Credentialing individual samples for proteogenomic analysis

Wei Zhao¹*, Jun Li², Rehan Akbani², Han Liang², Gordon B. Mills¹

¹Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

²Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

* To whom correspondence should be addressed:
Wei Zhao: Tel: (+1)713-563-4223, Email: wzhao3@mdanderson.org

Running Title: Credentialing samples for proteogenomic analysis

Key words: Proteomics, genomics, reverse-phase protein array, mass spectrometry, data quality
Credentialing samples for proteogenomic analysis

Abbreviations

AUC, Area Under Curve; FFPE, formalin-fixed paraffin-embedded; FDR, false discovery rate; MS, mass spectrometry; PMC: protein-mRNA correlation-based; ROC: Receiver Operating Characteristic; RNA-seq: RNA sequencing; RPPA, reverse phase protein arrays;

ACC, Adrenocortical carcinoma; BLCA, Bladder urothelial carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; CORE, Colon and rectum adenocarcinoma; DLBC, Lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and neck squamous cell carcinoma; KICH, Kidney chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LGG, Brain lower grade glioma, LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and paraganglioma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; SKCM, Skin cutaneous melanoma; STAD, Stomach adenocarcinoma; THCA, Thyroid carcinoma; THYM, Thymoma; UCEC, Uterine corpus endometrial carcinoma; UCS, Uterine Carcinosarcoma; UVM, Uveal melanoma
Credentialing samples for proteogenomic analysis

Summary

An integrated analysis of DNA, RNA and protein, so called proteogenomic studies, has the potential to greatly increase our understanding of both normal physiology and disease development. However, such studies are challenged by a lack of a systematic approach to credential individual samples resulting in the introduction of noise into the system that limits the ability to identify important biological signals. Indeed, a recent proteogenomic CPTAC study identified 26% of samples as unsatisfactory, resulting in a marked increase in cost and loss of information content. Based on a large-scale analysis of RNA-seq and proteomic data generated by reverse phase protein arrays (RPPA) and by mass spectrometry, we propose a protein-mRNA correlation-based (PMC) score as a robust metric to credential single samples for integrated proteogenomic studies. Samples with high PMC scores have significantly higher protein-mRNA correlation, total protein content and tumor purity. Our results highlight the importance of credentialing individual samples prior to proteogenomic analysis.
Introduction

mRNA amounts have long been used as a surrogate for amounts of its cognate protein, despite the fact that protein performs as the primary functional unit of biological process and it is thus critical to have an accurate estimate of protein levels. With advances in high-throughput genomic and proteomic technology, recent large-scale efforts have generated a wealth of transcriptomic and proteomic data on multiple platforms [1-5], however some key questions still remain about the quality of the proteomics data and of the concordance of proteomic and transcriptomic data [6]. Most studies observed only moderate correlations between protein and mRNA abundance; thus challenging the ability to impute protein levels and functions from RNA measurements [1, 7-9]. The rapidly increasing amount of available data provides an unprecedented opportunity to explore the relationship between the two molecular species in multiple tissue contexts. It is tempting to attribute the discrepancy between RNA and protein levels to multiple posttranscriptional processes that alter protein levels. Indeed, regulation of translation, protein modification, degradation and transport contribute substantially to protein levels that are unexplained by transcript amounts. However, the markedly higher correlations between RNA and protein in studies of cell lines [1, 7, 10, 11] where preanalytic conditions can be carefully controlled compared to patient samples where preanalytic characteristics and intra-tumoral heterogeneity are more likely to be problematic suggests that it is important to credential individual samples for proteogenomics analysis. Indeed, the complexity introduced by intra-tumoral heterogeneity, sample processing procedures, preanalytic processes, and technical constraints is largely underappreciated. Further, no systematic processes have been able to credential individual samples for proteogenomics analysis.

At present, transcript quantification technologies produce accurate and reproducible measurements of multiple RNA species. Orthogonal methods like RNA sequencing (RNA-seq) and hybridization-based platforms (i.e. DNA microarray) provide highly concordant results on fresh-frozen and even formalin-
fixed paraffin-embedded (FFPE) tissues [12]. Of note, this high degree of concordance only became possible as the result of a decade long effort to develop robust experimental and computational methodologies [13-17]. During this process, inter-platform concordance on matched samples was originally only modest (R~0.5 [18]) even when reproducibility using the same microarray platform had been achieved (R>0.9 [18]). Proteogenomic initiatives are now encountering a similar scenario with the added complexity of comparing two platforms encompassing proteomic and transcriptomic characteristics with intrinsic biases and different quality control metrics. Antibody-based functional proteomic profiling approaches, such as reverse-phase protein arrays (RPPA), can generate robust quantification using several hundred extensively validated antibodies (R>0.9 for replicates [7]). Non-targeted mass spectrometry (MS) technology has only been evaluated in small sets of samples for reproducibility [19, 20] with the performance exhibiting significant variability [21, 22]. Moreover, RNA-protein correlations have been shown to vary markedly across different tissues in multiple MS data sets [23]. Thus, interpretation of the observed discordance between mRNA and protein data requires extreme caution.

In addition to a set of technical constraints inherent to individual platforms, a basic assumption underlying integration of transcriptomic and proteomic data is that the samples analyzed are homogenous with the same amount and type of tumor and stroma, which may not hold even with rigorous sample collection protocols and is amplified by the large quantity of material required for technologies like MS. Moreover, it implies that simultaneous evaluation of quality control for RNA and protein data including across platform correlation is needed, while most if not all of the preanalytic characterization and quality control (QC) metrics only focus on individual data types. For instance, QC strategies for RPPA include visual examinations, correlation analysis on replicates in slides, and logistic models characterizing features such as dynamic range of signal intensity, signal-to-noise ratio, and dispersion of data distribution [24, 25]. MS-based studies incorporate different types of QC samples-
Credentialing samples for proteogenomic analysis

from a single peptide to a complex lysate-throughout the experimental workflow and select QC metrics in line with study designs to inform performance of both individual experiments and groups of experiments. QC metrics include the course of total ion current (TIC), the shape and number of peaks, and mass accuracy of specific spectra, amongst others. Here, we performed a large-scale analysis of transcriptomic data generated by RNA-seq and proteomic data generated by RPPA from 31 tumor types and by MS generated from breast and ovarian samples from The Cancer Genome Atlas (TCGA) by the Clinical Proteomic Tumor Analysis Consortium (CPTAC). We examined biological and technical factors that have the potential to influence protein-mRNA correlations. Finally we proposed a protein-mRNA correlation-based (PMC) score as a novel metric to credential individual samples for proteogenomic studies and demonstrated its utility in RPPA and MS data.

**Experimental Procedures**

**Description of gene expression, RPPA and MS data**

Level 3 RPPA data of TCGA samples of individual tumor types were obtained from TCPA database ([http://tcpaportal.org/tcpa/](http://tcpaportal.org/tcpa/)) and normalized by ‘replicates-based normalization’ (RBN, [1]), generating level 4 RPPA data. RNA-seq data of matched samples were obtained from NCI Genomic Data Commons ([https://portal.gdc.cancer.gov](https://portal.gdc.cancer.gov) [26]). Data generated by Illumina GA platform from UNC were adjusted to remove batch effects. The complete set contains 7303 TCGA samples from 31 tumor types that have both RNA-seq and RPPA data.

Level 4 RPPA data of 651 cell lines were downloaded from the MCLP data portal ([http://tcpaportal.org/mclp/](http://tcpaportal.org/mclp/)). Gene expression data of 1037 cell lines on Affymetrix U133+2 array platform were collected from CCLE ([http://www.broadinstitute.org/ccle](http://www.broadinstitute.org/ccle)). 369 cell lines have both mRNA expression data and RPPA data and were employed to investigate correlation of protein-mRNA pairs.
MS-based normalized global protein expression data from CPTAC breast and ovarian studies were retrieved from the CPTAC Data Portal (https://proteomics.cancer.gov/data-portal) and proteins detected in less than 50% of samples were removed. Paired MS and RNA-seq of 108 breast samples and 131 ovarian samples were analyzed.

**Correlation analysis between mRNA expression and RPPA data**

Level 4 RPPA data quantified 146 total proteins and 51 phosphoproteins and was compared with their mRNA counterparts. Antibodies targeting protein products of multiple genes were paired with the unique mRNA that provides highest correlation. For unique *cis* protein-mRNA pairs (*cis* represents matched pairs throughout), Pearson’s correlation was computed and was compared with *trans* protein-mRNA pairs (*trans* represents random pairs throughout) to determine statistical significance. ANOVA test and one-way hierarchical clustering analysis was performed to assess protein-mRNA correlation across tumor types.

**Identification of PMC markers for RPPA data**

Concordance of total proteins measured by RPPA and the mRNA counterparts of individual cancer type was analyzed and protein-mRNA pairs providing correlation coefficient of >0.6 in specific tissues were identified as disease-specific markers. 13 cancer types with >100 samples and >15 disease-specific markers were investigated with leave-one-out cross-validation to identify consensus PMC markers. Pearson’s correlation across PMC markers, aka PMC score, was computed for each sample.

**Association of PMC scores with preanalytic and clinical measurements**

Preanalytic measurements including within-gene and within-sample correlation on test protein-mRNA set (i.e. genes excluding PMC markers), total protein content and tumor purity were assessed in PMC-pass and PMC-fail samples by Student’s t-test. Total protein content was estimated by taking the median across all the antibodies for each sample [27]. Tumor purity of RNA samples was estimated using ABSOLUTE algorithm [28]. Receive Operating Characteristic (ROC) analysis was performed to quantify
the ability of using PMC scores to predict samples with low levels (<5% quantile) of protein-mRNA correlation, total protein content and tumor purity.

**PMC Analysis of MS and RNA-seq data**

PMC markers for CPTAC breast and ovarian samples were determined following the same pipeline as RPPA data and were analyzed by DAVID tools [29] for functional annotation. 28 breast samples were identified by the CPTAC study to have skewed distribution of protein abundance [9]. ‘Normal-like’ samples were identified by unsupervised hierarchical clustering of top 1000 differentially expressed genes of the remaining 80 samples and the true normal breast control samples. ROC analysis was performed to determine the cutoff threshold of PMC-pass and to assess the enrichment of two types of potential low-quality samples in the PMC-fail group.

**Results**

**Correlation analysis between mRNA and RPPA-based protein expression data**

We compared RPPA-based protein data for 7303 samples across 31 tumor types from TCGA study to RNA-seq-based mRNA expression data generated from matched samples, albeit from different but contiguous parts of the tumor. In an analysis of 146 matched (cis) total protein-mRNA pairs, 112 pairs demonstrated significant positive correlation compared to random protein-mRNA pairs at a false discovery rate (FDR) of 0.1 (Figure 1A-B), which is similar to the degree of correlation observed in our previous pan-cancer study of 11 tumor types [1]. The median correlation of 146 cis total protein-mRNA pairs was 0.336, which is similar to the CPTAC breast, ovarian and bowel data [9, 20, 30]. As expected, the 51 phosphorylated proteins assessed by RPPA showed much lower correlation with matched mRNAs (median Pearson’s R= 0.126). However, it is important to emphasize that although the median cis total protein-mRNA correlation was 0.336, the correlations for a subset of protein-mRNA pairs exceed 0.7 indicating that mRNA amount was likely the major determinant for protein amounts for these proteins.
Analysis of the same set of proteins in 369 cancer cell lines of 11 tumor lineages from the MD Anderson Cell Line Project (MCLP) [2, 7] demonstrated significantly higher correlation of matched total protein-mRNA pairs ($p=2.82 \times 10^{-6}$ median Pearson’s correlation = 0.514) (Figure 1A) despite the protein and RNA being isolated and analyzed at different centers. The higher correlation between proteins and RNAs in cell lines where sample management and preanalytic conditions could be carefully controlled compared to patient samples suggested that preanalytic conditions may contribute to decreased correlations between protein and RNA in patient samples. Further, in TCGA study, RPPA and RNA-seq were analyzed from separate but contiguous samples and thus heterogeneity in the tumor cells or in the amount and function of stromal cells could contribute to the decreased correlations observed.

The pattern of protein-mRNA correlations across TCGA samples showed significant association with tissue of origin ($p<1 \times 10^{-16}$, Figure 1C). While unsupervised clustering did not identify clear subgroups based on the patterns of cis protein-mRNA correlation (Figure 1D), three gynecological cancers – breast cancer (BRCA), uterine corpus endometrial carcinoma (UCEC) and serous ovarian carcinoma (OV) - and two lung cancers - lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD) co-clustered. Thyroid carcinoma (THCA), prostate adenocarcinoma (PRAD), pancreatic adenocarcinoma (PAAD), brain lower grade glioma (LGG) and cholangiocarcinoma (CHOL) had markedly lower concordance, with average cis protein-mRNA correlation <0.2. In THCA, PRAD, PPAD and LGG, ~50% of the total proteins measured by RPPA were correlated with cis mRNAs at levels comparable to the correlation with trans mRNAs (trans represents random pairs throughout), indicative of very low concordance (Supplementary Figure 1A). In contrast, OV and testicular germ cell tumors (TGCT) had higher concordance than other lineages (median $R = 0.44$ for OV and 0.40 for TGCT). Indeed, for ovarian samples, the cis protein-mRNA correlation was similar between tumors and cell lines (Supplementary Figure 1B). This could represent decreased heterogeneity of ovarian tumor samples or differences in preanalytic conditions. The large
tumor size of most ovarian cancers may have facilitated selection of samples with similar characteristics for both RNA and protein analyses.

We then examined features that contribute to cis protein-mRNA concordance. In pan-cancer analysis and analysis of individual tumor types, dynamic range (characterized by interquartile range) of protein expression was significantly associated with protein-mRNA correlation in all tissues except for THCA (Supplementary Figure 2). The dynamic range of mRNA amounts showed similar associations albeit to a lesser extent. Of note, the median expression level of protein or mRNA was not associated with protein-mRNA correlation. A high dynamic range, particularly of proteins, may decrease challenges with heterogeneity or preanalytic characteristics and increase the ability to detect mRNA-protein correlations.

Identification of correlation-based markers for sample quality control of RPPA data

Analysis of distribution of the cis total protein-mRNA correlations identified a 2-component Gaussian mixture model (Supplementary Figure 3A). This suggested the existence of a subset of highly correlated mRNA and proteins that could potentially be used for quality control of individual samples for integrated transcriptomic and proteomic investigation in terms of preanalytic characteristics. Therefore, we analyzed the set of proteins that were most highly correlated with cis mRNAs, and examined the utility of protein-mRNA correlation-based (PMC) scores as a metric for sample quality control. Across distinct tumor types, the sets of highly correlated proteins and mRNAs exhibited substantial differences consistent with alternative mechanisms of regulation of protein levels including dynamic range in different tissues. In the 21 cancer types with >100 samples available for analysis, the number of highly correlated protein-mRNA pairs (Pearson’s R > 0.6) ranged from 0 to 40 (Figure 2A, Supplementary Table 1). Analysis of diseases such as PRAD, colon and rectum adenocarcinoma (CORE), PAAD, stomach adenocarcinoma (STAD), kidney renal clear cell carcinoma (KIRC), THCA, and LGG yielded few highly
correlated protein-mRNA pairs, which may indicate limited power of this approach in these tumor lineages or alternatively marked challenges with preanalytic characteristics in these samples. Therefore, we integrated RNA-seq and RPPA data from 13 cancer types with >100 samples and in which >15 genes had highly correlated protein-mRNA levels, naming this the TCGA13 data set, and developed PMC markers in this pan-cancer cohort.

In TCGA13 data set, we performed a leave-one-out cross-validation analysis to identify a set of consensus PMC markers (Supplementary Figure 3B). Data of 12 tumor types were pooled as the training set and the remaining tumor type was left out as the test set. In the training set, highly correlated cis total protein-mRNA levels (Pearson’s correlation > 0.6) were used to select PMC markers. Then a PMC score, defined as the Pearson’s correlation across PMC markers for paired mRNA and protein data, was assigned to each sample in the training set. To determine the cutoff threshold for high quality samples, in the training set, permutation test was performed to estimate the empirical distribution of PMC scores on random protein and mRNA pairs. Samples with significantly higher PMC score than random pairs at FDR of 0.1 were designated high quality (PMC-pass). The same PMC markers and cutoff were then applied to the test set to estimate PMC score and to identify PMC-pass samples. The training-test procedure was applied to each disease type.

To assess the cross-validation performance, we compared two metrics, cis total protein-mRNA correlation for individual genes (within-gene correlation) and for individual samples (within-sample correlation), in the set of protein-mRNA pairs that were not selected as PMC markers, aka test protein-mRNA set, between PMC-pass and PMC-fail samples. In training and test sets, the two types of correlations were both significantly higher in PMC-pass compared to PMC-fail samples, which demonstrated that the leave-one-out analysis identified markers that allow robust improvement in protein-mRNA correlation across lineages (Supplementary Figure 4, Supplementary Table 2). In comparing PMC markers in the 13-fold cross-validation, substantial consensus was observed (Figure 2B).
While 34 genes were identified as PMC markers in at least one analysis, 30 genes were selected in >9 disease types and were defined as **consensus PMC markers** (Supplementary Table 3). Compared to the complete set of antibodies in RPPA, protein abundance of consensus PMC markers have a significantly greater dynamic range while exhibiting a similar median level (Supplementary Figure 5), which is consistent with the observed association between dynamic range and ability to detect protein-mRNA correlations.

**Consensus PMC markers identified samples with preanalytic challenges**

Based on the consensus PMC markers, we computed PMC scores on paired samples and on random protein-mRNA sample pairs. At FDR of 0.1, PMC scores of >0.386 were designated as PMC-pass (Figure 2C). Overall, 82.2% samples passed the test across cancer type. The pass rate ranged from 66.5% to 93.3% in different diseases (Figure 2D, Supplementary Table 4). To evaluate the performance of consensus PMC markers, we investigated within-gene and within-sample correlation, in test protein-mRNA pairs (genes that were assayed by RPPA and RNA-seq and were not selected as PMC markers). Correlation of random protein-mRNA pairs (**trans** within-gene correlation) and permutated sample pairs (**trans** within-sample correlation) were used as controls (Figure 3A-H, Supplementary Table 4). In TCGA13 pan-cancer data set, within-gene and within-sample correlation were significantly higher in PMC-pass samples (within-gene: $p=3.04\times10^{-10}$, within-sample: $p=1.21\times10^{-5}$), while no difference was observed in the control analyses. Next for individual disease types, we assessed within-gene and within-sample correlation as well as correlation of two types of **trans** protein-mRNA correlation. Again, while all tumor types had significantly higher **cis** within-gene protein-mRNA correlation and all except for SARC samples had significantly higher **cis** within-sample correlation in PMC-pass samples, no difference in **trans** correlation was observed in any tumor type.
Then we examined the relationship between PMC score and metrics that have previously been used for quality assessment in RPPA and RNA-seq data (Figure 3I-L, Supplementary Table 4). Total protein content defined as the overall levels across all antibodies in the RPPA set was adopted by TCGA study as a quality control measurement for RPPA data [1]. Tumor purity estimates based on somatic DNA copy number aberrations has been used as an indicator of tumor content to evaluate quality of DNA and RNA samples in TCGA [28]. In the pan-cancer dataset, total protein content and purity were both significantly higher in PMC-pass than PMC-fail samples (total protein content: p=0.00138, purity: p=0.0186). Within individual diseases, total protein content and purity were substantially higher in PMC-pass samples in 9/13 and 6/8 diseases respectively.

Consensus PMC markers developed from pooled patient cohorts provide a more comprehensive characterization of the complexity and diversity of cancer patients. Therefore, we predicted they would have more robust performance in identifying samples with favorable features than markers developed in a disease-specific context. To quantitatively compare the performance of consensus and disease-specific markers, we selected benchmarks of low-quality samples based on within-sample correlation, total protein content and purity. Samples with low level (<5% quantile) of the three measurements were considered likely be low-quality samples. With these criteria, Receiver Operating Characteristic (ROC) analysis demonstrated that the consensus PMC score was able to detect three types of low-quality samples, with Area Under Curve (AUC) of 0.793, 0.694 and 0.657 (Supplementary Figure 6A-C). Within individual tissues, compared to disease-specific PMC markers, consensus PMC markers provided more robust prediction of low-quality samples in most cancer types, especially for tumors where disease-specific markers have poor performance, such as ESCA, SARC and KIRP (Supplementary Figure 6D-F, Supplementary Table 5). ROC analysis of ability to predict samples with low within-sample correlation provided AUC >0.75 in 11/13 tumor types. ROC analysis of ability to predict samples with low total protein content and purity provided AUC>0.65 in 10/13 and 7/9 tumor types respectively.
Collectively, PMC analysis with 30 consensus marker genes and a cutoff of 0.386 following a proposed computational pipeline (Figure 4) could be used to identify samples with poor protein-mRNA correlation, low total protein content and low purity. The performance of consensus PMC markers was robust across 13 cancer types and was comparable or superior to markers developed for individual disease types.

**PMC markers for sample quality control of MS data**

We then assessed the performance of PMC analysis in a mass spectrometry (MS)-based protein analysis platform. The NCI CPTAC group employed MS with isobaric tags for relative and absolute quantitation (iTRAQ) to characterize global protein expression patterns in a subset of tumors from TCGA. We investigated protein-mRNA linkages in MS-based proteomic data from two cohorts, CPTAC BRCA and OV samples.

In the CPTAC BRCA study, MS data was generated for 111 breast samples, including 105 breast tumors, 3 tumor replicates and 3 normal breast control samples [9]. In more than half of the samples, MS detected protein products from 9204 genes, including 8688 genes with mRNA data available. In comparing protein profiles generated by MS and RPPA, median correlation of mRNA levels with cis total proteins quantified by RPPA for the 146 proteins assessed was 0.416 and by MS for the 8688 proteins assessed was 0.289 (Figure 5A). Overall, 118 proteins were measured by both RPPA and MS and provide comparable protein-mRNA concordance, including 28/30 RPPA consensus PMC markers (Figure 5B).

Intriguingly, unlike the antibody-based RPPA platform, low average mRNA or protein level was associated with poor protein-mRNA correlation, while the impact of protein dynamic range was more limited compared to that for RPPA correlations and was weakly and inversely associated with protein-mRNA correlation in MS analysis (Supplementary Figure 7A-C). We assessed protein-mRNA concordance of 28 RPPA consensus PMC markers that were detected by the orthogonal MS platform (Figure 5B).
Among them, 15 mRNA-protein pairs had highly correlated mRNA and protein levels as quantified by MS (R>0.6); another 8 pairs were moderately correlated (R>0.4); the other 5 pairs had low but statistically significant correlation (p<0.05).

We then identified MS-based PMC markers in CPTAC BRCA data following the same computational pipeline as RPPA analysis (Figure 4). 1068 genes were identified as having highly correlated protein and mRNA levels (R>0.6, Supplementary Table 6). Functional annotation clustering analysis showed no bias towards specific signaling pathways. The top clusters included oxidation-reduction process, mitochondrion, nucleotide binding, and cell-cell adhesion (Supplementary Table 7). PMC scores were computed for each sample and showed no significant difference across intrinsic subtypes or clinical/pathological subpopulations (Supplementary Table 8).

For most MS datasets, the number of protein-mRNA pairs assessed exceeds the sample size by one or two magnitudes. Therefore, instead of sample-based permutation, we employed supervised and unsupervised learning to estimate the cutoff for PMC-pass samples. In the CPTAC BRCA set, we adopted a supervised approach and used two types of low quality measurements as benchmarks. First, 28 samples had been identified by the CPTAC group to have skewed protein distribution which was potentially attributed to protein degradation [9]. Second, unsupervised clustering of the remaining tumors and three normal breast samples identified 10 tumor samples that clustered with the normal samples (Supplementary Figure 8). Given that none of the 10 tumors were characterized as normal-like by intrinsic subtype profiling by RNA-Seq, co-clustering with normal tissues was considered a potential indicator of low tumor content due to different parts of the tumor being analyzed for protein and RNA. Thus, in total, we defined 38 likely low-quality samples by these two criteria. PMC scores were determined using the 1068 MS-based markers. ROC analysis demonstrated that samples with low PMC scores were significantly enriched with the two classes of low-quality samples (Area Under Curve = 0.823) (Figure 6A-B). The cutoff threshold of PMC score was determined to allow 80% sensitivity to
identify high-quality samples, and identified 42/108 samples that fail the test, including 19 samples with skewed protein distribution and 9 samples that co-clustered with normal breast tissues. We then assessed within-gene and within-sample cis protein-mRNA correlations in PMC-pass and PMC-fail samples. PMC markers were removed from the total gene set generating a test set of 7620 protein-mRNA pairs. Within-gene and within-sample correlation were both significantly higher in PMC-pass samples (within-gene: p<1x10^{-16}, within-sample: p=4.08x10^{-7}), while the two types of trans correlations were not different between pass and fail samples (Figure 6C-F). Of the 108 tumor samples assessed, 63 samples were PMC-pass for both RPPA and MS data (dual PMC-pass). For the set of proteins quantified by both RPPA and MS that were not defined as either RPPA or MS PMC markers, inter-platform concordance of the proteins was significantly higher in the dual PMC-pass samples (median Pearson’s R=0.291) than in samples with at least one PMC-fail measurement (median Pearson’s R=0.166, p=0.02, Figure 6G). Similar to RPPA data, the PMC-pass set trended towards enrichment of samples with higher tumor purity (Figure 6H, p=0.066).

We further tested whether RPPA-based consensus PMC markers could be employed to credential MS samples. Based on the 28/30 RPPA-based consensus markers present in CPTAC BRCA data, a RPPA-based PMC score was computed for each sample. The RPPA-based score provided moderate correlation with MS-based score (R=0.497, Figure 7A) with a reasonable ability to identify low quality samples determined by MS-based markers (AUC=0.772, Figure 7B-C). Moreover, ROC analysis demonstrated that the RPPA-based score was able to detect samples with skewed protein distribution and normal-like samples (AUC=0.688, Figure 7D). Similar to MS-based markers, within-gene and within-sample cis correlations, but not trans correlations, were remarkably higher in RPPA-based PMC-pass samples and proteins quantified by MS and RPPA trended towards higher concordance between the two platforms in dual PMC-pass samples. (Figure 7E-I). Analysis of purity estimates showed no significant difference in
PMC pass and fail samples (Figure 7J). Importantly, the RPPA and MS analysis was performed on different portions of the tumor potentially contributing to the differences in scores.

We further tested the PMC score analysis in an independent MS data set from the CPTAC OV study. Samples from 106 patients had transcriptomic profiling by RNA-seq and proteomic profiling by MS, including 25 tumors analyzed by MS at two institutes. Both mRNA and protein expression were quantified for 7344 genes. We performed de novo identification of PMC markers following the same computational pipeline used in the CPTAC BRCA set (Figure 4) and adopted an unsupervised waterfall method [31] to estimate the cutoff threshold based on the inflection point in the PMC score curve, which classified 110/131 samples as the PMC-pass set (Supplementary Figure 9A). Consistent with the CPTAC BRCA set, the PMC-pass samples had significantly higher within-gene and within-sample cis protein-mRNA correlations (within-gene: p<1x10^{-16}, within-sample: 2.21x10^{-11}, Supplementary Figure 9B-E) and similar trans within-gene correlation. The trans within-sample correlation was higher in the PMC-pass set as well, but with lower significance (p=0.00022) than the within-sample cis correlation. A higher concordance of protein amounts determined by MS and RPPA was observed in dual PMC-pass samples (p=1.34x10^{-8}, Supplementary Figure 9F). Additionally, a higher purity score was associated with the PMC-pass set (p=0.0176, Supplementary Figure 9G). Therefore, in two independent MS data sets, samples with better correlated global protein-mRNA levels, higher purity and more concordant inter-platform proteomic quantification were identified by the PMC analysis.

Taken together, the PMC score appears to be a useful and robust metric to identify samples that are not well suited to proteogenomic analysis using either RPPA or MS data sets. The described PMC analysis approach could be applied to protein/mRNA data sets that have been generated, or even could serve as a low-cost solution to prequalify samples to be assayed by high-throughput technologies in proteogenomic studies (Figure 4).
Discussion

Recent large-scale multi-omics initiatives have characterized thousands of tumors expanding our understanding of information flow from DNA aberrations to RNA to protein. While emerging evidence is consistent with a complex linkage between mRNA and protein levels, there is a pressing need to minimize the effects of technical confounding factors and to evaluate sample quality for integrated proteogenomic studies. Further an ability to do this prior to performing sample preparation for RNA-Seq or MS would greatly decrease the costs and increase the information content obtained from such studies.

To address this issue, we analyzed RNA-seq and RPPA data across 31 tumor types from TCGA and confirmed that most cis protein-mRNA pairs have statistically significant but lower correlation in analysis of tumors as compared to cell line data where sample handling can be carefully controlled and heterogeneity is limited. This suggests that sample handling and heterogeneity of patient samples are likely contributing challenges to data interpretation. The same set of proteins quantified by MS and RPPA show similar protein-mRNA correlations in tumors consistent with the preanalytic challenges influencing both the RPPA and MS platforms. In the TCGA analysis different portions of the tumor (i.e. matched samples) were analyzed for RNA, RPPA and MS. The difference between tumor and cell lines underscores potential variation in ‘matched’ samples and the necessity to carefully review the preanalytic characteristics of clinical samples.

Intriguingly, different factors, likely independent of preanalytic characteristics, intrinsic to the analytic platform used likely contribute to protein-mRNA correlation in RPPA and MS sample sets. In RPPA data, dynamic range is significantly associated with protein-mRNA correlations, highlighting dynamic range as a critical feature to test for utility of antibodies in RPPA. While for MS, assay sensitivity is a more important determinant, as low mRNA or protein amounts are associated with remarkably worse
Credentialing samples for proteogenomic analysis

correlations. Indeed, previous MS and RPPA comparisons have shown that the antibody based RPPA platform has a greater sensitivity for at least a subset of proteins [32].

We reasoned that variation in correlations in a single sample for mRNA-protein pairs that were highly correlated across a sample set could be used to “qualify” individual samples as high quality including having similar tumor and stroma composition for proteogenomic analysis. Indeed, we were able to identify a set of proteins and RNAs that were highly correlated suggesting that protein amount is driven primarily by translation of existing RNAs. These protein-mRNA pairs were enriched for molecules that participate in normal cellular processes, a concept that the Human Proteome Map Project has also observed in normal human tissues [33, 34]. If this concept holds across multiple data sets, it would serve to identify samples that should be discarded from the analysis improving reliability of the data. We analyzed global protein-mRNA correlation as a metric for quality of paired samples because of its representative of two classes of preanalytic challenges: protein or RNA quality of individual samples and structural heterogeneity of matched sample pairs. Additional measurements such as total protein content and DNA-based tumor purity scores can also provide indirect information of sample quality. As expected, when we applied this approach to samples with matched RNA-seq and RPPA or MS data, samples with high PMC scores were associated with other correlates of high quality.

A lack of concordance in protein abundance measured by MS and RPPA has been observed in several proteomic studies [1, 7], which may be attributed to protein degradation, biological heterogeneity in the samples assayed by the two approaches or systematic difference between the two technologies. Our data revealed that it could be markedly improved by assessing samples credentialed by PMC analysis, which suggests that some of the discrepancies between the platforms are due to differences in sample quality or in composition of the samples assessed.

Protein-mRNA correlations represent an aggregation of distinct sources of variability. In addition to dependence of protein and mRNA levels on preanalytic challenges, samples from different tissues or
experimental batches and/or genes with markedly different dynamic range may exhibit opposite trends, thus leading to incorrect conclusions – an effect similar to Simpson’s paradox [23, 35]. In analysis of TCGA RPPA data set, a pan-cancer cohort that represents substantial diversity in tissue types and patient resources, we accounted for these potential confounding factors using computational approaches. To assess the effects of tissue types, we adopted a leave-one-out cross-validation approach and demonstrated the efficacy of PMC score both in the pan-cancer data set and within individual disease types. To address the challenges of batch effect, RBN-normalization was applied to all samples to ensure that dynamic range and median levels were comparable across batches. Dynamic range varies remarkably across different protein and mRNA amounts, which potentially would generate high across-gene correlation even when within-gene correlation is low [35]. Therefore, we assessed both within-gene and within-sample correlation in the test protein-mRNA set, and designed control approaches with trans protein-mRNA correlations. The data showed that when accounting for these factors, for most tumor types and samples, the PMC-pass set has superior performance in predicting high-quality samples.

In theory, PMC analysis is a simultaneous evaluation for sample quality of RNA and protein data as well as sample heterogeneity. Aberrations in any one of the components (i.e. quality of RNA or protein or sample heterogeneity) would result in a low PMC score and the approach is unable to distinguish the different causes of low PMC scores. RNA-seq data was used as the gold standard in our analysis for three reasons. First, the robustness of RNA-seq has been extensively validated. Second, specific to TCGA data set, a large proportion of mRNA samples were quantified by both microarray and RNA-seq and yielded highly concordant results [36]. Third, in TCGA13 dataset, ≥99.8% of samples have RNA integrity score (RIN) ≥7, and 50.9% of samples have RIN ≥8.5, suggesting that there was limited degradation of RNA in the TCGA sample set.
As described, PMC was validated for use after RNA and protein analysis has been performed. However, given the costs of each approach, it would be optimal to be able to prequalify samples. This could be done by analyzing a relatively small set of RNA and proteins (the consensus PMC marker is only 30 RNA protein pairs) using approaches such as Nanostring with barcoded antibodies[37] that allows concurrent analysis of RNA and protein or Nanostring combined with a Luminex or Multiple Reaction Monitoring (MRM) or similar platform (Figure 4). This could help to identify samples that are likely to provide high quality proteogenomics data. A more formal PMC analysis could be done following the generation of RNA-Seq and MS or RPPA data to eliminate any samples that failed later in the analysis pipeline.

The PMC score has some limitations. First, its performance varies in different tissues. In several tumor types, the overall protein-mRNA correlation is low potentially due to antibody set surveyed on the RPPA platform (Figure 2B), therefore, they were not included in our analysis. Second, PMC analysis based on RPPA consensus markers in RPPA data sets requires detection of 30 RNA and proteins for an efficient analysis. Third, a low PMC score can occur due to tissue heterogeneity or poor-quality protein or RNA analysis. PMC cannot distinguish between these mechanisms. Fourth, a surrogate quality control approach of analyzing the 30 PMC consensus markers on Nanostring or a similar platform that can assess both RNA and protein has not yet been developed. Fifth, setting cut-off values for high quality remains a challenge as it does with any continuous variable. When applying PMC analysis in MS data, the cutoff threshold was determined based on the assumption that the majority of samples are of high quality and provide high protein-mRNA correlation, so that the PMC score of matched sample pairs is significantly better than that of randomized pairs. Unfortunately the CPTAC BRCA MS data set had a limited sample size and many samples had evidence for preanalytic challenges including “skewed” quantification and likely normal contamination in the protein data set [9]. Nevertheless, it appears that the PMC score approach has the potential to greatly enrich the quality of proteogenomic analyses.
In conclusion, this study shed light on an untapped issue in quality assessment of proteogenomic studies through large-scale multi-platform transcriptomic and proteomic analysis, and provides a novel objective metric, PMC score, to simultaneously qualify paired RNA and protein samples for proteogenomic studies.

Acknowledgements

We acknowledge contributions from the TCGA Research Network and TCGA Pan-Cancer Analysis Working Group. This study was supported in part by grants from the U.S. National Institutes of Health (CA175486 to H.L. and CA209851 to H.L. and G.B.M., and CCSG grant CA016672); a grant from the Cancer Prevention and Research Institute of Texas (RP140462 to H.L.).

References

1. Akbani R, Ng PK, Werner HM, Shahmoradgoli M, Zhang F, Ju Z, Liu W, Yang JY, Yoshihara K, Li J, Ling S, Seviour EG, Ram PT, Minna JD, Diao L, Tong P, Heymach JV, Hill SM, Dondelinger F, Stadler N, Byers LA, Meric-Bernstam F, Weinstein JN, Broom BM, Verhaak RG, Liang H, Mukherjee S, Lu Y and Mills GB (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun* 5:3887. doi: 10.1038/ncomms4887

2. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Jr., de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR,
Morrisey MP, Sellers WR, Schlegel R and Garraway LA (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**:603-7. doi: 10.1038/nature11003

3. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C and Stuart JM (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**:1113-20. doi: 10.1038/ng.2764

4. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, Stefancsik R, Harsha B, Kok CY, Jia M, Jubb H, Sondka Z, Thompson S, De T and Campbell PJ (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**:D777-D783. doi: 10.1093/nar/gkw1121

5. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, Shen H, Omberg L, Chu A, Margolin AA, Van’t Veer LJ, Lopez-Bigas N, Laird PW, Raphael BJ, Ding L, Robertson AG, Byers LA, Mills GB, Weinstein JN, Van Waes C, Chen Z, Collisson EA, Cancer Genome Atlas Research N, Benz CC, Perou CM and Stuart JM (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**:929-44. doi: 10.1016/j.cell.2014.06.049

6. Liu Y, Beyer A and Aebersold R (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**:535-50. doi: 10.1016/j.cell.2016.03.014

7. Yi S, Lin S, Li Y, Zhao W, Mills GB and Sahni N (2017) Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat Rev Genet* **18**:395-410. doi: 10.1038/nrg.2017.8

8. Myhre S, Lingjaerde OC, Hennessy BT, Aure MR, Carey MS, Alsner J, Tramm T, Overgaard J, Mills GB, Borresen-Dale AL and Sorlie T (2013) Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Mol Oncol* **7**:704-18. doi: 10.1016/j.molonc.2013.02.018
9. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, Kawaler E, Mundt F, Krug K, Tu Z, Lei JT, Gatza ML, Wilkerson M, Perou CM, Yellapantula V, Huang KL, Lin C, McLellan MD, Yan P, Davies SR, Townsend RR, Skates SJ, Wang J, Zhang B, Kinsinger CR, Mesri M, Rodriguez H, Ding L, Paulovich AG, Fenyo D, Ellis MJ, Carr SA and Nci C (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**:55-62. doi: 10.1038/nature18003

10. Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenas C, Lundeberg J, Mann M and Uhlen M (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol* **6**:450. doi: 10.1038/msb.2010.106

11. Gholami AM, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M and Kuster B (2013) Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* **4**:609-20. doi: 10.1016/j.celrep.2013.07.018

12. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN and Perou CM (2014) Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**:419. doi: 10.1186/1471-2164-15-419

13. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND and Betel D (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **14**:R95. doi: 10.1186/gb-2013-14-9-r95

14. Liu S, Lin L, Jiang P, Wang D and Xing Y (2011) A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res* **39**:578-88. doi: 10.1093/nar/gkq817

15. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, Shaughnessy JD, Jr., Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, Chen M, Cheng J, Cheng J, Chou J, Davison TS, Delorenzi M, Deng Y, Devanarayan V, Dix DJ, Dopazo J, Dorff KC, Elloumi F, Fan J, Fan S, Fan X, Fang H, Gonzaludo N, Hess KR,
Credentialing samples for proteogenomic analysis

Hong H, Huan J, Irizarry RA, Judson R, Juraeva D, Lababidi S, Lambert CG, Li L, Li Y, Li Z, Lin SM, Liu G, Lobenhofer EK, Luo J, Luo W, McCall MN, Nikolsky Y, Pennello GA, Perkins RG, Philip R, Popovici V, Price ND, Qian F, Scherer A, Shi T, Shi W, Sung J, Thierry-Mieg D, Thierry-Mieg J, Thodima V, Trygg J, Vishnuvajjala L, Wang SJ, Wu J, Wu Y, Xie Q, Yousef WA, Zhang L, Zhang X, Zhong S, Zhou Y, Zhu S, Arasappan D, Bao W, Lucas AB, Berthold F, Brennan RJ, Buness A, Catalano JG, Chang C, Chen R, Cheng Y, Cui J, Czika W, Demichelis F, Deng X, Dosymbekov D, Eils R, Feng Y, Fostel J, Fulmer-Smentek S, Fuscoe JC, Gatto L, Ge W, Goldstein DR, Guo L, Halbert DN, Han J, Harris SC, Hatzis C, Herman D, Huang J, Jensen RV, Jiang R, Johnson CD, Jurman G, Kahlert Y, Khuder SA, Kohl M, Li J, Li L, Li M, Li QZ, Li S, Li Z, Liu J, Liu Y, Liu Z, Meng L, Madera M, Martinez-Murillo F, Medina I, Meehan J, Miclaus K, Moffitt RA, Montaner D, Mukherjee P, Mulligan GJ, Neville P, Nikolskaya T, Ning B, Page GP, Parker J, Parry RM, Peng X, Peterson RL, Phan JH, Quanz B, Ren Y, Riccadonna S, Roter AH, Samuelson FW, Schumacher MM, Shambaugh JD, Shi Q, Shippy R, Si S, Smalter A, Sotiriou C, Soukup M, Staedtler F, Steiner G, Stokes TH, Sun Q, Tan PY, Tang R, Tezak Z, Thorn B, Tsyganova M, Turpaz Y, Vega SC, Visintainer R, von Frese J, Wang C, Wang E, Wang J, Wang W, Westermann F, Willey JC, Woods M, Wu S, Xiao N, Xu J, Xu L, Yang L, Zeng X, Zhang J, Zhang L, Zhang M, Zhao C, Puri RK, Scherf U, Tong W, Wolfinger RD and Consortium M (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28:827-38. doi: 10.1038/nbt.1665

16. Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu PY, Wang M, Wang C, Thierry-Mieg D, Thierry-Mieg J, Kreil DP and Mason CE (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol 32:888-95. doi: 10.1038/nbt.3000

17. SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol 32:903-14. doi: 10.1038/nbt.2957

25
18. Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM and Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**:5676-84.

19. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou JY, Petyuk VA, Chen L, Ray D, Sun S, Yang F, Chen L, Wang J, Shah P, Cha SW, Aiyeatan P, Woo S, Tian Y, Gritsenko MA, Clauss TR, Choi C, Monroe ME, Thomas S, Nie S, Wu C, Moore RJ, Yu KH, Tabb DL, Fenyo D, Bafna V, Wang Y, Rodriguez H, Boja ES, Hiltke T, Rivers RC, Sokoll L, Zhu H, Shih Ie M, Cope L, Pandey A, Zhang B, Snyder MP, Levine DA, Smith RD, Chan DW, Rodland KD and Investigators C (2016) Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**:755-65. doi: 10.1016/j.cell.2016.05.069

20. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC and Nci C (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**:382-7. doi: 10.1038/nature13438

21. Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, Spiegelman CH, Zimmerman LJ, Ham AJ, Keshishian H, Hall SC, Allen S, Blackman RK, Borchers CH, Buck C, Cardasis HL, Cusack MP, Dodder NG, Gibson BW, Held JM, Hiltke T, Jackson A, Johansen EB, Kinsinger CR, Li J, Mesri M, Neubert TA, Niles RK, Pulsipher TC, Ransohoff D, Rodriguez H, Rudnick PA, Smith D, Tabb DL, Tegeler TJ, Varyiath AM, Vega-Montoto LJ, Wahlander A, Waldemarson S, Wang M, Whiteaker JR, Zhao L, Anderson NL, Fisher SJ, Liebler DC, Paulovich AG, Regnier FE, Tempst P and Carr SA (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol* **27**:633-41. doi: 10.1038/nbt.1546
Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC and Spiegelman C (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* 9:761-76. doi: 10.1021/pr9006365

23. Franks A, Airoldi E and Slavov N (2017) Post-transcriptional regulation across human tissues. *PLoS Comput Biol* 13:e1005535. doi: 10.1371/journal.pcbi.1005535

24. Ju Z, Liu W, Roebuck PL, Siwak DR, Zhang N, Lu Y, Davies MA, Akbani R, Weinstein JN, Mills GB and Coombes KR (2015) Development of a robust classifier for quality control of reverse-phase protein arrays. *Bioinformatics* 31:912-8. doi: 10.1093/bioinformatics/btu736

25. Hu J, He X, Baggerly KA, Coombes KR, Hennessy BT and Mills GB (2007) Non-parametric quantification of protein lysate arrays. *Bioinformatics* 23:1986-94. doi: 10.1093/bioinformatics/btm283

26. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA and Staudt LM (2016) Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* 375:1109-12. doi: 10.1056/NEJMp1607591

27. Liu W, Ju Z, Lu Y, Mills GB and Akbani R (2014) A comprehensive comparison of normalization methods for loading control and variance stabilization of reverse-phase protein array data. *Cancer Inform* 13:109-17. doi: 10.4137/CIN.S13329

28. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M and Getz G (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30:413-21. doi: 10.1038/nbt.2203

29. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4:P3.
30. Mitra S, Federico L, Zhao W, Dennison J, Sarkar TR, Zhang F, Takiar V, Cheng KW, Mani S, Lee JS and Mills GB (2016) Rab25 acts as an oncogene in luminal B breast cancer and is causally associated with Snail driven EMT. *Oncotarget* 7:40252-40265. doi: 10.18632/oncotarget.9730

31. The Cancer Cell Line Encyclopedia Consortium and The Genomics of Drug Sensitivity in Cancer Consortium (2015) Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 528:84-7. doi: 10.1038/nature15736

32. Mertins P, Yang F, Liu T, Mani DR, Petyuk VA, Gillette MA, Clauser KR, Qiao JW, Gritsenko MA, Moore RJ, Levine DA, Townsend R, Erdmann-Gilmore P, Snider JE, Davies SR, Ruggles KV, Fenyo D, Kitchens RT, Li S, Olvera N, Dao F, Rodriguez H, Chan DW, Liebler D, White F, Rodland KD, Mills GB, Smith RD, Paulovich AG, Ellis M and Carr SA (2014) Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol Cell Proteomics* 13:1690-704. doi: 10.1074/mcp.M113.036392

33. Kosti I, Jain N, Aran D, Butte AJ and Sirota M (2016) Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci Rep* 6:24799. doi: 10.1038/srep24799

34. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H and Pandey A (2014) A draft map of the human proteome. *Nature* 509:575-81. doi: 10.1038/nature13302
35. Fortelny N, Overall CM, Pavlidis P and Freue GVC (2017) Can we predict protein from mRNA levels? *Nature* **547**:E19-E20. doi: 10.1038/nature22293

36. The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**:61-70. doi: 10.1038/nature11412

37. Lee J, Geiss GK, Demirkan G, Vellano CP, Filanoski B, Lu Y, Ju Z, Yu S, Guo H, Bogatzki LY, Carter WE, Meredith RK, Krishnamurthy S, Ding Z, Beechem JM and Mills G (2018) Implementation of a Multiplex and Quantitative Proteomics Platform for Assessing Protein Lysates Using DNA-Barcoded Antibodies. *Mol Cell Proteomics*. doi: 10.1074/mcp.RA117.000291
Figure Legends

Figure 1. Correlation coefficient analysis of *cis* protein-mRNA pairs between RPPA and RNA-seq data sets. (A) Pearson’s correlation for 146 pairs of mRNA and matched (*cis*) total proteins, 51 pairs of mRNA and matched phosphorylated proteins, random paired (*trans*) mRNA and total proteins across all tumor types, and 142 pairs of mRNA and *cis* mRNA and total proteins in cell lines. Median correlation is labeled as dashed lines. (B) Bar plot of Pearson’s correlation of *cis* mRNA and proteins in all tumor types. (C) Box plot of Pearson’s correlation of *cis* mRNA and total proteins in 31 lineages. (D) Hierarchical clustering of tumor types based on *cis* protein-mRNA correlation coefficient.

Figure 2. Characterization of PMC markers and PMC-pass samples. (A) Comparison of disease-specific PMC markers in 21 cancer types with >100 samples. The numbers in the brackets indicate the sample size for specific tumor types. 13/21 diseases have >15 marker genes. (B) Comparison of PMC markers identified with leave-one-out cross-validation analysis. 30 protein/genes are selected as the consensus PMC markers. (C) Pearson’s correlation of matched protein-mRNA pairs and random sample pairs based on 30 consensus PMC markers in TCGA13 set. (D) Percentage of PMC-pass samples using consensus markers in TCGA13.

Figure 3. Analysis of consensus PMC scores and other quality control measurements in TCGA13 dataset. (A-D) Analysis of (A-B) within-gene, (C-D) *trans* within-gene, (E-F) within-sample, (G-H) *trans* within-sample correlation of 116 test protein-mRNA pairs (genes excluding PMC markers), and (I-J) total protein content and (K-L) purity in PMC-pass and PMC-fail samples determined by 30 consensus PMC markers in the pan-cancer TCGA13 set and in individual disease types.

Figure 4. Schematic workflow of PMC analysis for processing RPPA and MS data and for prequalify samples for high-throughput characterization.

Figure 5. Correlation coefficient analysis of *cis* total protein-mRNA pairs between RNA-seq and RPPA or MS breast cancer data sets. (A) Pearson’s correlation for 8688 *cis* protein-mRNA pairs detected by MS
and 146 cis protein-mRNA pairs across 105 matched breast tumors (MS data contain 3 additional replicates). Median value is labeled as dashed line. (B) Comparison of Pearson’s correlation of 146 pairs of cis mRNA and protein quantified by MS and RPPA. 118 proteins are measured by both MS and RPPA. RPPA-based consensus PMC markers are labeled with gene symbol.

**Figure 6.** Identification of MS-based PMC markers and analysis of PMC scores in CPTAC BRCA data. (A) PMC scores of 105 breast tumors and 3 replicates. Black line: cut point for PMC-pass samples. (B) ROC analysis of predicting two classes of low-quality samples using MS-based PMC score. 80% sensitivity was used as the cut point. (C-F) Analysis of (C) within-gene, (D) trans within-gene, (E) within-sample, (F) trans within-sample correlation of test protein-mRNA pairs (genes excluding PMC markers) in PMC-pass and PMC-fail samples determined by 1068 MS-based PMC markers. (G) Correlation of protein levels measured by MS and RPPA in 63 samples with both PMC-pass RPPA and MS data (dual PMC-pass) and in 45 samples with at least one PMC-fail data type. (H) Analysis of purity in PMC-pass and PMC-fail samples determined by MS-based markers.

**Figure 7.** Analysis of RPPA-based consensus PMC scores in CPTAC BRCA MS data. (A) Correlation of PMC scores using RPPA-based consensus markers and MS-based marker. (B) ROC analysis of predicting MS-based PMC-pass samples using RPPA-based consensus PMC score. (C) RPPA-based consensus PMC score. Black line: cut point for PMC-pass samples. (D) ROC analysis of predicting two classes of low-quality samples using RPPA-based consensus PMC score. (E-H) Analysis of (E) within-gene, (F) trans within-gene, (G) within-sample, (H) trans within-sample correlation of test protein-mRNA pairs (genes excluding RPPA-based consensus markers) in PMC-pass and PMC-fail samples determined by RPPA-based consensus markers. (I) Correlation of protein levels measured by MS and RPPA in 61 samples with both PMC-pass RPPA and MS data (dual PMC-pass) and in 47 samples with at least one PMC-fail data type. (H) Analysis of purity in PMC-pass and PMC-fail samples determined by RPPA-based consensus markers.
Figure 1

(A) Protein-mRNA correlation (Pearson's R)

(B) (C) (D)
Figure 2

(A) Number of markers

(B) Annexin1

(C) Frequency distribution with PMC score and 10% FDR

(D) Percentage of samples passing 10% FDR
Figure 3

(A) Trans within-gene correlation

(B) Within-gene correlation

(C) Trans within-gene correlation

(D) Within-sample correlation

(E) Trans within-sample correlation

(F) Within-sample correlation

(G) Trans within-sample correlation

(H) Within-sample correlation

(I) Total protein content

(J) Total protein content

(K) Purity

(L) Purity

- PMC-fail
- PMC-pass
RPPA data

Protein and mRNA levels of 30 consensus PMC markers (public database may be included)

QC-pass QC-fail

PMC-based quality assessment

PMC score > 0.386

No Yes

Integrated genomic analysis

Remove or Use with caution

Integrated genomic analysis

MS data

Protein and mRNA levels of 30 consensus PMC markers (public database may be included)

Pearson’s correlation > 0.6

Set-specific PMC markers

Integrated genomic analysis

Remove or Use with caution

Prequalification for high-throughput proteomic studies

Identification of PMC markers

Compute protein/mRNA correlation per gene

Quantification of protein/mRNA pairs of 30 consensus PMC marker

Integrated genomic analysis

Use with caution

characterization by high-throughput platform

Figure 4
Figure 5

(A) Pearson's correlation of protein and mRNA density.

(B) Pearson's R of MS and RNA-seq for various genes:
- ANXA1
- AR
- ASNS
- BCL2
- KIT
- CLDN7
- CCNB1
- CCNE1
- CDH1
- EGFR
- EPPK1
- ESR1
- FASN
- FN1
- G6PD
- GAB2
- GATA3
- ERBB2
- IGFBP2
- SERPINE1
- SRC
- SYK
- MSH6
- PREX1
- PRKCA
- EIF4EBP1
- CHEK2
- RPS6KA1
- TFRC

Legend:
- Cleaved
- Total
- Undetected
Figure 6

(A) Skewed distribution
(B) PMC score = 0.441
(C) Sens: 80.0%
(D) Spec: 73.7%
(E) PV+: 33.8%
(F) PV-: 18.2%
(G) Area under curve: 0.823

(A) PMC score
(B) Sensitivity

(C) p=1x10^-6
(D) p=0.281
(E) p=4.08x10^-7
(F) p=0.319

(G) p=0.02
(H) p=0.066
Figure 7

(A) RPPA-based PMC score vs. MS-based PMC score. R=0.497, p=4.53x10^-8

(B) Area under curve: 0.772

(C) Normal distribution of samples

(D) Normal distribution of samples

(E) Within-gene correlation

(F) Trans within-gene correlation

(G) Within-sample correlation

(H) Trans within-sample correlation

(I) Within-gene MS/RPA correlation

(J) Purity