

<b>Manuscript Number:</b>	GIGA-D-24-00168R2	
<b>Full Title:</b>	Mutation Impact on mRNA Versus Protein Expression across Human Cancers	
<b>Article Type:</b>	Research	
<b>Funding Information:</b>	National Institute of General Medical Sciences (R35GM138113)	Dr Kuan-lin Huang
	American Cancer Society (RSG-22-115-01-DMC)	Dr Kuan-lin Huang
<b>Abstract:</b>	<p><b>Background</b> 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.</p> <p><b>Results</b> 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 six 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.</p> <p><b>Conclusion</b> 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.</p>	
<b>Corresponding Author:</b>	Kuan-lin Huang, PhD Icahn School of Medicine at Mount Sinai New York, NY UNITED STATES	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>	Icahn School of Medicine at Mount Sinai	
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Yuqi Liu	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Yuqi Liu	
	Abdulkadir Elmas	
	Kuan-lin Huang, PhD	

<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>We would like to re-submit our manuscript titled "Mutation Impact on mRNA Versus Protein Expression across Human Cancers" (GIGA-D-24-00168), addressing all the editorial comments:</p> <p>1) Please include a citation to your new GigaDB dataset (including the DOI link) to your reference list, and cite this in the data availability section. Completed.</p> <p>2) Please structure your abstract ("Background - Results - Conclusions") Completed.</p> <p>3) Please submit the manuscript text without embedded figures. Please upload the figures separately in Editorial Manager (one file per figure in good resolution - please refer to the formatting instructions on our homepage). Completed.</p> <p>4) Please move URLs to the bibliography and cite them by reference number, rather than including them in the text directly (e.g. line 562, 570, 571). (We treat internet sources as citable objects). Completed.</p> <p>5) Please rename the "code availability" section as "Availability of supporting source code and requirements" and use this tabular format: Completed.</p> <p>6) For reference numbers in the text, please use square brackets ( e.g. [1]) instead of superscript numbers. Completed.</p> <p>7) Please also note the reviewer's comment below regarding the equation - although it looks fine in the Word document I'm looking at now, I believe this was just a conversion problem in the PDF. Yes it looked fine on our end, attached also the PDF we converted ourselves which looked fine.</p> <p>8) Please also ensure that your revised manuscript conforms to the journal style, which can be found in the Instructions for Authors on the journal homepage Completed.</p> <p>Sincerely and on behalf of the team,</p> <p>Kuan-lin Huang, Ph.D. Associate Professor of Genetics and Genomic Sciences &amp; Artificial Intelligence and Human Health</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes

<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>

1 **Mutation Impact on mRNA Versus Protein Expression across Human Cancers**

2

3 Yuqi Liu<sup>1\*</sup>, Abdulkadir Elmas<sup>1\*</sup>, Kuan-lin Huang<sup>1#</sup>

4

5 <sup>1</sup> Department of Genetics and Genomic Sciences, Department of Artificial Intelligence  
6 and Human Health, Center for Transformative Disease Modeling, Tisch Cancer Institute,  
7 Icahn Genomics Institute, Icahn School of Medicine at Mount Sinai, New York, NY  
8 10029, USA.

9 \* These authors contributed equally to this work.

10

11 ORCID: Yuqi Liu [0009-0006-3177-5417]; Abdulkadir Elmas [0000-0002-7999-5770];  
12 Kuan-lin Huang [0000-0002-5537-5817]

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

#Corresponding Author:

34

Kuan-lin Huang, Ph.D.

35

Departments of Genetics and Genomic Sciences & Artificial Intelligence and Human  
36 Health

37

Icahn School of Medicine at Mount Sinai

38

New York, NY 10029

39

Email: [kuan-lin.huang@mssm.edu](mailto:kuan-lin.huang@mssm.edu)

## 40 **ABSTRACT**

### 41 **Background**

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

49

### 50 **Results**

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

62

### 63 **Conclusion**

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

69

## 70 INTRODUCTION

71 Cancer arises from the acquisition of mutations that confer selective advantages. The  
72 majority of these mutations are thought to affect cellular functions by regulating the  
73 expression of gene products. For example, truncations can result in nonsense-mediated  
74 decay (NMD)[1], [2], which protects eukaryotic cells through degrading premature  
75 termination codon (PTC) bearing mRNA[3]. Additionally, a fraction of cancer mutations  
76 may uniquely affect protein abundance but not mRNA expression. However, previous  
77 studies characterizing genomic mutations affecting mRNA vs. protein levels have focused  
78 on germline variants as expression quantitative trait loci (eQTL)[4], [5], [6]. While other  
79 cancer studies have characterized the effect of somatic mutations on mRNA expression  
80 levels[7], [8], [9], it remains unclear how somatic mutations may affect protein abundance.  
81 The gap of knowledge is critical given that mRNA and protein levels are only moderately  
82 correlated[10], [11], [12], [13]. A myriad of factors, including cell state transition, signal  
83 delay, translation on demand, and cellular energy constraint, can lead to discrepancies  
84 between mRNA and protein levels[14]. Understanding protein-level consequences of  
85 cancer mutations is critical in identifying functionally important mutations and revealing  
86 their downstream mechanisms.

87 In recent years, advances in mass spectrometry (MS) technologies have generated a  
88 wealth of global proteomic profiles of primary tumor cohorts, many of which also have  
89 concurrent genomic and transcriptomic profiling[15], [16], [17], [18], [19], [20]. These  
90 proteogenomic datasets present ample opportunities to validate somatic mutations that  
91 show concordant impacts on downstream mRNA and protein levels. On the other hand,  
92 protein abundance may also be uniquely influenced by the efficiency of protein  
93 translation, transport, and degradation. Thus, proteogenomic analyses can reveal  
94 mutations that disproportionately impact protein abundances that may not be found using  
95 genomic analyses alone.

96 Herein, we conducted a systematic analysis to decode the relationship between somatic  
97 mutations vs. mRNA and protein levels using data from nearly a thousand cases across  
98 six cancer types in prospective and retrospective cohorts from the Clinical Proteomic  
99 Tumor Analysis Consortium (CPTAC). We identified mutations showing concordant

100 effects at both mRNA and protein expression levels *in cis*, as well as those that showed  
101 protein-specific effects. We further examined how mutations associated with expression  
102 changes may predict *in vitro* and *in vivo* functional effects measured by a massively  
103 parallel assays of variant effects (MAVE) of TP53[21]. Our results highlight the importance  
104 of pairing genomic and proteomic analyses to prioritize functionally important mutations.

## 105 **RESULTS**

### 106 **Mutation impacts on the mRNA and protein levels**

107 Following the study workflow (**Figure 1A**), we first sought to identify somatic mutations  
108 that may impact the corresponding gene's mRNA expression (somatic eQTL, termed  
109 seQTL below) and protein abundance (somatic pQTL, termed spQTL below) in primary  
110 tumor tissue samples. We performed a multiple regression analysis adjusted for age,  
111 gender, ethnicity, and TMT batch using the prospective CPTAC datasets that included  
112 matched DNA-Seq, RNA-Seq, and mass spectrometry (MS) global proteomics data of  
113 primary tumor samples across six cancer types (see **Methods, Figure 1B**), including  
114 breast cancer (BRCA)[19], 95 colorectal cancer (CRC)[16], 110 clear cell renal cell  
115 carcinoma (CCRCC)[15], 109 lung adenocarcinoma (LUAD)[17], 84 ovarian cancer  
116 (OV)[20], and 97 uterine corpus endometrial carcinoma (UCEC)[18], as well as  
117 proteogenomic datasets for additional, retrospective BRCA[11], CRC[13], and OV[12]  
118 cohorts from CPTAC for validation (**Figure S1A**). We focused on coding mutations given  
119 the coverage of the whole-exome sequencing (WES) data used in CPTAC studies; the  
120 analyses were further stratified for truncations, missense, and synonymous mutations  
121 given their likely different mechanisms of action in affecting levels of the mutated gene  
122 product.

123

124 Based on the statistical power achieved by these cohort sizes and to reduce false  
125 positives, we focused on genes with three or more samples affected by mutations in each  
126 functional class of missense, truncation, and synonymous within the cancer cohort,  
127 including 134, 13, and 15 genes tested in BRCA; 1360, 318, and 226 genes tested in  
128 CRC; 55, 12, and 4 genes tested in CCRCC; 94, 4, and 8 genes tested in LUAD; 134, 5,  
129 and 8 genes tested in OV; 2243, 273, and 196 genes tested in UCEC. We sought to

130 identify their seQTLs affecting *cis*-expression, i.e., expression of the mutation-affected  
131 genes. Using the multiple regression model (see **Methods**), we identified 74 gene-cancer  
132 seQTL pairs (FDR < 0.05), including 4 in BRCA, 47 in CRC, 7 in CCRCC, 3 in LUAD, 1  
133 in OV, and 12 in UCEC (**Figure 2A, Table S1**). Separated by the functional classes of  
134 mutations, 22 of those seQTLs are missense mutations, 12 are synonymous, and 40 are  
135 truncating. Top seQTLs showing up-regulation of gene expression are primarily  
136 missenses, including *SMARCA4* in LUAD, *WNT7B* in CRC, *TP53* in OV, and *FOXR2* in  
137 UCEC. Top candidates showing down-regulation of gene expression include *TP53* and  
138 *CDH1* truncations in BRCA, as well as *TP53* truncations in OV (**Figure 2B**).

139  
140 Using a similar multiple regression but modeling protein abundance as the dependent  
141 variable, we identified 103 significant gene-cancer spQTL pairs (FDR < 0.05), including 4  
142 in BRCA, 31 in CRC, 8 in CCRCC, 3 in LUAD, 2 in OV, and 55 in UCEC (**Figure 2A,**  
143 **Table S2**). Compared to the proportion of gene-mutation type evaluated in each cancer  
144 type, spQTLs showed significant enrichment for truncations (Fisher exact test p-value <  
145 0.05; **Figure 2A**), highlighting the persistent and more profound effect of truncations on  
146 protein abundance compared to mRNA levels. Among the identified spQTLs across  
147 cancer, 7 are missense and 96 are truncating. For example, truncating mutations of *NF1*  
148 and *ARID1A* in UCEC, and *YLPM1* in CCRCC are each associated with reduced protein  
149 level of the corresponding gene (**Figure 2B**). Notably, *TP53* missenses in OV, BRCA,  
150 LUAD, and UCEC are each significantly associated with increased protein expression in  
151 mutation carriers (**Figure 2B**).

152  
153 To verify these discoveries, we applied the same seQTL and spQTL analyses using  
154 retrospective CPTAC data (**Figure S1A**) that included independent cohorts of BRCA[11],  
155 CRC[13], and OV[12] primary tumors. While these cohorts afforded smaller sample sizes,  
156 8 seQTLs and 5 spQTLs were detected in both retrospective and prospective sets. The  
157 gene-cancer spQTL pairs showing strong validation in both datasets include *TP53*  
158 missense mutations and *CDH1* truncations in BRCA, and *TP53* truncations in CRC  
159 (**Figure S1B**).

160



### 161 **Mutations showing concordant effects at mRNA and protein levels**

162 We next examined the concordance of seQTL and spQTL associations for each gene-  
163 cancer type pair. As expected, for most (88.9%) of the significant seQTLs whose genes  
164 had sufficient observations at both the mRNA and protein levels, the identified  
165 associations showed the same directionality. However, we only identified 17 seQTLs  
166 (47.2%) that are also significant spQTLs at an FDR < 0.05, which we show as concordant  
167 QTLs (**Figure 3A, Table S3**). The effect sizes (in log fold change) of these gene-cancer  
168 pairs showing concordant seQTLs and spQTLs showed a high correlation between  
169 mRNA and protein (Pearson  $r = 0.90$ ,  $p$ -value <  $7.51E-7$ ).

170

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

182

### 183 **Protein-specific mutation impacts not observed at mRNA levels**

184 While most seQTLs and spQTLs show concordance, we postulate that certain mutations  
185 may uniquely affect protein abundance but not mRNA levels, which we term somatic  
186 protein-specific QTLs (spsQTLs). To identify spsQTLs, we applied two methods to  
187 stringently retain QTLs with discordant effects at mRNA and protein levels. First, applying  
188 a likelihood ratio test (LRT) between two regression models of protein level being  
189 predicted by mRNA level with or without the mutation term (see **Methods**)[4], 96  
190 candidate spsQTLs (FDR < 0.05) were identified. Second, complementing this LRT test

191 with an approach filtering for gene-cancer pair showing significant spQTL (FDR < 0.05)  
192 but not seQTLs (see **Methods**) [22], 86 candidate spsQTLs (FDR < 0.05) were identified.

193  
194 By overlapping candidate spsQTLs identified by both methods, we retained 83 spsQTLs,  
195 the majority (92.8%) of which are truncating mutations (**Figure 4A, Table S4**). Top  
196 spsQTLs associated with diminished protein expression include *NF1* truncations in UCEC,  
197 *PLEAHK5* truncations in CRC, and *MAP2K4* truncations in BRCA. The only spsQTLs that  
198 increase protein expression include *TP53* missense mutations in BRCA, LUAD, and  
199 UCEC. (**Figure 4B**). We further examined the discordance in mutation impacts on gene  
200 and protein expression levels (**Figure 4C**). While some of these truncations, such as *NF1*  
201 in UCEC and *MAP2K4* in BRCA, were often accompanied by lower-than-median mRNA  
202 expression in their respective tumor cohorts, their impacts were strikingly observed at  
203 diminished protein expression levels. We highlighted in **Figure S2A** spsQTLs where the  
204 affected gene's protein showed negative protein log fold-change (logFC) whereas the  
205 mRNA logFC is non-negative, including *CASP8* truncations in UCEC, *ARID1A* truncations  
206 in CRC and UCEC, and *ATM* truncations in LUAD and UCED. We also identified a set of  
207 spsQTLs truncations, where the logFC associated with a reduction in proteins is 15 times  
208 greater than mRNAs logFC (**Figure S2B**). These results suggest that NMD associated  
209 with these gene truncations are closely tied to the terminated translation but may not  
210 affect mRNA expression to the same degree [23].

211  
212 To complement the cross-tumor analyses, we also utilized the CPTAC samples with  
213 paired tumor-normal tissues to conduct paired differential expression tests for both protein  
214 and mRNA expression (**Figure 1A**). The paired sample sizes with proteomic data include  
215 17 in BRCA, 17 in UCEC, 84 in CCRCC, 100 in LUAD, 29 in CRC, and 10 in OV (**Figure**  
216 **1B**). Covariates including age at diagnosis, ethnicity, race, and sequencing operator are  
217 adjusted in the analysis. While this analysis had varied statistical power due to different  
218 normal tissue availabilities across cancer types, it served as an independent validation of  
219 spQTLs (**Table S5**). This paired tumor-normal analysis validated the protein-level impacts  
220 of several discordant spsQTLs (**Figure S3A**) as well as some concordant se/spQTLs  
221 (**Figure S3B**). For example, the validated discordant spsQTLs include truncations of

222 *SMAD4* and *SCRIB* in CRC as well as *NF1*, *GLYR1*, and *RASA1* in UCEC (**Figure S3A**).  
223 The validated concordant se/spQTLs include truncations of *YLPM1* and *PBRM1* in  
224 CCRCC, *SMARCA4* and *KEAP1* in LUAD, and *ESRP1* as well as *JAK2* in UCEC (**Figure**  
225 **S3B**).

226

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

228 Notably, *TP53* missenses are associated with higher protein expression in multiple cancer  
229 cohorts, in addition to the expected reduction in expression associated with truncations  
230 (**Figure 5A**). Such cis-effect of functional *TP53* missense mutations had previously been  
231 observed through immunohistochemistry (IHC[24]) or MS global proteomics  
232 experiments[25]. Here, we hypothesized that functional *TP53* missense mutations are  
233 more likely to show high levels of concurrent protein-level expression in the mutated  
234 tumor sample. To test this hypothesis, we compared gene and protein-level *TP53*  
235 expression from CPTAC with *TP53* mutation-level functional data from the *in vitro* and *in*  
236 *vivo* MAVe experiment conducted by Kotler et al[21], where they designed a p53 variants  
237 library to study the functional impact of those mutations.

238

239 We divided the *TP53* missense mutations from Kotler et al. into three categories: (1) *TP53*  
240 mutations with top 20% mRNA or protein expression in the prospective CPTAC cohorts,  
241 (2) the other *TP53* mutations observed across all CPTAC samples, and (3) the rest of the  
242 assayed *TP53* mutations from Kotler et al. For *in vitro* data, the number of tested  
243 mutations by each category is 32, 78, and 1,033, respectively. For *in vivo* data, the  
244 number of tested mutations by each category is 19, 10, and 381, respectively. We first  
245 compared the relative fitness score (RFS) measured from the *in vitro* assays<sup>17</sup>. While  
246 there may be a trend, we did not observe a significant difference between all the other  
247 mutations versus *TP53* missenses associated with either top 20% expression based on  
248 either mRNA (p-value = 0.090, Wilcoxon rank-sum test) or protein expression (p-value =  
249 0.720).

250

251 We next compared the *in vivo* enrichment scores across the same categories, and found  
252 that *TP53* missenses associated with top 20% protein expression showed significantly

253 higher enrichment score *in vivo* compared to that of other *TP53* missenses found in  
254 CPTAC (p-value = 0.016) or other experimentally-measured *TP53* mutations (p-value =  
255 3.23E-5, **Figure 5B, Table S6**). In comparison, *TP53* missenses associated with top 20%  
256 mRNA expression did not show a significant *in vivo* score difference to that of other *TP53*  
257 missenses found in CPTAC (p-value = 0.170). Kotler et al. observed that there was no  
258 significant correlation between enrichment score *in vivo* and RFS *in vitro*, which is  
259 consistent with our observations and may be explained by the different selective  
260 pressures between these settings *in vivo* and *in vitro*[21]. Finally, *TP53* missenses  
261 associated with top 20% protein expression (p-value = 5.91E-7) or top 20% mRNA  
262 expression (p-value = 2.38E-2) showed significantly higher prevalence than other CPTAC  
263 mutations based on counts from the International Agency for Research on Cancer (IARC)  
264 database[21] (**Figure 5B, Table S6**). Overall, these analyses suggested that protein-level  
265 consequences from primary tumor samples can aid the identification of functional  
266 mutations.

267

268

## 269 **DISCUSSION**

270

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

280

281 Integrating protein-level data identified nearly 47.2% seQTLs as concordant, significant  
282 spQTLs. The result demonstrates the capacity of proteomic data to validate genomic  
283 findings and potentially filter out noises that may arise for example due to the more

284 transient nature of transcription compared to translation. In addition to well-known tumor  
285 suppressors like *TP53* and *MSH3*, other gene mutations with concordant effects may also  
286 be “long tail” driver genes that will otherwise require large cohort sample sizes to discover.  
287 For example, *PBRM1*, which we found in CCRCC, is a subunit of the PBAF chromatin  
288 remodeling complex thought to be a tumor suppressor gene whose mutations may confer  
289 synthetic lethality to DNA repair inhibitors[26]. *ESRP1*, found in UCEC, is crucial in  
290 regulating alternative splicing and the translation of some genes during  
291 organogenesis[27]. Other less-studied genes we identified include *YLPM1* truncations  
292 associated with concordantly reduced *YLPM1* mRNA and protein expression levels in  
293 both CCRCC and UCEC. Analyzing the distribution of these gene mutations on NCI’s  
294 Genome Data Commons, we observed many other recurrent truncations (**Figure S4**),  
295 suggesting these mutations may represent some of the “long tail” driver mutations that  
296 warrant further investigation[28], [29].

297

298 By devising a specific pipeline to detect spsQTLs, our results showed that apart from  
299 mutations that influence protein level mediated by changes in mRNA level, many  
300 mutations are associated with disproportional aberrations at the protein level compared  
301 to mRNA changes, indicating post-transcriptional regulation. SpsQTLs were found to  
302 affect known driver genes such as *TP53* missenses, and truncations in *NF1*[30] and  
303 *MAP2K4*[31]. In most cases, protein molecules are more direct mediators of cellular  
304 functions and phenotypes than mRNAs[32]. Thus, the discordant effect between mRNA  
305 level and protein level discovered in our study highlights the importance of exploring  
306 disease mechanisms and developing treatments at the protein level.

307

308 One possible source of spsQTLs is the imperfect correlation between mRNA and protein  
309 expression in the affected genes. Additional statistical analyses revealed that this mRNA-  
310 protein correlations range widely across genes and cancer types (**Figure S5**). While  
311 genes harboring spsQTLs have lower mRNA-protein correlations in general than genes  
312 with concordant eQTL and pQTL, this is not the case for several discordant genes,  
313 including *MAP2K4* in BRCA and *PBRM1* in CCRCC (**Table S7**). Based on the number of  
314 mutations and genes identified, CRC and UCEC reached statistically significant

315 differences between concordant and all other expressed genes (Wilcoxon rank-sum tests,  
316  $p = 0.0056$  and  $p = 0.022$ , respectively); in CRC, mRNA-protein correlations also showed  
317 significant differences between discordant and all other expressed genes ( $p = 0.013$  and  
318  $p = 0.29$ , respectively); other cancer types likely did not reach statistical significance likely  
319 due to sufficient mutations identified. The imperfect correspondence between gene  
320 mRNA-protein correlations and mutation impacts further stresses the need to analyze and  
321 consider protein-specific impacts of mutations. **Table S7** provides complete mRNA-  
322 protein correlation data for all concordant/discordant eQTL/pQTLs in their respective  
323 cancer type for in-depth examination.

324

325 This study has several limitations. First, our findings do not distinguish between several  
326 potential mechanisms that could lead to discordant effects of mutations on gene and  
327 protein expression. One possibility is that the mutation affects the efficiency of translation,  
328 leading to changes in protein levels that are not reflected in mRNA levels. For example,  
329 accumulating evidence in recent years suggests that NMD is closely tied to the  
330 termination of translation[23], which may explain instances where some truncations afford  
331 much stronger associations with protein levels in our findings. But, in many cases, the  
332 mechanisms of how mutations may affect protein abundance may be context- and gene-  
333 specific and remain to be elucidated. For example, certain mutations may influence the  
334 binding of RNA binding proteins and the efficiency of translation, whereas others may  
335 alter post-translational modifications, such as phosphorylation or ubiquitination, which  
336 can impact protein stability or degradation without affecting transcription or translation  
337 rates. Second, the proteogenomic tumor cohorts used herein, while being some of the  
338 largest studies to date, still are limited in sample sizes and preclude sufficient statistical  
339 power to identify pQTLs at a single mutation level or reveal *trans* effects. Third, given the  
340 limitation of current omic technology and data, our findings do not resolve mutation impact  
341 on proteins at the temporal, spatial, or single-cell resolution, but provide candidate  
342 mutations to be investigated in future studies. Fourth, our regression models assume a  
343 linear relationship between mutations (one gene at a time), confounders, and expression,  
344 which may not capture more complex, nonlinear effects of mutations on multiple mRNA  
345 or protein expression. Future studies could explore non-linear regression models or

346 neural network approaches to better account for these effects. Fifth, we employed two  
347 complementary methods to confidently identify spsQTLs that represent true protein-  
348 specific regulatory events. However, the reliance on FDR thresholds could still limit the  
349 detection of spsQTLs with subtle effects. Alternative approaches, such as Bayesian  
350 models that account for prior biological knowledge or hierarchical modeling, could be  
351 considered in future analyses to improve the specificity of spsQTL detection. Additionally,  
352 while our method focuses on cis-acting mutations, potential trans-acting effects could be  
353 missed, a limitation that should be explored in larger datasets or by incorporating network-  
354 based analyses.

355

356 Finally, using *TP53* missense mutations as an example, we showed that protein-level  
357 expression can serve as an effective strategy to prioritize functional mutations. As DNA-  
358 Seq become ever more commonplace, many rare mutations are being identified and it  
359 remains challenging to accurately classify their functional impacts. Our data  
360 demonstrated that *TP53* missenses associated with high protein expression show  
361 significantly higher functional scores, particularly those measured *in vivo*. This protein-  
362 expression-based prioritization strategy can be particularly powerful when combined with  
363 high-throughput functional assays like using MAVE model systems that are typically *in*  
364 *vitro*. Considering that both MAVE and proteogenomic datasets of tumor cohorts are both  
365 expanding quickly in the next few years[33], [34], the combined approaches can help  
366 effectively pinpoint functional mutations for mechanistic and clinical characterization. The  
367 prioritized mutations based on protein-level consequences may also guide the selection  
368 of targeted therapy to advance precision medicine.

## 369 **METHODS**

### 370 **Proteogenomic datasets**

371

372 The prospective CPTAC data were downloaded and processed as described in the  
373 Method section of the work of Elmas et al. [35]. The overview table in **Figure 1A** of the  
374 dataset describes, for each cancer cohort, the sample size, female patient percentage,  
375 average cancer onset age, and tumor stage. Samples are normalized by their median  
376 absolute deviations (MAD), so that the MAD of all samples in the dataset is 1. Protein

377 markers with high fractions (greater than 20%) of missing values are filtered out. For the  
378 corresponding RNA-seq data, we used the log<sub>2</sub> normalization on the FPKM (fragments  
379 per kilobase of exon per million mapped fragments)-normalized RNA-seq counts and  
380 genes that have no expression in at least 90% of the samples were filtered out.

381  
382 The proteomics data used for validation were downloaded from the NCI CPTAC portal  
383 [36]. The dataset overview table in **Figure S1A** describe for each cancer cohort the  
384 sample size, female patient percentage, average cancer onset age, and tumor stage. The  
385 validation data are processed in the same way as the prospective data. The RNA-seq  
386 data sets of the three retrospective CPTAC cohorts were downloaded from the NCI  
387 CPTAC DCC portal[36]. The RNA expression was measured in Fragments Per Kilobase  
388 of transcript per Million mapped reads (FPKM)[37][38] and was further normalized by  
389 log<sub>2</sub>(FPKM+1).

#### 390 391 **pQTL and eQTL identification**

392  
393 For each cancer cohort, we identified pQTLs and eQTLs using the multiple linear  
394 regression model as implemented in the “limma” R package (v3.42.2)[39]. We also  
395 corrected confounding factors including age, gender, ethnicity, and TMT batch. The false  
396 discovery rate (FDR) was corrected from the p-values with the Benjamini-Hochberg  
397 procedure[40], ensuring that the identified QTLs are statistically robust. Somatic  
398 mutations are grouped at a gene level in the multiple regression model, similar to that  
399 implemented by our previously developed AeQTL tool[7]. Mutations are separately  
400 analyzed by their mechanisms of action, including nonsynonymous mutations that likely  
401 do not affect expression, missense mutations, and truncating mutations—including  
402 frameshift and in-frame indels, nonsense, splice site, and translation start site mutations.  
403 To improve statistical power, we focused our analysis on genes with three or more  
404 mutations in each cancer cohort and analyzed associations of mutations affecting *cis*-  
405 expression of the corresponding mRNA or protein products.

#### 406 407 **spsQTL identification**

408



409 We combined two complementary statistical methods to identify spsQTLs. In the first  
410 method adopted from Battle et al.[4], we compared the following two nested linear models  
411 using likelihood ratio test (LRT) with the “anova” function in R:

$$\begin{aligned} 412 \quad y &= \mu + \beta_0 x + \beta_1 p \\ 413 \quad y &= \mu + \beta_2 p \end{aligned}$$

414 where  $x$  is the genotype,  $y$  represents RNA level, and  $p$  is the protein level. By  
415 comparing these models using LRT and filtering results with an FDR less than 0.05, we  
416 identified candidate spsQTLs where the genotype (mutation) has a disproportionate  
417 impact on protein abundance independent of mRNA expression.  
418

419

420 In the second method adopted from Mirauta et al.[22], we selected QTLs where the  
421 spQTL FDR was less than 0.05 but the corresponding seQTL FDR was greater than 0.05  
422 as candidate spsQTLs, to specifically identify mutations that affect protein levels without  
423 influencing mRNA. We then overlapped these two lists of candidate spsQTLs obtained  
424 from two complementary methods to identify the final list of spsQTLs for downstream  
425 analyses.

426

#### 427 **mRNA-Protein correlation:**

428 To investigate the impact of mutations on mRNA and protein expression, we performed  
429 a comparative analysis across the six solid cancer types. For each cancer type, Pearson  
430 correlation coefficients were calculated for individual genes using paired mRNA and  
431 protein expression data. We analyzed three groups of genes we identified as showing  
432 variable impact on mRNA/protein level expressions: Concordant genes (with mutations  
433 showing concordant effects at both mRNA and protein levels in cis), Discordant genes  
434 (showing protein-specific effects), and Other genes (showing no concordant or protein-  
435 specific impact). Our aim was to test the hypothesis whether the mRNA-protein  
436 correlations of the Concordant/Discordant groups differed from the baseline genome-  
437 wide mRNA-protein correlations, indicating biological significance. To assess this, we  
438 employed two-sample Wilcoxon rank-sum test, comparing the mRNA-protein correlations  
439 for the Concordant/Discordant and Other gene groups within each cancer type. Pairwise  
440 comparisons were made between the Concordant and Other gene sets, as well as

441 between the Discordant and Other gene sets, demonstrating that the correlation  
442 coefficients for these groups were drawn from distinct population distributions with  
443 statistical significance at a p-value threshold of 0.05.

444

#### 445 **Tumor-normal differential expression analysis**

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

452

#### 453 **Figures**

454

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

461

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

472

473 **Figure 3. Gene mutations showing concordant impacts on gene and protein**

474 **expression levels.** (A) Overview of concordant QTLs as shown by their effect sizes in

475 log[Fold Change (FC)], where the gray line shows when the protein logFC equals RNA

476 logFC. Some of the top concordant QTLs were further labeled by cancer type and gene

477 name. (B) Examples of QTL with concordant effects at mRNA and protein expression

478 levels. For each gene, the plot on the left shows the corresponding mRNA levels of

479 mutation carriers vs. non-carriers in FPKM, and the plot on the right shows protein level

480 comparison in log ratio (MS TMT measurements) in the respective cancer type labeled

481 on top of each of the violin plots. The labeled mutations are the three mutations whose

482 carriers show the highest absolute expression differences of the mutated gene product

483 compared to the non-carriers.

484

485 **Figure 4. Gene mutations showing discordant impacts on gene and protein**

486 **expression levels.** (A) Overview of discordant QTLs identified by our statistical pipeline

487 as shown by their effect sizes in log[Fold Change (FC)], where the gray line shows when

488 the protein logFC equals RNA logFC. (B) Heatmaps of QTLs that are significant as either

489 seQTL or spQTL and that are shared across at least two cancer types. Brown box

490 indicates significant spsQTLs, and color indicates the effect size in log[Fold Change (FC)],

491 average protein expression of mutation carriers in log ratio from the MS TMT

492 quantifications. (C) Examples of QTL with discordant effects at mRNA vs. protein levels.

493 For each gene, the plot on the left shows the corresponding mRNA levels of mutation

494 carriers vs. non-carriers in FPKM, and the plot on the right shows protein level comparison

495 in log ratio (MS TMT measurements) in the respective cancer type labeled on top of each

496 of the violin plots. The labeled mutations are the three mutations whose carriers show the

497 highest absolute expression differences of the mutated gene product compared to the

498 non-carriers.

499

500 **Figure 5. Functional verification of *TP53* mutation associated with high mRNA or**

501 **protein levels using *in vitro* and *in vivo* data from a MAVE experiment.** (A) Percentile

502 of averaged expression associated with a given *TP53* mutation at the mRNA (x-axis) and

503 protein (y-axis) levels in the respective cancer cohort. *TP53* mutations are color coded by  
504 mutation type (left) and observed cancer type (right), respectively. (B) Violin plots  
505 comparing the in vitro functional score (RFS, top), in vivo enrichment score (middle), and  
506 IARC occurrences (bottom) for *TP53* mutations in the three groups defined by (1) *TP53*  
507 mutations with top 20% mRNA (left) or protein (right) expression in the prospective  
508 CPTAC cohorts, (2) the other *TP53* mutations observed across all CPTAC samples, and  
509 (3) the rest of the assayed *TP53* mutations from Kotler et al<sup>21</sup>.

510

## 511 **Supplementary Tables**

512

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

516

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

520

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

525

526 **Table S4. Significant somatic protein-specific QTLs (spsQTLs) identified by our**  
527 **statistical pipeline across six cancer types.** This table details the loci with mutations  
528 showing significant impacts on protein abundance not explained by mRNA levels, including  
529 summary statistics for eQTL/pQTL tests and the LRT and overlap test results.

530

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

535

536 **Table S6. Test statistics between the three groups of TP53 mutations.** The tested groups  
537 were defined by (1) TP53 mutations with top 20% mRNA (left) or protein (right) expression in  
538 the prospective CPTAC cohorts, (2) the other TP53 mutations observed across all CPTAC  
539 samples, and (3) the rest of the assayed TP53 mutations from Kotler et al. using *TP53*  
540 functional scores from Kotler et al.

541

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

544

### 545 **Supplementary Figures**

546

547 **Supplementary Figure 1. Overview of the retrospective cohorts** (A) Summary of the  
548 retrospective CPTAC proteogenomic cohorts used for the discovery analyses, including  
549 cancer type abbreviation, data source, sample size of tumor (T) and normal (N) tissues,  
550 female percentage, average onset age in years, and tumor stage distribution. (B) Volcano  
551 plots showing seQTLs associations in the six cancer types (left) and volcano plots  
552 showing spQTLs associations (right), where each dot denotes a gene-cancer pair  
553 included in the analysis. Top associated genes were further labeled. FC: log fold change.  
554 FDR: false discovery rate.

555

556 **Supplementary Figure 2. spsQTLs with strong effects.** (A) Examples of spsQTL  
557 whose effect sizes in mRNA level and protein level are in different direction. For each  
558 gene, the plot on the left shows the corresponding mRNA levels of mutation carriers vs.  
559 non-carriers in FPKM, and the plot on the right shows protein level comparison in log ratio  
560 (MS TMT measurements) in the respective cancer type labeled on top of each of the violin  
561 plots. The labeled mutations are the three mutations whose carriers show the highest  
562 absolute expression differences of the mutated gene product compared to the non-  
563 carriers. (B) Examples of spsQTL with a protein logFC and mRNA logFC ratio greater  
564 than 15

565

566 **Supplementary Figure 3. Overlapped of significant QTLs in cross-tumor analysis**  
567 **and matched tumor-normal analysis projected onto pQTL volcano plots based on**  
568 **cross-tumor analyses.** The plots were made separately for (A) discordant spsQTLs, and  
569 (B) concordant eQTL/pQTLs.

570

571 **Supplementary Figure 4. Example lollipop plots showing mutations for two genes that**  
572 **were identified as spsQTLs, including YLPM1 and ESRP1.** The number on each disc  
573 denotes the number of mutations in that position and the color of the disc represents the  
574 mutation type.

575

576 **Supplementary Figure 5. Correlation coefficients of Concordant vs. Discordant**  
577 **genes.** The violin plots depict the distribution of correlation coefficients between matched  
578 mRNA and protein expressions for Concordant (blue), Discordant (red), and Other genes  
579 (gray) across the six cancer types studied. Genes with notable correlations are labeled in  
580 each plot.

581

582

## 583 **DATA AND SOFTWARE AVAILABILITY**

### 584 **Data Availability**

585 Proteomic data for CPTAC-2/3 cohorts can be found on National Cancer Institute (NCI)  
586 Proteomic Data Commons (PDC) [41]. The studies used in the discovery cohorts and  
587 their PDC study IDs are: BRCA (PDC000120), CRC (PDC000116), CCRCC  
588 (PDC000127), LUAD (PDC000153), OV (JHU: PDC000110; PNNL: PDC000118), UCEC  
589 (PDC000125)

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

592 Genomic data, including DNA mutation and transcriptome profiling for all CPTAC-2/3  
593 cohorts used herein can be found on National Cancer Institute (NCI) Genome Data  
594 Commons (GDC) [42] (dbGaP Study Accession #: phs000892) and (dbGaP Study  
595 Accession #: phs001287) [43].

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

598 Supporting data of our analysis results and an archival copy of the corresponding code  
599 are available via the GigaScience repository, GigaDB [44].

600

### 601 **Availability of supporting source code and requirements**

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

604 Project home page: <https://github.com/Huang-lab/pQTL> [45]

605 Operating system(s): Platform independent

606 Programming language: R, Python, Jupyter Notebook

607 License: MIT

### 608 **ACKNOWLEDGEMENTS**

609 The authors wish to acknowledge CPTAC and its participating patients and families that  
610 generously contributed the data. This work was supported by NIH NIGMS  
611 R35GM138113, ACS RSG-22-115-01-DMC, and Mount Sinai funds to KH.

### 612 **DECLARATION OF INTERESTS**

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

### 616 **CONTRIBUTIONS**

617 K.H. conceived the research; Y.L and K.H. designed the analyses. Y.L. and A.E.  
618 developed the software and conducted the bioinformatics analyses, A.E. curated and  
619 preprocessed the datasets. Y.L., A.E., and K.H. wrote the manuscript. K.H. supervised  
620 the study. All authors read, edited, and approved the manuscript.

621

622

### 623 **REFERENCES**

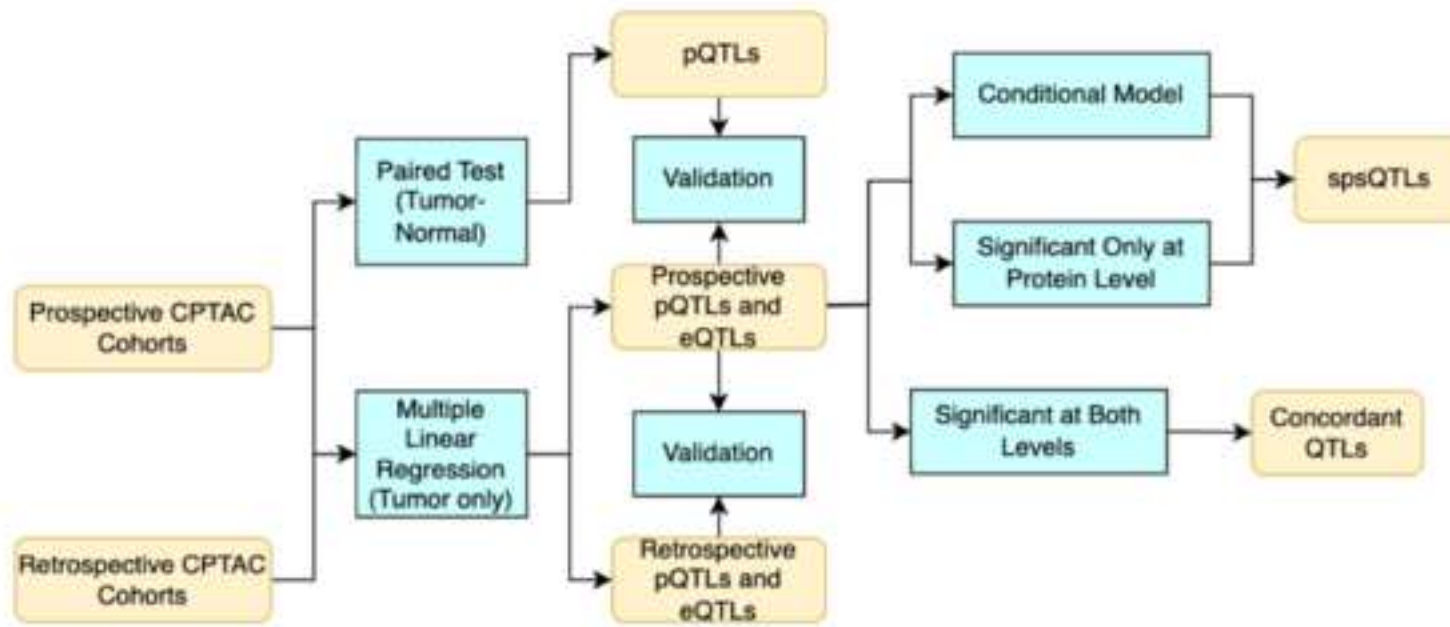
- 624 [1] T. Kurosaki, M. W. Popp, and L. E. Maquat, 'Quality and quantity control of gene  
625 expression by nonsense-mediated mRNA decay', 2019. doi: 10.1038/s41580-019-  
626 0126-2.
- 627 [2] Z. Wang *et al.*, 'Non-cancer-related pathogenic germline variants and expression  
628 consequences in ten-thousand cancer genomes', *Genome Med*, vol. 13, no. 1, 2021,  
629 doi: 10.1186/s13073-021-00964-1.
- 630 [3] R. G. H. Lindeboom, F. Supek, and B. Lehner, 'The rules and impact of nonsense-  
631 mediated mRNA decay in human cancers', *Nat Genet*, vol. 48, no. 10, 2016, doi:  
632 10.1038/ng.3664.
- 633 [4] A. Battle *et al.*, 'Impact of regulatory variation from RNA to protein', *Science*  
634 (1979), vol. 347, no. 6222, 2015, doi: 10.1126/science.1260793.
- 635 [5] C. Cenik *et al.*, 'Integrative analysis of RNA, translation, and protein levels reveals  
636 distinct regulatory variation across humans', *Genome Res*, vol. 25, no. 11, 2015, doi:  
637 10.1101/gr.193342.115.
- 638 [6] J. M. Chick *et al.*, 'Defining the consequences of genetic variation on a proteome-  
639 wide scale', *Nature*, vol. 534, no. 7608, 2016, doi: 10.1038/nature18270.
- 640 [7] G. Dong, M. C. Wendl, B. Zhang, L. Ding, and K. L. Huang, 'AeQTL: eQTL  
641 analysis using region-based aggregation of rare genomic variants', *Pac Symp*  
642 *Biocomput*, vol. 26, 2021, doi: 10.1142/9789811232701\_0017.
- 643 [8] R. Rabadán *et al.*, 'Identification of relevant genetic alterations in cancer using  
644 topological data analysis', *Nat Commun*, vol. 11, no. 1, 2020, doi: 10.1038/s41467-  
645 020-17659-7.
- 646 [9] J. Ding *et al.*, 'Systematic analysis of somatic mutations impacting gene expression  
647 in 12 tumour types', *Nat Commun*, vol. 6, 2015, doi: 10.1038/ncomms9554.
- 648 [10] G. Arad and T. Geiger, 'Functional impact of protein-RNA variation in clinical  
649 cancer analyses', *Molecular & Cellular Proteomics*, p. 100587, Jun. 2023, doi:  
650 10.1016/J.MCPRO.2023.100587.
- 651 [11] P. Mertins *et al.*, 'Proteogenomics connects somatic mutations to signalling in breast  
652 cancer', *Nature*, vol. 534, no. 7605, 2016, doi: 10.1038/nature18003.
- 653 [12] H. Zhang *et al.*, 'Integrated Proteogenomic Characterization of Human High-Grade  
654 Serous Ovarian Cancer', *Cell*, vol. 166, no. 3, 2016, doi: 10.1016/j.cell.2016.05.069.
- 655 [13] B. Zhang *et al.*, 'Proteogenomic characterization of human colon and rectal cancer',  
656 *Nature*, vol. 513, no. 7518, 2014, doi: 10.1038/nature13438.
- 657 [14] Y. Liu, A. Beyer, and R. Aebersold, 'On the Dependency of Cellular Protein Levels  
658 on mRNA Abundance', 2016. doi: 10.1016/j.cell.2016.03.014.
- 659 [15] D. J. Clark *et al.*, 'Integrated Proteogenomic Characterization of Clear Cell Renal  
660 Cell Carcinoma', *Cell*, vol. 179, no. 4, 2019, doi: 10.1016/j.cell.2019.10.007.
- 661 [16] S. Vasaikar *et al.*, 'Proteogenomic Analysis of Human Colon Cancer Reveals New  
662 Therapeutic Opportunities', *Cell*, vol. 177, no. 4, 2019, doi:  
663 10.1016/j.cell.2019.03.030.
- 664 [17] M. A. Gillette *et al.*, 'Proteogenomic Characterization Reveals Therapeutic  
665 Vulnerabilities in Lung Adenocarcinoma', *Cell*, vol. 182, no. 1, 2020, doi:  
666 10.1016/j.cell.2020.06.013.
- 667 [18] Y. Dou *et al.*, 'Proteogenomic Characterization of Endometrial Carcinoma', *Cell*,  
668 vol. 180, no. 4, 2020, doi: 10.1016/j.cell.2020.01.026.



- 669 [19] K. Krug *et al.*, ‘Proteogenomic Landscape of Breast Cancer Tumorigenesis and  
670 Targeted Therapy’, *Cell*, vol. 183, no. 5, 2020, doi: 10.1016/j.cell.2020.10.036.
- 671 [20] J. E. McDermott *et al.*, ‘Proteogenomic Characterization of Ovarian HGSC  
672 Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal  
673 Instability’, *Cell Rep Med*, vol. 1, no. 1, 2020, doi: 10.1016/j.xcrm.2020.100004.
- 674 [21] E. Kotler *et al.*, ‘A Systematic p53 Mutation Library Links Differential Functional  
675 Impact to Cancer Mutation Pattern and Evolutionary Conservation’, *Mol Cell*, vol.  
676 71, no. 1, 2018, doi: 10.1016/j.molcel.2018.06.012.
- 677 [22] B. A. Mirauta *et al.*, ‘Population-scale proteome variation in human induced  
678 pluripotent stem cells’, *Elife*, vol. 9, 2020, doi: 10.7554/ELIFE.57390.
- 679 [23] E. D. Karousis and O. Mühlemann, ‘Nonsense-mediated mRNA decay begins where  
680 translation ends’, *Cold Spring Harb Perspect Biol*, vol. 11, no. 2, 2019, doi:  
681 10.1101/cshperspect.a032862.
- 682 [24] A. M. Davidoff, P. A. Humphrey, J. Dirk Iglehart, and J. R. Marks, ‘Genetic basis for  
683 p53 overexpression in human breast cancer’, *Proc Natl Acad Sci U S A*, vol. 88, no.  
684 11, 1991, doi: 10.1073/pnas.88.11.5006.
- 685 [25] K. lin Huang *et al.*, ‘Spatially interacting phosphorylation sites and mutations in  
686 cancer’, *Nat Commun*, vol. 12, no. 1, 2021, doi: 10.1038/s41467-021-22481-w.
- 687 [26] R. M. Chabanon *et al.*, ‘PBRM1 deficiency confers synthetic lethality to DNA repair  
688 inhibitors in cancer’, *Cancer Res*, vol. 81, no. 11, 2021, doi: 10.1158/0008-  
689 5472.CAN-21-0628.
- 690 [27] Y. Vadlamudi, D. K. Dey, and S. C. Kang, ‘Emerging Multi-cancer Regulatory Role  
691 of ESRP1: Orchestration of Alternative Splicing to Control EMT’, *Curr Cancer  
692 Drug Targets*, vol. 20, no. 9, 2020, doi: 10.2174/1568009620666200621153831.
- 693 [28] J. Armenia *et al.*, ‘The long tail of oncogenic drivers in prostate cancer’, *Nat Genet*,  
694 vol. 50, no. 5, 2018, doi: 10.1038/s41588-018-0078-z.
- 695 [29] S. K. Loganathan *et al.*, ‘Rare driver mutations in head and neck squamous cell  
696 carcinomas converge on NOTCH signaling’, *Science*, vol. 367, no. 6483, 2020, doi:  
697 10.1126/science.aax0902.
- 698 [30] C. Philpott, H. Tovell, I. M. Frayling, D. N. Cooper, and M. Upadhyaya, ‘The NF1  
699 somatic mutational landscape in sporadic human cancers’, 2017. doi:  
700 10.1186/s40246-017-0109-3.
- 701 [31] Z. Xue *et al.*, ‘MAP3K1 and MAP2K4 mutations are associated with sensitivity to  
702 MEK inhibitors in multiple cancer models’, *Cell Res*, vol. 28, no. 7, 2018, doi:  
703 10.1038/s41422-018-0044-4.
- 704 [32] C. Buccitelli and M. Selbach, ‘mRNAs, proteins and the emerging principles of gene  
705 expression control’, 2020. doi: 10.1038/s41576-020-0258-4.
- 706 [33] N. J. Edwards *et al.*, ‘The CPTAC data portal: A resource for cancer proteomics  
707 research’, *J Proteome Res*, vol. 14, no. 6, 2015, doi: 10.1021/pr501254j.
- 708 [34] D. Kuang *et al.*, ‘MaveRegistry: a collaboration platform for multiplexed assays of  
709 variant effect’, *Bioinformatics*, vol. 37, no. 19, 2021, doi:  
710 10.1093/bioinformatics/btab215.
- 711 [35] A. Elmas, S. Tharakan, S. Jaladanki, M. D. Galsky, T. Liu, and K. lin Huang, ‘Pan-  
712 cancer proteogenomic investigations identify post-transcriptional kinase targets’,  
713 *Commun Biol*, vol. 4, no. 1, 2021, doi: 10.1038/s42003-021-02636-7.
- 714 [36] CPTAC Data Portal <https://cptac-data-portal.georgetown.edu/cptac/>

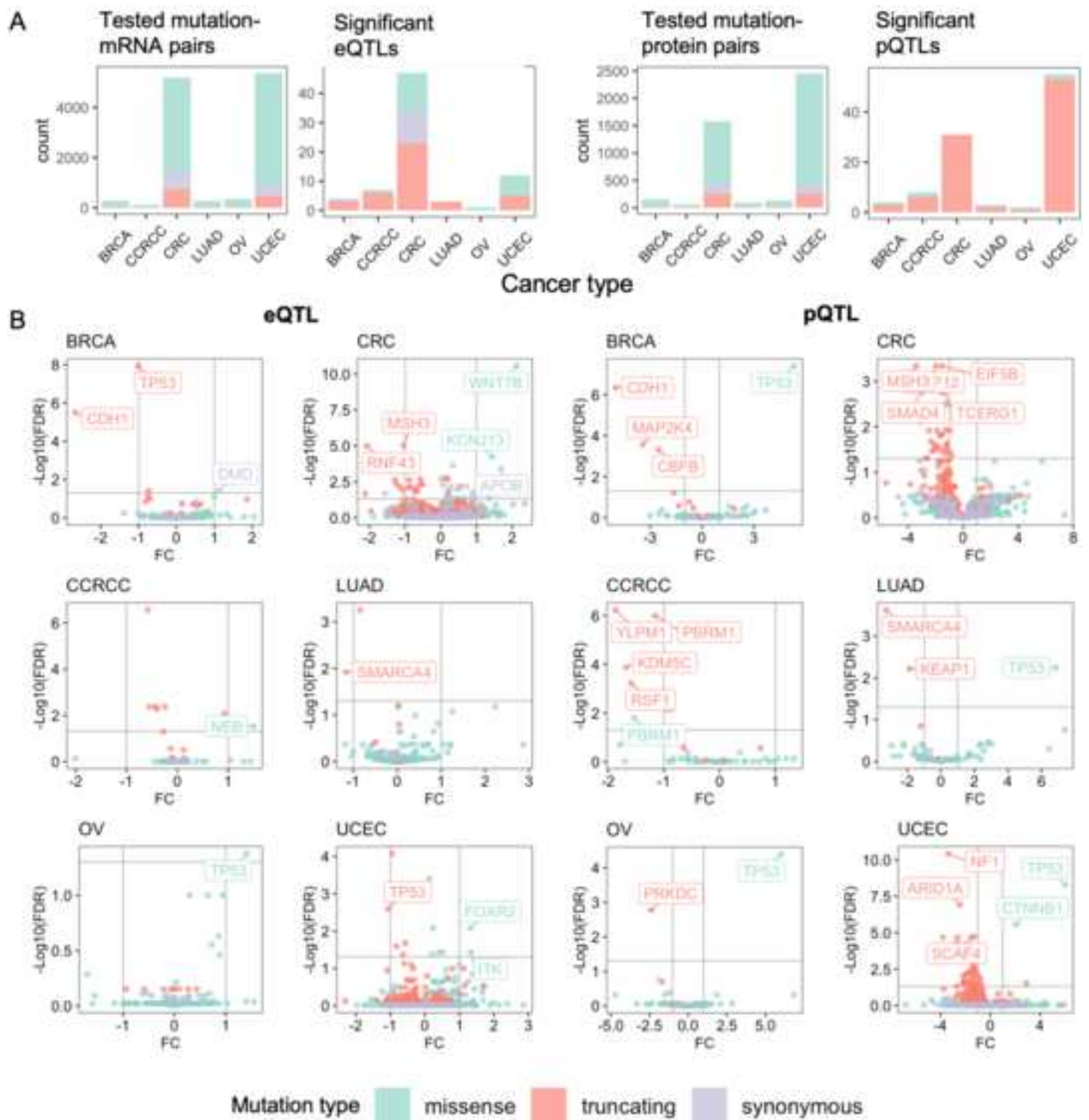
- 715 [37] National Cancer Institute - GDC documentation: FPKM  
716 [https://docs.gdc.cancer.gov/Data/Bioinformatics\\_Pipelines/Expression\\_mRNA\\_Pipeline/#fpkm](https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#fpkm)  
717
- 718 [38] cpta rna expression (Ding lab) [github repository] [https://github.com/ding-](https://github.com/ding-lab/cptac_rna_expression)  
719 [lab/cptac\\_rna\\_expression](https://github.com/ding-lab/cptac_rna_expression)’.
- 720 [39] M. E. Ritchie *et al.*, ‘limma powers differential expression analyses for RNA-  
721 sequencing and microarray studies.’, *Nucleic Acids Res*, vol. 43, no. 7, p. e47, Apr.  
722 2015, doi: 10.1093/nar/gkv007.
- 723 [40] Y. Benjamini and Y. Hochberg, ‘Controlling the False Discovery Rate: A Practical  
724 and Powerful Approach to Multiple Testing’, *J R Stat Soc Series B Stat Methodol*,  
725 vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
- 726 [41] National Cancer Institute – Proteomics Data Commons [https://cptac-data-](https://cptac-data-portal.georgetown.edu/cptacPublic/)  
727 [portal.georgetown.edu/cptacPublic/](https://cptac-data-portal.georgetown.edu/cptacPublic/)
- 728 [42] National Cancer Institute – Genomic Data Commons, CPTAC-2  
729 <https://portal.gdc.cancer.gov/projects/CPTAC-2>
- 730 [43] National Cancer Institute – Genomic Data Commons, CPTAC-3  
731 <https://portal.gdc.cancer.gov/projects/CPTAC-3>’.
- 732 [44] Y. Liu, A. Elmas, and K. Huang, ‘Supporting data for “Mutation Impact on mRNA  
733 Versus Protein Expression across Human Cancers”’. GigaScience Database. 2024.  
734 <https://doi.org/10.5524/102598>
- 735 [45] Protein expression quantitative trait loci (pQTLs): software and analytic code [github  
736 repository] 2024. <https://github.com/Huang-lab/pQTL>’.
- 737

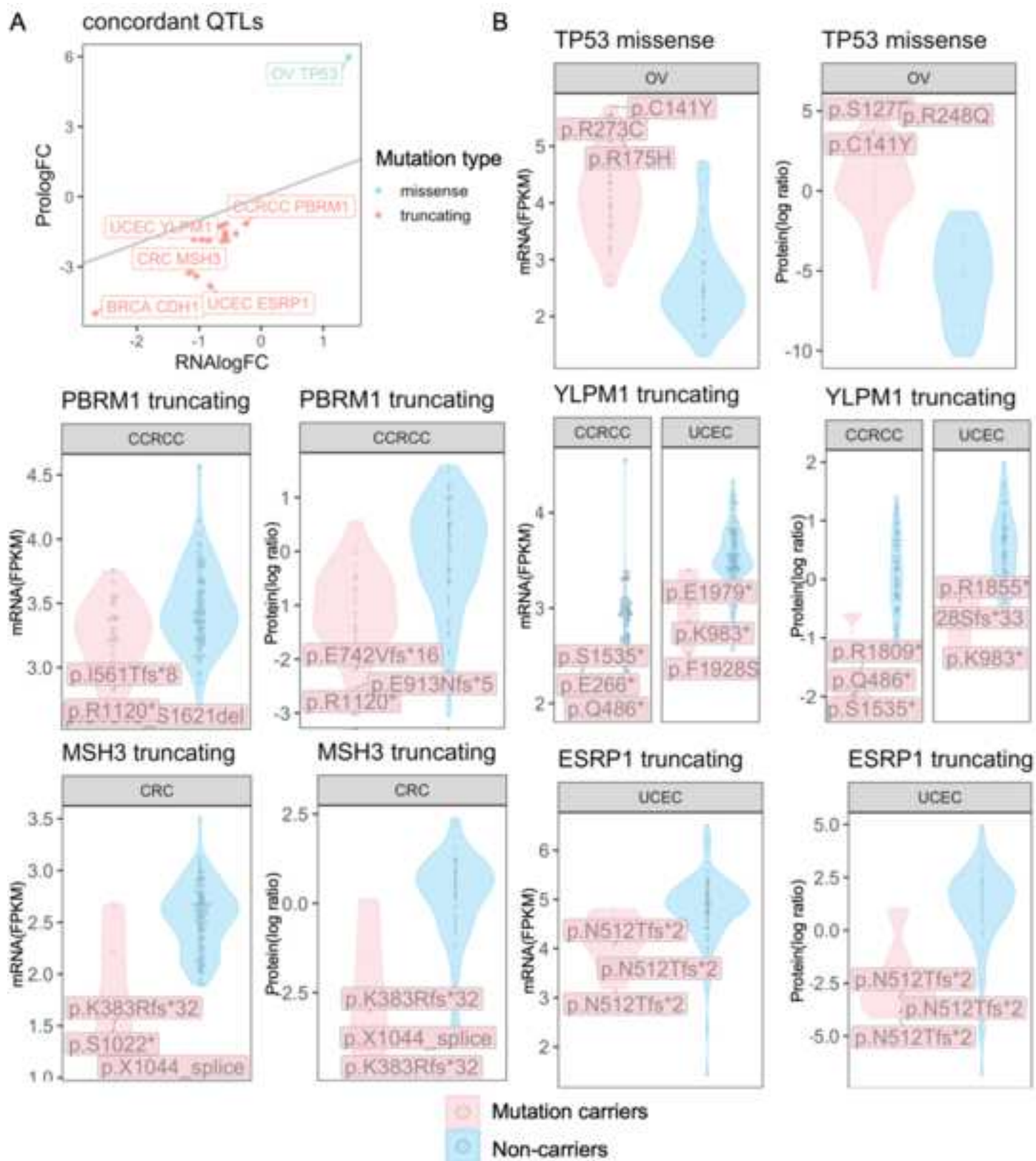
A



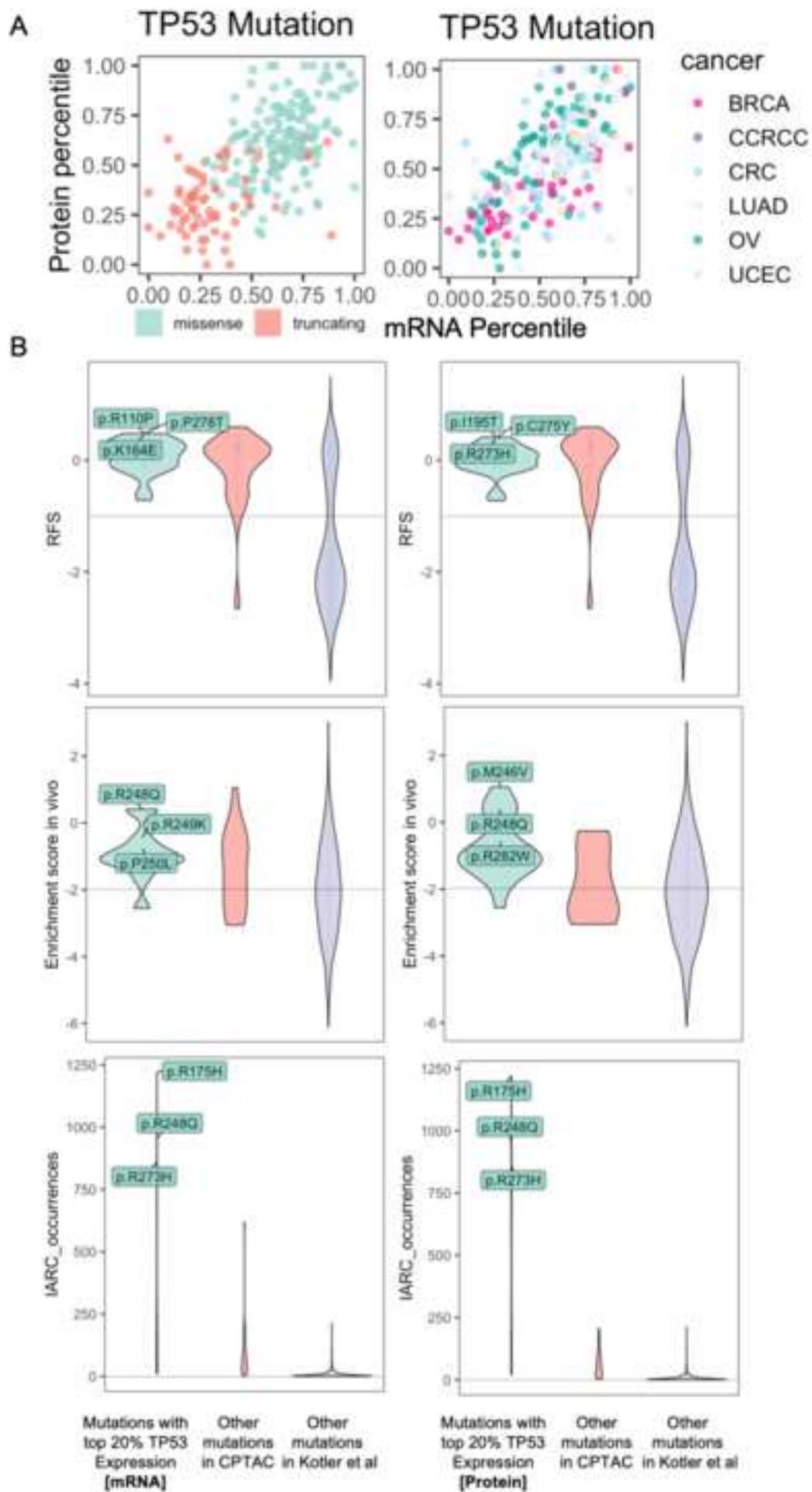
B

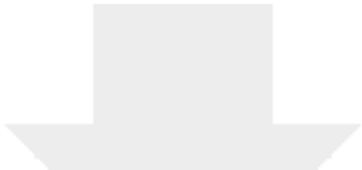
Cancer Type	Breast Cancer	Clear Cell Renal Cell Carcinoma	Colorectal Cancer	Lung Adenocarcinoma	Ovarian Cancer	Uterine Corpus Endometrial Carcinoma
Abbreviation	<b>BRCA</b>	<b>CCRCC</b>	<b>CRC</b>	<b>LUAD</b>	<b>OV</b>	<b>UCEC</b>
Data Source	Krug et al. 2020 (PMID: 33212010)	Clark et al. 2019 (PMID: 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%












Click here to access/download  
**Supplementary Material**  
SuppFigures.pdf







Click here to access/download  
**Supplementary Material**  
TableS1.eQTLs.xlsx

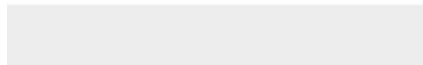


Click here to access/download  
**Supplementary Material**  
TableS2.pQTLs.xlsx





Click here to access/download  
**Supplementary Material**  
TableS3.concordant\_e.pQTLs.xlsx





Click here to access/download  
**Supplementary Material**  
TableS4.significant\_spsQTLs.xlsx





Click here to access/download  
**Supplementary Material**  
TableS5.DEP\_pairedTN\_stats.xlsx





[Click here to access/download](#)

**Supplementary Material**

[TableS6.TP53\\_mutation\\_test\\_statistics.xlsx](#)





[Click here to access/download](#)

**Supplementary Material**

[TableS7.CorCoefs.ConcordantAndDiscordant.xlsx](#)





Kuan-lin Huang, PhD  
Assistant Professor, Department of Genetics and Genomic Sciences  
Institute of Genomics and Multiscale Biology  
Icahn School of Medicine at Mount Sinai

1399 Park Avenue (Room 4-420C)  
Box 1498  
New York, NY 10029  
Phone: (212) 824-6134  
Email: [kuan-lin.huang@mssm.edu](mailto:kuan-lin.huang@mssm.edu)  
Web: [ComputationalOmicsLab.org](http://ComputationalOmicsLab.org)

Nov 21<sup>th</sup> 2024

Hans Zauner  
Editor, GigaScience

Dear Dr. Zauner,

We would like to re-submit our manuscript titled "*Mutation Impact on mRNA Versus Protein Expression across Human Cancers*" (GIGA-D-24-00168), addressing all the editorial comments:

- 1) Please include a citation to your new GigaDB dataset (including the DOI link) to your reference list, and cite this in the data availability section. **Completed.**
- 2) Please structure your abstract ("Background - Results - Conclusions") **Completed.**
- 3) Please submit the manuscript text without embedded figures. Please upload the figures separately in Editorial Manager (one file per figure in good resolution - please refer to the formatting instructions on our homepage). **Completed.**
- 4) Please move URLs to the bibliography and cite them by reference number, rather than including them in the text directly (e.g. line 562, 570, 571). (We treat internet sources as citable objects). **Completed.**
- 5) Please rename the "code availability" section as "Availability of supporting source code and requirements" and use this tabular format: **Completed.**
- 6) For reference numbers in the text, please use square brackets (e.g. [1]) instead of superscript numbers. **Completed.**
- 7) Please also note the reviewer's comment below regarding the equation - although it looks fine in the Word document I'm looking at now, I believe this was just a conversion problem in





Kuan-lin Huang, PhD  
Assistant Professor, Department of Genetics and Genomic Sciences  
Institute of Genomics and Multiscale Biology  
Icahn School of Medicine at Mount Sinai

1399 Park Avenue (Room 4-420C)  
Box 1498  
New York, NY 10029  
Phone: (212) 824-6134  
Email: [kuan-lin.huang@mssm.edu](mailto:kuan-lin.huang@mssm.edu)  
Web: [ComputationalOmicsLab.org](http://ComputationalOmicsLab.org)

the PDF. **Yes it looked fine on our end, attached also the PDF we converted ourselves which looked fine.**

8) Please also ensure that your revised manuscript conforms to the journal style, which can be found in the Instructions for Authors on the journal homepage **Completed.**

Sincerely and on behalf of the team,

Kuan-lin Huang, Ph.D.  
Associate Professor of Genetics and Genomic Sciences & Artificial Intelligence and Human Health  
Icahn School of Medicine at Mount Sinai  
New York, NY 10029