Molecular Systems Biology 7; Article number 548; doi:10.1038/msb.2011.81
Citation: ***Molecular Systems Biology* 7: 548**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Deep proteome and transcriptome mapping of a human cancer cell line

Nagarjuna Nagaraj[1], Jacek R Wisniewski[1], Tamar Geiger[1], Juergen Cox[1], Martin Kircher[2], Janet Kelso[2], Svante Pääbo[2] and Matthias Mann[1,*]

[1] Department for Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany and [2] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* Corresponding author. Department for Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. Tel.: +49 89 8578 2557; Fax: +49 89 8578 2219; E-mail: mmann@biochem.mpg.de

Received 15.7.11; accepted 29.10.11

While the number and identity of proteins expressed in a single human cell type is currently unknown, this fundamental question can be addressed by advanced mass spectrometry (MS)-based proteomics. Online liquid chromatography coupled to high-resolution MS and MS/MS yielded 166 420 peptides with unique amino-acid sequence from HeLa cells. These peptides identified 10 255 different human proteins encoded by 9207 human genes, providing a lower limit on the proteome in this cancer cell line. Deep transcriptome sequencing revealed transcripts for nearly all detected proteins. We calculate copy numbers for the expressed proteins and show that the abundances of >90% of them are within a factor 60 of the median protein expression level. Comparisons of the proteome and the transcriptome, and analysis of protein complex databases and GO categories, suggest that we achieved deep coverage of the functional transcriptome and the proteome of a single cell type.
Molecular Systems Biology **7**: 548; published online 8 November 2011; doi:10.1038/msb.2011.81
Subject Categories: Proteomics
Keywords: mass spectrometry; proteomics; RNA-Seq; systems biology; transcriptomics

Introduction

An inventory of the building blocks of a biological system is a prerequisite for a systems-wide understanding of its functions. For human genes this was enabled by the sequencing of the human genome, which yielded the unexpected result that the genome is comprised of a mere 20 000 protein-coding genes (Clamp *et al*, 2007). In contrast, the number of distinct transcripts has increased drastically due to the development of very deep—'next generation'—shotgun sequencing of transcriptomes, termed RNA-Seq (Mortazavi *et al*, 2008; Wang *et al*, 2009). Depending on the nature of the data and analysis criteria (Guttman *et al*, 2010; Haas and Zody, 2010; Trapnell *et al*, 2010), transcripts of between 8000 and 16 000 protein-coding genes expressed from a single cell type can be detected.

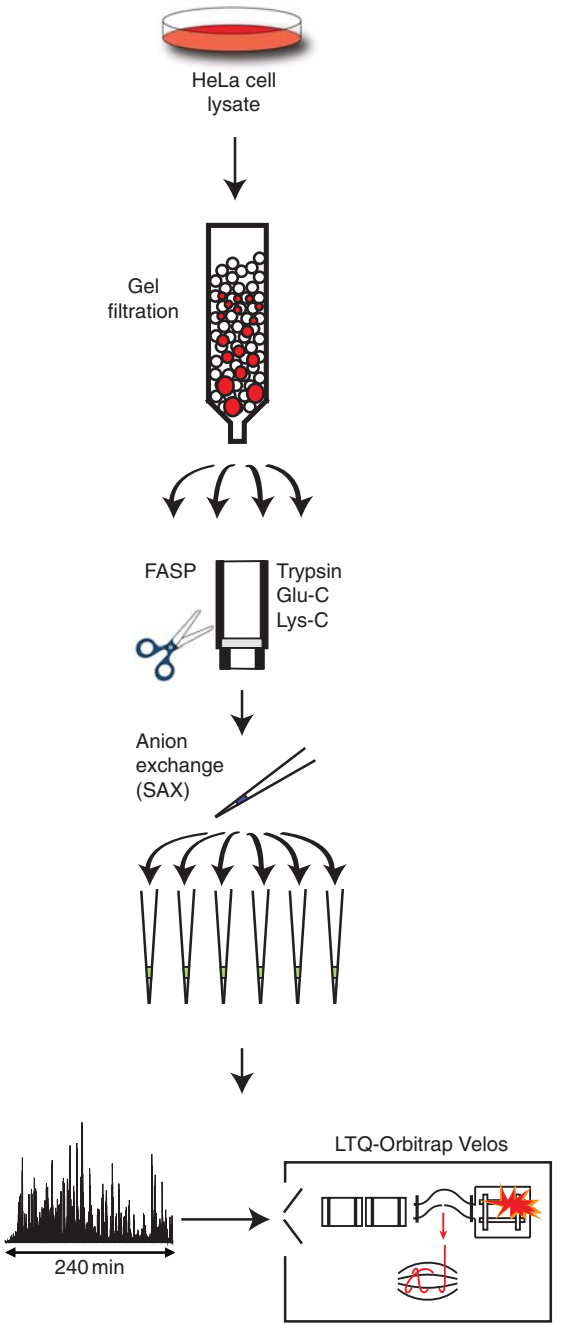
High-resolution mass spectrometry (MS)-based proteomics has improved at a rapid pace in recent years (Aebersold and Mann, 2003; Mallick and Kuster, 2010; Schwanhausser *et al*, 2011). These advances had allowed us to quantify an essentially complete proteome of the model organism yeast as judged by comparison with genomic tagging methods (de Godoy *et al*, 2008). In mammalian systems, in contrast, our depth of analysis in single cell types has typically been limited to 4000–6000 protein groups (proteins distinguishable by identified peptides) (Graumann *et al*, 2008; Lundberg *et al*, 2010; Wisniewski *et al*, 2009a). Here, we set out to explore a human proteome in the depth achievable with current technology and to compare it with the corresponding transcriptome.

Results and discussion

We chose to investigate HeLa cells, a human cervical carcinoma cell line, because it is widely used in research and because a cell line is a more homogeneous system compared with tissues. To achieve maximum proteome coverage while maintaining a reasonable measurement time, we investigated the effects of protein fractionation, proteolytic digestion, peptide fractionation and reverse phase chromatography on the number of proteins identified (Figure 1). We employed moderate fractionation at the protein level by gel filtration, digestion by three specific proteases, combined with pipette-based prefractionation at the peptide level by strong anion exchange (Wisniewski *et al*, 2009a) before online LC MS/MS analysis in 4 h gradients with relatively long columns (40 cm, 1.8 μm bead material). Peptide MS spectra as well as fragment MS/MS spectra were measured with high resolution and mass accuracy (Mann and Kelleher, 2008; Olsen *et al*, 2007; Olsen *et al*, 2009).

On the basis of initial results ('Experiment 1'), we generated a data set ('Experiment 2')—involving 72 fractions and a total measuring time of 288 h—which is the basis of all subsequent discussion. All data files were analyzed together in the MaxQuant computational proteomics environment (Cox and Mann, 2008). A total of 2 337 336 high-resolution fragmentation spectra, together with the corresponding high-accuracy precursor masses, were submitted to the Andromeda search engine (Cox *et al*, 2011). Median peptide score was 121, with only 6% below a score of 60 (Supplementary Figure S1) and

A



Exp. 1: 126 fractions
Exp. 2: 72 fractions

B

Experiment	Proteins	Peptides	MS measurement time
Exp 1	*10 596*	*187 006*	*21 days*
Exp 2	10 255	163 784	12 days

Figure 1 Deep proteomic analysis of HeLa cells. (**A**). Proteome preparation workflow included protein separation by gel filtration followed by three FASP digestions per fraction, followed by strong anion exchange fractionation. Each fraction was analyzed by LC MS/MS on an LTQ-Orbitrap Velos mass spectrometer. (**B**) Summary of protein and peptide identifications obtained in the two experiments.

the average identification of the fragmentation spectra was 43%. Average absolute mass deviation of the precursors was 1.2 and 4.8 p.p.m. for the matched fragment masses. This identified and quantified 163 784 peptides that have unique amino-acid sequence at a false discovery rate (FDR) of 1%, many of them fragmented multiple times (seven on average). Of these, 84 051 were from tryptic digestion, 52 108 from LysC and 44 704 from GluC. From these data, MaxQuant identified 10 255 proteins with 99% confidence (Figure 1B; Supplementary Table S1), providing a lower bound of the number of proteins expressed in HeLa cells. Trypsin digestion produces peptides in an ideal size range for MS/MS and, consequently, it yielded the highest number of identifications. Of the proteins identified after LysC digestion, 85% overlapped with the trypsin data set, and the GluC data only added another 5.2% of novel identifications. Less than 5% of all proteins were only identified by one peptide. Taken together, the three proteases resulted in >24% median sequence coverage of identified proteins.

The 10 255 proteins were mapped to 9207 Ensembl-annotated human protein-coding genes (Hubbard *et al*, 2002). These genes were equally distributed across the different human chromosomes with most and least number of genes identified in chromosomes 1 and 21, respectively (Supplementary Figure S2; Supplementary Table S2). Further, the MS/MS spectra were searched against the ENSEMBL database together with the GENSCAN predictions. This led to >1900 peptides mapping only to the GENSCAN predictions and not to the known ENSEMBL genes. We provide a list of the highest scoring of these peptides, as they may point to as yet unannotated exons (Supplementary Table S3).

To compare the proteome with the transcriptome and to evaluate the completeness of our results, we performed RNA-Seq on the same cells. Briefly, we acquired 50 million single-end 76 bp cDNA reads on the Illumina GAIIx platform. Reads were mapped to the human reference genome sequence and assembled into 49 000 unique transcripts (Trapnell *et al*, 2010) that mapped to 16 554 different protein-coding genes (Supplementary Table S4). The abundance of the non-filtered data expressed as Fragments Per Kilobase of exon per Million fragments mapped (FPKM) shows a bimodal distribution (Figure 2A) where about 33% of the transcripts have low signals below one FPKM. When excluding transcripts expressed at abundances lower than one FPKM, the number of genes identified was reduced to about 11 000, and genes corresponding to hundreds of low abundance proteins identified by MS were lost (Figure 2A). We therefore excluded transcripts for which the estimated abundance is lower than their 95% confidence interval (FPKM $>\Delta_{95}$ FPKM). Using this criterion, transcripts for 11 936 protein-coding genes were detected including a considerable number of transcripts in the low abundance region for which no proteins were detected. These include many genes that are not expected to be functionally relevant in HeLa cells, such as olfactory receptors (Supplementary Figure S3). The distribution of protein abundance values is broader than the filtered mRNA abundance distribution but has the same general shape (Figure 2B). Recently, the bimodal distribution of the transcriptome has been investigated in detail. The transcripts in the left part of the distributions appear to be present at less than one copy per cell and often code for functions not represented in the cell type

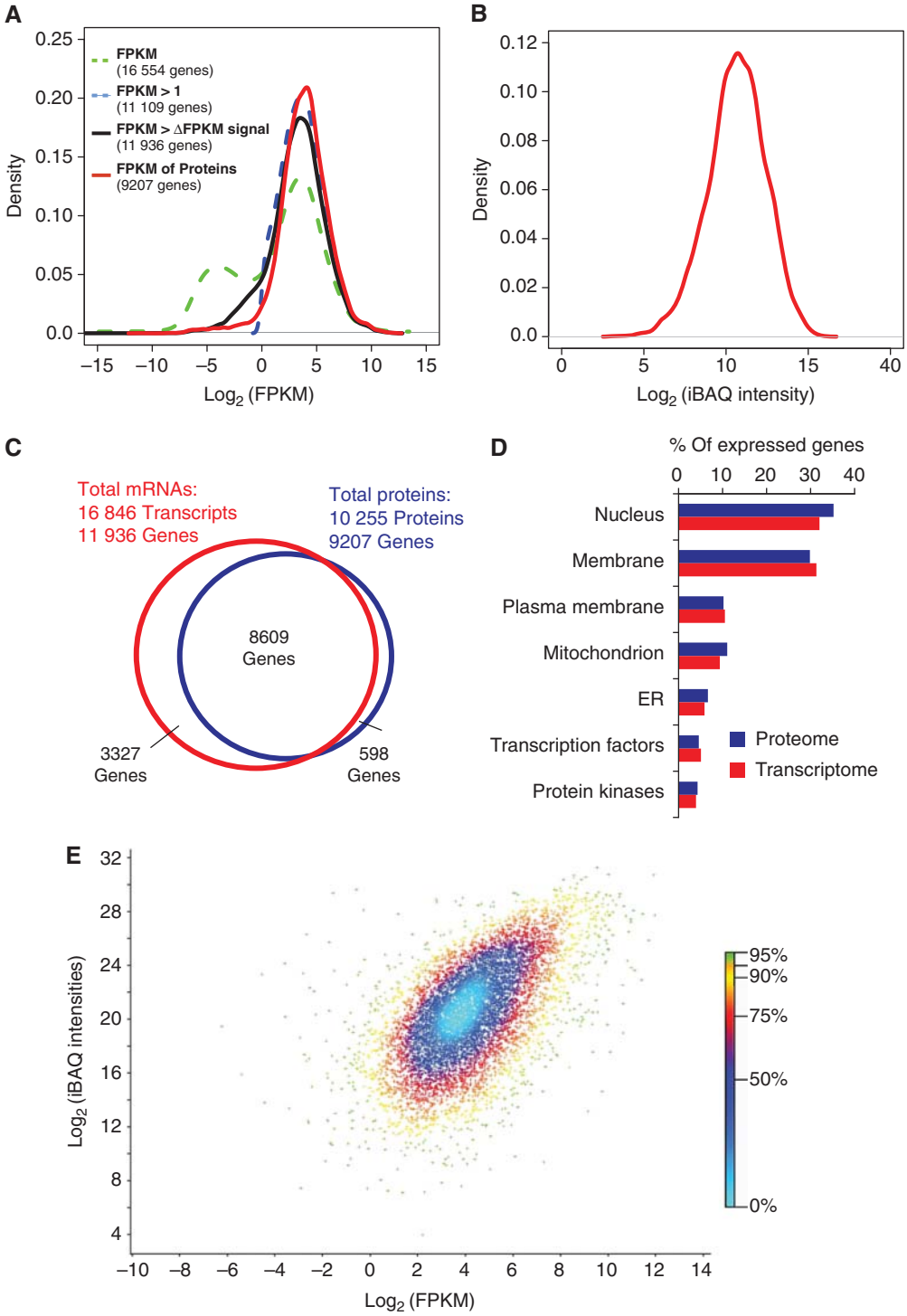


Figure 2 Comparison of proteomics and RNA-Seq data. (**A**) Distribution of FPKM data before filtration (green), filtered data with an FPKM threshold of 1 (blue) or based on the 95% confidence interval (ΔFPKM, black). FPKM values of the identified proteins are shown in red. (**B**) Distribution of abundance of proteins (iBAQ intensities) identified with FDR of 1%. (**C**) Venn diagram of the number of expressed genes on the mRNA level and on the protein level. (**D**) Proportions of proteins and transcripts annotated to various cellular compartments and molecular functions. (**E**) A density scatter plot of iBAQ intensities versus FPKM values. The color code indicates the percentage of points that are included in a region of a specific color.

(Hebenstreit *et al*, 2011). Therefore, it is possible that many of these transcripts are not expressed as proteins. Together, the data suggest that the detected proteome covers a very large part of the transcripts coding for functional proteins.

We compared the transcriptome and proteome on the basis of the ENSEMBL gene annotation. For 94% of genes for which a protein was identified by MS, a corresponding mRNA was detected (Figure 2C). Analysis of membrane proteins and

regulatory proteins is often challenging in proteomics but Gene Ontology (GO) analysis showed similar percentages of transcripts and proteins for these categories, demonstrating that there were no such biases in the proteomic data (Figure 2D). This is likely the result of essentially complete solubilization of the proteome in SDS in the FASP procedure (Wisniewski *et al*, 2009b) combined with the overall depth of analysis.

The MS signal of peptides identifying each protein can be used to estimate its absolute cellular abundance (de Godoy *et al*, 2008; Malmstrom *et al*, 2009; Silva *et al*, 2006) in a similar way that the FPKM is a proxy for the abundance of transcripts. To calculate the approximate abundance of each protein we used the iBAQ algorithm (Schwanhausser *et al*, 2011), which normalizes the summed peptide intensities by the number of theoretically observable peptides of the protein. These normalized protein intensities are translated to protein copy number estimates based on the overall protein amount in the analyzed sample. We obtained good agreement with independently determined absolute copy numbers of 37 HeLa proteins (Zeiler *et al*, 2011; Supplementary Table S5). FPKM-based transcript abundance values correlate well with iBAQ-based protein abundance values (Spearman's correlation 0.6; Figure 2E). The use of high-resolution MS and RNA-Seq may account for the fact that higher correlations between transcriptomes and proteomes are observed here than in previous studies (Cox and Mann, 2011; de Sousa Abreu *et al*, 2009; Maier *et al*, 2009), where technical imperfections in the quantification of both the proteome and the transcriptome are likely to have reduced their apparent correlations.

To assess the completeness of the detected proteome, we first inspected macromolecular complexes for which all core members are presumably functionally necessary. Most of such complexes, such as the proteasome, spliceosome, histone-modifying complexes and respiratory chain complexes were completely represented according to the Corum protein complex database (Supplementary Figure S4A). Mean proteome coverage of all Corum complexes was >95%, slightly less than the corresponding transcriptome coverage (96.5%). Sarcoglycan–sarcospan complex (normally expressed in the muscle), SNARE complexes (abundant in neuronal tissue), ITGA2b–ITGB3 complex (normally expressed in platelets) were among the complexes with lower coverage (20, 40 and 50%, respectively), likely due to cell type specificity. Even though only 5% of our HeLa cell population was in mitosis, we covered 61 of 63 proteins in a reference set of cell cycle-specific proteins (Jensen *et al*, 2006). Our data set also has a very high coverage of most metabolic pathways pertaining to basic cellular functions. Comprehensiveness of the proteome is difficult to determine by comparison with pathway databases because they contain cell type-specific proteins. Nevertheless, judged against the coverage of pathways achieved by deep-sequencing transcriptomics, the proteomics data were >90% complete (Supplementary Figure S4B). Together, the transcriptome and proteome data suggest that at least 10 000–12 000 genes are expressed in HeLa cells.

The iBAQ values determined above estimate the absolute amount of each protein, incorporating individual peptide signals in MS and normalized by the number of observable peptides of the protein. The 40 most abundant proteins comprised 25% of the proteome (Figure 3A; Supplementary Table S6) with filaminA, pyruvate kinase, enolase, vimentin and Hsp 60 contributing >1% each. The most abundant 600 proteins constitute 75% of HeLa cell proteome mass (sum of all iBAQ values). The individual contribution of each protein to the total mass in combination with the knowledge of number of cells in the initial sample was used to roughly estimate the absolute copy number of the proteins in HeLa cells. The ranked distribution of all individual proteins revealed that 90% of the quantified proteome is contained within a range of a factor of 60 above or below the median protein copy number of 18 000 molecules per cell (Figure 3B; Supplementary Table S7). The lower half of the proteome accounts for <2% of its total mass. The abundance distribution of the transcriptome is generally similar but its range is compressed compared with the proteome with 90% of the transcriptome contained in a 500-fold expression range and 2000 transcripts accounting for 75% of the total transcriptome mass.

The protein abundance values can also be used to estimate the proportional contribution of any individual protein, protein complex and protein class to the total proteome. For example, ribosomes, which are encoded by only 1% of human genes and for which we identified 195 different proteins contributed 6% to total protein mass in our data (Figure 3C). Similarly, the actin cytoskeleton, as classified by GO (Ashburner *et al*, 2000) annotation, contributes four-fold more to the proteome mass than expected from the number of genes and proteins and 'protein folding' is achieved by <2% of the identified proteome by numbers but requires 8% of proteome mass in line with the high abundance of heat-shock and similar proteins (Figure 3D). In contrast, integral membrane proteins account for 25% of the genome but contribute much less to the transcriptome and the proteome (7.6% of total protein mass). This presumably reflects the often cell type-specific functions of these proteins (Lundberg *et al*, 2010; Ramskold *et al*, 2009).

Structural proteins and proteins in basic cellular machineries are known to be much more abundant than regulatory proteins; however, the generality of this rule could not previously be evaluated. Ribosomal proteins indeed formed a tight cluster at the top end of the distribution of transcript and protein expression levels (Figure 3E). This was also true of the core components of the proteasome, but not its regulatory subunits, which were up to a factor of 100 less abundant. Interestingly, the abundance of cytoskeletal proteins extended over a broad range from the most abundant proteins and transcripts to the medium and low abundance parts of the distribution. Metabolic enzymes are likewise generally considered to be an abundant class of proteins, but we found that they extend over almost the entire distribution of the transcriptome and proteome expression (Figure 3F). Enolase was the protein with the highest expression value, while glycogen phosphorylase (muscle form) was expressed 100 000-fold less at the protein level and 10 000-fold less at the transcript level. Large differences in expression levels of different metabolic enzymes have also been observed in recent targeted proteomics experiments in yeast (Picotti *et al*, 2009). As expected, our data show that regulatory proteins such as

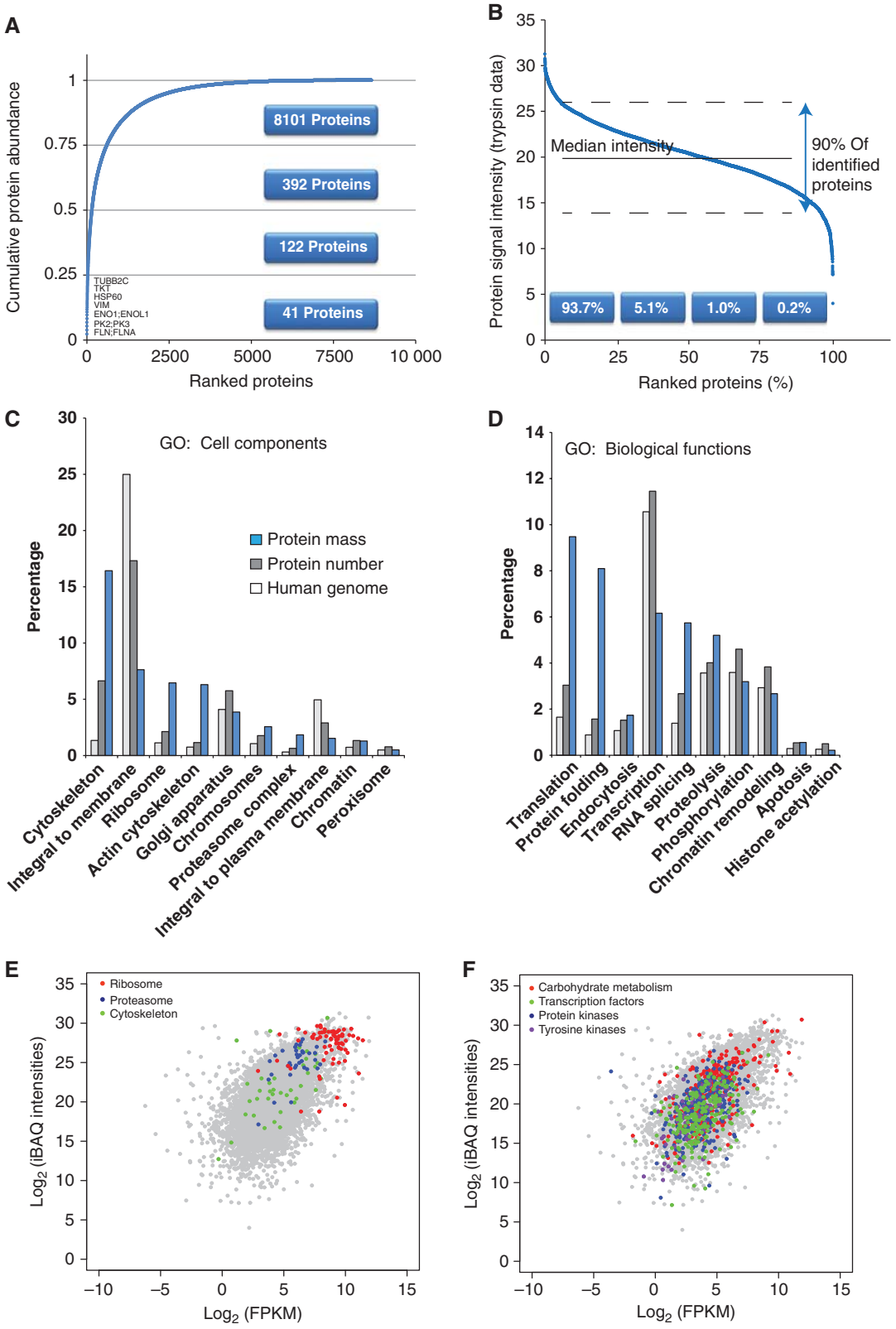


Figure 3 Quantitative analysis of expressed genes. (**A**) Cumulative protein mass from the highest to the lowest abundance proteins. (**B**) Ranked protein abundances from the highest to the lowest. (**C**) Gene ontology analysis of cellular compartments annotations including the percent of the annotated genes in the genome, the percent of the identified proteins and the percent of the protein mass that is attributed to these annotations. (**D**) Same as (C) but for Gene Ontology biological process annotations. (**E**) Scatter plot of iBAQ intensities versus FPKM values with highlighting of structural proteins and proteins in basic cellular machineries. (**F**) Same as (E) but highlighting of metabolic and regulatory proteins.

protein kinases and transcription factors have, on average, lower expression than the structural proteins discussed above. However, each of these categories spans a large expression range and surprisingly many of their members are in the top 25% of the proteome. Allowing these and similar comparisons of estimated expression levels of individual proteins and protein classes, as well as the corresponding transcripts, our data can provide starting points for systems biological modeling of the cell.

RNA-Seq already covers virtually the entire functional transcriptome. Ultra-deep mapping of the proteome is now also becoming possible with proteins identifiable for nearly all transcripts with an expected biological function in the cell type. Thus, both transcriptomics and proteomics are approaching completeness. Given the rapid technological progress in both fields, we predict that the required depth of 10 000–12 000 genes will be routinely reachable soon.

Materials and methods

HeLa cells lysate

Cell pellets were flash frozen in liquid nitrogen and stored at −80°C. Cells were lysed in a buffer consisting of 0.1 M Tris–HCl, pH 8.0, 0.1 M DTT, and 2% SDS at 99°C for 5 min. After chilling to room temperature, the lysates were sonicated using a Branson type sonicator and then were clarified by centrifugation at 16 100 *g* for 10 min. Protein content was determined using a Fluorescence Spectrometer.

Protein fractionation by gel filtration

In all, 0.100 ml of the cell lysate containing 10 mg of total protein was loaded onto a Superdex 200 10/300 GL column (GE Healthcare Bio-Sciences AB, Uppsala) equilibrated with TNS buffer composed of 0.1 M Tris–HCl, pH 8 buffer, 0.1 M NaCl and 0.2% SDS. Proteins were eluted with TNS buffer and 2 ml fractions were collected.

Protein digestion and peptide fractionation

Detergent was removed from the lysates and the proteins were digested with trypsin, LysC, or Gluc using the FASP protocol (Wisniewski *et al*, 2009b) using ultrafiltration units of nominal molecular weight cutoff of 30 000 (Cat No. MRCF0R030, Millipore). The eluted peptides were fractionated according to the previously described pipette tip protocol (Wisniewski *et al*, 2009a).

Mass spectrometry

The peptides were purified on StageTips (Rappsilber *et al*, 2007). Eluted peptides were separated on a reverse phase C_{18} column (40 cm long, 75 μm i.d., 1.8 μm beads, Dr Maisch GmbH, Germany) using the EASY-nLC system (Proxeon Biosystems now Thermo Fisher Scientific). MS analysis was performed using LTQ-Orbitrap Velos instrument (Thermo Fisher Scientific; Olsen *et al*, 2009). Data were acquired in data-dependent mode. The survey scans were acquired at a resolution of 30 000 at m/z=400 in the Orbitrap analyzer followed by up to 10 fragmentation events (HCD) in the collision cell. The fragment ions were also detected in Orbitrap analyzer resulting in high-resolution and high-accuracy fragmentation spectra.

RNA-seq

Total RNA was extracted from HeLa cell pellets using the RNAeasy Mini Spin columns protocol from Qiagen and an elution volume of 50 μl. RNA quality (RIN 10) and quantity (~1 μg/μl) were assessed using an Agilent RNA 6000 LabChip. The RNA extracts were stored at −80°C. The Illumina RNA-seq sample preparation protocol and kit (RS-100-0801) as well as the Illumina Paired End library preparation protocol and kit (PE-102-1001) were used for library preparation. Briefly, total RNA was enriched for poly-A tailed transcripts using magnetic beads with poly-T oligonucleotide coating. The enriched RNA was fragmented into small pieces using divalent cations and elevated temperature (94°C, 5 min). RNA fragments were copied into cDNA using a reverse transcriptase and random priming (Invitrogen SuperScript II). Second-strand synthesis was performed in the same reaction using RNaseH and DNA polymerase I. Overhangs were converted into blunt ends using T4 DNA polymerase (5′ overhang fill-in) and Klenow DNA polymerase (3′–5′ exonuclease activity). A deoxyadenosine was added to the 3′ end of the blunt and phosphorylated DNA fragments using the polymerase activity of Klenow fragment. T4 DNA ligase was used to ligate forked adapters and a gel length selection performed (~200 nt insert size). Molecules were then amplified with over-hanging primers that extend the adapters to their final length required for the sequencing.

The library was sequenced on two Illumina Genome Analyzer IIx lanes following vendor instructions for Multiplex Single Read sequencing and using 76 + 7 cycles. Protocols were followed except that an indexed φX174 control library was spiked into each lane, yielding about 1% of sequencing reads per lane. The φX174 control reads were aligned to the corresponding reference sequence to obtain a training data set for the base caller Ibis (Kircher *et al*, 2009), which was then used to generate base calls and quality scores.

Data availability

All RNA-seq sequence data is available from the European Nucleotide Archive (ENA) under the study accession ERP000959, and from ArrayExpress under accession number E-MTAB-823. All mass spectrometric raw files are uploaded to TRANCHE and can be accessed using the following hash codes: Hela_01_trypsin;phajxUWNFSW8gBCd3oQJ;Hash:dLuhvyddHELlkrXVJa1QYTHGOdFDttpFksh8iBqBT4kNyESmVFznzAtXe4qS + 9OCtJ//9y7DfdlcEIotcGCerr/ytCUAAAAAAAAAWwQ==;Hela_01_LysC;GRtGG4GkZoo6pYZEbyd0 Hash:r6G4xDnc8deuSSpRMDkYk7hJsjvuWrMFoJGenuTEdtYN3zMhGDXaOl/QheYipLUoe/37f1lrYS + GQhRgDH + K5gfKns4AAAAAAAAWNg==;Hela_01_GluC;34NGEzbCmXHXr09aPqOV;Hash:GGDWG1xveOYXVD5DkiSVybfbp41fzZzeNiDJdVCcOmmXaFjLTNdOzOIPO0aCXkvnInsZ2kO4hvq3WZ9IW + O8yenB + NQAAAAAAAAY7Q==Hela_02_trypsin;gfAYWK0ljixAdVddEQH5;Hash:6YBO0zZhlHORAXzJ; + UqC4i6tlnlLw5OAV5lOzkoW1dYVueWQD9M6k + 4YvQ/43iE7kalH + 3LPJT5wqq27TlG/zdXNJeAAAAAAAAAsfg==;Hela_02_LysC;hUU1ZRgB61kmdtEJHmX4;Hash:Bz9hlKJ5EaEq/rgoVH0 + fHehRgTSaCcD2;879Q1JnJm3d9sFaCpNgFnPPZT9WFu5K5mXKz8o1B9qaK7WBFxdFPu2ThkAAAAAAAAPmA==Hela_02_GluC;qEFG57NWsYggbpjHmQ5H;Hash:LEqiT5pWYpusY/SWaXJw8A3GcRAspRucqyb6L/nKSG9AywRpBL8hkBn8r + sZP3fXTWC2PoLNmhOpqkbg6lQR63GHeyAAAAAAAAAftQ==.

Gene and transcript quantification

Raw reads of two sequencing lanes were combined, adapters trimmed and reads shorter than 70 nt, or with more than five bases below a quality score of 15 (PHRED-scale), removed. The processed reads were aligned to the human reference genome (hg19/GRCh37 excluding additional haplotypes) using TopHat v1.0.13 (Trapnell *et al*, 2009) and transcripts and genes of the Ensembl (Hubbard *et al*, 2009) release 59 were quantified using Cufflinks v0.8.3 (Trapnell *et al*, 2010). This method allows up to 40 equally good mappings of a read. In cases where a read can be mapped to multiple transcripts, each transcript is assigned one per number of mappings in the quantification step. If >40 potential mapping locations are identified, then the read is not considered for quantification.

Data analysis

Raw files from MS analysis were processed using the MaxQuant computational proteomics platform (Cox and Mann, 2008) version 1.1.1.36. The peak list generated was searched against the IPI human

database (ipi.HUMAN.v3.68.fasta) with initial precursor and fragment mass tolerance set to 7 and 20 p.p.m., respectively. Peptides with minimum of six amino-acid length were considered with both the peptide and protein FDR set to 1%.

All MS data were mapped to gene identifiers obtained from Ensembl for comparison with the RNA-seq data. For the quantitative analysis, the iBAQ intensity and the FPKM values were used for proteome and transcriptome data, respectively.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Peter Bandilla for MS assistance, Birgit Nickel and Ayinuer Aximu for preparing the RNA-seq library and Marlis Zeiler for helpful discussions. This work was funded by the Max Planck Society, by the European Commission's 7th Framework Program PROteomics SPECification in Time and Space (PROSPECTS, HEALTH-F4-2008-201648), by the Munich Center for Integrated Protein Science (CIPSM) and by the Knut and Alice Wallenberg Foundation.

Author contributions: NN performed the MS analysis, analyzed the data and wrote the manuscript; JRW designed the experiment and prepared the samples; TG analyzed the data and wrote the paper; JC analyzed the data; MK, JK and SP performed the RNA-Seq and analyzed the data; MM initiated and supervised the work and wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

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

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci USA* **104:** 19428–19433

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26:** 1367–1372

Cox J, Mann M (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem* **80:** 273–299

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

de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455:** 1251–1254

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5:** 1512–1526

Graumann J, Hubner NC, Kim JB, Ko K, Moser M, Kumar C, Cox J, Scholer H, Mann M (2008) Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. *Mol Cell Proteomics* **7:** 672–683

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510

Haas BJ, Zody MC (2010) Advancing RNA-Seq analysis. *Nat Biotechnol* **28:** 421–423

Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7:** 497

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E *et al* (2002) The Ensembl genome database project. *Nucleic Acids Res* **30:** 38–41

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K *et al* (2009) Ensembl 2009. *Nucleic Acids Res* **37:** D690–D697

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443:** 594–597

Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10:** R83

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

Maier T, Guell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett* **583:** 3966–3973

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

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

Mann M, Kelleher NL (2008) Precision proteomics: the case for high resolution and high mass accuracy. *Proc Natl Acad Sci USA* **105:** 18132–18138

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628

Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **4:** 709–712

Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S (2009) A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* **8:** 2759–2769

Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138:** 795–806

Ramskold D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5:** e1000598

Rappsilber J, Mann M, Ishihama Y (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2:** 1896–1906

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

Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **5:** 144–156

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63

Wisniewski JR, Zougman A, Mann M (2009a) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J Proteome Res* **8:** 5674–5678

Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009b) Universal sample preparation method for proteome analysis. *Nat Methods* **6:** 359–362

Zeiler M, Straube WL, Lundberg E, Uhlen M, Mann M (2011) A protein epitope signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol Cell Proteomics* (e-pub ahead of print 30 September 2011; doi:10.1074/mcp.O111.009613)

