1   **Spatially Interacting Phosphorylation Sites and Mutations in Cancer**

2   Kuan-lin Huang[1#], Adam D. Scott[2], Daniel Cui Zhou[2], Liang-Bo Wang[2], Amila
3   Weerasinghe[2], Abdulkadir Elmas[1], Ruiyang Liu[2], Yige Wu[2], Michael C. Wendl[2], Matthew
4   A. Wyczalkowski[2], Jessika Baral[2], Sohini Sengupta[2], Chin-Wen Lai[3], Kelly Ruggles[4],
5   Samuel H. Payne[5], Benjamin Raphael[6], David Fenyö[4], Ken Chen[7], Gordon Mills[8], Li
6   Ding[2#]
7
8   [1]Department of Genetics and Genomics, Tisch Cancer Institute, Icahn Institute for
9   Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New
10   York, NY 10029
11   [2]Department of Medicine, McDonnell Genome Institute, Department of Genetics,
12   Siteman Cancer Center, Washington University in St. Louis, MO 63108, USA.
13   [3]Department of Pathology and Immunology, Washington University in St. Louis, MO
14   63108, USA.
15   [4]Center for Health Informatics and Bioinformatics, New York University School of
16   Medicine, New York, NY 10016, USA.
17   [5] Department of Biology, Brigham Young University, Provo, UT 84602, USA.
18   [6]Lewis-Sigler Institute, Princeton University, Princeton, NJ 08544, USA.
19   [7]Departments of Bioinformatics and Computational Biology, The University of Texas MD
20   Anderson Cancer Center, Houston, TX 77030, USA.
21   [8]Knight Cancer Institute, Oregon Health & Science University, Portland, OR 97239,
22   USA.
23
24   # Corresponding Authors:
25
26   Kuan-lin Huang, Ph.D.
27   Department of Genetics and Genomic Sciences
28   Icahn School of Medicine at Mount Sinai
29   New York, NY 10029
30   Email: kuan-lin.huang@mssm.edu
31
32   Li Ding, Ph.D.
33   Division of Oncology, Department of Medicine
34   McDonnell Genome Institute
35   Washington University School of Medicine
36   St. Louis, MO 63108
37   Email: lding@genome.wustl.edu

38 **ABSTRACT**

39   Advances in mass-spectrometry have generated increasingly large-scale proteomics
40   datasets containing tens of thousands of phosphorylation sites (phosphosites) that
41   require prioritization. We develop a bioinformatics tool called HotPho and systematically
42   discover 3D co-clustering of phosphosites and cancer mutations on protein structures.
43   HotPho identifies 474 such hybrid clusters containing 1,255 co-clustering phosphosites,
44   including RET p.S904/Y928, the conserved HRAS/KRAS p.Y96, and IDH1 p.Y139/IDH2
45   p.Y179 that are adjacent to recurrent mutations on protein structures not found by linear
46   proximity approaches. Hybrid clusters, enriched in histone and kinase domains,
47   frequently include expression-associated mutations experimentally shown as activating
48   and conferring genetic dependency. Approximately 300 co-clustering phosphosites are
49   verified in patient samples of 5 cancer types or previously implicated in cancer,
50   including CTNNB1 p.S29/Y30, EGFR p.S720, MAPK1 p.S142, and PTPN12 p.S275. In
51   summary, systematic 3D clustering analysis highlights nearly 3,000 likely functional
52   mutations and over 1,000 cancer phosphosites for downstream investigation and
53   evaluation of potential clinical relevance.

54

55 **KEYWORDS**

56
57   proteomics; genomics; phosphorylation sites; cancer
58

59 **INTRODUCTION**

60 Dysregulated phosphorylation of oncogenic proteins alters pathway activity and contributes to
61 tumor phenotypes[1,2]. Recent advances in mass-spectrometry have generated increasingly
62 large-scale proteomics datasets in multiple cancer types[3,4], each containing tens of thousands
63 of phosphosites that urgently require prioritization. Missense somatic mutations and
64 phosphorylations, independently or through mutual interactions, can affect the physicochemical
65 properties of the residue side chains and modulate protein functions or stability in oncogenic
66 pathways. Thus far, mutation and phosphorylation have been largely studied in isolation by
67 genomics and proteomics approaches. Integrated methodologies are required to reveal their
68 interactions and prioritize both types of events with functional significance.

69 Previous works highlighted the potential functionality of mutations that are linearly adjacent to
70 phosphosites in cancer driver genes[5–7], yet these studies did not consider the 3-dimensional
71 structures of proteins. We and others previously demonstrated that mutations in cancer genes
72 form 3-dimensional (3D) spatial clusters—defined by high local concentrations of mutations on
73 protein structures—enriched for functional missense mutations[8–10]. We hypothesize that co-
74 clustering mutations and phosphosites in spatial hotspots will also enrich for functional events of
75 both categories. Systematic analyses of mutations from sequencing data and phosphosites from
76 global proteomics data will enable us to investigate beyond currently-interrogated phosphosites
77 with available targeting antibodies and reveal functionalities of phosphosites.

78 Here, we report on the development and application of a bioinformatics tool called HotPho to
79 systematically discover spatial interactions of mutations and phosphosites. We find 474
80 significant hybrid clusters (defined as clusters containing both co-clustering phosphosites and
81 mutations) that prioritize 1,255 phosphosites and 2,938 mutations on protein structures from
82 large-scale proteomics and genomics data. Many co-clustering mutations are associated with
83 high functional scores, expression changes, and known recurrent/activating events that expose
84 genetic dependency; whereas many co-clustering phosphosites are found in kinase domains
85 and verified in primary tumor samples. We specifically prioritize phosphosites co-clustering with
86 activating mutations of BRAF, EGFR, and PIK3CA. Collectively, our approach of 3D spatial
87 clustering on protein structures systematically highlights likely functional mutations and
88 phosphosites for downstream investigation.

89   **RESULTS**

90   **HotPho Algorithm and Performance**

91   Extending beyond the originating framework of an earlier mutation-clustering tool we developed,

92   namely HotSpot3D[8], HotPho enables investigation of proximal and structural information of

93   phosphosites with their neighboring mutations and domains, both on a single protein structure

94   or co-crystallized binding partners in a protein complex (**Figure 1**). Briefly, all missense variants

95   and phosphosites are considered as nodes and their 3D distances as edges on an undirected

96   graph and the clusters are built up using the Floyd–Warshall shortest-paths algorithm

97   implemented by HotSpot3D[8] (**Methods**).

98   We demonstrated the capability of HotPho for identifying co-clustering cancer mutations and

99   phosphosites using data comprised of 225,151 unique phosphosites from PTMcosmos compiled

100   from multiple databases and CPTAC cancer proteomic cohorts[3,4] (**Methods**). We also included

101   791,637 missense mutations of 9,097 samples across 33 cancer types from a filtered set of

102   Multi-Center Mutation Calling in Multiple Cancers project (MC3) mutation calls from the TCGA

103   PanCanAtlas[11], taken in account their recurrence in the MC3 cohort. Both mutations and

104   phosphosites are mapped by HotPho and analyzed based on 5,950 processed human proteins

105   from UniProt[12] having at least one PDB structure.

106   To assess whether the co-clustering between aforementioned sets of mutations and

107   phosphosites is non-random, we analyzed the clusters against a set of permutated data as

108   follows: the original mutation backbone was maintained while phosphosites were randomly

109   populated 100 times, keeping the corresponding ratios of residue types of phosphosites

110   constant (**Methods**). We found a higher fraction of hybrid clusters in the original HotPho output

111   (8.1%) at the top 5% of the cluster closeness score compared to the null distribution from the

112   permutations (**Figure 1**). We defined the criteria of high-confidence clusters to have cluster

113   closeness scores within the top 5% of their respective cluster types and subsequently limited

114   our analyses to these clusters. In hybrid clusters, the 5% sensitivity corresponded to 97.4%

115   specificity in a receiver operating characteristic (ROC) curve analysis (AUC = 0.58,

116   **Supplementary Fig. 1A**).

117   We conducted a multitude of analyses to investigate the modality in the score distribution and

118   the implication of using the 5% threshold. First, while this threshold (cluster closeness score =

119   2.56) may permit false-positives if the simulated phosphosites only contain negatives, we

120   observed many of the clusters containing activating or recurrent mutations with cluster
121   closeness scores close to the threshold (**Supplementary Data 1**). It is possible that the spatial
122   distribution of cancer mutations and commonality phosphosite residues (i.e., serine, threonine,
123   and tyrosine) is not random and thus retaining these additional hybrid clusters is needed to
124   minimize false-negatives. Second, to resolve possible reasons underlying the multi-modal
125   distribution of cluster closeness scores, we compared the score distributions for 299 mutation-
126   enriched cancer driver genes[13] versus other genes. While hybrid clusters involving driver genes
127   showed a higher density at the higher-score mode, driver gene status did not guarantee high
128   scores (**Supplementary Fig. 1B**). The 5% score threshold showed a sensitivity = 0.17 and
129   specificity = 96.0% in distinguishing hybrid clusters with driver genes (**Supplementary Fig. 1C**).
130   Finally, we examined the score distribution using 200 bins on both the simulated vs. observed
131   clusters, finding multiple peaks and alternative thresholds, for example, thresholding using one
132   of the higher local minima retained only the top 2.28%, or the top 216 clusters (**Supplementary**
133   **Fig. 1D**). Cluster closeness scores for all identified clusters are provided herein to prioritize a
134   more stringent set of clusters (**Supplementary Data 1**).

135   **Co-clustering of Phosphosites and Mutations using HotPho**

136   HotPho generated a final high-confidence set of 906 mutation-only, 127 phosphosite-only, and
137   474 hybrid clusters based on the top 5% cluster closeness score threshold (**Methods,**
138   **Supplementary Data 1**). Top genes harboring each type of cluster varied (**Figure 2A**): MGAM,
139   SI, ERBB3, and LRRC4C each had at least 9 mutation-only clusters and such type of clusters
140   have been previously characterized[8–10]. Phosphosite-only clusters are found in fewer instances:
141   ANXA5, CLIP1, FLNB, GPI, HSPD1, PEBP1, and PTK2 each harbored two (**Supplementary**
142   **Fig. 2A**).

143   For subsequent analyses, we focused on investigating hybrid clusters and properties of this
144   entity found across 474 unique proteins (some proteins only form hybrid clusters with their
145   protein complex partners). Notably, the highest counts of hybrid clusters were found for genes
146   known for recurrent mutations, including TP53 (10 hybrid clusters), PIK3CA (8), CTNNB1 (6),
147   EGFR (6), and other genes involved in cancers, such as HIST1H2BC (6) and PLG (5) (**Figure**
148   **2A**). These clusters comprise a total of 1,255 phosphosites and 2,938 mutations. The
149   composition of co-clustered phosphosites differs across gene products; co-clustered tyrosines
150   are most commonly observed in PIK3CA and EGFR kinases, whereas serines are most
151   common in HIST1H2BC and HBG2 (**Figure 2A**). The top hybrid clusters of each protein—

152 identified by the highest cluster closeness score—may span mutations and phosphosites that
153 are far from one another in the linear distance (**Figure 2C, Supplementary Fig. 2B**).
154 Phosphosites prioritized in these clusters include CTNNB1 p.T40, EGFR p.T290, ERBB2
155 p.T733/T759, KIT p.Y578, and TP53 p.T284. We also compared the mutations in the hybrid
156 clusters to those found in a clustering analysis using only TCGA MC3 mutations, which
157 contained 9,403 clustered mutations. Among the 2,938 mutations found in the 474 hybrid
158 clusters, we found only 48 mutations not found by mutation-only clustering. The list of 48
159 mutations contained mutations of interest in PDE1B (5 mutations), SRSF7 (4 mutations), and
160 PTPN12 p.S275F/C that co-localized with p.S275 and co-clustered with p.S39/p.T40
161 (**Supplementary Data 2**).

162 Among the 1,255 co-clustered phosphosites, 291 sites directly overlap and 356 sites are
163 proximal (within 2 amino acid residues linearly) to their co-clustered mutations (**Supplementary
164 Data 1**, **Figure 2B**). The HotPho co-clustering analysis adds a substantial count of 608
165 phosphosites which are distant in terms of a linear sequence, yet close in 3D protein structure,
166 including the majority of the sites found on ACTB, HIST1H2BC, and ERBB2. Nearly half of the
167 clusters we identified can only be found by integrating 3D protein structure, demonstrating the
168 added value of 3D approaches for the discovery of spatial relationships between mutations and
169 phosphosites.

170 We then examined whether proteins containing hybrid clusters are enriched in specific biological
171 pathways curated by WikiPathways[14] and the NCI-Nature Pathway Interaction Database[15] using
172 Enrichr[16] (**Supplementary Data 3**, **Supplementary Fig. 3**, **Methods**). The most enriched NCI-
173 Nature pathways include PDGFR-beta, ErbB2/ErbB3, ErbB1, Hepatocyte Growth Factor
174 Receptor (c-Met), SHP2, Fc-epsilon receptor I, and mTOR signaling pathways (Fisher's Exact
175 Test, adjusted $P < 1E-12$), which is reaffirmed by the Focal Adhesion-PI3K-Akt-mTOR-signaling
176 pathway being one of the top enriched WikiPathways (adjusted $P = 8.8E-16$). These findings
177 suggest the possible involvement of hybrid clusters and co-clustering phosphosites in oncogenic
178 signaling pathways.

179 We further hypothesized that hybrid clusters would be enriched in functional domains related to
180 oncogenic processes[17]. Mapping residues to PFAM domains, we identified 26 PFAM domains
181 significantly enriched for mutations and phosphosites in hybrid clusters when comparing to the
182 background of all mapped mutations and phosphosites (**Supplementary Data 4**,
183 **Supplementary Fig. 3**, **Methods**). Domains of histone proteins, including centromere protein

184   Scm3, core histone H2A/H2B/H3/H4, and centromere kinetochore component CENP-T histone

185   fold showed the most significant enrichment (Fisher's Exact Test, FDR ≤ 1.53E-43). Another two

186   top PFAM domains are protein tyrosine kinase and protein kinase domains (**Supplementary**

187   **Fig. 3**). Specifically, we identified hybrid clusters in tyrosine kinase domains of tyrosine kinases

188   (TK), such as MET, FGFR2/3, and ERBB2/3, and in BRAF of the tyrosine kinase-like (TKL)

189   group. Other hybrid clusters involving sites at protein kinase domains included TGFBR1 of the

190   TKL group, MAP2K4 of the STE group, and AKT1/2 of the AGC group. Notably, some kinase-

191   domain clusters showed conserved mutation/phosphosite patterns across homologs, such as

192   FGFR2 and FGFR3 (**Supplementary Fig. 3**).

193

194   **Co-clustering Phosphosites Adjacent to Known Activating Mutations**

195   To prioritize candidate phosphosites, we first investigated phosphosites co-clustering with

196   known functional cancer mutations. We curated experimentally validated mutations from the

197   Cancer Biomarkers database with Cancer Genome Interpreter[18], OncoKB[19], and KinDriver[20],

198   collecting a total of 367 activating mutations (**Methods**). We found 29 hybrid clusters containing

199   90 of these activating mutations in 17 genes, suggesting the functional relevance of the 54 co-

200   clustering phosphosites (**Supplementary Data 5**). PIK3CA and EGFR are each involved in 4

201   hybrid clusters containing activating mutations and such clusters are also found in CTNNB1 (3),

202   KIT (3), BRAF (2), ERBB2 (2), KRAS (2), MET (2), and NRAS (2).

203   Phosphosites co-clustering with activating mutations are likely of functional relevance. We

204   specifically examined these clusters on protein structures (**Figure 3B, Supplementary Fig. 5**).

205   Both ERBB2 p.T733 and p.T759 are located adjacently to the activating mutation p.L755W.

206   NRAS phosphosite p.Y64 is co-clustered with two of the most recurrently mutated residues

207   p.G12 and p.Q61. Receptor tyrosine kinases, KIT, MET, and RET all harbor phosphorylated

208   tyrosine sites co-clustering with activating mutations. These prioritized phosphosites include KIT

209   p.Y578, MET p.Y1093/Y1159/Y1230, and RET p.Y928. Two hybrid clusters containing

210   activating mutations were found on a protein complex formed by PIK3CA/PIK3R1: PIK3R1

211   phospho-tyrosines p.Y470 and p.Y556 clustered with activating mutations PIK3CA p.N344G/M,

212   p.N345K, p.C420R, and PIK3R1 p.N564D. In the other hybrid cluster, PIK3R1 p.T463 clustered

213   with activating mutations PIK3CA p.E453K/Q. The co-clustering phosphosites next to known

214   activating mutations are promising targets for further investigation, along with their adjacent

215   mutations.

216   We hypothesized that phosphosites co-clustering with highly-recurrent mutations in a cancer
217   cohort might imply functionality in the specific cancer type. We calculated the frequency of each
218   of the co-clustering mutations within each of the TCGA cancer cohorts and identified their
219   spatially adjacent phosphosites (**Supplementary Data 6**). We found that co-clustering
220   phosphosites of the most recurrent mutations aggregate in proteins including CTNNB1, HRAS,
221   IDH1, KRAS, NRAS, PIK3CA, and TP53 (**Figure 3A, Supplementary Fig. 5**). In PIK3CA, we
222   identified p.T957 co-clustering with the highly recurrent p.H1047R that affects many gynecologic
223   cancer cases, including 13.8% of breast invasive carcinoma (BRCA), 7% of uterine
224   carcinosarcoma (UCS), and 5.8% of uterine corpus endometrial carcinoma (UCEC). In TP53,
225   phosphosites p.R249 and p.T284 co-cluster with p.R273C/H that affects 11.4% of brain lower-
226   grade glioma (LGG), 5.3% of UCS, and 3.8% of esophageal carcinoma (ESCA); TP53 p.T155
227   and P.S183 co-cluster with p.R175H that affects 8.3% of rectum adenocarcinoma (READ), 6.3%
228   of colon adenocarcinoma (COAD), 6% of ESCA, 3.7% of ovarian serous cystadenocarcinoma
229   (OV), and 3.5% of UCS.

230   Many phosphosites co-clustering with recurrent mutations were found in protein homologs.
231   IDH1 phosphosites p.Y135 and p.Y139 co-clustered with p.R132H, which is highly recurrent in
232   brain tumors (73.6% of LGG and 6.1% of glioblastoma multiforme [GBM]), as well as p.R132C
233   implicated in several cancer types (17.1% of cholangiocarcinoma [CHOL], 4.3% of acute
234   myeloid leukemia [LAML], 3.4% of LGG, and 3.2% of skin cutaneous melanoma [SKCM]). In its
235   homolog protein IDH2, p.Y179 co-clustered with p.R140Q affecting 6.5% of LAML (**Figure 3**).
236   For the Ras proteins, KRAS/HRAS/NRAS all harbor highly recurrent mutations for residues
237   p.G12/G13 that affect large fractions of pancreatic adenocarcinoma (PAAD), COAD, READ,
238   lung adenocarcinoma (LUAD), and UCEC. Each harbors overlapping yet distinct sets of co-
239   clustering phosphosites—KRAS p.S89/p.Y96, NRAS p.Y64, and HRAS p.Y32/T35/Y64/Y96—
240   warranting further investigation into their potentially shared and distinct signaling functions
241   across cancer types (**Figure 3**).

242

243   **<u>Functional Evidence for Co-clustering Mutations</u>**

244   We evaluated whether HotPho can effectively prioritize functional mutations in hybrid clusters by
245   comparing with functional scores predicted by VEST[21], Mutation Assessor[22], PolyPhen2[23],
246   SIFT[24], and a composite Eigen score composed of all these scores[25] (**Figure 4A**). Within
247   proteins harboring hybrid clusters, clustered mutations showed strikingly higher predicted

248   functional scores compared to other mutations in the same proteins (Wilcoxon Rank Sum Test,
249   P < 2.2e-16), supporting the view that co-clustered mutations should be prioritized.

250   To further demonstrate that hybrid clusters enrich for functional mutations, we examined
251   whether clustered mutations are associated with the protein or phosphoprotein changes, as
252   previously found for functional and pathogenic mutations[26,27]. Using the TCGA RPPA dataset for
253   each of the 33 cancer types, we conducted a differential expression analysis to search for
254   protein/phosphoprotein markers expressed at different levels in carriers of clustered mutations
255   (**Methods**), identifying 24 significant (FDR < 0.05, linear regression) gene-cancer associations
256   (**Figure 4B, Supplementary Data 7**). *TP53* mutations in hybrid clusters are significantly
257   correlated with higher p53 protein expression in 14 cancer types, most strikingly in UCEC,
258   BRCA, COAD, and OV, consistent with the previously reported *cis*-effect of functional TP53
259   missense mutations[28]. Clustered EGFR mutations are likewise associated with higher EGFR
260   protein and EGFR p.Y1068 expression in LGG, GBM, and LUAD, cancer types largely affected
261   by activating mutations of *EGFR*. We also found clustered *KIT* mutations to be associated with
262   higher c-Kit in TGCT and SKCM (**Figure 5B**).

263   At a single residue level, we noted clustered mutations showing high protein or phosphoprotein
264   expressions above the 95th percentile in the same cancer cohort (**Figure 4C**). These include
265   recurrent TP53 mutations p.R248Q/W, p.R273H/C/L, p.R175H, p.R282W/G, and p.R337C. Top
266   mutations in EGFR differ between brain and lung tumors: in LGG, EGFR p.R108K, p.R252C/P,
267   and p.R263I, which are adjacent to the phospho-threonine p.T290, are associated with high
268   EGFR protein and phosphoproteins. Many samples with top EGFR expression in GBM also
269   carry mutations in the same hybrid cluster, p.A289V/T and p.R252C/P, whereas in LUAD the
270   associated mutation is the recurrent EGFR p.L858R (4.3% of LUAD) co-clustering with
271   phosphosites p.Y869 and p.Y891. KIT mutation p.D816V and p.829P is associated with high c-
272   kit and it clustered with p.Y362 and p.Y823 (**Figure 4C**).

273   To further validate these findings, we conducted similar analyses of the mutational impact on
274   protein expression using global proteomics datasets from retrospective and prospective CPTAC
275   cohorts of breast, ovarian, and colorectal cancers (2 cohorts/cancer) each comprised of 78~126
276   samples (**Methods, Supplementary Data 8, Supplementary Fig. 6**). Given the limited sample
277   sizes, no associations passed multiple testing thresholds (FDR < 0.05) and only suggestive
278   associations (Wilcoxon rank-sum test, P < 0.05) are highlighted herein: we validated that TP53
279   co-clustering mutations associated with higher protein expression in all three cancer types. In

280 colorectal cancer, *KRAS* mutations in cluster 9458.0 affecting p.G12, p.G13, p.V14, and p.Q61

281 are associated with higher KRAS expression, whereas *HNF4A* mutations in cluster 7977.0 are

282 associated with low HNF4A expression. Other notable findings include that *ESR1* mutations in

283 cluster 1357.1 are associated with high ESR1 in ovarian cancer, whereas *AKT1* mutations in

284 cluster 756.0 (**Supplementary Fig. 6**) are associated with high AKT1 proteins.

285 Finally, given the potential functionality of co-clustering mutations, we characterized the

286 mutational landscape across cancer types in the TCGA MC3 dataset of ~10 thousand tumors[11].

287 By considering missense mutations (1) directly overlapping phosphosites, (2) proximal to

288 phosphosites, and (3) co-clustering with phosphosites, we noticed that considering co-clustering

289 mutations contribute significantly to the fraction of potentially functional mutations in many

290 cancer genes including *EGFR*, *KRAS*, and *PIK3CA* (**Supplementary Fig. 7**).

291

292 **<u>Functional Verification of Co-clustering Mutations</u>**

293 Finally, we test the hypothesis that hybrid clusters enrich for functional mutations, including

294 those shown to be activating and confer genetic dependency by cancer cells. To test for

295 activating mutations that confer clonal selection advantages, we assessed experimental data

296 from 1,054 somatic mutations in the Ba/F3 and MCF10A cells[29], including 549 found in genes

297 with hybrid clusters. Out of the 549 unique somatic mutations, 86 co-clustered with

298 phosphosites and 463 did not. There was a striking enrichment of activation in hybrid mutations

299 co-clustering with phosphosites. For mutations functionally assessed in Ba/F3, 77.6% (66/85) of

300 the co-clustering mutations were determined as activating compared to only 30.2% (138/457) of

301 the other mutations determined as activating (One-tailed Fisher's Exact Test, P = 2.66E-16,

302 **Figure 5A**). For mutations functionally assessed in MCF10A, 67.6% (46/68) of the co-clustering

303 mutations were activating compared to only 35.2% (146/415) of the other mutations being

304 activating (P = 5.03E-7, **Figure 5B**). Collectively 72 co-clustered phosphosites were determined

305 as activating in either Ba/F3 or MCF10A.

306 We then examined whether the co-clustered mutations show significant enrichment of activating

307 mutations compared to the other mutations on a gene-by-gene basis (**Figure 5A-B**,

308 **Supplementary Data 9**). Co-clustering mutations of PIK3CA are significantly enriched for

309 activating events in both cell lines (One-tailed Fisher's Exact Test, P ≤ 1.22E-3), with 33/35 co-

310 clustering mutations being validated in Ba/F3 and 21/22 in MCF10A; its binding partner PIK3R1

311 also shows suggestive enrichment in Ba/F3 (P = 0.072, 3/3). Significant enrichment of activating

312   mutations was also observed for EGFR (P = 1.63E-4, 10/11) and BRAF (P = 1.11E-3, 12/18)

313   co-clustering mutations in Ba/F3. In MCF10A, we also noted suggestive associations for BRAF,

314   where 7/18 co-clustering mutations are activating (P = 0.068), and ESR1, where 7/18 co-

315   clustering mutations are activating (P = 0.068). The enrichment of activating mutations in hybrid

316   clusters suggests structural adjacency to phosphosites implies functional significance in

317   oncogenes.

318   To evaluate the added predictive power of mutation functionality provided by phosphosite co-

319   clustering, we examined the relationships between mutation functionality versus co-clustering

320   mutation counts, mutation recurrence, and co-clustering with phosphosites (**Figure 5C-D**). First,

321   using a multivariate logistic regression model corrected for the mutated gene, we found that the

322   number of co-clustering mutations was not significantly associated with the mutation

323   functionality in either the BAF3 (P = 0.91) or MCF (P = 0.40) cell line (**Supplementary Fig. 9A**).

324   Second, using a multivariate logistic regression model corrected for the mutated gene, we found

325   that the recurrence of mutations in the TCGA MC3 cohort was significantly associated with the

326   mutation functionality in both the BAF3 (P = 2.8e-3) or MCF10A (P = 0.023) cell lines. But, when

327   adding the phosphosite co-clustering status to the regression model, the mutation functionality

328   was no longer associated with recurrence (P>0.21), but strongly associated with the co-

329   clustering status in both BAF3 (P = 1.4e-10) and MCF10A (P = 1.5e-4) cell lines (**Figure 5C-D**,

330   **Supplementary Fig. 9A**). Altogether, these results suggest that spatial co-clustering with

331   phosphosites may improve prediction of mutation functionality beyond the commonly used

332   mutation recurrence.

333   Next, we sought to test whether the co-clustered mutations may confer genetic dependency to

334   the mutated cancer cells. In this case, cancer cells with co-clustered mutations would show

335   higher vulnerability in a CRISPR knockout screen targeting the mutated genes than cells with

336   other mutations. To test this hypothesis, we utilized data using characterized by the CRISPR-

337   knockout screens in the Cancer Dependency Map (DepMap) project[30], where a negative

338   CERES dependency score indicates genetic dependency of the cancer cell. Within each of the

339   27 tested lineages, we carried out a Wilcoxon Rank Sum test between the cell lines with co-

340   clustered missense mutations versus those with other missense mutations (**Methods**). Strikingly,

341   cancer cell lines with co-clustered mutations showed significantly higher dependency (or more

342   vulnerability upon genetic knockout) than those with missenses in 14 lineages (Wilcoxon Rank

343   Sum test, FDR < 0.05), most notably lung, colorectal, skin, pancreas, and gastric cancer cells

344   (FDR ≤ 3.3E-7, **Figure 5E**). We also obtained similar results when comparing cell lines with co-

345   clustered missense mutations versus other non-synonymous mutations (**Supplementary Data**

346   **10**). Overall, these analyses showed that co-clustered mutations adjacent to phosphosites are

347   enriched for activating events and highlight genetic vulnerability of cancer cells.

348   **<u>Verification of Co-clustering Phosphosites</u>**

349   To verify co-clustering phosphosites, we sought evidence of these events being observed in the

350   CPTAC proteomic cohorts of prospective primary tumor samples of 123 breast invasive

351   carcinoma (BRCA), 83 ovarian carcinoma (OV), 97 colorectal adenocarcinoma (CRC), 103

352   uterine corpus endometrial carcinoma (UCEC), and 41 clear cell renal cell carcinoma (CCRCC).

353   Of the 1,255 co-clustered phosphosites, 259 were detected in at least one of the 5 cohorts

354   (**Figure 6A, Supplementary Data 11**). Some phosphosites may be cancer-type specific: we

355   uniquely observed BRAF p.S467/Y472 in BRCA and p.S465 in UCEC. For TP53, we uniquely

356   observed p.S149 in breast cancer versus TP53 p.T150 in ovarian cancer. Other phosphosites

357   are found in multiple cancer types (**Figure 6B**); for instance, AKT1 p.T308 was seen in both

358   breast and ovarian cancers. Notably, phosphosites (p.S29/Y30/T40/T41/T42) near the section-

359   terminus of the CTNNB1 protein were commonly seen in breast and ovarian cancers (p.T41

360   was also observed in UCEC), and p.S675 was detected in substantial numbers of samples in all

361   5 cancers (**Figure 6B**). Co-clustering tyrosine phosphosites, PIK3R1 p.Y452/556/580, on the

362   other hand, were observed in endometrial and renal cancer (p.Y580 was also observed in

363   colorectal cancer). As a cautionary note, given the different reference samples and mass

364   spectrometry runs for each cancer cohort, the cancer-specific phosphosites observed herein

365   require further validation. Nonetheless, the detection of the co-clustering phosphosites in

366   primary tumor samples further implicates their functionality in oncogenesis.

367   Finally, we conducted a systematic literature review of co-clustering phosphosites regulated or

368   implicated in cancer (**Methods**), finding 25 unique phosphosites across 18 proteins that were

369   experimentally linked to cancer (**Supplementary Data 12**). These include AKT1 p.T308 and

370   CTNNB1 p.T41 found in CPTAC patient tumors, as well as sites with known kinase regulations,

371   including EGFR p.S768/Y869/Y1016, ESR1 p.Y537, and TP53 p.S376/378. Other co-clustering

372   phosphosites showing functionality related to cancer include: MAPK1 p.S142 that was

373   previously shown to be critical to the ERK2 docking domain and its mutated form p.S142L

374   confers gain-of-function[31]; CTCF p.T374 that, along with a few nearby residues, were shown to

375   be phosphorylated during mitosis and to decrease its DNA-binding activity[32]; BRAF p.T599 and

376   p.S602 that are conserved from *C. elegans* to mammals and required for activation of the B-Raf

377 kinase[33,34], and RB1 (Rb) p.S567 that is uniquely phosphorylated by MAPK11 (p38), triggering

378 Rb-Hdm2 interaction and apoptosis[35]. These findings further validate the functionality of

379 selected co-clustering phosphosites HotPho identified and suggest other sites in hybrid clusters

380 may be prioritized for downstream investigations.

381

DISCUSSION

383 We describe the first systematic discovery of co-clustering mutations and phosphosites on 3D

384 protein structures, a feat enabled by a bioinformatics tool—HotPho. HotPho successfully

385 identifies likely functional mutational clusters and phosphosites in known cancer proteins

386 including EGFR, KIT, and KRAS/HRAS/NRAS, many of which are in kinase domains (**Figure 2**).

387 Co-clustering mutations in these clusters have higher predicted functional scores, increased

388 protein/phosphoprotein levels (**Figure 4**), and are experimentally validated as being functional

389 and confer genetic dependency by cancer cells (**Figure 5**). Concurrently, co-clustering

390 phosphosites show multiple characteristics supporting their contributions in oncogenesis,

391 including co-clustering with validated activating and recurrent mutations in multiple cancer types

392 (**Figure 3**) and are detected in patient tumors (**Figure 6**). Co-clustering events may represent

393 potential drivers and therapeutic targets.

394 Proteomics datasets, such as those generated by CPTAC, are quantifying increasingly larger

395 numbers of phosphoproteomes in cancer and other samples. The abundant phosphosites of

396 unknown significance (PUS) discovered in these datasets highlight the urgent need for

397 enhanced annotation and prioritization using approaches like HotPho. There are still significant

398 limitations to identifying functional hybrid clusters, as prioritization necessarily relies on known

399 mutations or functional domains. Thus, while we also discovered phosphosite-only clusters, we

400 cannot yet effectively determine their significance until we enhance our understanding of

401 functional phosphosites.

402 Our investigation supports the functional relevance of co-clustering phosphosites and mutations.

403 For example, we found that these phosphosites and mutations are enriched in functional

404 domains of kinases and histones, that co-clustering mutations tend to be functional active and

405 confer genetic dependency. Among the 1,255 co-clustering phosphosites, only 25 were

406 previously known to be associated with cancer (**Supplementary Data 12**). Hopefully, the

407   repertoire of characterized phosphosites will grow rapidly using combinations of high throughput

408   proteomics approaches, systematic *in silico* analysis, and experimental validation[36,37].

409   More, multiple co-clustering phosphosites were located in activation loops of kinase proteins,

410   including RET p.S904, MET p.Y1248, AKT1 p.T308, EGFR p.Y827/869, as well as the MAPK3-

411   regulated site MAPK8 p.Y185. Crystal structures revealed that PTPN12 p.S275 is found in the

412   Q loop that constitutes the pY+1 pocket demonstrating strong substrate specificity, and the

413   phospho-inhibitory mutant p.S275A significantly decreased the activity of PTPN12 toward all

414   three HER2 phospho-peptides[38]. Notably, this site also harbored a rare mutation p.S275C in the

415   TCGA MC3 mutation data, which only showed clusters when leveraging the adjacent

416   phosphosite information but was not found in the mutation-only clusters.

417   Spatially co-clustering phosphosites and mutations may interact and exhibit further associations

418   in patient samples. Currently, we only observed a handful of examples where a single tumor

419   sample carries both of the co-clustering phosphosites and mutations in existing quantitative

420   phosphoproteomics cohorts, precluding systematic investigations of their correlative

421   relationships (**Supplementary Fig. 9**). The growing cohort size of CPTAC and other cancer

422   global phosphoproteomic datasets will likely enable us to test the intriguing hypothesis that

423   samples carrying mutations will show disrupted regulation or mutual exclusivity of a co-

424   clustering phosphosite with sufficient statistical power.

425   In conclusion, we conduct a large-scale spatial clustering analysis between 225,151

426   phosphosites and 791,637 missense mutations using the HotPho tool. The resulting 474 hybrid

427   clusters help us discover 1,255 phosphosites co-clustering with mutations in human cancer,

428   dozens of which are adjacent to activating mutations and verified in patient tumor samples. Our

429   approach nominates phosphosites of likely functional significance for experimental validation

430   and may be expanded to investigate other post-translational modifications, such as acetylation

431   and glycosylation.

432   **METHODS**

433   **Data Sources**

434   **Phosphosites data**

435   We gathered 225,151 human phosphorylation sites from PTMcosmos, following a procedure

436   similar to our recently published study[39]. PTM sites from PTMcosmos were retrieved from

437   UniProt Knowledge Base (UniProtKB) version 2018.01, PhosphoSitePlus (snapshot on the date
438   2018-02-14), and CPTAC2 MS phosphoproteome data. A PTM site was included if it satisfied
439   either of the following: (1) included in UniProtKB and was reported in at least one publication or
440   by sequence similarity. (2) included in PhosphoSitePlus and was reported in at least one
441   publication or validated internally by Cell Signaling Technology. (3) included in CPTAC2
442   experiments and was detected in at least one of the samples. To match phosphosites between
443   multiple databases, we used transvar[40] to map amino acid residues on different protein isoforms
444   to their unique genomic positions.

Somatic mutation data

446   We used somatic mutations from the TCGA cohort as provided by the publicly-available MC3[11]
447   mutation annotation file (MAF) (syn7824274). These mutations were further filtered based on
448   flagged artifacts, hypermutators, and pathology to a driver discovery dataset of 9,097 samples
449   with 791,637 missense mutations, as described in the recent PanCanAtlas somatic driver
450   paper[25].

Known activating somatic mutations

452   We curated experimentally validated mutations identified as neutral or activating from multiple
453   databases and papers, including the Cancer Biomarkers database within the Cancer Genome
454   Interpreter[18], OncoKB[19], KinDriver[20], and ClinVar[41]. We subsequently required an activating
455   mutation to be seen in at least 2 of these sources, collecting a total of 367 activating mutations.

456   PDB structures

457   We used the GRCh37 assembly and Ensembl release 75 to preprocess residue pair data for all
458   human proteins in RCSB PDB as of 22 May 2017, which includes PDB structures of 6,002
459   genes.

460   Some chains or structures from PDB were filtered out due to the following types of artifacts in
461   the data file annotations, 1) chains with inconsistent PDB to UniProt coordinate ranges from
462   DBREF given any alterations from SEQADV length changes, 2) chains where SEQADV
463   describes REMARK 999, which indicates absent residues explained by free text, and 3)
464   structures in which site mismatches were identified (for example, where a Threonine should be
465   found according to the UniProt sequence, but a Valine was instead found at the position in the

466 PDB sequence) even after converting between the PDB and UniProt coordinates designated by
467 the DBREF line.

468

469 **Bioinformatics and Experimental Analyses**

470 Quality control of sites and structures

471 HotPho reads a site input file where each phosphosite must contain the HUGO symbol,
472 Ensembl transcript accession ID (ENST), its residue position within the given transcript, and
473 feature description. The sites are then combined and run through the HotSpot3D search step to
474 produce pairwise data between mutations and phosphosites, comprising mutation-mutation
475 pairs, mutation-site pairs, and site-site pairs. Even though HotPho calculates offsets in residue
476 numbers in PDB structures and transcripts, some offsets provided by structure uploaders
477 resulted in the erroneous mapping of residues. In the resulting pairwise files with phosphosites
478 (.musite and .sites files), we therefore filtered out the sites where the mapped residue on the
479 PDB structure differs from those documented in our original input phosphosite file, ultimately
480 retaining 785,867 mutation-mutation pairs, 376,614 mutation-site pairs (78,477 eliminated), and
481 1,010,011 site-site pairs (267,547 eliminated).

482 The HotPho algorithm and cluster discovery

483 HotPho extends beyond the originating HotSpot3D algorithm[8] and enables co-clusterings of
484 mutations and phosphosites on protein structures (**Figure 1**). Briefly, 3D distances of all
485 missense mutations and phosphosites were calculated using PDB structures, considering the
486 closest atoms on their respective amino acids on PDB structures, to identify proximity pairs.
487 Each potential cluster is then treated as an undirected graph G = (V,E), where V is a subset of
488 the input mutations and phosphosites and E is the set of proximal pairs from V. Considering v_i,
489 $v_j \in V$ for i, j $\in$ {1,2,…,N} where N is the number of vertices in V, the clusters are built up using
490 the Floyd–Warshall shortest-paths algorithm, initiated by the distance matrix of the edges, to
491 obtain the geodesics, $g_{i,j}$ between each v_i and v_j. For each $v_i \in V$ where i ≠ j, the cluster centrality,
492 $c(v_i)$, is then calculated as:

$$c(v_i) = \sum_{j=1}^{N} \frac{1}{2^{g_{i,j}}} \qquad (1)$$

493   For each of the cluster, the centroid is identified as the vertex showing the highest $c(v_i)$, and the

494   cluster closeness score (Cc) is calculated as:

$$Cc = \sum_{i=1}^{N} c(v_i) \qquad (2)$$

495   A high Cc score indicates a dense 3D cluster enriched in recurrent mutations and phosphosites

496   on the protein structure. In the pan-cancer study of mutation-only clusters, clusters with known

497   cancer proteins showed significantly higher Cc score than those without cancer-related proteins,

498   and a threshold at top 5% showed a notably significant difference between cancer- and non-

499   cancer-related proteins[8]. Here in our hybrid cluster analysis, we not only show the top 5%

500   clusters show sensitivity in distinguishing cancer driver genes vs. other genes, but also in

501   observed vs. randomly simulated clusters.

502   We then conducted clustering using HotPho with recurrence as the vertex type. The analysis

503   generated a total of 30,131 unfiltered clusters in 4989 unique genes, comprising 9,483 hybrid

504   clusters, 18,112 mutation-only clusters, and 2,536 site-only clusters. To resolve the multi-modal

505   distribution of cluster closeness scores, we further compared the score distributions for 299

506   mutation-enriched cancer driver genes[13] versus other genes using a ROC curve analysis

507   (**Supplementary Fig. 1B-C**).

508   Cluster benchmarking using permutation analysis

509   The hybrid clusters generated by HotPho was benchmarked by comparison to those obtained

510   through HotPho analyses using a combination of the TCGA MC3 mutation data and permutated

511   phosphosite data. Since we are interested in the hybrid clusters having high cluster closeness

512   (Cc) scores (ie. more closely packed clusters) we chose the top 100 genes having high CC

513   clusters. Next, for each of these genes, we found the number of phosphosites in the original

514   dataset which are covered by at least one structure. After that, we generated a permutated

515   phosphosites-dataset by randomly populating the sites at possible covered phosphosite residue

516   locations keeping the original residue ratios the same. HotPho clustering was performed for 100

517   such simulated phosphosites-datasets and the maintained TCGA MC3 mutation call backbone

518   given the non-random distribution and occurrence count of mutation calls. Finally, the clusters

519   from the original HotPho run and the simulated runs were compared focusing on the number of

520   clusters and their Cc score distribution. We further evaluated the sensitivity, specificity, and

521   ROC curves using different threshold of the Cc score (**Supplementary Fig. 1D**). Based on our

522 simulated results, we set the cluster closeness thresholds as the top 5% cluster closeness score

523 within each cluster type (e.g., hybrid, mutation-only, and phosphosite-only).

524

525 **Domain enrichment analysis**

526 We conducted a domain enrichment analysis of co-cluster phosphosites in PFAM domains

527 (Pfam 31.0 released Mar 2017)[42]. We evaluated domain enrichment of co-clustered

528 phosphosites using the Fisher test. Each 2x2 table was comprised of tallies of domain status

529 (current domain or not) versus co-clustered status (co-cluster with mutation or not). Although

530 this test is exact, we followed the general rule-of-thumb for table testing of only evaluating those

531 cases where there were at least 5 mutations and phosphosites in the domain. Resulting P-

532 values were corrected to FDR values using the standard Benjamini-Hochberg procedure.

533

534 **Mutational impact RPPA analysis**

535 Similar to our previous analyses of a different set of mutations[27], TCGA level-3

536 normalized RPPA expression data of the tumor samples were downloaded from

537 Firehose (2016/1/28 archive). The protein/phosphoprotein expression percentile of

538 individual proteins in each cancer cohort was calculated using the empirical cumulative

539 distribution function (ecdf), as implemented in R. Where there are at least 3 carriers

540 within each cancer type, we then applied the linear model to evaluate the

541 protein/phosphoprotein expression percentile between carriers of co-clustered

542 mutations and all other cancer cases. The age at initial diagnosis, gender, and ethnicity

543 are included as covariates to account for potential confounding effects. The resulting P

544 values were adjusted to FDR using the standard Benjamini-Hochberg procedure for

545 tests across all cancer types.

546

547 **Mutational impact proteome analysis**

548 We analyzed the effects of clustered mutations using samples from the CPTAC2

549 retrospective[3,4,43]   and   prospective   collection   (https://cptac-data-

550   portal.georgetown.edu/cptac/public). For each hybrid cluster, protein levels were

551   compared between samples with and without cluster mutations (Wilcoxon rank sum

552   test).

553

554   **Cancer cell dependency analyses**

555   Within each lineage, we carried out a Wilcoxon Rank Sum test between the cell lines with co-

556   clustered missense mutations versus those with either (1) other missense mutations, or (2)

557   other non-synonymous mutations. The p-values represent the alternative hypothesis that the

558   dependency score distribution of the cell lines with co-clustered mutations is located left (more

559   dependent, more vulnerable) of that of without co-clustered mutations, and they are multiple-

560   testing adjusted using the BH method for FDR.

561

562   **Literature reviews of cancer-associated phosphosites**

563   First, we confined our search space to 71 cancer genes with hybrid clusters by limiting

564   our search space to 299 cancer driver genes[25]. The abstracts of all publications

565   associated with a phosphosite were then retrieved from Europe PMC using their Digital

566   Object Identifier (DOI) or PubMed identifier (PMID). We determined a paper to be

567   cancer-related if its abstract contained the keyword 'tumor' and/or 'cancer'. We then

568   closely examined whether the exact co-clustering phosphosites identified by HotPho

569   showed any alterations on cancer-related phenotypes in these publications.

570   Additionally, we included all disease-associated sites in PhosphoSitePlus (snapshot on

571   the date 2018-02-14) that were connected to any type of cancer.

572   **Figure legends:**

573   Figure 1. HotPho workflow and performance benchmarks. (a) HotPho takes user-provided lists

574   of mutations and phosphosites as inputs, map them onto PDB protein structures, calculates

575   each of the pairwise distances, conducts clustering, and reports clusters with prioritized

576   mutations and phosphosites. (b) Comparison of HotPho results measured using phosphosite

577   and MC3 cancer mutation data vs. simulated data of randomly distributed phosphosites. The left

578   panel indicates the density of cluster closeness (CC) scores for all hybrid clusters in the HotPho

579   run and the simulated runs, where the vertical line indicates the top 5% score threshold. The

580   right panel shows the bar plot comparing the number of hybrid clusters passing the same 5%

581   threshold in the HotPho and simulated runs.

582

583   Figure 2. Hybrid clusters containing both phosphosites and mutations. (a) Left shows the counts

584   of hybrid clusters, mutation-only clusters, and site-only clusters in genes with at least two hybrid

585   clusters. Right barplot shows counts of each type of phospho-residue, being serine (S),

586   threonine (T), tyrosine (Y), or Arginine (R), found in hybrid clusters for each of the genes. (b)

587   Spatial interactions of co-clustered mutations and phosphosites. For each of the genes, we

588   counted how many of the co-clustered mutations and phosphosites are also directly overlapping

589   (Direct) or within 2 amino acid residues (Proximal), and without any of these apparent linear

590   relationships (Clustered) to phosphosites and mutations, respectively, in the same hybrid

591   clusters. (c) Phosphosite and mutations on the linear protein coordinate of top hybrid clusters as

592   defined by cluster closeness scores in each of the highlighted genes (excluding HIST1H4G and

593   H3F3A due to their top hybrid clusters included phosphosites from other proteins and excluding

594   MGAM due to its top clusters located at residue coordinate beyond the plotted range [centroid at

595   1514]). The Ensembl transcripts used for mapping of the protein coordinates are described in

596   Supplementary Data 1.

597

598   Figure 3. Hybrid clusters highlighting co-clustering phosphosites adjacent to recurrent mutations
599   across cancer types. (a) The frequency of co-clustering mutations within each TCGA cancer
600   type. For each of the cancer types shown in a distinct color, the mutation with the highest
601   recurrence is labeled along with its co-clustering phosphosites. The size of the dot to indicate
602   the pan-cancer score calculated by the averaging the frequencies of the mutation across the 32
603   cancer types in TCGA. (b) Selected hybrid clusters with activating mutations shown on 3D
604   protein structures obtained through PDB. Mutations are colored in shades of blue and
605   phosphosites are colored in shades of red.

606

607   Figure 4. Proteomic effects associated with co-clustered mutations. (a) Comparison of predicted
608   functional scores, including those provided by Mutation Assessor, PolyPhen-2, SIFT, VEST, and
609   an eigen score for mutations having different spatial interactions with phosphosites. The
610   interaction types are direct (directly overlapping), proximal (within 2 residues in the linear
611   distance), clustered (in hybrid clusters), and none of these interactions. (b) Plot showing cancer
612   types where the carrier of each gene's co-clustered mutation is associated with significantly
613   higher or lower expression of the corresponding protein/phosphoprotein marker in the RPPA
614   dataset. Each dot represents a gene-cancer association, where color depicts cancer type and
615   shape shows significance. (c) Expression percentile of the protein/phosphoprotein marker in
616   carriers of multiple types of mutations (direct, proximal, clustered, none) in the corresponding
617   genes. Mutations carried by the samples with greater than 97% marker expression are further
618   text-labeled.

619

620 Figure 5. Functional assessment of co-clustering mutations. Experimental validation data of co-
621 clustering somatic mutations were extracted from previous systematic assessments in (a) Ba/F3
622 and (b) MCF10A cell lines[29]; we evaluated the functionality of 1,054 somatic mutations in a cell
623 viability assay, where each of the evaluated mutations were assessed through one transfected
624 cell colony compared to control cell colonies. The number of activating mutations vs. other types
625 of mutations in each gene was then calculated for the set of co-clustering mutations adjacent to
626 phosphosites as discovered by HotPho and other non-clustering mutations. The asterisk
627 indicates the significance of the association (One-tailed Fisher's Exact Test, *** P < 0.005, **
628 0.005 <= p < 0.05, * 0.05 <= p < 0.1). The gene products showing significant associations
629 include EGFR (P = 1.63E-4) and BRAF (P = 1.11E-3) co-clustering mutations in Ba/F3, and
630 PIK3CA in Ba/F3 (P = 1.67E-5) and MCF10A (P = 0.0012). At a single mutation level, the
631 functional status (color codes) between co-clustered versus other mutations are further shown
632 for (c) Ba/F3 and (d) MCF10A cell lines against the mutation recurrence on the Y-axis,
633 demonstrating the additional predictive power of co-clustering status on mutation functionality.
634 (e) Dependency CERES score comparison of cell lines with co-clustered vs. other missense vs.
635 other non-synonomous mutations in the DepMap CRISPR screen dataset. The 10 lineages with
636 the highest numbers of co-clustered mutations are shown, including cell lines of the bile duct
637 (N=30), blood (N=44), central nervous system (N=61), colorectal (N=36), gastric (N=27), lung
638 (N=107), ovary (N=43), pancreas (N=34), skin (N=54), and uterus (N=26) tissues. The centre
639 line corresponds to the median; lower and upper hinges correspond to the first and third
640 quartiles (the 25th and 75th percentiles), respectively; the whiskers extend from the hinges to
641 the largest value no further than 1.5 * IQR from the respective hinge, where IQR (inter-quartile
642 range) is the distance between the first and third quartiles.

643 Figure 6. Verification of co-clustering phosphosites in primary tumors. (a) Verification of co-
644 clustering phosphosites of cancer proteins in patient tumor samples characterized by the
645 CPTAC prospective projects. Each dot represents the phosphosites being detected in the
646 cancer cohort colored-coded by cancer type. The size of the dot represents the standard
647 deviation of the phosphosite level in the respective cancer cohort. (b) Lolliplots showing number
648 of tumor samples (number in circle) where the co-clustering phosphosites of AKT1, CTNNB1,
649 TP53, and PIK3R1 proteins are detected in each cancer cohort out of 123 breast invasive
650 carcinoma (BRCA), 83 ovarian carcinoma (OV), 97 colorectal adenocarcinoma (CRC), 103
651 uterine corpus endometrial carcinoma (UCEC), and 41 clear cell renal cell carcinoma (CCRCC).

ACKNOWLEDGEMENTS

DATA AVAILABILITY

The TCGA somatic mutation data are available at the Genome Data Commons (GDC) [https://gdc.cancer.gov/about-data/publications/mc3-2017/]. CPTAC data are available at the CPTAC Data Portal [https://cptac-data-portal.georgetown.edu/]. The CPTAC datasets used in this study include the CPTAC2 prospective breast [https://cptac-data-portal.georgetown.edu/study-summary/S039], ovarian [https://cptac-data-portal.georgetown.edu/study-summary/S038], and colorectal [https://cptac-data-portal.georgetown.edu/study-summary/S037] cancer studies, as well as the CPTAC3 discovery discovery endometrial cancer [https://cptac-data-portal.georgetown.edu/study-summary/S053] and clear cell renal carcinoma [https://cptac-data-portal.georgetown.edu/study-summary/S050] studies. Phosphosite data from PTMCosmos are available at https://ptmcosmos.wustl.edu/.

CODE AVAILABILITY

HotPho is released within the package of HotSpot3D v1.7.0 and updated in later versions [available on GitHub: https://github.com/ding-lab/hotspot3d][44]. HotPho commands and all subsequent analyses scripts used for this manuscript are publicly available [https://github.com/ding-lab/HotPho_Analysis]. Analyses were conducted based on scripts written using the R programming language version 3.3.2 (2016-10-31).

AUTHOR CONTRIBUTIONS

L.D., K.H., and A.D.S. conceived the idea and designed the HotPho workflow. A.D.S. and A.W. developed the HotPho algorithm and conducted the simulation analysis. K.H. and L.D. designed analyses plans. K.H., D.C.Z., L.W., A.E., R.L., Y.W., J.B., S.S., A.D.S., and A.W. analyzed the data. K.H., D.C., M.A.W., C.L., and R.L. created figures. K.H. wrote the manuscript. L.D., K.C.,

681   M.C.W., K.R., S.H.P., B.R., D.F., and G.M. edited the manuscript. G.M. provided experimental

682   validation data. L.D. supervised research. All authors reviewed and approved the manuscript.

683   **COMPETING INTERESTS**

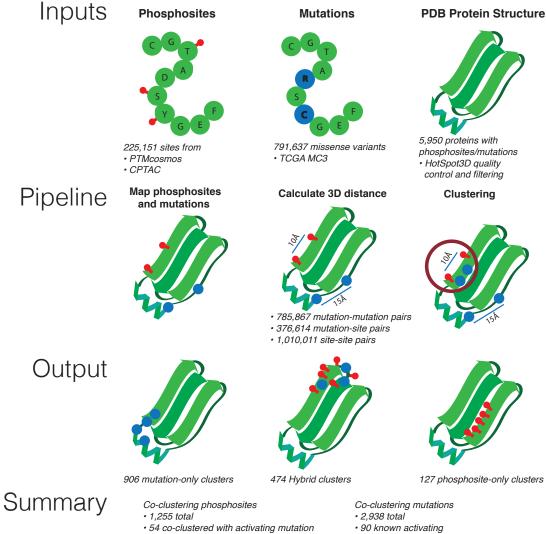684   The authors declare no competing interests.

685

REFERENCE

1. Sever, R. & Brugge, J. S. Signal transduction in cancer. *Cold Spring Harb. Perspect. Med.* (2015) doi:10.1101/cshperspect.a006098.

2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* (2011) doi:10.1016/j.cell.2011.02.013.

3. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **advance on**, (2016).

4. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755–765 (2016).

5. Reimand, J., Wagih, O. & Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* **3**, 2651 (2013).

6. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).

7. Krassowski, M. *et al.* ActiveDriverDB: Human disease mutations and genome variation in post-translational modification sites of proteins. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gkx973.

8. Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* (2016) doi:10.1038/ng.3586.

9. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U. S. A.* (2015) doi:10.1073/pnas.1516373112.

10. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. & Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol.* (2015) doi:10.1371/journal.pcbi.1004518.

11. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* (2018) doi:10.1016/j.cels.2018.03.002.

12. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkw1099.

13. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* (2018) doi:10.1016/j.cell.2018.03.033.

14. Slenter, D. N. *et al.* WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gkx1064.

15. Schaefer, C. F. *et al.* PID: The pathway interaction database. *Nucleic Acids Res.* (2009) doi:10.1093/nar/gkn653.

16. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

17. Porta-Pardo, E. & Godzik, A. E-Driver: A novel method to identify protein regions driving cancer. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu499.

18. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* (2018) doi:10.1186/s13073-018-0531-8.

19. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* (2017) doi:10.1200/po.17.00011.

20. Simonetti, F. L., Tornador, C., Nabau-Moreto, N., Molina-Vila, M. A. & Marino-Buslje, C. Kin-Driver: A database of driver mutations in protein kinases. *Database* (2014) doi:10.1093/database/bau104.

21. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* (2013) doi:10.1186/1471-2164-14-s3-s3.

22. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkr407.

23. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* (2013) doi:10.1002/0471142905.hg0720s76.

24. Ng, P. C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* (2002) doi:10.1101/gr.212802.

25. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* (2018) doi:10.1016/j.cell.2018.02.060.

26. Ding, J. *et al.* Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.* (2015) doi:10.1038/ncomms9554.

27. Huang, K. lin *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* (2018) doi:10.1016/j.cell.2018.03.039.

28. Davidoff, a M., Humphrey, P. a, Iglehart, J. D. & Marks, J. R. Genetic basis for p53 overexpression in human breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 5006–5010 (1991).

29. Ng, P. K. S. *et al.* Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* (2018) doi:10.1016/j.ccell.2018.01.021.
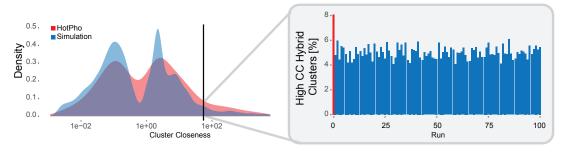
757   30.   Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* (2017)
758   doi:10.1016/j.cell.2017.06.010.

759   31.   Brenan, L. *et al.* Phenotypic Characterization of a Comprehensive Set of
760   MAPK1/ERK2 Missense Mutants. *Cell Rep.* (2016)
761   doi:10.1016/j.celrep.2016.09.061.

762   32.   Sekiya, T., Murano, K., Kato, K., Kawaguchi, A. & Nagata, K. Mitotic
763   phosphorylation of CCCTC-binding factor (CTCF) reduces its DNA binding
764   activity. *FEBS Open Bio* (2017) doi:10.1002/2211-5463.12189.

765   33.   Zhang, B.-H. Activation of B-Raf kinase requires phosphorylation of the conserved
766   residues Thr598 and Ser601. *EMBO J.* (2000) doi:10.1093/emboj/19.20.5429.

767   34.   Köhler, M. *et al.* Activation loop phosphorylation regulates B-Raf in vivo and
768   transformation by B-Raf mutants . *EMBO J.* (2016)
769   doi:10.15252/embj.201592097.

770   35.   Delston, R. B., Matatall, K. A., Sun, Y., Onken, M. D. & Harbour, J. W. P38
771   phosphorylates Rb on Ser567 by a novel, cell cycle-independent mechanism that
772   triggers Rb-Hdm2 interaction and apoptosis. *Oncogene* (2011)
773   doi:10.1038/onc.2010.442.

774   36.   Huang, K. *et al.* Regulated Phosphosignaling Associated with Breast Cancer
775   Subtypes and Druggability. *Mol. Cell. Proteomics* (2019)
776   doi:10.1074/mcp.ra118.001243.

777   37.   Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat.*
778   *Biotechnol.* (2020) doi:10.1038/s41587-019-0344-3.

779   38.   Li, H. *et al.* Crystal Structure and Substrate Specificity of PTPN12. *Cell Rep.*
780   (2016) doi:10.1016/j.celrep.2016.04.016.

781   39.   Dou, Y. *et al.* Proteogenomic Characterization of Endometrial Carcinoma. *Cell*
782   (2020) doi:10.1016/j.cell.2020.01.026.

783   40.   Zhou, W. *et al.* TransVar: A multilevel variant annotator for precision genomics.
784   *Nature Methods* (2015) doi:10.1038/nmeth.3622.

785   41.   Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant
786   variants. *Nucleic Acids Res.* **44**, D862-8 (2015).

787   42.   Finn, R. D. *et al.* The Pfam protein families database: Towards a more
788   sustainable future. *Nucleic Acids Res.* (2016) doi:10.1093/nar/gkv1344.

789   43.   Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal
790   cancer. *Nature* **513**, 382–7 (2014).

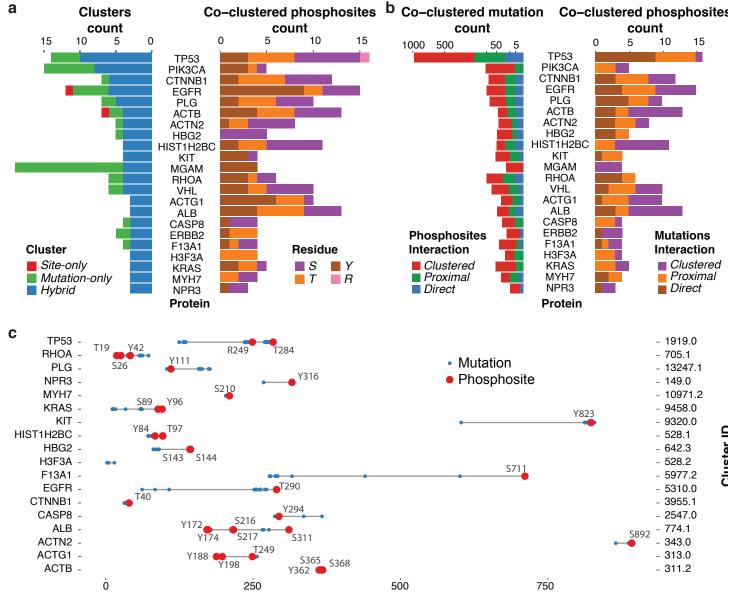791   44.   Niu, B. et al. HotSpot3D, *kuanlinhuang/hotspot3d,* doi:10.5281/zenodo.4539901,
792   (2021)

793

a

b Prioritized Mutation & Phosphosites