doi:10.1038/nature13319

Abbreviations

API	Application programming interface
CID	Collision induced dissociation
ESI	Electrospray ionization
FDR	False discovery rate
HCD	Higher energy CID
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
lincRNA	Long intergenic non-coding RNA
m/z	mass to charge ratio
MS	Mass Spectrometry
nESI	nano-ESI
NPC	Nuclear pore complex
PCA	Principle component analysis
PEP	Posterior error probability
PRM	Parallel reaction monitoring
PSM	Peptide spectrum match
PTM	Post translational modification
QTOF	Quadrupole time-of-flight
SRM	Selected reaction monitoring
TF	Transcription factor
TMT	Tandem mass tag
TUCP	Transcript of uncertain coding potential analysis of proteins in organs

Supplementary Methods

Source of human tissues and fluids

Tissue samples were obtained from the biobank of the Klinikum rechts der Isar and Faculty of Medicine of the TU München (BBTU). Healthy tissue samples were obtained by surgery and macroscopically dissected by an experienced pathologist. Tissue was snap frozen within the first 20 minutes after resection and stored in liquid nitrogen (-196°C) until usage. Before analysis, tissue type and tissue quality (no necrosis) was confirmed by a pathologist. For this purpose an HE stained slide was prepared from the actual sample to be used. Body fluids requiring no invasive procedures were provided by healthy volunteers. All patients/donors provided informed written consent, and the study was approved by the Ethics Committee of the Technische Universität München.

Preparation of protein extracts from tissues and body fluids

In order to generate a map of protein expression of all major tissues and organs in the human body ('human body map' project in ProteomicsDB), we analyzed 31 different tissues and body fluids from healthy volunteers (see above). An overview of all analyzed tissue and body fluid samples is provided in Supplementary Table 5. Fresh frozen tissue was thawed, cut into small pieces and extensively washed with precooled phosphate buffered saline. The tissue was homogenized in cold lysis buffer (50 mM Tris/HCl pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 150 mM NaCl, 0.8% Nonident P-40, 1 mM dithiothreitol [DTT] and 25 mM NaF) with freshly added protease (SIGMAFAST, Sigma-Aldrich, Germany) and phosphatase inhibitors (20 mM NaF, 1 mM sodium orthovanadate, 5 mM calyculin A; Sigma-Aldrich, Germany) using ceramic beads and an automated homogenizer (Precellys 24 Homogenisator, Peqlab, Germany) with 2x 10 second pulses at 5000 rpm. Homogenates were clarified by centrifugation for 15 minutes at 10,000 xg at 4°C and protein concentration was determined by the Bradford method. The supernatant was stored at -80 °C.

Approximately 30 µl of body fluids (cerumen, saliva, ascites etc.) were boiled with 2x NuPAGE sample buffer (Life Technologies, Germany) prior to LDS-PAGE and in-gel digestion.

Protein separation and in-gel digestion

100 µg protein extract from tissues or 30 µl from body fluid extracts were reduced and alkylated by 10 mM DTT at 55 °C for 10 min followed by the addition to 55 mM chloroacetamide and incubation at room temperature for 30 min and subsequently denatured at 95 °C for 10 min. Samples were then separated via a 4–12% NuPAGE gel (Life Technologies, Germany) and cut into 24 slices prior to in-gel digestion using trypsin (Promega, Germany), LysC (Wako Chemicals, Japan) or chymotrypsin (Roche, Germany). In-gel digestion was performed according to standard procedures [1] and peptides were analyzed by LC-MS/MS.

Chymotrypsin in-solution digestion and hydrophilic strong anion exchange chromatography (hSAX) peptide separation

In-solution digestion of testis protein extract was performed according to standard high urea procedures. Briefly, the lysate was denatured in 8 M urea, 0.1 M Tris/HCl, subsequently diluted to 2 M urea followed by protein digestion with chymotrypsin. After overnight digestion at 30 °C, peptides were concentrated and purified on C_{18} StageTips as described [2]. Peptide separation using hSAX was essentially performed as described [3] and the resulting 48 fractions (1 minute collection time per fraction) were dried down and subsequently analysed by LC-MS/MS.

Liquid chromatography tandem mass spectrometry (LC-MS/MS)

LC-MS/MS was performed by coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent, Dublin,

CA) to an Orbitrap Elite instrument (Thermo Scientific, Bremen, Germany). Peptides from an in-gel digest were delivered to a trap column (100 µm×2 cm, packed in-house with Reprosil-Pur C_{18}-GOLD 5 µm resin, Dr. Maisch, Ammerbuch, Germany) at a flow rate of 5 µL/min in 100% solvent A (0.1% formic acid, FA, in HPLC grade water). After 10 min of loading and washing, peptides were transferred to an analytical column (75 µm×40 cm, packed in-house with Reprosil-Gold C_{18}, 3 µm resin, Dr. Maisch, Ammerbuch, Germany) and separated at a flow rate of 300 nL/min using a 110 min gradient from 4% to 32% solvent B (solvent A: 0.1% FA, 5% DMSO in HPLC grade water; solvent B: 0.1% FA, 5% DMSO in acetonitrile). Both solvent A and B contained 5% DMSO to boost the nESI response and the MS signal of peptides [4]. The eluent was sprayed via stainless steel emitters (Proxeon) at a spray voltage of 2.2 kV and a heated capillary temperature of 275°C. The Orbitrap Elite instrument was operated in data-dependent mode, automatically switching between MS and MS2. Full scan MS spectra (*m/z* 360 – 1300) were acquired in the Orbitrap at 30,000 (*m/z* 400) resolution using an automatic gain control (AGC) target value of 1e6 charges. Tandem mass spectra of up to 15 precursors were generated in the multipole collision cell by using higher energy collisional dissociation (HCD) (AGC target value 2×10^4, normalized collision energy of 30%) and analyzed in the Orbitrap at a resolution of 15,000. Precursor ion isolation width was set to 2.0 Th, the maximum injection time for MS/MS was 100 ms and dynamic exclusion was set to 30 s. Internal calibration was performed on-the-fly using a DMSO-related lock mass (*m/z* 401.922718, $[C_6H_{10}O_{14}S_3]^+$)[4].

Measurements using an Orbitrap Velos used the same LC system, column and gradient as described and similar data acquisition parameters. The main differences are that the Orbitrap Velos used HCD of the top 10 precursor ions at an AGC target value of 3×10^4 and Orbitrap readout at a resolution of 7,500. Measurements using a LTQ Orbitrap XL used the same LC system, column and gradient as described and similar data acquisition parameters. The main differences are that the LTQ Orbitrap XL used CID fragmentation with normalized collision energy of 35% of the top 10 precursor ions at an AGC target value

of 5×10^3 and ion trap readout. Supplementary Table 5 provides an overview of which tissues were measured on which platform.

Retrieval of MS-based proteomics data from public repositories and other sources

Supplementary Table 1 provides a detailed overview of all projects and their data sources. Alongside published and unpublished data generated at the TU München (including tissue and body fluid proteomes, Supplementary Table 5), and datasets obtained from individual laboratories (Parag Mallick, Stanford School of Medicine; Hanno Steen, Harvard Medical School, Boston; Tamar Geiger and Mathias Mann, Max-Planck-Institute of Biochemistry, Martinsried; Shabaz Mohammed, Javier Munoz and Albert Heck, Utrecht University; Magnus Berle, Haukeland University Hospital, Oslo; Andrew Emili, University of Toronto; Roman Zubarev, Karolinska Insitute, Stockholm), MS-based proteomics datasets were downloaded from public repositories and servers including ProteomeXchange[5,6], PeptideAtlas[7,8], Tranche[9,10], MassIVE[11], CPTAC data portal[12,13], Broad Institute's proteomics FTP server[14], SCOR[15,16], and PRIDE[5,17]. Raw data files of all published studies and obtained from the public domain are made available through ProteomicsDB [4,15,18-70]. Two exceptions are the data sets from the CPTAC consortium because CPTAC requires permission to download/use their data and we want to make sure that CPTAC policy is respected. In addition, there is a small amount of data (10% of the total) that originate from (ongoing, unpublished) projects in the TUM lab and which are outside the scope of the current manuscript. These will also be made available for download once the respective manuscripts have been accepted for publication. Still, even for these projects, we show the peptides/spectra in ProteomicsDB for the purpose of protein identification (without disclosing experimental details and sample annotations).

Peptide identification using Mascot and Maxquant/Andromeda

Raw MS data files from Orbitrap-type instruments (including LTQ-FT) were processed in

parallel with two different data processing pipelines and search engines, Mascot [71] and Maxquant/Andromeda [72,73].

The following protein sequence databases were used for peptide identification: (1) the UniProtKB complete human proteome (download date: 05 Sep 2012; 86,725 sequences) contains the UniProtKB/Swiss-Prot complete human proteome (20,225 canonical sequences of protein-coding genes and 16,624 manually curated isoform sequences) and 49,876 UniProtKB/TrEMBL sequences, the latter representing all predicted protein sequences from Ensembl (except fragments) that were found to be absent from the UniProtKB/Swiss-Prot complete proteome (for a detailed description, see also [74,75]); (2) the cRAP database (common Repository of Adventitious Proteins; download date: 05 Sep 2012; 113 sequences) contains common laboratory proteins, proteins added by accident through dust or physical contact; and proteins used as molecular weight or mass spectrometry quantitation standards [76]. The UniProtKB complete human proteome and cRAP database were concatenated and every database search was performed against both databases.

In the Mascot workflow, raw MS data files were processed using Mascot Distiller (version 2.3.2, Matrix Science, UK) with processing parameters separately optimized for high and low resolution tandem mass spectra. Data processing comprised peak picking, de-isotoping and charge deconvolution of fragment ions. The resulting peaklist files were searched using the Mascot search engine (version 2.4.1, Matrix Science, UK) against the UniProtKB complete human proteome and the cRAP database. The target-decoy option of Mascot was enabled (on-the-fly search against a decoy database with reversed protein sequences) and search parameters included a precursor tolerance of 10 ppm and a fragment tolerance of 0.5 Da for CID spectra and 0.05 Da for HCD spectra. Enzyme specificity was set to trypsin, LysC, GluC, or chymotrypsin (as applicable), and up to two missed cleavage sites were allowed. The Mascot ^{13}C option, which accounts for the mis-assignment of the monoisotopic precursor peak, was set to 1 and the following variable modifications were included by default: oxidation of Met as well as acetylation of the protein amino-terminus.

Other variable and fixed modifications (such as phosphorylation of Ser, Thr and Tyr or acetylation of Lys as well as iTRAQ, TMT, SILAC or carbamidomethylation of Cys) were set as provided in the corresponding publications. Mascot search results were processed using the Mascot Percolator stand-alone software to obtain posterior error probabilities (PEPs) [77,78].

In the Maxquant workflow, raw MS data files were processed by Maxquant (version 1.3.0.3) for peak detection and quantification [72]. MS/MS spectra were searched against the same set of target and decoy databases as described for Mascot using the Andromeda search engine [73]. Proteases, variable and fixed modifications were specified as above. 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. No FDR cut-off was specified (see below), but a minimum peptide length of 6 amino acids was required. In few cases, where Maxquant (version 1.3.03) was unable to process the raw files, a newer version of Maxquant (version 1.4.05) was used.

In two cases, where raw MS data were unavailable (Burkhart_Blood_2012, Didangelos_MCP_2011), the available data was processed only through the Mascot pipeline, and, if required, peaklist files were converted into Mascot generic format (mgf) prior to database search. A special case represents the PeptideAtlas spectrum library (release December 2012) containing spectra of 253,693 distinct peptides from high and low resolution tandem mass spectra[60]. In order to make this resource available in ProteomicsDB, all spectra representing peptide identifications from high resolution MS and MS/MS measurements (including their respective modifications) were parsed and converted into mgf format and subjected to Mascot database search.

Identification of peptides of lincRNAs and transcripts of unknown coding potential (TUCPs)

In order to assess the presence and abundance of peptides representing translation products of lincRNAs and TUCPs, we searched 23 tissue datasets of the human body map (adrenal gland, anus, cervix uteri, esophagus, gall bladder, kidney, liver, lung, nasopharynx, oral cavity, ovary, pancreas, placenta, prostate, salivary glands, seminal vesicle, spleen, stomach, testis, thyroid gland, tonsils, uterus [post-menopause], uterus [pre-menopause]) as well as a deep HeLa proteome dataset[55] (excluding GluC data) separately against two lincRNA/TUCP sequence databases each appended to the human UniProtKB and cRAP database using Mascot with default parameters. The following two lincRNA/TUCP reference catalogues were used to search for translation products: (1) The Ensembl human non-coding map (GRCh37; download date: 02 Nov. 2013) comprises 13,564 non coding genes and (2) the Broad Institute's human body map of 21,487 lincRNAs and TUCPs (14,281 non coding transcripts as well as 7,206 novel transcripts with potential coding capacity, TUCPs) [79]. The Broad Institute utilized RNA sequencing and transcript abundance estimations to identify and characterise lincRNAs across 24 tissues and cell types. The BED files of Broad Institute's lincRNA transcripts were converted to FASTA using BEDTools [80]. The FASTA nucleotide sequences were translated to protein sequences by translating all three reading frames using custom python scripts.

To minimize the level of uncertainty in assigning identified peptides to lincRNAs/TUCPs, the identified peptides were as rigorously filtered as peptide identifications from standard database searches (see above) and, in addition, a Mascot delta score of >10 was required to assure that the best match for the spectrum is significantly better (10x) than the second best match[81]. Identified lincRNA peptides were further subjected to a BLAST search against all sequences in the UniProtKB complete human proteome database using parameters for short sequences (blastp -word_size 2 -matrix PAM30 -seg "no" -evalue 20000 -comp_based_stats 0). Peptides with identical sequence matches (and isobaric sequence variants thereof) in the UniProtKB complete human proteome were categorically rejected.

Identified peptides representing translation products of lincRNAs or TUCPs were imported into ProteomicsDB and the corresponding links are provided in Supplementary Table 3. To reduce redundancy on peptide and transcript level, identified transcripts from the Ensembl and Broad Institute's database search were grouped together according to their matched peptides as same-sets and sub-sets of peptide identifications.

Generation of reference peptide spectrum libraries

Reference peptides for uncertain protein-coding genes[82] were manually selected based on MS/MS evidence for these genes. Reference peptides for cytokines were selected for 361 proteins associated with the keyword 'cytokine' in UniProtKKB/Swiss-Prot using peptides present in SRMAtlas with lengths from 7 to 25 amino acids. Reference peptides for proteins with weak or no evidence in ProteomicsDB (at the time of synthesis) were selected based on the availability of PSMs or, for proteins not observed in ProteomicsDB, a set of peptides with good MS properties was derived by bioinformatic prediction using PeptideSieve[83].

Reference peptides for uncertain genes were synthesized by resin based solid-phase peptide synthesis at the authors labs at the TU München [84] or by SPOT synthesis technology [85,86] at the authors labs at JPT. After synthesis, the reference peptide sets were pooled (up to 1000 peptides per pool) and appropriately diluted in LC solvent A (1:100-1:10,000) before LC-MS/MS analyses using an Orbitrap Velos or Elite (see above). To generate reference spectrum libraries for both CID and HCD spectra, the peptide pools were analyzed at least once with each fragmentation type. Reference peptide spectra were processed using Mascot and Maxquant as described above and imported into ProteomicsDB. Available reference spectra and their links into ProteomicsDB are provided in Supplementary Table 2.

Data analysis in ProteomicsDB

ProteomicsDB utilize the main memory as the primary data storage. This reduces disk seek when querying data and, ultimately, results in faster data retrieval. Additionally, SAP HANA is specifically optimized for in memory operations which reduces the need of indexing tables and on top allows direct operations on compressed columns. This allows ProteomicsDB to show data and run queries using the entire database in real-time without the requirement of pre-assembled builds.

All relevant information from search result files of Mascot (dat file) and Maxquant ("combined/txt" folder, apl and res files) were imported into ProteomicsDB. In particular, ProteomicsDB comprises the up to top 10 best matches for each experimental spectrum both for Mascot and for Andromeda. Storing these PSMs allows to compute delta scores between the top two (or n) best matches which is useful, for instance, for computing false localization rates for phosphopeptides.

All projects imported into ProteomicsDB were manually annotated in order to provide high quality meta information such as project, experiment and sample descriptions in a computer-readable format. The annotations comprise experimental design of imported studies as well as biological and biochemical workflows. Where available, experiment annotations were translated into controlled vocabulary of open ontologies, remaining annotations were incorporated into a ProteomicsDB controlled vocabulary. The following ontologies are used in ProteomicsDB: Brenda tissue ontology (release date: 20 Dec 2012) [87], PSI-MS (version 3.44.0), UO (version 1.2) and sepCV (version 1.0). Identifiers, gene names, sequences etc. in ProteomicsDB are based on UniprotKB (download date: 05 Sep 2012), which, for instance, allows the localization of identified proteins to their chromosomal location (Fig. 2a).

False discovery rate (FDR) control using a global and a local FDR filter

Target and decoy peptide spectrum matches (PSMs; including lower ranking hits) were imported into ProteomicsDB irrespective of their scores or posterior error probability. In

order to control FDR on spectrum and peptide level, we applied a two-step process in which we first controlled the FDR at 1% for PSMs generated by each LC-MS/MS experiment using a global target-decoy approach[88]. Because this can still result in the retention of a considerable number of false matches, peptide identifications had to pass a length dependent Mascot or Andromeda score threshold of 5% local FDR as an additional filter (Fig. 1d and Extended Data Fig. E1 and E2). To this end, peptide identifications of the same length from all experiments in ProteomicsDB were sorted in bins of 1 score points and the target-decoy FDR was calculated for each bin individually ('local score and length dependent FDR'). Given the data in ProteomicsDB at the time of writing, this resulted in Mascot and Andromeda cutoffs for a 5% local FDR as provided in Supplementary Table 1. Comparison to 27 published high-throughput studies shows that our scoring scheme is in line with the often-used 1% protein FDR criterion (Supplementary Table 1) and avoids one unsolved issue of target-decoy searching that artificially high protein FDRs are generated when analyzing very large data sets [89]. See below for a discussion on protein FDR.

Analysis of protein expression

For a detailed comparison of five different label-free protein quantification approaches and normalization, see Supplementary Notes. Protein abundance was estimated for UniProtKB/Swiss-Prot sequences using the iBAQ approach (intensity based absolute protein quantification) [90]. To this end, peptide intensities for a given protein in a sample were obtained from Maxquant, summed up and divided by the number of observable peptides (length 6 to 30, no missed cleavage). In order to be able to compare protein abundances across multiple samples, experiments and projects, the iBAQ protein intensities were normalized based on the total sum of all protein intensities. Unless otherwise stated, the displayed protein abundance values were $\log_{10}$ transformed and right-shifted by 10 $\log_{10}$ units into positive numerical space. It is important to note that ProteomicsDB aggregates any isoform specific abundance information at the gene locus

level. Isoform-specific abundance information is currently not calculated as isoforms are usually under-sampled in MS-based proteomics. Protein expression analysis including normalization, hierarchical cluster analysis and principle component analysis was performed using R (v 2.12.1; [91]). Cluster analyses using a variety of common algorithms and metrics were performed to group the tissues and body fluids on the basis of protein expression pattern.

Principle component analysis (PCA) of proteome profiles of tissues and cell lines (Fig. 3b, upper panel) was performed on 6094 proteins showing significant expression differences (ANOVA, Benjamini-Hochberg adjusted *p*-value of 0.05) between the tissue groups (ovary: 3 tissues, 4 cell lines; colon: 3 tissues, 8 cell lines; kidney: 1 tissue, 8 cell lines; lung: 4 tissues, 11 cell lines). PCA of proteome profiles of tissues samples (colon: 3; breast: 2; liver: 2; kidney: 1; ovary: 3; lung: 4; prostate: 2) was performed on all 10,710 proteins (Fig. 3b, lower panel).

Hierarchical clustering of the top100 proteins per tissue (Fig. 3c, Supplementary Table 4) as well as protein kinases and transcription factors (TF; Fig. 4a, Supplementary Table 4) was performed on log transformed normalized iBAQ intensities using euclidean distance and complete linkage. For the analysis of tissue-specific kinase and TF expression, human protein kinases and TFs were retrieved from UniProtKB, resulting in protein expression profiles for 310 protein kinases (Pfam protein kinase domain), 557 transcription factors (GO term ‘transcription factor’) and 39 proteins annotated as kinases as well as transcription factors (Supplementary Table 4).

Tissue-specific proteins (Fig. 4b, Supplementary Table 5) were calculated based on normalized iBAQ values for 47 tissues. Median protein expression values were used in all cases where multiple datasets for the same tissue were available. Proteins were considered as tissue-specific if their expression in a given tissue was at least ten-fold higher than the average expression in the remaining tissues. Classification and functional enrichment analysis of protein lists (such as the core proteome, missing proteins or tissue

specific proteins) were performed using the DAVID Bioinformatics Database[92] as well as ReviGO[93].

The stoichiometry of protein complexes (nuclear pore complex [NPC], core proteasome; Fig. 5c, Extended Data Fig. E9, Supplementary Table 8) was reconstructed using normalized iBAQ values, and complex stoichiometry was calculated relative to either a particular subunit (nuclear pore complex: subunit Nup43) or the median subunit expression across all subunits (core proteasome). NPC compositions are based on data of triplicate shotgun experiments of HeLa nuclear extracts from a recent study of Ori *et al.* [56], and the proteasome composition was determined for 109 different samples (29 tissues from the human body map, 80 different cell line samples from two recent studies profiling the proteomes of 11 and 59 cell lines, respectively [50,94].

Genome-wide comparison of protein and transcript abundance levels across twelve tissues

For the systematic genome-wide comparison of protein and mRNA abundances across multiple tissues (Fig. 5a, Supplementary Table 6), we downloaded quantitative transcriptomics data (RNASeq) from Human Protein Atlas[95] (download date: Nov 11, 2013). Transcript expression data (abundances expressed as fragments per kilobase per million, FPKM) was extracted from the RNA-Seq data for 12 tissues (adrenal gland, esophagus, kidney, ovary, pancreas, prostate, salivary gland, spleen, stomach, testis, thyroid gland, uterus). Normalized iBAQ values were exported from ProteomicsDB for theses tissues. In cases where multiple full proteomes were available for a single tissue or organ, the median protein expression values were used. Proteins were mapped to transcripts using BioMart [96] resulting in 6104 transcripts/proteins with expression data. For the comparison of protein and transcript abundances, protein and mRNA expression was re-scaled using z-score transformation (Extended Data Fig. E7a). Translation rate and transcript abundance are the major determinants of protein expression, while transcript and protein half-life only play a

minor role [90]. We calculated the ratio between mRNA and protein abundance for all 12 tissues as a proxy for the amount of protein which can be synthesized from a given mRNA. The median ratio across all 12 tissues was then used to predict protein abundance from mRNA expression. The mRNA and protein expression values as well as median protein/mRNA ratios are provided in Supplementary Table 6.

Elastic net analysis

In order to predict proteins involved in drug resistance and sensitivity, we modelled proteome and drug-response profiles for 24 FDA-approved drugs and 35 cell lines to identify known and potential protein markers for drug sensitivity and resistance. As described previously [50], we employed elastic net regression [97] using full cellular proteome data of 135 experiments corresponding to 35 cell lines and 10,825 proteins (Supporting Table 7). In cases where multiple full cellular proteomes were available for a single cell line, the median protein expression values were used. Drug activity levels were obtained from the Cancer Cell Line Encyclopedia (CCLE) resource [98]. All features were regressed to fit a Gaussian model of drug activity area for each drug.

Elastic net definition

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

$$\sum_{i=1}^{n} X_{ij} = 0 and \sum_{i=1}^{n} X_{ij}^{2} = 1 \quad [1]$$

Naïve elastic net criterion was then defined by

$$L(\lambda_1, \lambda_2, \beta) = \left|y - X\beta\right|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| \quad [2]$$

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

$$\hat{\beta}\ NaiveElasticNet = \arg\min_{\beta}\{L(\lambda_1, \lambda_2, \beta)\} \quad [3]$$

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

$$\hat{\beta}\, ElasticNet = (1+\lambda_2)\,\hat{\beta}\, NaiveElasticNet \quad [4]$$

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

$$\alpha = \frac{\lambda_1}{(\lambda_1 + \lambda_2)} \quad [5]$$

Then the elastic net penalty can be defined as

$$(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}\left|\beta_j\right| \quad [6]$$

Implementation

Elastic net analysis was conducted by the R package "glmnet" [99,100], 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 dataset were randomly selected to identify protein markers. The procedure was repeated 1000 times for each drug. The final signature of markers for a drug consisted of all features that frequently appear in any of the 1000 runs and represent a consistent weight in different signatures. The weight was defined as

$$\omega = \frac{F[\beta > 0] - F[\beta < 0]}{N} \quad [7]$$

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 drug activity area, sensitive marker) or -1 (negatively correlated with activity area, 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

$$e = \beta_j \omega_j^2 D[x_j] \quad [8]$$

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

Supplementary Notes

Building the human proteome

The human genome represents a (more or less) static definition of the potential protein inventory of every cell in the human body (in health and disease). In contrast, the proteomes of cells are very divers and highly complex. Protein expression typically spans 4-5 orders of magnitude (see below and [21]) in cell lines (and presumably tissues) and more than 10 orders of magnitude in body fluids [101]. The molecular complexity is even higher as proteins are often expressed as splice variants and/or processed and chemically modified on the co- or post-translational level.

In order to identify at least one protein per protein coding gene, the strategy towards a mass spectrometry based draft of the human proteome focused on the assumption that the selection of datasets has to cover the breadth and depth of the protein repertoire in human samples. To this end, we selected studies on (mostly cancer) cell lines which analyzed in depth a single cell line or a large number of divers cell lines (e.g. [50,94]). To capture tissue and body fluid specific proteins, we analyzed 31 tissues and body fluids in-house and added similar datasets from the public domain, often from rarely analyzed or not readily accessible anatomical components. To address low abundance proteins such as kinases or transcription factors, we included ~1,300 affinity purifications including kinome enrichments[50], HDAC enrichments[20] and protein-protein-interaction studies[43]. Similarly, we included numerous PTM studies, which increased the coverage of proteins that are highly modified and/or of low abundance (e.g. [37,49]). However, with the progression of assembling data in ProteomicsDB, it became clear that numerous protein groups were underrepresented (e.g. membrane proteins, keratin associated proteins, cytokines), and we hence tuned the focus on datasets which may fill these gaps. For instance, to increase the chance of identifying keratin-associated proteins as well as other proteins with only few tryptic peptides, we re-analyzed numerous tissues (including hair follicles) using proteases different from trypsin (namely chymotrypsin and/or LysC). To further increase the coverage

of the human proteome, we pursued a data-driven approach and systematically assembled publically available and in-house data from experiments in which missing protein coding genes have been identified in previous studies (e.g. 'Cellzome_adopted' project in ProteomicsDB).

False-discovery rate control in ProteomicsDB

To control false-discovery rate across all samples in ProteomicsDB and find a reasonable compromise between sensitivity and specificity of the filtered data, we applied a two-stage filtering approach based on a classical target-decoy search strategy. The results of the two-stage filtering approach for the two separate search engines Mascot and Maxquant/Andromeda are depicted in Fig. 1d (Mascot only) and Extended Data Fig. E1.

The first filtering stage comprises the straightforward removal of all PSMs below 1% PSM FDR on the level of single LC-MS/MS experiments (Extended Data Fig. E2a). The second filtering step is performed on the aggregated data in ProteomicsDB (and not individual samples or experiments) and in fact comprises more than one element (Extended Data Fig. E2b). First, we categorically reject peptides shorter than 7 amino acids. Second, we consider every peptide length individually in the FDR calculation. We found this to be necessary because PSMs for short peptides contribute a lot more false positive identifications than longer peptides. Third, within each length bin, we apply a 5% local (!) FDR that rejects all PSMs below the respective search engine score threshold making sure that low scoring PSMs are removed (we note that this is much more stringent than a global FDR criterion). The resulting 5% local length and score dependent thresholds are depicted in Fig. 1d and Extended Data Fig. E1 and provided in Supplementary Table 1. As an example, an Andromeda score of 100 (35 for Mascot) is required for peptides of length 7. This threshold is a high hurdles for a 7 amino acid peptide and, as a result, the majority of all PSMs of length 7 are in fact rejected. Collectively, the above measures filter out a lot of false positives that would creep in using the classical global FDR approach.

We refrained from calculating protein FDR measures as the concept of a protein FDR is problematic (see below). Instead, we used the above filtering criteria and compared the results to many (i.e. 27) reports from the literature that report 1% protein FDR data. This comparison shows that the results are consistent (Supplementary Table 1).

Protein false-discovery rate

The estimation of protein false discovery rates in very large proteomics experiments is a challenge for which no satisfactory solution has been found yet. We discuss this in three sections below: i) a brief account of the current state of the discussion in the field on the topic of protein FDR, ii) the presentation of data showing that the classical protein FDR approach does not work for large data collections and iii) plus a new idea that might alleviate the issue in the future.

i) State of protein FDR discussion in the field

We have discussed the matter extensively with numerous colleagues in the field and the opinions on how to approach the challenge of estimating protein FDR vary a lot. One view (perhaps extreme but shared by many) is that protein FDRs are actually not very meaningful because proteomics measures peptides not proteins and the definition of a 'decoy protein' is quite problematic (for an interesting thought experiment please see http://www.matrixscience.com/blog.html). Another (perhaps also extreme) view is that one should lower the peptide/PSM FDR to whatever degree necessary to reach a desired protein FDR (e.g. [60]). The downside of this approach is that a lot of valid identifications are removed rendering this approach more conservative than necessary). Then there is the 'middle ground' represented by e. g. the MAYU and SPIRE approaches (or the combination thereof called IPM [102]) that attempt to overcome the scaling issues associated with the classical protein FDR approach. Unfortunately, IPM and similar approaches are also not as independent of scale as one might have hoped (data not shown).

ii) Failure of classical protein FDR for large data sets:

The principle issues with the concept of a protein FDR aside (see above), estimating protein FDRs in large data sets is difficult. The rapidly saturating number of true positive identifications and the slow but steady growth of false positive identifications when aggregating more and more experiments leads to the paradoxical situation that, eventually, all true and false positive proteins will have been observed resulting in a protein FDR of up to 100%. In other words, performing say 1,000 experiments of very high individual quality would lead to an aggregated data set of very poor quality. Common sense would rightfully state that both cannot be true at the same time. As there is no reason to suspect that mass spectrometry based proteomics is fundamentally flawed (as countless publications illustrate the power of the technology in measuring biological systems), the paradox most likely arises from shortcomings of the classical protein FDR model itself. We investigated this issue in some detail (Extended Data Fig. E2c) and the data show that the classical FDR model does not scale and is therefore inappropriate for the estimation of protein FDR in large data sets. Briefly, the aggregation of a few data sets initially leads to an increase of protein IDs at a particular protein FDR of say 1% (perhaps arguing that the classical protein FDR model is appropriate for relatively small studies). However, when adding further experiments, the number of protein IDs at 1% FDR begins to drop and continues to do so with an unclear endpoint. One would have to conclude from such a behaviour that proteins that were confidently identified early in the data aggregation would actually be false (because they do not survive in the subsequent aggregation steps). This would either mean that MS based proteomics is a flawed technique for protein ID (very unlikely, see above) or that the protein FDR model is invalid (more likely).

iii) Inspired by these results, we investigated another approach for estimating protein FDR ('picked' target-decoy approach)

The 'picked' target-decoy approach is based on the observation that when aggregating hundreds to thousands of proteomic experiments, all target and decoy entries in a given

protein sequence database eventually accumulate PSMs (Extended Data Fig. E2c). Briefly, in the 'picked' target decoy approach, for each identified target and decoy protein hit the sum of the scores (Mascot or Maxquant) of the unique best scoring peptides was calculated. Subsequently for each protein the sum score of its target version was compared to the sum score of its decoy version. If the sum score of the decoy version was higher than that of the target version the protein was counted as a decoy hit and the target version was disregarded. If the sum score of the target version was higher than that of the decoy version the protein was counted as a target hit and the decoy version was disregarded. Finally the protein FDR was calculated using the remaining target and decoy hits in the same way as for the classical target-decoy approach. This approach is similar to the decoy fusion approach proposed recently [103] and automatically corrects for the propensity of large proteins to accumulate more random hits than small proteins. It is also completely symmetrical and does not in any implicit way favor decoy or target hits.

Although the approach looks promising in terms of dealing with the scaling issues of protein FDR estimation (as the loss of proteins with aggregating more and more data can be stopped), we are not yet confident enough at this stage to propose this new idea as a generally applicable protein FDR estimate. This is because this new model does not address the conceptual issues of a protein FDR (see discussion above) and we cannot be sure yet that we understand all possible parameters that might influence the outcome. One further angle from which the problem could be investigated is the multiplicity of protein identifications between experiments. Current methods for estimating protein FDR do not properly (or at all) acknowledge the fact that true positive identifications often reproduce between experiments while false positive identifications do not. Reproducibility of a protein observation should constitute a strong benefit but how this can be best incorporated in protein FDR estimates is not yet clear. Yet another approach would be the systematic use of synthetic peptide reference standards for all (proteotypic) peptides of a protein, its isoforms and post-translational modifications [104]. This would require the synthesis and measurement of approx. 2 million peptides, which is entirely feasible using today's peptide

synthesis and measurement technology. For now, we refer the issue to future research but the large quantities of data in ProteomicsDB and similar databases will prove useful for this purpose.

Estimating and normalizing protein abundance in ProteomicsDB

Protein abundance in cells is often expressed as an approximated measure of concentration, such as copies per cell. However, this metric requires detailed knowledge about a couple of parameters including but not limited to cell number, size and morphology, and cannot be readily translated to tissues, extracellular structures or body fluids for which other concentration measures are required. One pragmatic approach to overcome this issue is to express the abundance of a protein as a fraction of all proteins detected and quantified in a sample. This normalization on the sum of all protein abundances avoids the difficulties associated with concentration measures. However, this approach is not particularly sensitive to the scaling of copy numbers by different parameters such as cell volume, shape, number, and potentially many others. This may lead to some inaccuracies, i.e. overestimating/underestimating proteins that have some relationship with a particular parameter (e.g. histones scaling with the number of genomes, ribosomes scaling with the cell volume). Nonetheless, this pragmatic approach enables the straightforward comparison of protein expression across a wide range of different types of samples.

In order to provide a consistent and reasonably accurate estimation of protein abundance enabling the comparison of protein expression between samples of an experiment or project as well as across multiple different projects, we systematically investigated five intensity-based approaches for the estimation of protein abundances. We did not consider spectral-counting based approaches as these are associated with less accurate protein abundance estimations [105,106].

The metrics investigated in detail are top3 intensity [107], top3 divided by the number of observable peptides, iBAQ [90], average intensity [108] and sum of all peptide intensities. We

compared the performance of the five protein abundance metrics with accurate copy number estimates from a large-scale absolute protein abundance data set (based on the AQUA technology). To this end, we retrieved reference protein copy numbers for U2-OS cells from Beck *et al.* [21], and calculated copy numbers for the five abundance metrics using a large-scale U2-OS cell line data set acquired by Geiger *et al.* [94] (Extended Data Fig. E5a). While all five metrics perform reasonably well (median fold error <2.5), the two approaches which take the number of observable peptides into account (iBAQ, top3 intensity divided by the number of observable peptides) are in better agreement with the original copy number estimates (Extended Data Fig. E5e). Note that the observed median fold errors are in line with the quantitative accuracy estimated by Beck *et al.* using a bootstrapping approach for their AQUA quantified proteins. When comparing the quantitative accuracy of the two most frequently employed approaches, iBAQ and top3, as a function of the number of quantified peptides (Extended Data Fig. E5f) as well as protein length (Extended Data Fig. E5g), it becomes obvious that iBAQ is slightly less biased for low abundant proteins and small proteins than the top3 approach.

To further investigate the effect of the normalization based on the sum of all protein abundance measures, we compared two data sets with considerably different proteome coverage (e. g. samples measured on different instruments). We analyzed which combination of metric and normalization appropriately re-scales the protein expression values using a set of nine different cell lines analyzed on two different mass spectrometers in the study of Moghaddas Gholami *et al.* [50]. As exemplified for the Colo-205 cell line dataset in Extended Data Fig. E5b for all five protein abundance metrics, the differences in intensity distribution before normalization primarily reflect the differences in sensitivity of the instruments. The chosen normalization ensures that what is an abundant protein in one data set is also an abundant protein in the other (by virtue of each protein being expressed as the fraction of the total; Extended Data Fig. E5c). The main difference after normalization is thus the number of identified proteins rather than the resulting (relative) abundances. The histograms show that the data is very well aligned on the high abundance side of the

histogram. For the low abundance proteins this visualization is somewhat misleading because the x-axis is in log10 scale, so the right half of the histogram actually covers >99% of the total intensity. The Q-Q plots for the same data are a more appropriate visualization for this purpose and show that the data is well aligned over approximately 4.5 orders of magnitude (Extended Data Fig. E5d). As a consequence, one should confine the interpretation of the proteome profiles to abundances of >500-1,000 copies per cell (or an appropriate equivalent measure for body fluids), which is in line with e.g. the study of Beck *et al.* [21]. Extended Data Fig. E6 a-c depicts the remaining eight cell line proteomes using iBAQ quantification before and after normalization (a) and the corresponding Q-Q plots (b).

Following the rationale that normalization should enable the comparison of disparate data sets, including the comparison of samples, experiments and projects based on label-free and stable isotope labeling quantification, we further extended our analysis and investigated how the normalization affects the comparison of label-free and label-based approaches. Note that the MS1 intensity of light, medium or heavy labeled peptides was utilized to calculate the protein abundance estimate for every SILAC or dimethyl labeling channel separately. Similarly, for isobaric quantification, the corresponding MS2 reporter ion intensity (for iTRAQ or TMT) was used to calculate protein abundance estimates for the quantification channels separately. As exemplified in Extended Data Fig. E6c based on the comparison of data from SILAC-labeled and label-free analyzed MCF-7 cell digests from two different studies [34,94], the protein abundance of MS1 based label-free and label-based quantification approaches exhibit a similar dynamic range of 4-5 orders of magnitude and can be as comparably re-scaled. In contrast, as exemplified using data from MCF-7 cell lines based on label-free [94] and iTRAQ quantification [42], quantification based on MS2 reporter ion signals shows drastically different intensity distribution characteristics with respect to dynamic range and symmetry of the distribution (Extended Data Fig. E6d). We concluded that, although desirable, the quantitative comparison of protein abundances between MS1 (label-free, SILAC and dimethyl labeling) and MS2 based approaches (iTRAQ, TMT) is not possible without introducing gross errors (or re-scaling artefacts). In

light of these results, we decided to keep MS1 and MS2 quantification throughout this study as well as in ProteomicsDB separately. We also keep affinity data separate from shotgun data as the data structures are not comparable. In addition, we allow users to separate data from tissues/cell lines/body fluids in the expression analysis of proteins in ProteomicsDB.

The distribution of 347 samples depicted in Extended Data Fig. E6e underscores that the re-scaling using total sum normalization actually results in very similar abundance distributions for samples with MS1-level quantification.

Last, but not least, a commonly employed approach to illustrate that derived protein abundance metrics actually reflect protein copy numbers is to compare the abundance of proteins with a known stoichiometry within complexes. To this end, we compared normalized iBAQ values with the reported composition of nuclear pore complexes from a recent study [56] (NPC). As depicted in Extended Data Fig. E9a, the stoichiometry of the NPC components derived from normalized iBAQ values of triplicate measurements of HeLa nuclear extracts from the same study are in very good agreement with the reported stoichiometries from a corresponding AQUA experiment (median fold error < 32%). Note that the components of the stable Nup107 sub-complex even show a median fold error of less than 10%.

In summary, while the assumption behind our normalization approach will affect the accuracy of the data is detail, the results of the technical analysis as well as the biological data shown in Fig.s 3, 4 and 5 indicate that the overall approach is appropriate and enables a reasonable biological interpretation of the protein expression profiles.

Analysis of protein expression and drug sensitivity

We have previously shown that protein expression can be correlated to drug sensitivity [26] and applied it here to discover sensitivity/resistance markers for 24 drugs in 35 human cancer cell lines using drug sensitivity data provided by the cancer cell line encyclopedia (CCLE)[38]. For instance, analysis of the EGFR kinase inhibitors Erlotinib and Lapatinib

identified a number of common proteins associated with drug sensitivity or resistance (Fig. 5b, Extended data Fig. E8, Supplementary Table 7). The primary target (EGFR) as well as annexin A1 (ANXA1, a direct EGFR substrate), and EGFR interacting proteins at stress fibers (PDLIM, KRT5, KRT14) all indicate drug sensitivity while high expression of ANXA6 or S100A4 renders cells less responsive. Consequently, knock-down of ANXA6 in BT549 cells has been shown to sensitize cells for lapatinib [39] and addition of S100A4 to cells in culture has been shown to stimulate EGFR and to promote metastasis [40]. We note that high expression of S100 proteins is often associated with resistance against kinase inhibitors (Supplementary Table 7), suggesting that S100 overexpression may be a general molecular resistance mechanism. The data further suggest that similar effects can be postulated for the Zn-finger protein THAP2, the NAD(P) transhydrogenase NNT and Dermcidin (DCD). Likewise, high expression of MED11 (part of the RNA Pol II mediator complex), IFI35 (an interferon induced protein of unknown function), HECTD1 (a E3 ubiquitin ligase) and CDRT1 (an orphan F-box protein) should promote drug sensitivity but their molecular connections to EGFR are not clear at present. In light of a recent report showing increased phosphorylation of HECTD1 upon EGF treatment[41], it is tempting to speculate that a HECTD1/CDRT1 complex may be involved in regulating the stability of EGFR via the ubiquitin/proteasome system.

Calculation of experimental proteotypicity

Experimental proteotypicities for peptides are provided in Supplementary Table 9. Briefly, experimental proteotypicity of a single peptide counts how often this particular peptide was observed when the protein was identified. The cumulative proteotypicity describes how often a protein was identified using either one of the top most frequently identified peptides. Example: one particular peptide may have been observed in 50% of all cases that the protein was identified. Another peptide may have also been observed in 50% of all cases. However, these two peptides are not necessarily observed together. The cumulative

proteotypicity of the two peptides may therefore vary from 50% (both peptides observed together every time a protein is identified) to 100% (both peptides never observed together when a protein is identified). The experimental proteotypicity of a single peptide is very useful e.g. for selecting peptides for SRM/MRM experiments. The cumulative proteotypicity can be used in a similar fashion (i.e. deciding about how many AQUA peptides to synthesize for a given protein) but also offers an explanation why e.g. the top3 intensity method (see above) works well for quantifying a protein (because it turns out that very few peptides are responsible for almost all identifications of a particular protein). An additional use is that the detection of a proteotypic peptide gives confidence in protein identifications based on single/few peptides (either because a particular protein may not produce more than one/few MS compatible peptides upon protease digestion or the protein is of low abundance in which case the detection of a proteotypic peptide is much more likely than the detection of any other peptide).

Supplementary References

1 Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Analytical chemistry* **68**, 850-858 (1996).

2 Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature protocols* **2**, 1896-1906, doi:10.1038/nprot.2007.261 (2007).

3 Ritorto, M. S., Cook, K., Tyagi, K., Pedrioli, P. G. & Trost, M. Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes. *Journal of proteome research* **12**, 2449-2457, doi:10.1021/pr301011r (2013).

4 Hahne, H. *et al.* DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nature methods* **10**, 989-991, doi:10.1038/nmeth.2610 (2013).

5 Hermjakob, H. & Apweiler, R. The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert review of proteomics* **3**, 1-3, doi:10.1586/14789450.3.1.1 (2006).

6 *ProteomeXchange*, <http://www.proteomexchange.org/> (

7 Desiere, F. *et al.* The PeptideAtlas project. *Nucleic acids research* **34**, D655-658, doi:10.1093/nar/gkj040 (2006).

8 *PeptideAtlas*, <http://www.peptideatlas.org/> (

9 Hill, J. A., Smith, B. E., Papoulias, P. G. & Andrews, P. C. ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. *Journal of proteome research* **9**, 2809-2811, doi:10.1021/pr1000972 (2010).

10 *Proteome Commons (Tranche)*, <https://www.proteomecommons.org/tranche/downloads.jsp> (

11 *MassIVE*, <http://proteomics.ucsd.edu/ProteoSAFe/datasets.jsp> (

12 Tao, F. 1st NCI annual meeting on Clinical Proteomic Technologies for Cancer. *Expert review of proteomics* **5**, 17-20, doi:10.1586/14789450.5.1.17 (2008).

13 *Clinical Proteomic Tumor Analysis Consortium Data Portal*, <https://cptac-data-portal.georgetown.edu/cptacPublic/> (

14 *Broad Institute Proteomics FTP server*, <ftp://ftp.broadinstitute.org/distribution/proteomics/public_datasets/> (

15 Phanstiel, D. H. *et al.* Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nature methods* **8**, 821-827, doi:10.1038/nmeth.1699 (2011).

16 *Stem Cell Omics Repository*, <http://scor.chem.wisc.edu/> (

17 *PRIDE PRoteomics IDEntifications database*, <http://www.ebi.ac.uk/pride/> (

18 Addona, T. A. *et al.* A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. *Nature biotechnology* **29**, 635-643, doi:10.1038/nbt.1899 (2011).

19 Arabi, A. *et al.* Proteomic screen reveals Fbw7 as a modulator of the NF-kappaB pathway. *Nature communications* **3**, 976, doi:10.1038/ncomms1975 (2012).

20 Bantscheff, M. *et al.* Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nature biotechnology* **29**, 255-265,

doi:10.1038/nbt.1759 (2011).

21 Beck, M. *et al.* The quantitative proteome of a human cell line. *Molecular systems biology* **7**, 549, doi:10.1038/msb.2011.82 (2011).

22 Bergamini, G. *et al.* A selective inhibitor reveals PI3Kgamma dependence of T(H)17 cell differentiation. *Nature chemical biology* **8**, 576-582, doi:10.1038/nchembio.957 (2012).

23 Berle, M. *et al.* Quantitative proteomics comparison of arachnoid cyst fluid and cerebrospinal fluid collected perioperatively from arachnoid cyst patients. *Fluids and barriers of the CNS* **10**, 17, doi:10.1186/2045-8118-10-17 (2013).

24 Bhattacharjee, M. *et al.* A multilectin affinity approach for comparative glycoprotein profiling of rheumatoid arthritis and spondyloarthropathy. *Clinical proteomics* **10**, 11, doi:10.1186/1559-0275-10-11 (2013).

25 Burgener, A. *et al.* Comprehensive proteomic study identifies serpin and cystatin antiproteases as novel correlates of HIV-1 resistance in the cervicovaginal mucosa of female sex workers. *Journal of proteome research* **10**, 5139-5149, doi:10.1021/pr200596r (2011).

26 Burkhart, J. M. *et al.* The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* **120**, e73-82, doi:10.1182/blood-2012-04-416594 (2012).

27 Casado, P. *et al.* Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Science signaling* **6**, rs6, doi:10.1126/scisignal.2003573 (2013).

28 Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293-1307, doi:10.1016/j.cell.2012.02.009 (2012).

29 Cho, C. K. *et al.* Quantitative proteomic analysis of amniocytes reveals potentially dysregulated molecular networks in Down syndrome. *Clinical proteomics* **10**, 2, doi:10.1186/1559-0275-10-2 (2013).

30 Dawson, M. A. *et al.* Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* **478**, 529-533, doi:10.1038/nature10509 (2011).

31 Didangelos, A. *et al.* Extracellular matrix composition and remodeling in human abdominal aortic aneurysms: a proteomics approach. *Molecular & cellular proteomics : MCP* **10**, M111 008128, doi:10.1074/mcp.M111.008128 (2011).

32 Farina, A. *et al.* Bile carcinoembryonic cell adhesion molecule 6 (CEAM6) as a biomarker of malignant biliary stenoses. *Biochimica et biophysica acta*, doi:10.1016/j.bbapap.2013.06.010 (2013).

33 Fernandez-Saiz, V. *et al.* SCFFbxo9 and CK2 direct the cellular response to growth factor withdrawal via Tel2/Tti1 degradation and promote survival in multiple myeloma. *Nature cell biology* **15**, 72-81, doi:10.1038/ncb2651 (2013).

34 Geiger, T., Madden, S. F., Gallagher, W. M., Cox, J. & Mann, M. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer research* **72**, 2428-2439, doi:10.1158/0008-5472.CAN-11-3711 (2012).

35 Giansanti, P., Stokes, M. P., Silva, J. C., Scholten, A. & Heck, A. J. Interrogating cAMP-dependent kinase signaling in jurkat T cells via a protein kinase A targeted immune-precipitation phosphoproteomics approach. *Molecular & cellular proteomics : MCP* **12**, 3350-3359, doi:10.1074/mcp.O113.028456 (2013).

36 Guo, H. *et al.* Integrative network analysis of signaling in human CD34(+) hematopoietic progenitor cells by global phosphoproteomic profiling using TiO2

enrichment combined with 2D LC-MS/MS and pathway mapping. *Proteomics* **13**, 1325-1333, doi:10.1002/pmic.201200369 (2013).

37 Hahne, H. *et al.* Proteome wide purification and identification of O-GlcNAc-modified proteins using click chemistry and mass spectrometry. *Journal of proteome research* **12**, 927-936, doi:10.1021/pr300967y (2013).

38 Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150**, 1068-1081, doi:10.1016/j.cell.2012.08.011 (2012).

39 Hennrich, M. L., Groenewold, V., Kops, G. J., Heck, A. J. & Mohammed, S. Improving depth in phosphoproteomics by using a strong cation exchange-weak anion exchange-reversed phase multidimensional separation approach. *Analytical chemistry* **83**, 7137-7143, doi:10.1021/ac2015068 (2011).

40 Hennrich, M. L., van den Toorn, H. W., Groenewold, V., Heck, A. J. & Mohammed, S. Ultra acidic strong cation exchange enabling the efficient enrichment of basic phosphopeptides. *Analytical chemistry* **84**, 1804-1808, doi:10.1021/ac203303t (2012).

41 Iwata, K. *et al.* The human oligodendrocyte proteome. *Proteomics* **13**, 3548-3553, doi:10.1002/pmic.201300201 (2013).

42 Johansson, H. J. *et al.* Retinoic acid receptor alpha is associated with tamoxifen resistance in breast cancer. *Nature communications* **4**, 2175, doi:10.1038/ncomms3175 (2013).

43 Joshi, P. *et al.* The functional interactome landscape of the human histone deacetylase family. *Molecular systems biology* **9**, 672, doi:10.1038/msb.2013.26 (2013).

44 Kentsis, A. *et al.* Urine proteomics for profiling of human disease using high accuracy mass spectrometry. *Proteomics. Clinical applications* **3**, 1052-1061, doi:10.1002/prca.200900008 (2009).

45 Kliemt, S. *et al.* Sulfated hyaluronan containing collagen matrices enhance cell-matrix-interaction, endocytosis, and osteogenic differentiation of human mesenchymal stromal cells. *Journal of proteome research* **12**, 378-389, doi:10.1021/pr300640h (2013).

46 Kruse, U. *et al.* Chemoproteomics-based kinome profiling and target deconvolution of clinical multi-kinase inhibitors in primary chronic lymphocytic leukemia cells. *Leukemia* **25**, 89-100, doi:10.1038/leu.2010.233 (2011).

47 Larance, M., Ahmad, Y., Kirkwood, K. J., Ly, T. & Lamond, A. I. Global subcellular characterization of protein degradation using quantitative proteomics. *Molecular & cellular proteomics : MCP* **12**, 638-650, doi:10.1074/mcp.M112.024547 (2013).

48 Maier, S. K. *et al.* Comprehensive identification of proteins from MALDI imaging. *Molecular & cellular proteomics : MCP* **12**, 2901-2910, doi:10.1074/mcp.M113.027599 (2013).

49 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods* **10**, 634-637, doi:10.1038/nmeth.2518 (2013).

50 Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell reports* **4**, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).

51 Munoz, J. *et al.* The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *The EMBO journal* **31**, 3079-3091, doi:10.1038/emboj.2012.166 (2012).

52 Munoz, J. *et al.* The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Molecular systems biology* **7**, 550,

doi:10.1038/msb.2011.84 (2011).

53 Muraoka, S. *et al.* In-depth membrane proteomic study of breast cancer tissues for the generation of a chromosome-based protein list. *Journal of proteome research* **12**, 208-213, doi:10.1021/pr300824m (2013).

54 Nagaprashantha, L. D. *et al.* Proteomic analysis of signaling network regulation in renal cell carcinomas with differential hypoxia-inducible factor-2alpha expression. *PloS one* **8**, e71654, doi:10.1371/journal.pone.0071654 (2013).

55 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548, doi:10.1038/msb.2011.81 (2011).

56 Ori, A. *et al.* Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Molecular systems biology* **9**, 648, doi:10.1038/msb.2013.4 (2013).

57 Marimuthu, A. *et al.* SILAC-based quantitative proteomic analysis of gastric cancer secretome. *Proteomics. Clinical applications* **7**, 355-366, doi:10.1002/prca.201200069 (2013).

58 Papachristou, E. K. *et al.* The shotgun proteomic study of the human ThinPrep cervical smear using iTRAQ mass-tagging and 2D LC-FT-Orbitrap-MS: the detection of the human papillomavirus at the protein level. *Journal of proteome research* **12**, 2078-2089, doi:10.1021/pr301067r (2013).

59 Paulo, J. A. *et al.* Proteomic analysis (GeLC-MS/MS) of ePFT-collected pancreatic fluid in chronic pancreatitis. *Journal of proteome research* **11**, 1897-1912, doi:10.1021/pr2011022 (2012).

60 Farrah, T. *et al.* The state of the human proteome in 2012 as viewed through PeptideAtlas. *Journal of proteome research* **12**, 162-171, doi:10.1021/pr301012j (2013).

61 Pirmoradian, M. *et al.* Rapid and deep human proteome analysis by single-dimension shotgun proteomics. *Molecular & cellular proteomics : MCP* **12**, 3330-3338, doi:10.1074/mcp.O113.028787 (2013).

62 Prewitz, M. C. *et al.* Tightly anchored tissue-mimetic matrices as instructive stem cell microenvironments. *Nature methods* **10**, 788-794, doi:10.1038/nmeth.2523 (2013).

63 Rhee, H. W. *et al.* Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328-1331, doi:10.1126/science.1230593 (2013).

64 Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular & cellular proteomics : MCP* **12**, 2341-2353, doi:10.1074/mcp.O113.028142 (2013).

65 Shiromizu, T. *et al.* Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project. *Journal of proteome research* **12**, 2414-2421, doi:10.1021/pr300825v (2013).

66 Werner, T. *et al.* High-resolution enabled TMT 8-plexing. *Analytical chemistry* **84**, 7188-7194, doi:10.1021/ac301553x (2012).

67 Wu, Z. *et al.* Quantitative chemical proteomics reveals new potential drug targets in head and neck cancer. *Molecular & cellular proteomics : MCP* **10**, M111 011635, doi:10.1074/mcp.M111.011635 (2011).

68 Wu, Z., Moghaddas Gholami, A. & Kuster, B. Systematic identification of the HSP90 candidate regulated proteome. *Molecular & cellular proteomics : MCP* **11**, M111

016675, doi:10.1074/mcp.M111.016675 (2012).

69 Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499**, 79-82, doi:10.1038/nature12223 (2013).

70 Farrah, T. *et al.* A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Molecular & cellular proteomics : MCP* **10**, M110 006353, doi:10.1074/mcp.M110.006353 (2011).

71 Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567, doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (1999).

72 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

73 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794-1805, doi:10.1021/pr101065j (2011).

74 *UniProt release 2011_05*, <http://www.uniprot.org/news/2011/05/03/release> (

75 *UniProtKB FAQs: What are complete proteomes?*, <http://www.uniprot.org/faq/15> (

76 *The Global Proteome Machine > common Repository of Adventitious Proteins*, <http://www.thegpm.org/cRAP/index.html> (

77 Brosch, M., Yu, L., Hubbard, T. & Choudhary, J. Accurate and sensitive peptide identification with Mascot Percolator. *Journal of proteome research* **8**, 3176-3181, doi:10.1021/pr800982s (2009).

78 Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* **4**, 923-925, doi:10.1038/nmeth1113 (2007).

79 Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25**, 1915-1927, doi:10.1101/gad.17446611 (2011).

80 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

81 Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell. Proteomics* **10**, M110 003830, doi:M110.003830 [pii] n10.1074/mcp.M110.003830 (2011).

82 Lane, L. *et al.* Metrics for the Human Proteome Project 2013-2014 and strategies for finding missing proteins. *Journal of proteome research* **13**, 15-20, doi:10.1021/pr401144x (2014).

83 Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nature biotechnology* **25**, 125-131, doi:nbt1275 [pii] 10.1038/nbt1275 (2007).

84 Hahne, H. & Kuster, B. A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J. Am. Soc. Mass Spectrom.* **22**, 931-942, doi:10.1007/s13361-011-0107-y (2011).

85 Wenschuh, H. *et al.* Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Biopolymers* **55**, 188-206, doi:10.1002/1097-0282(2000)55:3<188::AID-BIP20>3.0.CO;2-T (2000).

86 Frank, R. & Overwin, H. SPOT synthesis. Epitope analysis with arrays of synthetic peptides prepared on cellulose membranes. *Methods in molecular biology* **66**, 149-169, doi:10.1385/0-89603-375-9:149 (1996).

87 Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research* **39**, D507-513, doi:10.1093/nar/gkq968 (2011).

88 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).

89 Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* **8**, 2405-2417, doi:10.1074/mcp.M900317-MCP200 (2009).

90 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, doi:10.1038/nature10098 (2011).

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

92 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

93 Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* **6**, e21800, doi:10.1371/journal.pone.0021800 (2011).

94 Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & cellular proteomics : MCP* **11**, M111 014050, doi:10.1074/mcp.M111.014050 (2012).

95 Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & cellular proteomics : MCP* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).

96 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).

97 Zou & Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* **67**, 199-320 (2005).

98 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

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

100 Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* **39**, 1-13 (2011).

101 Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Molecular & cellular proteomics : MCP* **1**, 845-867 (2002).

102 Higdon, R. *et al.* IPM: An integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *Journal of proteomics* **75**, 116-121, doi:10.1016/j.jprot.2011.06.003 (2011).

103 Zhang, J. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics : MCP*

11, M111 010587, doi:10.1074/mcp.M111.010587 (2012).

104 Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nature biotechnology* **31**, 557-564, doi:10.1038/nbt.2585 (2013).

105 Malmstrom, J. *et al.* Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **460**, 762-765, doi:10.1038/nature08184 (2009).

106 Ahrne, E., Molzahn, L., Glatter, T. & Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **13**, 2567-2578, doi:10.1002/pmic.201300135 (2013).

107 Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Molecular & cellular proteomics : MCP* **5**, 144-156, doi:10.1074/mcp.M500230-MCP200 (2006).

108 Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology* **4**, 222, doi:10.1038/msb.2008.61 (2008).