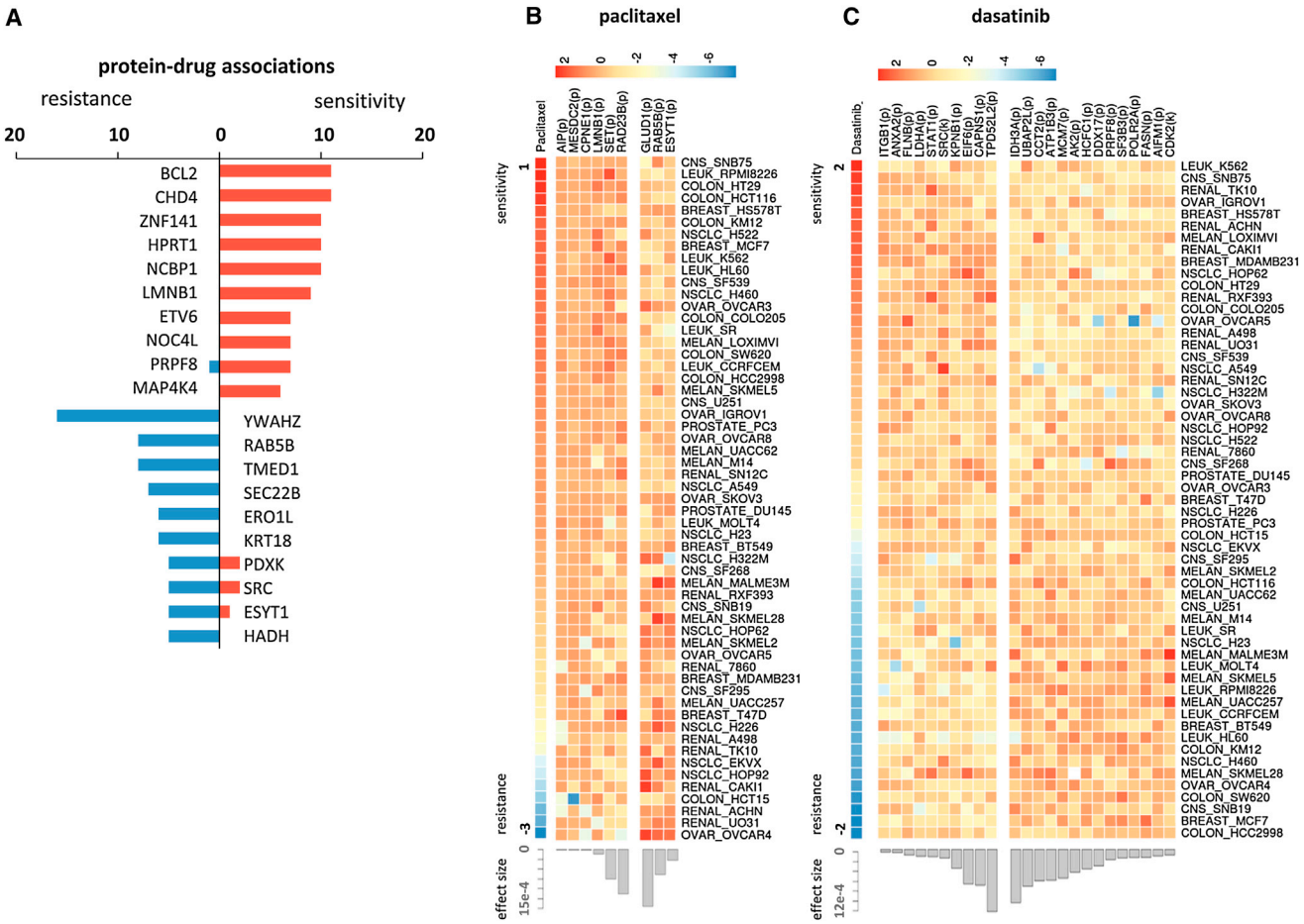

Figure 6. Elastic Net Modeling Reveals Drug Sensitivity and Resistance Signatures

(A) Proteins recurrently and significantly associated with drug resistance and sensitivity.

(B) Heatmaps of highly significant elastic net features associated with response to dasatinib (right) and paclitaxel (left) for the most resistant (blue) and sensitive (red) cell lines. For each cell line, features are at the top of the heatmap followed by expression features (blue corresponds to low expression, red to high expression). To the bottom of each feature is a bar indicating the absolute value of the effect size.

See also Figure S6.

The NCI-60 Proteome Database Enables Access to This Comprehensive Resource

This current proteome resource of 59 commonly employed cancer cell lines enables a wide range of analyses, which are beyond the scope of this study. For instance, the data can be used to obtain reference expression profiles for proteins of interest both within and across a range of cell line proteomes. It may facilitate selecting the appropriate cell line for the study of a particular biological phenomenon of interest without the need for additional experiments such as RNA deep sequencing (Danielsson et al., 2013), or might alternatively be used as reference for targeted proteome analyses (Picotti and Aebersold, 2012). To enable efficient use of these data by the scientific community, the data were incorporated into a database, which is accessible via a user-friendly web interface. The database contains protein expression profiles along with details about protein and peptide identification information, including the matched tandem mass spectra. The data are cross-referenced to Uniprot, Ensemble, and several other major resources to facilitate additional information browsing.

Conclusions

In the present study, we provide a comprehensive resource of NCI-60-wide protein expression profiles for more than 10,000 proteins, including more than 350 protein kinases. Our results indicate that, despite the moderate depth of proteome coverage per single cell line, the broad biological information contained in transcriptomics and proteomics data is similar, but each technique provides complementary information. While transcriptomics enables genome-wide investigations of mRNAs and has proven very useful, it can by nature not be utilized to study post-transcriptional processes of cellular regulation, such as protein expression, posttranslational modifications, protein degradation, protein interaction, or protein activities, which are areas largely confined to proteome analysis (some of which were highlighted here). Although the proteomic data acquired in this study

did not reach the same degree of genome coverage as the mRNA profiles, the proteomic data appeared to be particularly powerful for the identification of mechanisms by which cancer cells evade potent targeted inhibitors or broad chemotherapeutic compounds. Using the experimental and bioinformatic methods employed here, we anticipate that proteomics will play an increasing role in molecular profiling of cancer, and we are making our data available to the community via a database so that the data can be broadly utilized in research aimed at understanding and fighting cancer.

EXPERIMENTAL PROCEDURES

Protein Preparation and Kinobead Affinity Purification

Cell pellets obtained from DTP of the NCI were lysed and kinobead pull-downs were performed and prepared for in-gel digestion as previously described (Wu et al., 2011, 2012). For full proteomes, 50 μg from each kinobead flow-through were reduced, alkylated, and separated via an LDS-PAGE gel. It is of note that the kinobead procedure does not result in significant kinase depletion from the flow-through (Bantscheff et al., 2007). In-gel trypsin digestion was performed according to standard procedures (Shevchenko et al., 1996).

Protein Identification and Quantification

Nanoflow LC-MS/MS analyses of tryptic peptides were conducted on an Eksigent nanoLC-Ultra 1D+ (Eksigent) coupled to an LTQ Orbitrap XL ETD or Orbitrap Elite mass spectrometer (Thermo Scientific). The mass spectrometers were operated in data-dependent mode and raw MS spectra were processed using Maxquant (version 1.3.0.3; Cox and Mann, 2008). Tandem mass spectra were searched with Andromeda (Cox et al., 2011) against the IPI human database (version 3.68; 87,061 sequences) and a maximum false discovery rate (FDR) of 1% for peptides and proteins was required. Protein abundance was estimated based on summed peptide intensities from proteome profiling experiments (but not deep proteomes and kinomes), and label-free quantification was used for comparisons between samples (Luber et al., 2010).

Statistical and GO Enrichment Analyses

Statistical analysis of quantified proteins was performed using R (version 2.12.1; Team, 2012). Differential expression was assessed via ANOVA and p values were corrected for multiple hypothesis testing to control the FDR at 5% (Benjamini and Hochberg, 1995). Cluster analyses including hierarchical clustering and PCA were performed using a variety of algorithms and metrics. Classification and functional enrichment as well as pathway membership were analyzed using BiNGO (Maere et al., 2005) and Ingenuity Pathway Analysis (Ingenuity Systems). The kinome dendrogram was an adapted and reproduced courtesy of Cell Signaling Technology using modified version of the of Human Kinome application of the Tripod project (http://tripod.nih.gov/).

Comparison of Proteomics and Transcriptomics

Normalized gene expression data for NCI-60 cell lines were obtained from the Gene Expression Omnibus (series accession number GSE32474; Barrett et al., 2009; Pfister et al., 2009). Significant differences were identified applying a Bayesian approach using the limma package (Bioconductor 2.7; Smyth, 2004). A threshold of an adjusted p value < 0.05 was used to identify significant changes. CIA (Culhane et al., 2003; Dolédec and Chessel, 1994) was used to analyze statistical relationships between protein and gene expression patterns.

Protein-Drug Associations

Elastic net regression (Zou and Hastie, 2005) was used to identify associations between proteins and drug response across NCI-60 cell lines. Protein-drug associations have been assessed for 108 FDA-approved drugs, and drug activity levels were obtained from Cellminer (Reinhold et al., 2012). Proteomic data, including full proteome data (8,113 proteins), kinome profiling (220 kinases), NCI-60 mutation data (from the Cellminer database), as well as the tissue type, were used as input variables.

For further details, please refer to the Extended Experimental Procedures.

ACCESSION NUMBERS

The MS raw files associated with this manuscript are available for download (https://www.proteomicsdb.org/proteomicsdb/#projects/35). Processed MS files (MaxQuant output files) are available for download (http://wzw.tum.de/proteomics/nci60).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures and six figures and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.07.018.

ACKNOWLEDGMENTS

We gratefully acknowledge Fiona Pachl for mass spectrometric analyses, Guillaume Médard and Thomas Kuehne for programming, and Michaela Kroetz-Fahning, Andrea Hubauer, and Andreas Klaus for excellent laboratory assistance.

Received: February 14, 2013
Revised: May 23, 2013
Accepted: July 18, 2013
Published: August 8, 2013

REFERENCES

Abaan, O.D., Polley, E.C., Davis, S.R., Zhu, Y.J., Bilke, S., Walker, R.L., Pineda, M., Gindin, Y., Jiang, Y., Reinhold, W.C., et al. (2013). The exomes of the NCI-60 Panel: a genomic resource for cancer biology and systems pharmacology. Cancer Res. *73*, 4372–4382.

Advani, R.H., Buggy, J.J., Sharman, J.P., Smith, S.M., Boyd, T.E., Grant, B., Kolibaba, K.S., Furman, R.R., Rodriguez, S., Chang, B.Y., et al. (2013). Bruton tyrosine kinase inhibitor ibrutinib (PCI-32765) has significant activity in patients with relapsed/refractory B-cell malignancies. J. Clin. Oncol. *31*, 88–94.

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C., et al. (2007). Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat. Biotechnol. *25*, 1035–1044.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., et al. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. *37*(Database issue), D885–D890.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol. Syst. Biol. *7*, 549.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Blower, P.E., Chung, J.H., Verducci, J.S., Lin, S., Park, J.K., Dai, Z., Liu, C.G., Schmittgen, T.D., Reinhold, W.C., Croce, C.M., et al. (2008). MicroRNAs modulate the chemosensitivity of tumor cells. Mol. Cancer Ther. *7*, 1–9.

Boyd, M., and Paull, K. (1995). Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. Drug Dev. Res. *34*, 91–109.

Bromberg, J.F., Horvath, C.M., Wen, Z., Schreiber, R.D., and Darnell, J.E., Jr. (1996). Transcriptionally active Stat1 is required for the antiproliferative effects

of both interferon alpha and interferon gamma. Proc. Natl. Acad. Sci. USA *93*, 7673–7678.

Burkard, T.R., Planyavsky, M., Kaupe, I., Breitwieser, F.P., Bürckstümmer, T., Bennett, K.L., Superti-Furga, G., and Colinge, J. (2011). Initial characterization of the human central proteome. BMC Syst. Biol. *5*, 17.

Bussey, K.J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W.C., Kuo, W.L., Gwadry, F., Ajay, Kouros-Mehr, H., Fridlyand, J., et al. (2006). Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Mol. Cancer Ther. *5*, 853–867.

Cohen, P. (2002). Protein kinases—the major drug targets of the twenty-first century? Nat. Rev. Drug Discov. *1*, 309–315.

Costa, C., Germena, G., Martin-Conte, E.L., Molineris, I., Bosco, E., Marengo, S., Azzolino, O., Altruda, F., Ranieri, V.M., and Hirsch, E. (2011). The RacGAP ArhGAP15 is a master negative regulator of neutrophil functions. Blood *118*, 1099–1108.

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.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. *10*, 1794–1805.

Culhane, A.C., Perrière, G., and Higgins, D.G. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics *4*, 59.

Danielsson, F., Wiking, M., Mahdessian, D., Skogs, M., Ait Blal, H., Hjelmare, M., Stadler, C., Uhlen, M., and Lundberg, E. (2013). RNA deep sequencing as a tool for selection of cell lines for systematic subcellular localization of all human proteins. J. Proteome Res. *12*, 299–307.

Dolédec, S., and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw. Biol. *31*, 277–294.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA *102*, 14338–14343.

Ehrich, M., Turner, J., Gibbs, P., Lipton, L., Giovanneti, M., Cantor, C., and van den Boom, D. (2008). Cytosine methylation profiling of cancer cell lines. Proc. Natl. Acad. Sci. USA *105*, 4844–4849.

Ferlini, C., Cicchillitti, L., Raspaglio, G., Bartollino, S., Cimitan, S., Bertucci, C., Mozzetti, S., Gallo, D., Persico, M., Fattorusso, C., et al. (2009). Paclitaxel directly binds to Bcl-2 and functionally mimics activity of Nur77. Cancer Res. *69*, 6906–6914.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., et al. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117–122.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell. Proteomics *11*, M111 014050.

Georgitsi, M., Karhu, A., Winqvist, R., Visakorpi, T., Waltering, K., Vahteristo, P., Launonen, V., and Aaltonen, L.A. (2007). Mutation analysis of aryl hydrocarbon receptor interacting protein (AIP) gene in colorectal, breast, and prostate cancers. Br. J. Cancer *96*, 352–356.

Gnesutta, N., and Minden, A. (2003). Death receptor-induced activation of initiator caspase 8 is antagonized by serine/threonine kinase PAK4. Mol. Cell. Biol. *23*, 7838–7848.

Gnesutta, N., Qu, J., and Minden, A. (2001). The serine/threonine kinase PAK4 prevents caspase activation and protects cells from apoptosis. J. Biol. Chem. *276*, 14414–14419.

Hanash, S., and Taguchi, A. (2010). The grand challenge to decipher the cancer proteome. Nat. Rev. Cancer *10*, 652–660.

Hermeking, H. (2003). The 14-3-3 cancer connection. Nat. Rev. Cancer *3*, 931–943.

Holbeck, S.L., Collins, J.M., and Doroshow, J.H. (2010). Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. Mol. Cancer Ther. *9*, 1451–1460.

Ikediobi, O.N., Davies, H., Bignell, G., Edkins, S., Stevens, C., O'Meara, S., Santarius, T., Avis, T., Barthorpe, S., Brackenbury, L., et al. (2006). Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol. Cancer Ther. *5*, 2606–2612.

Kang, B.H., and Altieri, D.C. (2006). Regulation of survivin stability by the aryl hydrocarbon receptor-interacting protein. J. Biol. Chem. *281*, 24721–24727.

Knapp, S., Arruda, P., Blagg, J., Burley, S., Drewry, D.H., Edwards, A., Fabbro, D., Gillespie, P., Gray, N.S., Kuster, B., et al. (2013). A public-private partnership to unlock the untargeted kinome. Nat. Chem. Biol. *9*, 3–6.

Lane, N., and Martin, W. (2010). The energetics of genome complexity. Nature *467*, 929–934.

Lemeer, S., Zörgiebel, C., Ruprecht, B., Kohl, K., and Kuster, B. (2013). Comparing immobilized kinase inhibitors and covalent ATP probes for proteomic profiling of kinase expression and drug selectivity. J. Proteome Res. Published online March 28, 2013.

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., Desmedt, C., Sotiriou, C., Szallasi, Z., Iglehart, J.D., et al. (2010). Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. Nat. Med. *16*, 214–218.

Liu, H., D'Andrade, P., Fulmer-Smentek, S., Lorenzi, P., Kohn, K.W., Weinstein, J.N., Pommier, Y., and Reinhold, W.C. (2010). mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. Mol. Cancer Ther. *9*, 1080–1091.

Lorenzi, P.L., Reinhold, W.C., Varma, S., Hutchinson, A.A., Pommier, Y., Chanock, S.J., and Weinstein, J.N. (2009). DNA fingerprinting of the NCI-60 cell line panel. Mol. Cancer Ther. *8*, 713–724.

Lorenzo, H.K., and Susin, S.A. (2007). Therapeutic potential of AIF-mediated caspase-independent programmed cell death. Drug Resist. Updat. *10*, 235–255.

Luber, C.A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M., and Mann, M. (2010). Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. Immunity *32*, 279–289.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol. *6*, 450.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448–3449.

Mallick, P., and Kuster, B. (2010). Proteomics: a pragmatic perspective. Nat. Biotechnol. *28*, 695–709.

Malmström, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E.W., and Aebersold, R. (2009). Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. Nature *460*, 762–765.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Neal, C.L., Yao, J., Yang, W., Zhou, X., Nguyen, N.T., Lu, J., Danes, C.G., Guo, H., Lan, K.H., Ensor, J., et al. (2009). 14-3-3zeta overexpression defines high risk for breast cancer recurrence and promotes cancer cell survival. Cancer Res. *69*, 3425–3432.

Niemantsverdriet, M., Wagner, K., Visser, M., and Backendorf, C. (2008). Cellular functions of 14-3-3 zeta in apoptosis and cell adhesion emphasize its oncogenic character. Oncogene *27*, 1315–1319.

Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W.C., Waltham, M., Kouros-Mehr, H., Bussey, K.J., Lee, J.K., Espina, V., et al. (2003).

Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc. Natl. Acad. Sci. USA *100*, 14229–14234.

Nord, K.H., Magnusson, L., Isaksson, M., Nilsson, J., Lilljebjörn, H., Domanski, H.A., Kindblom, L.G., Mandahl, N., and Mertens, F. (2010). Concomitant deletions of tumor suppressor genes MEN1 and AIP are essential for the pathogenesis of the brown fat tumor hibernoma. Proc. Natl. Acad. Sci. USA *107*, 21122–21127.

Pál, C., Papp, B., and Hurst, L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics *158*, 927–931.

Park, E.S., Rabinovsky, R., Carey, M., Hennessy, B.T., Agarwal, R., Liu, W., Ju, Z., Deng, W., Lu, Y., Woo, H.G., et al. (2010). Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. Mol. Cancer Ther. *9*, 257–267.

Pfister, T.D., Reinhold, W.C., Agama, K., Gupta, S., Khin, S.A., Kinders, R.J., Parchment, R.E., Tomaszewski, J.E., Doroshow, J.H., and Pommier, Y. (2009). Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Mol. Cancer Ther. *8*, 1878–1884.

Picotti, P., and Aebersold, R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat. Methods *9*, 555–566.

Pierrat, B., Correia, J.S., Mary, J.L., Tomás-Zuber, M., and Lesslauer, W. (1998). RSK-B, a novel ribosomal S6 kinase family member, is a CREB kinase under dominant control of p38alpha mitogen-activated protein kinase (p38alphaMAPK). J. Biol. Chem. *273*, 29661–29671.

Rae, J.M., Creighton, C.J., Meck, J.M., Haddad, B.R., and Johnson, M.D. (2007). MDA-MB-435 cells are derived from M14 melanoma cells—a loss for breast cancer, but a boon for melanoma research. Breast Cancer Res. Treat. *104*, 13–19.

Reinhold, W.C., Sunshine, M., Liu, H., Varma, S., Kohn, K.W., Morris, J., Doroshow, J., and Pommier, Y. (2012). CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. *72*, 3499–3511.

Rhee, I., and Veillette, A. (2012). Protein tyrosine phosphatases in lymphocyte activation and autoimmunity. Nat. Immunol. *13*, 439–447.

Roschke, A.V., Tonon, G., Gehlhaus, K.S., McTyre, N., Bussey, K.J., Lababidi, S., Scudiero, D.A., Weinstein, J.N., and Kirsch, I.R. (2003). Karyotypic complexity of the NCI-60 drug-screening panel. Cancer Res. *63*, 8634–8647.

Rubinstein, L.V., Shoemaker, R.H., Paull, K.D., Simon, R.M., Tosini, S., Skehan, P., Scudiero, D.A., Monks, A., and Boyd, M.R. (1990). Comparison of in vitro anticancer-drug-screening data generated with a tetrazolium assay versus a protein assay against a diverse panel of human tumor cell lines. J. Natl. Cancer Inst. *82*, 1113–1118.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nat. Genet. *24*, 236–244.

Schirle, M., Heurtier, M.A., and Kuster, B. (2003). Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics *2*, 1297–1305.

Schirle, M., Bantscheff, M., and Kuster, B. (2012). Mass spectrometry-based proteomics in preclinical drug discovery. Chem. Biol. *19*, 72–84.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., Major, S., Morita, D., Chary, K.K., Reimers, M.A., Scherf, U., Kahn, A., Dolginow, D., et al. (2007). Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol. Cancer Ther. *6*, 820–832.

Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal. Chem. *68*, 850–858.

Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Soloaga, A., Thomson, S., Wiggin, G.R., Rampersaud, N., Dyson, M.H., Hazzalin, C.A., Mahadevan, L.C., and Arthur, J.S. (2003). MSK2 and MSK1 mediate the mitogen- and stress-induced phosphorylation of histone H3 and HMG-14. EMBO J. *22*, 2788–2797.

Stinson, S.F., Alley, M.C., Kopp, W.C., Fiebig, H.H., Mullendore, L.A., Pittman, A.F., Kenney, S., Keller, J., and Boyd, M.R. (1992). Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. Anticancer Res. *12*, 1035–1053.

Subramanian, S., and Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics *168*, 373–381.

Team, R.D.C. (2012). R: A Language and Environment for Statistical Computing.

Weinstein, J.N. (2006). Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. Mol. Cancer Ther. *5*, 2601–2605.

Wu, Z., Doondeea, J.B., Gholami, A.M., Janning, M.C., Lemeer, S., Kramer, K., Eccles, S.A., Gollin, S.M., Grenman, R., Walch, A., et al. (2011). Quantitative chemical proteomics reveals new potential drug targets in head and neck cancer. Mol. Cell. Proteomics *10*, M111.011635.

Wu, Z., Moghaddas Gholami, A., and Kuster, B. (2012). Systematic identification of the HSP90 candidate regulated proteome. Mol. Cell. Proteomics *11*, M111.016675.

Yamanashi, Y., Okada, M., Semba, T., Yamori, T., Umemori, H., Tsunasawa, S., Toyoshima, K., Kitamura, D., Watanabe, T., and Yamamoto, T. (1993). Identification of HS1 protein as a major substrate of protein-tyrosine kinase(s) upon B-cell antigen receptor-mediated signaling. Proc. Natl. Acad. Sci. USA *90*, 3631–3635.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B *67*, 199–320.

Cell PRESS
Open ACCESS

EXTENDED EXPERIMENTAL PROCEDURES

Protein Preparation from Cell Pellets

Cell line pellets provided by the Drug Therapeutics Program (DTP) of the National Cancer Institute (NCI) were lysed with 1x compound pulldown (CP) buffer (50 mMTris/HCl pH 7.5, 5% Glycerol) supplemented with 0.8% Nonidet P-40 and freshly added protease (SIGMAFAST, Sigma-Aldrich, Germany) and phosphatase inhibitors (20 mMNaF, 1 mM sodium orthovanadate, 5 mMcalyculin A; Sigma-Aldrich, Germany). Homogenates were centrifuged at 6000 x g at 4°C for 10 min followed by ultracentrifugation at 4°C for 1h at 145,000 x g, supernatants were collected and aliquots were frozen in liquid nitrogen and stored at −80°C until further use. Protein concentration in lysates was determined by the Bradford assay.

Kinobead Affinity Purification

Kinobead pulldowns were performed as described (Wu et al., 2011, 2012). Briefly, cell lysates were diluted with equal volumes of 1x CP buffer containing protease and phosphatase inhibitors. Lysates were further diluted if necessary to a final protein concentration of 5 mg/ml using 1 x CP buffer supplemented with 0.4% Nonidet P-40 followed by incubation with kinobeads at 4°C for 4 hr. Subsequently, beads were washed with 1 x CP buffer and collected by centrifugation. Bound proteins were eluted with 2 xNuPAGE® LDS Sample Buffer (Invitrogen, Darmstadt, Germany) and eluates were reduced and alkylated by 10 mM DTT (dithiothreitol) and 55 mM IAA (iodoacetamide). Samples were then run into a 4%–12% NuPAGE gel (Invitrogen, Darmstadt, Germany) for about 1 cm to concentrate the sample prior to in-gel tryptic digestion. In-gel trypsin digestion was performed according to standard procedures (Shevchenko et al., 1996).

Full Proteome and Deep Proteome Separation

50 μg flow-through from each kinobead pulldown were reduced and alkylated by 10 mM DTT and 55 mM IAA and sequentially denatured at 95°C for 10 min. Samples were then separated via a 4%–12% NuPAGE gel (Invitrogen, Darmstadt, Germany) and cut into 12 slices for the full proteome and 24 slices for the deep proteome experiments prior to in-gel tryptic digestion. In-gel trypsin digestion was performed according to standard procedures (Shevchenko et al., 1996).

LC-MS/MS Analysis

Nanoflow LC-MS/MS was performed by coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent, Dublin, CA) to a LTQ Orbitrap XL ETD or Orbitrap Elite mass spectrometer (Thermo Scientific, Bremen, Germany).

Full proteome and kinobead eluates were analyzed on the LTQ Orbitrap XL ETD mass spectrometer, while for the deep proteome profiles, the more sensitive Orbitrap Elite mass spectrometer was employed. Tryptic peptides were dissolved in 20μl 0.1% formic acid and 10 μl was injected for each analysis. Peptides were delivered to a trap column (100 μmi.d. × 2 cm, packed with 5μm C18 resin, Reprosil PUR AQ, Dr. Maisch, Ammerbuch, Germany) at a flow rate of 5 μL/min in 100% buffer A (0.1% FA in HPLC grade water). After 10 min of loading and washing, peptides were transferred to an analytical column (75μmx40 cm, C18 Reprosil PUR AQ, 3μm, Dr. Maisch, Ammerbuch, Germany) and separated either using a 210 min gradient (kinobead eluates), 110 min gradient (full proteomes) or 60 min gradient (deep proteome) from 2% to 35% of buffer B (0.1% FA in acetonitrile) at 300 nL/minute flow rate.

The mass spectrometers were operated in data dependent mode, automatically switching between MS and MS2. Precursor masses selected for MS2 were dynamically excluded from fragmentation for 10 s (full proteome and kinobead eluates) and 120 s (deep proteome), respectively. Full scan MS spectra were acquired in the Orbitrap at 60,000 resolutions. Internal calibration was performed using the ion signal $(Si(CH_3)_2O)_6H^+$ at m/z 445.120025 present in ambient laboratory air. Tandem mass spectra were acquired using collision-induced dissociation (CID) for kinobead and full proteome experiments and higher energy collision induced dissociation (HCD) for deep proteome experiments.

Peptide and Protein Quantification and Identification

Raw MS spectra were processed by Maxquant (version 1.3.0.3) for peak detection and quantification (Cox and Mann, 2008). MS/MS spectra were searched against the IPI human database human (v. 3.68; 87,061 sequences) using the Andromeda search engine (Cox et al., 2011) enabling contaminants and the reversed versions of all sequences with the following search parameters: Carbamidomethylation of cysteine residues as fixed modification and Acetyl (Protein N-term) andOxidation (M) as variable modifications. Trypsin was specified as proteolytic enzyme with up to 2 missed cleavages. Mass accuracy of the precursor ions was determined by the time-dependent recalibration algorithm of Maxquant, and fragment ion mass tolerance was set to of 0.6 Da and 20 ppm for CID and HCD, respectively. The maximum false discovery rate (FDR) for proteins and peptides was 0.01 and a minimum peptide length of 6 amino acids was required.

Statistical Analysis

Statistical analysis of quantified proteins was performed using R (v 2.12.1; Team, 2012). We normalized the protein intensities by the summed peptide intensities over the number of theoretically observable peptides of the protein (Schwanhausser et al., 2011). For comparison between samples we used label-free quantification with a minimum of two ratio counts to determine the normalized

protein intensity (Luber et al., 2010). To investigate the data distribution and ensure the appropriate application of statistical tools, normal quantile-quantile plots were created for all protein intensities in each cell line and variance-mean dependencies were visually verified. Statistical significance of differential expression was assessed by performing a multiple t test (ANOVA) on proteins that are quantified in at least 5 out of the 59 cell lines. We accepted or rejected the null hypothesis on the basis of P-values computed for the multiple t test at a specified significance level. P-values were adjusted for multiple testing to control the False Discovery Rate (FDR) at 5%. For multiple testing adjustments, we calculated the FDR using the algorithm of Benjamini and Hochberg (Benjamini and Hochberg, 1995). P-values, with appropriate multiple testing adjustment to control the False Discovery Rate (FDR) at 5% allowed us to identify differentially expressed proteins. Cluster analyses using a variety of algorithms and metrics were performed to group the cell lines on the basis of protein expression pattern. Hierarchical clustering of proteins was performed on logarithmized intensities after filtration of the data to have at least 5 valid values across all cell lines and z-score normalization of the data, using Spearman Rank correlation with Ward metric.

GO Enrichment/Pathway Analysis

Classification and functional enrichment analysis of the identified proteins were performed using BiNGO (Maere et al., 2005) and DAVID Bioinformatics Database (Huang da et al., 2009) for the biological process (BP), molecular function (MF) and cellular component (CC). Pathway membership of the identified proteins were analyzed by the Ingenuity Pathway Analysis (IPA) tool (Ingenuity Systems) for their functional significance and in the context of biological association networks.

Comparison of Proteomics and Transcriptomics

Normalized gene expression data for NCI-60 cell lines were obtained at GEO (Barrett et al., 2009) through the series accession number GSE32474 (Pfister et al., 2009). Significant differences were identified applying a Bayesian approach using the limma package (Bioconductor 2.7; Smyth, 2004). A threshold of an adjusted p-value ≤ 0.05 was used to identify significant changes.

We used co-inertia analysis to analyze statistical relationships between protein and gene expression patterns (Culhane et al., 2003; Dolédec and Chessel, 1994). Principal component analysis was used to visualize expression profiles in each data set. In co-inertia plot (like with PCA) a protein that is particularly highly expressed in a certain cell line will be located in the direction of this cell line. The farther away toward the outer margin of the plot in this direction it is displayed, the stronger the association. Overall similarity of the data sets is captured by the RV-coefficient ($0 < RV < 1$) that is a commonly used matrix correlation (Robert and Escoufier, 1976). In CIA, the RV is calculated as the co-inertia (sum of eigenvalues of a co-inertia analysis) divided by the square root of the product of the square inertias (sum of the eigenvalues) from the individual PCA. Much like a correlation coefficient, the stronger the joint trends between two data sets agree, the closer to 1 the RV score becomes. A zero RV score indicates no similarity.

Coinertia Analysis

Co-inertia analysis (CIA) is a multivariate approach that can identify co-relationships in multiple data sets by finding successive principal axes of maximum co-variance. It was first introduced applying ecological data (Dolédec and Chessel, 1994) using co-inertia as a measure of co-structure between two data matrices. When the matrices are centered, co-inertia is a sum of square covariance. Culhane and co-workers demonstrated the efficiency of CIA on cross platform comparisons of gene expression data (Culhane et al., 2003). CIA is often used in combination with principal component analysis (PCA) to visualize relationships. The mathematical basis of CIA, following the notation of Dolédec and coworkers (Dolédec and Chessel, 1994) is summarized as below.

Let X and Y be the original data tables, with n rows, and respectively p and q columns. The two statistical triplets produced by the ordination methods performed on the data sets are denoted (X, D_n, D_p) and (Y, D_n, D_q), with D_n and D_p being diagonal matrices containing row and column weights for X, and D_n and D_q diagonal matrices containing row and column weights for Y. After diagonalization let u and v be a pair of eigenvectors for (X, D_n, D_p) and (Y, D_n, D_q), respectively. The projection of the multidimensional space associated with X onto vector u generates n coordinates in a column matrix:

$$\alpha = XD_p u. \quad \text{(Equation 1)}$$

The projection of the multidimensional space associated with table Y on to vector v generates n coordinates in a column matrix:

$$\varphi = YD_q v. \quad \text{(Equation 2)}$$

Co-inertia associated with the pair of vectors u and v can be written as

$$H(u,v) = \alpha^t D\varphi. \quad \text{(Equation 3)}$$

If the initial data tables are centered, then the co-inertia is the covariance between the two new scores:

$$Cov(\alpha,\varphi) = Corr(\alpha,\varphi)\sqrt{\eta_1(u)\eta_2(v)}. \quad \text{(Equation 4)}$$

With $\eta_1(u)$denoting the projected inertia on to vector u (i.e., the variance of the new scores on u), $\eta_2(v)$the projected inertia on to vector v (i.e., the variance of the new scores on v), and $Corr(\alpha, \varphi)$ the correlation between the two coordinate systems. A CIA axis associated with a pair of eigenvectors u and v will maximize $Cov(\alpha,\varphi)$.

Elastic Net Analysis

We used the described approach of the elastic net (Zou and Hastie, 2005). Proteomic data including proteome profiling (8113 proteins), kinomes (220 protein kinases), and 221 mutations of 21 cancer genes were used as input variables. Drug activity levels expressed as 50% growth-inhibitory levels (GI50) were determined by the DTP (http://dtp.nci.nih.gov/) at 48 hr using the sulforhodamine B assay (Rubinstein et al., 1990). Standardized drug activity data (Z scores) were obtained from Cellminer (Reinhold et al., 2012). In our study, we considered only FDA approved drugs (108 drugs) and all features were regressed to fit a Gaussian model of GI50 values for each drug.

Elastic Net Definition

Features data were standardized before passing to elastic net. Let X be a n × p matrix of input features, where n is the number of cell lines and p is the number of features. For all features in X:

$$\sum_{i=1}^{n} X_{ij} = 0 \; and \; \sum_{i=1}^{n} X_{ij}^2 = 1. \qquad \text{(Equation 5)}$$

Naive elastic net criterion was then defined by

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|, \qquad \text{(Equation 6)}$$

where λ_1 and λ_2 are non-negative values. The right side of the above equation is the ordinary least square regression criterion with combination of two widely used penalties (RIDGE and LASSO). The naive elastic net estimator is the solution that satisfies

$$\widehat{\beta}\ NaiveElasticNet = {}^{\text{argmin}}_{\beta}\{L(\lambda_1, \lambda_2, \beta)\}. \qquad \text{(Equation 7)}$$

A scaling factor is then added to the naive elastic net to prevent double shrinking of combining RIDGE and LASSO in Equation 6.

$$\widehat{\beta}\ ElasticNet = (1 + \lambda_2)\widehat{\beta}\ NaiveElasticNet. \qquad \text{(Equation 8)}$$

In addition, elastic net mixing parameter α was defined to control the optimization of λ_1 and λ_2

$$\alpha = \frac{\lambda_1}{(\lambda_1 + \lambda_2)}. \qquad \text{(Equation 9)}$$

Then the elastic net penalty can be defined as

$$(1 - \alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|. \qquad \text{(Equation 10)}$$

Implementation

Elastic net was conducted by R package "glmnet" (Friedman et al., 2010; Simon et al., 2011), 4 fold cross validation was used to optimize α. Potential α values were restricted from 0.001 to 0.2 in order to control the number of final markers retained in each run. Under the optimized α, 75% of cell lines across the whole data set were randomly selected to identify biomarkers. The procedure was repeated 100 times for each drug. The final signature of markers for a drug consisted of all features that frequently appear in any of the 100 runs (mean of frequency + 2 × standard deviation) and represent a consistent weight in different signatures. The weight was defined as

$$\omega = \frac{F[\beta > 0] - F[\beta < 0]}{N}, \qquad \text{(Equation 11)}$$

where F stands for the frequency of an event and N is the number of signatures containing the feature. Only features with ω equals 1 (positively correlated with GI50, sensitive marker) or -1 (negatively correlated with GI50, resistance marker) was selected as potential markers. In addition, effect size was introduced to assess the efficiency of a feature resulting in drug resistance or sensitivity. Effect size is defined as

$$e = \beta_j \omega_j^2 D[x_j], \quad \text{(Equation 12)}$$

where D is the standard deviation of feature across all cell lines.

SUPPLEMENTAL REFERENCES

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18*(*Suppl 1*), S96–S104.

Liscovitch, M., and Ravid, D. (2007). A case study in misidentification of cancer cell lines: MCF-7/AdrR cells (re-designated NCI/ADR-RES) are derived from OVCAR-8 human ovarian carcinoma cells. Cancer Lett. *245*, 350–352.

Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Appl. Stat. *25*, 257–265.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. *39*, 1–13.

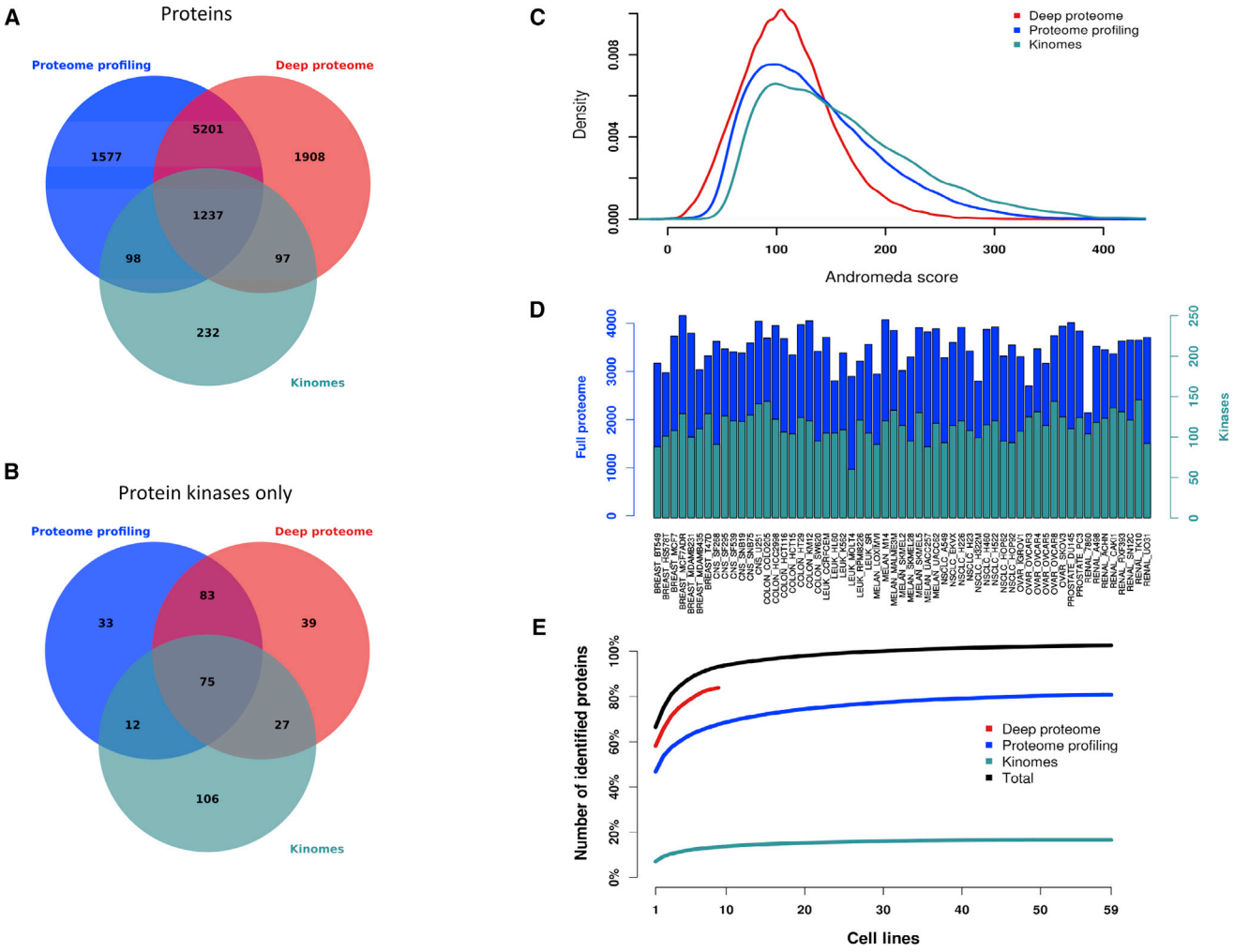

Figure S1. Protein Identifications, Related to Figure 1

(A) Venn diagram of proteins identified across all three experiments. The proteome profiling of 59 NCI-60 cell lines contributed 8,113 protein identifications. The deep proteome analysis of nine representative cell lines was performed on a more sensitive mass spectrometry (MS) platform resulting in increased proteome coverage and 8,443 proteins. Around 63% of the protein identifications overlapped with the proteome profiles of all NCI-60 panel cell lines.

(B) Venn diagram of protein kinase identifications. The kinase profiling approach based on the kinobeads technology identified 220 protein kinases and numerous ATP- and purine-binding proteins as well as protein kinase interaction partners (Bantscheff et al., 2007). The kinase profiling contributed 232 (2%) exclusive protein identifications including 106 protein kinases, which otherwise would remain undetected.

(C) Peptide identification (Andromeda) score distributions. Overall, 221,977 distinct peptides were identified. Average Andromeda identification score was 134, with only 7% below a score of 60. The average number of peptides per protein was 16 (median 10), leading to an average sequence coverage of 34%. Only 8% of the complete proteome was identified by a single peptide.

(D) Protein and kinase identifications in individual cell lines of the NCI-60 panel. Individual proteome profiles yielded 2,100 to 4,100 proteins per cell line (median 3,551) and kinobead experiments between 60 to 146 kinases.

(E) Number of proteins identified in increasing number of cell lines. Each individual cell line contributed to an incremental number of protein identifications. In total 10,350 non-redundant proteins were identified.

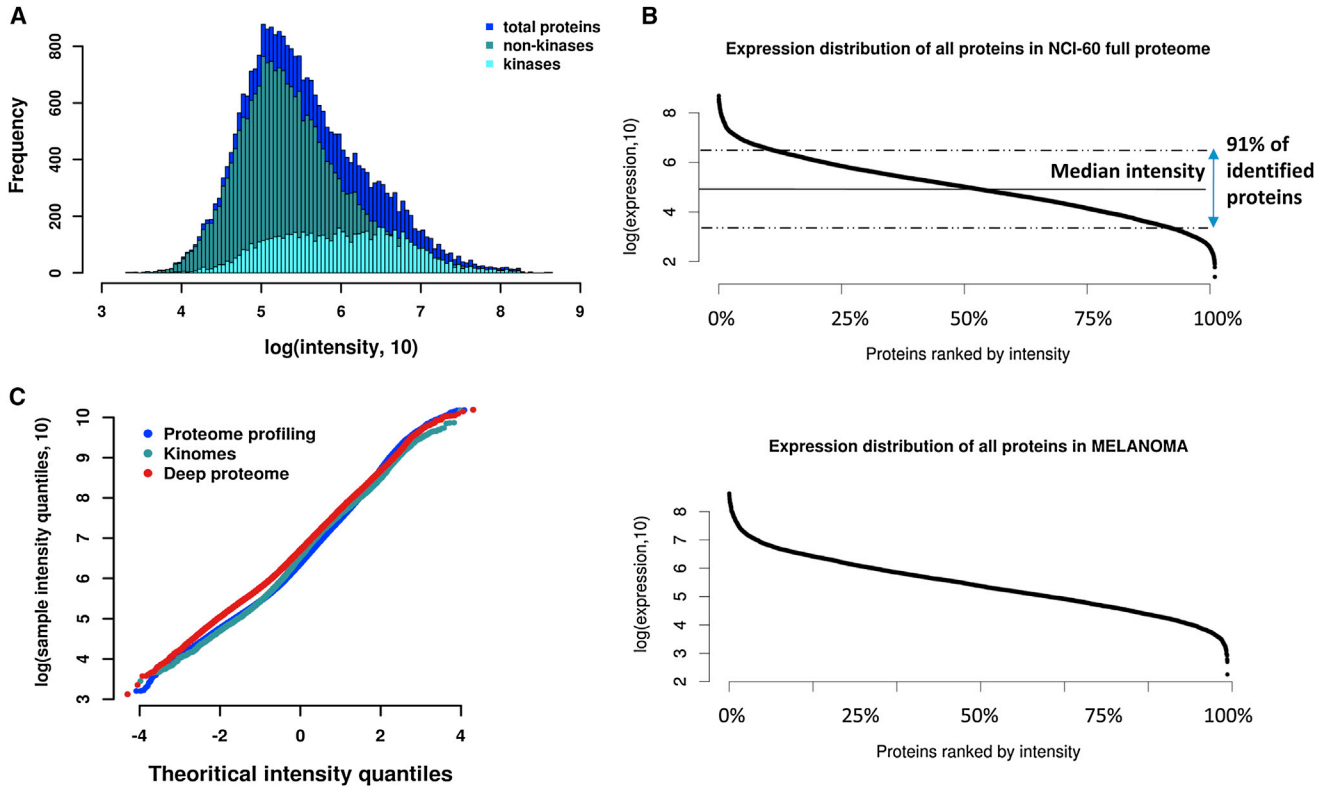

Figure S2. Protein Quantifications, Related to Figure 2

(A) Kinase MS signal intensities. Shown are MS signal intensity distributions for proteins identified from kinobead purifications. Kinobeads not only capture protein kinases but also binds a defined sub-proteome consisting of other ATP- and purine-binding proteins as well as protein kinase interactions partners (Bantscheff et al., 2007) and identified 232 proteins not covered by complete proteome approaches (Figures S1A and S1B). About 15% of all proteins identified on kinobeads are protein kinases. However, based on the total mass spectrometric signal, kinases account for around 50% of the total captured protein.

(B) Ranked expression distribution of all detected proteins. The rank plots show the median absolute expression value of each protein of all the NCI-60 cell lines (top) and an arbitrary tissue group (bottom; Melanoma). The ranked distribution of all individual proteins revealed that 91% of the quantified proteome is contained within a range of a factor of 10 above or below the median intensity.

(C) Normal quantile-quantile plot. Q-Q plot showing the intensity data distributions of NCI-60 cell lines on the vertical axes versus the standard normal distribution on the horizontal axes. The linearity of the data points suggest that the data have a normal (or close to normal) distribution.

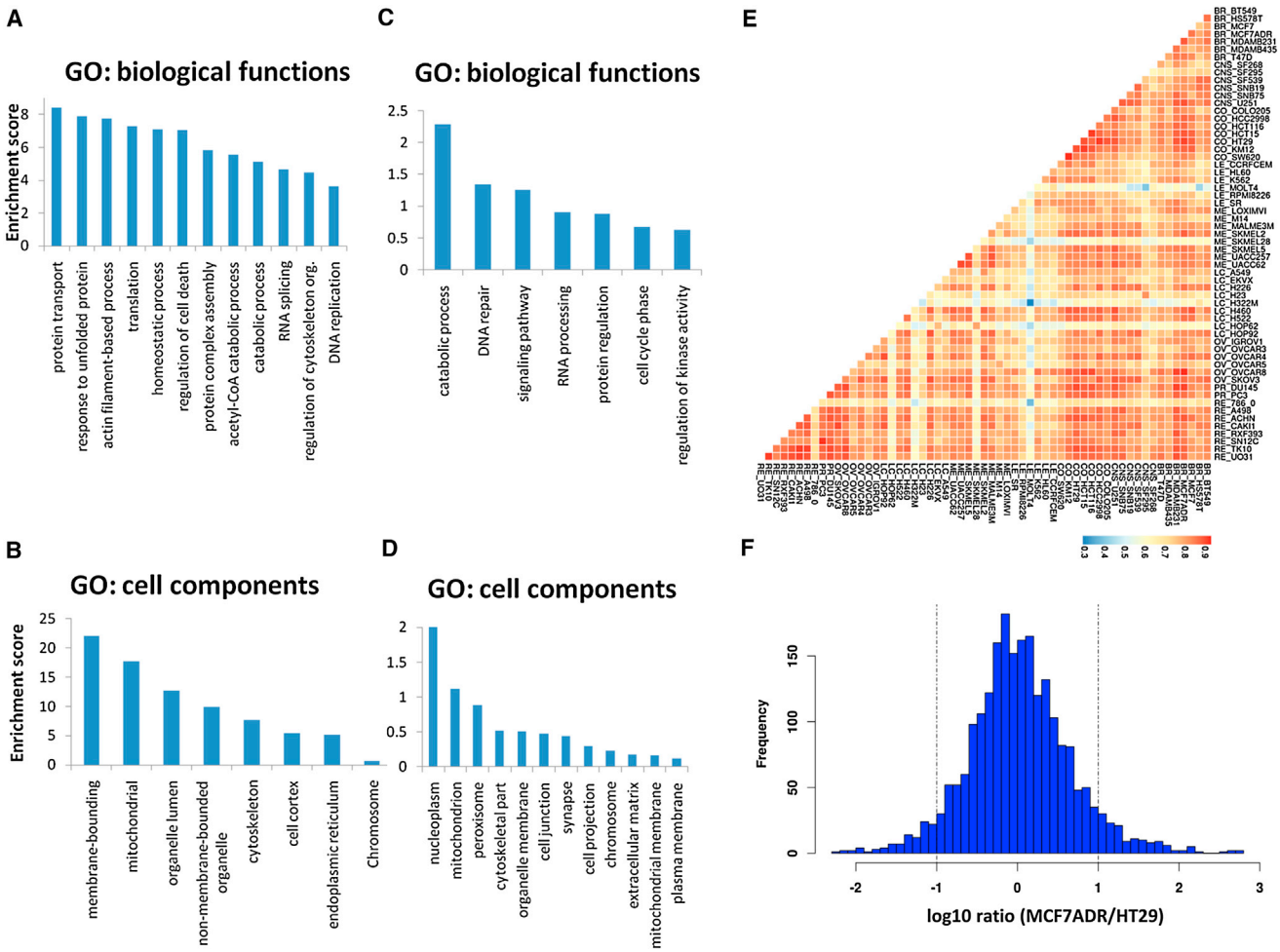

Figure S3. Gene Ontology Enrichment and Correlation Analysis between Cell Lines, Related to Figure 2

(A and B) Go categories significantly enriched among the 10% most abundant proteins of the NCI-60 panel.

(C and D) Go categories significantly enriched among the 10% least abundant proteins. Enrichment score of biological functions and cellular components is measured by modified Fisher exact test (Hosack et al., 2003).

(E) Protein abundance correlation of all versus all cell lines. The heatmap depicts the correlation of protein abundance between cell lines (red: high, blue: low). Overall, the correlation is in the range of 0.29 to 0.92. LE_MOLT4 and LC_H322M cells exhibit the lowest correlation (R = 0.29) whereas the proteomes of BR_MCF7ADR and OV_OVCAR8 were the most similar ones (R = 0.92). This is not surprising given that BR_MCF7ADR (originally named MCF-7/AdrR cells, later re-designated NCI/ADR-RES) was recently found to be derived from OVCAR8 (Liscovitch and Ravid, 2007).

(F) Protein abundance differences of two arbitrarily selected cell lines. This plot shows that the protein abundance of around 94% of all proteins of BR_MCF7ADR and CO_HT29 cells (expressed as logarithmic fold change) is within an order of magnitude (between −1 and 1 on log10 scale).

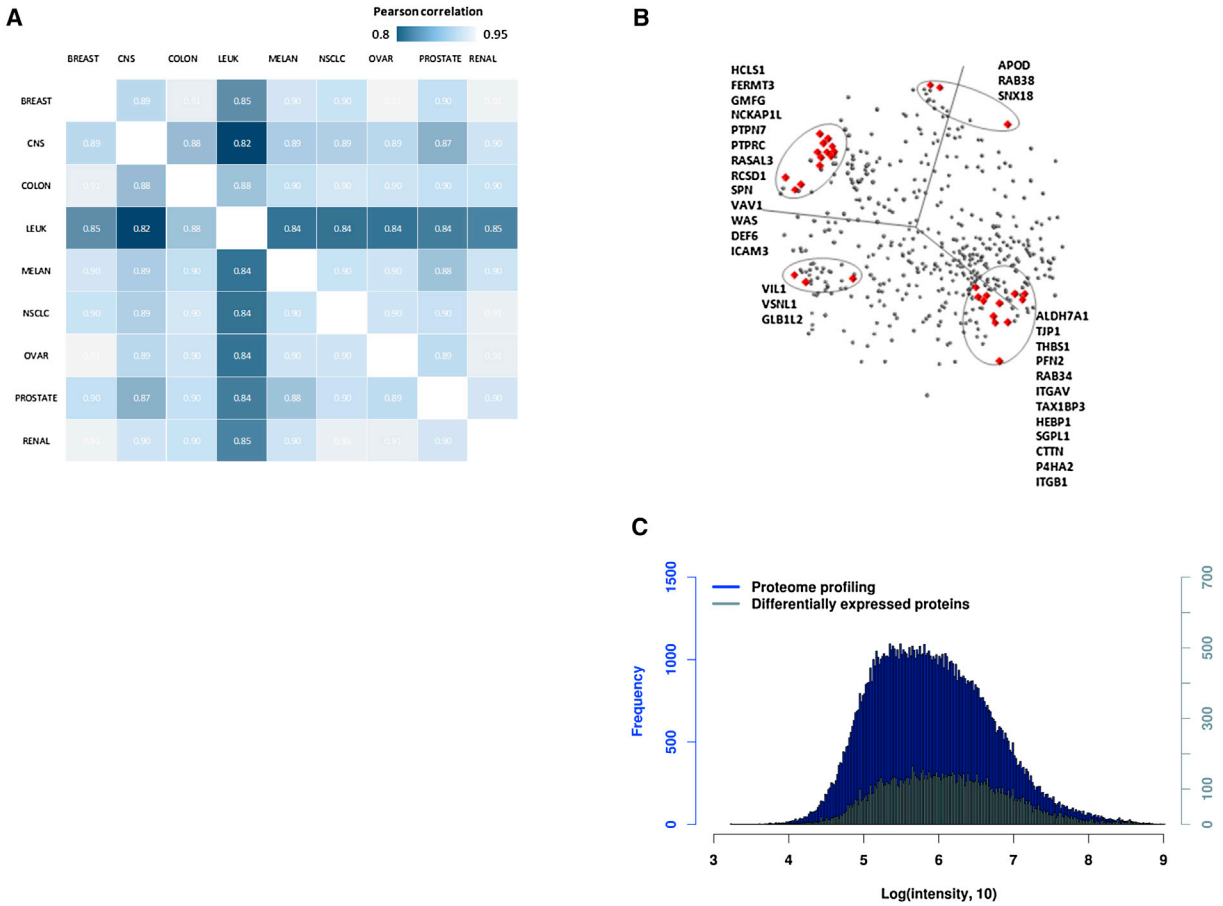

Figure S4. Protein Abundance Correlation between Tissue Groups and Differential Expression Analysis, Related to Figures 2 and 3

(A) Pearson correlation of average protein abundances highlights global differences and similarities between tissue groups of NCI-60 cell lines. In particular, leukemia cell lines ('LEUK') differ considerably from other tissue groups.

(B) Principal component analysis (PCA) of all differentially expressed proteins. The farther away from the centroid a protein is displayed, the higher the abundance in a particular set of cell lines. Top differentially expressed proteins in each direction are highlighted.

(C) Histogram of differentially expressed versus all identified proteins. Overall, this shows that the majority of differentially expressed proteins is of high abundance, which is consistent with the fact that these can be more accurately quantified across the entire NCI-60 panel.

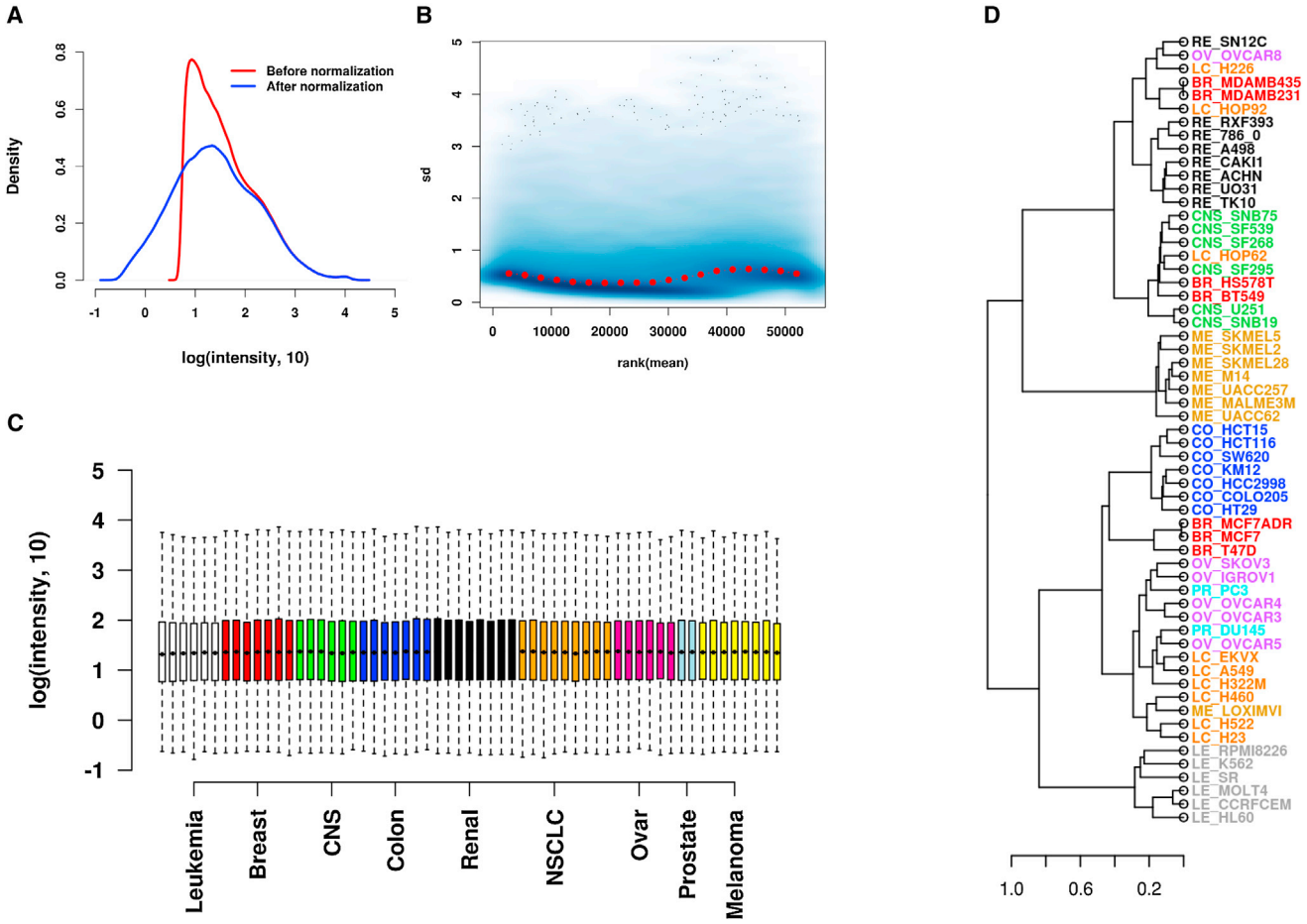

Figure S5. Comparative Analysis of Proteome and Transcriptome Profiles, Related to Figure 5

(A) Comparison of intensity data distributions of transcriptomic data. Shown is the histogram of gene intensities before and after normalization.

(B) Standard deviation versus mean relationship for the normalized NCI-60 data. Each data point represents a protein. The red dots depict the running median estimator (window-width 10%). Within each window, the median may be considered a pooled estimator of the standard deviation, and the curve given by the red line is an estimate of the systematic dependence of the standard deviation on the mean (Huber et al., 2002). After variance stabilization, this should approximately be a horizontal line.

(C) Box plots of the intensity distributions from the entire NCI-60 panel. Whiskers represent the most extreme data point.

(D) Unsupervised hierarchical clustering of cell lines based on 16,669 transcriptome profiles. Dendrograms show an average linkage hierarchical clustering using Spearman rank correlation with Ward metric. The corresponding HCL results for the proteomic data is presented in Figure 3C.

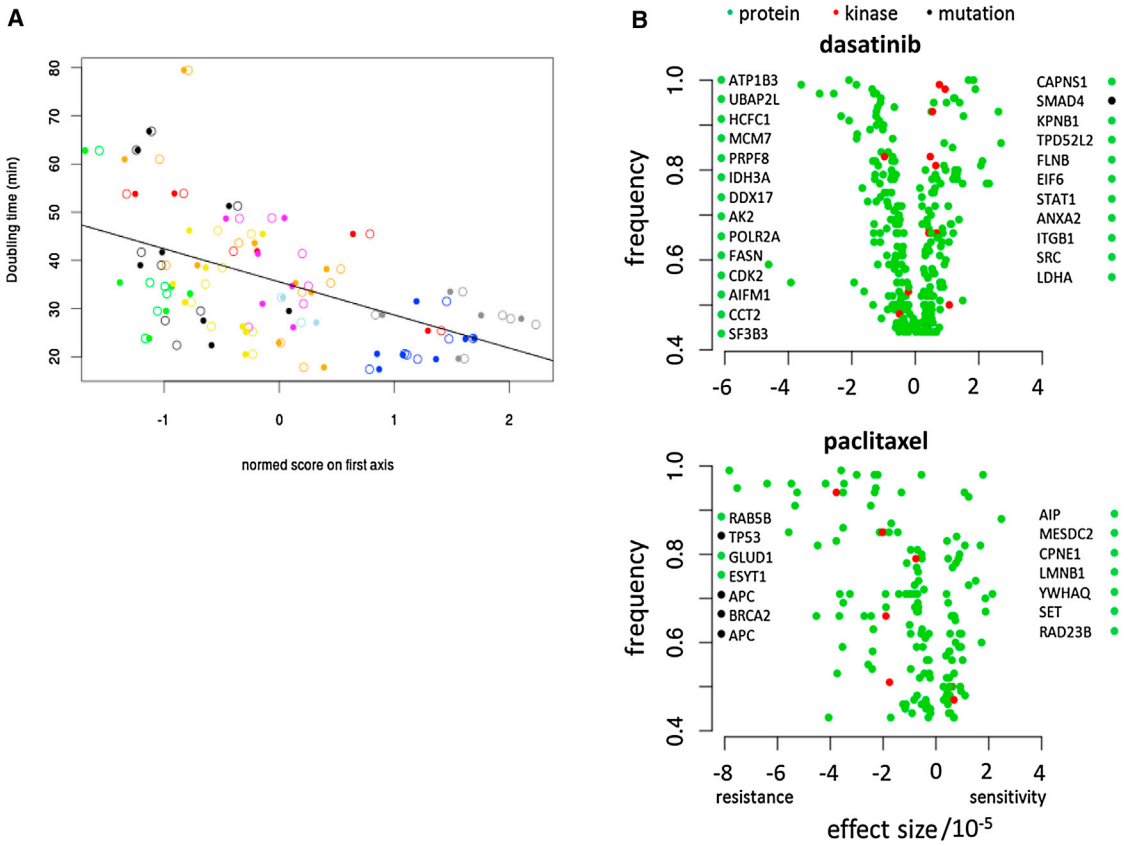

Figure S6. Coinertia Axes and Drug Feature Associations, Related to Figures 5 and 6

(A) Statistically significant relationship between doubling time and co-inertia analysis projection. Doubling times are plotted against the NCI-60 cell line co-ordinates of the horizontal axis of the CIA plot in Figure 5D. Filled circles represent mRNA sample space (Spearman rank correlation, p = 4.7E-05), open circles represent protein sample space (Spearman rank correlation, p = 6.4E-06). The CIA plot separates cell lines both in mRNA and protein space according to their growth rate (or doubling time). This correlation is statistically significant.

(B) Drug–feature associations identified by elastic net analysis. Drug–feature associations of dasatinib and paclitaxel identified by the elastic net procedure are plotted according to their frequency and effect size. Associations from mutation, proteome and kinome data are color-coded in black, green and red, respectively. Features outside the boundaries are sorted by frequency and plotted as outliers.