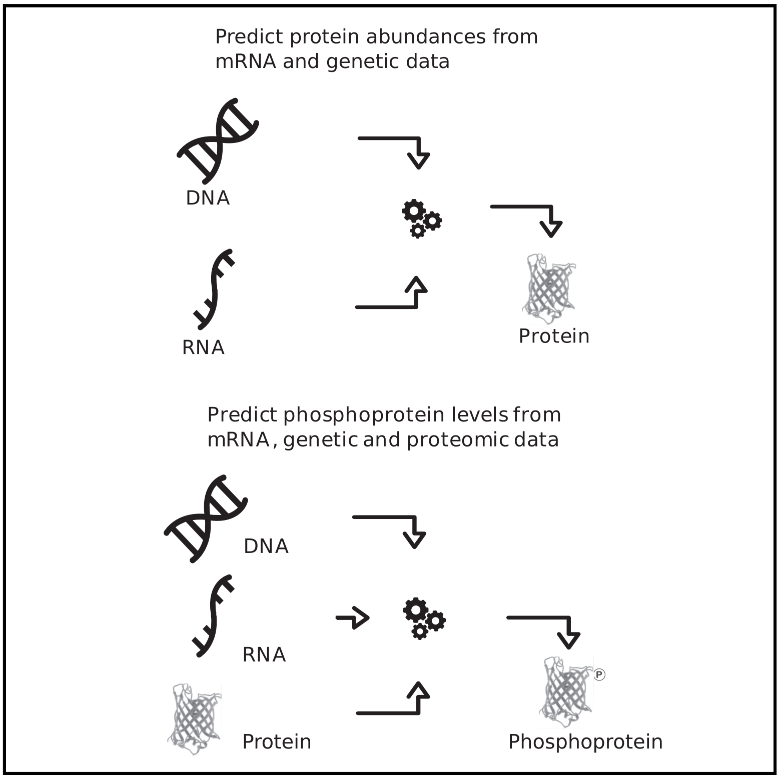
Cell Systems

Report

Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics

Graphical Abstract


Predict protein abundances from mRNA and genetic data
DNA
RNA
Protein
Predict phosphoprotein levels from mRNA, genetic and proteomic data
DNA
RNA
Protein
Phosphoprotein

Highlights

- Crowdsourcing of methods to predict (phospho)proteins from DNA and RNA
- Unbiased assessment uncovers best practices
- Proteins in metabolic pathways and complexes best and worst predicted, respectively

Authors

Mi Yang, Francesca Petralia, Zhi Li, ..., Pei Wang, David Fenyö, Julio Saez-Rodriguez

Correspondence

david@fenyolab.org (D.F.), julio.saez@bioquant.uni-heidelberg.de (J.S.-R.)

In Brief

A major manifestation of cancer is the alteration of protein measurements. However, proteins are harder and more expensive to measure than genes and transcripts. To address this problem, we crowdsourced it via the NCI-CPTAC DREAM proteogenomics challenge. We provided participants data to build models to predict protein and phosphorylation levels from genomic and transcriptomic data in cancer patients. We then asked participants to use such models to predict unseen (phospho)protein data from given genomic and transcriptomic data in other patients. This experiment allowed us to assess the predictive performance of the proposed methods in an unbiased and “double-blinded” manner. We found that ensemble methods perform better, and we identified which proteins and biological processes are easier or harder to predict. In general, performance was limited, suggesting that (phospho)proteomic cannot be replaced, at least yet, by genomic and transcriptomic profiling.

Yang et al., 2020, Cell Systems *11*, 186–195
August 26, 2020 © 2020 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.06.013

CellPress
OPEN ACCESS

Cell Systems

Report

Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics

Mi Yang,[1,2,28,29] Francesca Petralia,[3,28] Zhi Li,[4,5,28] Hongyang Li,[6] Weiping Ma,[3] Xiaoyu Song,[7,8] Sunkyu Kim,[9] Heewon Lee,[9] Han Yu,[10] Bora Lee,[11] Seohui Bae,[11,12] Eunji Heo,[11,13] Jan Kaczmarczyk,[14] Piotr Stępniak,[14] Michał Warchoł,[14] Thomas Yu,[15] Anna P. Calinawan,[3] Paul C. Boutros,[16,17,18,19,20,21,22] Samuel H. Payne,[23]

(Author list continued on next page)

[1]Faculty of Biosciences, Heidelberg University, 69120 Heidelberg, Germany
[2]Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Faculty of Medicine, 52074 Aachen, Germany
[3]Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[4]Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY 10016, USA
[5]Department of Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, New York, NY 10016, USA
[6]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[7]Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[8]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[9]Department of Computer Science and Engineering, Korea University, Seongbuk-gu, Seoul, Republic of Korea
[10]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA
[11]Deargen, Daejeon 34051, Republic of Korea
[12]Department of Biological Science, Department of Bio-Brain Engineering, KAIST, Daejeon, Republic of Korea
[13]Department of AI, KAIST, Daejeon 34141, Republic of Korea
[14]Ardigen, Kraków 30-394, Poland
[15]Sage Bionetworks, Seattle, WA 98121, USA
[16]Ontario Institute of Cancer Research, Toronto, ON M5G 0A3, Canada
[17]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada

(Affiliations continued on next page)

SUMMARY

Cancer is driven by genomic alterations, but the processes causing this disease are largely performed by proteins. However, proteins are harder and more expensive to measure than genes and transcripts. To catalyze developments of methods to infer protein levels from other omics measurements, we leveraged crowdsourcing via the NCI-CPTAC DREAM proteogenomic challenge. We asked for methods to predict protein and phosphorylation levels from genomic and transcriptomic data in cancer patients. The best performance was achieved by an ensemble of models, including as predictors transcript level of the corresponding genes, interaction between genes, conservation across tumor types, and phosphosite proximity for phosphorylation prediction. Proteins from metabolic pathways and complexes were the best and worst predicted, respectively. The performance of even the best-performing model was modest, suggesting that many proteins are strongly regulated through translational control and degradation. Our results set a reference for the limitations of computational inference in proteogenomics.

A record of this paper's transparent peer review process is included in the Supplemental Information.

INTRODUCTION

The central dogma of molecular biology describes the two-step process, transcription and translation, by which the information in genes flows into proteins: DNA → RNA → protein. Proteins can be further modified post-translationally to regulate cellular function. The processes of transcription and translation are regulated in numerous ways. Understanding these regulations and how they are altered in tumors, holds the promise to advance cancer research and treatment (Alfaro et al., 2014). Dysregulated protein activity—including kinase signaling and chromatin acetylation—is most directly assessed with measurements of proteins and their post-translational modifications. Therefore, proteomics holds important complementary value to the genomic and transcriptomic characterization of tumors.

Boris Reva,[3] NCI-CPTAC-DREAM Consortium, Emily Boja,[24] Henry Rodriguez,[24] Gustavo Stolovitzky,[25] Yuanfang Guan,[6] Jaewoo Kang,[9] Pei Wang,[3] David Fenyö,[4,5,*] and Julio Saez-Rodriguez[2,26,27,30,*]

[18]Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON M5S 1A8, Canada
[19]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA
[20]Department of Urology, University of California, Los Angeles, CA 90095, USA
[21]Institute for Precision Health, University of California, Los Angeles, CA, USA
[22]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA
[23]Department of Biology, Brigham Young University, Provo, UT 84604, USA
[24]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Bethesda, MD 20892, USA
[25]IBM Research, IBM Thomas J Watson Research Center, Yorktown Heights, NY 10598, USA
[26]European Molecular Biology Laboratory-The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK
[27]Institute for Computational Biomedicine, Heidelberg University Hospital and Heidelberg University, Faculty of Medicine, Bioquant Heidelberg, Hedelberg 69120, Germany
[28]These authors contributed equally
[29]Present address: Division of Oncology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA 94305, USA
[30]Lead Contact
*Correspondence: david@fenyolab.org (D.F.), julio.saez@bioquant.uni-heidelberg.de (J.S.-R.)
https://doi.org/10.1016/j.cels.2020.06.013

The relationship between mRNA transcripts and proteins is fundamental to our understanding and application of molecular biology (Crick, 1958). With the rise of comprehensive cellular measurement of both transcripts and proteins, this relationship has been extensively explored (Payne, 2015; Vogel and Marcotte, 2012; Liu et al., 2016). Since gene expression is regulated at both transcription and translation, transcript and protein levels are not necessarily expected to be correlated, and the relative strength of the transcriptional and translational regulation varies widely between genes. Genes with little translational regulation will have a highly correlated transcript and protein levels, but the vast majority of proteins have a non-negligible translational regulation. Measurements of transcript levels provide information about how many transcripts are available for translation at that moment. In contrast, protein level measurements provide a historical record of the protein production and degradation. Therefore, differences in degradation rates between transcripts and proteins will decrease the correlation. Given the different regulations of transcripts and proteins, it is neither surprising nor controversial that for some genes transcript and protein levels are highly correlated, but for the others they are not. In this challenge, we assess how much information about the differences in transcript and protein regulation is contained in the transcript quantities of other genes. Many groups have attempted at measuring the quantitative correlation between transcript and protein levels (Gygi et al., 1999; Nagaraj et al., 2011) or computationally predicted protein levels from transcript levels (Wilhelm et al., 2014) with only modest success (Fortelny et al., 2017). Proteogenomic analyses of large cohorts of primary tumors have repeatedly shown low correlations between transcript and protein levels, with large inter-gene variability (Zhang et al., 2014, 2016; Mertins et al., 2016; Sinha et al., 2019). Given the importance of proteomics data for understanding cancer phenotypes (Alfaro et al., 2014) and the costs of mass spectrometry-based proteomic experiments, it is of great interest to assess and improve the predictability of protein levels based on other data. This would enable discovery analysis of data from many large cohorts with only DNA or mRNA data. For all these reasons, it is crucial to understand to what extent protein levels can be inferred from transcript levels.

To improve the performance of using gene copy number and transcript levels in predicting protein and phosphorylation levels, we launched a community-based collaborative competition (Saez-Rodriguez et al., 2016): the NCI-CPTAC DREAM proteogenomics challenge in November 2017 (https://www.synapse.org/ProteogenomicsChallenge). In this challenge, participants applied different computational methods to proteogenomic data generated by the clinical proteomic tumor analysis consortium (CPTAC) to predict protein and phosphorylation levels based on genomic and transcriptomic data. Participants had access to quantitative measurements of gene copy number, transcript, protein, and phosphorylation levels for thousands of genes for two cancer cohorts. In addition to the omics data, participants were invited to use prior information from existing databases, such as protein-protein interactions and physicochemical properties to improve performance.

Here, we present the outcome of the NCI-CPTAC DREAM proteogenomics challenge, where we investigated results from the winning team and created a consensus algorithm from the top performers. We revealed factors influencing protein predictability and biological pathways associated with the top predictor genes. Finally, we applied the consensus algorithm to additional ovarian tumor data (Cancer Genome Atlas Research Network, 2011) and showed that inferred phosphorylation levels predict overall survival and capture disease related mechanisms better than the corresponding protein level, illustrating the value of the methods produced in the challenge.

Our unbiased assessment of predictive methods suggests that, while these methods can provide additional insight and might work well for specific proteins, we are far from being able to predict proteomics from transcriptomics and genomics globally.

RESULTS

Challenge Design

We posed the following two questions: (1) How accurately can protein levels be predicted from transcriptome and gene copy number variation data (Figure 1A; the proteomics sub-challenge)? (2) How accurately can phosphoprotein levels be

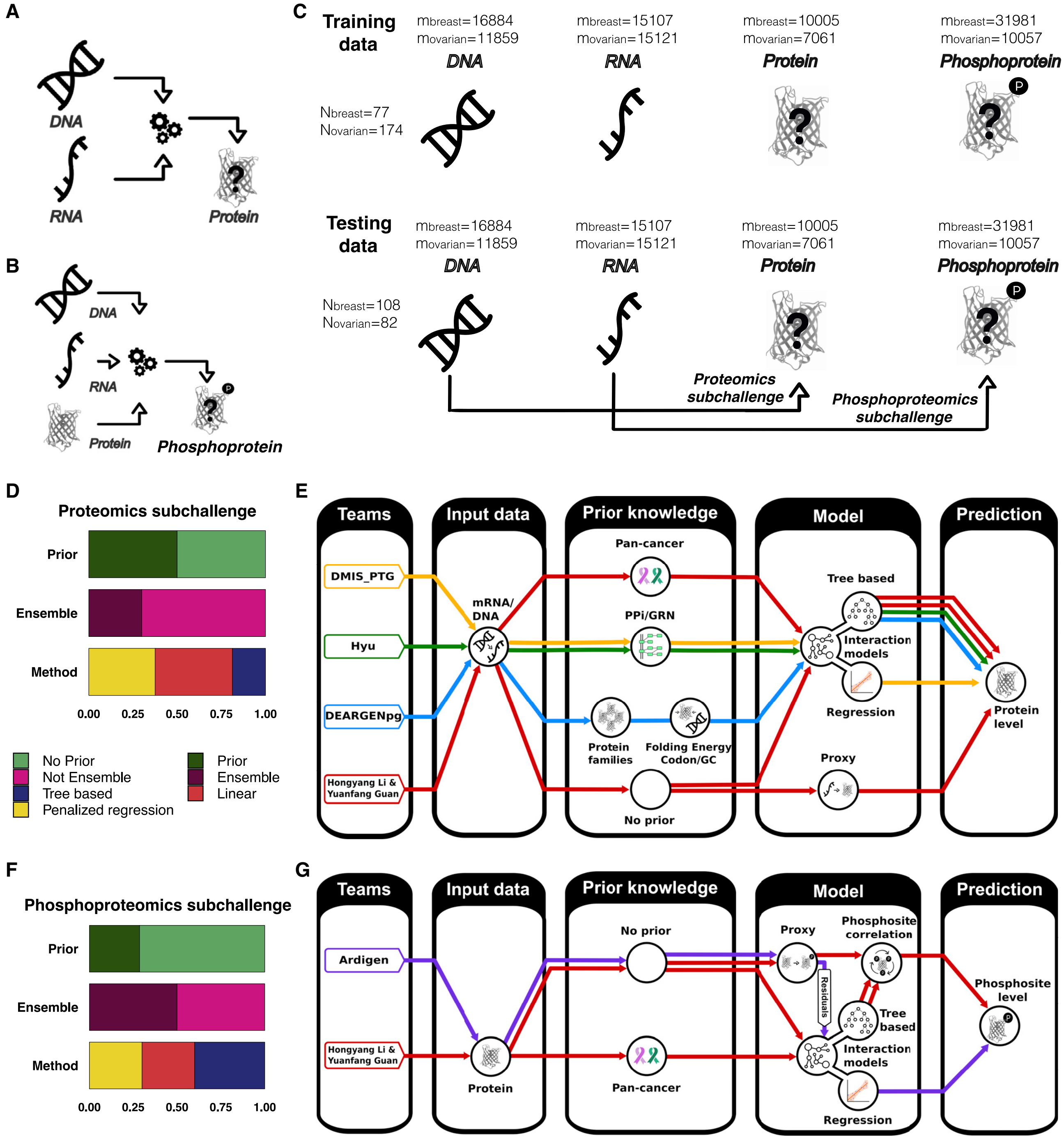


Figure 1. Challenge Overview and Best Performing Methods

(A) Protein prediction task: participants are given gene copy numbers and transcript levels to predict protein levels for breast and ovarian tumors.

(B) Phosphoprotein prediction task: participants are given gene copy numbers, transcript levels and protein levels to predict phosphoprotein level data for the breast and ovarian tumors.

(C) Graphic of data involved in the challenge and sample size for breast and ovarian data sets.

(D) Characteristics of models implemented for the prediction of protein levels.

(E) Protein prediction task: the methods used by each of the four performing teams for the protein prediction task: (1) Team Hongyang Li and Yuanfang Guan used an ensemble of the transcript level as proxy, a random forest regression model and a transfer model using data from both tissues to train the random forest; (2) Team Hyu selected features using KEGG pathways and protein-protein interactions (human protein reference database) to train a random forest regression model; (3) Team DEARGENpg built models on protein families and used them to train an ensemble of random forest, XGboost and gradient boost models, including additional features such as codon bias, GC count and protein folding energy; (4) Team DMIS_PTG used least absolute shrinkage and selection operator

(legend continued on next page)

predicted from protein levels, transcript levels, and gene copy number (Figure 1B; the phosphoproteomics sub-challenge)? The CPTAC proteogenomic challenge had multiple rounds of competition: two leaderboard rounds, a final round, and a collaborative phase. For these challenges, genomics, transcriptomics, proteomics, and phosphoproteomics data from the CPTAC discovery collection of 77 breast (Mertins et al., 2016) and 174 ovarian (Zhang et al., 2016) tumors were utilized to train the models (Figure 1C; Table S1). All proteomics and phosphoproteomics data were processed using the CPTAC common data analysis pipeline (Rudnick et al., 2016) (STAR Methods). An overview of the data in terms of the missing data rate and the batch effect is contained in Figures S1 and S2. In order to assess the predictive performance of different methods, the CPTAC confirmatory data were utilized as a validation set. These data included 108 breast and 82 ovarian cancer samples (Figure 1C). Since the test data were not published at the time of the competition, Docker containers were utilized to store pre-trained models and to perform prediction of test data (Guinney and Saez-Rodriguez, 2018). The predictive performance of each job was assessed using the Docker system, and the corresponding score was released to participants so that they could further improve their models throughout the leaderboard rounds. In the final round, participants evaluated the predictive performance of their models on the confirmatory breast and ovarian cancer data. In order to limit overfitting, during the final round, a maximum number of three submissions per team was allowed. The final-round submissions were scored against a held-out dataset and the best-performing teams were identified. In the collaborative phase, top-performing teams of the final round built an ensemble prediction model.

Methods and General Outcome of the Challenge

A total number of 30 and 15 teams participated in the proteomics and phosphoproteomics prediction sub-challenges, respectively. Among the teams participating in the proteomics challenge, 25 teams performed prediction of both breast and ovarian cancer data in the final round, while all the teams submitted jobs for both cancer types for the prediction of phosphoproteome levels. The predictive performance of different models was evaluated using Pearson's correlation and mean squared error between observed proteome or phosphoproteome levels versus predicted values across different patients in the testing data. As the final score, the average metric across all proteins was considered. A variety of models including ensemble methods, linear regression, and methods taking into account prior information were utilized (Figures 1D–1G). Most teams utilized prior information for the prediction of proteomics data and preferred linear versus non-linear models (Figure 1D). A different trend was observed for the prediction of phosphorylation levels where most teams opted for non-linear models and no-prior information (Figure 1F). Figures 1E, 1G, and S3 illustrate the models utilized by the top four performers in the two challenges. The same team, *Hongyang and Yuangfang*, achieved the best predictive performance in both challenges. Specifically, for the prediction of proteome data, the best-performing method was built as an ensemble of the following models: (1) a model using the corresponding transcript level as a proxy, (2) a model accounting for protein-protein interactions, and (3) a model borrowing information across tumor types (Figure 1E; STAR Methods). Similarly, for phosphorylation level prediction, they utilized an ensemble algorithm based on the following models: (1) a model using protein level as proxy, (2) a model accounting for the interaction between proteins, (3) a model accounting for correlation between nearby phosphorylation sites, and (4) a model borrowing information across cancer types (Figure 1G; STAR Methods). The winning team applied quantile normalization on the input features of the training and testing data within the Docker system (STAR Methods). Other high-performing teams utilized information from existing biological pathways, protein-protein interaction databases, and physicochemical properties of transcripts and proteins. During the collaborative phase, the top five high-performing teams built an ensemble model combining their models (Figures 1E and 1G).

Prediction of Protein Levels

For breast cancer data, 26 teams participating in the competition submitted a total of 35 models, while for the ovarian cancer data, 25 teams participating in the challenge submitted a total number of 44 different models during the final round. The average prediction performance across different algorithms in terms of correlation metric was R = 0.26 (SD = 0.17) for breast cancer and R = 0.29 (SD = 0.18) for ovarian cancer (Figure 2A). The best-performing algorithm resulted in an average correlation between actual and predicted values of 0.51 (Figures 2A and S4) for breast cancer and R = 0.53 for ovarian cancer (Figure 2A). Most performing teams utilized ensemble algorithms to predict proteomics data (Figures 2A and S3) and preferred linear over non-linear models. However, the top-performing team utilized a non-linear model based on the random forest (Figure 2A). Indeed, linear models and penalized linear regression performed worse than ensemble methods in the prediction of protein levels (Wilcoxon test $p < 0.05$) (Figure 2A). In addition, algorithms utilizing prior knowledge such as protein-protein interaction database performed better in the prediction of protein levels (Wilcoxon test $p < 0.05$). From the post challenge survey, all teams found transcript levels more predictive than gene copy number data in the proteome prediction challenge, and protein levels more predictive than all other data for phosphoproteome prediction (Table S2). The top team performed better than a simple model (R = 0.53 versus R = 0.47) utilizing the corresponding transcript level as a proxy for protein-level prediction (hereafter, the proxy model). This highlights the utility of a multivariate model in predicting proteome based on genomic data (Figure S3). Specifically, the best-performing

(LASSO) with features selected based on gene prediction networks, CORUM protein complexes and protein-protein interactions. See STAR Methods for more details.

(F) Characteristics of models implemented for the prediction of phosphorylation levels.

(G) The methods used by each team for the phosphoproteome prediction task: (1) Team Hongyang Li and Yuanfang Guan used protein level as a proxy, the correlation among phosphorylation sites on a given protein, and the pan-cancer models to train a random forest regression model; (2) Team Ardigen used the protein level as proxy to train a least-angle regression (LARS) model. See STAR Methods for more details.

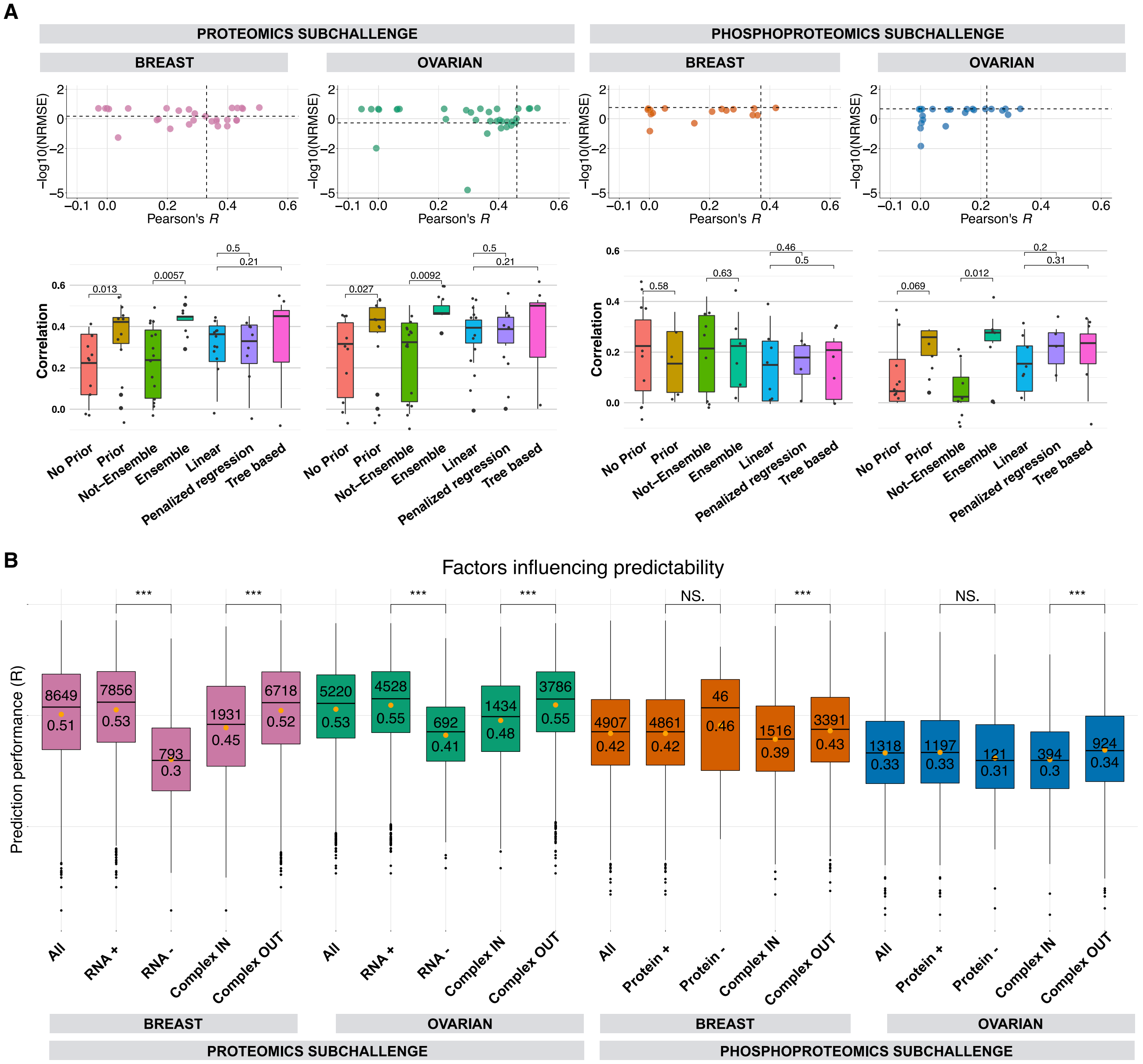


Figure 2. Factors Influencing Predictability and Commonly Predictive Genes

(A) Top: performance of submitted algorithms (dot-plot) compared to the baseline machine learning model (dashed line) for breast and ovarian cancer based on proteomics sub-challenge and phosphoproteomics sub-challenge. Bottom: Performance of teams participating in the challenge based on correlation metric for stratified based on the type of algorithms utilized. p values were computed using a Wilcoxon test.

(B) proteomics and phosphoproteomics prediction tasks: the best performer's prediction performance is shown for: (1) all proteins (*all*); (2) subset of proteins or phosphoproteins for which the corresponding transcript/protein is measured (*RNA+/Protein+*); (3) subset of proteins or phosphoproteins for which the corresponding transcript/protein is missing (*RNA-/Protein-*); (4) subset of proteins belonging to a protein complex from CORUM database (*Complex IN*); (5) subset of proteins not belonging to a protein complex (*Complex OUT*). In each boxplot, above the median, the number of proteins is indicated, and below is the average prediction performance or correlation.

team utilized an ensemble algorithm borrowing information across different cancer types when predicting the protein level of a given protein (STAR Methods). This joint analysis resulted in better predictive performance compared to the cancer-specific models for both breast and ovarian cancers. In fact, for ovarian cancer data the ovarian-specific model resulted in an average correlation of 0.44; while for breast cancer data the breast cancer-specific model resulted in an average correlation of 0.38.

Prediction of Phosphoprotein Levels

For breast data, 15 teams participated in the challenge with 52 models submitted, resulting in an average performance of R = 0.17 (SD = 0.16) and best performance of R = 0.42 (Figures 2A and S4). For ovarian cancer data, 15 teams submitted 32 models, with an average performance of R = 0.11 (SD = 0.11) and best performance of R = 0.33. Hence, predictive performance was markedly lower for the prediction of phospho proteomics compared with the prediction of proteomics data (Figure 2A). Contrary to the prediction of protein levels, ensemble methods and algorithms considering prior information did not perform better in the prediction of phosphorylation levels. However, this result might be due to the lack of statistical power given the smaller number of teams participating in the prediction of phosphorylation levels (Figure 2A).

The performance of the teams participating in the two subchallenges was compared to baseline models implemented by the challenge organizers to predict protein levels based on transcript levels and phosphorylation levels from protein levels (Supplemental Notes). Team Hongyang Li and Yuanfang Guan (University of Michigan) achieved the best performance in both prediction challenges (Figure 2A; Supplemental Notes; Figures S5 and S6) and outperformed both baseline models. All models submitted by the participants in Docker containers are freely available for public reuse at https://www.synapse.org/ProteogenomicsChallenge. Predicted protein and phosphorylation levels are available for visualization in an R Shiny application (**ProteoExplorer**).

Global Insights

Factors Influencing Protein and Phosphorylation Prediction Performance

When considering proteins with available transcript levels (4,528 proteins for ovarian and 7,856 proteins for breast), the winning team reached an average correlation of 0.53 and 0.55 for breast and ovarian tumors, respectively (Figure 2B). Expectedly, the levels of proteins whose corresponding transcripts were observed in all samples were better predicted compared with those with missing transcript levels (Figure 2B, mean R = 0.53 versus 0.30 and 0.55 versus 0.41, *p < 0.001*, Wilcoxon test, for breast and ovarian tissues, respectively). In addition, proteins which were not in a protein complex were better predicted than those belonging to a protein complex, as defined by the CORUM database (Figure 2B, mean R = 0.52 versus 0.45 and 0.55 versus 0.48, p < 0.001, Wilcoxon test, for breast and ovarian tissues, respectively). Indeed, proteins forming complexes can be co-regulated through post-transcriptional mechanisms, including degradation (Gonçalves et al., 2017) (e.g., a component of a complex that is available in higher amounts than the other components is degraded faster, making the levels of complex components more robust to transient variation of transcript level). We observed the same phenomenon with phosphoproteins, where the best predicted ones are not in protein complexes (mean R = 0.39 versus 0.43 and R = 0.30 versus 0.34 *P* < 0.001, Wilcoxon test, for breast and ovarian tissues, respectively). Finally, how well protein levels can be predicted are correlated between breast and ovarian tissues (R = 0.63, $p < 10^{-10}$, Student's t test, Supplemental Notes). No association was found between predictability of a given protein and the protein level, as shown in Figure S7. The correlation of prediction performance between breast and ovarian tissues is R = 0.63 ($p < 1 \times 10^{-10}$, Figure S8).

In order to assess the effect of protein and gene half-life on predictive performance, we leveraged transcript and protein half-life estimations from two previous publications (Tani et al., 2012), (Zecha et al., 2018). Genes overlapping with ovarian and breast cancer data were categorized into four groups depending on transcript and protein half-life: long-long, short-long, short-short, and long-short. Despite the low correlation between protein half-life and predictive performance (Figure S9A: R = 0.13 and 0.15 in ovarian and breast cancer, respectively), we observed greater performance in genes with longer protein half-lives (Figure S9B), regardless of the half-lives of transcripts and the cancer type (*p < 0.001*, Wilcoxon test, for breast and ovarian tissues, respectively).

We assessed the biological insights in the best-performing method. In particular, for each protein, predictive features were selected based on the importance score resulting from random forest (STAR Methods). Then, common predictors of protein level were identified by considering the predictive features selected across a large number of proteins. We found that genes commonly predictive of protein levels may be more essential for cancer cell development (STAR Methods; Figures S10 and S11), and potentially involved in survival outcome, especially in ovarian cancer tumors (STAR Methods; Table S3).

Pathway Analysis of Top Predicted Proteins in Ovarian Cancer

Subsequently, we assessed whether the most predictable proteins in ovarian cancer were enriched in any biological mechanisms. For this purpose, we considered the canonical pathways from the MSigDB database (Liberzon et al., 2011), and specifically 1,101 pathways that contained at least 5 proteins observed in both the training and testing data. Based on the ensemble model, proteins were ordered from the most predictable to the least predictable using the correlation between the predicted value and true value as metric. Then, an enrichment analysis was performed to identify pathways overrepresented at the top of this list. Most of the pathways enriched at the top of the list were metabolic pathways (Figure 3A; Tables S4 and S5). Although the number of enriched pathways was similarly based on the proxy model and ensemble method at 10% false discovery rate (FDR) (i.e., 37 and 41 pathways enriched, respectively), the ensemble method outperformed the proxy model in terms of predictive performance (Figure 3B). Accordingly, proteins mapping to enriched pathways were predicted with better accuracy based on the ensemble model rather than the proxy model (Tables S4 and S5) (Benjamini-Hochberg adjusted p value from t test < 10%). An interesting pathway was the "KEGG ECM interaction pathway," whose proteins were predicted with higher accuracy based on the ensemble method compared to the proxy model (Figure 3). Several studies demonstrated the role of extracellular matrix in cell proliferation, migration, and apoptosis (Pickup et al., 2014). Although in ovarian cancer, dysregulation and loss of ECM components are observed, it is still unclear how tumor cells influence ECM remodeling (Cho et al., 2015). Based on the best-performing model, we found a total of 337 unique genes regulating proteins in "KEGG ECM interaction pathway." In particular, these predictors were enriched for "KEGG ECM-receptor interaction," "KEGG focal adhesion," "PI3K-Akt signaling pathway," and "regulation of actin cytoskeleton" adjusted $p = 1 \times 10^{-5}$).

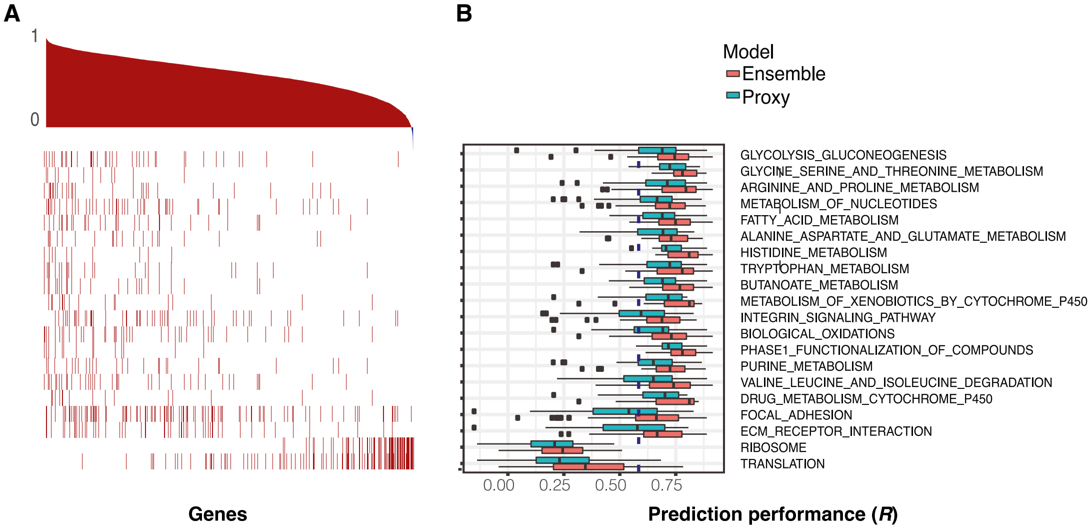


Figure 3. Most Predictable Pathways based on Ensemble Method for Ovarian Cancer

(A) List of pathways enriched in the set of most predictable and least predictable proteins in the community phase ensemble model. A subset of enriched pathways at 10% FDR for ovarian data. Correlation between predicted and true value for each protein, with proteins ordered in decreasing order of correlation. Then, for each pathway, the position of the genes included in the pathway is shown.

(B) Comparison between proxy and ensemble model in predicting proteins contained in each of the enriched pathways. Only proteins with corresponding transcript data were considered. The dashed line shows the average correlation between predicted and true value across all proteins.

Survival-Enriched Pathways in Phosphorylation Prediction

A total number of 1,318 phosphoproteins in 82 ovarian cancer patients was predicted by the proxy and ensemble models in the phosphoproteomics prediction challenge. Based on the ensemble model, we predicted the phosphorylation levels among additional 105 ovarian tumors whose protein but not phosphorylation measurements were available in CPTAC. We evaluated the performance of predicted phosphorylation levels in identifying pathways associated with overall survival (OS) (STAR Methods). Neither measured phosphoproteomics data from the training set nor measured proteomics data from the training and prediction set identified any KEGG pathway associated to OS (FDR < 0.05). However, predicted phosphorylation scores based on the ensemble model identified one KEGG pathway, vascular muscle contraction, as being associated with survival in the prediction data. This pathway also showed the trend of significance in the real protein and phosphoprotein data ($p < 0.05$) but did not reach the FDR cutoff (STAR Methods). None of the pathways were identified based on the proxy model.

A total of 25 predicted phosphoproteins from the vascular muscle contraction pathway were available in our prediction dataset (Figure S12). Based on these phosphoproteins, we further performed principal component analysis (PCA) and created a survival score by combining the 2nd and 3rd principal components (PCs), which were found associated with OS. This PC-based signature was significantly associated with survival (*HR* = 1.32, p = 0.003). Stratified by high versus low score, the Kaplan-Meier curve separated the patient population (rank score p = 0.001; HR=3.2, p = 0.002). Among the top contributing phosphoproteins of the PC-score, PRKCB resulted in a strong negative loading (Figure S12). PRKCB generally functions as a tumor suppressor (Antal et al., 2015) and is therefore associated with better outcome if overexpressed (Figure S12). In addition, PPP1R14A resulted in a positive loading (Figure S12), and thus, an increased level of this phosphoprotein was associated with a poorer outcome (Figure S12). Of note, similar analyses were performed for the proteomics prediction challenge, but no pathways were found associated with OS.

DISCUSSION

In this study, we performed an unbiased assessment of computational models to predict protein and phosphorylation levels from transcript levels and gene copy number. The broad context is an open-ended hypothesis generating omics experiments in which we are attempting to quantitatively measure as many genes as possible, without limiting ourselves with a hypothesis. The overarching question is how much information about the proteome is contained in the transcriptome, and we show that this information varies substantially for different genes and, in general, it is quite modest. Therefore, we conclude that it is important to measure the proteome, i.e., the functional gene products directly, to generate most useful hypotheses. After having generated hypotheses for small sets of genes, these should be investigated in targeted functional studies to validate the results of the proteomics results.

By crowdsourcing this question, we were able to extensively sample computational approaches for answering it. Thus, the achieved performance is unlikely to improve significantly by applying other algorithms and represents a probable approximation to the actual capacity to predict protein levels, at least for the datasets we have used. The best-performing methods were tree- and ensemble-based models. External information, such as functional pathways, interaction networks, codon count, GC percentage, protein folding energy, and transcription factors, was also useful but not critical for good performance. Further information, such as transcript and protein half-lives, and the effect of codons on protein-to-transcript ratio, which have been shown to be informative (Eraslan et al., 2019), were not used by the participants. The best-performing teams mainly used transcript levels to predict protein levels and found no benefit by including gene copy number data. For the phosphoproteomics prediction task, protein level was the most effective data type to predict phosphorylation levels, and the inclusion of gene copy numbers or transcript levels did not provide any additional improvement over using protein levels only. Another key feature of the winning team was to borrow information across cancer data to increase the effective sample size, and thus the performance of the predictive model. This is important when dealing with relatively small sample size and a high number of features (''large p small n problem'').

Overall, the prediction performances for most participants were below the baseline methodology, implementing a single task machine learning algorithm without using prior knowledge, and some teams even performed close to random (Figure 1E). The overall low performance could be due to multiple reasons, including the small sample size compared to the number of features, or largely simply because a significant part of the biology is not captured by the available molecular data, which did not include quantitative measurements of all the molecules involved in the regulation of a cell. In addition, we had no data capturing the dynamics of the underlying processes, such as rates of translation and rates of degradation of transcripts and proteins. Whether using additional molecular data can improve results, remains open and should be addressed in future work.

The relationship between transcript and protein levels have been extensively studied. During highly dynamic phases, such as cellular differentiation or stress response, the transcript-protein correlation is weak (Liu et al., 2016), while in steady-state conditions, protein levels are largely determined by transcript levels. Additional mechanisms, including differences in half-lives and post-transcriptional regulation (Haider and Pal, 2013) also largely determine the correlations, but our understanding of this dynamic process is still far from complete.

We observed that proteins for which the corresponding transcripts have been measured were better predicted compared with those which have not (Figure 2A). In contrast, we did not observe a better predictive performance for high abundance proteins (Figure S7). Proteins not belonging to complexes were better predicted than those belonging to a protein complex (Figure 2A). This can be explained by the fact that degradation rates of proteins can be different if they are in a complex, e.g., a pair of proteins can be stabilized when in a complex, and therefore, their protein levels will be correlated (Gonçalves et al., 2017; Ori et al., 2016)(Liu et al., 2016). This additional regulation through degradation renders the protein less likely to follow the transient variation of transcript level. Despite the globally modest performance, some proteins could be predicted accurately. In some cases, the transcript levels themselves are good predictors; here, a direct use of RNA as proxy of proteins seems adequate. For others, the model predicted well and better than the RNA proxy; here is where the approaches developed within the challenge are most useful. In contrast, proteins subjected to tight control, such as those in complexes, are harder to predict.

Using results from the challenge, we were able to identify metabolic and extracellular matrix pathways enriched in the set of most predictable proteins. Highly connected nodes are more pleiotropic and more likely to be essential than other nodes (Macneil and Walhout, 2011). Therefore, we explored the feature importance in the best-performing model and identified the most predictive genes of protein levels. We found that commonly predictive genes were more essential and predictive of patient survival for ovarian cancer (Supplemental Notes). Predicted phosphorylation scores using an ensemble model significantly identified the KEGG vascular muscle contraction pathway to be associated with patient survival for ovarian cancer. We further built a score based on this pathway, which was significantly associated with survival (HR = 1.32, p = 0.003).

A drawback of this study is given by the limited coverage of proteomic profiling. In fact, even for the deep mass spectrometry-based proteomics data, the coverage in terms of proteins measured is limited and it is therefore difficult to quantify specific splice forms. Because of the limited coverage, we chose to quantify proteins at the gene level to avoid unreliable quantitation of splice forms. For transcripts, the RNA-seq data could give some splice form quantitation, but because the mass spectrometry data do not support quantifying splice forms, we chose to use the more reliable gene-level quantitation for both transcripts and proteins.

In summary, we leveraged crowdsourcing to establish an unbiased framework for protein level prediction. We explored to what extent protein levels can be predicted based on transcript levels and gene copy number. We made the best-performing models available for reuse and provided the challenge predictions as a resource for understanding protein-transcript relationships and further development of methods by the community. Methods derived from this challenge could be used for predicting protein level on patient samples for which only transcript level measurements are available. We found that, despite their limitations, the resulting models can derive meaningful insights, at least at an aggregated level, such as new associations between pathways and survival from predicted proteomic data. Thus, we suggest scientists to run these methods on transcriptomic data to obtain large-scale estimated proteogenomic characterization of tumor samples, that could help better understand cancer biology and identify novel biomarkers to predict survival and for patient stratification.

Key Changes Prompted by Reviewer Comments

Based on reviewers' comments, we have implemented a number of major changes to the manuscript. First, we refocused the paper on the challenge itself. For this purpose, we included a detailed description of predictive models utilized by different teams and their performance (Figures 1C, 1D, 1F, and 2A), and we revised the corresponding text accordingly. We also included more details on the data utilized in the challenge in the STAR Methods and commented on missing data rate (Figure S1). Further, we removed downstream analysis based on cell line datasets.

We also expanded some of our analyses. In particular, we commented on the essentiality of common predictors and included Figures S10 and S11. To gain insight on factors influencing predictability, we added Figure S7, which shows: (1) the association between predictive performance and protein levels and (2) the overall distribution of correlation metric across all measured proteins. The corresponding text was added to the section Global Insights. We added Figure S9 to show the effect of RNA and protein half-lives on the prediction of proteomics-based transcriptomics data. Corresponding text was added to the section Global Insights.

Finally, we toned down the claim of common predictors of protein expression being predictive of patient survival and only mentioned it briefly in the Introduction section. We have improved the readability of Figure S3. We have moved the figure concerning the survival analysis based on predicted phosphorylation levels to supplementary figures (Figure S12).

For context, the complete transparent peer review record is included within the Supplemental Information.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - Description of Challenge Data
  - Training Data
  - Team Ardigen Method

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.06.013.

CONSORTIA

The members of the NCI-CPTAC-DREAM Consortium are Tunde Aderinwale, Ebrahim Afyounian, Piyush Agrawal, Mehreen Ali, Alicia Amadoz, Francisco Azuaje, John Bachman, Seohui Bae, Sherry Bhalla, Emily Boja, Paul C. Boutros, Anna P. Calinawan, José Carbonell-Caballero, Priyanka Chakraborty, Kumardeep Chaudhary, Yonghwa Choi, Yoonjung Choi, Cankut Çubuk, Sandeep Kumar Dhanda, Joaquín Dopazo, Laura L. Elo, David Fenyö, Ábel Fóthi, Olivier Gevaert, Kirsi Granberg, Russell Greiner, Yuanfang Guan, Eunji Heo, Marta R. Hidalgo, Vivek Jayaswal, Hwisang Jeon, Minji Jeon, Jan Kaczmarczyk, Sunil V Kalmady, Yasuhiro Kambara, Jaewoo Kang, Keunsoo Kang, Tony Kaoma, Harpreet Kaur, Hilal Kazan, Devishi Kesar, Juha Kesseli, Daehan Kim, Keonwoo Kim, Sang-Yoon Kim, Sunkyu Kim, Sajal Kumar, Bora Lee, Heewon Lee, Hongyang Li, Zhi Li, Yunpeng Liu, Roland Luethy, Weiping Ma, Swapnil Mahajan, Mehrad Mahmoudian, Arnaud Muller, Petr V. Nazarov, Hien Nguyen, Matti Nykter, Shujiro Okuda, Sungsoo Park, Samuel H. Payne, Francesca Petralia, Gajendra Pal Singh Raghava, Jagath C. Rajapakse, Tommi Rantapero, Boris Reva, Henry Rodriguez, Hobin Ryu, Julio Saez-Rodriguez, Francisco Salavert, Sohrab Saraei, Ruby Sharma, Ari Siitonen, Artem Sokolov, Xiaoyu Song, Piotr Stępniak, Gustavo Stolovitzky, Kartik Subramanian, Veronika Suni, Tomi Suomi, Léon-Charles Tranchevent, Salman Sadullah Usmani, Tommi Välikangas, Roberto Vega, Pei Wang, Michał Warchoł, Mi Yang, Han Yu, Thomas Yu, and Hua Zhong.

ACKNOWLEDGMENTS

We would like to thank Nicolas Palacio and Emanuel Gonçalves for their suggestions and ideas in improving the manuscript. We would also like to thank the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of proteogenomics, for providing the data sets used in this challenge and making it freely available to the public. The challenge was supported financially by the CPTAC. This research work is supported by the National Institutes for Health (NIH), National Cancer Institute (NCI) CPTAC awards U24 CA210972 and U24 CA210993, and a contract 16X230 from Leidos Biomedical Research.

AUTHOR CONTRIBUTIONS

M.Y., F.P., Z.L., W.M., D.F., P.C.B., E.B., H.R., and J.S.-R. organized the challenge. M.Y., F.P., X.S., H.L., Y.G., and Z.L. performed the post challenge analysis and wrote the manuscript. M.Y. developed the Shiny application. M.Y. and F.P. tested baseline solutions for the challenge questions. H.L. and Y.G. developed the best-performing algorithms of the challenge. F.P. organized the collaborative round. S.K., H.L., H.Y., B.L., S.B., E.H., J.K., P.S., M.W., Y.G., and J.K. participated in the collaborative round and the writing of their methods. Z.L. processed the data used in the challenge. M.Y. assisted in processing the challenge data. T.Y. implemented the challenge infrastructure. P.C.B. assisted in challenge design. P.C.B., S.H.P., H.R., Y.G., and G.S. commented on the manuscript and provided meaningful feedback. P.W. provided meaningful feedback at all stages. H.R. initiated the challenge. D.F. and J.S.-R. supervised the project. NCI-CPTAC-DREAM consortium participated in the challenge and submitted their predictions.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 22, 2019
Revised: March 12, 2020
Accepted: June 29, 2020
Published: July 24, 2020

REFERENCES

Alfaro, J.A., Sinha, A., Kislinger, T., and Boutros, P.C. (2014). Onco-proteogenomics: cancer proteomics joins forces with genomics. Nat. Methods *11*, 1107–1113.

Antal, C.E., Hudson, A.M., Kang, E., Zanca, C., Wirth, C., Stephenson, N.L., Trotter, E.W., Gallegos, L.L., Miller, C.J., Furnari, F.B., et al. (2015). Cancer-associated protein kinase C mutations reveal kinase's role as tumor suppressor. Cell *160*, 489–502.

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Cho, A., Howell, V.M., and Colvin, E.K. (2015). The extracellular matrix in epithelial ovarian cancer - a piece of a puzzle. Front. Oncol. *5*, 245.

Crick, F.H. (1958). On protein synthesis. Symp. Soc. Exp. Biol. *12*, 138–163.

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallström, B.M., Uhlén, M., Asplund, A., Pontén, F., Wieland, T., Hopf, T., et al. (2019). Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol. Syst. Biol. *15*, e8513.

Fortelny, N., Overall, C.M., Pavlidis, P., and Freue, G.V.C. (2017). Can we predict protein from mRNA levels? Nature *547*, E19–E20.

Gonçalves, E., Fragoulis, A., Garcia-Alonso, L., Cramer, T., Saez-Rodriguez, J., and Beltrao, P. (2017). Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. Cell Syst. *5*, 386–398.e4.

Guinney, J., and Saez-Rodriguez, J. (2018). Alternative models for sharing confidential biomedical data. Nat. Biotechnol. *36*, 391–392.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19*, 1720–1730.

Haider, S., and Pal, R. (2013). Integrated analysis of transcriptomic and proteomic data. Curr. Genomics *14*, 91–110.

Kanehisa, M. (2008). The KEGG database. In 'In Silico' Simulation of Biological Processes, Novartis Foundation., G. Bock, and J.A. Goode, eds. (Wiley), pp. 91–103.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database–2009 update. Nucleic Acids Res *37*, D767–D772.

Kosti, I., Jain, N., Aran, D., Butte, A.J., and Sirota, M. (2016). Cross-tissue analysis of gene and protein expression in normal and cancer tissues. Sci. Rep. *6*, 24799.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell *165*, 535–550.

Macneil, L.T., and Walhout, A.J.M. (2011). Gene regulatory networks and the role of robustness and Stochasticity in the control of gene expression. Genome Res. *21*, 645–657.

McFarland, J.M., Ho, Z.V., Kugener, G., Dempster, J.M., Montgomery, P.G., Bryan, J.G., Krill-Burger, J.M., Green, T.M., Vazquez, F., Boehm, J.S., et al. (2018). Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat. Commun. *9*, 4610.

Menden, M.P., Wang, D., Mason, M.J., Szalai, B., Bulusu, K.C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., et al. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat. Commun. *10*, 2674.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol *17*, 47.

Park, S., Kim, J.M., Shin, W., Han, S.W., Jeon, M., Jang, H.J., Jang, I.S., and Kang, J. (2018). BTNET: boosted tree based gene regulatory network inference algorithm using time-course measurement data. BMC Syst. Biol. *12*, 20.

Payne, S.H. (2015). The utility of protein and mRNA correlation. Trends Biochem. Sci. *40*, 1–3.

Pickup, M.W., Mouw, J.K., and Weaver, V.M. (2014). The extracellular matrix modulates the hallmarks of cancer. EMBO Rep *15*, 1243–1253.

Rudnick, P.A., Markey, S.P., Roth, J., Mirokhin, Y., Yan, X., Tchekhovskoi, D.V., Edwards, N.J., Thangudu, R.R., Ketchum, K.A., Kinsinger, C.R., et al. (2016). A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. J. Proteome Res. *15*, 1023–1032.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., and Mewes, H.W. (2008). Corum: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. *36*, D646–D650.

Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L., Meyer, P., Norman, T., and Stolovitzky, G. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat. Rev. Genet. *17*, 470–486.

Sales, G., Calura, E., Cavalieri, D., and Romualdi, C. (2012). Graphite - a bioconductor package to convert pathway topology to gene network. BMC Bioinformatics *13*, 20.

Sales, G., Calura, E., and Romualdi, C. (2019). metaGraphite-a new layer of pathway annotation to get metabolite networks. Bioinformatics *35*, 1258–1260.

Sinha, A., Huang, V., Livingstone, J., Wang, J., Fox, N.S., Kurganovs, N., Ignatchenko, V., Fritsch, K., Donmez, N., Heisler, L.E., et al. (2019). The proteogenomic landscape of curable prostate cancer. Cancer Cell *35*, 414–427.e6.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. *22*, 947–956.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Wang, H., Zhang, Y., and Du, Y. (2013). Ovarian and breast cancer spheres are similar in transcriptomic features and sensitive to fenretinide. BioMed Res. Int. *2013*, 510905.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.

Zecha, J., Meng, C., Zolg, D.P., Samaras, P., Wilhelm, M., and Kuster, B. (2018). Peptide level turnover measurements enable the study of proteoform dynamics. Mol. Cell. Proteomics *17*, 974–992.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.-Y., Petyuk, V.A., Chen, L., Ray, D., et al. (2016). Integrated Proteogenomic characterization of human high-grade serous ovarian cancer. Cell *166*, 755–765.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Testing breast and ovarian proteome and phosphoproteome	This paper	https://www.synapse.org/ProteogenomicsChallenge
Training breast and ovarian proteome and phosphoproteome	Mertins et al., 2016; Zhang et al., 2016	https://cptac-data-portal.georgetown.edu/cptacPublic
Software and Algorithms		
Python	Python Software Foundation	http://www.python.org
R version 3.4.0	R Foundation for Statistical Computing	https://www.r-project.org/
Code Repository	This paper	https://github.com/Sage-Bionetworks/NCI-CPTAC-DREAM-Proteogenomics-Challenge/tree/master/Subchallenges_2_3
Proteome Estimator	This paper	https://github.com/Sage-Bionetworks/NCI-CPTAC-DREAM-Proteogenomics-Challenge/tree/master/Subchallenges_2_3/ProteoEstimator
ProteoExplorer	This paper	https://heidelberg.shinyapps.io/proteoexplorer/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Julio Saez-Rodriguez (julio.saez@bioquant.uni-heidelberg.de)

Materials Availability

This study did not generate new materials.

Data and Code Availability

The challenge website can be found at https://www.synapse.org/ProteogenomicsChallenge

A Shiny application ProteoExplorer is at https://heidelberg.shinyapps.io/proteoexplorer/

Codes and python module ProteoEstimator are freely available for download at https://github.com/Sage-Bionetworks/NCI-CPTAC-DREAM-Proteogenomics-Challenge

METHOD DETAILS

Description of Challenge Data

The NCI-CPTAC DREAM challenge utilized the CPTAC discovery data (TCGA) from breast and ovarian tumor samples as training data (Mertins et al., 2016; Zhang et al., 2016); while the CPTAC confirmatory data was used as testing data for performance evaluation (NCI-CPTAC data portal). Both discovery and confirmatory data included proteomics, phosphoproteomics, transcriptomics (mRNA) and copy number alteration data (CNA). Sample size varied between different platforms due to the availability and quality of original tumor samples at the time of the study. Mass-Spectrometry based proteomic and phosphoproteomic characterization of these tumor samples yield quantitative measurements of thousands of proteins and tens of thousands of phosphorylation sites, which served as the target to be predicted in the sub-challenges.

Training Data

Breast Cancer

- Proteome: 10005 proteins for 77 patients
- Phosphoproteome: 31981 phosphosites for 77 patients
- CNA: 16884 genes for 77 patients
- mRNA: 15107 genes for 77 patients

Ovarian Cancer

- Proteome from PNNL: 7061 proteins for 84 patients
- Proteome from JHU: 7061 proteins for 122 patients
- Phosphoproteome: 10057 phosphosites for 69 patients
- CNA: 11859 genes for 559 patients
- mRNA (Array): 15121 genes for 569 patients
- mRNA (RNA-seq): 15121 genes for 294 patients

Training proteomics and phosphoproteomics data of breast and ovarian tumors were downloaded from the CPTAC data portal (https://cptac-data-portal.georgetown.edu/cptacPublic/) and processed by the common data analysis pipeline from CPTAC. For breast proteome, 105 (77 passed QC) tumors from different patients were analyzed at the Broad Institute. The log ratios of protein levels were calculated, including only peptides mapping unambiguously to a single gene. Breast tumor samples ('TCGA-AO-A12B', 'TCGA-AO-A12D', 'TCGA-C8-A131' assayed in duplicate for quality control purposes) were mean aggregated in the uploaded training data (https://cptac-data-portal.georgetown.edu/cptac/s/S015). For ovarian proteome, 206 samples were collected from 174 unique patients (84 from Pacific Northwest National Laboratory (PNNL), 122 from Johns Hopkins University and 32 measured by both centers). We provided the participants with both proteome collections for training to cover the maximum number of samples for the Proteomics sub-challenge. However, for the phosphoproteomics sub-challenge, ovarian phosphoproteome of 69 patients were measured exclusively by PNNL.

Breast cancer RNA-seq data with 1212 samples was downloaded from:http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/BRCA/20160128/gdac.broadinstitute.org_BRCA.mRNAseq_Preprocess.Level_3.2016012800.0.0.tar.gz

Ovarian Microarray data was downloaded from Zhang et al. (Zhang et al., 2016) and Breast cancer RNA-seq data was downloaded from broad firehose. The main reason for providing participants microarray data was that sample coverage was greater between microarray and proteome data compared to RNAseq data. However, only RNA-seq data was measured for the CPTAC confirmatory collection (testing data). CNA data was directly downloaded from the two corresponding CPTAC publications (Mertins et al., 2016; Zhang et al., 2016). Non-unique gene IDs were median aggregated.

Testing Data

CNA, RNA-seq, Proteome and Phosphoproteome from the CPTAC confirmatory collected patients were provided as testing data for Proteomics and Phosphoproteomics sub-challenges. Since during the first and second round of the challenge only ovarian cancer testing data was available, predictive performance was evaluated on ovarian cancer data only. In particular, the same testing data was utilized in both rounds to allow participants to improve the prediction of their proposed algorithms. This testing data based on ovarian cancer included:

- Proteome data containing 7061 proteins for 20 patients
- Phosphoproteome data containing 10057 phosphosites for 20 patients
- CNA data containing 11859 genes for 20 patients
- mRNA(RNA-seq) containing 15121 genes for 20 patients

During the final round, predictive performance was evaluated based on both breast and ovarian cancer data. The breast cancer data included

- Proteome data containing 10005 proteins for 108 patients
- Phosphoproteome data containing 31981 phosphosites for 108 patients
- CNA data containing 16884 genes for 108 patients
- mRNA data containing 15107 genes for 108 patients

while the ovarian cancer data included

- Proteome data containing 7061 proteins for 62 patients
- Phosphoproteome data containing 10057 phosphosites for 62 patients
- CNA data containing 11859 genes for 62 patients
- mRNA(RNA-seq) data containing 15121 genes for 62 patients

Quality of Proteomics and Phosphoproteomics Data

For both ovarian cancer and breast cancer tissues, proteome and phosphoproteome of training data were acquired using mass spectrometry experiments with isobaric mass tags (iTRAQ: Isobaric Tags for Relative and Absolute Quantification); while TMT (Tandem Mass Tags) platform was used to produce testing datasets. All of the proteome and phosphoproteome intensity tables were quantified with the CPTAC common data analysis pipeline (CDAP). As the output of this quantification pipeline, log ratios of protein levels between regular samples and reference samples in the same multiplex were calculated, including only peptides mapping unambiguously to a single gene. With the multiplex design of this experiment, substantial missing values were produced in the final data matrix. For example in the training data, proteomic and phosphoproteomic data of ovarian cancer presented 17.96% and 58.18% missing values. Similarly, proteomics and phosphoproteomics of breast cancer presented 8.90% and 49.65% missing values, respectively (Figure S1). Proteins or phosphosites with higher missing values are much less informative in either training prediction model or evaluating performance, and features with missing rate over 30% were removed from the data sets in the challenge.

Description of CPTAC Common Data Analysis Pipeline (CDAP)

Raw files retrieved from the CPTAC Data Coordinating Center were verified with MD4 checksums. File conversion was performed with the National Institute of Standards and Technology (NIST)- developed converter ReAdW4Mascot2.exe. MS-GF+ was used for peptide identification. Training MS1 data processing used for label-free quantitation was calculated with NIST-ProMS, a program developed by NIST. The proteome database was integrated from RefSeq H. sapiens (build 37) and the sequence for S. scrofa (porcine) trypsinogen. The resulting output for the dream challenge is a matrix where each row represents genes/genes_phospho site and columns represent tumor samples. Detailed documentation is also available at NCI-CPTAC CDAP portal (cptac-data-portal.georgetown.edu).

Batch Effect

Different techniques utilized to generate the CPTAC discovery and confirmatory data resulted in a batch effect between training and testing data for both proteomics and phospho-proteomics data. In order to remove this batch effect, different techniques were implemented. First, each sample in the training and testing data was median normalized (samplewise). However, this simple normalization was not able to eliminate the batch effect in the data (Figure S2). As the next normalization strategy, after sample-wise normalizing of both CPTAC discovery and confirmatory data, genes in the confirmatory data were normalized to have the same mean and standard deviation of the genes in the discovery data (Genewise). This simple normalization could substantially eliminate the batch effect between the two datasets (Figure S2). Given the satisfying performance of Genewise normalization, this strategy was adopted during the challenge; confirmatory data was normalized accordingly and uploaded on Synapse. However, given that the discovery data was uploaded on Synapse months before confirmatory data became available, the raw and original version of discovery data was available on Synapse. In summary, to eliminate any batch effect in the data, participants needed to apply a sample-wise normalization on discovery data.

Best Performing Methods

Winning Method. For the protein prediction task, the winning team used a weighted average of three models (Figure 1C) which will be referred to as protein proxy model, interaction model and pan-cancer model.

Protein Proxy Model. This model is based on the observation that protein and transcript levels are correlated, and simply uses the transcript level of a given gene as a proxy for its protein level. Missing values positions are replaced with the gene average across non-missing samples. This model has several limitations including that it assumes no differential translational regulation and degradation, and it disregards interaction between genes.

Interaction Model. Since different genes are regulated differently, individual models were built for each gene using random forests with maximum depth of 5 and 100 trees. The response variables for training are the non-missing observations across all samples, and the values for all genes are used as the training features to take into account gene-gene interactions.

Pan-Cancer Model. The performance of the interaction model is limited by the sample size. The training data only contain 77 and 174 tumors for breast and ovarian, respectively. This is a relatively small sample size, but when combining all the samples from both tissues, a better performance was achieved as the majority of genes have similar regulation across different tumor tissues (Kosti et al., 2016; Wang et al., 2013).

For the phosphorylation prediction task, the proxy model was changed to use protein levels instead of transcript levels and a fourth model was added (Figure 1D).

Phosphorylation Proxy Model. This model is based on the observations that protein and phosphorylation levels are correlated albeit only modestly, and simply uses the protein level of a given gene as a proxy for its phosphorylation level. This model assumes that for any given gene a constant fraction of the proteins is phosphorylated.

Phosphosite Correlation Model. The levels of multiple phosphorylation sites from the same protein are not independent. The biological rationale behind this model is that if a protein is phosphorylated, it is likely that multiple phosphosites are phosphorylated simultaneously. In addition, for technical limitations it is sometimes not possible to distinguish two phosphosites that are very close in the linear sequence so that they are in the same peptide after digestion and no fragment peaks are observed from fragmentation between them. Therefore, a phosphorylation site is correlated with other phosphorylation sites on the same protein. The winning team utilized this and calculated the weighted average prediction from all phosphorylation sites of the same gene as the multi-site prediction.

Other Best Performing Methods for Proteomics Sub-challenge

We next sought to further improve the prediction by organizing a collaboration round between the top ranked teams (Figures 1C and 1D; Supplemental Notes). In addition to the winning team, these included:

Team HYU trained a Random Forest using transcript levels as proxy. For each protein, cross-validation was performed to build individual prediction models. The missing values present in the transcript data were imputed using k-Nearest Neighbors (kNN) with k=10. Features were standardized and filtered based on KEGG signaling pathway and human PPI network from Human Protein Reference Database (HPRD) and their correlation with the responses. When a protein is present in the networks, all neighbors within a distance of two in either KEGG pathway or PPI network were selected. The set of features was further expanded by important genes from the pathways of mRNA surveillance, RNA degradation, RNA polymerase, basal transcription factors, cell cycle, protein processing in endoplasmic reticulum, and microRNAs in cancer. These pathways are assumed to play important roles in regulating gene expression.

Team DEARGENpg predicted protein levels by building a model on a group of proteins. For a single protein, there might be only a small number of samples available in the training set, but for a group of proteins, the data can be combined, partially overcoming the issue of small sample size. A group was defined by a set of proteins whose coding genes show high correlation with each other, and the group size was limited to 16 proteins for breast and 10 proteins for ovarian, yielding 628 and 707 protein groups, respectively. Each group was trained using several types of features: (i) the copy number and transcript level for a protein in a group, (ii) Pearson score-based features, (iii) gene metadata related to protein levels; codon bias, GC count and folding energy of each protein, and (iv) clustering label (STAR Methods). And an ensemble model was constructed using the average of XGboost, Extra Tree and Random Forest.

Team DMIS_PTG trained models for each protein using LASSO regression. Features were selected based on PPI networks, biological pathways and LASSO. The following PPI networks have been used: (i) Protein-Protein interaction network (BioGRID) that include protein and genetic interactions; (ii) Gene predictory network: a task-specific gene predictory network was built based on the given transcript level data using Boosted Tree based gene predictory NETwork (BTNET(Park et al., 2018)); (iii) Protein complex network (CORUM): the team assumed that the proteins in a single complex shared edges with each other, and the relationship was treated as a network. These three networks were combined into a single union network. Finally, only genes directly connected to a target protein in the final network were selected as input features for LASSO.

Second Best Performing Method for Phosphoproteomics Sub-challenge

Team Ardigen used a two-stage approach where the model used in the first stage captured the main effect and the model in the second stage captured corrections to the main effect (e.g. interactions among genes), by fitting to the residuals of the first stage model. To this end, the model in the first stage was in most cases (see Supplemental Information for details) the phosphoprotein proxy model, whereas in the second stage a forward-feature selection procedure was applied using all protein levels. The forward-feature selection procedure was terminated when the cross-validation score of the resulting model dropped with adding a new feature. Least Angle Regression (LARS) was used as the second stage model. In the final solution an ensemble of models over different cross-validation fold splits was used.

Ensemble Method - Proteomics Sub-challenge

To improve the prediction performance, an ensemble algorithm based on the models of the top four teams (i.e., Team Hongyang Li and Yuanfang Guan, Team DMIS_PTG, Team DEARGENpg and Team HYU) was derived. By analyzing the 5-fold cross validation results of these models on the training data, the prediction correlation of each protein was calculated. For each protein, the correlation scores were used as the stacking weights of these top four models (hereafter referred to as the individual ensemble model). To estimate the overall performance, the average correlation of all proteins was calculated and used as the weights for all proteins (hereafter referred to as the global ensemble model). For ovarian cancer, we observed a significant improvement (i.e., correlation equal to 0.56) of the global ensemble model, compared with the best performer in the challenge final round (i.e., correlation equal to 0.53), $P < 0.05$. There was also a reduction of NRMSE from 0.19 to 0.18. However, the improvement of the global ensemble model was very marginal in breast cancer with a correlation change from 0.505 to 0.506, $P > 0.99$.

Ensemble Method - Phosphoproteomics Sub-challenge

Within the collaborative phase for phosphoproteomics sub-challenge, we used an ensemble model based on the models developed by the winning team and Ardigen. The final solution is obtained as a weighted average of four models: three of the models used by the winning team (Protein proxy model, Interaction model, and Pan-cancer model) and the Ardigen's model. The four weights of these models are determined by optimization for the holdout set score (we used 20% randomly selected samples as the holdout set). During the collaborative phase we introduced slight modifications to the models: (i) the Phosphosite correlation model is disregarded from the final solution as it does not improve the score; (ii) cancer type is included as a binary variable for the ovarian cancer to the Pan-cancer model (for the breast cancer it does not improve the score and is not included); (iii) Ardigen's model modifications are discussed in the Supplemental Information (removal of the CNA and mRNA variables from the forward-selection procedure is the most important modification). For both cancer types we observed a significant improvement with respect to the final round of the challenge. The final prediction correlations are 0.37 for the ovarian cancer and 0.48 for the breast cancer (NRMSEs are 0.20 and 0.16, respectively).

Team HYU Method

The prediction of protein levels based on RNA/DNA information is treated as a regression problem where the protein levels of a particular protein is modeled as function of RNA/DNA levels of all the proteins.

Prediction of protein expression can be challenging due to complex regulations at multiple levels (e.g., post-transcriptional, post-translational), and mRNA measures are not always available for corresponding proteins. To address these problems, predictive models were built relying on two assumptions. First, the expression of a protein is modeled as a function of genes belonging to the same biological pathways or interacting with such pathways. Second, high order interactions are considered in the regression model for certain proteins due to complex pattern of regulations. Specifically, features (RNAs) are selected based on the Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa, 2008) pathway database and the Human Protein Reference Database (HPRD) human protein-protein interaction (PPI) network(Keshava Prasad et al., 2009). Random Forests was utilized to model the level of each protein as a function of the features.

Only RNA-seq data was used as predictors to predict protein level. Two sources of protein level data are available for some ovarian cancer patients. In such cases, the average of the two measures was used as response variable. For both breast cancer and ovarian cancer data sets, missing values were present in RNA data and were imputed with the k-Nearest Neighbors algorithm (kNN) using k=10. RNA-seq data was gene-wisely normalized to z score (i.e., mean zero and unit variance).

Features were filtered based on signaling pathways from the KEGG database and human PPI network information. Since inclusion of non-informative features can greatly affect the prediction performance, our purpose was to exclude as many irrelevant genes as possible based on prior biological knowledge, while retaining most important predictors.

The networks of KEGG signaling pathways were obtained through R package graphite(Sales et al., 2012; Sales et al., 2019), while the human PPI network was obtained from HPRD(Keshava Prasad et al., 2009). Given the two network topologies, all genes belonging to the second order neighborhood of a given protein in either network were selected. This set was further refined by adding genes in pathways of mRNA surveillance, RNA degradation, RNA polymerase, basal transcription factors, cell cycle, protein processing in endoplasmic reticulum and microRNAs in cancer, which are assumed to play important roles in regulating translation. Given the large number of genes selected based on this criteria, the set of features was refined based on the PPI network. In particular, only genes with degrees greater or equal to 10 in the PPI network were considered. Finally, this set was modified by adding genes whose Pearson's correlation coefficient with the protein of interest was greater than 0.3.

Random Forest based models (RF) were compared with a baseline method using 5-fold cross validation, in which the RNA level is directly used as a proxy of corresponding protein level. For all proteins RF delivered the best performance, or which do not have corresponding RNA data available, Random Forests were used. The baseline model was used for the remaining proteins.

Team DEARGENpg Method

To predict protein level based on RNA data, proteins were first grouped in different clusters and a predictive model for each group of proteins was utilized. This strategy borrowing information across proteins was adopted in order to increase the power of the predictive model. Specifically, two different grouping methods were utilized, one based on proteins correlation and the other based on alphabetically divided group. During the community phase, the method used to group proteins was further improved.

Pearson Score Based Selected Features

For both CNA and mRNA data, the Pearson's correlation between a group's vectorized protein level and each feature was computed, and then only the top 300 features (genes) with highest correlation were considered in the predictive model.

Gene Metadata

As metadata,codon bias count, GC count and folding energy of each protein and corresponding genes were considered.

(i) Codon count based on codon bias: top five codons which were best-expressed in humans were selected. Information about the percentage of the selected codons in each gene sequence were considered. This information is important since genes containing codons that are frequently used in the specific organism can be more stably translated to proteins.
(ii) GC count: When SNP occurs at G or C base position, it usually does not lead to mutation. Also AT-rich sequences within the gene could cause premature transcriptional termination and reduced mRNA levels, which affects protein levels. Based on these biological backgrounds, DEARGENpg hypothesized that the GC content in the gene sequence will eventually affect protein levels and calculated the ratio of GC-base in the entire sequence.
(iii) Folding energy: Another important factor for stable protein levels is folding energy of protein. Lower folding energy is better for proteins to be folded into stable secondary structure. The propensity of the 5-end of the mRNA, called Kozak sequence in eukaryotes cases, plays a major role in protein folding and since that Kozak sequence is available for human. The percentage of similarity between the 5'-end region of each gene and the human kozak sequence was calculated and included that percentage into the gene metadata.

Clustering Label

Samples were clustered based on CNA and RNA data. CNA and RNA data were combined into one matrix and then collapsed to two dimensions using principal components. Using this reduced data, samples were clustered into 2 clusters via K-means clustering. For breast cancer data, clustering result lead to similar classification as PAM50 breast subtypes label of each patient (i.e., basal-like & HER2 and Luminal A & Luminal B).

Grouping Using Simplified Final Round Method

During the community phase, a simplified version of the model was used. Protein groups were identified only for making protein group, protein-protein correlation based grouping method was used, and not the alphabetical group. For each protein, only the top 350 features with a correlation higher than 0.1 were considered in the regression modelInstead of selecting top 300 features with the highest correlation, a correlation threshold of 0.1 was utilized and all features with the maximum feature number as 350. Also, unlike the previous method of stacking three methods (XGboost, random forest, tree regressor) for model training, the average of three methods was used. Every group was trained using several types of features: (i) CNA/mRNA level of coding gene of protein in a group, (ii) Pearson score based features, (iii) gene metadata related to protein levels; codon bias, GC count and folding energy of each protein, and (iv) clustering label.

Team DMIS_PTG Method

In the Proteomics sub-challenge, protein level was predicted based on genomics and transcriptomics (i.e. RNAseq and copy number variation data). Since the number of features was much larger than the number of samples, feature selection was necessary to build a reliable prediction model. Specifically, features were selected as follows:

(i) a biological union network was constructed based on protein-protein interaction database, inferred gene predictory networks, and protein complex information. This final network was utilized in order to select features for each protein. In particular, only genes directly related to a target protein in the union network were considered as features. For this set of genes, both CNV and RNA data was utilized as predictors.
(ii) MSigDB database(Liberzon et al., 2011) was considered to derive meta genes. For each pathway, the median mRNA values across genes in the pathway was considered as feature.
(iii) Although prior biological knowledge was utilized to reduce the number of features, the number of samples was still insufficient to avoid the curse of dimensionality issues caused by the high-dimensional feature space. To address this, a penalized linear regression such as LASSO was utilized to model each protein as a function of features.

It is important to notice that, for each protein, a different set of features was selected. The final model based on LASSO regression was estimated based on 5-fold cross validation.

First, Z-score normalization was applied to the mRNA expression values. Between microarray and RNAseq data, to train the prediction models, microarray data was utilized since it involved a higher number of samples.

Considering the insufficient number of samples, using all the CNA and mRNA levels as features was infeasible. A combined biological network was first constructed using the following resources:

- Protein-Protein interaction network from BioGRID(Stark et al., 2006)
- Gene predictory network: task-specific gene predictory network was constructed based on mRNA data using BTNET(Park et al., 2018) - a Boosted Tree based gene predictory NETwork. BTNET trains a tree-based model and considers the feature importance in the model as the weight of the predictory relationships between a feature gene and a target gene.
- Protein complex network: Some proteins gather and form a single protein complex with each protein complex having its own unique function. Protein complex information from the CORUM database(Ruepp et al., 2008) was leveraged to form a protein complex network. In particular, the network was built by connecting all proteins in the same protein complex.

The three networks were combined into a single network by aggregating edges which were present in at least one network. Next, genes directly connected to a target protein in the combined network were used as feature in the regression model. In particular, both CNA and mRNA levels of the selected genes were considered as features. The average of the number of the selected gene features for all protein models was approximately 20 for both cancer types. As mentioned previously, in addition to raw gene expression and copy number variation data, gene-set level information was considered as a feature. In particular, for each pathway, the median mRNA values across genes in the pathway was considered as feature. For this analysis, only pathways in the MSigDB database with all genes observed in the RNAseq data were considered. This filtering criteria resulted in 700 pathways being included in the final predictive model. Finally, the LASSO algorithm was utilized to model each protein as a function of the selected features.

Team Ardigen Method

In the phosphoproteomics sub-challenge, the task was to predict phosphoprotein levels based on genomic, transcriptomic and proteomic data. Due to a large number of variables that could be used for this purpose (above 30 000) and a small number of training samples (around 100), as well as possible batch effects, model generalization was a serious issue, a two-stage procedure was used.

In the first stage, given a target phosphoprotein, two models relating directly to the target phosphoprotein were considered: (i) a model on its parent protein level (equivalent to phosphoproteomic proxy model in the main text) and (ii) a model on its parent protein and its mRNA level (we apply LARS). Of the two models, the one which yields a higher 5-fold cross-validation (CV) score was selected. In the absence of the parent protein level among the variables, the best of the following three univariate models was selected: (i) a model on the mRNA level corresponding to the target phosphoprotein, (ii), (iii) models on the protein level with the highest correlation to the target phosphoprotein level selected among proteins known as transcription factors (labeled "hc: TFs"; http://www.tfcheckpoint.org/) (ii) and kinases (labeled "hc: kinases"; http://kinase.com/human/kinome/) (iii).

In the second stage, a model was fit to the residuals of the first stage model. All protein levels were considered as variables and a forward feature selection procedure was used. At each step, the variable with the highest correlation was added to the residuals of the current model. The procedure terminated when a decrease in the CV score was observed. Note that we may have different variables in the model for each train-test split of the cross-validation procedure. Finally, an ensemble (arithmetic average) of models trained per cross-validation fold splits was used. For ovarian (breast) cancer, an ensemble obtained by 13 (7) runs of the procedure with 5 folds was used, which yielded 13 x 5 (7 x 5) models used to make the prediction. For imputing missing values, mean imputation was used.

During the challenge, the same hierarchical two-stage approach was used. However, in the collaborative phase, several modifications were introduced, which improved the model performance: (i) the test-set mRNA levels have been normalized to have the same mean and standard deviation as in training set; (ii) in the second stage of the procedure, CNA and mRNA variables were removed and only protein levels were considered; (iii) the auxiliary model used by the team to predict generalization properties of stage 1 models was disregarded and the best of the considered models was selected (this model was built on the JHU and PNNL results relating to Proteomics sub-challenge); (iv) only the CV score was used for model selection (in the challenge, models with more variables were penalized); (v) all CNA variables we1re disregarded (also from the first stage of the procedure).

Baseline Methods

For predicting proteomics data, we used RNAseq for breast and microarray for ovarian, and trained a standard Elastic net model with 5 fold cross validation and $\alpha = 0.5$. For each protein, we repeated over 10 iterations, and at each iteration, we obtained an optimal λ parameter. Finally, we fit the average lambda (over the 10 iterations) to all available samples and saved the model.

For the prediction of phospho-proteomics levels, a random forest resulted in a better performance compared to lasso model and, therefore, was considered as the benchmark. Specifically, each phosphosite was modeled as a function of the protein level of all the proteins via random forest. To reduce the dimensionality of the problem, proteins were first selected based on the training data before being utilized in a random forest model. In particular, for each phospho-site, only the top 200 proteins more correlated to the phospho-site were considered in the predictive model. In the random forest model, a total number of 1,000 trees were considered and at each node a total number of N=p variables with p being the total number of predictors were sampled and proposed for the splitting rule. The predictive model was trained on the training data and, then, directly utilized to predict the phospho-level of samples in the test data.

Scoring

To evaluate the performance of different methods, the Pearson's correlation between predicted and actual values was chosen as the primary metric for the following reasons: (i) since in precision medicine patients are stratified based on biomarker level, the relative level of a specific protein (if used as a biomarker) across patients is important. (ii) Metrics which emphasize effect size (such as sum of residuals) require the training and test features normalized in the same way. However, since users do not necessarily have access to the training data, the Pearson's correlation is more appropriate to evaluate the true performance of a prediction model.

Proteomics Sub-challenge: Prediction of Protein Levels Based on mRNA Levels

These models were evaluated in two novel, unpublished held-out datasets of ovarian and breast cancer. Only proteins with less than 30% missing values in both training and testing data were considered for the analysis (i.e. 5220 proteins for ovarian and 8649 proteins for breast). For model evaluation, first, the Pearson's correlation between observed and predicted levels across all samples was computed for each protein. Then, the mean of correlations across proteins in the test dataset was considered as the final evaluation score. When a tie was observed, NRMSE for all proteins was utilized to select the best performing team.

Phosphoproteomics Sub-challenge: Prediction of Phosphosite Levels Based on Protein Levels

Only phosphosites with less than 30% missing values in both training and testing data were considered in the challenge (i.e.,1318 phosphosites for ovarian and 4907 phosphosites for breast). We first computed the Pearson's correlation between observed and predicted phosphosite levels across all samples for each phosphoprotein. We then took the mean correlations of phosphoproteins in the test data set as the final evaluation score. If there is a tie, we further used NRMSE for all phosphoproteins to select the winner.

Tie Breaking

Bootstrapping was utilized in order to identify the best performing team. Specifically, from each team, we obtained a *p-by-n* matrix *X* of predictive values, with *n* being the number of samples and *p* the number of variables. For each bootstrap iteration, we sampled with replacement *n* columns of X. We then computed the two metrics (i.e., correlation and NRMSE) using the bootstrapped set of samples. We performed 1,000 bootstrap iterations and assessed significance using two criteria:

1. Confidence Intervals. Specifically, for each team, 95% confidence interval was computed across bootstrap iterations. Then, two teams were declared statistically different when the corresponding confidence intervals did not overlap.

2. Bayes Factor. Given two teams, we estimated the Bayes Factor using the same strategy as in (Menden et al., 2019).

Proteomics Sub-challenge

The four teams scoring the best performance based on both Ovarian and Breast Data, i.e., *Hongyang Li and Yuanfang Guan*, *DMIS_PTG*, *Hyu* and *DEARGENpg*, were considered to assess the top performing team. In particular, the top performer (*Hongyang Li and Yuanfang Guan*) was the same for both Ovarian and Breast data. Therefore, we were interested in assessing whether *Hongyang Li and Yuanfang Guan* significantly outperformed other teams. Figure S5 shows the 95% confidence intervals for the four best scoring teams in Proteomics sub-challenge. As shown, for ovarian cancer data, a tie was observed between teams *Hongyang Li and Yuanfang Guan* and *DMIS_PTG*. In fact, the two teams resulted in overlapping 95% confidence intervals of both metrics (i.e., correlation and NRMSE) On the other hand, the Bayes factor always favored the best scoring team (BF=infinity). In fact, the correlation metric of *Hongyang Li and Yuanfang Guan* was greater than the correlation of the other teams for all bootstrapped samples. In conclusion, based on ovarian data, a tie was observed between *Hongyang Li and Yuanfang Guan* and *DMIS_PTG* based on 95% confidence intervals. For Breast cancer data, as shown by the 95% confidence intervals in Figure S5, *Hongyang Li and Yuanfang Guan* significantly outperformed the other teams. Also, the Bayes factor always favored *Hongyang Li and Yuanfang Guan* over other teams (BF=infinity). In conclusion, for this sub-challenge the best performing team was *Hongyang Li and Yuanfang Guan*.

Phosphoproteomics Sub-challenge

For this sub-challenge, *Hongyang Li and Yuanfang Guan* and *Ardigen* were the best performing teams based on both ovarian and breast data, with *Hongyang Li and Yuanfang Guan* being the best performer. Figure S6 shows 95% confidence intervals of these two teams in phosphoproteomics sub-challenge for ovarian (a, b) and breast (c,d) data. As shown, for both cancer types, *Hongyang Li and Yuanfang Guan* significantly outperformed the other team in terms of both correlation and NRMSE metric. Also the Bayes factor favored team *Hongyang Li and Yuanfang Guan* over the other team *(BF=infinity)*. In conclusion, *Hongyang Li and Yuanfang Guan* were the best performers in the phosphoproteomics sub-challenge.

Pathway Analysis of Protein Prediction Performance

In order to identify higher predictable pathways, we implemented the following test. Proteins were sorted from the most predictable to the least predictable based on the prediction correlation. Specifically, let $\{g_1, g_2, \ldots, g_G\}$ be the sequence of sorted proteins. For each pathway M containing n_M proteins, the following statistics were computed:

$$KL(i) = \sum_{j:(g_j \in M)\&(j \le i)} \frac{1}{n_M} - \sum_{j:(g_j \notin M)\&(j \le i)} \frac{1}{G - n_M}; \quad i \in \{1, \ldots, G\} \qquad \text{(Equation 1)}$$

The final test statistics was derived as the maximum deviation from zero of the sequence $\{KL(i)\}_{i=1}^{G}$. The significance of pathway enriched was assessed using permutation techniques. Specifically, for each permutation π, proteins were randomly reordered and Equation 1 was computed. By repeating this step $\Pi = 10,000$ times, we obtained scores $\left\{\{KL(i,\pi)\}_{i=1}^{G}\right\}_{\pi=1}^{\Pi}$ which were utilized to derive p-values and adjust for multiple comparison. Before deriving false discovery rates, scores $\{KL(i)\}_{i=1}^{G}$ and $\left\{\{KL(i,\pi)\}_{i=1}^{G}\right\}_{\pi=1}^{\Pi}$ were properly adjusted for pathway size using the strategy illustrated by Subramanian et al., 2005.

Identifying the Common Regulators

The team performing the best in both challenges utilized a method based on random forest to predict the abundance of each protein as a function of gene expression data. For each predictor, random forest returns a measure of importance defined as the summation of node impurities across all nodes utilizing a particular predictor for the splitting rule divided by the total number of trees in the random forest model. In order to derive a final list of regulators, a proper cutoff value for importance scores must be derived. This cut-off can be derived via permutation techniques. Let $\{I_{j\to k}\}$ be the importance score corresponding to the regulatory event ($j \to k$). To derive a proper cutoff value for importance scores, we utilize the following permutation-based procedure:

(a) For $b \in \{1, \cdots, B\}$, with B being the number of permutations:

(a.1) For any target protein k, we first permute its sample order and fit random forest to predict the abundance of protein k based on the expression of all genes. Repeating this procedure for each permutation, we obtain a list of importance scores $\{I^b_{j\to k}\}_{b=1}^{B}$

(b) For each threshold ι, we compute

$$f(\iota) = \frac{\frac{1}{B}\sum_{i=1}^{B}\sum_j 1\left(I^b_{j\to k} > \iota\right)}{\sum_j 1\left(I_{j\to k} > \iota\right)} \qquad \text{(Equation 2)}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to one if event × occurs and zero otherwise. $f(\iota)$ can serve as an approximation of the false discovery rate (FDR). In the following numerical studies, we use $\iota 0 = \min\{\iota : f(\iota) \le 0.001\}$ and declare an edge between j and k if $I_{j\to k} > \iota 0$.

Essentiality of Common Protein Predictors

We considered the top 83 common predictors of 10% proteins in breast and ovarian tissues, and checked their essentiality with the BROAD DepMap's Combined RNAi (McFarland et al., 2018) and CRISPR-Cas9 Avana screen (Meyers et al., 2017) (Figures S10 and S11).

Top common predictors were enriched in the set of genes associated with cancer survival. We reasoned that genes regulating a larger number of proteins were more likely to be essential for cancer cell progression (Macneil and Walhout, 2011) and thus having larger impact on disease outcomes. Subsequently, genes were ordered based on the number of proteins that they regulated, and the top 10% (i.e., 824 for breast and 516 for ovarian cancer) were selected for downstream analysis. Among the 824 top predictors of breast cancer, 96 showed association with patient survival outcomes in Cox survival analysis adjusted for age, gender and cancer subtype (FDR < 25%), which corresponded to a 1.28 fold (P = 0.038, Fisher's exact test) enrichment compared to the rest of the genome (Table S3). Similarly, among the 516 common predictors in ovarian tumors, after correcting for age, stage and grade, 6 out of the 516 predictors were found associated with overall survival, whereas only 14 out of the 14605 non-common predictor genes were correlated (12 fold enrichment; $P = 3.96\text{x}10^{-5}$, Fisher's exact test).

Survival Pathway Enrichment Analysis with Real and Predicted Data

We first conducted Cox proportional hazard regression to assess the association between each (phospho)protein with overall survival (OS). Then, we leveraged the KEGG pathway database to compare if the p-values for (phospho)proteins in a pathway were significantly smaller than those outside the pathway via one-sided Wilcoxon rank test. This comparison was conducted among top 55 well-predicted KEGG pathways, selected by comparing the correlations between true phosphoprotein and prediction scores in independent confirmatory data. At FDR < 0.05 control, neither real phosphoprotein data (from 69 training set) nor real protein data (from 105 prediction set) identified any significant KEGG pathway; predicted phosphoprotein scores in 105 prediction data, using ensemble model trained in 69 samples, significantly identified a KEGG pathway associated with survival. This pathway was at *P* < 0.05 in 69 training set for association between real phosphoprotein and OS and between real protein and OS. None of the pathways were identified based on the proxy prediction model. The significant pathway: ‘‘Vascular muscle contraction’’ contained a total of 25 phosphoproteins in our prediction data set. We further performed PCA among the 25 phosphoproteins, and evaluated the associations between the top 3 principal components (PCs). The top 3 PCs explained 65%, 10% and 8% of the variation in the data with the 2nd and 3rd PCs being associated with OS. A PC-based survival score was created by combining the 2nd and 3rd PCs (Figure S12).