ARTICLE

doi:10.1038/nature13319

Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm[1,2]*, Judith Schlegl[2]*, Hannes Hahne[1]*, Amin Moghaddas Gholami[1]*, Marcus Lieberenz[2], Mikhail M. Savitski[3], Emanuel Ziegler[2], Lars Butzmann[2], Siegfried Gessulat[2], Harald Marx[1], Toby Mathieson[3], Simone Lemeer[1], Karsten Schnatbaum[4], Ulf Reimer[4], Holger Wenschuh[4], Martin Mollenhauer[5], Julia Slotta-Huspenina[5], Joos-Hendrik Boese[2], Marcus Bantscheff[3], Anja Gerstmair[2], Franz Faerber[2] & Bernhard Kuster[1,6]

Proteomes are characterized by large protein-abundance differences, cell-type- and time-dependent expression patterns and post-translational modifications, all of which carry biological information that is not accessible by genomics or transcriptomics. Here we present a mass-spectrometry-based draft of the human proteome and a public, high-performance, in-memory database for real-time analysis of terabytes of big data, called ProteomicsDB. The information assembled from human tissues, cell lines and body fluids enabled estimation of the size of the protein-coding genome, and identified organ-specific proteins and a large number of translated lincRNAs (long intergenic non-coding RNAs). Analysis of messenger RNA and protein-expression profiles of human tissues revealed conserved control of protein abundance, and integration of drug-sensitivity data enabled the identification of proteins predicting resistance or sensitivity. The proteome profiles also hold considerable promise for analysing the composition and stoichiometry of protein complexes. ProteomicsDB thus enables navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

The large-scale interrogation of biological systems by mass-spectrometry-based proteomics provides insights into protein abundance, cell-type and time-dependent expression patterns, post-translational modifications (PTMs) and protein–protein interactions, all of which carry biological information that is best investigated at the protein level. Perhaps surprisingly, it is still not clear which of the 19,629 human genes annotated in Swiss-Prot[1] (20,493 in UniProt) are translated into proteins. Therefore, major efforts are underway to identify these genes, including the Human Proteome Project (HPP), which aims to broadly characterize the human proteome, the Human Protein Atlas project (HPA), which seeks to generate antibodies for all human proteins, and the ProteomeXchange consortium, which facilitates the gathering and sharing of proteomic data[2–4]. Despite the fact that a plethora of individual human proteomic studies exist, only a few systematic efforts to assemble and characterize human proteomes have been reported so far[5,6]. In part this is because most proteomic data do not reside in public repositories, proteomic data annotation is often sketchy, and the data generation and processing platforms are of varying capability, performance and maturity. Importantly, there is also a notable challenge in making 'big data' (that is, large amounts of data that are difficult or impossible to process using traditional technology and algorithms) more widely accessible to the scientific community, because the development of scalable analysis tools is only in its infancy.

Assembly of the proteome in ProteomicsDB

Here we present a draft of the human proteome assembled using data from 16,857 liquid chromatography tandem-mass-spectrometry (LC-MS/MS) experiments involving human tissues, cell lines, body fluids, as well as data from PTM studies and affinity purifications. We also present the analysis of the assembled data, in ProteomicsDB, an in-memory database designed for the real-time analysis of big data (https://www.proteomicsdb.org). For this study (Fig. 1a), we combined data available from repositories and otherwise contributed by colleagues (60% of total) with published as well as new data from the authors' laboratories (40% of total; Supplementary Table 1 and Supplementary Information). To maximize proteome coverage, we reprocessed all experiments using MaxQuant[7] and Mascot[8], and the resulting 1.1-billion peptide spectrum matches (PSMs) were imported into ProteomicsDB. The database (Fig. 1b) comprises a public repository, a web interface featuring several data views and analysis tools, and an application programming interface (API). At the heart of ProteomicsDB is an 'in-memory' computational resource commanding 2 terabytes (TB) of random access memory (RAM) and 160 central processor units (CPUs), which enables the storage of all data in the main memory, all of the time. This makes computational tasks very efficient, illustrated by the capability to display and annotate any of the approximately 71-million currently identified peptide-mass spectra in real time (Extended Data Fig. 1). Controlling the quality of peptide and protein identifications is important but exactly how this is best accomplished is still debated in the community[9,10]. For the current assembly of the proteome, we relied on high resolution mass-spectrometry data to keep false identifications low. We applied a two-step filtering process, first by controlling the false discovery rate (FDR) at 1% for PSMs generated by each LC-MS/MS experiment using a global target-decoy approach[11]. Peptide identifications then had to pass a length-dependent Mascot or Andromeda score threshold of 5% local FDR on the total aggregated data and we categorically rejected all peptides shorter than seven amino acids (Extended Data Figs 1 and 2, and Supplementary Information). Comparison to 27 published studies shows that this scheme is in line with the often-used 1% protein FDR criterion (Supplementary Table 1) and avoids the unsolved issue of artificially high protein FDRs when analysing large data sets[12] (Extended Data Fig. 2 and Supplementary Information).

Proteomic annotation of the genome

At the time of writing, ProteomicsDB held protein evidence for 18,097 of the 19,629 human genes annotated in Swiss-Prot (92%) as well as

[1]Chair of Proteomics and Bioanalytics, Technische Universität München, Emil-Erlenmeyer Forum 5, 85354 Freising, Germany. [2]SAP AG, Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany. [3]Cellzome GmbH, Meyerhofstraße 1, 69117 Heidelberg, Germany. [4]JPT Peptide Technologies GmbH, Volmerstraße 5, 12489 Berlin, Germany. [5]Institute of Pathology, Technische Universität München, Trogerstraße 18, 81675 München, Germany. [6]Center for Integrated Protein Science Munich, Germany.
*These authors contributed equally to this work.

©2014 Macmillan Publishers Limited. All rights reserved

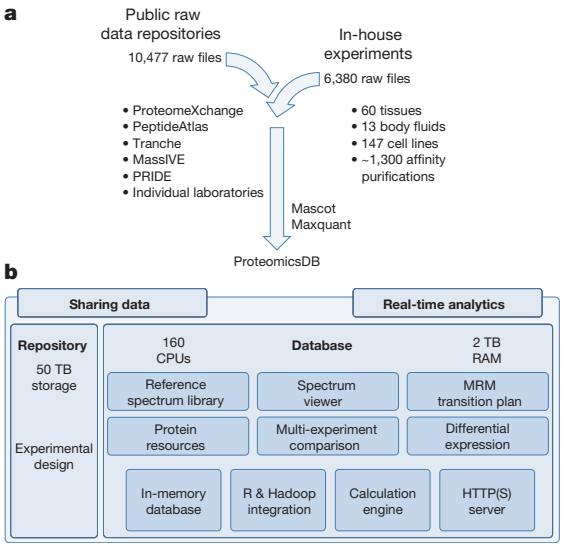


Figure 1 | Strategy for the assembly of the human proteome. **a**, Experimental workflow for the identification and quantification of proteins. **b**, Structure and features of ProteomicsDB. ProteomicsDB consists of a repository part for raw-data storage and an in-memory database designed for the storage, analysis and visualization of proteomic data sets. Fast computation on large data sets is backed by 160 CPUs and 2 TB of RAM.

19,376 out of 86,771 protein isoforms listed in UniProt (22%; Supplementary Table 1). Chromosomes were evenly covered with the notable exceptions of chromosome 21 and the Y chromosome (Fig. 2a). The former contains many proteins with few mass-spectrometry-compatible tryptic peptides. As 257 human proteins (not counting isoforms) do not produce any such peptides, this renders trypsin—as the most frequently used protease in proteomics—ineffective. As a result, alternative proteases or top down sequencing approaches will have a part to play in the eventual completion of the human proteome (Extended Data Fig. 3a)[13,14]. To facilitate this, ProteomicsDB provides a tool predicting the best protease or combinations thereof for any protein which can also be valuable when systematically mapping PTMs.

We next attempted to estimate the size of the protein-coding genome based on UniProt protein evidence categories. ProteomicsDB currently covers 97% of the 13,378 genes with annotated evidence on protein and 84% (of 5,531) with evidence on transcript level. The overlap with proteins detected by antibodies in the HPA project is 93% (of 15,156 HPA proteins) providing independent evidence that these genes exist as proteins. Conversely, proteomic coverage of genes inferred from homology (52% of 159), genes marked as predicted (64% of 72) or uncertain (56% of 489), was considerably lower, suggesting that the protein-coding human genome may be several hundred genes smaller than anticipated previously. Still, we were able to validate the identification of 36 of the uncertain genes (out of 44 tried)[15] using reference spectra from synthetic peptides (Supplementary Table 2). Among the identified uncertain genes were three lincRNAs (Extended Data Fig. 3). This unexpected result prompted us to search approximately 9-million tandem-mass-spectrometry spectra from tissues and cell lines against 13,564 lincRNA sequences from Ensembl and 21,487 lincRNAs and TUCPs (transcripts of uncertain coding potential) from the Broad Institute[16]. This returned 430 high-quality peptides (no homology to UniProt sequences) from 404 lincRNAs and TUCPs (Supplementary Table 3). There was no apparent bias in chromosomal location or biological source, and the abundance distribution of translated lincRNA peptides was broadly similar to that of peptides from ordinary proteins (Extended Data Fig. 3). To our knowledge, this is the largest number of lincRNA and TUCP translation products with direct peptide evidence reported to date[17], arguing that translation of such transcripts is more common than anticipated previously[18–20]. The biological significance of translated lincRNAs and TUCPs is not clear at present. These may constitute proteins 'in evolution' representing hitherto undiscovered biology[21] or arise by stochastic chance marking such proteins as 'biological noise'.

Core proteome and missing proteome

Aggregation of the data used for building the draft proteome shows that proteome coverage rapidly saturates at approximately 16,000 –17,000 proteins, which is similar to transcriptome coverage obtained by RNA sequencing (RNA-seq). Addition of human-tissue and body-fluid data each led to small but noticeable contributions not provided by cell lines. The same is true when adding PTM or affinity data to shotgun proteomic data (Extended Data Fig. 4). When comparing five of the largest data sets in ProteomicsDB[22–25], the existence of a human core proteome[26] of approximately 10,000–12,000 ubiquitously expressed proteins can be postulated, the primary function of which is the general control and maintenance of cells (Extended Data Fig. 4 and Supplementary Table 4). The low abundance range of the core proteome is enriched in proteins with regulatory functions. The observed proteome saturation implies that adding more shotgun data will not considerably increase coverage, although it would increase confidence in individual proteins. Instead, it is likely that the 'missing proteome' (Fig. 2b and Supplementary Table 4) will have to be identified by more focused experimentation. It is also possible that a considerable part of the missing proteome constitutes (pseudo)genes that are no longer expressed. G-protein-coupled receptors (GPCRs) are underrepresented in ProteomicsDB and the respective transcripts are also notoriously absent in RNA-seq data[27]. Earlier work suggests that more than half of the 853 human GPCRs have lost their function over the course of human evolution and may be considered obsolete[28]. Similarly, a large number of functionally uncategorized proteins are annotated pseudogenes, potentially further reducing the number of (actual) protein-coding genes. Cytokines may be underrepresented because of experimental issues as small, secreted proteins can

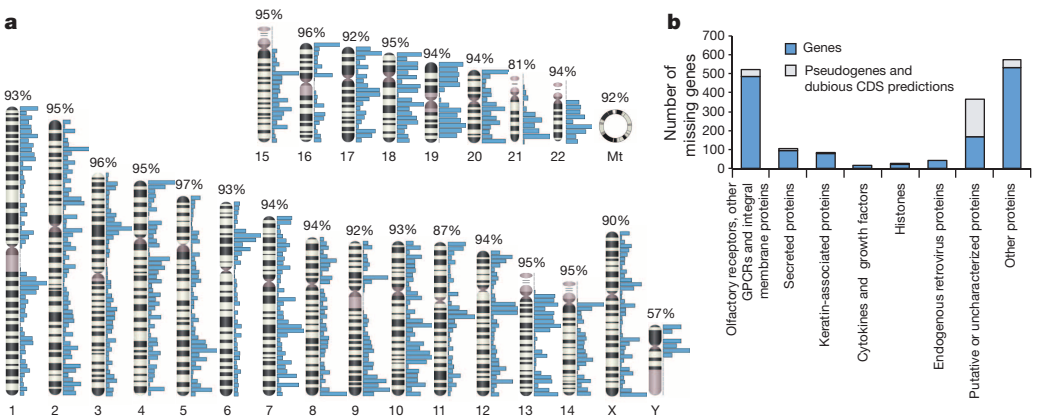


Figure 2 | Characterization of the human proteome. **a**, Chromosomal coverage of the 18,097 proteins identified in this study exceeds 90% in all but three cases. Blue bars indicate the density of proteins in a particular chromosomal region. **b**, Gene ontology analysis of the 'missing' proteome' identifies GPCRs, secreted and keratin-associated proteins as the major protein classes underrepresented in proteomic experiments. CDS, coding sequence; Mt, mitochondrial DNA.

©2014 Macmillan Publishers Limited. All rights reserved

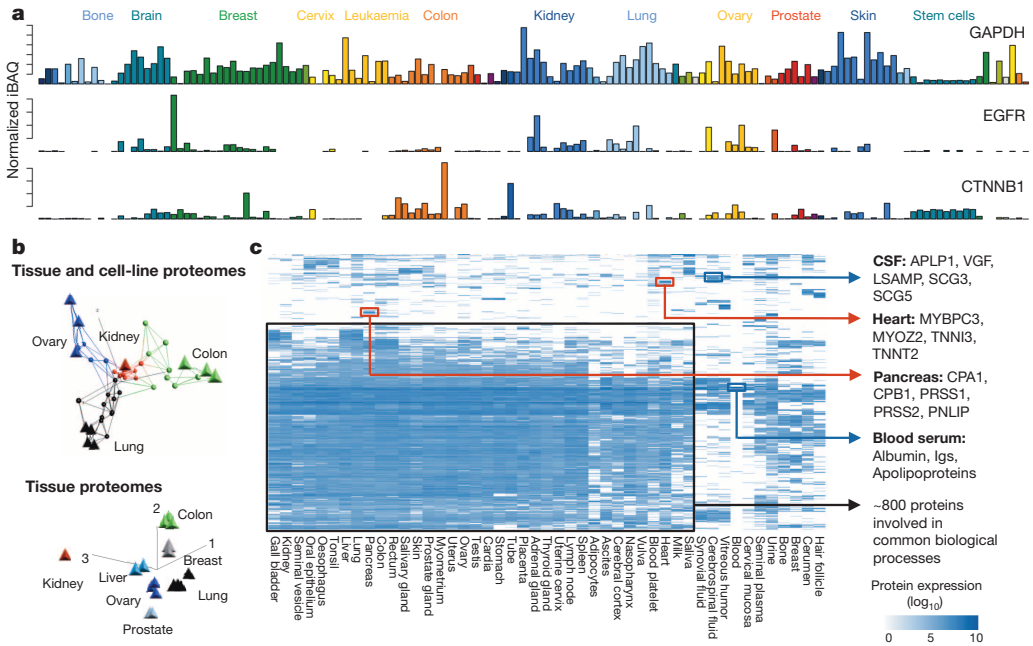


Figure 3 | Global protein expression analysis. **a**, Protein expression in different tissues and cell lines, showing that levels of housekeeping (GAPDH), signalling (EGFR) and tumour-associated (CTNNB1) proteins can vary substantially between tissues (grouped by colour). **b**, PCA showing that cell lines (circles) retain protein-expression characteristics of their respective primary tissue (triangles) and that proteomes of different organs are more diverse. **c**, Hierarchical clustering of the 100 most highly expressed proteins from each of 47 tissues and body fluids. Despite the presence of a large group of common proteins, clusters of organ and fluid-selective proteins with respective biological functions can readily be identified.

still be difficult to obtain from the supernatants of cells, the intercellular space of tissues or from body fluids. To fill the remaining gaps in the human proteome, ProteomicsDB provides a facility to engage the community by 'adopting' a missing protein; that is, to provide mass-spectrometric evidence for its existence. In addition, we have synthesized and identified 435 peptides for all 273 cytokines as well as 3,539 further peptides for proteins not yet well covered and have made their tandem-mass-spectrometry spectra available in ProteomicsDB so that any identification of such proteins in the future may be validated using the synthetic reference standard (Supplementary Table 2).

Functional proteome–expression analysis

We have generated proteome profiles of 27 human tissues and body fluids (human body map) complemented with publically available data (Supplementary Tables 1 and 5) to begin to analyse human proteomes in functional terms. To normalize the disparate data sets in Proteomics DB, we found the intensity-based absolute-protein-quantification method (iBAQ) to be appropriate (Extended Data Fig. 5 and Supplementary Information)[29–31]. A simple common task is to compare the expression level of a single protein across many biological sources (Fig. 3a). Although housekeeping proteins such as GAPDH (glyceraldehyde-3-phosphate dehydrogenase) show high (and sometimes extreme) expression throughout biological sources, high levels of the proto-oncogene EGFR (epidermal growth factor receptor) are mostly confined to cancerous tissue; for example, breast cancer tissue. Similarly, β-catenin, a member of the Wnt pathway, is highly expressed in colon cancer cells, where the protein participates in the development of the malignancy. Principle component analysis (PCA) of protein abundances in 42 proteomes shows that protein expression in a particular tissue and its corresponding cell lines is broadly similar and that there are more substantial differences between tissues of different organs (Fig. 3b). This result is important for the interpretation of data presented below and also contributes to the ongoing discussion regarding the suitability of cell lines as model systems for studying human biology. A comparative analysis of the 100 most highly expressed proteins in each of 47 human organs and body fluids (Fig. 3c) revealed that approximately 70% of these proteins are found in all organs and body fluids but show expression differences of up to five orders of magnitude (Supplementary Table 4). Interestingly, even the most highly abundant proteins in a tissue or fluid often point to molecular processes associated with the respective biological specialization; myofibrillar proteins, including troponins, are abundant in the heart, proteases in the pancreas and neuronal proteins in cerebrospinal fluid.

Similar observations can be made when investigating proteins forming functional classes such as protein kinases or transcription factors (Fig. 4 and Supplementary Table 4). Akin to core proteomes, some of the 349 detected kinases and 557 transcription factors are broadly expressed, but others seem to be confined to few organs where they drive more specific processes. For example, the kinases HCK, ZAP70, LCK, JAK3, TXK and FGR are found in a tight cluster of kinases in the spleen and all have important roles in the biology of immune cells. This is 'mirrored' by transcription factors in the same cluster with strong ties to immunity, including the NF-κB system (REL, PRKCH, NFKBIE) and Toll-like receptor signalling (SIGIRR, IRF5, ARRB2, NLRC4). It is noteworthy that many of the proteins in the spleen cluster are also highly

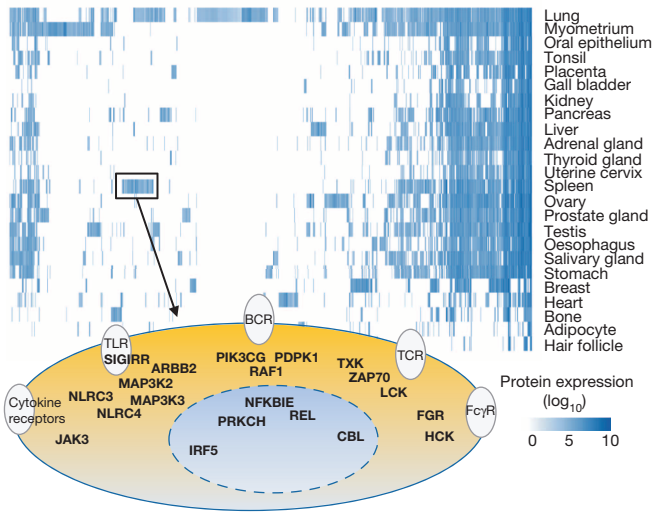


Figure 4 | Functional protein expression analysis. Quantitative expression analysis of 906 kinases and transcription factors across 24 tissues (top panel) identifies organ-selective signatures indicative of the underlying biology. The highlighted cluster in spleen contains the kinases LCK, ZAP70 and JAK and the transcription factors SIGIRR, NFKBIE and NLRC3 with strong links to the immune system (bottom panel). Yellow oval represents a cell; blue oval represents the nucleus.

©2014 Macmillan Publishers Limited. All rights reserved

expressed in the lung, a primary entry point for human pathogens. The number of proteins that are exclusively or preferentially detected in a particular organ is surprisingly small, and gene ontology analysis invariably highlights organ-specific biology (Extended Data Fig. 6). For example, adipocytes are rich in proteins involved in lipid storage, platelets in growth factors, and placenta in proteins relating to hormonal regulation and pregnancy (Supplementary Table 5). The above shows that even disparate, though high-quality proteomic data can be used to construct protein expression maps across an entire complex organism. A recent report has shown that this is feasible in mice[32] but to our knowledge, organism-wide proteome-expression profiling has not been described in humans before. In addition, the identification of a considerable number of proteins with no ascribed function but exclusive (or high) expression in particular organs implies a functional role. The contextual information provided in ProteomicsDB may thus provide guidance for the eventual identification of the biological role of these orphan proteins.

Integration and utility of proteomes

Many further uses of protein-expression profiling can be envisaged, of which we can only outline a few here. We have compared messenger RNA (RNA-seq)[27] and protein (iBAQ, this study) expression profiles for 12 human tissues (Extended Data Fig. 7, Supplementary Table 6 and Supplementary Information) and clear correlations are observed in all cases but the Spearman's rank correlation coefficients are rather moderate and somewhat poorer than those previously reported for cell lines. This is likely to be due to the fact that tissues generally comprise a mixture of cell types, connective tissue and blood. Both mRNA and protein levels vary greatly between tissues as one might expect; however, the ratio of protein and mRNA levels is remarkably conserved between tissues for any given protein (Fig. 5a, top panel)[33]. It has been shown previously[31] that the translation rate constant is one dominant factor determining protein abundance in cell lines. Using the ratio of protein to mRNA levels as a proxy for translation rates, our data show

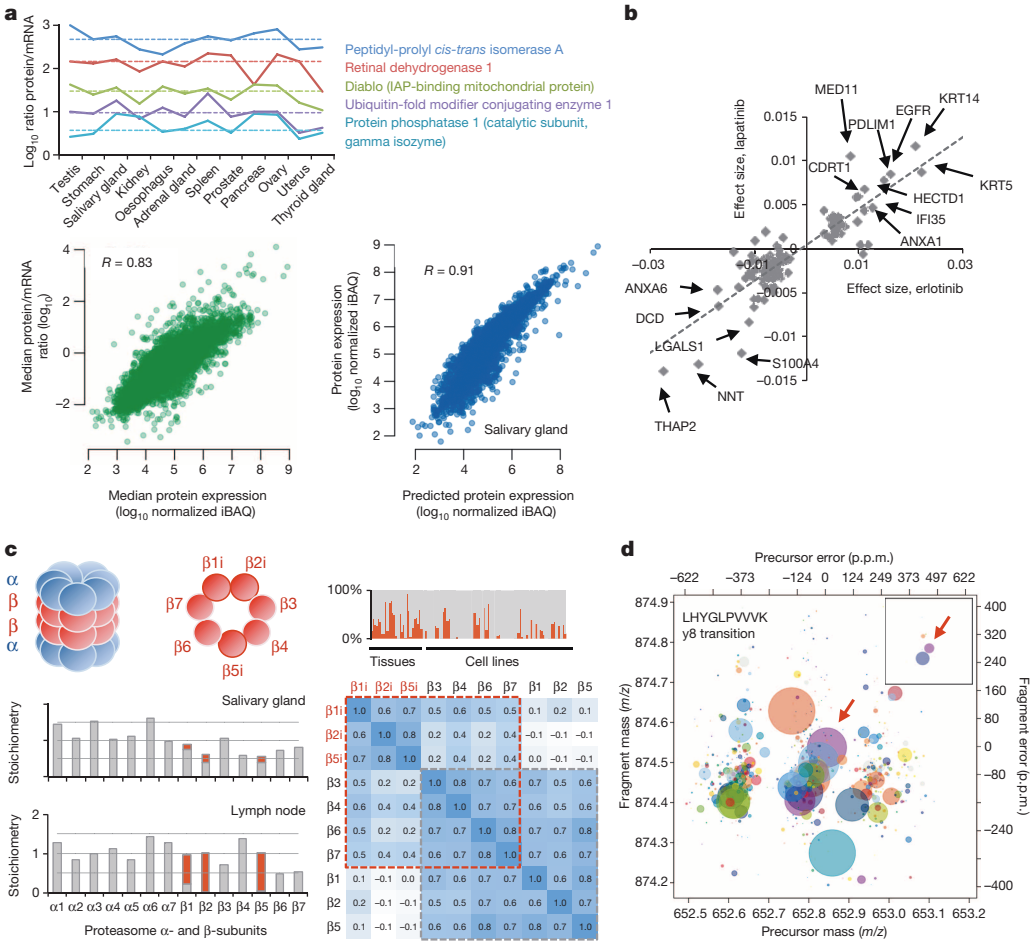


Figure 5 | Integration and utility of large proteomic data collections. **a**, Analysis of mRNA and protein levels across 12 organs shows that the protein/mRNA ratio is largely conserved (top panel). The median translation rates of all transcripts across all tissues correlate well with protein abundance (bottom-left panel), leading to the ability to predict individual protein levels from the respective mRNA levels (bottom-right panel). **b**, Elastic net analysis for the identification of drug sensitivity (positive-effect-size) or resistance (negative-effect-size) markers against the EGFR kinase inhibitors erlotinib and lapatinib in cancer cell lines. **c**, Analysis of the composition and stoichiometry of the proteasome. Top-left panel, schematic structure of the 'constitutive' proteasome and the 'immunoproteasome' (marked by the suffix 'i'). Middle-left and bottom-left panels, stoichiometry derived by iBAQ of the constitutive proteasome (grey) and the immunoproteasome (red) in the salivary gland and the lymph node. Top-right panel, expression analysis of the β1 subunit across more than 100 tissue and cell-line proteomes reveals that many cells express both forms of the proteasome. Bottom-right panel, expression correlation analysis of all β subunits across the said tissues and cell lines showing strong co-expression of the β1i, β2i and β5i subunits as well as all other β-subunits but no correlation with the expression of the corresponding β1, β2 and β5 subunits. **d**, ProteomicsDB enables the computation of molecular interferences in selected reaction-monitoring experiments (SRM) from experimental data. The transition of the target peptide LHYGLPVVVK (y8 fragment ion, β-catenin) is marked with an arrow. All other circles in the plot are interfering SRM transitions of other peptides found in ProteomicsDB that fall within the same mass tolerance of the experiment (here, 0.7 Da). The size of each circle indicates the severity of the interference. The inset shows that interference can be substantially reduced by the use of high-resolution fragment-ion data (here, 0.04 Da) and confining the analysis to the tissue from which a sample is derived (here, a colon sample).

©2014 Macmillan Publishers Limited. All rights reserved

that this is also true for human tissues and that the ratio is similar in every tissue (Fig. 5a, bottom left panel). It therefore appears that the translation rate is a fundamental, encoded (constant) characteristic of a transcript, suggesting that the actual amount of protein in a given cell is primarily controlled by regulating mRNA levels. Having learned the protein/mRNA ratio for every protein and transcript, it now becomes possible to predict protein abundance in any given tissue with good accuracy from the measured mRNA abundance (Fig. 5a, bottom right panel, and Extended Data Fig. 7).

We have shown previously that protein expression can be correlated to drug sensitivity[24]. Here we used drug-sensitivity data provided by the cancer cell line encyclopedia[34] (CCLE) to discover sensitivity and resistance markers for 24 drugs in 35 human cancer cell lines (Supplementary Table 7). For the EGFR kinase inhibitors erlotinib and lapatinib the primary target (EGFR) as well as annexin A1 (ANXA1, a direct EGFR substrate), and EGFR interacting proteins at stress fibres (PDLIM, KRT5, KRT14) all indicate drug sensitivity, whereas high expression of ANXA6 or S100A4 renders cells less responsive (Fig. 5b and Extended Data Fig. 8). Assuringly, knockdown of ANXA6 in BT549 cells has been shown to sensitize cells to lapatinib[35] and addition of S100A4 to cells in culture has been shown to stimulate EGFR and to promote metastasis[36]. High expression of S100 proteins is often associated with resistance against kinase inhibitors (Supplementary Table 7), suggesting that this may constitute a general molecular resistance mechanism. Similar effects can be postulated for other proteins (Supplementary Notes) and in light of a recent report showing increased phosphorylation of HECTD1 on EGF treatment[37], it is tempting to speculate that a HECTD1–CDRT1 E3 ubiquitin ligase–orphan F-box protein complex may be involved in regulating the stability of EGFR via the ubiquitin–proteasome system.

The composition and stoichiometry of protein complexes is typically analysed by affinity purification coupled to mass-spectrometry-based protein analysis and it emerges that protein expression profiling may also have potential for this purpose[38]. We found that stoichiometries measured by iBAQ for the nuclear pore complex agreed well with a prior study using absolute protein quantification by spiked peptide standards (Extended Data Fig. 9 and Supplementary Table 8)[39]. Using the proteasome as an example, we explored its composition and stoichiometry heterogeneity across cell lines and tissues (Fig. 5c). The constitutive core proteasome consists of 2×7 non-catalytic α- and 2×7 catalytic β-subunits but, for example, an 'immunoproteasome' has been identified in which the β1, 2 and 5 subunits are replaced by homologous proteins (β1i, β2i and β5i) in immune cells[40,41]. Our analysis shows that the proteasome in the salivary gland is primarily of the constitutive type and that lymph nodes almost exclusively contain the immunoproteasome (Fig. 5c, left panel). The same analysis across more than 100 cell-line and tissue samples (Fig. 5c, right panel) reveals that the immunoproteasome is surprisingly widely expressed, including in tissues for which no primary immunological function would be expected. In addition, the data imply that the molecular composition and stoichiometry of proteasomes is heterogeneous and cell-type-dependent. Correlation analysis of the expression of all β-subunits (Fig. 5c, bottom right panel) strongly suggests that the β1, 2 and 5 subunits and their respective immunoproteasome counterparts are expressed independently (no correlation). In contrast, it seems that the remaining subunits (β3, 4, 6, 7) are co-expressed with either group.

Proteomic data collections can be valuable data mines for post-translational modification analysis or developing proteome technology. ProteomicsDB currently contains 81,721 unique phosphorylated peptides representing 11,025 human genes, demonstrating that more than half of all human proteins are substrates of kinases. Similarly, there are 29,031 unique ubiquitinylated peptides from 5,769 proteins representing substrates of ubiquitin ligases as well as 16,693 acetylated peptides from 7,098 proteins that are substrates of acetylases. Our analysis also detected amino-terminal peptides for 7,977 proteins and carboxy-terminal peptides for 6,778 proteins confirming a large number of translation start and stop sites (Extended Data Fig. 10a). We expect that the PTM branch of ProteomicsDB will grow rapidly over time and help to build a future version of the human proteome that provides more direct links between protein expression and activity.

So-called 'proteotypic' peptides[42] have proven useful as quantification standards in targeted proteomic measurements, which are increasingly employed to develop clinical biomarker assays[43]. ProteomicsDB enabled us to determine the proteotypicity of ~approximately 500,000 peptides and to expand the concept to chemically labelled peptides (Extended Data Fig. 10b, Supplementary Table 9 and Supplementary Information). The 71-million peptide-precursor and 18-billion peptide-fragment ion measurements enables the computational assessment of the specificity of targeted measurements ahead of the actual experiment. Exemplified by the peptide LHYGLPVVVK of the proto-oncogene β-catenin (Fig. 5d and Extended Data Fig. 10c), mining of ProteomicsDB revealed a large number of potentially interfering peptides that may distort the quantification of the target peptide. Interference can be substantially reduced by high-resolution instruments[44] and by limiting the allowed interferences to the tissue in question (Fig. 5d, inset). We anticipate that the combination of experimental proteotypicity, interference estimation and high-resolution instrumentation will provide for more robust targeted proteomic assays in the future.

Discussion

Here we have shown that an extensive draft of the human proteome can be assembled from disparate but high-quality proteomic data. We have outlined some of the many applications that can be envisaged for its use and some of the biological insights that may be generated by mining the proteome. Similar to the evolution of the human genome projects, the eventual completion of the human proteome will take further time and effort but will also lead to substantial improvements in technology, which are still needed. One issue to address is proteome coverage and resolution. While DNA and RNA sequencing technologies have attained single-nucleotide resolution, the amino-acid coverage of proteins is still limited, which currently impairs our ability to detect protein variants, such as differential splice products, PTMs, mutations or isoforms in a systematic fashion. A related challenge is to improve the ability to sample a proteome comprehensively; that is, 'all proteins, all the time'. Another important area of future research concerns overcoming the uncertainties associated with peptide and protein identification by sequence-database searching[45]. ProteomicsDB and similar resources have a part to play in these challenges as the data assembled will enable the development of computational tools and laboratory reagents facilitating proteome-wide discovery experiments, multiplexed quantitative protein assays, as well as general exploration of the human proteome.

METHODS SUMMARY

Proteomic data were downloaded from public repositories, contributed by individual laboratories and specifically generated for this study by the authors' laboratories. For the specifically generated data, human tissue specimens were obtained from the Biobank of the Technische Universität München following approval of the study by the local ethics committee. Samples were collected within the first 30 min after resection, macroscopically resected by an experienced pathologist, snap frozen and stored in liquid nitrogen until use. Body fluids requiring no invasive procedures were provided by volunteers. Proteins were extracted under denaturing conditions and either separated by LDS-PAGE followed by in-gel protease digestion or digested in solution in the presence of chaotropic agents. Synthetic peptides were produced by solid-phase chemistry following the standard Fmoc strategy and used without purification. Peptides were separated by ultra-high-pressure liquid chromatography and analysed on Orbitrap mass spectrometers using either resonance-type or beam-type collision-induced dissociation. For peptide identification, tandem mass spectra were processed in parallel using Mascot Distiller and MaxQuant with Andromeda[7], and searched against UniProt and/or a custom build fasta-formatted sequence file containing lincRNA sequences. Search results and tandem mass spectra were imported into ProteomicsDB (https://www.proteomicsdb.org) and filtered at 1% PSM FDR and 5% local peptide-length-dependent FDR. For bioinformatic analysis, data were extracted from ProteomicsDB using HANA Studio and further processed using custom python scripts and statistics programme R. Gene ontology analysis was

©2014 Macmillan Publishers Limited. All rights reserved

performed using David (http://david.abcc.ncifcrf.gov) and REViGO (http://revigo.irb.hr/). See Supplementary Information for details.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 November 2013; accepted 11 April 2014.

1. UniProt. C. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41,** D43–D47 (2013).
2. Paik, Y. K. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature Biotechnol.* **30,** 221–223 (2012).
3. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnol.* **28,** 1248–1250 (2010).
4. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnol.* **32,** 223–226 (2014).
5. Farrah, T. *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J. Proteome Res.* **13,** 60–75 (2014).
6. Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* **11,** 492–500 (2012).
7. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372 (2008).
8. 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 (1999).
9. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22,** 1111–1120 (2011).
10. Higdon, R. *et al.* IPM: An integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *J. Proteomics* **75,** 116–121 (2011).
11. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnol.* **24,** 1285–1292 (2006).
12. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8,** 2405–2417 (2009).
13. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548 (2011).
14. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480,** 254–258 (2011).
15. Lane, L. *et al.* Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.* **13,** 15–20 (2014).
16. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–1927 (2011).
17. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
18. Bánfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22,** 1646–1657 (2012).
19. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251 (2013).
20. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
21. Flintoft, L. Non-coding RNA: Ribosomes, but no translation, for lincRNAs. *Nature Rev. Genet.* **14,** 520 (2013).
22. 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. *Mol. Cell. Proteomics* **11,** M111.014050 (2012).
23. Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature Methods* **10,** 634–637 (2013).
24. Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **4,** 609–620 (2013).
25. 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. *J. Proteome Res.* **12,** 2414–2421 (2013).
26. Schirle, M., Heurtier, M. A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2,** 1297–1305 (2003).
27. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13,** 397–406 (2014).
28. Hughes, G. M., Teeling, E. C. & Higgins, D. G. Loss of olfactory receptor function in hominin evolution. *PLoS ONE* **9,** e84714 (2014).
29. Ahrné, E., Molzahn, L., Glatter, T. & Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **13,** 2567–2578 (2013).
30. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549 (2011).
31. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
32. Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell. Proteomics* **12,** 1709–1722 (2013).
33. Low, T. Y. *et al.* Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.* **5,** 1469–1478 (2013).
34. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
35. Koumangoye, R. B. *et al.* Reduced annexin A6 expression promotes the degradation of activated epidermal growth factor receptor and sensitizes invasive breast cancer cells to EGFR-targeted tyrosine kinase inhibitors. *Mol. Cancer* **12,** 167 (2013).
36. Klingelhöfer, J. *et al.* Epidermal growth factor receptor ligands as new extracellular targets for the metastasis-promoting S100A4 protein. *FEBS J.* **276,** 5936–5948 (2009).
37. Argenzio, E. *et al.* Proteomic snapshot of the EGF-induced ubiquitin network. *Mol. Syst. Biol.* **7,** 462 (2011).
38. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150,** 1068–1081 (2012).
39. Ori, A. *et al.* Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.* **9,** 648 (2013).
40. Hisamatsu, H. *et al.* Newly identified pair of proteasomal subunits regulated reciprocally by interferon gamma. *J. Exp. Med.* **183,** 1807–1816 (1996).
41. Nandi, D., Jiang, H. & Monaco, J. J. Identification of MECL-1 (LMP-10) as the third IFN-gamma-inducible proteasome subunit. *J. Immunol.* **156,** 2361–2364 (1996).
42. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnol.* **25,** 125–131 (2007).
43. Domon, B. Considerations on selected reaction monitoring experiments: implications for the selectivity and accuracy of measurements. *Proteomics Clin. Appl.* **6,** 609–614 (2012).
44. Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteomics* **11,** 1709–1723 10.1074/mcp.O112.019802 (2012).
45. Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nature Biotechnol.* **31,** 557–564 (2013).

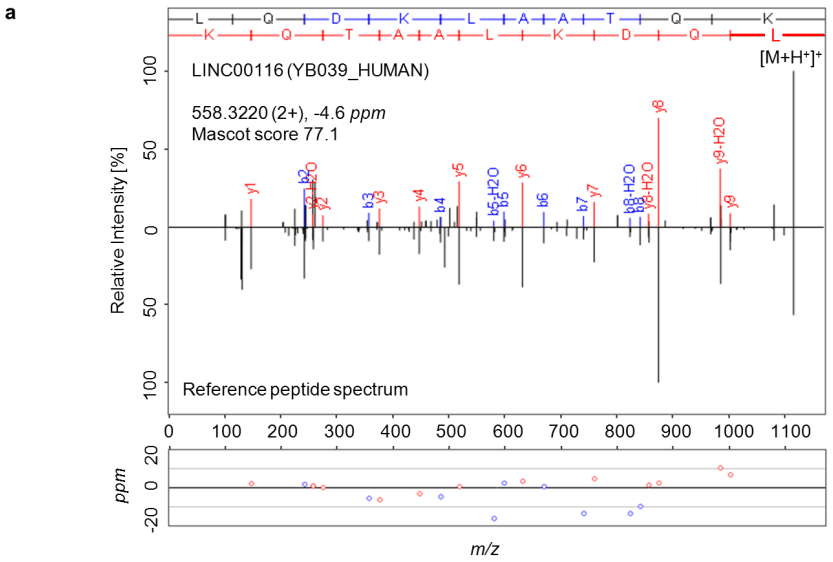
Supplementary Information is available in the online version of the paper.

Acknowledgements The authors wish to thank all originators of the mass-spectrometry-data used in this study for making their data available. We are grateful to P. Mallick, J. Cottrell and M. Schirle for conceptual discussions, to F. Pachl, S. Heinzlmeir, S. Klaeger, S. Maier, D. Helm, B. Ferreia, M. Frejno, H. Koch, M. Mundt, J. Zecha, D. Zolg, E. Gillmeier, B. Ruprecht, K. Kramer, G. Medard and X. Ku of TUM for the annotation of experiments, and to Y. Morad, A. Niadzelka, E. Kny, H. Cossmann, D. Schikora of SAP and V. Wichnalek, A. Klaus, M. Kroetz-Fahning, T. Schmidt of TUM for technical assistance.

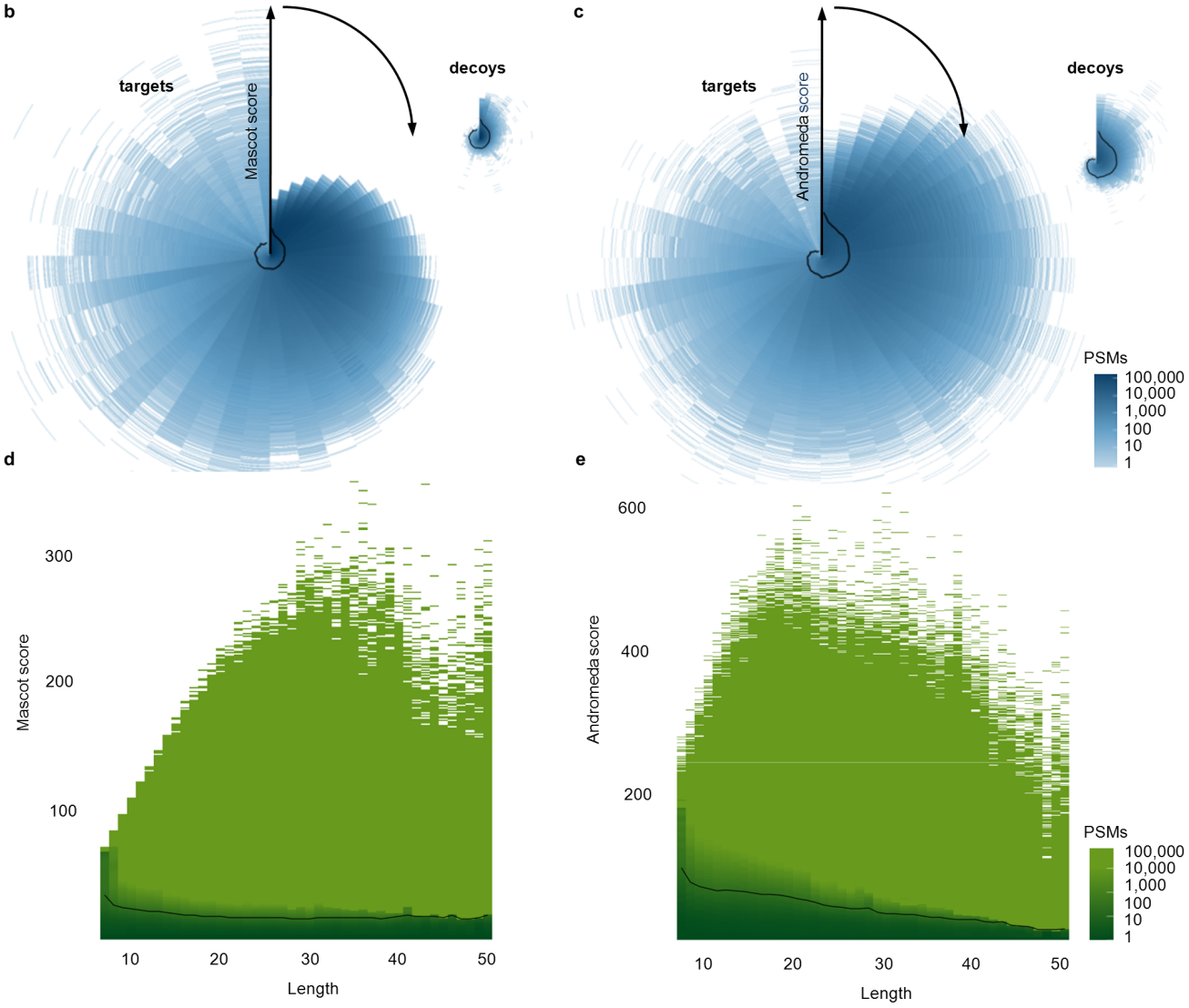
Author Contributions M.W., J.S., M.L., E.Z., L.B., J.-H.B., S.G., A.G., H.H., A.M.G. and B.K. designed ProteomicsDB. H.H., K.S., U.R., M.M. and J.S.-H. performed experiments. M.W., H.H., A.M.G., M.M.S., H.M., T.M., S.L. and B.K. performed data analysis. H.W., M.B., F.F. and B.K. conceptualized the study. M.W., H.H., A.M.G. and B.K. wrote manuscript.

Author Information Mass-spectrometry data are available from ProteomicsDB (https://www.proteomicsdb.org) and ProteomeXchange (http://proteomecentral.proteomexchange.org; dataset identifier PXD000865). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.K. (kuster@tum.de).

©2014 Macmillan Publishers Limited. All rights reserved

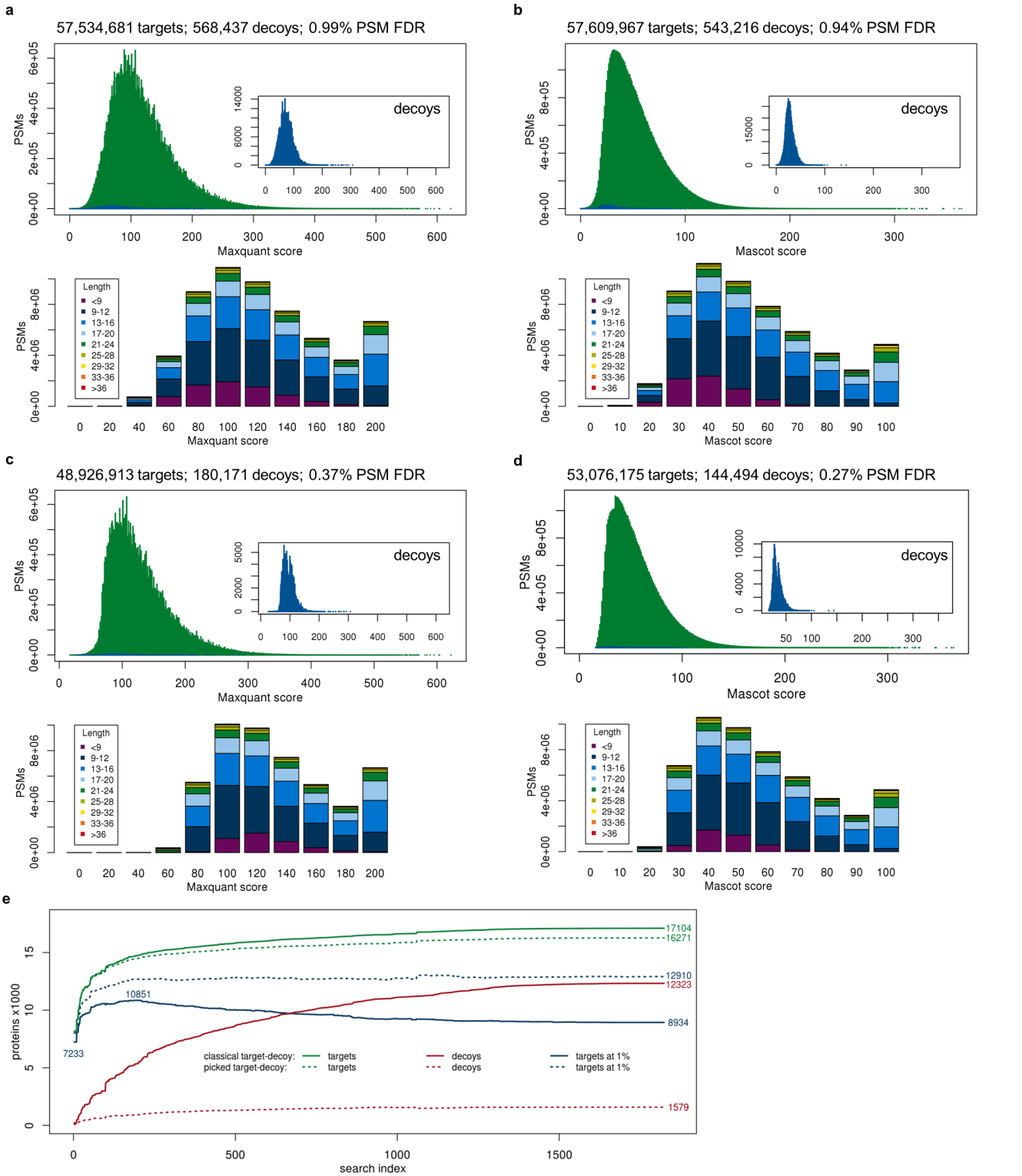


b+	Seq.		y+
114.09	1	L	10
242.14	2	Q	9	1002.55
357.17	3	D	8	874.49
485.27	4	K	7	759.47
598.35	5	L	6	631.37
669.39	6	A	5	518.29
740.43	7	A	4	447.25
841.47	8	T	3	376.21
969.53	9	Q	2	275.17
	10	K	1	147.11



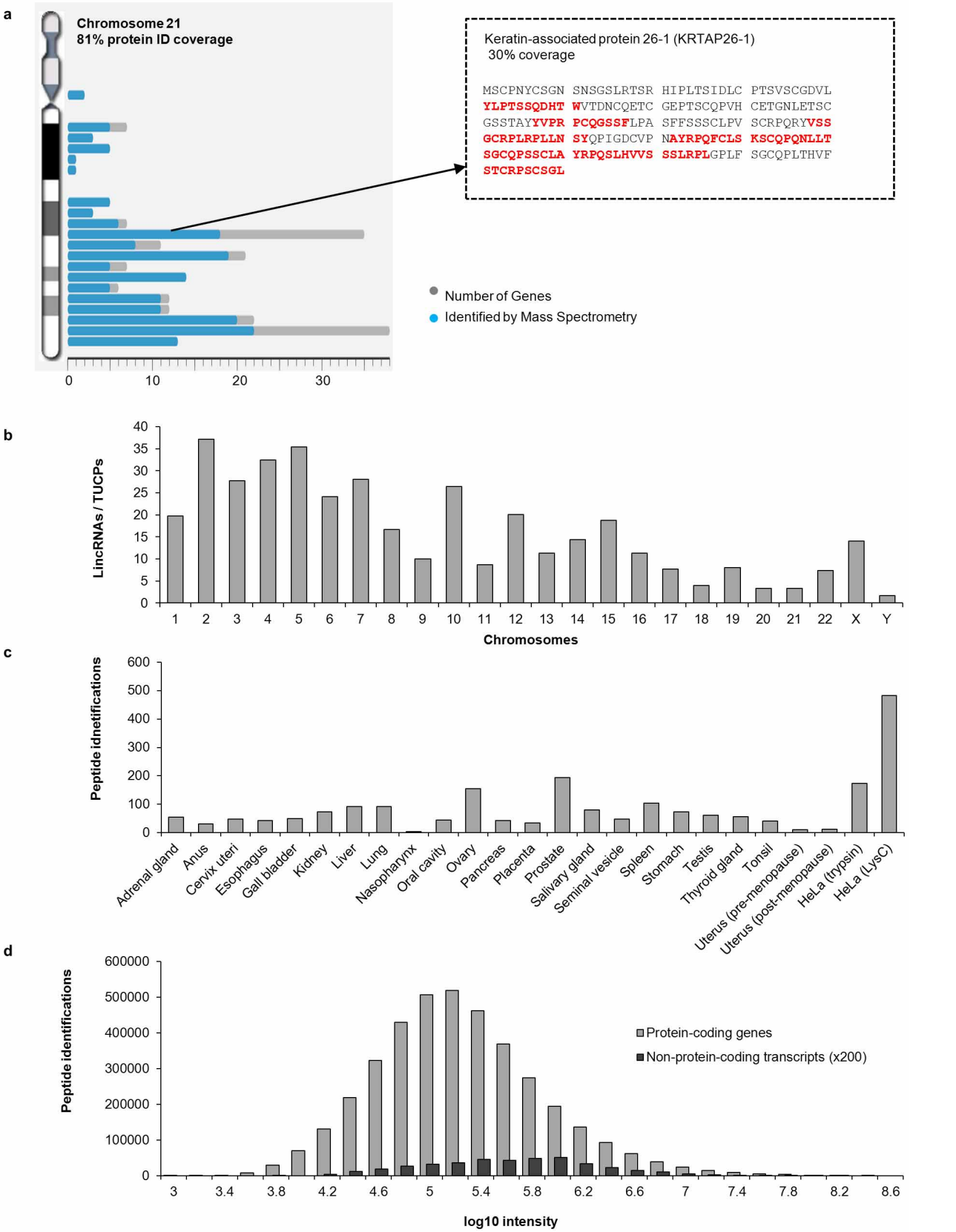
Extended Data Figure 1 | Peptide and protein identifications. **a**, Spectrum viewer enabling access to more than 70-million annotated tandem mass spectra of endogenous peptides and synthetic reference standards in real time. **b**, Peptide length and score distribution for targets and decoys for the search engine Mascot. It is of note that the peptide- and protein-identification criteria followed a two-step process. First, for each LC-MS/MS run, we applied a global 1% target-decoy false discovery rate (FDR) cut on the level of peptide spectrum matches (PSMs, not shown); second, we applied a peptide-length-dependent local FDR cut of 5% for all PSMs and the results are depicted here. **c**, Same as in **a** but for the search engine Andromeda. **d**, **e**, Heat maps showing FDRs as a function of search engine score and peptide length. Solid lines indicate the 5% local FDR.

©2014 Macmillan Publishers Limited. All rights reserved



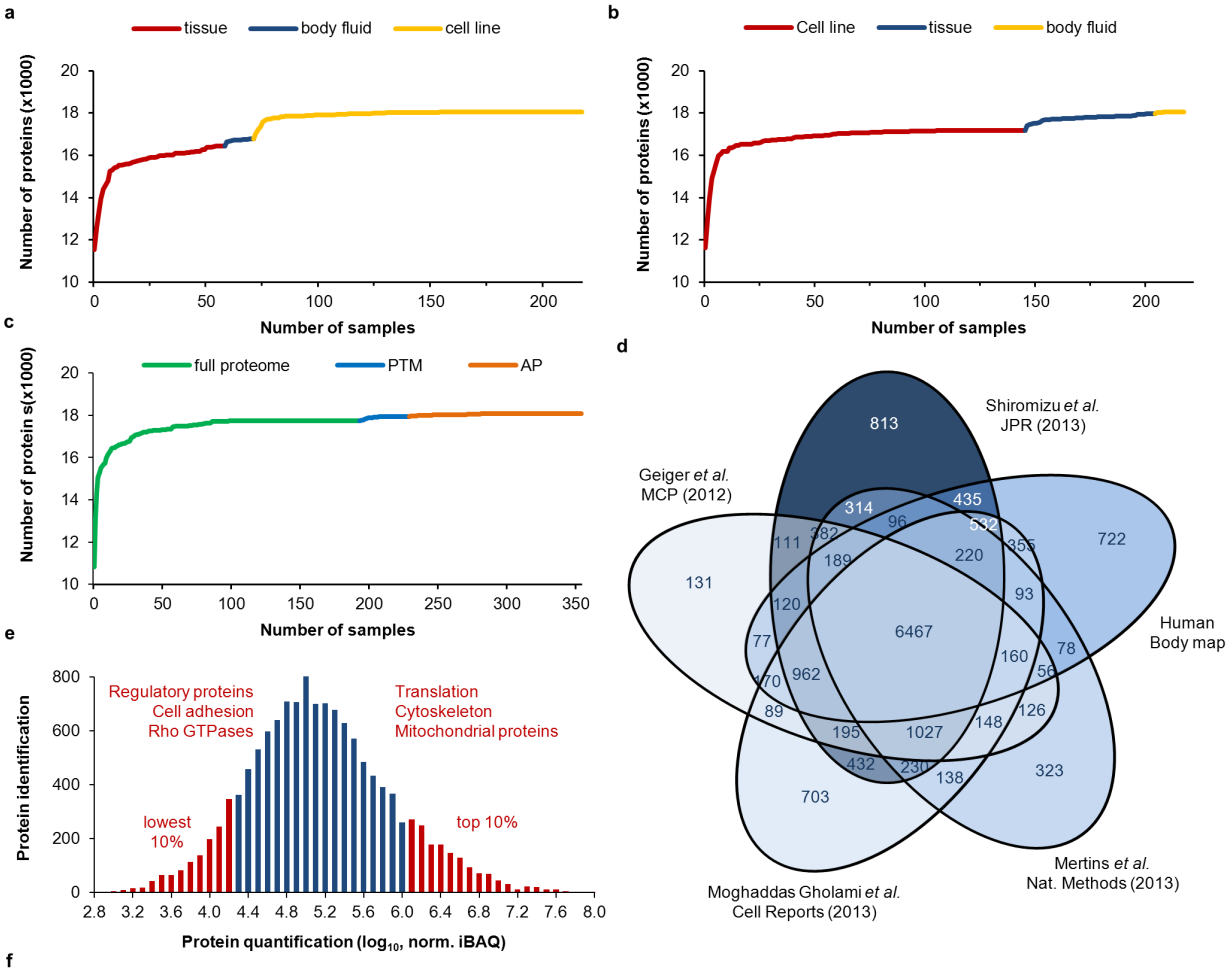
Extended Data Figure 2 | Protein-identification quality in very large data sets. a, First filtering step. The first step filters every LC-MS/MS run at 1% PSM FDR. Top panel, score distribution for target and decoy PSMs following 1% PSM FDR filtering for Maxquant identifications. Bottom panel, the binned peptide-length distribution for target PSMs. **b**, Same as **a** but for Mascot identifications. **c**, Second filtering step. Same as **a**, but this time applying an additional 5% local length- and score-dependent FDR on the total aggregated data for Maxquant identifications in ProteomicsDB. It is apparent that the second filtering step improves the FDR about threefold and removes most PSMs shorter than 9 amino acids. **d**, Same as **c** but for Mascot identifications in ProteomicsDB. **e**, Comparative analysis of protein FDR characteristics of two different approaches based on Mascot analysis. In the classical target-decoy approach, aggregation of large quantities of data leads to accumulation of large numbers of decoy proteins and a concomitant loss of true target proteins when filtering the data at 1% protein FDR. The alternative 'picked' target-decoy method does not suffer from this scaling problem and maintains a constant decoy rate (and therefore lower protein FDR) but at the expense of lower sensitivity of target protein detection compared to the classical target-decoy approach. Please refer to the Supplementary Information for details and a discussion on the topic. Note that the two protein FDR methods were not used in this manuscript. Instead, we used the criteria shown in **a** and **b**.

©2014 Macmillan Publishers Limited. All rights reserved



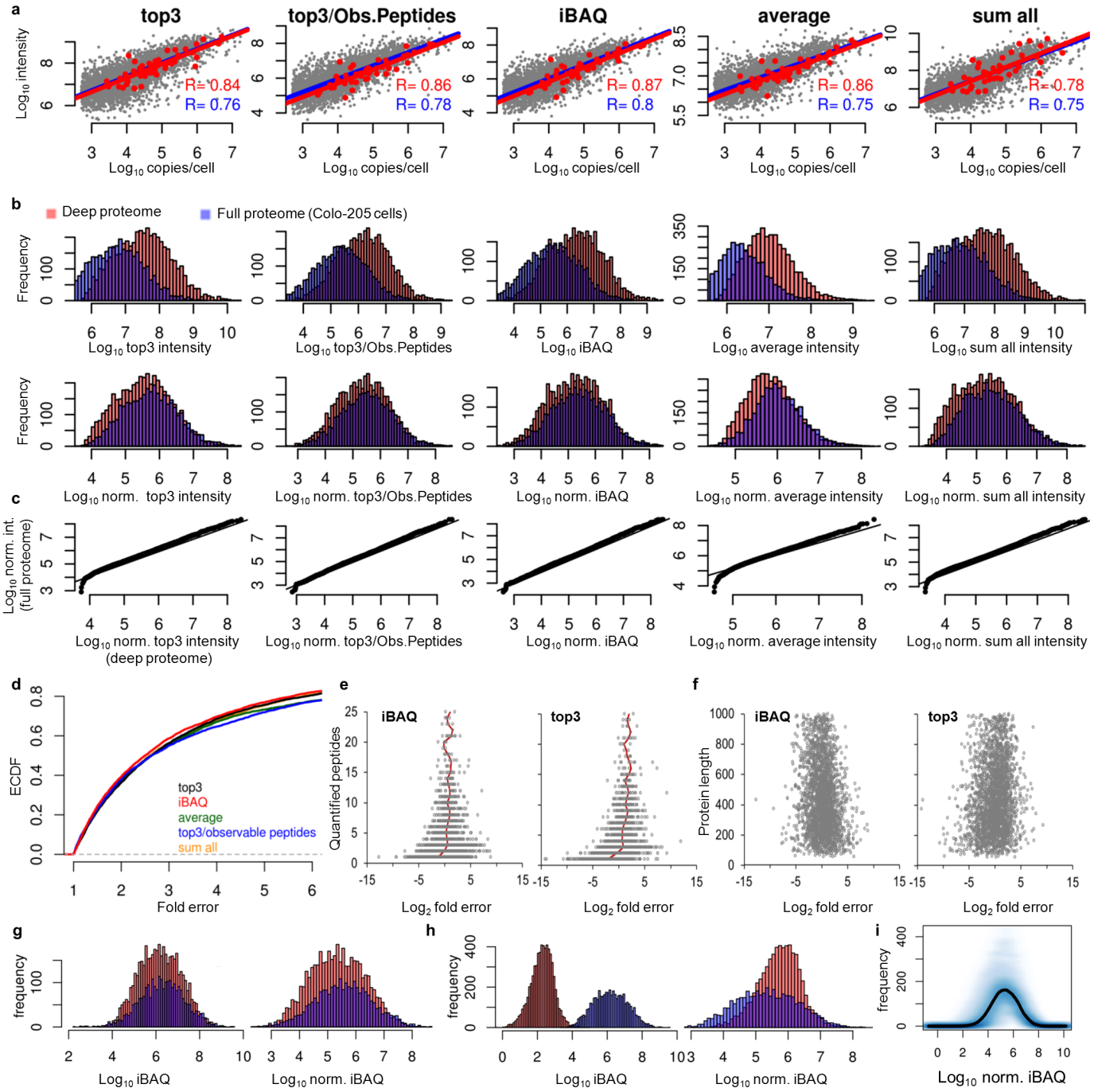
Extended Data Figure 3 | Further characterization of the proteome. **a**, Some proteins are refractory to identification using tryptic digestion because they do not generate sufficient—or any—peptides that are within the productive mass range of a mass spectrometer typically used for bottom-up proteomics. This can be improved by the use of alternative proteases; for example, chymotrypsin as shown here for one of the many keratin-associated proteins localized on chromosome 21 (detected chymotryptic peptides in red). **b**, **c**, Translation of lincRNAs is rare but does exist and can be identified (**b**) across all chromosomes as well as (**c**) in many tissues and in HeLa cells. **d**, Peptide-intensity distribution of protein-coding genes and non-coding transcripts. Interestingly, the abundance of translated lincRNAs is broadly similar to that of classical proteins.

©2014 Macmillan Publishers Limited. All rights reserved



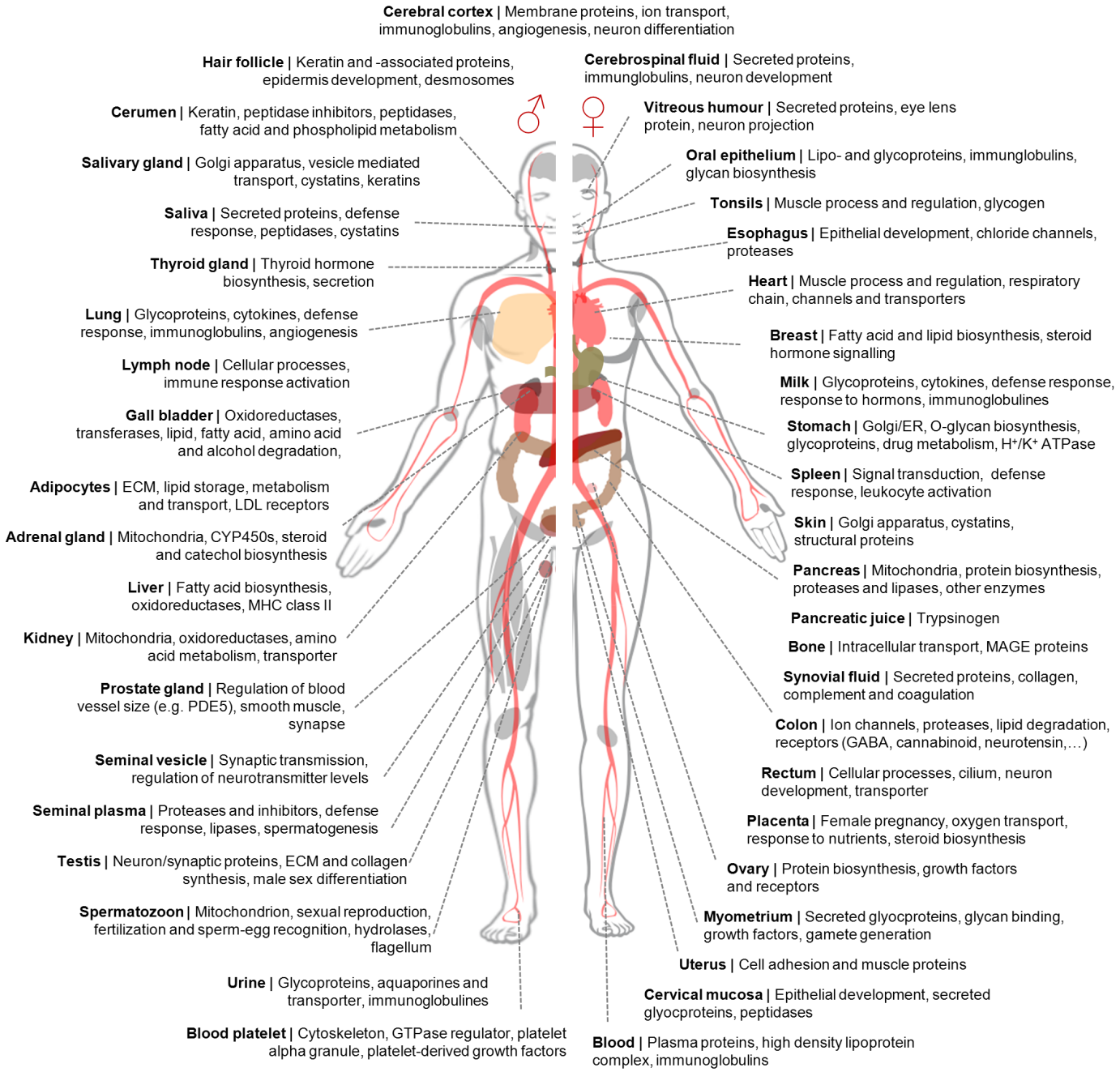
Extended Data Figure 4 | Further characterization of the proteome. **a**, Proteome coverage rapidly saturates with the addition of shotgun proteomic data. Tissue proteomes saturate at ~approximately 16,000 proteins, but both body fluids and cell lines add small but noticeable numbers of proteins not covered in the tissues (see also **b** and **c** for a different ordering of samples). This indicates that proteome coverage is likely not to increase much more by merely adding high-throughput data (although it may increase confidence in protein identifications and will probably also increase sequence coverage). **b**, Same plot as **a** but different ordering of samples. **c**, Saturation plots showing that PTMs and affinity purifications each contribute distinctly to the coverage of the proteome. **d**, Comparison of five large-scale projects suggesting that a 'core proteome' of 10,000–12,000 ubiquitously expressed proteins exists. Ellipses represent the corresponding publications. **e**, Abundance distribution of the 'core proteome' based on the normalized iBAQ method. The most highly expressed 10% of proteins are dominated by proteins relating to energy production and protein synthesis. The least abundant 10% of proteins are enriched in proteins with regulatory functions. **f**, Tree-view summary of Gene Ontology (GO) term analysis for the proteins constituting the 'core proteome', showing that the core proteome is mainly concerned with biological processes relating to the homeostasis and life cycle of cells. The colours represent the broader categories of the treemap.

©2014 Macmillan Publishers Limited. All rights reserved



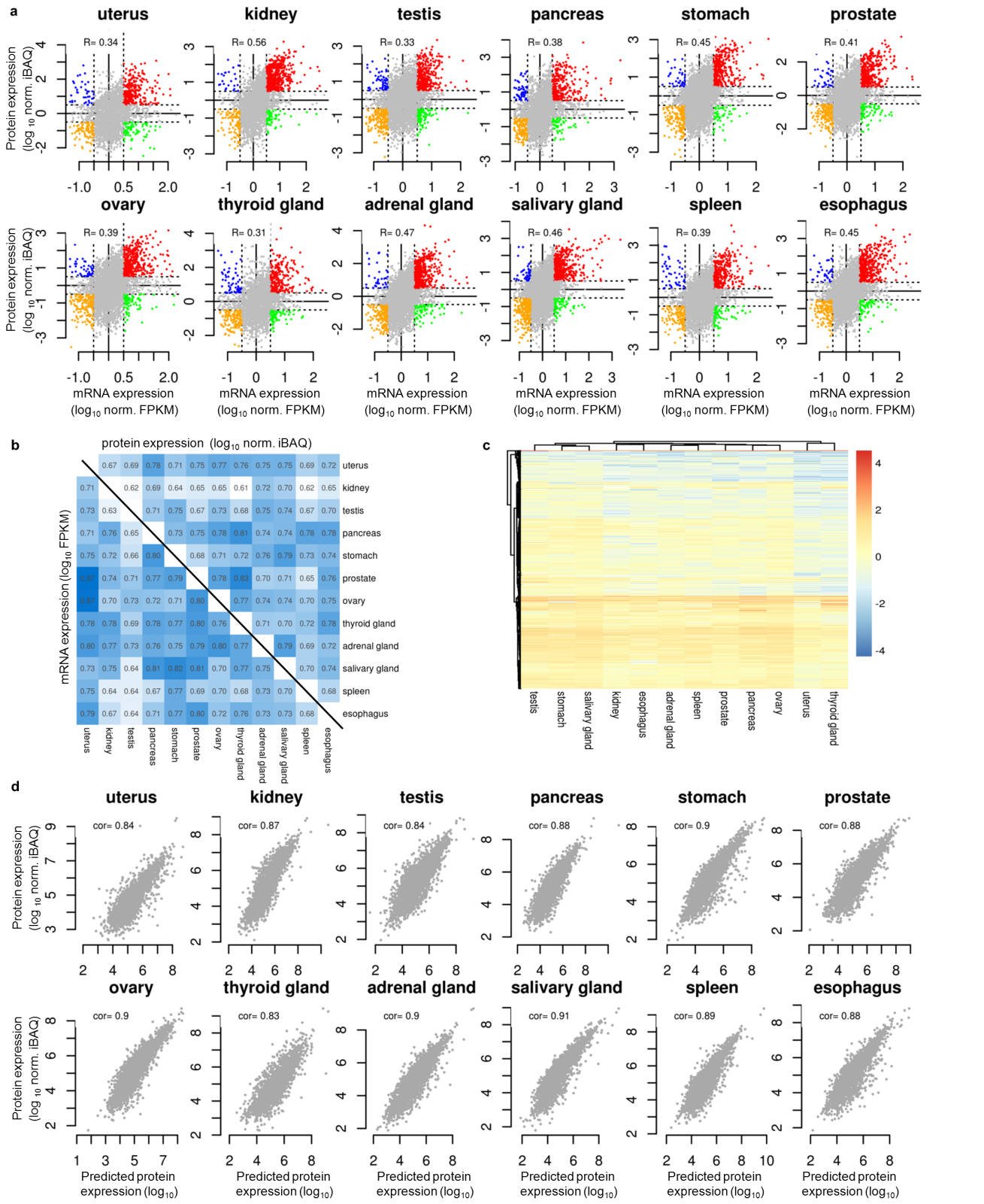
Extended Data Figure 5 | Comparative analysis of five intensity-based label-free absolute-quantification approaches. **a**, Linearity of intensity (U2-OS cell line data from ref. 22) and copies per cell for absolute protein quantification (AQUA)-quantified proteins (red dots, red regression line; same cell line[30]) and derived copy-number estimates (grey dots, blue regression line; from the same study). **b**, Total sum normalization re-scales intensity distributions of Colo-205 cell digests measured on two different mass spectrometers (Orbitrap Elite data in red, LTQ Orbitrap XL data in blue[24]). **c**, Quantile-quantile (Q-Q) plots of the normalized data presented in **b** illustrating good alignment of data across 4.5 orders of magnitude. **d**, Empirical cumulative density function (ECDF) of error distributions derived from **a** showing that all five methods have merit. **e**, Comparison of the fold error of iBAQ and top3 as a function of the number of quantified peptides. **f**, Same as **e** but for protein length. When peptide numbers are low, iBAQ shows errors that are slightly smaller in magnitude compared to the top3 method. **g**, Comparison of iBAQ and total sum normalized iBAQ for heavy SILAC-labelled MCF-7 cell digests (red bars[32] and label-free quantified MCF-7cell digests (same as MCF-7 deep proteome in **a**; blue bars) before (left panel) and after normalization (right panel) showing no influence of the presence of the SILAC label on quantification results. **h**, Comparison of iBAQ and total sum normalized iBAQ for iTRAQ reporter-ion-intensity-based quantification (red bars; MCF-7 cell digest[46]) and label-free quantified MCF-7 cell digests (blue bars; same as **a** and **c**) before (left panel) and after normalization (right panel). The intensity-distribution characteristics of iTRAQ and label-free measurements are too different to allow for comparative analyses of MS1- and MS2-based quantification data. **i**, Normalized iBAQ distributions of 347 cell-line and tissue proteomes (all MS1 quantified) available in ProteomicsDB showing the general applicability of MS1-based quantification across many sources of biological material.

©2014 Macmillan Publishers Limited. All rights reserved



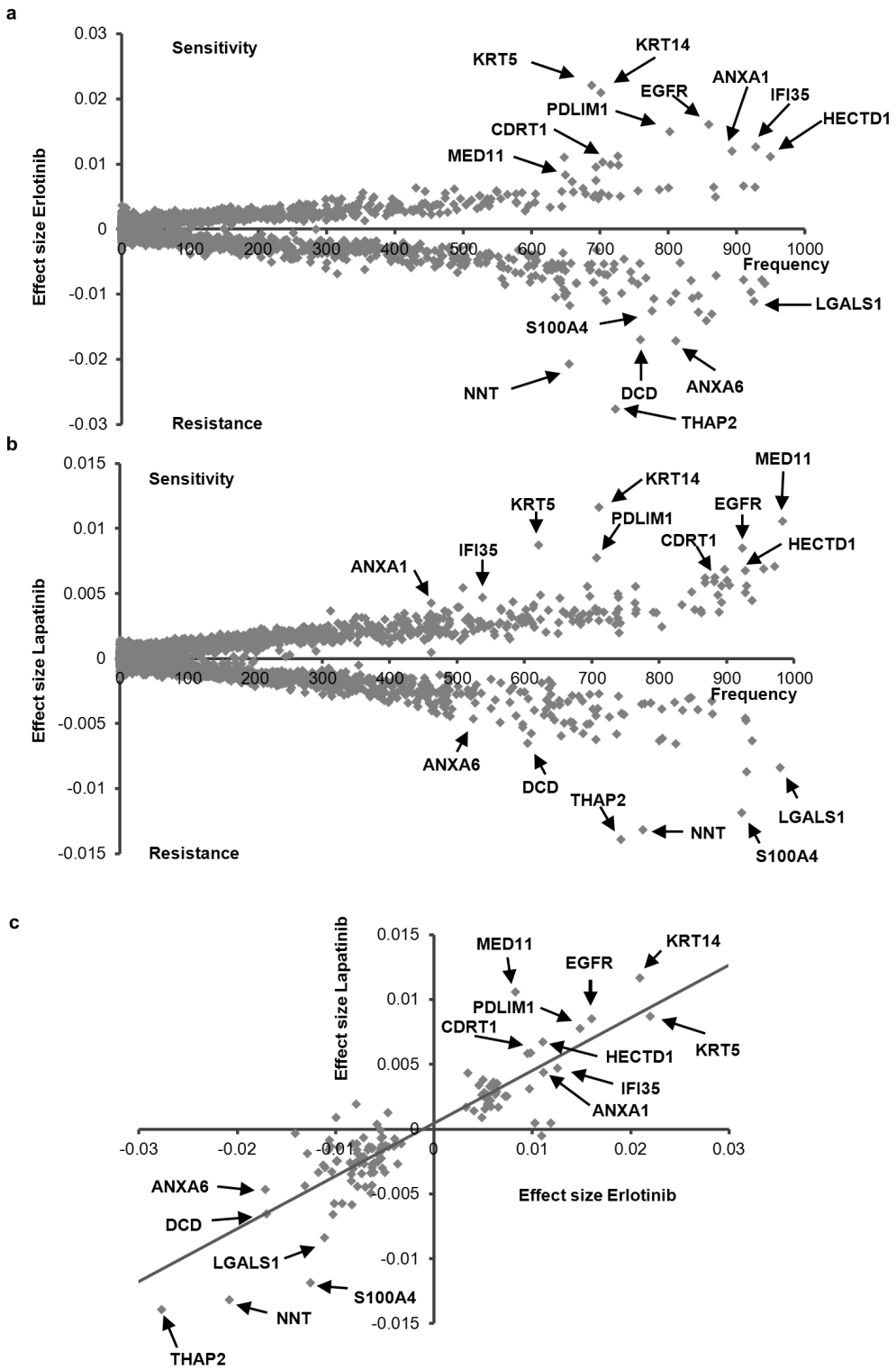
Extended Data Figure 6 | Functional protein-expression analysis. Gene ontology analysis of proteins with expression levels 10-fold above average in a particular organ or body fluid invariably highlights protein signatures with direct organ-related functional significance.

©2014 **Macmillan Publishers Limited. All rights reserved**



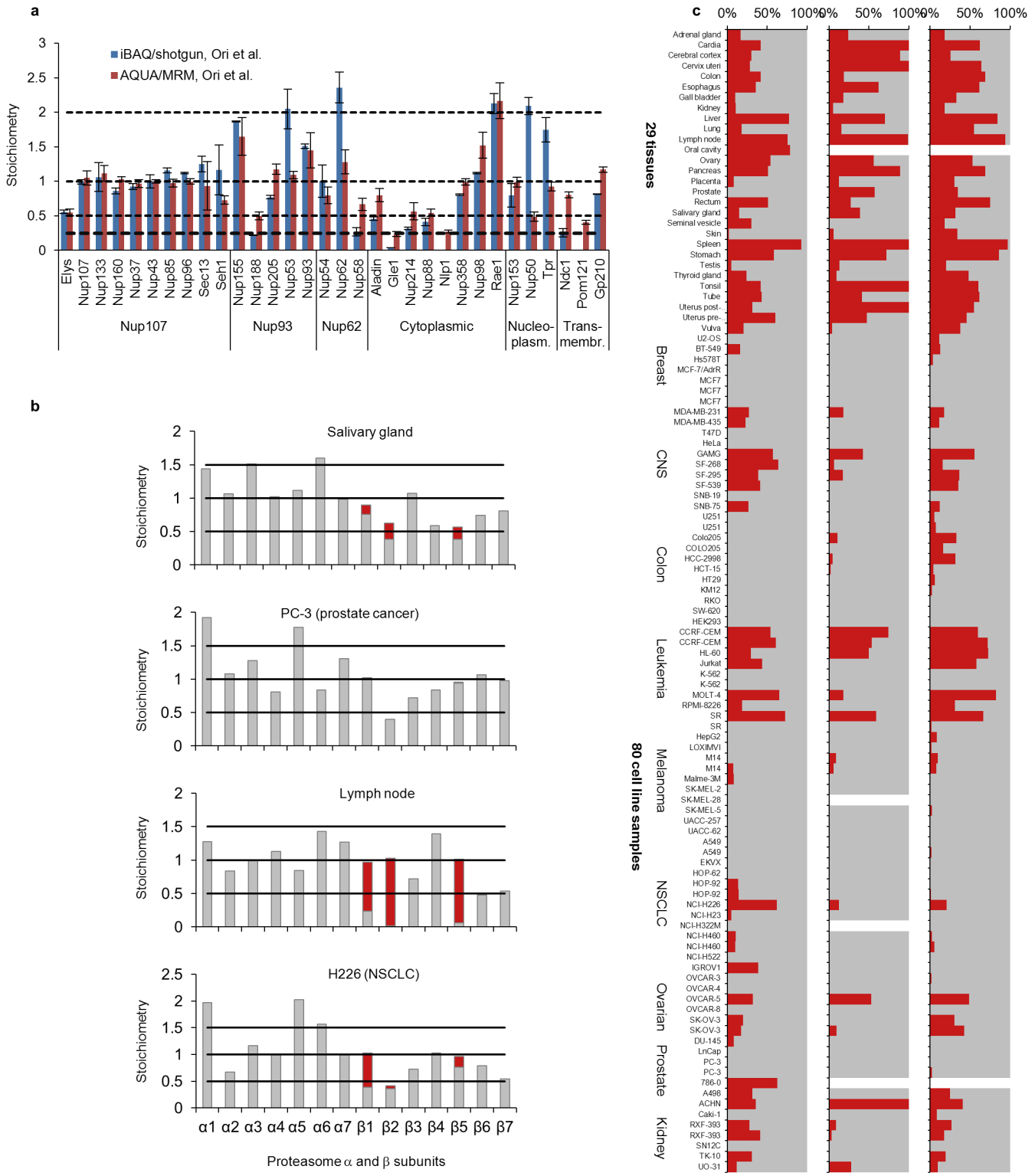
Extended Data Figure 7 | Protein- versus mRNA-expression analysis. **a**, Comparison of mRNA and protein expression of 12 human tissues showing the general rather poor correlation of protein and mRNA levels, implying the widespread application of transcriptional, translational and post-translational control mechanisms of protein-abundance regulation. Spearman correlation coefficients vary from 0.41 (thyroid gland) to 0.55 (kidney). 'Corner proteins' (0.5 logs to either side of zero) are marked in colours. **b**, Clustering of mRNA expression (left triangle) and protein expression (right triangle) across the 12 tissues does not reveal tissues with common profiles suggesting that the transcriptomes and proteomes of human tissues are quite different from each other. **c**, The ratio of protein and mRNA level for a protein is approximately constant across many tissues. The heat map shows proteins and tissues clustered according to their protein/mRNA ratio. **d**, Protein abundance can be predicted from mRNA levels. Using the median ratio of protein/mRNA across 12 tissues, it is possible to predict protein levels from mRNA levels for every tissue with a good correlation coefficient, underscoring the importance of the translation rate (and mRNA levels) on protein expression.

©2014 Macmillan Publishers Limited. All rights reserved



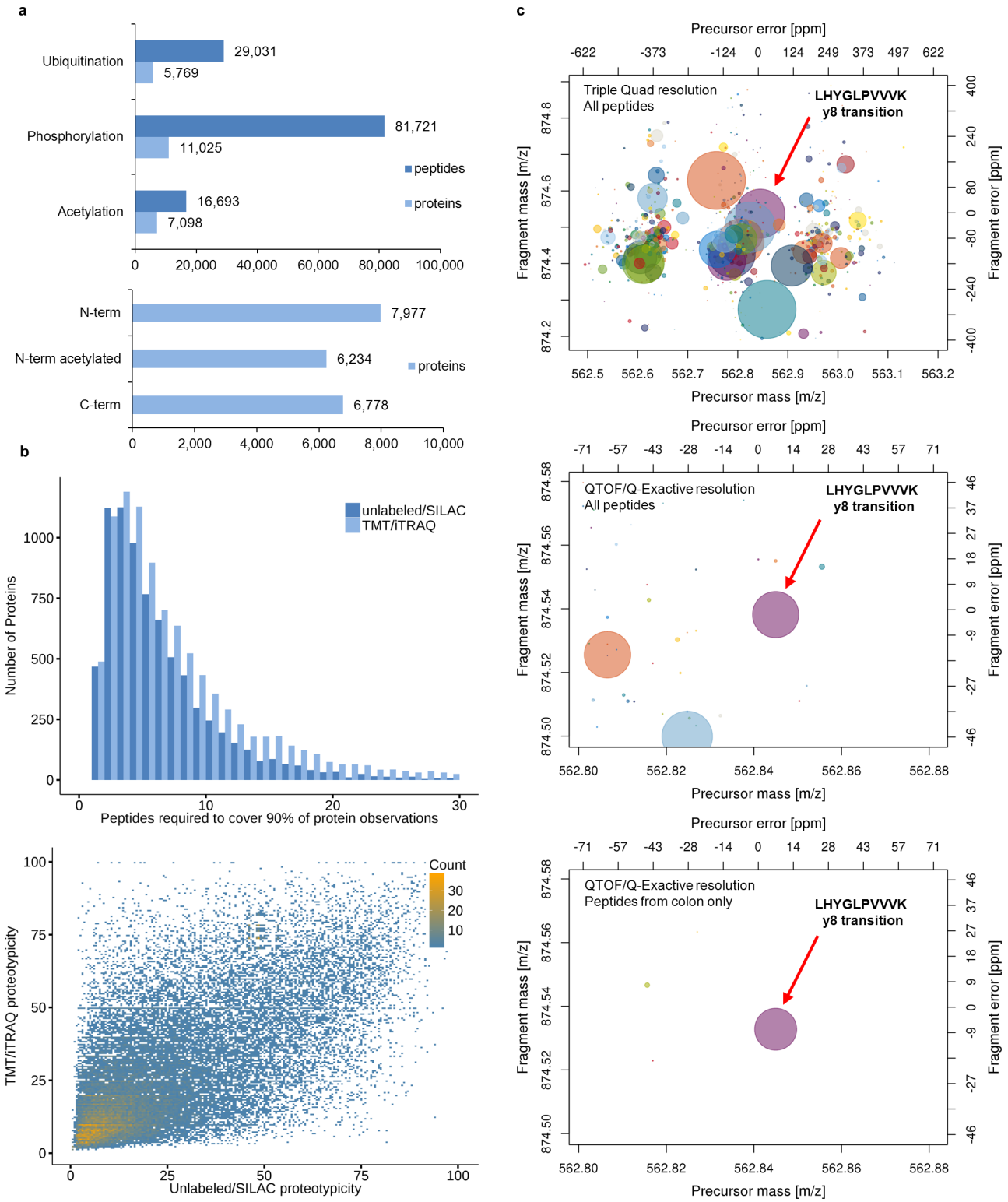
Extended Data Figure 8 | Protein markers for drug sensitivity and resistance. **a**, Elastic net analysis of protein expression and drug sensitivity for the EGFR kinase inhibitor erlotinib. Positive-effect-size values indicate that high protein expression is associated with drug sensitivity. Negative-effect-size values indicate that high protein expression is associated with drug resistance. **b**, Same as in **a** but for the EGFR kinase inhibitor lapatinib. **c**, Correlation analysis of the elastic net effect sizes for erlotinib and lapatinib (proteins with elastic net frequencies of less than 600 are not shown for clarity). Proteins in the top-right quadrant are common markers for drug sensitivity (including EGFR as the primary target of both drugs). Proteins in the bottom-left quadrant are common markers for drug resistance (including S100A4, a known resistance marker for lapatinib). Proteins that are strong markers for sensitivity or resistance are annotated in each plot and most proteins can be easily placed into EGFR signalling and regulation pathways.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Protein complex composition and stoichiometry from shotgun proteomic data. **a**, Stoichiometry of the nuclear pore complex (NPC) reconstructed from shotgun proteomics data. To illustrate that normalized iBAQ values from shotgun experiments actually reflect protein copy numbers, we reconstructed the stoichiometry of the NPC (blue bars, data from nuclear extracts of HeLa cells[39]; error bars indicate standard deviation from triplicate experiments) and compared it to the stoichiometry determined in the same study using AQUA peptides and SRM experiments (red bars). Note that most of the time, the stoichiometries are in very good agreement between the methods and the stoichiometries reported in the literature. **b**, Stoichiometry of the α- and β-subunits of the proteasome reconstructed from shotgun proteomics data (examples). β-subunits of the constitutive proteasome are indicated in grey, immunoproteasome subunits (β1i, β2i, β5i) are indicated in red. Note that PC-3 cells are devoid of the immunoproteasome, whereas cells in the lymph node almost exclusively express this version of the molecular machine. **c**, Systematic assessment of the fraction of βi subunits (red bars) and β-subunits (grey bars) across 29 tissue samples and 80 cell-line samples (tissue data from human body map (this study), cell-line data from[22,24]). Note that many cell lines and tissues contain both versions of the proteasome and the data also suggest that further forms of the proteasome with different subunit compositions may exist.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Examples for the analytical utility of large mass-spectrometry-based data collected in ProteomicsDB. **a**, Enumeration of post-translational modifications and protein termini. **b**, Computation of proteotypic peptides. Generally the same one to five peptides are identified every time a protein is identified (top panel) making proteotypic peptides useful for assessing protein identification and as reagents for targeted mass-spectrometry measurements. We note that the proteotypicity of a peptide strongly depends on the presence or absence of a chemical modification (bottom panel, here tandem mass tags (TMT) or isobaric tags for relative and absolute quantification (iTRAQ)). **c**, Analysis of the selectivity of SRM transitions. The top panel shows the y8 transition of the peptide LHYGLPVVVK (β-catenin, marked with an arrow) in a slice of the precursor and fragment-ion window of 0.7 Da and 0.7 Da, respectively, typically employed on triple-quadrupole mass spectrometers. The size of the circle represents the relative intensity of the y8 fragment in a full tandem mass spectrum of this peptide. All other circles are interfering peptides (extracted from the entire ProteomicsDB) that have precursor and fragment ions in the same *m/z* window and with varying intensities (circle size). Interference can be reduced by using high-resolution mass spectrometry (middle panel) and confining the analysis to the tissue in question (here, a colon sample, bottom panel). Such interference plots in conjunction with the proteotypicity of peptides can be valuable for the design of targeted proteomic experiments.

©2014 Macmillan Publishers Limited. All rights reserved

46. Johannsson, H. J. *et al.* Retinoic acid receptor alpha is associated with tamoxifen resistance in breast cancer. *Nature Commun.* **4,** 2175 (2013).

©2014 **Macmillan Publishers Limited. All rights reserved**