PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Identifying the genes impacted by cell proliferation in proteomics and transcriptomics studies

Marie Locard-Paulet, Oana Palasca, Lars Juhl Jensen*

Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

* lars.juhl.jensen@cpr.ku.dk

OPEN ACCESS

Citation: Locard-Paulet M, Palasca O, Jensen LJ (2022) Identifying the genes impacted by cell proliferation in proteomics and transcriptomics studies. PLoS Comput Biol 18(10): e1010604. https://doi.org/10.1371/journal.pcbi.1010604

Editor: Attila Csikász-Nagy, Pázmány Péter Catholic University: Pazmany Peter Katolikus Egyetem, HUNGARY

Received: June 16, 2022

Accepted: September 26, 2022

Published: October 6, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010604

Data Availability Statement: All the scripts and input tables associated to this study are available on Zenodo.org (10.5281/zenodo.6346643) under a BSD2-Clause "Simplified" license.

Abstract

Hypothesis-free high-throughput profiling allows relative quantification of thousands of proteins or transcripts across samples and thereby identification of differentially expressed genes. It is used in many biological contexts to characterize differences between cell lines and tissues, identify drug mode of action or drivers of drug resistance, among others. Changes in gene expression can also be due to confounding factors that were not accounted for in the experimental plan, such as change in cell proliferation. We combined the analysis of 1,076 and 1,040 cell lines in five proteomics and three transcriptomics data sets to identify 157 genes that correlate with cell proliferation rates. These include actors in DNA replication and mitosis, and genes periodically expressed during the cell cycle. This signature of cell proliferation is a valuable resource when analyzing high-throughput data showing changes in proliferation across conditions. We show how to use this resource to help in interpretation of *in vitro* drug screens and tumor samples. It informs on differences of cell proliferation rates between conditions where such information is not directly available. The signature genes also highlight which hits in a screen may be due to proliferation changes; this can either contribute to biological interpretation or help focus on experiment-specific regulation events otherwise buried in the statistical analysis.

Author summary

Nowadays, one can routinely measure how thousands of genes and proteins are regulated using so-called omics technology. This is used in many areas of biology, for example, to explore the differences between cancer cell lines and to understand what drugs do to the cells in our body. Interpreting the results of these experiments is challenging: it often results in a list of hundreds of regulated genes, which makes it difficult to pinpoint specific genes for follow up with further studies. Here, we combined data sets from two omics technologies—proteomics and transcriptomics—of more than a thousand cancer cell lines growing at different speed. We calculated the correlation of all their genes to how fast the cells were growing, to find genes that correlate reproducibly in both proteomics and transcriptomics data. These constitute what we call a "proliferation signature", which can tell us how fast the cells are growing in proteomics and transcriptomics experiments,

Funding: MLP, OP and LJJ are supported financially by the Novo Nordisk Foundation (Grant agreement NNF14CC0001). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

where this cannot be easily measured. Furthermore, these signature genes can be regulated not because of the specific treatment or disease of interest, but because of changes in cell growth that were not accounted for in the experimental plan. This resource helps target selection in screens by revealing experiment-specific regulation events, otherwise buried in a long gene list.

Introduction

Nowadays, high-throughput proteome profiling allows relative quantification of thousands of proteins across samples. It is used in many biological contexts to characterize differences between cell lines and tissues, determine drug mode of actions, identify drivers of drug resistance, to name a few. While this reveals meaningful gene regulations across numerous conditions, these results can be confounded by secondary effects of a given treatment (or biological context). For example, a change in cell proliferation is a common undesired side effect of biological treatment and a well acknowledged confounding factor that influences results without being the intended effect of a given treatment [1]. Indeed, differences in cell growth rates correlate with the proportion of cells in each phase of the cell cycle: less proliferative cells have longer G1 or G2 phases than more proliferative cells. Consequently, slower-growing cell cultures will have more cells in G1 and G2 phase and fewer in S and M phase [2], and S and M phase-specific proteins will thus be less abundant in the lysates.

Genes highly expressed in proliferative cells have been used as proliferation markers by pathologists and researchers for many years [3–5]. Their expression indeed often correlates with the proportion of cells in S and M phase in a given sample and can strongly correlate with tumor progression and prognosis [6]. Nevertheless, there is to our knowledge no study that determines which proteins confound hypothesis-free high-throughput data analysis by correlating with cell proliferation, and the overall impact of cell growth rate on the transcriptome and the proteome remains to be determined.

In this work, we first define a pseudo-proliferation index based on transcriptomics and proteomics data for cells with known proliferation rate. We use this to analyze even larger datasets to identify a list of genes that correlate with cell proliferation at both transcript and protein level. These genes constitute a cell proliferation signature that is a valuable resource to identify and analyze datasets where proliferation is affected. We illustrate this in the context of proteomics cancer classification and drug screens [6,7], where identifying these signature genes allows to quickly discard less relevant changes that may be explained by change in cell proliferation and focus on genes that are regulated in a more context-specific manner.

Results and discussion

Pseudo-proliferation index derived from transcriptomics and proteomics data

Cell doubling times, or growth rates (*growth rate = ln(2)/doubling time*), are rarely provided alongside proteomics and transcriptomics data, so calculating correlation between gene relative quantities and cell growth rates is only possible for a limited number of publicly available data sets. For this reason, we defined a list of proliferation markers for which relative abundances reflect cell proliferation at protein and transcript level that would then be used to calculate an index for relative cell proliferation in datasets with no growth rates reported. The NCI60 cell lines [8] have been extensively characterized with high-throughput proteomics [9–

12] and transcriptomics [13–15] and their doubling times are publicly available from the Developmental Therapeutics Program (DTP) website (dtp.cancer.gov/discovery_development/nci-60/cell_list.htm; update of the 05/08/15). Gholami *et al.* [10] and Guo *et al.* [11] obtained pellets from DTP and lysed them directly, while Frejno *et al.* [9] obtained the cell lines from DTP and followed the DTP recommendations for *in vitro* growth. We used these data sets to identify proliferation markers that would reproducibly correlate with cell growth rates in proteomes.

We calculated the Pearson correlation with growth rates for each of the 3,645 protein groups quantified in at least two of the four NCI60 proteome data sets. Among these, we found nine human proteins that were reported as proliferation markers in the literature [3,16]. Most of these are transcribed at specific phases of the cell cycle [17] (green line in Fig 1a, and colored in S1 Fig). Although not referenced as cycling in Cyclebase v3.0, MCM3, MCM7 and MYBL2 have been shown to be expressed in a cell cycle-dependent fashion in single-cell transcriptomics [18] where MCM3/7 and MYBL2 expression peaks in G1 and G2, respectively. In the same study, CCND1 is found cyclic at protein but not transcript level, peaking in G1.

Fig 1a shows that the expression of most of these proliferation markers correlate strongly with NCI60 cell growth rates. We hypothesized that other cycling genes could be good markers of cell proliferation, and that increasing the number of genes used to estimate cell proliferation would be more robust to missing values and quantification uncertainties. Among the genes known to cycle at transcript level according to [17,19], eighteen were quantified in minimum two of the NCI60 proteomics data sets (colored points in S1a Fig). These proteins form complexes with other subunits that were not identified as cycling at RNA level but could correlate with cell growth rates; examples include the DNA polymerases A complex known to bind the cycling primases PRIM1 and PRIM2, or members of the replication factor C (RFC5 was not detected in [19] so its cycling status is unknown) (empty circles in S1a Fig).

Since we wanted to estimate relative cell proliferation in transcriptomics as well as in proteomics data, we also analyzed two transcriptomics data sets of the NCI60 cell lines grown according to the DTP recommandations [14,15] (S1b Fig). Fig 1b shows the correlation of the selected genes with cell growth rates in the transcriptome (horizontal axis) and the proteome (vertical axis). The periodic genes with the strongest correlation peak in G1/S and S phase at transcript and protein level, respectively [17]. From these data, we defined a set of potential proliferation markers containing the genes presenting high correlation with cell growth rates both in the transcriptomics and proteomics data sets (Fig 1b, grey area in the top-right corner). We compared pseudo-proliferation indexes calculated as the mean signal of:

- proliferation markers referenced in the literature (PCNA, MCM2–7, PLK1 and MKI67).
- proliferation markers referenced in the literature and genes known to cycle at transcript level (FEN1, RRM1, RRM2, CDK1, RPA2, RFC4, RFC2, PRIM2).
- all the above plus the known interacting non-cycling subunits POLA1, RFC3, RPA1, RPA3, RFC5, SMC3, STAG2, SMC1A.

We compared how the resulting pseudo-proliferation indexes correlated with cell growth rates in the proteomics NCI60 data sets (Fig 1c). As expected, the more genes were included in the proliferation markers list, the stronger the correlation. Based on these results, we decided to include the proliferation markers, periodic genes, and subunits of cycling complexes to calculate pseudo-proliferation index (all proteins in the top-right corner of Fig 1b). We performed the same comparison using the median instead of the mean of relative signals of proliferation markers. This led to lower Pearson correlations with cell growth rates and more variability between proteomics data sets (S2 Fig).

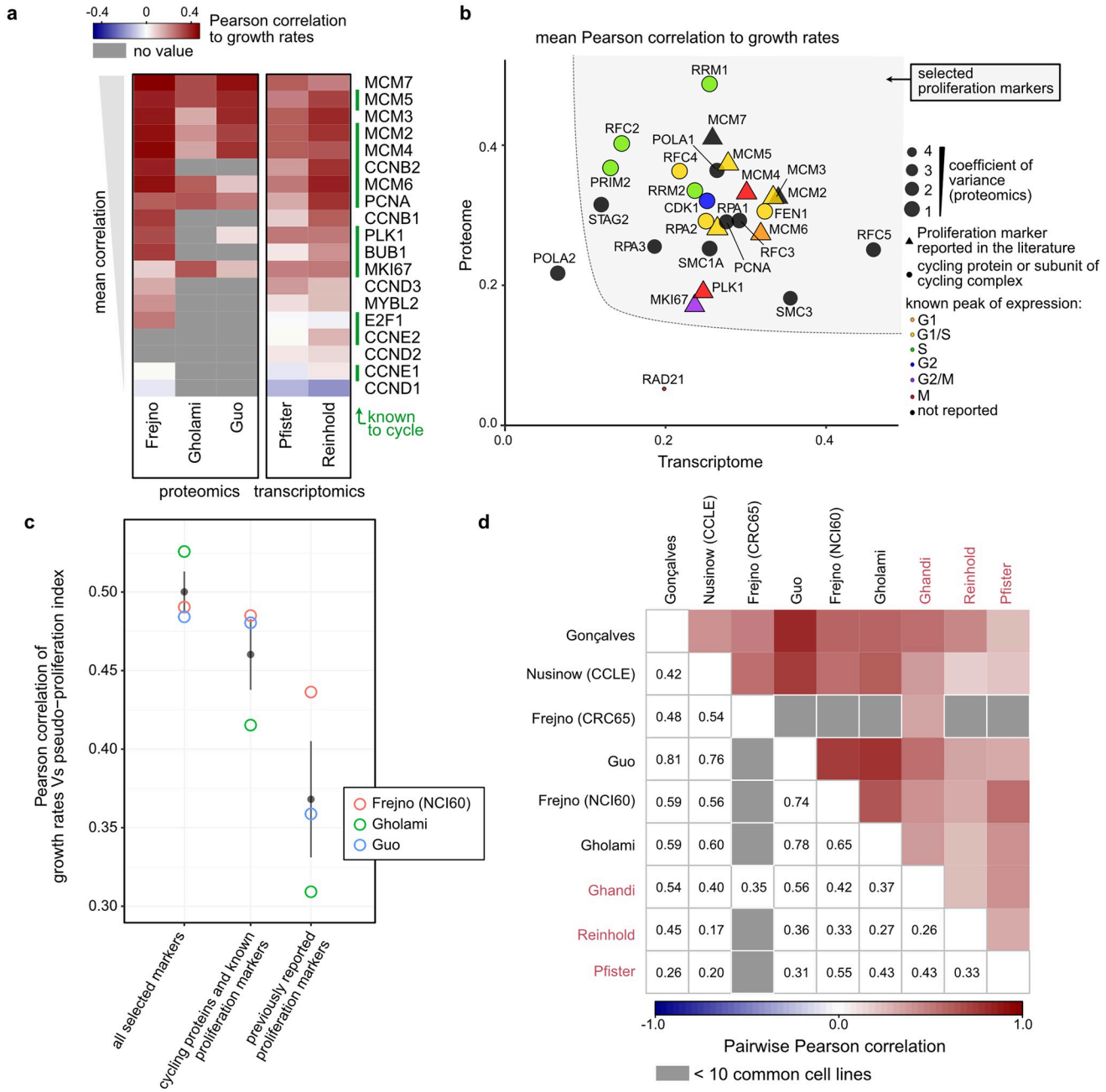

Fig 1. Calculation of pseudo-proliferation index. a) Pearson correlation with growth rates of the NCI60 cell lines that follow the Developmental Therapeutics Program (DTP)'s growing instructions for proliferation markers referenced in the literature. The proteins that cycle in Cyclebase 3.0 are indicated by green bars on the right ("known to cycle"), and the data set names are indicated in the bottom. **b)** Set of known markers, cycling genes and complex-associated subunits considered as proliferation marker for pseudo-proliferation index calculation. Mean Pearson correlation with growth rates in the datasets presented in (a) in the proteome (vertical axis) and transcriptome (horizontal axis). The point size is proportional to the inverse of the coefficient of variance in the proteomics data, proteins present in less than 2 data sets were excluded, as well as the cycling CDC27 due to its negative correlation with growth rates in the proteomics data sets. Periodic genes are color-coded by the phase of their expression peak, proliferation markers reported in the literature are indicated by triangles. The selected proliferation markers are indicated by the grey area. **c)** Pearson correlations between pseudo-proliferation index and growth rates in the proteomics data sets presented in (a) using the mean signal of the proliferation markers as selected in (b) (grey area), all the previously reported proliferation markers, or the previously reported proliferation markers and cycling genes with the exclusion of RAD21. Grey points and bars are mean and confidence intervals across data sets. **d)** Pairwise Pearson correlation between the pseudo-proliferation indexes calculated in the different data sets (proteomics and transcriptomics in black and red, respectively) Pairwise comparisons with less than 10 cell lines were excluded (in grey).

https://doi.org/10.1371/journal.pcbi.1010604.g001

This data-driven approach was used to estimate relative cell proliferation on proteomics data sets with no growth rates reported: the proteomes of the CRC65 cancer cell lines [9]; the Cancer Cell Line Encyclopedia (CCLE) that comprises the CRC65, NCI60 and other cell lines [12,20]; and the recently published Pan-Cancer panel [21] (S3 Fig shows the cell lines present in each panel). For each data set, we first calculated the pseudo-proliferation indexes, and next the correlations of each protein to this proxy for cell proliferation. Gene set enrichment analysis (GSEA) showed that proteins involved in chromatin remodeling, DNA replication and chromosome organization were highly correlated to pseudo-proliferation index (S4 Fig). These results were similar to those of a GSEA performed on proteins ranked by their correlation to growth rates when available ("NCI60 only"), which confirmed that pseudo-proliferation index reflects the proliferative state of cells and can be used as an estimation of relative cell growth rates. We further controlled that the same Gene Ontology (GO) terms were reproducibly enriched across larger data sets of non-NCI60 cell lines. With the same approach, we calculated pseudo-proliferation index at RNA level in data sets containing the NCI60 and CCLE cells transcriptome [13–15] (S2 Table). Pseudo-proliferation indexes were highly consistent across proteomes (0.42 to 0.81) as well as between proteomes and transcriptomes (Fig 1d), although the growth rates of the same cell line can vary between data sets due to differences in experimental conditions and cell passages [22].

Identification of a proliferation gene signature

Using pseudo-proliferation index, we could identify which protein quantities correlated with cell proliferation rates in the six proteomics data sets presented above (S5a Fig). We filtered out the proteins that were detected in less than two data sets and calculated the mean of Pearson correlations to pseudo-proliferation index across data sets.

We benchmarked our approach with three sets of genes expected to be highly expressed in proliferative cells (*i.e.* gold standards) either because they are known to be expressed in a cell-cycle-dependent fashion or because they were reported to be expressed under the control of a transcription factor only active on S-phase entry:

- B1: 48 genes known to be periodically expressed in synchronized cell cultures [23].
- B2: 382 genes compiled from two lists of proposed E2F transcription factor targets [19,24,25].
- Cyclebase 3.0: 570 periodically-expressed genes (https://cyclebase.org/) [17].

We ranked the proteins (excluding the proliferation markers used to calculate pseudo-proliferation index in the first place) by decreasing absolute mean of correlation to pseudo-proliferation index and counted the number of proteins belonging to each of the three gold standard sets (Fig 2a). As expected, these gold standards were enriched for the proteins most strongly correlated with cell growth rates (left of the horizontal axis). We determined a cutoff for correlation with pseudo-proliferation index: ≥ 0.344 (Fig 2a). The exact same strategy was applied with three transcriptomic data sets to determine a transcriptomics confidence threshold of ≥ 0.567 (Fig 2b). In both analyses, we calculated gene correlations with randomized pseudo-proliferation index (50 iterations) to check that all the gene signatures had a FDR under 0.1% (see material and methods).

Fig 2c shows gene correlations to pseudo-proliferation index at transcript and protein level. Overall, transcripts presented a higher mean correlation with pseudo-proliferation index than the proteins these were translated to, and the distribution of Pearson correlations to pseudo-proliferation index was wider at transcript than protein level. This indicates post-translational

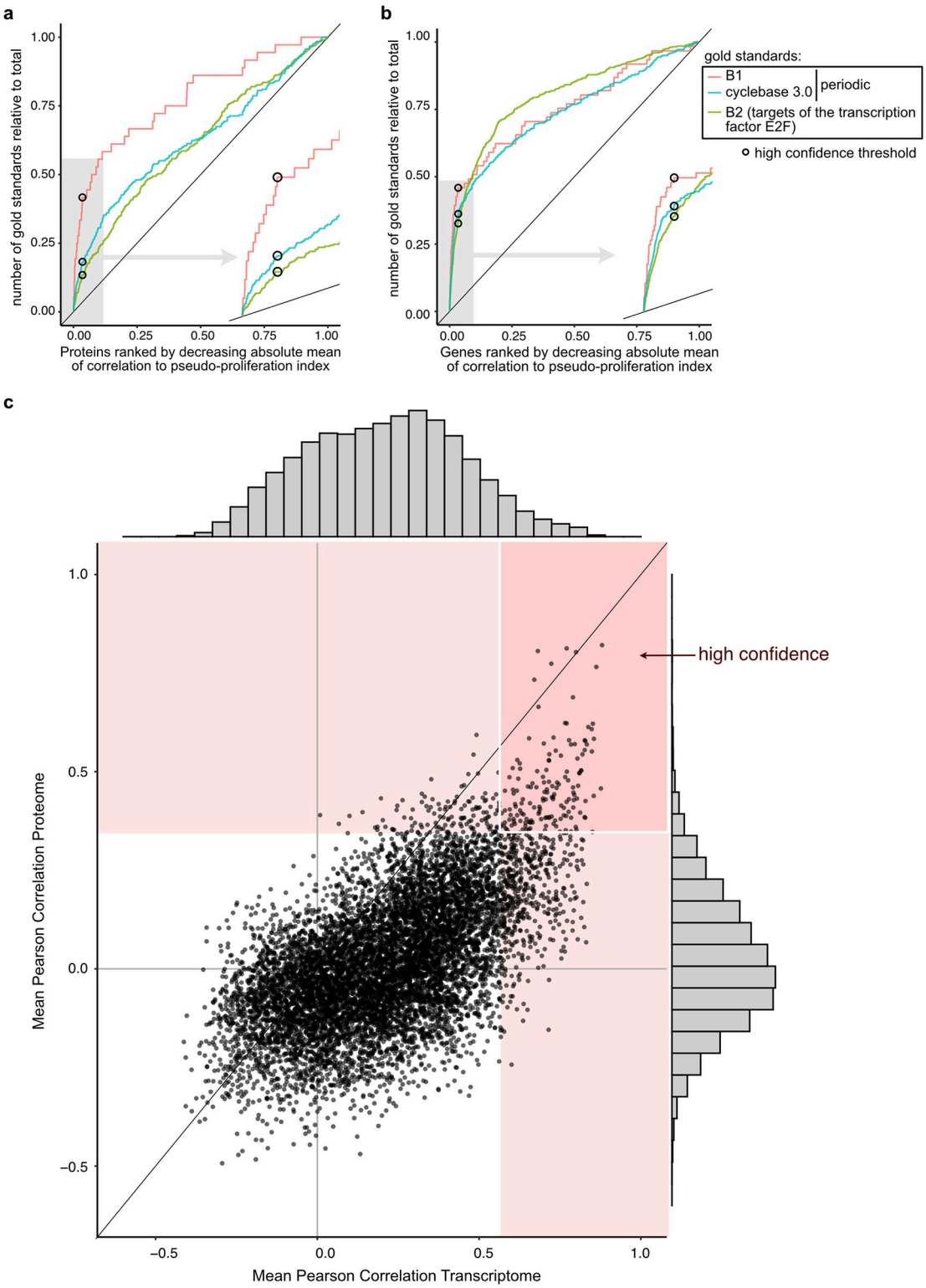

Fig 2. Signature genes of cell proliferation. a-b) Definition of the cutoff for correlation with pseudo-proliferation index with three sets of gold standards in the proteomes (a) and the transcriptomes (b). Proteins/genes were ranked by decreasing absolute Pearson correlation to pseudo-proliferation index (horizontal axis) and the vertical axis presents the cumulative number of gold standards for each set. Proteins/genes quantified in less than 3 and 2 data sets were excluded in (a) and (b), respectively. **c)** Scatter plot of the mean Pearson correlation to pseudo-proliferation index at protein (vertical axis) and transcript (horizontal axis) level

across all data sets. The red areas contain the proteins above the threshold in the proteome and/or transcriptome and the rectangle with white borders indicates the final list of proliferation signature genes defined in this study. The point distribution in the proteomes and transcriptomes are presented on the sides of the plot.

https://doi.org/10.1371/journal.pcbi.1010604.g002

adjustment of protein quantities and/or post-transcriptional regulatory processes. Gene correlations to pseudo-proliferation index at protein and transcript level are available in S3 Table. We defined a threshold for a signature of cell proliferation constituted of 157 genes that correlate with pseudo-proliferation index at transcript as well as protein level.

Fig 3 shows the physical interactions between the proliferation signature genes according to the STRING physical interaction subnetwork. Each node (gene) is colored with its Pearson correlation with pseudo-proliferation index in each data set (ring). These are involved in DNA replication and mitosis. As expected, we find back all the genes used for calculating the pseudo-proliferation index (circled in black) except STAG2 and PLK1, which were just under

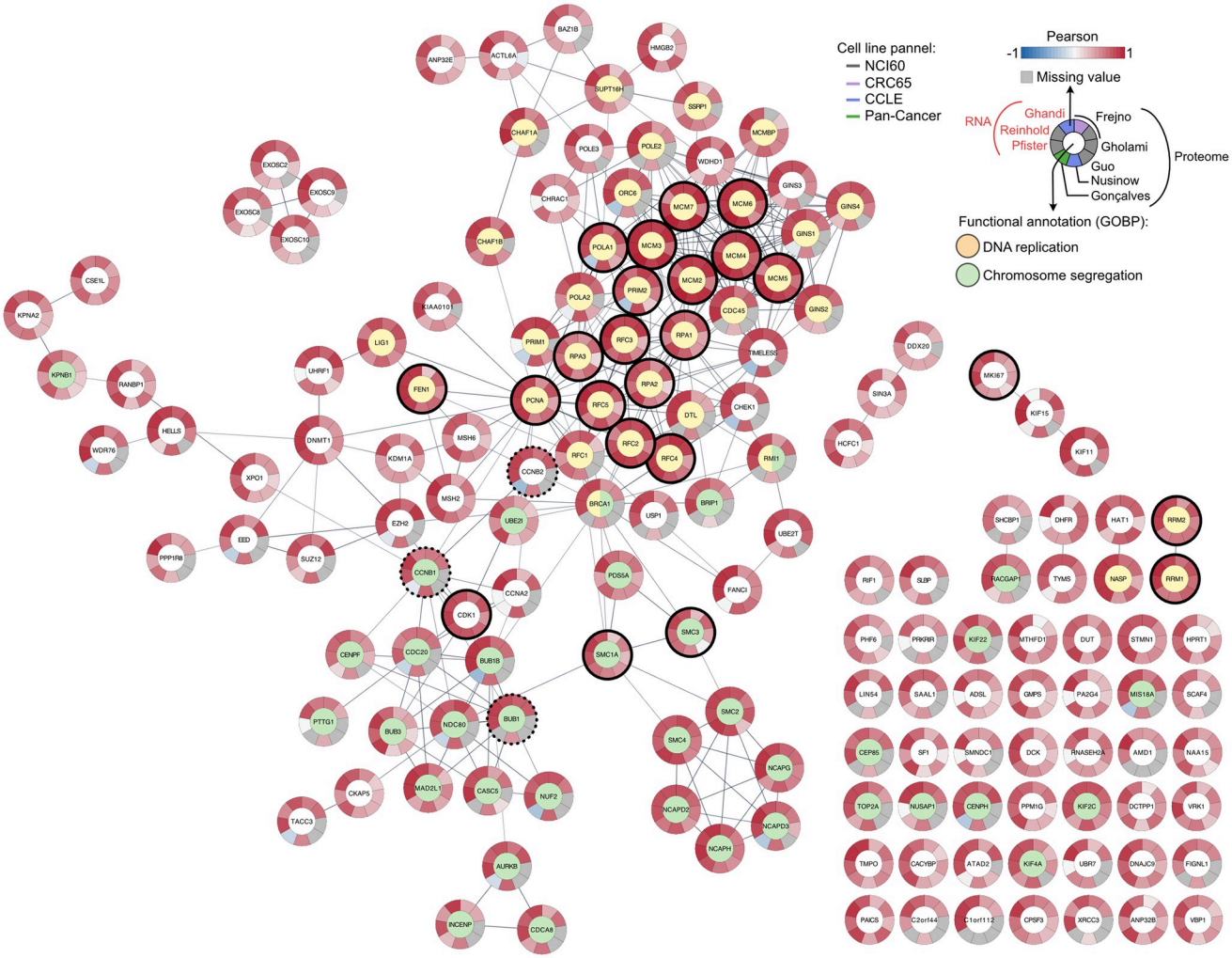

Fig 3. Proliferation signature. STRING subnetwork of physical interactions (score ≥ 0.7) corresponding to the proliferation signature as selected in Fig 2. The genes used to calculate the pseudo-proliferation index and the known proliferation markers not included for pseudo-proliferation index calculation are highlighted by black solid and dashed borders, respectively. The nodes are color-coded by selected gene annotations of biological processes. External ring are the Pearson correlations for each data set independently.

https://doi.org/10.1371/journal.pcbi.1010604.g003

the threshold in the proteomics data (and the transcriptomics data for STAG2). Some of the proliferation markers previously described in the literature were also found in the proliferation signature, such as BUB1 and CCNB1/2 (dashed black borders in Fig 3). These genes were not included in the refined list of proliferation markers used for calculating pseudo-proliferation index because they did not consistently correlate with NCI60 growth rates, but they strongly correlate with relative cell proliferation when integrating more cell lines and more data sets. Although the selected set of proliferation markers used to calculate pseudo-proliferation index mainly consists of genes involved in DNA repair, many genes of the proliferation signature are involved in other parts of the cell division cycle. Fig 3 shows that many of them are involved in M-phase processes such as chromosome segregation. This indicates that our strategy for selecting signature genes was not biased towards S-phase functions but retrieved genes for which expression corelated with cell proliferation for reasons that are yet to determine.

Use case 1: Proliferation signature in drug screens

Many drugs affect cell proliferation, thereby decreasing the proportion of cells actively dividing in samples. This can confuse data analysis when investigating drug mode of action because many of the genes regulated upon treatment are in fact correlated with cell proliferation. A recently published paper provides the proteomes of five cell lines after 53 drug treatments [7]. In many experiments, the proliferation signature was enriched for the proteins that were downregulated after treatment, suggesting that the drug treatments reduced cell proliferation rates.

After brefeldin A [26] treatment, 10% of the downregulated proteins (q-value ≤ 0.05) were proliferation signature genes (p-value $< 10^{-15}$; Fig 4a). Brefeldin A disassembles the Golgi complex and induces endoplasmic reticulum (ER) stress. It is usually used as potent inhibitor of cell secretion. Consequently, Brefeldin A treatment reduces cell proliferation, which is very visible when labelling signature genes in the volcano plot Fig 4b (orange dots): most of them are shifted towards the left of the volcano. Labelling them facilitates data analysis by: 1) highlighting global fold-change shifts that can be due to proliferation increase or decrease as a consequence of drug treatment and 2) disregarding protein regulations due to proliferation changes if these are not the main focus of the experiment to concentrate on more direct consequences of drug treatment.

Docetaxel treatment impacts cell proliferation specifically in A549 cells (lung carcinoma epithelial cells) where 47% of the proteins significantly downregulated (q-value ≤ 0.05) were signature genes (p-value $< 10^{-15}$, Fig 4c). The volcano plot corresponding to this experiment is presented Fig 4d. Docetaxel is a taxane that interferes with microtubule growth by binding to the β-subunit of tubulin. It is used in the treatment of many cancers. Fig 4e shows the STRING network of functional associations of the proteins significantly downregulated in Fig 4d (grey box). Most of these genes are functionally connected in a "hairball" that contains all but one signature gene. Some of these hits are involved in microtubule remodeling, but others are downregulated because of a reduction of cell proliferation of the A549 cells upon treatment. Examples of the latter include RRM2, which catalyzes the biosynthesis of deoxyribonucleotides, and the chromatin-assembly factors CHAF1A and CHAF1B. Labeling the proliferation signature facilitates the identification of proteins potentially more relevant to the drug treatment (grey nodes outside of the hairball). For example, the Microtubule-associated tumor suppressor 1 (ATIP3, coded by the gene *MTUS1*). *MTUS1*-deficiency is associated with increased microtubule dynamics [27], which is the opposite of docetaxel-induced microtubule stabilization. In breast cancer, ATIP3 was found significantly downregulated in taxane-sensitive tumors [28]. It is an interesting therapeutic target for breast cancer [29]. Caspase 2 (*CASP2*)

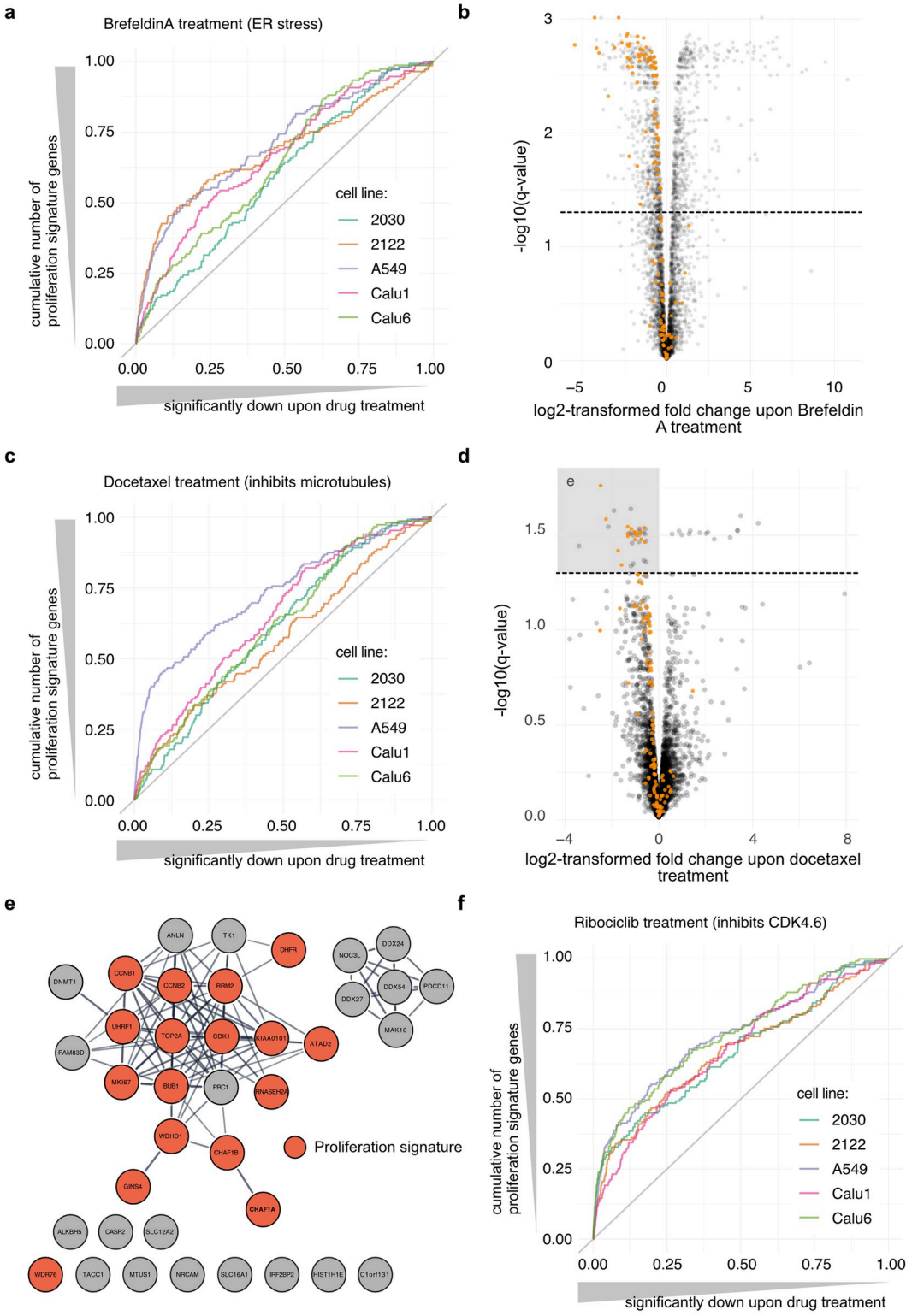

Fig 4. Proliferation signature in the context of drug treatment. a) Enrichment of the proliferation signature in the proteomes of cells treated with Brefeldin A. Proteins were ranked by significance of down-regulation according to Ruprecht *et al.* [7] (*q*-value) (horizontal axis) and the vertical axis presents the cumulative number of signature genes for each cell line. "2030" and "2122" correspond to the NCIH-2030 and NCIH-2122 cell lines, respectively. **b)** Volcano plot for A549 cells treated with Brefeldin A. Proliferation signature genes are highlighted in orange. The dashed line corresponds to a *q*-value of

0.05. **c)** Enrichment of the proliferation signature in the proteomes of cells treated with Docetaxel as in (a). **d)** Volcano plot for A549 cells treated with Docetaxel as presented in (b). **e)** Significantly down-regulated proteins (grey square in (d)) are presented in a STRING network of functional associations (score $\geq$ 0.7). Proliferation signature genes are highlighted in orange. **f)** Enrichment of the proliferation signature in the proteomes of cells treated with Ribociclib as in (a).

https://doi.org/10.1371/journal.pcbi.1010604.g004

has been shown to cleave the Microtubule-associated protein tau (coded by the gene *MAPT*) that promotes microtubule assembly and stability and potentially competes with taxanes for microtubule binding. It is associated with resistance to taxanes in several cancers [30–32]. The Transforming acidic coiled-coil-containing protein 1 (TACC1) is also involved in microtubule regulation [33]. The Nucleolar complex protein 3 homolog (*NOC3L*), protein MAK16 homolog, RRP5 homolog (*PDCD11*) and the ATP-dependent RNA helicases DDX24/27/54 are RNA-binding proteins. Although there is no obvious known association of these proteins with docetaxel treatment and/or microtubule regulation, these downregulated proteins may inform on docetaxel impact on A549 cells.

In other cases, such as ribociclib treatment the same genes are not to be set aside but reflect the drug mode of action. Ribociclib inhibits CDK4/6 activity and thereby prevents progression through the G1/S checkpoint, blocking cells in G1 phase. This results in a high enrichment of the proliferation signature in negatively regulated genes (Fig 4f), which is highly relevant for data interpretation.

Use case 2: Proliferation signature in the context of cancer prognostic and classification

The proliferation signature can also be useful for analysis of *in vivo* samples and patient data, for example in the context of cancer since most tumors are characterized by an increased proliferation rate. Many of the signature genes reported here are indeed reported prognostic markers in the context of cancer. This can be highly relevant since these genes may be significantly regulated because of the presence of more dividing cells in certain tumor samples. Other genes/proteins may be more appropriate for targeted therapy.

The recently published meta-analysis of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [6] identified proteins which relative quantities are correlated with tumor grade or stage in patient samples. The proliferation signature genes identified in this study were not enriched in proteins associated with tumor stage (Fig 5a). Proteins strongly correlating with tumor grade, however, were enriched in the proliferation signature in lung adenocarcinoma (LUAD), uterine endometrial carcinoma (UCEC), and pediatric glioma, but not in clear cell renal cell carcinoma (CCRCC) and ovarian serous adenocarcinoma (OV) (Fig 5b). This is in agreement with the GO-term enrichment presented in Monsivais *et al.* [6], where "cell cycle process" and "DNA replication" are strongly enriched in the proteins the most associated with cancer grades in LUAD, glioma and UCEC.

In the lung adenocarcinoma and pediatric glioma data, the proteins the most associated with cancer grade include a high number of signature genes of cell proliferation that may not be the best candidates for targeted treatment. Fig 5c shows the proteins correlation with grade (vertical axis), with signature genes highlighted in orange. With such figure, it is possible to quickly identify proteins that are specifically correlated with high tumor grade but not associated with the high proliferative state of aggressive lung tumors.

In lung adenocarcinoma, the three proteins the most correlated to cancer grade belonged to the proliferation signature: the U3 ubiquitin-protein ligase UHRF1, Kinesin-like protein KIF11 and the well-known proliferation marker MKI67 FHA domain-interacting nucleolar phosphoprotein. The Anillin actin binding protein (*ANLN*) was the top hit amongst non-

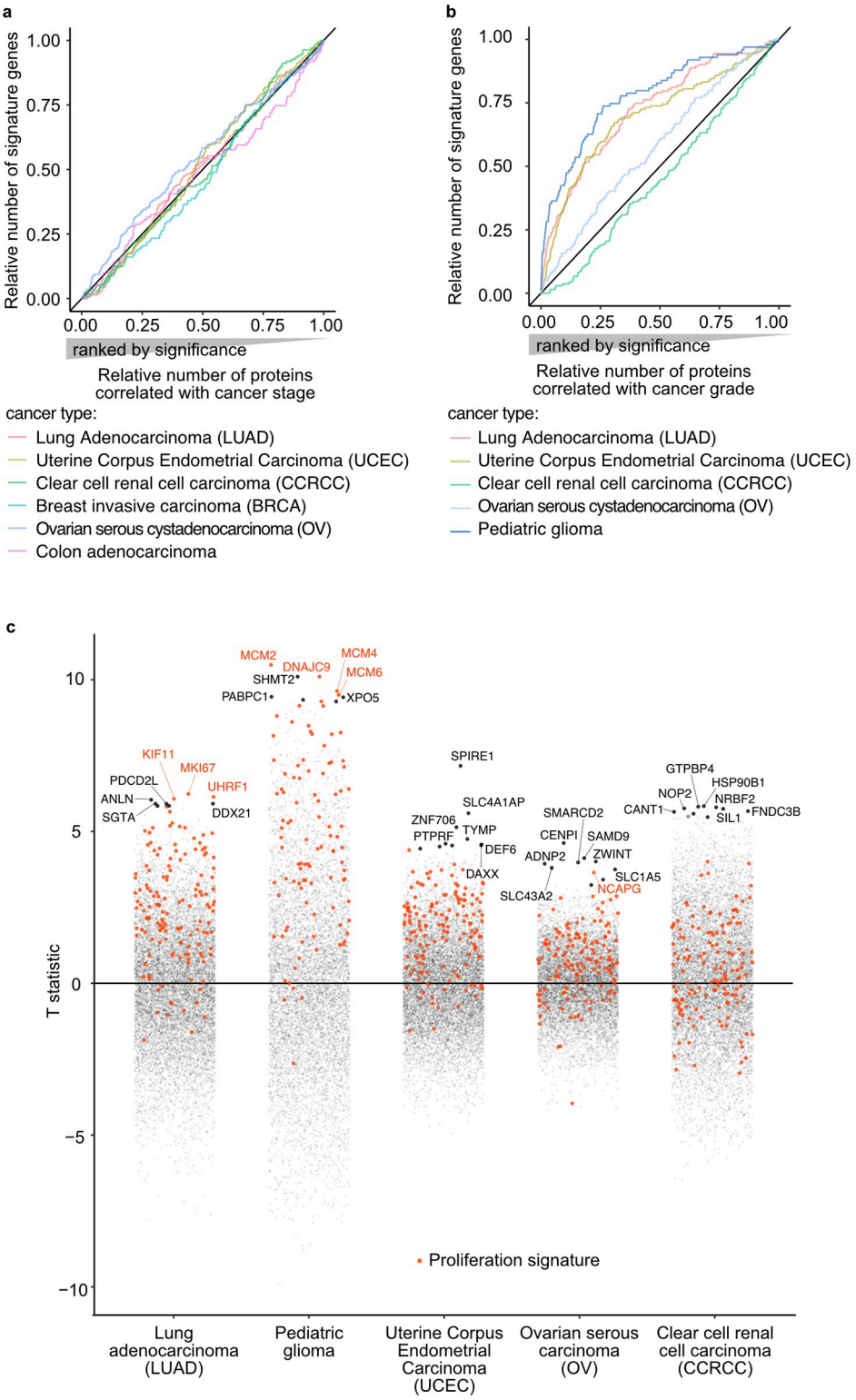

Fig 5. Proliferation signature in the context of cancer grade. a-b) Enrichment of the proliferation signature in proteins associated with cancer stage (a) and grade (b). Proteins were ranked by significance of correlation according to Monsivais *et al.* [6] (*p*-value of Pearson correlation) (horizontal axis) and the vertical axis presents the cumulative number of signature genes for each cancer type. **d)** Proteins T-statistic provided by Monsivais *et al.* [6] for the analysis of cancer grades (positive = high correlation with cancer grade) for each cancer type (horizontal axis). Each point

corresponds to a protein, signature genes are highlighted in orange. The seven top hits for each cancer type are indicated by their gene names.

https://doi.org/10.1371/journal.pcbi.1010604.g005

signature genes. It is highly expressed in lung cancer cell lines and tumor samples compared to normal tissues [34], and *ANLN* high expression is a predictive marker of poor outcome for patients with lung adenocarcinoma in TCGA [35]. Anillin activates cellular migration of lung cancer cells *in* vitro [34] and increases tumor growth and metastasis in breast cancer through induction of mesenchymal to epithelial trans-differentiation [36].

76% of the signature genes identified in this study were correlated to cancer grade with a *p*-value under 0.01 in gliomas (Fig 5c). In such context, it is particularly important to acknowledge that these genes may be regulated because of differences in cell growth rates. The signature genes MCM2/4/6, and the heat shock co-chaperone and histone chaperone DNAJC9 [37] were amongst the five proteins the most correlated with glioma grades. In the figure, these surround the mitochondrial serine hydroxymethyltransferase (SHMT2) (ranked 3rd), which could be a more interesting hit for targeted therapy. It participates to the synthesis of glycine by catalyzing serine-to-glycine conversion. Glycine is a key resource for proliferative cells, and elevated concentration of glycine in IDH-mutated glioma tumors has been associated with aggressive glioma and is predictive of shortened patient survival [38]. While SHMT2 expression is not directly correlated with glycine concentration in gliomas [38,39] it has been shown to favor cancer cells adaptation to poorly vascularized tumor micro-environments in the context of ischemic glioma [39], and to be associated with poor prognostic in glioma [40].

Highlighting the proliferation signature in the analysis of LUAD and pediatric glioma allowed to quickly focus on proteins more directly associated with cancer grades in these contexts. This illustrates the advantage of taking these signature genes into consideration when analyzing proteomics data of patient samples.

Conclusion

Here, we calculated a pseudo-proliferation index that we used as proxy for relative cell proliferation at transcript and protein level to define high-confidence thresholds for identifying a set of genes correlated with cell proliferation rates. We combined the transcriptomics and proteomics analysis to provide a final list of genes constituting a proliferation signature. The S3 Table provides the correlations to pseudo-proliferation index for 10,600 genes/proteins quantified in the data sets that were used for this analysis. With this list of signature genes, anybody can identify in their data sets the genes/proteins correlated with cell proliferation like contaminants are routinely flagged using the CRAPome [41].

We showed examples of high-throughput data analysis where labelling the proliferation signature facilitates data interpretation. It informs on the potential impact of differences in cell proliferation in a given experimental set up–which is for example strongly down-regulated upon treatment with drugs blocking the cell cycle–and can inform on differences in cell proliferation in tissue samples such as tumors.

We also show how cell growth rate can be a confounding factor that results in down- or up-regulation of many genes in *in vitro* drug screens and tumor samples. Flagging these confounders among the most regulated genes allows to quickly identify other regulated hits that could be more relevant in the context of the experiment. Such analyses still require strong knowledge of the biological context and molecular regulation at play, but the genes correlated with cell proliferation rates are not all annotated as being involved in replication of cell-cycle-related processes. Thus, our refined list of proliferation signature genes is an invaluable

resource for interpreting data where changes in cell growth rates/proliferation is a confounding factor.

The strategy that we describe here to identify the proliferation gene signature is straightforward and can be applied to many other types of confounding factors. The only requirement, which can be very limiting, is the availability of several high-dimensional data sets on samples where the confounding factor of interest can be quantified. We believe that taking such gene signatures into consideration should become part of the high-throughput data analysis routine and will facilitate data interpretation in many biological contexts.

Materials and methods

Retrieval and pre-processing of proteomics data

The raw data from Gholami *et al.* [10] were retrieved from the PRIDE proteomeXchange repository PXD005946 and searched against the Human reviewed protein database (download 12/03/2021 from Uniprot.org) with MaxQuant v1.6.17.0. The mqpar.xml and the fasta file associated with the search are available on Zenodo.org (10.5281/zenodo.6346643). The proteinGroups.txt table was filtered to remove the reverse sequences and potential contaminants identified with the contaminant database included in MaxQuant. We kept only the protein groups with minimum one unique peptide and a q-value ≤ 0.01 (6,900 protein groups). We further removed the samples with more than 70% of missing values. LFQ was utilized for correlation calculation after variance stabilizing normalization (vsn) [42].

The data from Guo *et al.* [11] were retrieved from the S1E Table provided in the paper (3,171 protein groups with no missing value) and normalized using vsn before correlation calculation.

The normalized iBAQ quantification from Frejno *et al.* [9] was retrieved from the supplementary Data 3 available with the paper. The tables for Trypsin, GluC and Trypsin digestion of the CRC65 cells were filtered to remove the reverse sequences and potential contaminants identified with the contaminant database included in the original MaxQuant search. We kept only the protein groups with minimum one unique peptide and a q-value ≤ 0.01 (9,744 and 7,271 protein groups in the trypsin and GluC dataset for the NCI60 cells, respectively and 11,308 for the CRC65 digested with trypsin). We further removed the protein groups with more than 50% of missing values. We also removed the protein "PLIB" in the trypsin dataset due to bad annotation. For the analysis of the NCI60 cell lines, we took the protein groups mean signal from the trypsin and the GluC data sets.

The normalized TMT quantification from Nusinow *et al.* [12] was retrieved from the supplementary data available on https://gygi.hms.harvard.edu/publications/ccle.html ("Protein Quantitation (TSV Format)"). The tables were filtered to remove the protein groups with more than 50% of missing values.

The data from Gonçalves *et al.* [21] were retrieved from the S1 Table provided in the paper (6,692 protein groups with a minimum of 2 quantified peptides) and normalized using median subtraction before correlation calculation. We removed the protein groups quantified in less than 10 cell lines (6,451 protein groups remaining), and the four following cell lines from the analysis because their names were too similar and could create mismatch between the different data sets: "TT", "T-T", "KMH-2" and "KM-H2".

Proteome inter-data set matching

Since the searches were performed on each proteomics data set independently, the same protein can be labelled differently in the search outputs (*i.e.* belong to different protein groups, split across several isoforms. . .). We retrieved the protein groups corresponding to the same

protein in different data sets. We first combined variants/isoforms signal by keeping their mean values. Then, we matched and renamed them across data sets according to the mapping table that is provided as S1 Table. In the cases where several rows of a given data set were mapped to the same homogenized protein group ID, we kept the mean value per sample. If several accessions of a given data set corresponded to a unique accession in another data set, we favored the homogenized protein group ID with the highest number of matching protein groups across data sets. In cases of tie, we kept the one with the least "combined" accessions (several accessions corresponding to the homogenized accession in a given data set).

Proteomics proliferation signature

For each data set independently, we calculated the mean of signal of proteins of the MCM complex (MCM2, MCM7, MCM3, MCM4, MCM5 and MCM6), CDK1, PCNA, PLK1, RPA2, RRM1, RRM2, RFC4, RFC2, FEN1, MKI67, PRIM2, POLA1, RPA1, RPA3, RFC5, RFC3, SMC1A, SMC3, STAG2 to generate a pseudo-proliferation index. The gene names and corresponding Uniprot accessions are provided in S4 Table.

For each protein group quantified in a minimum of 10 cell lines, we calculated its Pearson correlation to cell lines pseudo-proliferation index and to the growth rates calculated based on doubling time (available on dtp.cancer.gov/discovery_development/nci-60/cell_list.htm—Last Updated: 05/08/15). Missing values were replaced in each data set with the 1% quantile. We excluded the protein groups only quantified in one data set and calculated the mean of Pearson correlations. The absolute mean of Pearson correlation to pseudo-proliferation index was utilized to rank the protein groups. We performed the same analysis after randomization of the cell lines' pseudo-proliferation index (50 iterations); the distribution of the resulting absolute mean of Pearson correlations across data sets allowed us to define FDR thresholds: 0.1% FDR was obtained for an absolute mean of correlation to pseudo-proliferation index ≥ 0.189 in the proteomics data. To define a confidence threshold, we benchmarked the list of proteins ranked by decreasing absolute Pearson correlation to pseudo-proliferation index with three gene lists of gold standards: B1 [23], B2 [19,24,25], and the periodic genes described in Cyclebase 3.0 [16]. Their cumulative count in the ranked list of proteins was utilized to select Pearson correlation values corresponding to high enrichment of gold standards.

Retrieval and processing of transcriptomics data

The processed data for the Affymetrix NCI60 dataset (Pfister *et al.*) [14] was obtained from the GEO NCBI portal using the GeoQuery R package (v.2.60.0) [43]. Probesets of the HGU133-Plus2 chip were mapped to Ensembl genes using the custom annotation provided by BrainArray [44]. The mapping file for probesets to Ensembl transcripts was obtained from the BrainArray version 25 download page (brainarray.mbni.med.umich.edu), and Ensembl transcript-gene mapping was retrieved using the R package biomaRt [45] v2.48.3. Probeset intensities were averaged across replicates of the same cell line. Only gene-specific probesets were considered, probesets mapping to multiple genes were excluded. When multiple probesets corresponded to the same gene, the one with the highest mean signal across all cell lines was selected to represent the gene. In total, 16,608 Ensembl genes (of which 15,541 having the biotype "protein coding genes") were uniquely mapped to probesets on the chip for the 59 NCI60 cell lines.

Raw fastq files corresponding to the NCI60 RNA-Seq profiling (Reinhold *et al.*) [15] were obtained from the European Nucleotide Archive (project accession PRJNA433861). The raw sequence reads were trimmed using Trimmomatic v038 [46], using the adapter file "TruSeq3-PE-2.fa", and with the following parameters: "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15

MINLEN:36". Transcript abundance estimates were then obtained using salmon v1.4.0 [47] in "quant" mode with the default parameters against the Human GRCh38 cDNA set obtained from the Ensembl release 103 [48]. Gene-level abundance estimates were summarized using the R package tximport v.1.20.0 [49], and upper quartile normalization was performed with the calcNormFactors function from the edgeR package v. 3.34.1 [50]. Finally, expression levels were obtained for 57,937 Ensembl genes (21,391 having the biotype "protein coding genes") for the same 59 cell lines profiled in Pfister *et al.*.

For the CCLE dataset (Ghandi *et al.*) [13], normalized gene expression levels in TPM (transcripts per million) units were obtained from the DepMap portal (depmap.org/portal, "CCLE_expression_full.csv"). We used the original Ensembl gene identifiers provided in the files: 51,832 Ensembl genes (19,790 protein coding) across 1,026 cell lines.

Transcriptomics proliferation signature

For each dataset independently, the pseudo-proliferation index was obtained as described for the proteomics datasets, by averaging the expression levels of the selected proliferation markers. For each gene, in each dataset we computed the correlations with the pseudo-proliferation index calculated for the dataset, as well as correlation with growth rates using the NCI60 cell lines doubling times when available. We selected the genes quantified in at least two datasets and calculated the mean of Pearson correlations. We performed the same analysis after randomization of the cell lines' pseudo-proliferation index (50 iterations) to define FDR thresholds: 0.1% FDR was obtained for an absolute mean of correlation to pseudo-proliferation index ≥ 0.251 in the transcriptomics data.

Mapping between Ensembl gene identifiers and UniprotKB swissport accessions has been performed using biomaRt [45]. For the integration of the RNA data with the proteome, we removed the genes from the transcriptome that matched to more than 6 protein groups in the proteome (2 genes) and reported the values of each gene from the transcriptome if they matched the same protein group (25 genes).

Gene set enrichment and GO term redundancy reduction

Gene set enrichments were performed with R v4.0.3 (R-project.org/) and RStudio v1.3.1093 (rstudio.com/) on a x86_64-apple-darwin17.0 (64-bit) running macOS Big Sur 10.16, using the packages clusterProfiler v3.18.1 [51] and org.Hs.eg.db v 3.12.0. The protein accessions were ordered by decreasing Pearson correlation with growth rates or proliferation index. We ran the function gseGO() with the following parameters: ont = "ALL", keyType = "UNIPROT", minGSSize = 6, maxGSSize = 800, pvalueCutoff = 0.05, verbose = TRUE, OrgDb = "org.Hs.eg.db", eps = 0, pAdjustMethod = "BH". The output summary was used to make the S4 Fig that presents GSEA on data sets with only NCI60 cell lines (first 2 panels) or without any NCI60 cell (last panel). We then simplified the output to reduce GO terms redundancy globally: we calculated the pairwise Jaccard indexes between all pairs of GO terms identified across data sets. Pairs of GO terms with a Jaccard index ≥ 0.5 were considered similar and only the one with the lowest enrichment q-value in any data set was kept for plotting. S4 Fig shows the 80 biological processes with the lowest absolute q-value (minimum value across all data sets and enrichments).

Functional annotations and networks

The two gene/protein networks presented in this paper were generated with Cytoscape v 3.9.1 [52]. GO term annotations were retrieved with the StringApp v 1.7.0 [53] and the donut visualization of Pearson correlations was performed with Omics Visualizer v 1.3.0 [54].

Proliferation signature in the context of drug treatment

The proteomics analyses of drug-treated cells were found in the supplementary Data 1 of Ruprecht *et al.* [7]. We counted the cumulative number of proteins subjected to statistical analysis by the authors and with a negative fold change upon drug treatment with Ribociclib (10,000 nM in all cell lines), Brefeldin A (100 nM, 30 nM, 100 nM, 100 nM, 30 nM for NHI-2030, NHI-2122, A549, Calu1 and Calu6, respectively) and Docetaxel (30 nM, 3 nM, 1 nM, 10 nM, 3 nM for NHI-2030, NHI-2122, A549, Calu1 and Calu6, respectively). Volcano plots were drawn with the data from Ruprecht *et al.* [7], proliferation signature genes were mapped to protein groups if minimum one of the proteins in the protein groups had a gene name corresponding to a signature gene.

Proliferation signature in the context of cancer tissue samples

The proteomics analyses of tumor tissues were found in Montsivais *et al.* [6] (Supplementary Data 2 and 3 for correlation with grade and stages, respectively). Proliferation signature genes were mapped to protein groups if minimum one of the proteins in the protein groups had a gene name corresponding to a signature gene.

Supporting information

S1 Fig. Selection of proliferation markers for calculating pseudo-proliferation index in proteomics and transcriptomics data. Volcano plots showing the mean correlation of proteins (a) or transcripts (b) to growth rates in the NCI60 data sets (horizontal axis) and the -log10(coefficient of variance) across all the data sets (vertical axis). Proteins quantified in less than 3 data sets were excluded in (a). Proteins/genes of interest are highlighted, and proliferation markers identified from literature search are indicated with triangles. These were color coded based of their expression peak according to Santos *et al.* [17].
(TIFF)

S2 Fig. Robustness of pseudo-proliferation index with different sets of proliferation markers. Pearson correlations between pseudo-proliferation index and growth rates in the proteomics data sets containing NCI60 cells presented in (Fig 1a) using the median (left panel) or mean (right panel) signal of the three sets of proliferation markers as selected in (Fig 1b) (grey area), all the previously reported proliferation markers, or the previously reported proliferation markers and cycling genes with the exclusion of RAD21. Grey points and bars are mean and confidence intervals across data sets. The right panel is the same as Fig 1c, it is reported here for direct comparison with the left panel.
(TIFF)

S3 Fig. Cell lines in the different proteomics data sets. Number of cell lines in the proteomics data sets used in the study. The total number of cell lines in the data sets are indicated in the left-hand side bar plot (color-coded by the cell line panel). The cell lines present in multiple data sets are indicated by the bar plot on the top: number of protein groups detected in the data sets indicated by a dot on the dot plot. Each data set is identified by the first author's name.
(TIFF)

S4 Fig. Comparison of the biological functions enriched in proteins strongly correlated with cell lines growth rates or pseudo-proliferation index. For each data set, genes (with the exception of the genes used for calculating pseudo-proliferation index) were ranked based on their correlation to growth rates or correlation to pseudo-proliferation index (left and right

panel, respectively). Gene set enrichments were performed using the "gseGO" function from the R package clusterProfiler v 3.18.1, resulting *p*-values are indicated in each tile, as well as color-coded normalized enrichment scores (NES). Only the annotations from biological processes are included, they are ordered by decreasing maximum NES per data set (top 80 enriched GO terms, see material and methods for a detailed description of the procedure used to reduce GO redundancy). Data sets are labeled based on the first author's name, enrichments were performed independently on the NCI60 cell lines or cell lines with no reported doubling time ("NCI60 only" and "NCI60 excluded", respectively).
(TIFF)

S5 Fig. Proteomics data and Brefeldin A treatment. a) Protein coverage of the proteomics data sets used in the study (after isoform removal and accessions homogenization—see material and methods). These are identified by the first author's name (left). The data set "Frejno" contained two independent MS searches of different cell line panels, we kept them separated. The total number of protein groups detected in the data sets are indicated in the bar plot on the right-hand side (color-coded by the cell line panel). The protein groups identified in multiple data sets are indicated by the bar plot on the bottom: number of protein groups detected in the data sets indicated by a dot on the dot plot. **b)** Volcano plots for each cell line treated with Brefeldin A. Genes constituting the proliferation signature are highlighted in orange. The dashed line corresponds to a *q*-value of 0.05.
(TIFF)

S1 Table. Accession mapping between the proteomics data sets. The proteomics data sets did not all group the proteins the same way (due to differences in peptide coverage). This table indicate which groups were mapped to which accession in this study.
(XLSX)

S2 Table. Pseudo-proliferation indexes calculated in each data set. Pseudo-proliferation indexes calculated for each proteomics and transcriptomics data set.
(XLSX)

S3 Table. Gene/protein correlations to pseudo-proliferation indexes. Gene and protein Pearson correlation to pseudo-proliferation indexes in each data set, mean across data sets and associated FDR. This table contains the final list of signature genes (TRUE values in the column "Signature gene").
(XLSX)

S4 Table. Genes/proteins used for calculating pseudo-proliferation index. List of genes and protein accessions selected for calculating the pseudo-proliferation index.
(XLSX)

Author Contributions

Conceptualization: Marie Locard-Paulet, Oana Palasca.

Formal analysis: Marie Locard-Paulet, Oana Palasca.

Funding acquisition: Lars Juhl Jensen.

Investigation: Marie Locard-Paulet, Oana Palasca.

Methodology: Marie Locard-Paulet, Oana Palasca, Lars Juhl Jensen.

Supervision: Lars Juhl Jensen.

Visualization: Marie Locard-Paulet.

Writing – original draft: Marie Locard-Paulet.

Writing – review & editing: Marie Locard-Paulet, Oana Palasca, Lars Juhl Jensen.

Reference

1. Polymenis M. Proteins associated with the doubling time of the NCI-60 cancer cell lines. Cell Division. 2017; 12(1). https://doi.org/10.1186/s13008-017-0032-y PMID: 28855958
2. Chao HX, Fakhreddin RI, Shimerov HK, Kedziora KM, Kumar RJ, Perez J, et al. Evidence that the human cell cycle is a series of uncoupled, memoryless phases. Molecular Systems Biology. 2019; 15 (3). https://doi.org/10.15252/msb.20188604 PMID: 30886052
3. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. Nature Reviews Cancer. 2006; 6(2):99–106. https://doi.org/10.1038/nrc1802 PMID: 16491069
4. Perou CM, Jeffrey SS, Van De Rijn M, Rees CA, Eisen MB, Ross DT, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proceedings of the National Academy of Sciences. 1999; 96(16):9212–7. https://doi.org/10.1073/pnas.96.16.9212 PMID: 10430922
5. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics. 2000; 24(3):227–35. https://doi.org/10.1038/73432 PMID: 10700174
6. Monsivais D, Vasquez YM, Chen F, Zhang Y, Chandrashekar DS, Faver JC, et al. Mass-spectrometry-based proteomic correlates of grade and stage reveal pathways and kinases associated with aggressive human cancers. Oncogene. 2021; 40(11):2081–95. https://doi.org/10.1038/s41388-021-01681-0 PMID: 33627787
7. Ruprecht B, Di Bernardo J, Wang Z, Mo X, Ursu O, Christopher M, et al. A mass spectrometry-based proteome map of drug action in lung cancer cell lines. Nature Chemical Biology. 2020; 16(10):1111–9. https://doi.org/10.1038/s41589-020-0572-3 PMID: 32690943
8. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer. 2006; 6(10):813–23. https://doi.org/10.1038/nrc1951 PMID: 16990858
9. Frejno M, Meng C, Ruprecht B, Oellerich T, Scheich S, Kleigrewe K, et al. Proteome activity landscapes of tumor cell lines determine drug responses. Nature Communications. 2020; 11(1). https://doi.org/10.1038/s41467-020-17336-9 PMID: 32686665
10. Gholami AM, Hahne H, Wu Z, Florian, Meng C, Wilhelm M, et al. Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Reports. 2013; 4(3):609–20. https://doi.org/10.1016/j.celrep.2013.07.018 PMID: 23933261
11. Guo T, Luna A, Rajapakse VN, Koh CC, Wu Z, Liu W, et al. Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. iScience. 2019; 21:664–80. Epub 2019/11/17. https://doi.org/10.1016/j.isci.2019.10.059 PMID: 31733513.
12. Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER, Kalocsay M, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell. 2020; 180(2):387–402.e16. https://doi.org/10.1016/j.cell.2019.12.023 PMID: 31978347
13. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019; 569(7757):503–8. https://doi.org/10.1038/s41586-019-1186-3 PMID: 31068700
14. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, et al. Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Molecular Cancer Therapeutics. 2009; 8(7):1878–84. https://doi.org/10.1158/1535-7163.MCT-09-0016 PMID: 19584232
15. Reinhold WC, Varma S, Sunshine M, Elloumi F, Ofori-Atta K, Lee S, et al. RNA Sequencing of the NCI-60: Integration into CellMiner and CellMiner CDB. Cancer Research. 2019; 79(13):3514–24. https://doi.org/10.1158/0008-5472.CAN-18-2047 PMID: 31113817
16. Juríková M, Danihel Ľ, Polák Š, Varga I. Ki67, PCNA, and MCM proteins: Markers of proliferation in the diagnosis of breast cancer. Acta Histochemica. 2016; 118(5):544–52. https://doi.org/10.1016/j.acthis.2016.05.002 PMID: 27246286
17. Santos A, Wernersson R, Jensen LJ. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Research. 2015; 43(D1):D1140–D4. https://doi.org/10.1093/nar/gku1092 PMID: 25378319

18. Mahdessian D, Cesnik AJ, Gnann C, Danielsson F, Stenström L, Arif M, et al. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Nature. 2021; 590(7847):649–54. https://doi.org/10.1038/s41586-021-03232-9 PMID: 33627808

19. Jensen LJ, Jensen TS, De Lichtenberg U, Brunak S, Bork P. Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature. 2006; 443(7111):594–7. https://doi.org/10.1038/nature05186 PMID: 17006448

20. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483 (7391):603–7. https://doi.org/10.1038/nature11003 PMID: 22460905

21. Goncalves E, Poulos RC, Cai Z, Barthorpe S, Manda SS, Lucas N, et al. Pan-cancer proteomic map of 949 human cell lines. Cancer Cell. 2022. Epub 2022/07/16. https://doi.org/10.1016/j.ccell.2022.06.010 PMID: 35839778.

22. Dawson KB, Madoc-Jones H, Field EO. Variations in the Generation Times of a Strain of Rat Sarcoma Cells in Culture. Exp Cell Res. 1965; 38:75–84. Epub 1965/04/01. https://doi.org/10.1016/0014-4827(65)90429-5 PMID: 14281207.

23. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, et al. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. Molecular Biology of the Cell. 2002; 13(6):1977–2000. https://doi.org/10.1091/mbc.02-02-0030 PMID: 12058064

24. Balciunaite E, Spektor A, Lents NH, Cam H, Te Riele H, Scime A, et al. Pocket protein complexes are recruited to distinct targets in quiescent and proliferating cells. Mol Cell Biol. 2005; 25(18):8166–78. Epub 2005/09/02. https://doi.org/10.1128/MCB.25.18.8166-8178.2005 PMID: 16135806.

25. Cam H, Balciunaite E, Blais A, Spektor A, Scarpulla RC, Young R, et al. A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell. 2004; 16(3):399–411. Epub 2004/11/05. https://doi.org/10.1016/j.molcel.2004.09.037 PMID: 15525513.

26. Chardin P, McCormick F. Brefeldin A. Cell. 1999; 97(2):153–5. https://doi.org/10.1016/s0092-8674(00)80724-2

27. Molina A, Velot L, Ghouinem L, Abdelkarim M, Bouchet BP, Luissint A-C, et al. ATIP3, a Novel Prognostic Marker of Breast Cancer Patient Survival, Limits Cancer Cell Migration and Slows Metastatic Progression by Regulating Microtubule Dynamics. Cancer Research. 2013; 73(9):2905–15. https://doi.org/10.1158/0008-5472.CAN-12-3565 PMID: 23396587

28. Rodrigues-Ferreira S, Nehlig A, Moindjie H, Monchecourt C, Seiler C, Marangoni E, et al. Improving breast cancer sensitivity to paclitaxel by increasing aneuploidy. Proceedings of the National Academy of Sciences. 2019; 116(47):23691–7. https://doi.org/10.1073/pnas.1910824116 PMID: 31685623

29. Haykal MM, Rodrigues-Ferreira S, Nahmias C. Microtubule-Associated Protein ATIP3, an Emerging Target for Personalized Medicine in Breast Cancer. Cells. 2021; 10(5):1080. https://doi.org/10.3390/cells10051080 PMID: 34062782

30. Mimori K, Sadanaga N, Yoshikawa Y, Ishikawa K, Hashimoto M, Tanaka F, et al. Reduced tau expression in gastric cancer can identify candidates for successful Paclitaxel treatment. Br J Cancer. 2006; 94 (12):1894–7. Epub 2006/05/25. https://doi.org/10.1038/sj.bjc.6603182 PMID: 16721363.

31. Rouzier R, Rajan R, Wagner P, Hess KR, Gold DL, Stec J, et al. Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer. Proc Natl Acad Sci U S A. 2005; 102(23):8315–20. Epub 2005/05/26. https://doi.org/10.1073/pnas.0408974102 PMID: 15914550.

32. Wagner P, Wang B, Clark E, Lee H, Rouzier R, Pusztai L. Microtubule Associated Protein (MAP)-Tau: a novel mediator of paclitaxel sensitivity in vitro and in vivo. Cell Cycle. 2005; 4(9):1149–52. Epub 2005/08/17. https://doi.org/10.4161/cc.4.9.2038 PMID: 16103753.

33. Gergely F, Karlsson C, Still I, Cowell J, Kilmartin J, Raff JW. The TACC domain identifies a family of centrosomal proteins that can interact with microtubules. Proceedings of the National Academy of Sciences. 2000; 97(26):14352–7. https://doi.org/10.1073/pnas.97.26.14352 PMID: 11121038

34. Suzuki C, Daigo Y, Ishikawa N, Kato T, Hayama S, Ito T, et al. ANLN Plays a Critical Role in Human Lung Carcinogenesis through the Activation of RHOA and by Involvement in the Phosphoinositide 3-Kinase/AKT Pathway. Cancer Research. 2005; 65(24):11314–25. https://doi.org/10.1158/0008-5472.CAN-05-1507 PMID: 16357138

35. Long X, Zhou W, Wang Y, Liu S. Prognostic significance of ANLN in lung adenocarcinoma. Oncology Letters. 2018. https://doi.org/10.3892/ol.2018.8858 PMID: 30008873

36. Wang D, Naydenov NG, Dozmorov MG, Koblinski JE, Ivanov AI. Anillin regulates breast cancer cell migration, growth, and metastasis by non-canonical mechanisms involving control of cell stemness and differentiation. Breast Cancer Research. 2020; 22(1). https://doi.org/10.1186/s13058-019-1241-x PMID: 31910867

37. Hammond CM, Bao H, Hendriks IA, Carraro M, Garcia-Nieto A, Liu Y, et al. DNAJC9 integrates heat shock molecular chaperones into the histone chaperone network. Mol Cell. 2021; 81(12):2533–48 e9. Epub 2021/04/16. https://doi.org/10.1016/j.molcel.2021.03.041 PMID: 33857403.

38. Tiwari V, Daoud EV, Hatanpaa KJ, Gao A, Zhang S, An Z, et al. Glycine by MR spectroscopy is an imaging biomarker of glioma aggressiveness. Neuro-Oncology. 2020; 22(7):1018–29. https://doi.org/10.1093/neuonc/noaa034 PMID: 32055850

39. Kim D, Fiske BP, Birsoy K, Freinkman E, Kami K, Possemato RL, et al. SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance. Nature. 2015; 520(7547):363–7. https://doi.org/10.1038/nature14363 PMID: 25855294

40. Wang B, Wang W, Zhu Z, Zhang X, Tang F, Wang D, et al. Mitochondrial serine hydroxymethyltransferase 2 is a potential diagnostic and prognostic biomarker for human glioma. Clin Neurol Neurosurg. 2017; 154:28–33. Epub 2017/01/21. https://doi.org/10.1016/j.clineuro.2017.01.005 PMID: 28107674.

41. Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, et al. The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. Nature Methods. 2013; 10(8):730–6. https://doi.org/10.1038/nmeth.2557 PMID: 23921808

42. Huber W, Von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002; 18 (Suppl 1):S96–S104. https://doi.org/10.1093/bioinformatics/18.suppl_1.s96 PMID: 12169536

43. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007; 23(14):1846–7. Epub 2007/05/15. https://doi.org/10.1093/bioinformatics/btm254 PMID: 17496320.

44. Dai M. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research. 2005; 33(20):e175–e. https://doi.org/10.1093/nar/gni179 PMID: 16284200

45. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009; 4(8):1184–91. Epub 2009/07/21. https://doi.org/10.1038/nprot.2009.97 PMID: 19617889.

46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

47. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. 2017; 14(4):417–9. https://doi.org/10.1038/nmeth.4197 PMID: 28263959

48. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Research. 2021; 49(D1):D884–D91. https://doi.org/10.1093/nar/gkaa942 PMID: 33137190

49. oneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2016; 4:1521. https://doi.org/10.12688/f1000research.7563.2 PMID: 26925227

50. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 2016; 5:1438. https://doi.org/10.12688/f1000research.8987.2 PMID: 27508061

51. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16(5):284–7. Epub 2012/03/30. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463.

52. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003; 13 (11):2498–504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

53. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. Journal of Proteome Research. 2019; 18(2):623–32. https://doi.org/10.1021/acs.jproteome.8b00702 PMID: 30450911

54. Legeay M, Doncheva NT, Morris JH, Jensen LJ. Visualize omics data on networks with Omics Visualizer, a Cytoscape App. F1000Research. 2020; 9:157. https://doi.org/10.12688/f1000research.22280.2 PMID: 32399202