nature biotechnology

ANALYSIS

https://doi.org/10.1038/s41587-019-0037-y

Multi-omic measurements of heterogeneity in HeLa cells across laboratories

Yansheng Liu[1,2*], Yang Mi[3,4,18], Torsten Mueller[5,18], Saskia Kreibich[6,18], Evan G. Williams[5], Audrey Van Drogen[5], Christelle Borel[7], Max Frank[5], Pierre-Luc Germain[8], Isabell Bludau[5], Martin Mehnert[5], Michael Seifert[9,10], Mario Emmenlauer[11], Isabel Sorg[11], Fedor Bezrukov[7], Frederique Sloan Bena[12], Hu Zhou[13], Christoph Dehio[11], Giuseppe Testa[8,14], Julio Saez-Rodriguez[4,15], Stylianos E. Antonarakis[7,12,16], Wolf-Dietrich Hardt[6] and Ruedi Aebersold[5,17*]

Reproducibility in research can be compromised by both biological and technical variation, but most of the focus is on removing the latter. Here we investigate the effects of biological variation in HeLa cell lines using a systems-wide approach. We determine the degree of molecular and phenotypic variability across 14 stock HeLa samples from 13 international laboratories. We cultured cells in uniform conditions and profiled genome-wide copy numbers, mRNAs, proteins and protein turnover rates in each cell line. We discovered substantial heterogeneity between HeLa variants, especially between lines of the CCL2 and Kyoto varieties, and observed progressive divergence within a specific cell line over 50 successive passages. Genomic variability has a complex, nonlinear effect on transcriptome, proteome and protein turnover profiles, and proteotype patterns explain the varying phenotypic response of different cell lines to *Salmonella* infection. These findings have implications for the interpretation and reproducibility of research results obtained from human cultured cells.

Many technical factors can contribute to poor reproducibility, but the complexity inherent in biological systems also poses a major challenge. The effect of genomic and environmental perturbations on the molecular makeup and phenotypic response of a cell or organism, and how genetically different cells or organisms react to identical perturbations, remains largely unknown.

Several recent studies have highlighted problems in human cancer cell lines, such as cell line misidentification, cross-contamination and poor annotation, that could impair the reproducibility of results obtained from these cell lines between laboratories[1–4]. Consequently, short tandem repeat and single nucleotide polymorphism profiles have been proposed to authenticate cells[4,5]. However, the extent to which genotypic variability induces proteotype and phenotype variations in the same cell line cultured in different laboratories is unknown. Unsystematic observations suggest that many cultured cell lines might be genomically unstable[6,7], and a recent report shows that cancer cell lines may undergo rapid genetic diversification as a result of positive clonal selection that is highly sensitive to culture conditions[8]. In such cases, even careful experimentation cannot assure reproducibility of research results.

HeLa cells present an important example of human cancer cells that have widely influenced biological studies. More than 100,000 publications have used or directly referenced HeLa cells. However, owing to extensive genome instability during passaging and transfer between laboratories, HeLa cells have been reported to contain a very large number of genomic variants[7,9–12]. The currently widely used HeLa variants include HeLa CCL2, the 'original' HeLa cell line; HeLa S3 (also called CCL2.2), the third clone isolated from an early HeLa culture; and HeLa Kyoto. Whole-genome sequencing of HeLa Kyoto[13] and HeLa CCL2[14] has been performed, and notable variation in sequence, copy number and chromosomes were reported between the two[14]. Furthermore, different HeLa Kyoto clones were reported to vary in terms of mRNA expression between laboratories[7]. However, little is known about how such genomic variability affects the proteomes or cellular phenotypes[15] of the different HeLa strains between laboratories, the effect of successive passages within a stock-derived line on its molecular makeup, and how these would affect a biological research outcome.

Here we perform a system-wide analysis of HeLa cell line variants collected from 13 laboratories to investigate biological variation across these lines. We use SWATH (sequential-window acquisition of all theoretical fragments) mass spectrometry (SWATH-MS)[16–20]

[1]Department of Pharmacology, Yale University School of Medicine, New Haven, CT, USA. [2]Yale Cancer Biology Institute, Yale University, West Haven, CT, USA. [3]Heidelberg University, Faculty of Biosciences, Heidelberg, Germany. [4]Joint Research Center for Computational Biomedicine (JRC-COMBINE), Faculty of Medicine, RWTH Aachen University, Aachen, Germany. [5]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [6]Institute of Microbiology, ETH Zurich, Zurich, Switzerland. [7]Department of Genetic Medicine and Development, University of Geneva Medical School, and University Hospitals of Geneva, Geneva, Switzerland. [8]IEO, European Institute of Oncology IRCCS, Milan, Italy. [9]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany. [10]National Center for Tumor Diseases, Dresden, Germany. [11]Biozentrum, University of Basel, Basel, Switzerland. [12]Service of Genetic Medicine, University Hospitals of Geneva, Geneva, Switzerland. [13]Department of Analytical Chemistry and CAS Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. [14]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [15]Institute for Computational Biomedicine, Heidelberg University, Faculty of Medicine, Bioquant Heidelberg, Germany. [16]iGE3 Institute of Genetics and Genomics of Geneva, Geneva, Switzerland. [17]Faculty of Science, University of Zurich, Zurich, Switzerland. [18]These authors contributed equally: Yang Mi, Torsten Mueller, Saskia Kreibich. *e-mail: yansheng.liu@yale.edu; aebersold@imsb.biol.ethz.ch

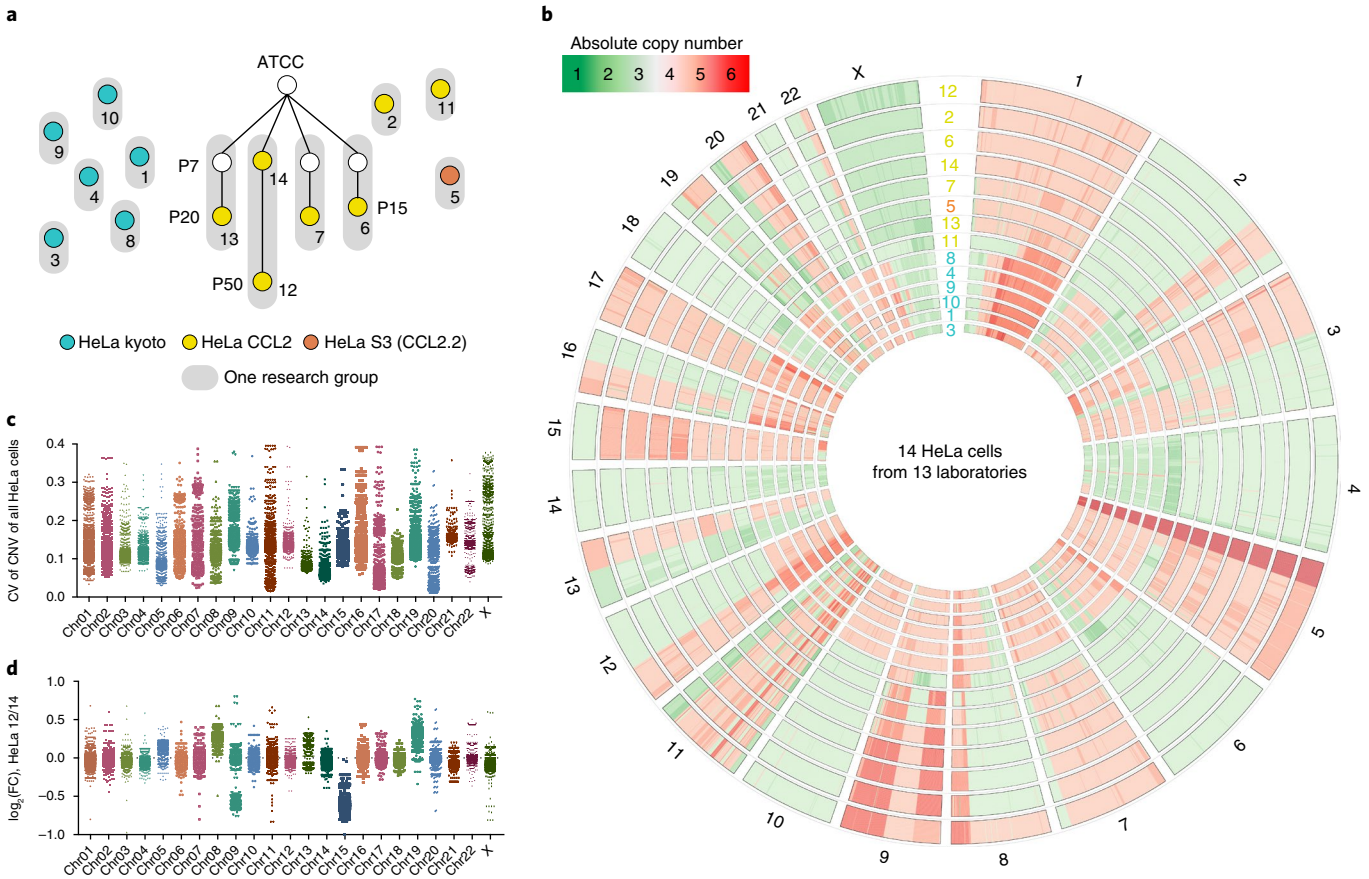


Fig. 1 | HeLa cell lines from different laboratories showed varied and evolving genotypes. a, HeLa cell variants were collected from 13 laboratories and arbitrarily numbered from HeLa 1 to HeLa 14. ATCC, American Type Culture Collection; P, passages. **b**, Circus plot of raw absolute copy numbers across all 14 HeLa cell lines. **c**, Coefficient of variation (CV) distribution of all genes encoded by different chromosomes among all cell lines. **d**, Log of fold change (FC) values for gene copy number changes between HeLa 12 and HeLa 14 at each chromosome.

to quantify steady-state protein profiles and relate these to copy number profiles, transcript profiles, protein turnover rates[21–23] and phenotypic characteristics such as cell doubling time and a variable response to *Salmonella* infection. The results highlight the biological complexities of commonly used human cancer cell lines and provide an important basis for discussing how experimental data obtained from these lines should be reported and interpreted.

Results

HeLa cell variants from different laboratories have different and rapidly evolving genotypes. We surveyed HeLa cells used in 13 laboratories (Fig. 1). Four laboratories shared the seventh passage (P) 7 of an initial HeLa CCL2 (the original version of HeLa, purchased from the American Type Culture Collection (ATCC)). Three of these laboratories separately cultured cells until P15 or P20, a stage at which cells are generally regarded as still acceptable for biology research (HeLa 6, 7, and 13) (Fig. 1a and Supplementary Table 1). The fourth group provided both P7 (HeLa 14) and P50 cells (HeLa 12), which allowed us to identify molecular alterations occurring after 3 months of continuous culture. In total the 14 HeLa cell variants comprised 7 HeLa CCL2 lines, 1 HeLa S3 line and 6 HeLa Kyoto lines. To reduce experimental bias, we centrally cultured all the cell lines for an additional three passages under the same conditions (Supplementary Note 1).

First, we evaluated HeLa cell heterogeneity by measuring gene copy number variation (CNV) by array comparative genomic hybridization (aCGH) (Fig. 1b). We discovered pervasive CNV differences organized by domains, large chromosomal segments, and even whole chromosomes. Particularly notable were ploidy changes in chromosomes (Chr) 1, 2, 6, 9, 10, 17, 19, 21, 22 and X. (Fig. 1b). On average, the genomes of all cells tested had an overall hypertriploid state, as reported[13,14]. However, HeLa CCL2 lines had 1.87 times as many genes with two copies and 0.7 times as many genes with three copies compared to the Kyoto lines (Supplementary Fig. 1). Even within CCL2 and Kyoto groups, significant CNVs could be observed, albeit at a smaller scale—for example, at the distal region on Chr8. Moreover, HeLa 11 deviated in many chromosomes from other HeLa CCL2 cells. Overall, we observed widespread DNA dosage variation among HeLa cells, although Chr13, 14, 18 and 20 were more stable than the other chromosomes (Fig. 1c). Notably, in comparison to P7, the P50 HeLa cells gained or depleted entire chromosome copies or large chromosome domains (Fig. 1d). Examples are copy gains of a whole Chr8 and loss of Chr15, as well as a two-thirds gain of Chr 19 and a partial loss of Chr9. The comparison to P20 cells further reveals progressive accumulation of CNV differences with passaging (Supplementary Fig. 2).

In summary, we observed, as a likely consequence of genomic instability or clonal selection, a considerable degree of large-scale CNV across HeLa cells used in different laboratories, even among strains with the same annotation.

Diversity of gene expression patterns at steady state. We then analyzed how the diverse CNV patterns influence steady-state gene expression between HeLa strains (Fig. 2). Collectively, we quantified transcripts for 11,365 genes with an average number of reads per kilobase of transcript per million mapped reads (RPKM) >1

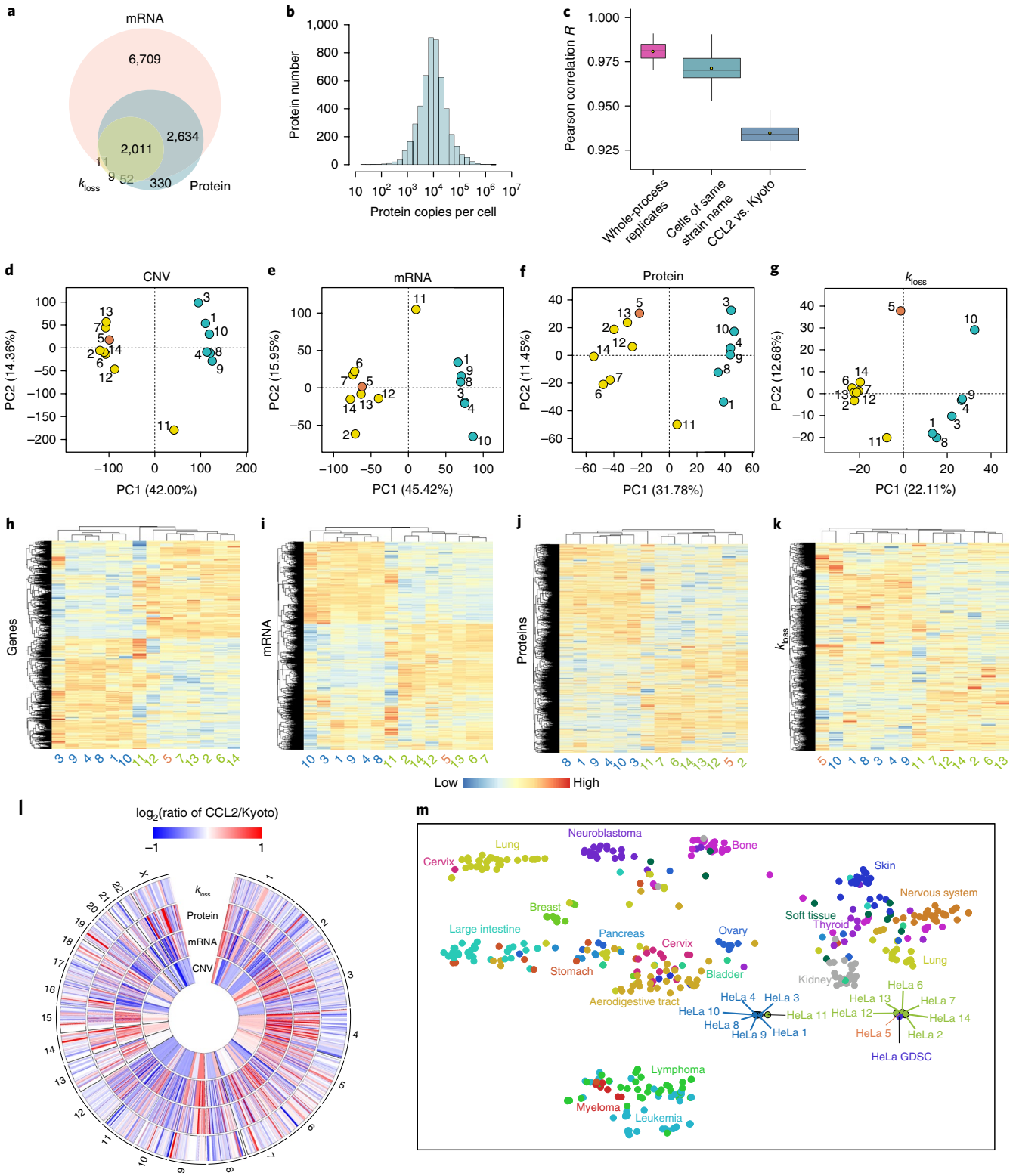


Fig. 2 | Heterogeneous transcriptome, proteome and protein turnover profiles between HeLa cell lines across laboratories. a, Expression of genes measured at each layer. **b**, Label-free absolute quantification of averaged protein copies profiled per HeLa cell ($n = 5{,}030$ proteins). **c**, Quantitative reproducibility of SWATH-MS for whole-process replicates and HeLa cells from different laboratories ($n = 11$, 60 and 60 observations for the red, green and blue boxes, respectively). The borders of the box represent the 25th and 75th quartiles, the bar within the box represents the median, the dot denotes the mean, and whiskers represent the range (Methods). **d**–**g**, PCA of CNV, transcripts ($N = 11{,}365$), proteins ($N = 5{,}030$), and k_{loss} ($N = 2{,}084$) of all HeLa cell lines. k_{loss} is protein loss rate, a proxy for protein turnover rate. **h**–**k**, HCA of the data. **l**, HeLa CCL2/Kyoto ratios. Fold changes greater than 2 are shown in dark blue or red. The data are proteome centric; i.e., data matched to available proteomic identifications. **m**, t-SNE analysis of our HeLa transcriptomic data combined with 1,001 molecularly annotated human cancer cell lines from Iorio et al.[26]. The published HeLa cell in the GDSC panel (i.e., HeLa_GDSC, strain identity not mentioned) clusters with our CCL2 cells, indicating that it is likely to be a HeLa CCL2 cell line.

(Fig. 2a). Using SWATH-MS[17–20], we consistently quantified 5,030 proteins across all samples (1% peptide and protein false discovery rate (FDR) controlled by PyProphet[17]; Methods and Supplementary Fig. 3). The absolute label-free quantification determined the number of protein copies to center at 10,000 and to span from 100 to >10^6 copies per cell (Fig. 2b). Using pulsed stable isotope labeling with amino acids in culture (pSILAC) in combination with SWATH-MS[23], we further quantified the proxy turnover rate (k_{loss}; Methods) of a consistent set of 2,084 proteins. Taken together, our results present a well-matched dataset for studying gene expression in HeLa cells (Fig. 2a and Supplementary Fig. 4). As a technical assessment, the sample-to-sample Pearson correlation suggested that SWATH-MS achieved a high and consistent reproducibility that is sufficient to distinguish whole process replicates and HeLa cells from different laboratories, with minimal quantification bias against low-abundance proteins (Fig. 2c, Supplementary Fig. 5 and Supplementary Note 2).

For a global comparison, we performed principal component analysis (PCA) and unsupervised hierarchical clustering analysis (HCA) at each omics layer (Fig. 2d–k). Both analyses suggested, first, that HeLa CCL2 and Kyoto variants differ substantially at every level of gene expression; second, that HeLa S3 (HeLa 5) cells are systematically closer to HeLa CCL2 than to Kyoto variants; and third, that HeLa 11 resides between CCL2 and Kyoto groups in PCA, although it is closer to the CCL2 cluster in HCA at all levels. We thus confirmed all the HeLa cells, including HeLa 11, as bona fide HeLa cell lineages by the deep mapping of single nucleotide variants (SNV) extracted from our RNA-seq dataset relative to the COSMIC HeLa reference[24,25] (Supplementary Fig. 6).

To understand the CCL2–Kyoto difference, we first noted that the transcriptomic and proteomic CCL2/Kyoto ratios at the respective gene loci largely followed CNV imbalance (Fig. 2l and Supplementary Fig. 7), whereas the k_{loss} ratios showed a higher degree of variation. We then benchmarked the CCL2–Kyoto mRNA expression difference to the transcriptomic variation of human cancer cell lines of the GDSC panel[26] by both *t*-distributed stochastic neighbor embedding (t-SNE; Fig. 2m) and PCA (Supplementary Fig. 8). This analysis demonstrates that HeLa CCL2 and Kyoto groups are as distinct from each other as are cancer cell lines from different tissue types (Supplementary Fig. 8). Similar observations were made using published protein abundance and turnover datasets in skin fibroblast cells discordant for trisomy 21 (Supplementary Fig. 9 and Supplementary Note 3).

Gene expression patterns evolve with cell passages. HeLa cells are immortalized cell lines. However, it is not clear how much functional variation will be introduced during passaging of cells, considering its unstable genome[7,9–12]. We compared the mRNA-seq data acquired from three biological replicates each of HeLa 12 and 14. Strikingly, 731 transcripts (~6.4% of confidently profiled transcripts) showed significantly altered expression between HeLa 14 and 12 (adjusted $P < 0.01$ by edgeR[27], fold change > 2) (Fig. 3a). Among these 731, we measured the protein levels of 166. They showed a notably positive mRNA–protein correlation (Spearman's $\rho = 0.47$, Fig. 3b). Thus, the differential transcript profiles between P7 and P50 also significantly affect the proteome. For example, the integrin-mediated signal pathway, negative regulation of endopeptidase activity and inflammatory response are consistently regulated (adjusted $P < 0.05$ for all processes) at both mRNA and protein levels (Supplementary Fig. 10). These results argue that investigations focusing on these processes can lead to inconsistent results if HeLa cells of quite different generations, such as HeLa 14 or 12, are used.

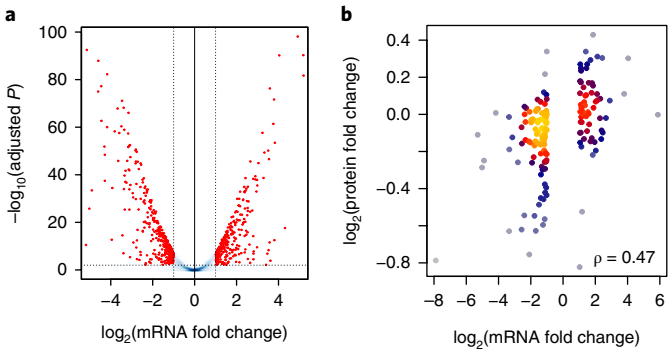
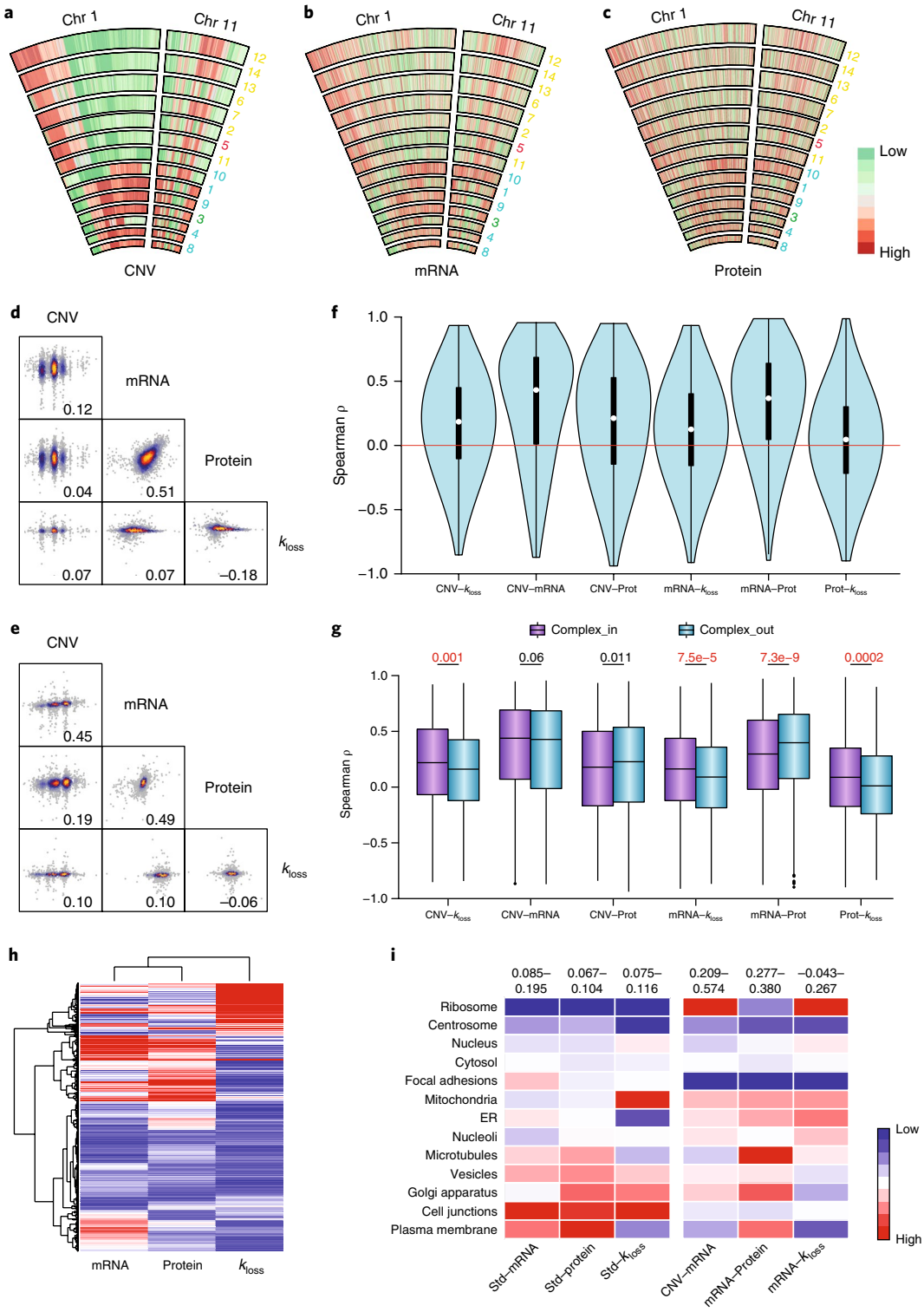


Fig. 3 | Gene expression comparison between HeLa 12 and 14 representing 3-month passaging. a, Volcano plot of 3 vs. 3 biological replicates by transcriptomics data. 415 transcripts showed increased expression levels, whereas 316 transcripts showed decreased expression levels in HeLa 12 (P50) compared to HeLa 14 (P7). *P* values were calculated by edgeR tests. The red dots denote the significantly altered gene expressions (BH adjusted $P < 0.01$; fold-change > 2). **b**, Spearman correlation analysis between quantitative data of mRNA and protein fold changes ($n = 166$ mRNA–protein pairs), showing Spearman's ρ. The blue-to-yellow color scale denotes increasing data density by Kernel Density Estimation (Methods).

Global processes shape proteotypes between HeLa cell variants. The multilayered HeLa dataset allowed us to analyze global proteome expression control mechanisms and how the cells reconcile the large-scale CNVs with their essential metabolic needs. Indeed, numerous CNVs leading to significant mRNA changes did not translate into corresponding protein levels (for example, those of Chr1 and Chr11 in Fig. 4a–c). Moreover, the total variance from mRNA to protein and to proteostasis levels followed a reduction trend (Supplementary Fig. 11).

Fig. 4 | Global processes affecting HeLa proteotypes. a–c, Quantitative differences in CNV and mRNA and protein levels between HeLa cells after normalizing the data to the mean of the respective values. Chromosome 1 and 11 are shown as representative examples. **d**, Across-gene Spearman's correlation between CNV, mRNA, protein and k_{loss} values using absolute-scale data for HeLa 1 (a HeLa Kyoto line). **e**, Within-gene Spearman's correlation between layers using the relative quantification data between HeLa 1 and HeLa 14 (the ATCC CCL2). **f**, Gene-specific, cross-cell-line Spearman's ρ between layers. The outer violin curve denotes the kernel density of the dataset, the thick black bar in the center represents interquartile range, the thin black line represents 95% confidence intervals, and the white dot denotes the median. **g**, Cross-layer correlations between those proteins that are annotated in any stable complex in the CORUM database ("complex_in" group) and those that are not ("complex_out" group); Prot, protein. Box borders represent the 25th and 75th quartiles, bar within the box represents the median, and whiskers represent the range (Methods). Two-sided Wilcoxon test, $P = 1.102 \times 10^{-3}$, 6.373×10^{-2}, 1.096×10^{-2}, 7.483×10^{-5}, 7.285×10^{-9}, 1.770×10^{-4} (approximate values shown at top). **h**, Gene expression variability calculated by s.d. from the average at each level ($n = 2{,}011$ genes) suggests regulation preferences. **i**, Organelle perspective of gene expression variability and Spearman's correlation. Std, standard deviation of the relative quantification. For **d**, **e**, **f** and **i**, all proteome-centric data points were included (i.e., we visualized $n = 2{,}002$ CNV-k_{loss} pairs, $n = 4{,}530$ CNV–mRNA pairs, $n = 4{,}853$ CNV–protein pairs, $n = 1{,}964$ mRNA-k_{loss} pairs, $n = 4{,}524$ mRNA–protein pairs, and $n = 2{,}002$ protein-k_{loss} pairs). For **g**, these numbers are further divided into a complex_in group ($n = 818$, 1,450, 1,487, 806, 1,449 and 818; purple boxes) and a complex_out group ($n = 1{,}184$, 3,080, 3,366, 1,158, 3,075 and 1,184; blue boxes).

To understand post-transcriptional regulation in this context, we correlated the absolute-scale data from HeLa 1 (a HeLa Kyoto cell) as an example between omics layers in an across-gene manner (Fig. 4d). We also correlated the relative fold changes between HeLa 1 and HeLa 14 (a CCL2 line) in a within-gene manner[28,29] (Fig. 4e). The mRNA–protein correlation was quantitatively strong for both absolute and relative scales (Spearman's $\rho = 0.51$ and 0.49), reinforcing the previous notion that protein abundances at steady state are primarily determined by mRNA levels[28,30]. Notably, CNVs strongly determined the mRNA levels when we considered the relative difference between two cell lines ($\rho = 0.45$), but only weakly determined mRNA absolute copies within one cell line ($\rho = 0.12$). Similar trends were found for protein levels ($\rho = 0.19$ vs. 0.04). This illustrates the importance of considering both across-gene and within-gene analyses. The absolute protein–k_{loss} correlation ρ was −0.18 (Fig. 4d), confirming that highly abundant proteins are less strongly regulated by protein degradation rates than proteins expressed at lower levels[31]. Summarizing the gene-specific Spearman ρ across 14



HeLa variants, the CNV–k_{loss} and mRNA–k_{loss} correlations were 0.16 and 0.11, respectively, supporting the notion that, for many genes, protein turnover functions as a buffering step, shaping the quantitative proteome between HeLa cells (Fig. 4f). Using protein complex annotation in the CORUM database[32], we verified the maintenance of protein complex stoichiometries as an efficient buffering mechanism[23,28,33,34] (Fig. 4g and Supplementary Note 4).

We next analyzed the biological relevance of variations in mRNA, protein and protein turnover (Fig. 4h). The majority (~75%) of genes showed a consistent extent of variation between mRNA and proteins, whereas k_{loss} particularly strongly affected a subset of genes. Furthermore, we used a recently established subcellular atlas to distribute cross-layer correlations according to the protein organelle locations[35] (Fig. 4i). Notably, we found that the mRNA–protein correlation for ribosomal proteins was low whereas mRNA–k_{loss} correlation was highest, demonstrating that ribosomes are tightly controlled at the protein level. Indeed, compared to CNV and mRNA, about 40% of the cytosolic ribosome proteome showed a reversed expression pattern between HeLa CCL2 and Kyoto cells (Supplementary Fig. 12). Moreover, both k_{loss} variation and mRNA–k_{loss} correlation of mitochondrial proteins were among the highest of the observed values, reinforcing our previous observation that the turnover of the mitochondrial proteome can be important in buffering aneuploidy stress[23]. Conversely, there was a very weak mRNA–k_{loss} correlation for plasma membrane proteins, suggesting a weaker role of protein turnover in shaping membrane proteome variability between HeLa cells. The above results suggest that the proteotypic variability is controlled by a multitude of global processes, including the control of protein complex stoichiometry and organelle-specific proteostasis.

HeLa proteotypic variability tightly links to phenotypic variability. To better understand the consequences of molecular heterogeneity between HeLa strains, we analyzed their phenotypes. Direct molecular imaging analysis showed that the cells were morphologically different in many aspects, such as the texture contrast of their actin structures (Supplementary Fig. 13). Moreover, cell doubling time was strikingly different between HeLa strains, with extremes being 17.5 and 32.3 h under identical culture conditions (Fig. 5a). HeLa Kyoto cells grew faster than CCL2 cells (averaged doubling time, 21.1 vs. 28.2 h, respectively; $P = 0.018$, Welch's t-test), an observation that has not been documented previously. Intriguingly, 63 proteins annotated to cell cycle could be used to distinguish HeLa Kyoto and CCL2 lines, respectively (Fig. 5b). For example, the absolute protein copy number of cyclin-dependent kinases 1, 2 and 7 (CDK1, CDK2 and CDK7) were on average 57.7, 46.8 and 81.2% higher in the Kyoto than in the CCL2 group (Supplementary Fig. 14). Such an observation might help to establish a possible 'proteotype–phenotype' link explaining cell doubling time differences between HeLa cell variants.

Besides these readily apparent phenotypes, we sought to inspect the response consistency of the HeLa cells to the same stimulus or perturbation[36]. We selected mimics of *Let7*, which is a highly conserved microRNA (miRNA) that has been shown to play a central role in development and tumor suppression, to transfect all the HeLa cell lines[37,38] (Supplementary Fig. 15). After a 72-h incubation, we quantified the abundance ratio of 5,030 proteins between *Let7*-treated and control cells for each cell line. Notably, the proteome-wide abundance ratios again separated HeLa CCL2 and Kyoto strain groups. The only exception was HeLa 11, which clustered between groups (Fig. 5c). SWATH-MS detected 107 of the top 500 most likely *Let7* gene targets according to TargetScan[39] (Supplementary Fig. 16). In transfected cells, ~70% of these targets showed a decrease in protein abundance, indicating that the *Let7* treatment was effective. In particular, ten targets showed remarkably strain-dependent regulation upon *Let7* transfection in CCL2 and Kyoto. Thus, the HeLa cell lines tested showed a varied response to a specific perturbation such as *Let7* treatment.

Expression of *Let7* microRNAs is downregulated upon *Salmonella* Typhimurium (*S*.Tm) infection in HeLa cells, which indicates a role of *Let7* in mediating the eukaryotic host response[38]. Using the ATCC HeLa CCL2 (i.e., HeLa 14), we found that *Let7* mimics interfered with *S*.Tm infection by impairing the docking step of *S*.Tm[40,41] (Supplementary Fig. 17). We repeated the infection experiment for all except one of the HeLa cells transfected with *Let7* (Fig. 5d,e; only HeLa 6 was not included; Methods). We found that the same concentration of *Let7* produced similar protective effect in most HeLa Kyoto cells, whereas the CCL2 cell lines showed a higher degree of phenotypic variability. HeLa 11 again showed an exceptional pattern in which the infection rate was not modulated by *Let7* (Fig. 5d). To explore this, we compared wild-type *Salmonella* and a noninvasive mutant (*S*.Tm$^{\Delta T1}$) for their abilities to infect HeLa 11, HeLa 3 and HeLa 14 cells over a range of multiplicities of infection (MOI) from 4 to 250. Notably, HeLa 11 showed lower overall infection rates at a given MOIs (Fig. 5f). Consistent with these results, we found that, compared to other CCL2 cells, HeLa 11 had the lowest expression of the Arp2/3 complex, which has a well-established role in bacterial internalization by host cells, acting by initiating membrane ruffles[40–42]. Remarkably, all seven of the protein subunits of the Arp2/3 complex followed the same pattern (Fig. 5g), which could explain the reduced *S*.Tm infection in HeLa 11. We used the STRING database to perform gene set enrichment for the 178 proteins that were differentially expressed between HeLa 11 and other CCL2 cells (linear model test, adjusted $P < 0.05$), and again the membrane-bound vesicle (GO:0031988)[43] was most significantly enriched (adjusted $P = 3.61 \times 10^{-12}$, Fig. 5h). Finally, we examined docking[41] by counting the *S*.Tm bacteria that remained bound to the respective HeLa cells after extensive washing. We confirmed a less pronounced *S*.Tm docking phenotype in HeLa 11 compared to HeLa 14, which may further indicate that HeLa 11 has a different membrane topology or composition (Fig. 5i). Such membrane topology effects are known to affect the rate of *S*.Tm docking to host cells[44]. These results indicate that the different molecular response to *Let7* of the tested HeLa cells directly affected a bacterial infection phenotype.

Overall, our data show that the set of proteins that constitute the cell line specific expression response provide useful information about the molecular basis underlying the observed phenotypic differences. We make the entire dataset available to the community online at https://HelaProt.shinyapps.io/Crosslab/ (Supplementary Figs. 18 and 19 and Supplementary Note 5).

Discussion

Although this study focuses on HeLa cells, previous studies suggest that the findings are likely to generalize to other cell lines with unstable genomes. One example is the MCF-7 human breast cancer cell line. Genetic variability in MCF-7 was recently found to be undetectable in routine authentication, and more in-depth systems analyses may be needed to fully appreciate the extend of this variation[45]. Altered ploidy and signaling pathways has been shown in a triple-negative MCF-7 variants[46]. Recently, 27 MCF-7 strains were cultured in the presence of 321 anticancer compounds, and >75% of compounds that strongly inhibited certain clones were completely inactive in others[8]. Genomic studies on other cell lines—for example, HEK293—have also uncovered dynamic changes of aneuploidy in response to cellular manipulations[47]. Together, these studies present multiple lines of evidence that underline the importance of documenting identity and molecular heterogeneity of commonly used, genomically unstable cell lines in research reports.

On the basis of the present findings, we suggest initial, minimal and easy to implement measures to minimize the effects of the observed variability on the publication of research results. First, we

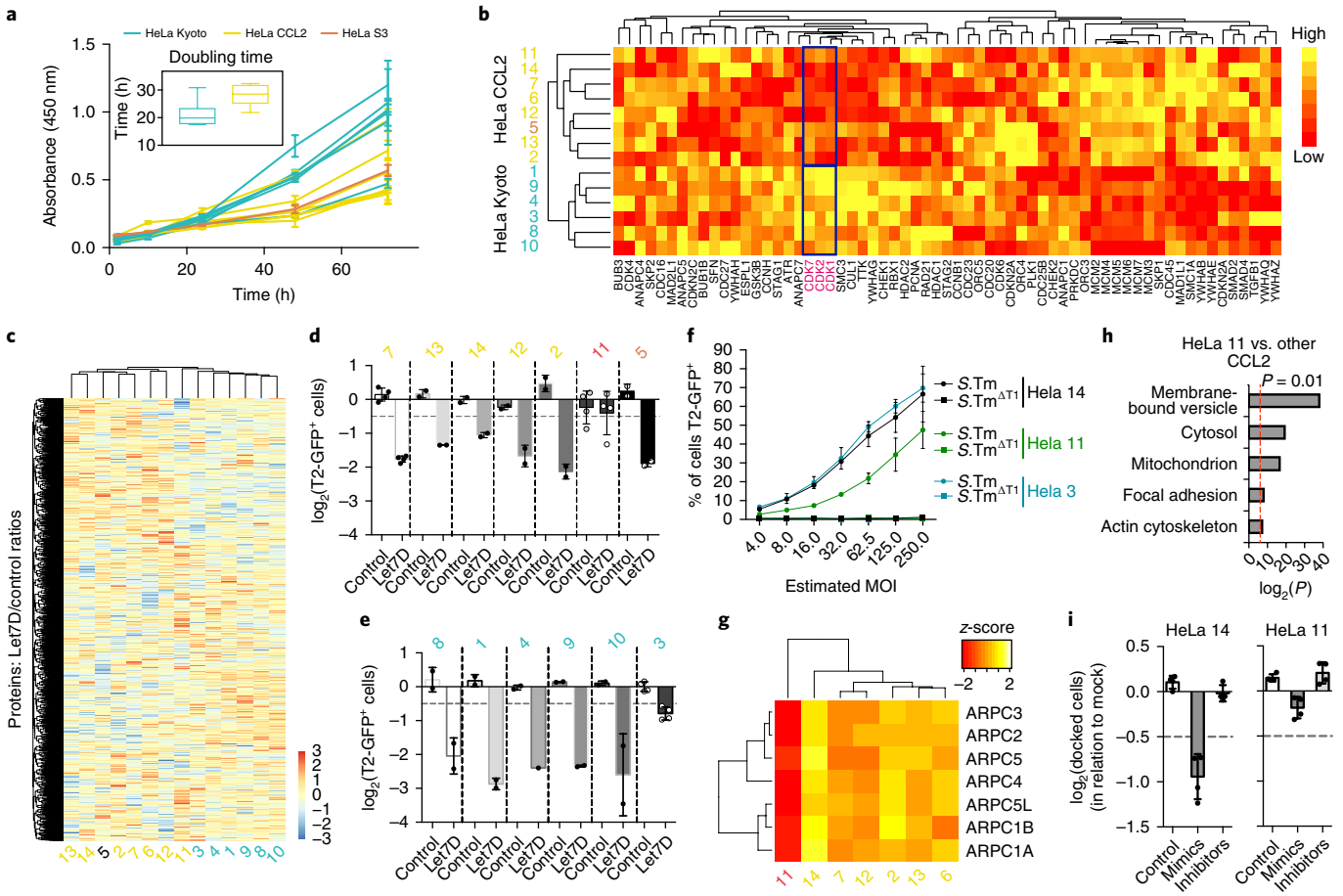


Fig. 5 | Proteotypes of HeLa cells tightly link to phenotypes. a, Cell doubling time measurements and comparisons (inset, box plots) of HeLa Kyoto and HeLa CCL2 lines ($n = 3$ biological replicates), respectively. Error bars denote mean (center) ± s.d. **b**, Expression of 63 proteins annotated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) to be involved in cell cycle process between HeLa CCL2 and Kyoto groups. All CDKs are upregulated in the Kyoto group, including the significant upregulations of CDK 1, 2 and 7 outlined in blue. **c**, HCA of all 5,030 quantified proteins and their ratios between *Let7*-treated vs. control cells for each cell line. **d**, *Salmonella* Typhimurium infection rates quantified in HeLa CCL2 cell lines after transfection with *Let7* mimics, which is expected to decrease the infection rate. Let7D, the sequence form D of let7 microRNA precursor. T2, type-three secretion system 2 in *S.* Tm. T2-GFP+ cells indicate those cells with positive S.Tm infection (Methods). See Supplementary Fig. 15. Error bars, s.d. **e**, Infection rates in HeLa Kyoto cells; error bars, s.d. **f**, A reduced infection rate in HeLa 11 at the respective MOIs as compared to HeLa 14 and HeLa 3; error bars, s.d. **g**, Expression of the Arp2/3 complex among all HeLa CCL2 cells, with expression in HeLa 11 being lowest. **h**, Gene Ontology processes enrichment analysis for the 178 proteins that were differentially expressed between HeLa 11 and other CCL2 cells. *P* values calculated by Fisher's exact test with correction for multiple testing. **i**, HeLa 11 has a less pronounced docking phenotype than HeLa 14. Mimics and inhibitors denote Let7 mimics and hairpin inhibitors, respectively. In **d**–**f** and **i**, $n = 3{,}150$ initial seeding cells; error bars denote mean ± s.d.

strongly suggest that all future HeLa related studies should at least annotate whether CCL2 or Kyoto cell lines were used. Our results demonstrate that HeLa CCL2 and Kyoto cells are consistently and notably different at every level, including cell morphology, doubling time, karyotype, steady-state mRNA expression, protein expression and protein turnover rate. The gene expression variance between CCL2 and Kyoto is as large as that between many other cell lines originating from different tissues. Furthermore, the difference between HeLa CCL2 and Kyoto also results in distinct proteomic responses to *Let7* transfection.

Second, we suggest the use of early and clearly annotated passages of cancer cell lines. In this study we detected substantial divergence of HeLa CCL2 cells after 3 months of continuous culture that resulted in a substantial accumulation CNV changes and differential expression of ~6% of genes. This suggests that even during studies of moderate duration the molecular makeup of the cells may gradually change, resulting in important proteomic, and therefore biochemical, changes. We further recommend that researchers investigate, record and document the sources of the resident cell lines and report detailed cell culture protocols. In the case of collaborative projects involving several groups, both cell passages and protocols should be kept consistent among groups[48]. One example might be the Good Cell Culture Practice (GCCP) guideline in toxicology community[49].

Third, we suggest that important observations derived from single cancer cell lines be repeated in different samples of the cell line, in different cell models or in different laboratories. In addition to alternative individual cancer cell lines, cell line panels of a particular cancer type, primary cells, organoid cultures[50] or induced pluripotent stem cells are worthwhile models to consider[51].

Finally, we hope that our study fuels the community discussions and consideration toward a new level of cell authentication by documenting the precise molecular makeup of the cells, whether carrying the same name or not, used for a study. This step goes beyond authentication based on short tandem repeat or single nucleotide polymorphism profiles. Transcript profiles could be used to document the basic state of the cells used in an experiment[8]. In this study we found that complex phenotypes such as resistance to *S.*Tm

infection could be linked to proteotype patterns. Because of the tight relationship between proteotype and phenotype, we also suggest fast proteome profiling of the cells used in a particular study by a cost-effective label-free approach and appending of the resulting data to publications as an informative documentation of the biochemical state of the cells. This is technically feasible via the SWATH-MS technology, for example, and not prohibitively expensive.

Here we quantified ~5,000 proteins (roughly half of the protein species expressed in HeLa cells) and revealed proteotype heterogeneity between different HeLa cell variants at high resolution. SWATH-MS shows comparable reproducibility to transcriptomics and can distinguish biological signals from technical noise in whole-process replicates. The pSILAC-SWATH–based protein turnover rate determination also achieved fairly high reproducibility in differentiating HeLa Kyoto and CCL2 cells. The HeLa cells from different laboratories essentially present a cell line panel in which the gene dosages are substantially different, but the sequence variances are minimal or modest between cells. The thousands of dosage-changing events observed at each layer provide the opportunity to explore gene dosage effects during gene expression. The inclusion of protein-specific k_{loss} data that cannot be predicted by CNV, mRNA and total protein levels strongly supports protein complex stoichiometry control as a key post-translational regulation in the face of aneuploidy. Also, proteome and proteostasis dynamics of different cellular compartments appeared to vary among HeLa cells. These conclusions would not be transparent if only one layer of omics data was acquired.

Our multilayered dataset not only demonstrates the high degree of heterogeneity of HeLa cells used in different laboratories but also provides a resource for understanding genotype–phenotype relationships in cancer cells.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0037-y.

Received: 29 April 2018; Accepted: 21 November 2018;
Published online: 18 February 2019

References

1. Capes-Davis, A. et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127**, 1–8 (2010).
2. Zhao, M. et al. Assembly and initial characterization of a panel of 85 genomically validated cell lines from diverse head and neck tumor sites. *Clin. Cancer Res.* **17**, 7248–7264 (2011).
3. Lorsch, J. R., Collins, F. S. & Lippincott-Schwartz, J. Fixing problems with cell lines. *Science* **346**, 1452–1453 (2014).
4. Yu, M. et al. A resource for cell line authentication, annotation and quality control. *Nature* **520**, 307–311 (2015).
5. Almeida, J. L., Cole, K. D. & Plant, A. L. Standards for cell line authentication and beyond. *PLoS Biol.* **14**, e1002476 (2016).
6. Muff, R. et al. Genomic instability of osteosarcoma cell lines in culture: impact on the prediction of metastasis relevant genes. *PLoS One* **10**, e0125611 (2015).
7. Frattini, A. et al. High variability of genomic instability and gene expression profiling in different HeLa clones. *Sci. Rep.* **5**, 15377 (2015).
8. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
9. Bottomley, R. H., Trainer, A. L. & Griffin, M. J. Enzymatic and chromosomal characterization of HeLa variants. *J. Cell Biol.* **41**, 806–815 (1969).
10. Nelson-Rees, W. A., Hunter, L., Darlington, G. J. & O'Brien, S. J. Characteristics of HeLa strains: permanent vs. variable features. *Cytogenet. Cell Genet.* **27**, 216–231 (1980).
11. Macville, M. et al. Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer Res.* **59**, 141–150 (1999).
12. Rutledge, S. What HeLa cells are you using? *The Winnower* https://doi.org/10.15200/winn.143896.65158 (2014).
13. Landry, J. J. et al. The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)* **3**, 1213–1224 (2013).
14. Adey, A. et al. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).
15. Williams, E. G. et al. Systems proteomics of liver mitochondria function. *Science* **352**, aad0189 (2016).
16. Gillet, L. C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).
17. Rosenberger, G. et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).
18. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031 (2014).
19. Röst, H. L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).
20. Röst, H. L. et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13**, 777–783 (2016).
21. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
22. Jovanovic, M. et al. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347**, 1259038 (2015).
23. Liu, Y. et al. Systematic proteome and proteostasis profiling in human trisomy 21 fibroblast cells. *Nat. Commun.* **8**, 1212 (2017).
24. Fasterius, E. et al. A novel RNA sequencing data analysis method for cell line authentication. *PLoS One* **12**, e0171435 (2017).
25. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
26. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
27. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
28. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
29. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20 (2017).
30. Lundberg, E. et al. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6**, 450 (2010).
31. Claydon, A. J. & Beynon, R. Proteome dynamics: revisiting turnover with a global perspective. *Mol. Cell. Proteomics* **11**, 1551–1565 (2012).
32. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).
33. Stingele, S. et al. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol. Syst. Biol.* **8**, 608 (2012).
34. Dephoure, N. et al. Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife* **3**, e03023 (2014).
35. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
36. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
37. Roush, S. & Slack, F. J. The let-7 family of microRNAs. *Trends Cell Biol.* **18**, 505–516 (2008).
38. Schulte, L. N., Eulalio, A., Mollenkopf, H. J., Reinhardt, R. & Vogel, J. Analysis of the host microRNA response to *Salmonella* uncovers the control of major cytokines by the let-7 family. *EMBO J.* **30**, 1977–1989 (2011).
39. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, 05005 (2015).
40. Misselwitz, B. et al. RNAi screen of *Salmonella* invasion shows role of COPI in membrane targeting of cholesterol and Cdc42. *Mol. Syst. Biol.* **7**, 474 (2011).
41. Kreibich, S. et al. Autophagy proteins promote repair of endosomal membranes damaged by the *Salmonella* type three secretion system 1. *Cell Host Microbe* **18**, 527–537 (2015).
42. Criss, A. K. & Casanova, J. E. Coordinate regulation of *Salmonella enterica* serovar Typhimurium invasion of epithelial cells by the Arp2/3 complex and Rho GTPases. *Infect. Immun.* **71**, 2885–2891 (2003).
43. Cossart, P. & Helenius, A. Endocytosis of viruses and bacteria. *Cold Spring Harb. Perspect. Biol.* **6**, a016972 (2014).
44. Misselwitz, B. et al. Near surface swimming of *Salmonella* Typhimurium explains target-site selection and cooperative invasion. *PLoS Pathog.* **8**, e1002810 (2012).
45. Kleensang, A. et al. Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. *Sci. Rep.* **6**, 28994 (2016).
46. Leung, E., Kim, J. E., Askarian-Amiri, M., Finlay, G. J. & Baguley, B. C. Evidence for the existence of triple-negative variants in the MCF-7 breast cancer cell population. *Biomed. Res. Int.* **2014**, 836769 (2014).
47. Lin, Y. C. et al. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun.* **5**, 4767 (2014).

48. Geraghty, R. J. et al. Guidelines for the use of cell lines in biomedical research. *Br. J. Cancer* **111**, 1021–1046 (2014).
49. Pamies, D. & Hartung, T. 21st century cell culture for 21st century toxicology. *Chem. Res. Toxicol.* **30**, 43–52 (2017).
50. Lancaster, M. A. & Knoblich, J. A. Organogenesis in a dish: modeling development and disease using organoid technologies. *Science* **345**, 1247125 (2014).
51. Drubin, D. G. & Hyman, A. A. Stem cells: the new "model organism". *Mol. Biol. Cell.* **28**, 1409–1411 (2017).

Acknowledgements

We thank G. Rosenberger, A. Beyer, B. Collins and S. Nikolaev for discussions. We thank L. Reiter, R. Bruderer and O. Rinner from Biognosys AG for sharing their thoughts about cell line proteome analysis from a commercial perspective. We thank H. Zhang and J. Chen from Johns Hopkins University, D. Pflieger and O. Filhol-Cochet from CEA Grenoble, M. Riwanto from University Hospital Zurich, U. Greber and M. Suomalainen from the University of Zurich, C. Arrieumerlou from the University of Basel (through InfectX), M. Beck and M.-T. Mackmull from the European Molecular Biology Laboratory, C. Jorgensen and J. Worboys from the Cancer Research UK Manchester Institute, M. Peter and C. Barnes from ETH Zurich, and A. Venkitaraman and C. Williams from the University of Cambridge for providing us their HeLa cells.

The work was supported by the SystemsX.ch project PhosphoNetX PPM (to R.A.), TargetInfectX (to C.D.), the Swiss National Science Foundation (grant 3100A0-688 107679 to R.A.), the European Research Council (ERC-20140AdG 670821 to R.A.), the JRC for Computational Biomedicine (which was partially funded by Bayer AG, to J.S.-R.), the Swiss National Science Foundation (grant 163180 to S.E.A.), the European Research Council (grants AdG 249968 to S.E.A. and 616441-DISEASEAVATARS to G.T.), the Umberto Veronesi Foundation (fellowship to P.-L.G.), the ERA-NET Neuron Program (P.-L.G.), Regione Lombardia (Ricerca Indipendente 2012 to G.T.) and the Italian Ministry of Health (Ricerca Corrente to G.T.) E.G.W. was supported by an NIH F32 Ruth Kirchstein Fellowship (F32GM119190).

Author contributions

Y.L. and R.A. designed and supervised the whole project. Y.L., Y.M., E.G.W., P.-L.G., M.F., I.B., M.S., M.E. and F.B. analyzed the data and performed the bioinformatics analysis. Y.M. developed the HeLa Proteome website. T.M. performed the pSILAC experiment. S.K. and Y.L. performed the *Let7* experiment. S.K. performed the *S.*Tm infection experiment. A.V.D., C.B, I.S., C.D. and H.Z. established and cultured the cell lines. Y.L. and M.M. performed the mass spectrometry experiments. I.B. performed pyProphet analysis. F.S.B. generated CNV data. M.S. processed the CNV data. C.B. generated RNA-seq data. M.F. performed sequence variation analysis. F.B. and P.-L.G. analyzed RNA-seq data. M.E. analyzed the microscopy phenotypic data. G.T. and J.S.-R. supervised data interpretation. S.E.A. supervised the genomics data generation. W.-D.H. supervised all the microbiology experiments and provided critical inputs. Y.L., E.G.W. and R.A. wrote the paper.

Competing interests

R.A. holds shares of Biognosys AG, which operates in the field covered by the article.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0037-y.

Correspondence and requests for materials should be addressed to Y.L. or R.A.

Methods

Collection of cells. Frozen HeLa cells were prepared in each laboratory and sent to the coordinating laboratory at ETH Zurich in dry ice for centralized culturing. A uniform protocol was used at each site to prepare the cells for shipment. Cells collected from a 15-cm dish were transferred to 1 mL of the medium used for freezing (70% volume of DMEM, 20% FBS, 10% DMSO). Cells were then placed at –20 °C for 2–3 h, transferred to –80 °C for 24 h, and stored in liquid nitrogen until delivery. The frozen state of the delivered cells and presence of dry ice were confirmed upon arrival. Additionally, two aliquots of cell pellets were prepared at each site. The HeLa strains were then centrally cultured in the coordinating laboratory using standard culture methods according to ATCC. HeLa cells were tested upon arrival in the central laboratory for mycoplasma using the GATC Biotech service (Germany) and confirmed to be negative. For central culturing, the same researcher cultured the cells using the same culture medium and protocol. We did not notice routine symptoms of microbial contamination in any of the cell lines used upon careful daily inspection of the cultures under a microscope.

Culture conditions were 5% CO_2, 37 °C; Dulbecco's modified Eagle medium (DMEM, Gibco, 41965-039). To make the complete growth medium, FBS (Sigma Aldrich, F7524) was added to a final concentration of 10%, together with a penicillin/streptomycin/glutamine solution (Gibco, 10378016). We used 15-cm dishes (Nunclon 100 × 15 mm, Airvent). Detailed protocols are provided in Supplementary Note 1. We used 80% confluency to determine cell subculture frequency. Cells were seeded based on a split ratio of 1:2 to 1:4. The labeling of cell cultures was done through both the initials of the laboratory head (HeLa identity unknown to the technician) and an assigned number. Experts in the respective multi-omic techniques, blinded to the cell identities, carried out the measurements at each layer of gene expression, thus contributing to a multi-layer dataset of related cultured cells.

Determination of cell doubling time. The cell doubling time for each HeLa strain was determined using a cell counting CCK-8 kit (Dojindo Laboratories, Japan). Cells were seeded in triplicate at a density of 2,800 cells per well of a 96-well plate, and samples were prepared for counting at five different time points: 2 h, 11 h, 24 h, 48 h and 72 h. The final doubling time was then calculated (using http://www.doubling-time.com/compute.php). The entire experiment was repeated on two cell lines. Cell doubling time differences between whole-process replicates were all less than 2 h.

Phenotypic characterization of HeLa cells. HeLa cell plates were imaged with Molecular Devices ImageXpress microscopes, using MetaXpress v5.1 software. Imaging settings were adjusted for the highest exposure that did not incur overexposure, with 14-bit dynamic range and laser-based focusing. We imaged nine sites per well in a 3 × 3 grid with no spacing and no overlap. Three channels where imaged, with filters for DAPI for the nucleus, GFP for bacteria, and RFP for F-actin. Robotic plate handling was used to load and unload plates (Thermo Scientific).

Array-CGH analysis. Array-CGH (aCGH) analyses were performed using the Agilent Human Genome CGH Microarray Kit G3 180K (Agilent Technologies, Palo Alto, CA, USA) with 13 kb overall median probe spacing. The control was a DNA pool from 7 diploid individuals. Labeling and hybridization were performed following the protocols provided by the manufacturer.

aCGH processing and gene copy number detection. aCGH measurements of 173,540 genomic probes were used to compute an aCGH $\log_2$-ratio profile that compares the DNA copy number of each probe in a specific HeLa cell line to normal diploid DNA. aCGH profiles were sorted according to the chromosomal locations of probes and further segmented into chromosomal regions of constant copy number using DNAcopy[52] (R package DNAcopy with settings smooth.region = 3, outlier.SD.scale = 0.5, smooth.SD.scale = 0.25). Copy number values of individual genes (30,237 known canonical genes of hg19/GRCh37) were determined by mapping chromosomal location of genes to obtained aCGH segments. If a gene was covered by a whole segment, then its copy number value was set to the segment-specific $\log_2$ ratio. If a gene was covered by more than one adjacent segments (breakpoints within a gene), then its copy number value was set to a weighted average of $\log_2$ ratios of involved segments according to their overlap with the gene. Most parts of the HeLa genomes are known to be triploid[13,14]. This was also reflected in our aCGH profiles, but the position of the closest peak to the triploid state varied among the different HeLa cell lines. Therefore, we further aligned the obtained gene copy number measurements to ensure that this peak was located at the triploid state for each cell line.

RNA extraction, library preparation and mRNA sequencing. Total RNA was collected using the TRIzol reagent (Life Technologies) following the manufacturer's instructions. RNA quality was verified on an Agilent 2100 Bioanalyzer (Agilent) and quantity was measured on a Qubit instrument (Life Technologies).

Libraries were prepared with 4 µg of total RNA using TruSeq RNA kit (Illumina) according to the manufacturer's instructions. Libraries were sequenced on an Illumina HiSeq2000 machine as 100-bp single-end reads. The reads were aligned to the hg19 human genome using TopHat (v2.1.1)[53] with standard configurations (no more than two mismatches allowed). The numbers of reads for the genes are calculated using the GENCODE v24 release. Only uniquely mapped reads were included.

Deep analysis of single nucleotide variants and cell line authentication. The RNA sequencing (RNA-seq) results of all HeLa cell lines as well as one Hek293 cell line were analyzed for small sequence variants (i.e., single nucleotide variants (SNV) and small insertions/deletions (indels)). Raw RNA-seq reads for the HEK293 cell line were publicly available and obtained through the NCBI sequence read archive[54] (accession SRX1300887, downloaded 17 January 2017 from https://www.ncbi.nlm.nih.gov/sra?term=SRP064410). All RNA-seq reads were manually checked with the FastQC[55] tool (version 0.11.4). Then reads were quality trimmed with Trimmomatic[56] (version 0.35) and a second quality check was performed with FastQC. Alignment to the GRCh38 reference genome was performed with STAR[57] (version 2.4.2a).

Single nucleotide variants were called individually for every cell line, according to a best practices recommendation within the GATK framework[58] (https://www.broadinstitute.org/gatk/guide/article?id=3891). The pipeline marks duplicate reads in the alignment file with the Picard Tools (http://broadinstitute.github.io/picard, version 2.0.1) MarkDuplicates function. Then the following GATK[58] (version 3.7) tools are sequentially employed: SplitNCigarReads, HaplotypeCaller, VariantFiltration. Variants were filtered according to the following criteria: Fisher Strand value above 30, Quality by Depth value greater than 2 and clusters no more than two variants within a 35-base-pair window.

Pairwise cell line concordance was determined as previously described[24]. Briefly, for every cell line pair, concordance marked the fraction of identical variant calls among all variant calls between the cell lines.

SNV mapping to the HeLa cell line dataset in the COSMIC database. Previous research has demonstrated the possibility of authenticating cell lines by means of comparing SNV profiles derived from RNA-seq to publicly available SNV profiles derived from genomic sequencing. To authenticate the cell lines used in this study as HeLa, their SNV profiles were compared with the SNV profiles of the COSMIC cell lines project[25] as previously described[24]. Mutation data were obtained using the GRCh38 assembly with release v85. An overview of the cell line characterization can be found at https://cancer.sanger.ac.uk/cell_lines/sample/overview?id=1298134.

Pairwise cell line concordance and concordance with COSMIC profiles was determined using in-house R scripts. Briefly, for every cell line variant, calls from COSMIC were subset to the genomic loci where variants were found by RNA sequencing. The average coverage of the COSMIC loci was 34.3%, which is in accordance with the expected coverage of genomic loci that can be achieved by RNA-seq variant calling[59]. Concordance marks the proportion of the overlapping variant calls that are consistent (i.e., the two alleles are identical).

Transfection of Let7d mimics to HeLa cells. Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific) was used to deliver the microRNA human *Let7d* mimics (cat. no. 4464066) as *Let7* treatment, Negative Control microRNA (4464058), and Positive Control for transfection efficiency (*Kif11*; 4390824) provided by Life Technologies Europe (Zug, Switzerland) to all the HeLa strains expect for one GFP-positive stain (HeLa 6). The ratio of RNAiMax and DMEM was kept at 1:500 (v/v). The initial seeding concentration was 40 cells/µL for all strains and the working concentration of miRNAs was 20 nM. The 6-well plate containing 2 mL final transfection medium per well and the 96-well plate containing 100 µL DMEM medium were used for perturbed proteomic measurements and *S.*Tm infection, respectively, 72 h after the transfection.

***Salmonella* Typhimurium (*S.*Tm) infection.** After transfection, HeLa cells in 96-well plate format were cultured for 72 h before *S.*Tm infection using the pipeline described before[40]. Cells were infected for 20 min with *S.*Tm, incubated for 3 h and 40 min in medium with 400 µg/ml gentamicin, fixed by 4% PFA containing 4% sucrose, and stained with DAPI and DY-547–phalloidin. All liquid-handling steps, including the infection, fixation and staining, were performed manually. The high-throughput image acquisition was performed using a Molecular Devices ImageXpress microscope (10× S Fluor). During the internalization of bacteria into the host cells, they first form an early SCV (*Salmonella* containing vacuole), which then matures and acidifies over the course of infection, and the acidification serves as the main trigger to induce the expression of SPI-2 (*Salmonella* pathogenicity island 2) with its T2 system (type-three secretion system 2). Since the *Salmonella* strain used harbors a gfp reporter under the control of a T2 promoter, the *Salmonella* of the late stage SCV will be green fluorescent (i.e., T2-GFP+) and can be detected in the microscope. A CellProfiler-based image analysis pipeline was applied to determine the infection rate of *S.*Tm for each cell line of the control and *Let7d* mimic transfection conditions.

We tested whether *Let7* mimics would have a phenotype regarding the docking step of *Salmonella* infection (Supplementary Note 6). To do so, cells were infected with *S.*Tm$^{\Delta 4}$ (a noninvasive *S.*Tm mutant that lacks the four main SPI-1 effectors SopE, SopE2, SipA and SopB) at an MOI of 125 for 6 min at 37 °C and 5% CO_2.

This noninvasive mutant strain allowed us to measure the binding capacity of *S.* Tm. Afterwards, the cells were washed three times with 60 µL DMEM/10% FCS and fixed with 60 µL 4% PFA. To visualize bound *S.*Tm$^{\Delta T}$, we performed immunofluorescence staining using a primary anti-LPS antibody (BD, 226601) and a FITC-bound secondary antibody (Jackson ImmunoResearch, 111-095-144). Afterwards, cells were permeabilized and then nuclei were stained with DAPI (Sigma-Aldrich, D9542).

Pulsed SILAC experiment. For the pSILAC experiment, SILAC DMEM High Glucose medium (GE Healthcare) lacking L-arginine and L-lysine was first supplemented with light or heavy isotopically labeled lysine and arginine, 10% dialyzed FBS (PAN Biotech), and 1% penicillin/streptomycin mix (Gibco). Specifically, 146 mg/L of heavy L-lysine ($^{13}C_6{}^{15}N_2$) and 84 mg/L of arginine ($^{13}C_6{}^{15}N_4$) (Chemie Brunschwig AG) and the same amount of corresponding unlabeled amino acids (Sigma-Aldrich)[21] were supplemented respectively to configure heavy and light SILAC medium. Additionally, 400 mg/L L-proline (Sigma-Aldrich) was also added to SILAC medium to prevent potential arginine-to-proline conversion. HeLa variants were first cultured on 15-cm cell culture dishes in pre-prepared light SILAC medium and stabilized in culture for 3–4 d. Upon release of cells by 0.25% trypsin/EDTA, cells were counted using a Neubauer hemocytometer. Subsequently, six 10-cm dishes were prepared for each cell variant with a seeding density of 1.5×10^6 cells per plate, corresponding to three time points with two replicates each. The cell culture plates were incubated for 14 h, at 5% CO_2 and 37 °C, overnight. Cells were then washed three times with PBS at 37 °C. The medium was then replaced by heavy SILAC (K8R10) medium. Cells were harvested and counted in two biological replicates at four different time points (0 h, 1 h, 4.5 h and 11 h). Two dishes of whole-process replicate were prepared at each time. The cell pellets were snap frozen in liquid nitrogen after removal of the PBS and stored at −80 °C.

Protein extraction and in-solution digestion. HeLa cells and cell pellets harvested from shipped tubes, centrally cultured conditions, *Let7d* treated and control experiments, and the pSILAC experiment were suspended in 10 M urea lysis buffer and complete protease inhibitor cocktail (Roche) and ultrasonically lysed at 4 °C for 2 min with two rounds using a VialTweeter device (Hielscher-Ultrasound Technology). The mixtures were centrifuged at 18,000*g* for 1 h to remove insoluble material. Protein in the supernatant was quantified by Bio-Rad protein assay. Protein samples were reduced with 10 mM tris-(2-carboxyethyl)-phosphine (TCEP) for 1 h at 37 °C and 20 mM iodoacetamide in the dark for 45 min at room temperature. All samples were further diluted 1:6 (v/v) with 100 mM NH_4HCO_3 and were digested with sequencing-grade porcine trypsin (Promega) at a protease/protein ratio of 1:25 overnight at 37 °C. Digestion was carried out in a 96-well plate format to increase experimental reproducibility. Amounts of purified peptides were determined using a Nanodrop ND-1000 (Thermo Scientific) and 1 µg of peptides were injected in each LC-MS run.

Shotgun proteomics on TripleTOF mass spectrometer. The peptides digested from cell lysate derived from the first time point samples in pSILAC experiment were measured on an SCIEX 5600 TripleTOF mass spectrometer operated in DDA mode[16,60,61] by SCIEX Analyst v1.7. The mass spectrometer was interfaced with an Eksigent NanoLC Ultra 1Dplus HPLC system. Peptides were directly injected onto a 20-cm PicoFrit emitter (New Objective, self-packed to 20 cm with Magic C18 AQ 3-µm 200-Å material) and then separated using a 90-min gradient from 5–35% (buffer A: 0.1% (v/v) formic acid, 2% (v/v) acetonitrile; buffer B: 0.1% (v/v) formic acid, 100% (v/v) acetonitrile) at a flow rate of 300 nL/min. MS1 spectra were collected in the 360–1,460 *m/z* range with 250 ms per scan. The 20 most intense precursors with charge state 2–5 that exceeded 250 counts per second were selected for fragmentation, and MS2 spectra were collected in the 50–2,000 *m/z* range for 100 ms. The precursor ions were dynamically excluded from reselection for 20 s.

SWATH mass spectrometry. Normal proteome, *Let7d*-treated cells and controls, and pSILAC samples were all measured by SWATH-MS. The same LC-MS/MS systems used for shotgun measurements on SCIEX 5600 TripleTOF was also used for SWATH analysis[16,60,61]. For normal proteome samples and pSILAC samples, a 90-min LC gradient was used, whereas a 60-min gradient was used for the two biological replicates of the *Let7d* experiment. Specifically, in the present SWATH-MS mode, the SCIEX 5600+ TripleTOF instrument was specifically tuned to optimize the quadrupole settings for the selection of 64 variable-width precursor ion selection windows. The 64-variable window schema was optimized based on a normal human cell lysate sample, covering the precursor mass range 400–1,200 *m/z*. The effective isolation windows were 399.5–408.2, 407.2–415.8, 414.8–422.7, 421.7–429.7, 428.7–437.3, 436.3–444.8, 443.8–451.7, 450.7–458.7, 457.7–466.7, 465.7–473.4, 472.4–478.3, 477.3–485.4, 484.4–491.2, 490.2–497.7, 496.7–504.3, 503.3–511.2, 510.2–518.2, 517.2–525.3, 524.3–533.3, 532.3–540.3, 539.3–546.8, 545.8–554.5, 553.5–561.8, 560.8–568.3, 567.3–575.7, 574.7–582.3, 581.3–588.8, 587.8–595.8, 594.8–601.8, 600.8–608.9, 607.9–616.9, 615.9–624.8, 623.8–632.2, 631.2–640.8, 639.8–647.9, 646.9–654.8, 653.8–661.5, 660.5–670.3, 669.3–678.8, 677.8–687.8, 686.8–696.9, 695.9–706.9, 705.9–715.9, 714.9–726.2, 725.2–737.4, 736.4–746.6, 745.6–757.5, 756.5–767.9, 766.9–779.5, 778.5–792.9, 791.9–807, 806–820, 819–834.2, 833.2–849.4, 848.4–866, 865–884.4, 883.4–899.9, 898.9–919, 918–942.1, 941.1–971.6, 970.6–1,006, 1,005–1,053, 1,052–1,110.6, 1,109.6–1,200.5 (containing 1 *m/z* for the window overlap). SWATH MS2 spectra were collected from 50 to 2,000 *m/z*. The collision energy was optimized for each window according to the calculation for a charge 2+ ion centered on the window with a spread of 15 eV. An accumulation time (dwell time) of 50 ms was used for all fragment-ion scans in high-sensitivity mode, and for each SWATH-MS cycle a survey scan in high-resolution mode was also acquired for 250 ms, resulting in a duty cycle of ~3.45 s. A 100-fmol β-galactosidase standard digest (SCIEX) was injected between each pair of runs to monitor instrument performance and tune the mass accuracy of MS1 and MS2 signals in a real-time manner throughout the sample acquisition.

SWATH-MS data extraction of protein expression data. With the exception of the pSILAC dataset, all SWATH-MS datasets were analyzed and identified by OpenSWATH software[19] searching against a previously established SWATH assay library that contains mass spectrometric assays for 10,000 human proteins[18]. Profile-mode .wiff files from shotgun data acquisition were centroided and converted to mzML format using the AB Sciex Data Converter v.1.3 and converted to mzXML format using MSConvert v.3.04.238 before OpenSWATH analysis. OpenSWATH from OpenMS (Git OpenMS/develop@ 4bca6fc) was first used to the peak groups from all individual SWATH maps with statistical control (see below) and then aligned between SWATH maps using a novel TRIC (transfer of identification confidence)[20]. For large-scale targeted proteomics, protein FDR control needs specific attention and should be equally important as in shotgun proteomics[17,18]. Therefore, to pursue a strict statistical quality control of peptide- and protein-level identification, we used the newly developed PyProphet extended version[17]. This version combines the set of scores from OpenSWATH for each peptide query to a single discriminant score using semisupervised learning to best separate decoys from high-scoring targets. Particularly for all the label-free samples, PyProphet was run to estimate *q*-values[17] for all runs (run-specific context) and in a global fashion (global context). A strategy applying two steps of filtering was used. For proteins accepted by PyProphet at FDR <1% in the global context, the sets of peak groups detected at 1% FDR in the run-specific context were included for quantification. This criteria yielded 4,335 proteins. For proteins accepted with an FDR <5% in the global context, only those peak groups detected at 1% FDR and also identified in >25% of the total MS runs were accepted. The requantification feature in OpenSWATH was enabled for the filtered protein list. In total 50,225 peak groups were identified, corresponding to 46,951 unique peptides (43,521 tryptic peptides) assigned to 5,030 unique SwissProt proteins.

To quantify the protein abundance levels across samples, we summed up the most abundant peptides for each protein (i.e., top three peptide groups based on intensity were used for those proteins identified with more than three proteotypic peptide signals whereas all the peptides were summarized for other proteins). This allows a reliable estimate of global protein level changes[60,62]. The protein expression data matrix was $\log_{10}$ transformed and quantile normalized for statistical and bioinformatics analysis.

SWATH-MS data extraction of pSILAC data. The centroid-converted mzML files from shotgun analysis of the first time point samples in pSILAC experiment were searched against different engines including using the iPortal platform[63] to establish the sample-specific library for pSILAC data. Nine published HeLa runs included in the Pan-Human Library[18] were also used to generate this library. iPortal used the iProphet schema[64] to integrate the search results from X!Tandem, Omessa, Myrimatch and Comet at a peptide-level FDR of 1% by target–decoy strategy. The Xinteract option was "-dDECOY_ -OAPdlIw". Specially, peptide tolerances at MS and MS/MS level were set to be 50 ppm and 0.1 Da, respectively. Up to two missed trypsin cleavages were allowed. Oxidation at methionine was set as the variable modification whereas carbamidomethylation at cysteine was set as the fixed modification. The iPortal identification result finally contained 3,973 proteins and 42,236 peptides at 1% protein FDR cutoff.

The light version of the consensus spectral library was generated using SpectraST[65]. Then the spectrast2tsv.py function in OpenSWATH[19] was used to generate the light and heavy MS assays as the final library (including the decoy transitions), which was constructed from top six most intense y-ion fragments with Q3 range from 400 to 1,200 *m/z* excluding those falling in the precursor SWATH window. This final library was used for targeted data analysis of SWATH maps. OpenSWATH analysis was run with target FDR of 1% and extension FDR of 5%[19] (quality cutoff to still consider a feature for alignment) and aligned by TRIC[20], whereas requantified data points were discarded for protein turnover calculation.

Determining protein turnover rates. In pSILAC, the quantification of heavy and light signals of proteins at different time points after pulse labeling permits the quantification of protein-specific turnover rates. The protein turnover rate was determined similarly to that in our previous study[23]. The rates of loss of the light isotope (k_{loss}) were directly calculated from the output data matrix generated by OpenSWATH, following the methods previously described by Pratt et al.[66]: specifically, we modeled the relative isotope abundance (RIA), defined as the signal intensity in the light channel divided by the sum of light and heavy intensities, onto

an exponential decay model assuming a null heavy intensity (RIA = 1) at time 0; i.e., $RIA(t) = e^{-k_{loss}t}$.

We used nonlinear least-squares estimation[23] to perform the fit and to further perform a weighted average of the k_{loss} value of each peptide in the protein, which ensured giving more weight to peptides carrying robust information[23].

Boisvert et al.[67] reported a significant recycling of the unlabeled amino acid from the light protein degradation in HeLa cells; they had to use a 0.8 offset to calculate the protein degradation rate (k_{deg}) from k_{loss}. Therefore, we simply used a direct proxy of turnover rate—i.e., $\log_2(k_{loss})$—whenever applicable, to perform the cross-cell and multi-omics comparisons and illustrations[31]. This proxy parameter essentially avoids the data distortion due to the possible different light amino acid recycling speed and inaccuracy of cell doubling time determination. Only those k_{loss} values assayed in every cell sample were accepted for cross-comparison.

Estimation of absolute protein copies. HeLa cells are a heavily investigated cell line for which many research resources are available. In a study published by Zeiler et al.[68], the absolute protein copy numbers in a HeLa cell line of unknown identity were determined using Protein Epitope Signature Tag (PrEST) and SILAC-based absolute quantification[68]. For all anchor PrEST proteins reported by Zeiler et al.[68], we covered the full dynamic range in this study. This enabled us to correlate 23 HeLa anchor proteins with their respective copy numbers reported by Zeiler et al.[68] to our summed top three peptide SWATH-MS intensity estimates for each protein. The high Pearson correlation coefficient (average $R = 0.872$) allowed us to directly utilize the correlation equation in each cell line to infer the protein copy numbers for all the protein identified by OpenSWATH.

Other bioinformatic analyses. To calculate the Kyoto/CCL2 ratios at all levels, HeLa 11 was excluded because of its deviating genome dosage type (Fig. 1b), so there were six cell variants in each group. All statistical tests used are two-sided. In the box plots, the whiskers represent the range from minimal to maximal values. All error bars in Fig. 5 and the Supplementary Figures denote s.d.

Cellular compartment annotation was done by mapping protein identifies to a recently established subcellular map of the human proteome[35]. The David bioinformatics resource v6.8 (https://david.ncifcrf.gov) was used to extract the protein annotations for other organelles of interest that are not covered by the subcellular atlas[35]. Protein complex information was extracted from the CORUM database[32] and mapped through SwissProt ID. To compare HeLa cells at each omics layer, we performed simple quantile normalization for the CNV, mRNA, protein and k_{loss} data, respectively, and visualized their variation through principal component analysis (PCA) (Fig. 2d–g) and unsupervised hierarchical clustering analysis (HCA, Fig. 2h–k). To combine the RNA-seq datasets from this study and the GDSC cell lines available[26], we used the $\log_{10}$(RPKM) data and centered the HeLa and GDSC cells (for those tissue types represented by at least five GDSC cancer cell lines) individually, then combined them (by bind) for quantile normalization before all t-SNE and PCA analyses. The t-SNE plots are done using the R package Rtsne version 0.13. As parameters, we used a perplexity of 20 and 500 iterations. PCA was done using the R package FactoMineR version 1.39; we chose 5 as the number of principal components. For differential gene expression analysis between HeLa 11 and other cells, we fitted a linear model using the R package limma version 3.30.13 and then corrected by false discovery rate with adjusted P-value at 0.05. The gene set enrichment analysis was performed using the STRING database[69] (https://string-db.org). For the circle plots, we used R package RCircos version 1.2.0 and initialized the cytoband with UCSC.HG19.Human.CytoBandIdeogram. For HCA, data are centered in both gene and sample dimensions before clustering and heat map visualization. Heat maps are made using R package pheatmap version 1.0.8. For the website, we used the R packages shiny version 1.0.5, shinydashboard version 0.6.1 and shinyBS. The colored scatterplots from blue-to-yellow (Figs. 3b and 4d,e) were visualized by "heatscatter" function in R "LSD" package using a two-dimensional Kernel Density Estimation.

To generate the box plots in the figures and supplementary figures, we used the boxplot function (R package "graphics") with default parameters. This means the central rectangle in the plot spans the first quartile to the third quartile (the interquartile range, or IQR). The bold line within the box represents the median of each dataset. The single dot within the box, if present, denotes the mean value of the dataset. The whiskers are defined as $\min(\max(x), Q\_3 + 1.5 \times IQR)$ for the upper whisker and $\max(\min(x), Q\_1 - 1.5 \times IQR)$ for the lower whisker. Here, Q_1 and Q_3 represent the 25th and 75th percentiles, respectively, and IQR = Q_3 – Q_1. Outliers were defined based on whisker positions.

For violin plot in Fig. 4f, we used vioplot (version 0.2) with default settings. This is the standard violin plot, which uses a combination of the box plot and kernel density as the violin curve to show the distribution of the data. The thick black bar in the center represents the interquartile range, the thin black line extended from it represents the 95% confidence intervals, and the white dot is the median.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

RNA-seq data are available on GEO (GSE111485). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[70] partner repository with the dataset identifier PXD009273. The full dataset is available at https://HelaProt.shinyapps.io/Crosslab/.

References

52. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
53. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
54. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
55. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2018).
56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
57. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome AnalysisToolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
59. Cirulli, E. T. et al. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome. Biol.* **11**, R57 (2010).
60. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786 (2015).
61. Collins, B. C. et al. Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods* **10**, 1246–1253 (2013).
62. Ludwig, C., Claassen, M., Schmidt, A. & Aebersold, R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol. Cell. Proteomics* **11**, M111.013987 (2012).
63. Kunszt, P. et al. iPortal: the Swiss grid proteomics portal: requirements and new features based on experience and usability considerations. *Concurr. Comput.* **27**, 433–445 (2015).
64. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10**, M111.007690 (2011).
65. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
66. Pratt, J. M. et al. Dynamics of protein turnover, a missing dimension in proteomics. *Mol. Cell. Proteomics* **1**, 579–591 (2002).
67. Boisvert, F. M. et al. A quantitative spatial proteomics analysis of proteome turnover in human cells. *Mol. Cell. Proteomics* **11**, M111.011429 (2012).
68. Zeiler, M., Straube, W. L., Lundberg, E., Uhlen, M. & Mann, M. 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* **11**, O111.009613 (2012).
69. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
70. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).

nature research

Corresponding author(s): Ruedi Aebersold

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Commercial Software: SCIEX Analyst v1.7, MetaXpress v5.1
Data analysis	Data in this study were analyzed by public softwares: R v3.3, TopHat v2.1.1, GENCODE v24, FastQC version 0.11.4, Trimmomatic version 0.35, STAR version 2.4.2a, Picard version 2.0.1, GATK version 3.7OpenSWATH (OpenMS develop@ 4bca6fc), SpectraST 5.0, AB Sciex Data Converter v.1.3, MSConvert v.3.04.238, PyProphet v2.0, David Bioinformatics v6.8, pheatmap version 1.0.8, shinydashboard version 0.6.1, RCircos version 1.2.0, FactoMineR version 1.39, limma version 3.30.13, Rtsne version 0.13

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-seq data are available on GEO (GSE111485).
The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE 69 partner repository with the dataset identifier PXD009273.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical methods were not used to predetermine sample size. Sample size for each experiment is stated in figure captions and data descriptions.
Data exclusions	No data were excluded.
Replication	All attempts at replication were successful and are presented. Specifically the sample-to-sample Pearson correlation was used.
Randomization	The samples (Hela cells) are randomized by names in the beginning of experiments being done.
Blinding	Investigators are blinded to the group information during data collection, but not to data analysis, because classification and comparison are needed.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Anti-LPS (BD, catalogue number 226601, Salmonella O Antiserum, Factor 5) FITC coated secondary antibody: Fluorescein (FITC) AffiniPure Goat Anti-Rabbit IgG (H+L) (Jackson ImmunoResearch, catalogue number 111-095-144) DAPI: 4',6-Diamidino-2-phenylindole by Sigma-Aldrich. Catalogue number D9542
Validation	Anti-LPS: https://www.bd.com/en-us/offerings/capabilities/microbiology-solutions/stains-and-reagents/bd-antigens-and-antisera. FITC: https://www.jacksonimmuno.com/catalog/products/111-095-144 DAPI: https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma/Product_Information_Sheet/d9542pis.pdf

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa cells at Ruedi Aebersold lab were purchased from ATCC. Furthermore, the following labs provided HeLa cells: Hui Zhang (Johns Hopkins University), Odile Filhol-Cochet (CEA Grenoble), Meliana Riwanto (University Hospital Zurich), Urs Greber (University of Zurich), Cécile Arrieumerlou lab (University of Basel), Martin Beck (European Molecular Biology Laboratory), Claus Jorgensen (Cancer Research UK Manchester Institute), Matthias Peter (ETH Zurich), Ashok Venkitaraman (University of Cambridge), Wolf-Dietrich Hardt (ETH Zurich), Christoph Dehio (University of Basel), Stylianos E. Antonarakis (University of Geneva Medical School).
Authentication	The cell authentication was performed by SNV profiles.
Mycoplasma contamination	All the cells before any proteomic studies were confirmed to be mycoplasma negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.