

Mutation impact on mRNA versus protein expression across human cancers

Yuqi Liu [†], Abdulkadir Elmas [†], and Kuan-lin Huang ^{†*}

Department of Genetics and Genomic Sciences, Department of Artificial Intelligence and Human Health, Center for Transformative Disease Modeling, Tisch Cancer Institute, Icahn Genomics Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

*Correspondence address. Kuan-lin Huang, Departments of Genetics and Genomic Sciences & Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY 10029. Email: kuan-lin.huang@mssm.edu

[†]These authors contributed equally to this work.

Abstract

Background: Cancer mutations are often assumed to alter proteins, thus promoting tumorigenesis. However, how mutations affect protein expression—in addition to gene expression—has rarely been systematically investigated. This is significant as mRNA and protein levels frequently show only moderate correlation, driven by factors such as translation efficiency and protein degradation. Proteogenomic datasets from large tumor cohorts provide an opportunity to systematically analyze the effects of somatic mutations on mRNA and protein abundance and identify mutations with distinct impacts on these molecular levels.

Results: We conduct a comprehensive analysis of mutation impacts on mRNA- and protein-level expressions of 953 cancer cases with paired genomics and global proteomic profiling across 6 cancer types. Protein-level impacts are validated for 47.2% of the somatic expression quantitative trait loci (seQTLs), including *CDH1* and *MSH3* truncations, as well as other mutations from likely “long-tail” driver genes. Devising a statistical pipeline for identifying somatic protein-specific QTLs (spsQTLs), we reveal several gene mutations, including *NF1* and *MAP2K4* truncations and *TP53* missenses showing disproportional influence on protein abundance not readily explained by transcriptomics. Cross-validating with data from massively parallel assays of variant effects (MAVE), *TP53* missenses associated with high tumor *TP53* proteins are more likely to be experimentally confirmed as functional.

Conclusion: This study reveals that somatic mutations can exhibit distinct impacts on mRNA and protein levels, underscoring the necessity of integrating proteogenomic data to comprehensively identify functionally significant cancer mutations. These insights provide a framework for prioritizing mutations for further functional validation and therapeutic targeting.

Keywords: Proteogenomics, cancer mutations, protein expression, quantitative trait loci (QTL), variant effects

Introduction

Cancer arises from the acquisition of mutations that confer selective advantages. Most of these mutations are thought to affect cellular functions by regulating the expression of gene products. For example, truncations can result in nonsense-mediated decay (NMD) [1, 2], which protects eukaryotic cells through degrading premature termination codon (PTC) bearing mRNA [3]. Additionally, a fraction of cancer mutations may uniquely affect protein abundance but not mRNA expression. However, previous studies characterizing genomic mutations affecting mRNA versus protein levels have focused on germline variants as expression quantitative trait loci (eQTL) [4–6]. While other cancer studies have characterized the effect of somatic mutations on mRNA expression levels [7–9], it remains unclear how somatic mutations may affect protein abundance. The gap of knowledge is critical given that mRNA and protein levels are only moderately correlated [10–13]. A myriad of factors, including cell state transition, signal delay, translation on demand, and cellular energy constraint, can lead to discrepancies between mRNA and protein levels [14]. Understanding protein-level consequences of cancer mutations is critical in identifying functionally important mutations and revealing their downstream mechanisms.

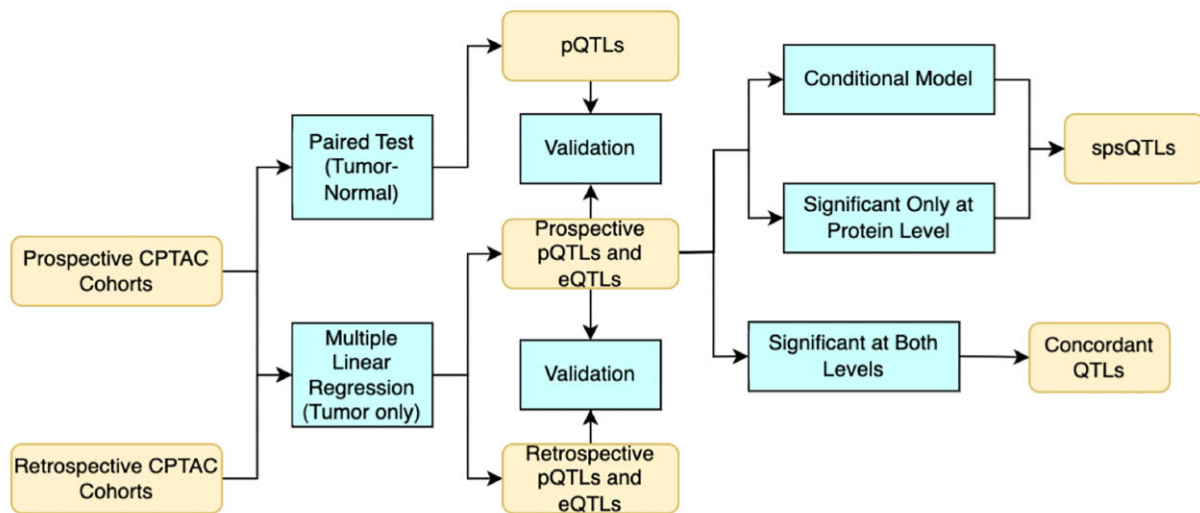
In recent years, advances in mass spectrometry (MS) technologies have generated a wealth of global proteomic profiles of primary tumor cohorts, many of which also have concurrent genomic and transcriptomic profiling [15–20]. These proteogenomic datasets present ample opportunities to validate somatic mutations that show concordant impacts on downstream mRNA and protein levels. On the other hand, protein abundance may also be uniquely influenced by the efficiency of protein translation, transport, and degradation. Thus, proteogenomic analyses can reveal mutations that disproportionately impact protein abundances that may not be found using genomic analyses alone.

Herein, we conducted a systematic analysis to decode the relationship between somatic mutations versus mRNA and protein levels using data from nearly a thousand cases across 6 cancer types in prospective and retrospective cohorts from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). We identified mutations showing concordant effects at both mRNA and protein expression levels in *cis*, as well as those that showed protein-specific effects. We further examined how mutations associated with expression changes may predict *in vitro* and *in vivo* functional effects measured by a massively parallel assay of variant effects (MAVE) of *TP53* [21]. Our results highlight the importance of pairing ge-

Received: May 17, 2024. Revised: September 13, 2024. Accepted: December 5, 2024

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

A



B

Cancer Type	Breast Cancer	Clear Cell Renal Cell Carcinoma	Colorectal Cancer	Lung Adenocarcinoma	Ovarian Cancer	Uterine Corpus Endometrial Carcinoma
Abbreviation	BRCA	CCRCC	CRC	LUAD	OV	UCEC
Data Source	Krug et al. 2020 (PMID: 33212010)	Clark et al. 2019 (PMID: 31675502)	Vasaikar et al. 2019 (PMID: 31031003)	Gillette et al. 2020 (PMID: 32649874)	McDermott et al. 2020 (PMID: 32529193)	Dou et al. 2020 (PMID: 32059776)
Sample Size (Tumors/Normals)	T: 115 N: 18	T: 110 N: 84	T: 95 N: 100	T: 109 N: 102	T: 84 N: 19	T: 97 N: 20
Female %	100%	25.2%	57.4%	34.6%	100%	100%
Average Onset (yr)	60.4	60.6	65.2	62.7	59.1	63.7
Tumor Stage	1: 3% 2: 60.4% 3: 26.9% NA: 9.7%	1: 41.8% 2: 15.5% 3: 34.5% 4: 8.2%	1: 10.2% 2: 40.6% 3: 41.1% 4: 8.1%	1: 53.5% 2: 27.5% 3: 18.5% 4: 0.5%	1: 1% 2: 1% 3: 72.8% 4: 15.5% NA: 9.7%	1: 76.1% 2: 6.8% 3: 14.5% 4: 2.6%

Figure 1: Overview of the study workflow and proteogenomic cohorts. (A) Study workflow to identify eQTLs, pQTLs, concordant QTLs (between mRNA and protein levels), and spsQTLs showing disproportional effects on protein expression. (B) Summary of the prospective CPTAC proteogenomic cohorts used for the discovery analyses, including cancer type abbreviation, data source, sample size of tumor (T) and normal (N) tissues, female percentage, average onset age in years, and tumor stage distribution.

omic and proteomic analyses to prioritize functionally important mutations.

Results

Mutation impacts on the mRNA and protein levels

Following the study workflow (Fig. 1A), we first sought to identify somatic mutations that may impact the corresponding gene's mRNA expression (somatic eQTL, termed seQTL below) and protein abundance (somatic pQTL, termed spQTL below) in primary tumor tissue samples. We performed a multiple regression analysis adjusted for age, gender, ethnicity, and Tandem Mass Tag (TMT) [22] batch using the prospective CPTAC datasets that included matched DNA sequencing (DNA-seq), RNA sequencing (RNA-seq), and MS global proteomics data of primary tumor samples across 6 cancer types (see Methods, Fig. 1B), including 115 breast cancer

(BRCA) [19], 95 colorectal cancer (CRC) [16], 110 clear cell renal cell carcinoma (CCRCC) [15], 109 lung adenocarcinoma (LUAD) [17], 84 ovarian cancer (OV) [20], and 97 uterine corpus endometrial carcinoma (UCEC) [18] samples, as well as proteogenomic datasets for additional, retrospective BRCA [11], CRC [13], and OV [12] cohorts from CPTAC for validation (Supplementary Fig. S1A). We focused on coding mutations given the coverage of the whole-exome sequencing (WES) data used in CPTAC studies; the analyses were further stratified for truncations, missense, and synonymous mutations given their likely different mechanisms of action in affecting levels of the mutated gene product.

Based on the statistical power achieved by these cohort sizes and to reduce false positives, we focused on genes with 3 or more samples affected by mutations in each functional class of missense, truncation, and synonymous within the cancer cohort, including 134, 13, and 15 genes tested in BRCA; 1,360, 318, and 226 genes tested in CRC; 55, 12, and 4 genes tested in CCRCC; 94, 4, and 8 genes tested in LUAD; 134, 5, and 8 genes tested in OV; and 2,243,

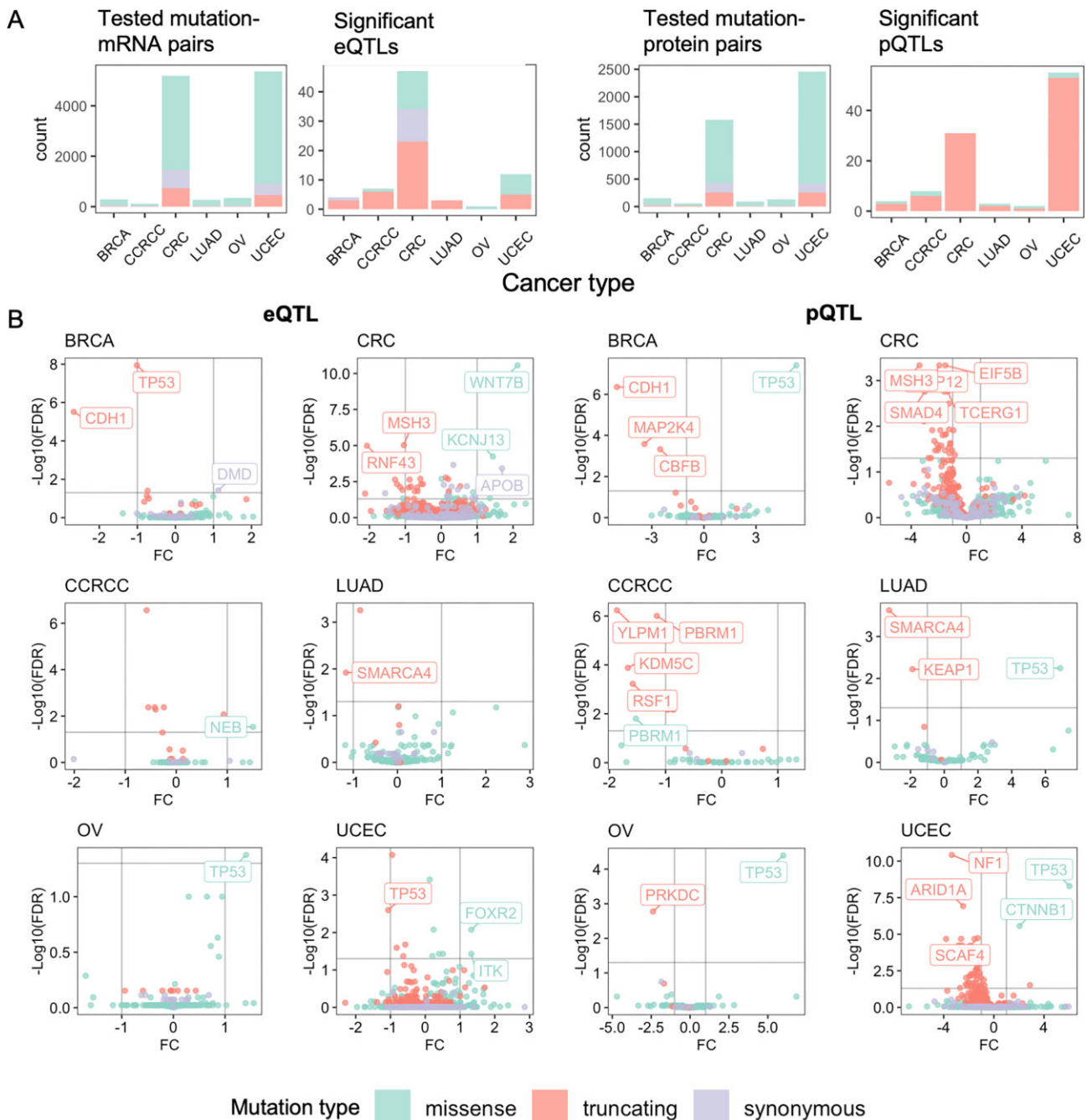


Figure 2: Gene mutations identified as cis seQTLs and spQTLs across 6 adult cancer types. (A) Overview of the somatic mutation QTLs identified in different cancer types and mutation types, including missense (green), truncating (orange), and synonymous (purple) mutations. For both eQTLs and pQTLs, the panel on the left shows the counts of the mutation–gene pairs included in analyses, and the figure on the right shows the counts of the significant eQTLs and pQTLs. (B) Volcano plots showing seQTL associations in the 6 cancer types (left) and volcano plots showing spQTL associations (right), where each dot denotes a gene–cancer pair included in the analysis. Top associated genes were further labeled. FC: mRNA/protein expression log fold change; FDR: false discovery rate.

273, and 196 genes tested in UCEC. We sought to identify their seQTLs affecting cis-expression (i.e., expression of the mutation-affected genes). Using the multiple regression model (see Methods), we identified 74 gene–cancer seQTL pairs (false discovery rate [FDR] < 0.05), including 4 in BRCA, 47 in CRC, 7 in CCRCC, 3 in LUAD, 1 in OV, and 12 in UCEC (Fig. 2A, Supplementary Table S1). Separated by the functional classes of mutations, 22 of those seQTLs are missense mutations, 12 are synonymous, and 40 are truncating. Top seQTLs showing upregulation of gene expression are primarily missenses, including SMARCA4 in LUAD, WNT7B in

CRC, TP53 in OV, and FOXR2 in UCEC. Top candidates showing downregulation of gene expression include TP53 and CDH1 truncations in BRCA, as well as TP53 truncations in OV (Fig. 2B).

Using a similar multiple regression but modeling protein abundance as the dependent variable, we identified 103 significant gene–cancer spQTL pairs (FDR < 0.05), including 4 in BRCA, 31 in CRC, 8 in CCRCC, 3 in LUAD, 2 in OV, and 55 in UCEC (Fig. 2A, Supplementary Table S2). Compared to the proportion of gene mutation type evaluated in each cancer type, spQTLs showed significant enrichment for truncations (Fisher exact test $P < 0.05$;

Fig. 2A), highlighting the persistent and more profound effect of truncations on protein abundance compared to mRNA levels. Among the identified spQTLs across cancer, 7 are missense and 96 are truncating. For example, truncating mutations of *NF1* and *ARID1A* in UCEC and *YLPM1* in CCRCC are each associated with reduced protein level of the corresponding gene (Fig. 2B). Notably, *TP53* missenses in OV, BRCA, LUAD, and UCEC are each significantly associated with increased protein expression in mutation carriers (Fig. 2B).

To verify these discoveries, we applied the same seQTL and spQTL analyses using retrospective CPTAC data (Supplementary Fig. S1A) that included independent cohorts of BRCA [11], CRC [13], and OV [12] primary tumors. While these cohorts afforded smaller sample sizes, 8 seQTLs and 5 spQTLs were detected in both retrospective and prospective sets. The gene–cancer spQTL pairs showing strong validation in both datasets include *TP53* missense mutations and *CDH1* truncations in BRCA and *TP53* truncations in CRC (Supplementary Fig. S1B).

Mutations showing concordant effects at mRNA and protein levels

We next examined the concordance of seQTL and spQTL associations for each gene–cancer type pair. As expected, for most (88.9%) of the significant seQTLs whose genes had sufficient observations at both the mRNA and protein levels, the identified associations showed the same directionality. However, we only identified 17 seQTLs (47.2%) that are also significant spQTLs at an FDR < 0.05, which we show as concordant QTLs (Fig. 3A, Supplementary Table S3). The effect sizes (in log fold change) of these gene–cancer pairs showing concordant seQTLs and spQTLs showed a high correlation between mRNA and protein (Pearson $r = 0.90$, $P < 7.51E-7$).

In different cancer types, genes whose mutation impacts on gene and protein expressions are concordant include well-known drivers of the disease, including *TP53* missense mutations in OV, *CDH1* truncations in BRCA, and *MSH3* truncations in CRC. Upregulation of mutated *TP53* in OV is the only association found for genes affected by missense mutations. The 16 other concordant se/spQTLs are all truncations associated with reduced expression and highlight some “long-tail” driver genes, including *PBRM1* in CCRCC, *YLPM1* in CCRCC/UCEC, and *ESRP1* in UCEC (Fig. 3B). The concordant QTLs with truncating mutation can likely be explained by NMD, which reduces gene expression and in turn diminishes the expression of the corresponding proteins [3]. Compared to the substantially higher counts of seQTL associations (Fig. 2A, B), these concordant se/spQTL effects validate mutation impacts on the gene product.

Protein-specific mutation impacts not observed at mRNA levels

While most seQTLs and spQTLs show concordance, we postulate that certain mutations may uniquely affect protein abundance but not mRNA levels, which we term somatic protein-specific QTLs (spsQTLs). To identify spsQTLs, we applied 2 methods to stringently retain QTLs with discordant effects at mRNA and protein levels. First, applying a likelihood ratio test (LRT) between 2 regression models of protein level being predicted by mRNA level with or without the mutation term (see Methods) [4], 96 candidate spsQTLs (FDR < 0.05) were identified. Second, complementing this LRT test with an approach filtering for a gene–cancer pair showing significant spQTL (FDR < 0.05) but not seQTLs (see Methods) [23], 86 candidate spsQTLs (FDR < 0.05) were identified.

By overlapping candidate spsQTLs identified by both methods, we retained 83 spsQTLs, the majority (92.8%) of which are truncating mutations (Fig. 4A, Supplementary Table S4). Top spsQTLs associated with diminished protein expression include *NF1* truncations in UCEC, *PLEAHK5* truncations in CRC, and *MAP2K4* truncations in BRCA. The only spsQTLs that increase protein expression include *TP53* missense mutations in BRCA, LUAD, and UCEC (Fig. 4B). We further examined the discordance in mutation impacts on gene and protein expression levels (Fig. 4C). While some of these truncations, such as *NF1* in UCEC and *MAP2K4* in BRCA, were often accompanied by lower-than-median mRNA expression in their respective tumor cohorts, their impacts were strikingly observed at diminished protein expression levels. We highlighted in Supplementary Fig. S2A spsQTLs where the affected gene’s protein showed negative protein log fold-change (logFC), whereas the mRNA logFC is nonnegative, including *CASP8* truncations in UCEC, *ARID1A* truncations in CRC and UCEC, and *ATM* truncations in LUAD and UCEC. We also identified a set of spsQTLs truncations, where the logFC associated with a reduction in proteins is 15 times greater than mRNA’s logFC (Supplementary Fig. S2B). These results suggest that NMD associated with these gene truncations is closely tied to the terminated translation but may not affect mRNA expression to the same degree [24].

To complement the cross-tumor analyses, we also utilized the CPTAC samples with paired tumor-normal tissues to conduct paired differential expression tests for both protein and mRNA expression (Fig. 1A). The paired sample sizes with proteomic data include 17 in BRCA, 17 in UCEC, 84 in CCRCC, 100 in LUAD, 29 in CRC, and 10 in OV (Fig. 1B). Covariates including age at diagnosis, ethnicity, race, and sequencing operator are adjusted in the analysis. While this analysis had varied statistical power due to different normal tissue availabilities across cancer types, it served as an independent validation of spQTLs (Supplementary Table S5). This paired tumor-normal analysis validated the protein-level impacts of several discordant spsQTLs (Supplementary Fig. S3A) as well as some concordant se/spQTLs (Supplementary Fig. S3B). For example, the validated discordant spsQTLs include truncations of *SMAD4* and *SCRIB* in CRC as well as *NF1*, *GLYR1*, and *RASA1* in UCEC (Supplementary Fig. S3A). The validated concordant se/spQTLs include truncations of *YLPM1* and *PBRM1* in CCRCC, *SMARCA4* and *KEAP1* in LUAD, and *ESRP1* as well as *JAK2* in UCEC (Supplementary Fig. S3B).

Functional evidence of *TP53* missenses associated with high protein expression

Notably, *TP53* missenses are associated with higher protein expression in multiple cancer cohorts, in addition to the expected reduction in expression associated with truncations (Fig. 5A). Such cis-effect of functional *TP53* missense mutations had previously been observed through immunohistochemistry (IHC [25]) or MS global proteomics experiments [26]. Here, we hypothesized that functional *TP53* missense mutations are more likely to show high levels of concurrent protein-level expression in the mutated tumor sample. To test this hypothesis, we compared gene- and protein-level *TP53* expression from CPTAC with *TP53* mutation-level functional data from the *in vitro* and *in vivo* MAVe experiment conducted by Kotler et al. [21], who designed a p53 variants library to study the functional impact of those mutations.

We divided the *TP53* missense mutations from Kotler et al. [21] into 3 categories: (i) *TP53* mutations with top 20% mRNA or protein expression in the prospective CPTAC cohorts, (ii) the other *TP53* mutations observed across all CPTAC samples, and (iii) the rest

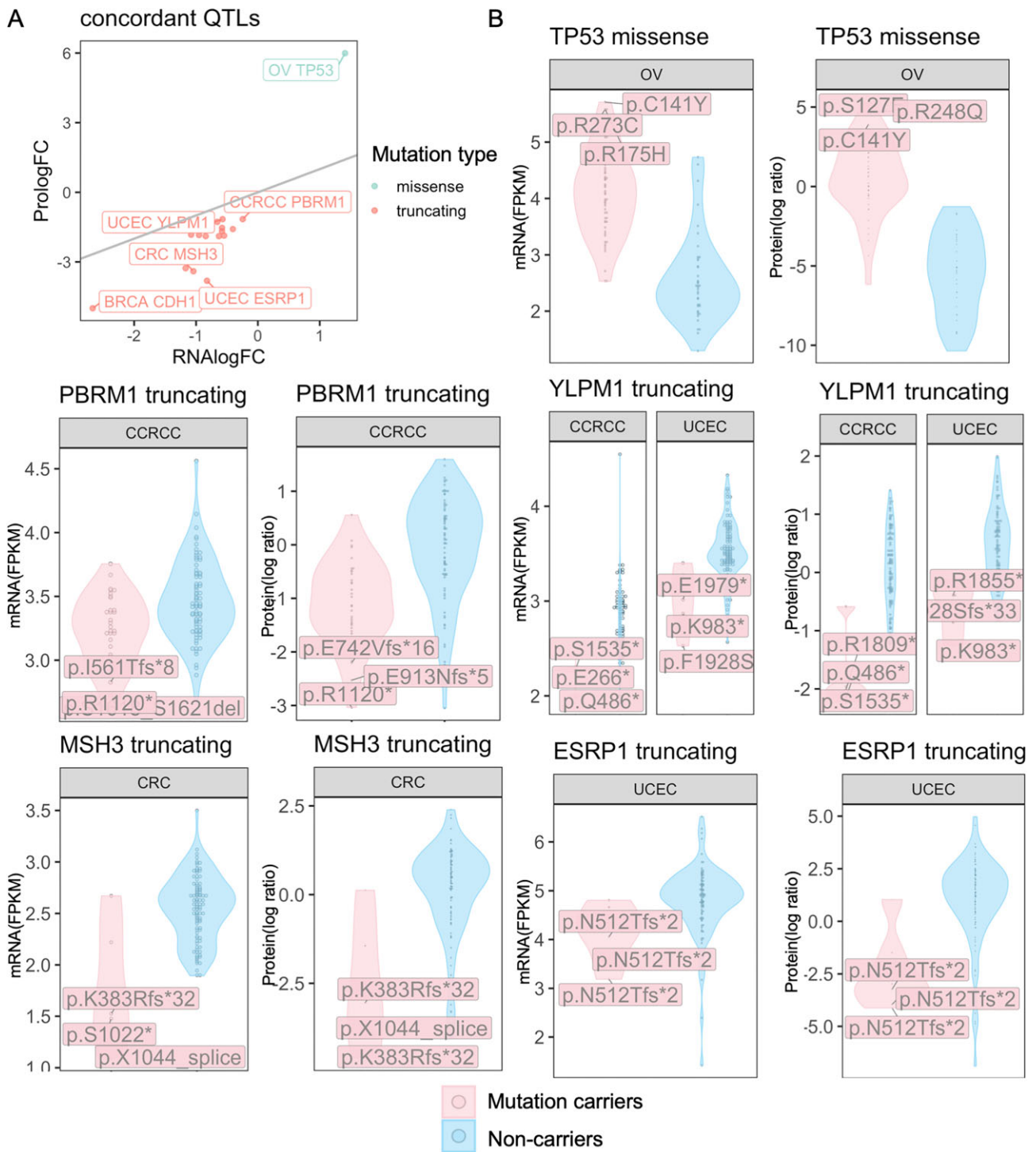


Figure 3: Gene mutations showing concordant impacts on gene and protein expression levels. (A) Overview of concordant QTLs as shown by their effect sizes in log(fold change (FC)), where the gray line shows when the protein logFC equals RNA logFC. Some of the top concordant QTLs were further labeled by cancer type and gene name. (B) Examples of QTL with concordant effects at mRNA and protein expression levels. For each gene, the plot on the left shows the corresponding mRNA levels of mutation carriers versus noncarriers in FPKM, and the plot on the right shows protein level comparison in log ratio (MS TMT measurements) in the respective cancer type labeled on top of each violin plot. The labeled mutations are the 3 mutations whose carriers show the highest absolute expression differences of the mutated gene product compared to the noncarriers.

of the assayed TP53 mutations from Kotler et al. [21]. For *in vitro* data, the number of tested mutations by each category is 32, 78, and 1,033, respectively. For *in vivo* data, the number of tested mutations by each category is 19, 10, and 381, respectively. We first compared the relative fitness score (RFS) measured from the *in vitro* assays [21]. While there may be a trend, we did not observe

a significant difference between all the other mutations versus TP53 missenses associated with either top 20% expression based on either mRNA ($P = 0.090$, Wilcoxon rank-sum test) or protein expression ($P = 0.720$).

We next compared the *in vivo* enrichment scores across the same categories and found that TP53 missenses associated with

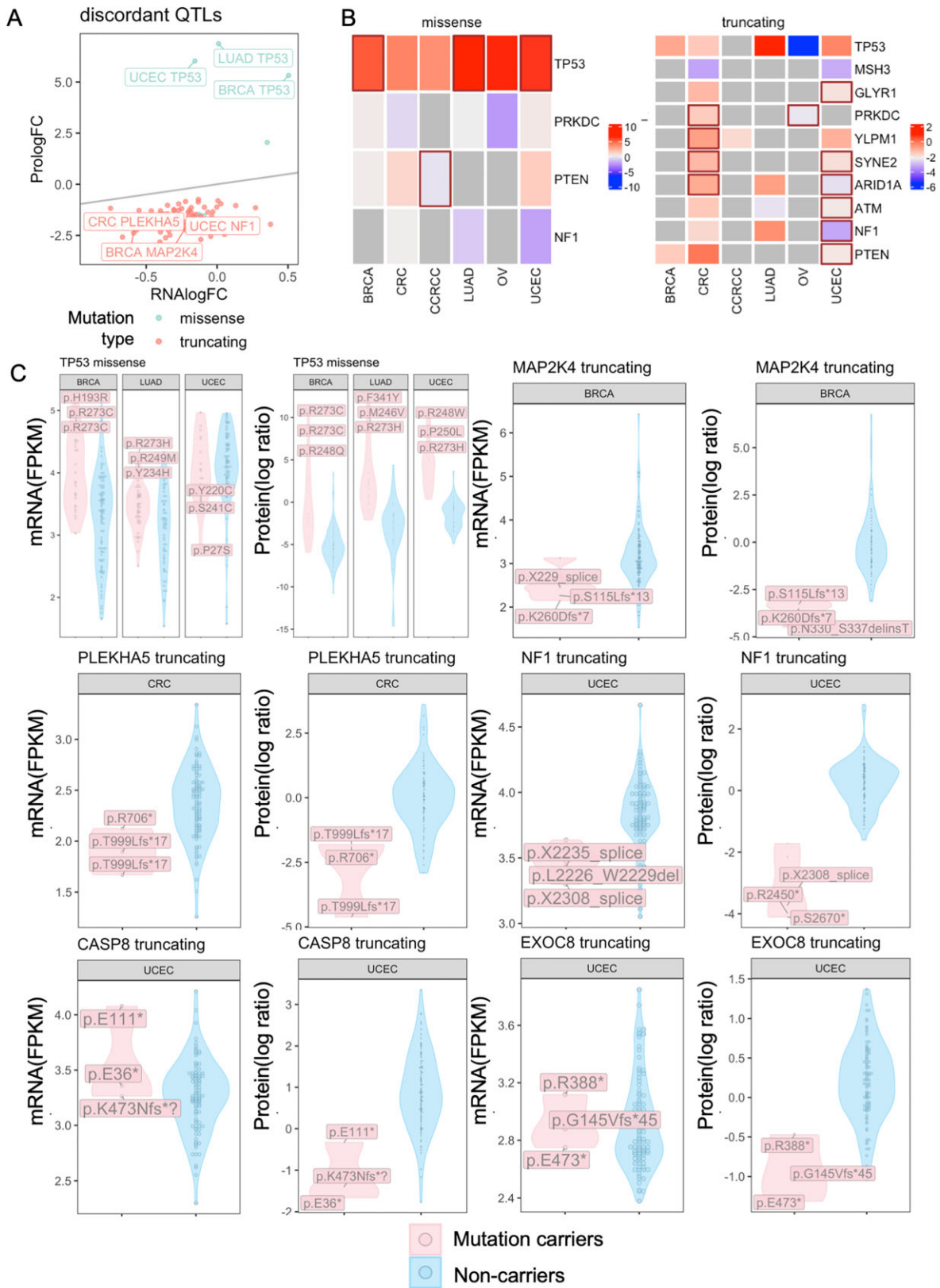


Figure 4: Gene mutations showing discordant impacts on gene and protein expression levels. (A) Overview of discordant QTLs identified by our statistical pipeline, as shown by their effect sizes in logFC, where the gray line shows when the protein logFC equals RNA logFC. (B) Heatmaps of QTLs that are significant as either seQTL or spQTL and that are shared across at least 2 cancer types. Brown box indicates significant spsQTLs, and color indicates the effect size in logFC and average protein expression of mutation carriers in log ratio from the MS TMT quantifications. (C) Examples of QTL with discordant effects at mRNA versus protein levels. For each gene, the plot on the left shows the corresponding mRNA levels of mutation carriers versus noncarriers in FPKM, and the plot on the right shows protein-level comparison in log ratio (MS TMT measurements) in the respective cancer type labeled on top of each violin plot. The labeled mutations are the 3 mutations whose carriers show the highest absolute expression differences of the mutated gene product compared to the noncarriers.

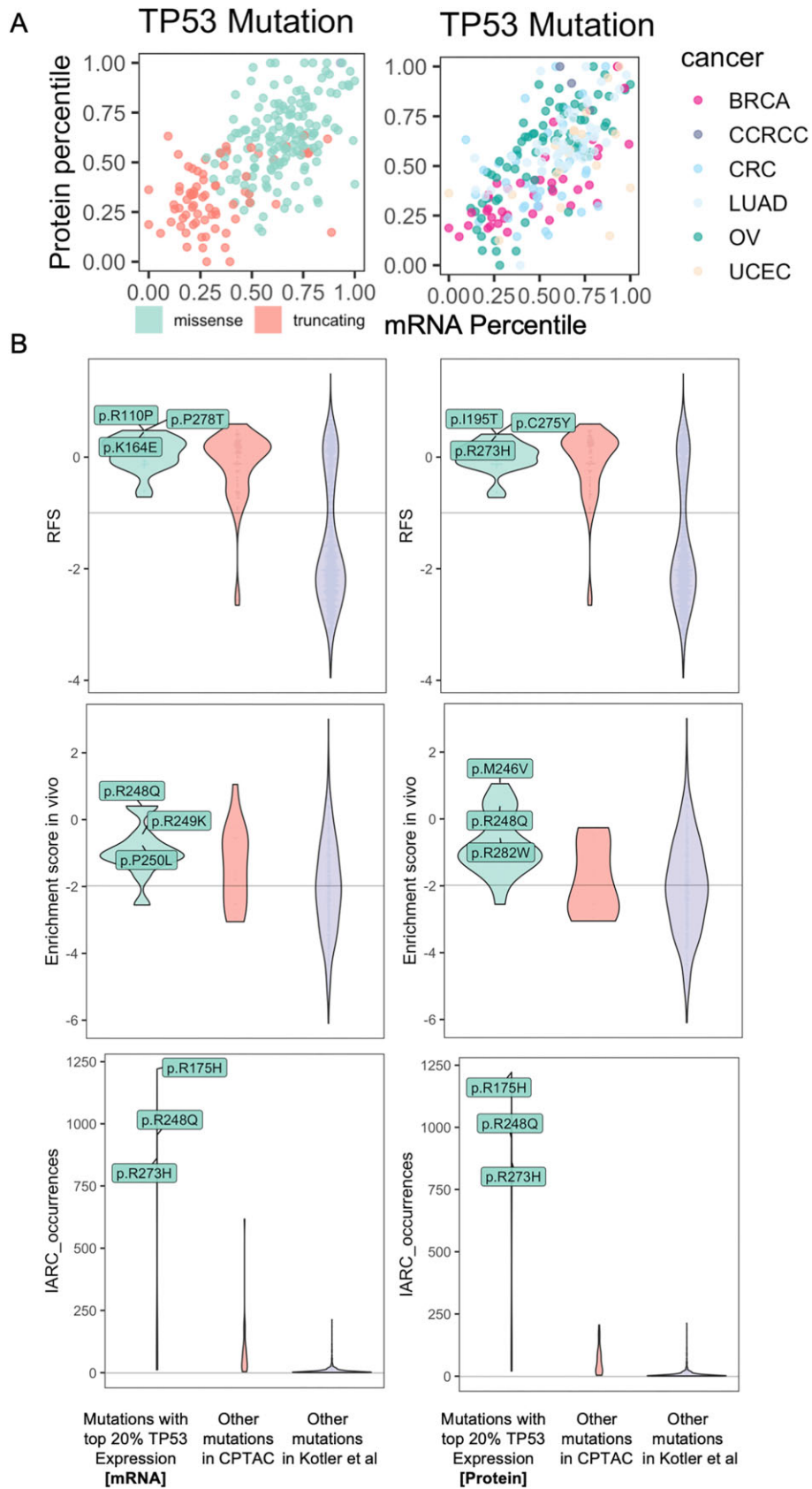


Figure 5: Functional verification of TP53 mutation associated with high mRNA or protein levels using *in vitro* and *in vivo* data from a MAVE experiment. (A) Percentile of averaged expression associated with a given TP53 mutation at the mRNA (x-axis) and protein (y-axis) levels in the respective cancer cohort. TP53 mutations are color-coded by mutation type (left) and observed cancer type (right), respectively. (B) Violin plots comparing the *in vitro* functional score (RFS, top), *in vivo* enrichment score (middle), and IARC occurrences (bottom) for TP53 mutations in the 3 groups defined by (i) TP53 mutations with top 20% mRNA (left) or protein (right) expression in the prospective CPTAC cohorts, (ii) the other TP53 mutations observed across all CPTAC samples, and (iii) the rest of the assayed TP53 mutations from Kotler et al. [21].

top 20% protein expression showed a significantly higher enrichment score *in vivo* compared to that of other TP53 missenses found in CPTAC ($P = 0.016$) or other experimentally measured TP53 mutations ($P = 3.23E-5$, Fig. 5B, Supplementary Table S6). In comparison, TP53 missenses associated with top 20% mRNA expression did not show a significant *in vivo* score difference to that of other TP53 missenses found in CPTAC ($P = 0.170$). Kotler et al. [21] observed that there was no significant correlation between enrichment score *in vivo* and RFS *in vitro*, which is consistent with our observations and may be explained by the different selective pressures between these settings *in vivo* and *in vitro*. Finally, TP53 missenses associated with top 20% protein expression ($P = 5.91E-7$) or top 20% mRNA expression ($P = 2.38E-2$) showed significantly higher prevalence than other CPTAC mutations based on counts from the International Agency for Research on Cancer (IARC) database [21] (Fig. 5B, Supplementary Table S6). Overall, these analyses suggested that protein-level consequences from primary tumor samples can aid the identification of functional mutations.

Discussion

Herein, we analyzed how somatic mutations affect mRNA and protein levels using matched genomic, transcriptomic, and global proteomic data from 953 cases across 6 solid cancer types. We first investigated the mutation impacts at the mRNA level and protein level, finding that although most seQTLs have the same direction of effect as spQTLs, less than half of them are also significant at the protein level. We also studied the concordant or discordant relationship between seQTL versus spQTLs, finding several spQTLs that have disproportional effects on protein. Finally, we conducted analyses to provide functional validation [21] for our findings of TP53 missenses associated with high protein expression.

Integrating protein-level data identified nearly 47.2% seQTLs as concordant, significant spQTLs. The result demonstrates the capacity of proteomic data to validate genomic findings and potentially filter out noises that may arise, for example, due to the more transient nature of transcription compared to translation. In addition to well-known tumor suppressors like TP53 and MSH3, other gene mutations with concordant effects may also be “long-tail” driver genes that will otherwise require large cohort sample sizes to discover. For example, PBRM1, which we found in CCRCC, is a subunit of the PBAF chromatin remodeling complex thought to be a tumor suppressor gene whose mutations may confer synthetic lethality to DNA repair inhibitors [27]. ESRP1, found in UCEC, is crucial in regulating alternative splicing and the translation of some genes during organogenesis [28]. Other less-studied genes we identified include YLPM1 truncations associated with concordantly reduced YLPM1 mRNA and protein expression levels in both CCRCC and UCEC. Analyzing the distribution of these gene mutations on NCI’s Genome Data Commons, we observed many other recurrent truncations (Supplementary Fig. S4), suggesting these mutations may represent some of the “long-tail” driver mutations that warrant further investigation [29, 30].

By devising a specific pipeline to detect spsQTLs, our results showed that apart from mutations that influence protein level mediated by changes in mRNA level, many mutations are associated with disproportional aberrations at the protein level compared to mRNA changes, indicating posttranscriptional regulation. SpsQTLs were found to affect known driver genes such as TP53 missenses and truncations in NF1 [31] and MAP2K4 [32]. In most cases, protein molecules are more direct mediators of cellular functions and phenotypes than mRNAs [33]. Thus, the discor-

dant effect between mRNA level and protein level discovered in our study highlights the importance of exploring disease mechanisms and developing treatments at the protein level.

One possible source of spsQTLs is the imperfect correlation between mRNA and protein expression in the affected genes. Additional statistical analyses revealed that these mRNA–protein correlations range widely across genes and cancer types (Supplementary Fig. S5). While genes harboring spsQTLs have lower mRNA–protein correlations in general than genes with concordant eQTL and pQTL, this is not the case for several discordant genes, including MAP2K4 in BRCA and PBRM1 in CCRCC (Supplementary Table S7). Based on the number of mutations and genes identified, CRC and UCEC reached statistically significant differences between concordant and all other expressed genes (Wilcoxon rank-sum tests, $P = 0.0056$ and $P = 0.022$, respectively); in CRC, mRNA–protein correlations also showed significant differences between discordant and all other expressed genes ($P = 0.013$ and $P = 0.29$, respectively); other cancer types likely did not reach statistical significance likely due to sufficient mutations identified. The imperfect correspondence between gene mRNA–protein correlations and mutation impacts further stresses the need to analyze and consider protein-specific impacts of mutations. Supplementary Table S7 provides complete mRNA–protein correlation data for all concordant/discordant eQTL/pQTLs in their respective cancer type for in-depth examination.

This study has several limitations. First, our findings do not distinguish between several potential mechanisms that could lead to discordant effects of mutations on gene and protein expression. One possibility is that the mutation affects the efficiency of translation, leading to changes in protein levels that are not reflected in mRNA levels. For example, accumulating evidence in recent years suggests that NMD is closely tied to the termination of translation [24], which may explain instances when some truncations afford much stronger associations with protein levels in our findings. But, in many cases, the mechanisms of how mutations may affect protein abundance may be context and gene specific and remain to be elucidated. For example, certain mutations may influence the binding of RNA-binding proteins and the efficiency of translation, whereas others may alter posttranslational modifications, such as phosphorylation or ubiquitination, which can impact protein stability or degradation without affecting transcription or translation rates. Second, the proteogenomic tumor cohorts used herein, while being some of the largest studies to date, still are limited in sample sizes and preclude sufficient statistical power to identify pQTLs at a single mutation level or reveal *trans* effects. Third, given the limitation of current omic technology and data, our findings do not resolve mutation impact on proteins at the temporal, spatial, or single-cell resolution but provide candidate mutations to be investigated in future studies. Fourth, our regression models assume a linear relationship between mutations (1 gene at a time), confounders, and expression, which may not capture more complex, nonlinear effects of mutations on multiple mRNA or protein expression. Future studies could explore nonlinear regression models or neural network approaches to better account for these effects. Fifth, we employed 2 complementary methods to confidently identify spsQTLs that represent true protein-specific regulatory events. However, the reliance on FDR thresholds could still limit the detection of spsQTLs with subtle effects. Alternative approaches, such as Bayesian models that account for prior biological knowledge or hierarchical modeling, could be considered in future analyses to improve the specificity of spsQTL detection. Additionally, while our method focuses on *cis*-acting mutations, potential *trans*-acting

effects could be missed, a limitation that should be explored in larger datasets or by incorporating network-based analyses.

Finally, using TP53 missense mutations as an example, we showed that protein-level expression can serve as an effective strategy to prioritize functional mutations. As DNA-seq becomes ever more commonplace, many rare mutations are being identified, and it remains challenging to accurately classify their functional impacts. Our data demonstrated that TP53 missenses associated with high protein expression show significantly higher functional scores, particularly those measured *in vivo*. This protein expression-based prioritization strategy can be particularly powerful when combined with high-throughput functional assays like using MAVE model systems that are typically *in vitro*. Considering that both MAVE and proteogenomic datasets of tumor cohorts are both expanding quickly in the next few years [34, 35], the combined approaches can help effectively pinpoint functional mutations for mechanistic and clinical characterization. The prioritized mutations based on protein-level consequences may also guide the selection of targeted therapy to advance precision medicine.

Methods

Proteogenomic datasets

The prospective CPTAC data were downloaded and processed as described in the Method section of the work of Elmas et al. [36]. The overview table in Fig. 1A of the dataset describes, for each cancer cohort, the sample size, female patient percentage, average cancer onset age, and tumor stage. Samples are normalized by their median absolute deviations (MAD), so that the MAD of all samples in the dataset is 1. Protein markers with high fractions (greater than 20%) of missing values are filtered out. For the corresponding RNA-seq data, we used the log₂ normalization on the FPKM (fragments per kilobase of exon per million mapped fragments)-normalized RNA-seq counts, and genes that have no expression in at least 90% of the samples were filtered out.

The proteomics data used for validation were downloaded from the NCI CPTAC portal [37]. The dataset overview table in [Supplementary Fig. S1A](#) describes for each cancer cohort the sample size, female patient percentage, average cancer onset age, and tumor stage. The validation data are processed in the same way as the prospective data. The RNA-seq datasets of the 3 retrospective CPTAC cohorts were downloaded from the NCI CPTAC DCC portal [37]. The RNA expression was measured in FPKM [38, 39] and was further normalized by log₂(FPKM + 1).

pQTL and eQTL identification

For each cancer cohort, we identified pQTLs and eQTLs using the multiple linear regression model as implemented in the “limma” R package (v3.42.2) [40]. We also corrected confounding factors, including age, gender, ethnicity, and TMT batch. The FDR was corrected from the *P* values with the Benjamini-Hochberg procedure [41], ensuring that the identified QTLs are statistically robust. Somatic mutations are grouped at a gene level in the multiple regression model, similar to that implemented by our previously developed AeQTL tool [7]. Mutations are separately analyzed by their mechanisms of action, including nonsynonymous mutations that likely do not affect expression, missense mutations, and truncating mutations—including frameshift and in-frame indels, nonsense, splice site, and translation start site mutations. To improve statistical power, we focused our analysis on genes with 3 or more mutations in each cancer cohort and analyzed associations

of mutations affecting cis-expression of the corresponding mRNA or protein products.

spsQTL identification

We combined 2 complementary statistical methods to identify spsQTLs. In the first method adopted from Battle et al. [4], we compared the following 2 nested linear models using the LRT with the “anova” function in R:

$$p = \mu + \beta_0g + \beta_1r$$

$$p = \mu + \beta_2r$$

where *g* is the genotype, *r* represents RNA level, and *p* is the protein level. By comparing these models using the LRT and filtering results with an FDR less than 0.05, we identified candidate spsQTLs where the genotype (mutation) has a disproportionate impact on protein abundance independent of mRNA expression.

In the second method adopted from Mirauta et al. [23], we selected QTLs where the spQTL FDR was less than 0.05 but the corresponding seQTL FDR was greater than 0.05 as candidate spsQTLs to specifically identify mutations that affect protein levels without influencing mRNA. We then overlapped these 2 lists of candidate spsQTLs obtained from 2 complementary methods to identify the final list of spsQTLs for downstream analyses.

mRNA-protein correlation

To investigate the impact of mutations on mRNA and protein expression, we performed a comparative analysis across the 6 solid cancer types. For each cancer type, Pearson correlation coefficients were calculated for individual genes using paired mRNA and protein expression data. We analyzed 3 groups of genes we identified as showing a variable impact on mRNA/protein-level expressions: concordant genes (with mutations showing concordant effects at both mRNA and protein levels in *cis*), discordant genes (showing protein-specific effects), and other genes (showing no concordant or protein-specific impact). Our aim was to test the hypothesis whether the mRNA-protein correlations of the concordant/discordant groups differed from the baseline genome-wide mRNA-protein correlations, indicating biological significance. To assess this, we employed a 2-sample Wilcoxon rank-sum test, comparing the mRNA-protein correlations for the concordant/discordant and other gene groups within each cancer type. Pairwise comparisons were made between the concordant and other gene sets, as well as between the discordant and other gene sets, demonstrating that the correlation coefficients for these groups were drawn from distinct population distributions with statistical significance at a *P* value threshold of 0.05.

Tumor-normal differential expression analysis

We conducted this analysis in the prospective CPTAC cohorts with paired tumor-adjacent tissue-normal samples. For each cancer cohort, we paired the tumor and normal samples from the same patient and performed a differential protein/mRNA expression analysis to identify differentially expressed proteins with the “limma” package. Demographic factors and batch effects, including age, ethnicity, race, and sequencing operator, are adjusted in the multiple regression model.

Availability of Supporting Source Code and Requirements

Project name: Protein expression quantitative trait loci (pQTLs): software and analytic code

Project homepage: <https://github.com/Huang-lab/pQTL> [42]
 Operating system(s): Platform independent
 Programming language: R, Python, Jupyter Notebook
 License: MIT

Additional Files

Supplementary Fig. S1. Overview of the retrospective cohorts. (A) Summary of the retrospective CPTAC proteogenomic cohorts used for the discovery analyses, including cancer type abbreviation, data source, sample size of tumor (T) and normal (N) tissues, female percentage, average onset age in years, and tumor stage distribution. (B) Volcano plots showing seQTL associations in the 6 cancer types (left) and volcano plots showing spQTL associations (right), where each dot denotes a gene–cancer pair included in the analysis. Top associated genes were further labeled. FC: log fold change; FDR: false discovery rate.

Supplementary Fig. S2. spsQTLs with strong effects. (A) Examples of spsQTL whose effect sizes in mRNA level and protein level are in a different direction. For each gene, the plot on the left shows the corresponding mRNA levels of mutation carriers versus non-carriers in FPKM, and the plot on the right shows protein-level comparison in log ratio (MS TMT measurements) in the respective cancer type labeled on top of each of the violin plots. The labeled mutations are the 3 mutations whose carriers show the highest absolute expression differences of the mutated gene product compared to the noncarriers. (B) Examples of spsQTL with a protein logFC and mRNA logFC ratio greater than 15.

Supplementary Fig. S3. Overlap of significant QTLs in cross-tumor analysis and matched tumor-normal analysis projected onto pQTL volcano plots based on cross-tumor analyses. The plots were made separately for (A) discordant spsQTLs and (B) concordant eQTL/pQTLs.

Supplementary Fig. S4. Example lollipop plots showing mutations for 2 genes that were identified as spsQTLs, including YLPM1 and ESRP1. The number on each disc denotes the number of mutations in that position, and the color of the disc represents the mutation type.

Supplementary Fig. S5. Correlation coefficients of concordant versus discordant genes. The violin plots depict the distribution of correlation coefficients between matched mRNA and protein expressions for concordant (blue), discordant (red), and other genes (gray) across the 6 cancer types studied. Genes with notable correlations are labeled in each plot.

Supplementary Table S1. List of expression quantitative trait loci (eQTLs) identified across 6 cancer types. This table provides details on the gene mutations associated with mRNA expression levels, including statistical test results, mutation type, *P* values (adjusted), and effect sizes.

Supplementary Table S2. List of protein quantitative trait loci (pQTLs) identified across 6 cancer types. This table provides details on the gene mutations associated with protein abundance levels, including statistical test results, mutation type, *P* values (adjusted), and effect sizes.

Supplementary Table S3. Concordant expression and protein quantitative trait loci (eQTLs and pQTLs) identified across 6 cancer types. This table includes information on the gene mutations, identified cancer types, and their impact on both mRNA and protein expression levels, demonstrating loci with consistent effects across both molecular layers.

Supplementary Table S4. Significant somatic protein-specific QTLs (spsQTLs) identified by our statistical pipeline across 6 cancer types. This table details the loci with mutations showing sig-

nificant impacts on protein abundance not explained by mRNA levels, including summary statistics for eQTL/pQTL tests and the LRT and overlap test results.

Supplementary Table S5. Summary statistics for differentially expressed proteins (DEPs) identified in paired tumor-normal (TN) samples across 6 cancer types. This table includes the test statistics of protein expression differences between tumor and normal tissues harboring the specific mutation.

Supplementary Table S6. Test statistics between the 3 groups of TP53 mutations. The tested groups were defined by (i) TP53 mutations with top 20% mRNA (left) or protein (right) expression in the prospective CPTAC cohorts, (ii) the other TP53 mutations observed across all CPTAC samples, and (iii) the rest of the assayed TP53 mutations from Kotler et al. using TP53 functional scores from Kotler et al.

Supplementary Table S7. Pearson's correlation coefficient tests between paired mRNA and protein expressions for each concordant and discordant gene, within each cancer cohort.

Acknowledgments

The authors thank CPTAC and its participating patients and families who generously contributed the data.

Author Contributions

Yuqi Liu (Data curation [supporting], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Visualization [lead], Writing – original draft [lead]), and Abdulkadir Elmas (Data curation [lead], Formal analysis [supporting], Methodology [supporting], Writing – review & editing [supporting]). Kuan-lin Huang (Conceptualization [lead], Supervision [lead], Writing – original draft [supporting], Writing – review & editing [lead]).

Funding

This work was supported by NIH NIGMS R35GM138113, ACS RSG-22-115-01-DMC, and Mount Sinai funds to K.H.

Data and Software Availability

Data Availability

Proteomic data for CPTAC-2/3 cohorts can be found on National Cancer Institute (NCI) Proteomic Data Commons (PDC) [43]. The studies used in the discovery cohorts and their PDC study IDs are BRCA (PDC000120), CRC (PDC000116), CCRCC (PDC000127), LUAD (PDC000153), OV (JHU: PDC000110; PNNL: PDC000118), and UCEC (PDC000125).

The studies used in the validation cohorts and their PDC study IDs are BRCA (PDC000173), CRC (PDC000111), and OV (JHU: PDC000113; PNNL: PDC000114).

Genomic data, including DNA mutation and transcriptome profiling for all CPTAC-2/3 cohorts used herein, can be found on National Cancer Institute (NCI) Genome Data Commons (GDC) [44] (dbGaP Study Accession #: phs000892) and (dbGaP Study Accession #: phs001287) [45].

Data for TP53 MAVE assays can be downloaded from the Supplementary Information from Kotler et al. [21].

Supporting data of our analysis results and an archival copy of the corresponding code are available via the GigaScience repository, GigaDB [46].

Competing Interests

K.H. is a cofounder and board member of a non-for-profit 501(c)(3) organization, Open Box Science, from which he does not receive any compensation and has no competing financial interests with this work. All other authors declare no competing interests.

References

1. Kurosaki T, Popp MW, Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol* 2019;20:406–20. <https://doi.org/10.1038/s41580-019-0126-2>.
2. Wang Z, Fan X, Shen Y, et al. Non-cancer-related pathogenic germline variants and expression consequences in ten-thousand cancer genomes. *Genome Med* 2021;13(1). <https://doi.org/10.1186/s13073-021-00964-1>.
3. Lindeboom RGH, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet* 2016;48(10):1112–18. <https://doi.org/10.1038/ng.3664>.
4. Battle A, Khan Z, Wang SH, et al. Impact of regulatory variation from RNA to protein. *Science* 2015;347(6222):664. <https://doi.org/10.1126/science.1260793>.
5. Cenik C, Cenik ES, Byeon GW, et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res* 2015;25(11):1610–21. <https://doi.org/10.1101/gr.193342.115>.
6. Chick JM, Munger SC, Simecek P, et al. Defining the consequences of genetic variation on a proteome-wide scale. *Nature* 2016;534(7608):500–5. <https://doi.org/10.1038/nature18270>.
7. Dong G, Wendl MC, Zhang B, et al. AeQTL: eQTL analysis using region-based aggregation of rare genomic variants. *Pac Symp Biocomput* 2021;26:172–83. https://doi.org/10.1142/9789811232701_0017.
8. Rabadán R, Mohamedi Y, Rubin U, et al. Identification of relevant genetic alterations in cancer using topological data analysis. *Nat Commun* 2020;11(1). <https://doi.org/10.1038/s41467-020-17659-7>.
9. Ding J, McConechy MK, Horlings HM, et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat Commun* 2015;6. <https://doi.org/10.1038/ncomms9554>.
10. Arad G, Geiger T. Functional impact of protein-RNA variation in clinical cancer analyses. *Mol Cell Proteomics* 2023;22:100587. <https://doi.org/10.1016/J.MCPRO.2023.100587>.
11. Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 2016;534(7605):55–62. <https://doi.org/10.1038/nature18003>.
12. Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of Human high-grade serous ovarian cancer. *Cell* 2016;166(3):755–65. <https://doi.org/10.1016/j.cell.2016.05.069>.
13. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;513(7518):382–87. <https://doi.org/10.1038/nature13438>.
14. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. *Cell* 2016;165(3):535–50. <https://doi.org/10.1016/j.cell.2016.03.014>.
15. Clark DJ, Dhanasekaran SM, Petralia F, et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* 2019;179(4):964–83. <https://doi.org/10.1016/j.cell.2019.10.007>.
16. Vasaikar S, Huang C, Wang X, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* 2019;177(4):1035–49. <https://doi.org/10.1016/j.cell.2019.03.030>.
17. Gillette MA, Satpathy S, Cao S, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* 2020;182(1):200–25. <https://doi.org/10.1016/j.cell.2020.06.013>.
18. Dou Y, Kawaler EA, Cui Zhou D, et al. Proteogenomic characterization of endometrial carcinoma. *Cell* 2020;180(4):729–48. <https://doi.org/10.1016/j.cell.2020.01.026>.
19. Krug K, Jaehnig EJ, Satpathy S, et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* 2020;183(5):1436–56. <https://doi.org/10.1016/j.cell.2020.10.036>.
20. McDermott JE, Arshad OA, Petyuk VA, et al. Proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability. *Cell Reports Medicine* 2020;1(1):100004. <https://doi.org/10.1016/j.xcrm.2020.100004>.
21. Kotler E, Shani O, Goldfeld G, et al. A systematic p53 mutation library links differential functional impact to cancer mutation pattern and evolutionary conservation. *Mol Cell* 2018;71(1):178–90. <https://doi.org/10.1016/j.molcel.2018.06.012>.
22. Zecha J, Satpathy S, Kanashova T, et al. TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Mol Cell Proteomics* 2019;18:1468–78. <https://doi.org/10.1074/mcp.TIR119.001385>.
23. Mirauta BA, Seaton DD, Bensaddek D, et al. Population-scale proteome variation in human induced pluripotent stem cells. *eLife* 2020;9:e57390. <https://doi.org/10.7554/ELIFE.57390>.
24. Karousis ED, Mühlemann O. Nonsense-mediated mRNA decay begins where translation ends. *Cold Spring Harb Perspect Biol* 2019;11(2):a032862. <https://doi.org/10.1101/cshperspect.a032862>.
25. Davidoff AM, Humphrey PA, Iglehart JD, et al. Genetic basis for p53 overexpression in human breast cancer. *Proc Natl Acad Sci* 1991;88(11):5006–10. <https://doi.org/10.1073/pnas.88.11.5006>.
26. Huang K-L, Scott AD, Cui Zhou D, et al. Spatially interacting phosphorylation sites and mutations in cancer. *Nat Commun* 2021;12(1). <https://doi.org/10.1038/s41467-021-22481-w>.
27. Chabanon RM, Morel D, Eychenne T, et al. PBRM1 deficiency confers synthetic lethality to DNA repair inhibitors in cancer. *Cancer Res* 2021;81(11):2888–902. <https://doi.org/10.1158/0008-5472.CAN-21-0628>.
28. Vadlamudi Y, Dey DK, Kang SC. Emerging multi-cancer regulatory role of ESRP1: orchestration of alternative splicing to control EMT. *Curr Cancer Drug Targets* 2020;20(9):654–65. <https://doi.org/10.2174/1568009620666200621153831>.
29. Armenia J, Wankowicz SAM, Liu D, et al. The long tail of oncogenic drivers in prostate cancer. *Nat Genet* 2018;50(5):645–51. <https://doi.org/10.1038/s41588-018-0078-z>.
30. Loganathan SK, Schleicher K, Malik A, et al. Rare driver mutations in head and neck squamous cell carcinomas converge on NOTCH signaling. *Science* 2020;367(6483):1264–69. <https://doi.org/10.1126/science.aax0902>.
31. Philpott C, Tovell H, Frayling IM, et al. The NF1 somatic mutational landscape in sporadic human cancers. *Hum Genomics* 2017;11(1):13. <https://doi.org/10.1186/s40246-017-0109-3>.
32. Xue Z, Vis DJ, Bruna A, et al. MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Res* 2018;28(7):719–29. <https://doi.org/10.1038/s41422-018-0044-4>.
33. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* 2020;21:630–44. <https://doi.org/10.1038/s41576-020-0258-4>.

34. Edwards NJ, Oberti M, Thangudu RR, et al. The CPTAC data portal: a resource for cancer proteomics research. *J Proteome Res* 2015;14(6):2707–13. <https://doi.org/10.1021/pr501254j>.
35. Kuang D, Weile J, Kishore N, et al. MaveRegistry: a collaboration platform for multiplexed assays of variant effect. *Bioinformatics* 2021;37(19):3382–83. <https://doi.org/10.1093/bioinformatics/btab215>.
36. Elmas A, Tharakan S, Jaladanki S, et al. Pan-cancer proteomic investigations identify post-transcriptional kinase targets. *Commun Biol* 2021;4(1):1112. <https://doi.org/10.1038/s42003-021-02636-7>.
37. CPTAC Data Portal. <https://cptac-data-portal.georgetown.edu/cptac/>. Accessed 21 December 2024.
38. National Cancer Institute—GDC documentation: FPKM. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#fpkm. Accessed 21 December 2024.
39. cptac rna expression (Ding lab) [GitHub repository]. https://github.com/ding-lab/cptac_rna_expression. Accessed 21 December 2024.
40. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43(7):e47. <https://doi.org/10.1093/nar/gkv007>.
41. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol* 1995;57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
42. Protein expression quantitative trait loci (pQTLs): software and analytic code [GitHub repository]. 2024. <https://github.com/Huang-lab/pQTL>. Accessed 21 December 2024.
43. National Cancer Institute—Proteomics Data Commons. <https://cptac-data-portal.georgetown.edu/cptacPublic/>. Accessed 21 December 2024.
44. National Cancer Institute—Genomic Data Commons, CPTAC-2. <https://portal.gdc.cancer.gov/projects/CPTAC-2>. Accessed 21 December 2024.
45. National Cancer Institute—Genomic Data Commons, CPTAC-3. <https://portal.gdc.cancer.gov/projects/CPTAC-3>. Accessed 21 December 2024.
46. Liu Y, Elmas A, Huang K. Supporting data for “Mutation Impact on mRNA versus Protein Expression across Human Cancers.”. GigaScience Database. 2024. <https://doi.org/10.5524/102598>.