Computational and Structural Biotechnology Journal 19 (2021) 4759–4769

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Protein-gene Expression Nexus: Comprehensive characterization of human cancer cell lines with proteogenomic analysis

Daejin Hyung [a], Min-Jeong Baek [a], Jongkeun Lee [a], Juyeon Cho [a], Hyoun Sook Kim [a], Charny Park [a,*], Soo Young Cho [a,b,*]

[a] National Cancer Center, 323 Ilsan-ro, Goyang-si, Gyeonggi-do 10408, Republic of Korea
[b] Department of Molecular and Life Science, Hanyang University, Ansan 15588, Republic of Korea

ARTICLE INFO

Article history:
Received 28 February 2021
Received in revised form 13 August 2021
Accepted 14 August 2021
Available online 17 August 2021

Keywords:
Proteogenomics
Single amino acid variation
Phosphoproteomics
Systems biology
Cancer cell line

ABSTRACT

Researchers have gained new therapeutic insights using multi-omics platform approaches to study DNA, RNA, and proteins of comprehensively characterized human cancer cell lines. To improve our understanding of the molecular features associated with oncogenic modulation in cancer, we proposed a proteogenomic database for human cancer cell lines, called Protein-gene Expression Nexus (PEN). We have expanded the characterization of cancer cell lines to include genetic, mRNA, and protein data of 145 cancer cell lines from various public studies. PEN contains proteomic and phosphoproteomic data on 4,129,728 peptides, 13,862 proteins, 7,138 phosphorylation site-associated genomic variations, 117 studies, and 12 cancer. We analyzed functional characterizations along with the integrated datasets, such as cis/trans association for copy number alteration (CNA), single amino acid variation for coding genes, post-translation modification site variation for Single Amino Acid Variation, and novel peptide expression for noncoding regions and fusion genes. PEN provides a user-friendly interface for searching, browsing, and downloading data and also supports the visualization of genome-wide association between CNA and expression, novel peptide landscape, mRNA-protein abundance, and functional annotation. Together, this dataset and PEN data portal provide a resource to accelerate cancer research using model cancer cell lines. PEN is freely accessible at http://combio.snu.ac.kr/pen.

1. Introduction

In the past decade, the number of large-scale multi-omics studies in human cancer models has increased rapidly [1–3]. The integration of these datasets has revolutionized biology and led to the emergence of systems-based approaches to advance our understanding of biological processes. Human cancer cell lines are valuable tumor model systems, which are widely used in target research and drug discovery. The systematic study of cancer cell lines provided by multi-omics platforms is founded on novel insights for many different molecular features, including genetic variation, expression abundance, and alterations in cellular signaling. Recently, proteogenomics, which integrates proteomics with genomics and transcriptomics, has emerge as a promising approach to tumor profiling with the potential to advance basic, translational, and clinical research. To accelerate the understanding of human cancer, several proteogenomic studies have been proposed for various tumor types [4]. In particular, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) brings together expertise in proteomics, genomics, cancer biology, and clinical chemistry, while creating open community resources that are widely used in cancer research [4].

Comprehensive characterization of human cancers is still sparse in proteogenomic databases, but raw proteomic data for diverse cancer types are available. Many databases, such as PRIDE ProteomeXchange, ProteomicsDB, and GEO, have been developed for open resources and public storage of proteomic and genomic datasets [3,5–7]. These databases usually contain a large amount of raw multi-omics data. Other studies have provided novel insights into protein and mRNA expression. The Human Proteome Map portal contains data on protein expression of 17,000 human genes from 17 adult tissues, six primary hematopoietic cells, and seven fetal tissues [8]. Wilhelm et al. proposed a database of mass-spectrometry-based protein expression, called ProteomicsDB, and proposed a comparison of mRNA and protein expression in 12 human tissues [9]. The cis/trans effect was compared to functional

* Corresponding authors at: National Cancer Center, 323 Ilsan-ro, Goyang-si, Gyeonggi-do 10408, Republic of Korea (S.Y. Cho).
E-mail addresses: charn78@ncc.re.kr (C. Park), sooycho@ncc.re.kr (S.Y. Cho).

https://doi.org/10.1016/j.csbj.2021.08.022

knockdown data in the library of the LINC database to identify candidate driver genes [10]. CellMiner provides a casual multi-omics resource and user interface for performing integrative, cross-database analyses of proteogenomic data for the NCI-60 cell lines [1], and ATLANTiC offers an interactive web application to explore protein abundance and drug response across the NCI-60 and CRC65 cell lines [2]. However, protein alterations and their associated genomic factors, and phosphoprotein regulation with its corresponding amino acid changes remain largely unexplored.

The proteogenomic approach proposes novel variants in an unbiased and accurate manner. The cis/trans association with copy number alteration (CNA) may provide oncogenic drivers in tumors, which often undergo modifications such as single amino acid variation (SAAV) [11–12]. Accordingly, several tools, including JUMPg, IPAW, and ProGeo-neo, have been developed to analyze pipelines for SAAV identification in proteogenomic data [13–15]. Comprehensive proteogenomics resources are necessary to understand their function in the context of various human cancer models.

Even though proteogenomics in human cancer has become the driving force in uncovering oncogenic effects, we still lack a comprehensive and integrated database for CNA profiles, gene expression patterns, phosphorylation profiles, as well as SAAV, novel, and fusion peptide profiles across various cancer types. Here, we introduce the Protein-gene Expression Nexus (PEN), which consolidates an extensive dataset of paired proteomics and genomics studies. The user interface is fully constructed with a dedicated proteomics browser and novel viewers that enable users to examine the mRNA-protein relationships with readily accessible expression correlation information. We describe the main characteristics of cancer proteogenomics in the following sections.

2. Materials and methods

2.1. Dataset selection

We constructed proteomics and phosphoproteomics datasets for eight gastric cancer cell lines and collected proteomics and phosphoproteomics datasets on human cancer cell lines from PRIDE, ProteomeXchange, and ProteomicsDB [3,5,7]. We removed conditional studies such as gene KO/transgenic, drug test, and other environmental factors. The proteomics datasets included 192,754,270 spectra, 138 cell lines, and 17 tissues. Additionally, we collected the phosphorylation profile of proteomics in cancer cell lines from the same databases (58 cell lines; 6,4631,573 spectra). Tissue information of the selected cell lines was annotated based on the primary site informed in the Cancer Cell Line Encyclopedia (CCLE) and COSMIC database. The transcriptomes and CNAs for cell lines were particularly useful for the comprehensive characterization of various cancer cell line types. The matched transcriptome and CNA data for cell lines used the processed output only from Depmap [16], not the raw sequence data.

2.2. Processing of proteomics data

The proteomics and proteogenomics datasets, including 85 label-free and 32 labeled studies, were searched using a cancer reference database, which is a customProDB-based tumor-specific protein database that supports the identification of novel peptides, SAAVs, and fusion genes [17]. We constructed a cancer reference database with a non-redundant custom protein sequence, built from UniProtKB-SwissProt, single nucleotide variations, non-coding genes, novel junctions, and fusion sequences. More information on genomic data resources is presented in Table 1.

In proteomic data, the PEN pipeline consisted of four major steps (Fig. 1 and Table 2): (i) The obtained fragmentation spectra were searched against the cancer reference database using MS-GF+, which is a searching tool that performs peptide identification by scoring MS/MS spectra against peptides derived from a protein sequence database [18]. Short peptides with fewer than six residues were removed, and a 1% peptide-level false detection rate (FDR) cutoff was used during peptide identification. Up to two mismatches were allowed in the search process to identify novel peptides. Peptide identification was performed with MS-GF + and the labeled data contained modification parameter. The same method was used in another step of our pipeline for processing label-free and labeled data. The fixed modifications in the labeled data were, for TMTdata, TMT-6plex on lysine residues and N-terminal amines, and carbamidomethylated cysteine; for iTRAQ-4plex, iTRAQ-4plex on lysine residues and N-terminal amines, and carbamidomethylated cysteine; for iTRAQ-8plex, iTRAQ-8plex on lysine residues and N-terminal amines, and carbamidomethylated cysteine; and for SILAC labeled data, 13C [6]15 N [4] on arginine residues, and carbamidomethylated cysteine. Finally, the variable modification for TMT, iTRAQ-4plex, iTRAQ-8plex, and SILAC labeled data was methionine oxidation. (ii) Peptide identification and data interpretation were performed using the percolator, which is an algorithm for improving the rate of confident peptide identification from a collection of tandem mass spectra [19], setting the FDR to 1% at all levels (protein, phosphopeptide, to spectrum matching). In step i, ii and iii, we used the cut-off thresholds and parameters of the CPTAC common data analysis pipeline [4]. (iii) The peptides from label-free studies identified with the known proteins from UniProtKB-SwissProt were used for protein expression quantitation using the Trans-Proteomic Pipeline's algorithm (StPeter) [20]. In labeled studies, we used the Libra and ASAPRatio algorithms for protein expression quantitation [20]. This procedure yielded 13,862 known proteins for 4,129,728 peptides. We constructed an expression pattern for 13,862 proteins from 124 cell lines. To harmonize the different studies, we used CCLE normalization methods [21]. All the proteome and transcriptome values were transformed with log2 ratio. Then, the proteins and genes values of the cell lines were mean centered to remove errors. Although mRNA and protein expression tend not to perfectly correlate, Nusinow et al. have found that RNA/protein correlation, as well as that of biological replicates, significantly improves with mean centered normalization procedures [22]. (iv) The remaining spectra were used to predict the novel SAAV and fusion peptides, using the two-stage FDR strategy [23]. Previous studies have found that the two-stage FDR control was more stringent than the separate FDR control strategy [24–26]. In the first stage, the proteomic data were searched against the reference protein database, and the confidently identified spectra with 1% FDR were removed. In the second stage, the remaining spectra were searched against the variant protein sequence database. The FDR estimation for variant peptides was based on the search results from the second stage. Extra filtration was performed using BLAST search for fusion peptides. We removed mapping peptide to known proteins

Table 1
Database list for searching and novel peptide analysis. The cancer reference database was constructed by CustomPRoDB.

Type	Database	Number of sequence or variation
Protein sequence	Uniprot (2019.02.01)	191,406
Variation	CanProVar 2.0	877,018
Variation	COSMIC release 87	1,312,882
Variation	CCLE	683,219
Pseudogene sequence	GENCODE release 19	79,484
lncRNA sequence	GENCODE release 19	71,694
lncRNA sequence	LNCipedia org v5.2	335,049
Fusion sequence	Fusion cancer	48,522

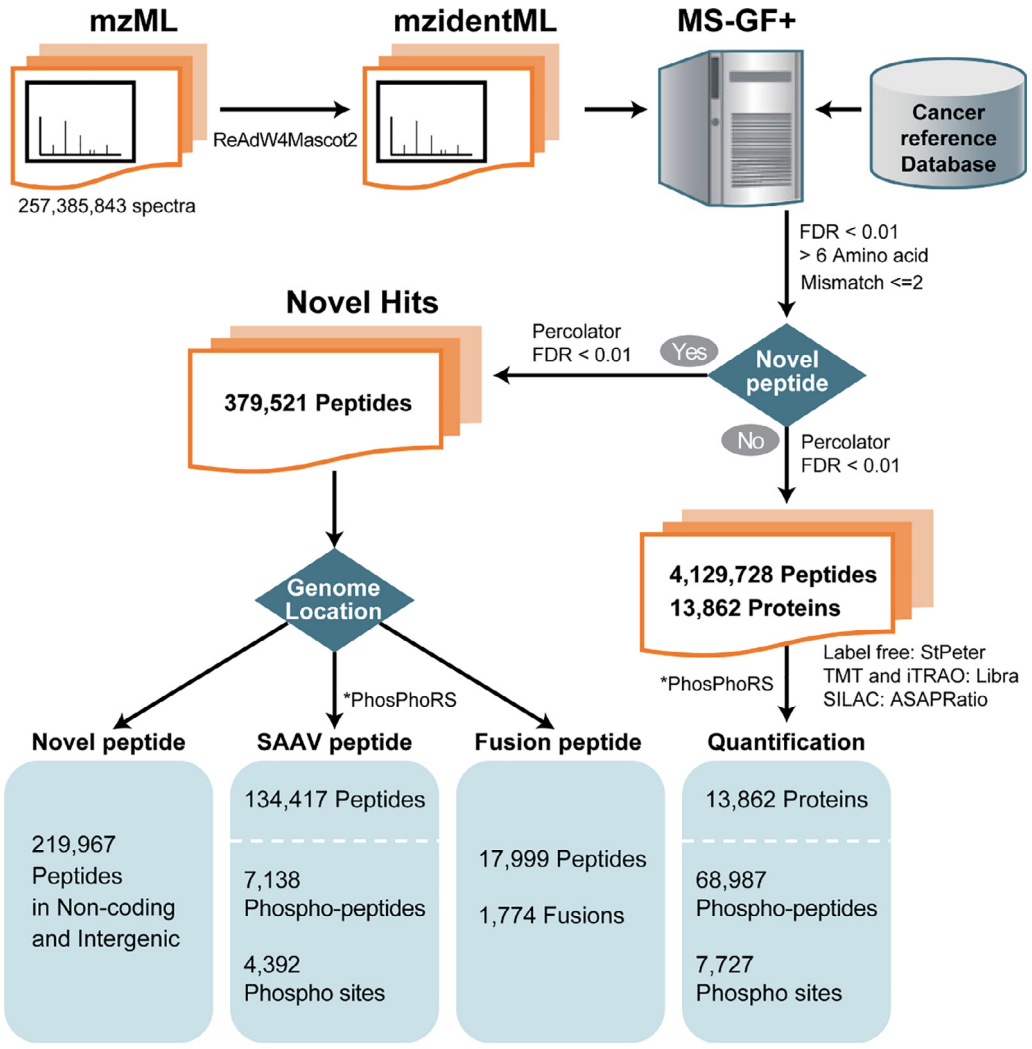


Fig. 1. Proteogenomic analysis pipeline in PEN. All spectra were converted and mapped to the cancer reference database by MS-GF + . Up to two mismatches were allowed and short peptides with fewer than six residues were removed. One percent peptide-level false discovery rate cutoff was used during the peptide identification. The protein-level expression was estimated using Trans-Proteomic Pipelines' quantification algorithms. Novel peptides were classified by single amino acid variation, noncoding region, and fusion peptide. *performed only using phosphoproteomics dataset.

(E-value < 0.05), which are non-redundant protein sequences from UniPort, Ensemble, Refseq, and Gencode.

In phosphoproteomics studies, the PEN pipeline consisted of two major steps (Fig. 1b): (i) The obtained fragmentation spectra were searched against the cancer reference database using MS-GF+ [18]. The phosphopeptide identification pipeline was the same as that of the proteomic analytic pipeline. (ii) Phosphorylation sites were estimated using PhosphoRS [27]. Phosphorylation patterns for 7,138 proteins from 57 cell lines were constructed.

3. Results

3.1. System overview

The PEN includes publicly available genomic, proteomic, and phosphoproteomic data, which have become the principal resources of information on protein diversity and expression. PEN contains a compilation of 115 studies on 145 cell lines in 12 cancer types.

3.2. User interface

The PEN website incorporates various user-friendly features (Fig. 2). Most menus are self-evident except for the PEN site, for which detailed instructions are available on the help page. Basic searches can be performed for cancer type, cell line, and gene names. The search window suggests plausible keywords and supports autocompletion.

The search output for the query consists of (i) basic information including GeneRIF information, (ii) relevant studies, (iii) cell lines in the selected cancer type with a link to PEN available, and (iv) comprehensive analysis for proteogenomics in tumor type, cell line, and gene names. A search using PEN can produce quantitative comparisons of proteogenomic analysis results, as explained in the following section.

3.3. Quantitative comparison of expression versus cancer driver genomic alterations

The CNA-driven genes responsible for tumorigenesis are *cis*-associated, whereas *trans*-associated genes are thought to further abet the development of cancer [10,28–29]. Correlation between CNA and protein expression is strong evidence of cancer function [28–29]. Herein, we compiled a variety of CNA-expression correlations and integrated them to help users identify reliable oncogenic functions. Further, the correlation coefficient was calculated using the significantly altered genes, which were identified with GISTIC2

Table 2
Number of cell lines, peptide and protein in each analysis step.

Analysis Pipeline	# of cell lines	# of peptide	# of protein
Platform			
- Proteomics	138	–	–
- Phosphoproteomics	58	–	–
Mapping MS-GF+			
(step i)	138	113,467,727	
- Proteomics	58	21,652,328	
- Phosphoproteomics			
Peptide identification and data interpretation with percolator			16,143
(step ii)			9,385
- Proteomics	138	4,502,111	
- Phosphoproteomics	58	76,125	
Quantification for protein expression			13,862
(step iii)			7,727
- Proteomics	124	4,129,728	
- Phosphoproteomics	57	68,987	
Single Amino Acid Variation			11,171
(step iv)			4,392
- Proteomics	137	134,417	
- Phosphoproteomics	57	7,138	
Mapping to non-coding (Proteomics) (step iv)	137	219,967	–
Mapping to fusion (Proteomics) (step iv)	95	17,999	937
Quantitative comparison of CNA	100	–	7,003
Quantitative comparison of expression	70	–	10,838
Total number	145	4,502,111	16,724

from matched cancer types in TCGA (FDR < 0.1), and mRNA/protein expression from the cell lines, at an adjusted *p*-value < 0.1. The 'p.adjust()' command in R was used to calculate adjusted p-values from a set of p-values using a number of adjustment procedures. The CNA profiles of the cell lines were obtained from nine cancer types, and 18,985 significantly altered genes were selected from the TCGA data portal [30]. The *cis*-associated genes were identified by correlation between CNA and the expression of significantly altered genes, and the *trans*-association genes were identified by correlations between CNA of significantly altered genes and other gene expression. We used the Spearman's rank correlation, which is robust to different normalization methods between genomics and proteomics. The 3,546 *cis*-associated genes were identified using positive correlation (R > 0.5, adjusted *p*-value < 0.05). A total of 17,234 *trans*-associated genes were identified, CNAs of genes and expression of other genes showed correlation (|R|>0.5, adjusted *p*-value < 0.05).

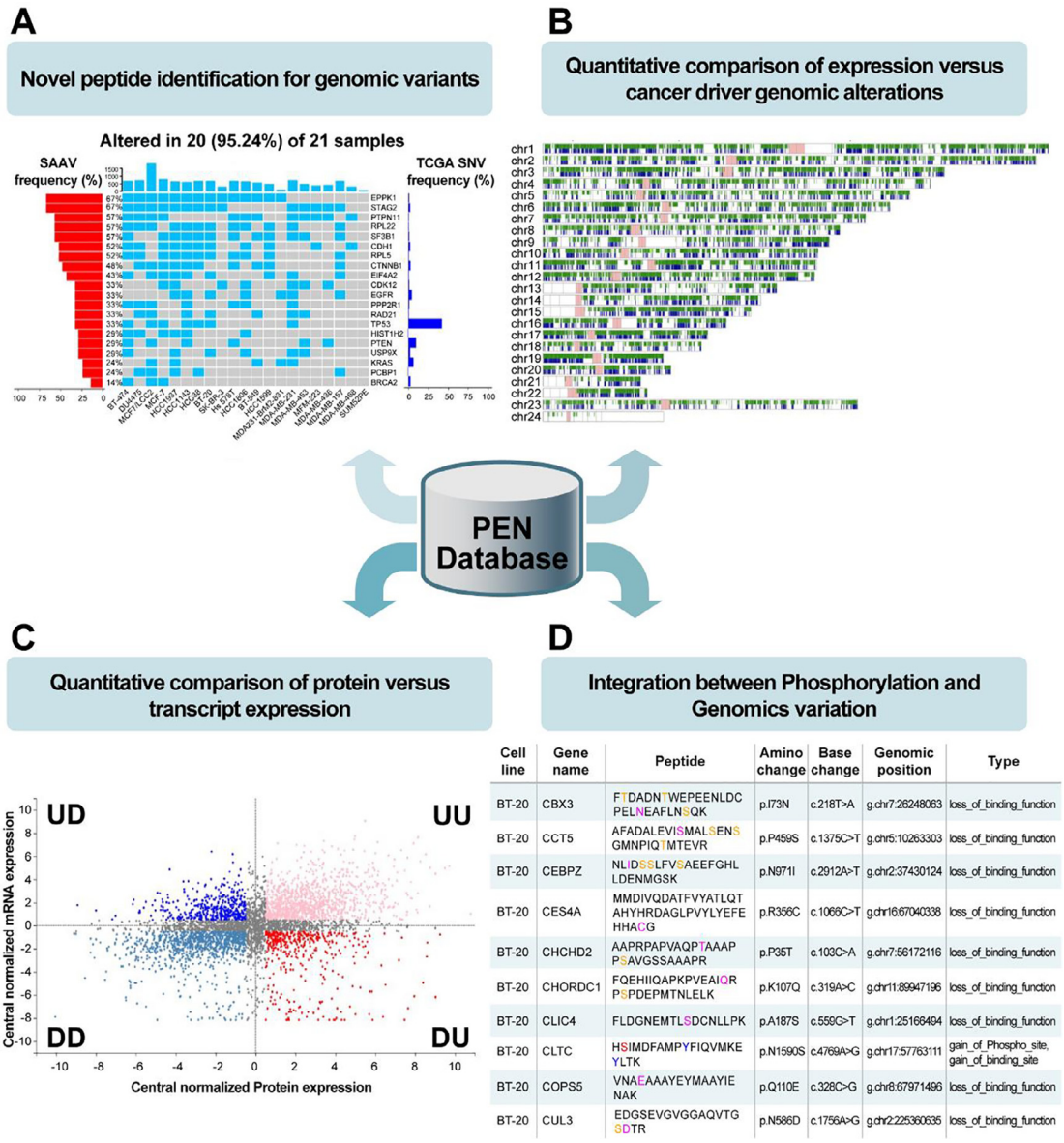
The cis/trans association is visually represented in a karyotype plot, as shown in Fig. 2. The karyotype map shows the global correlation for mRNA (colored green) and protein (colored blue) within each tissue. The scatter plot of CNA and expression can be displayed by clicking each cytoband to examine the cell line-dependent correlation of expression. The source of cis/*trans*-association is also indicated to help users identify consensus oncogene expression (or tumor-suppressor genes), which are more likely to be genuine events. For example, frequency of *CD274* (PD-L1) amplification in over 100,000 patients is rare in most solid tumors [31], but amplification and overexpression of *CD274* in lymphoma has been reported and lymphoma patients present CNA in *CD274* resulting in high response rates to PD-1/PD-L1 blockade [31]. We found *cis*-association of *CD274* in lymphoma cell lines, but not in the other cell lines. As another example of *cis*-association, a significant correlation was found between the mRNA and CNA of *CCND1*, in liver cancer. *CCND1* and its neighbor on 11q13.3 are amplified in human hepatocellular carcinoma and plays a role in tumor differentiation [32].

We compared expression with the coding region mutation on the mRNA and protein level. The association between genomic mutation and mRNA/protein expression of 100 cell lines was estimated, this matched the cell lines in accordance with their genomic and mRNA/protein expression (transcriptome and proteome). We estimated the occurrence of *cis*-events for 652 genes (FDR < 0.001). All the information on the association between mutation and expression can be found by searching in PEN. The estimations of association were performed using MatrixEQTL in R packages [33].

3.4. Novel peptide identification

Novel peptides can be of significant value for researchers studying the biological roles of oncogenes [34–35]. The SAAV, noncoding region, and fusion peptide were characterized using a novel peptide identification pipeline. SAAV peptides related to the translation of mutated genes can be readily identified with the PEN pipeline in multiple cell lines. We yielded 134,417 SAAV peptides for 137 cell lines and annotated them with clinically related mutations, such as ClinVar [36]. The SAAV-oncoplot was specifically designed to examine the SAAV distribution in tissue and SAAV sites in genes with oncoplot information in an intuitive and interactive user interface (Fig. 3A). The number of SAAV sites per cell line is displayed on the top panel, the frequency of SAAV sites in tissue in the left panel, and the mutation frequency in TCGA genomic studies in the right panel. The SAAVs in the main panel highlight the corresponding amino acid changes per cell line and are indicated in turquoise color. The user may search the SAAV profile for interesting genes and filter with clinically significant annotations using ClinVar. The detailed information for SAAV sites for each cell lines can be explored with a lollipop plot and table. The length of each SAAV site was also reflected in the frequency of each cell.

Novel peptides can be categorized into noncoding regions and fusion sequences. Novel peptides are encoded from a range of supposedly noncoding regions, including pseudogenes, 5′ or 3′ untranslated regions of mRNAs, lncRNAs, and intergenic and intronic sequences. We identified 219,967 novel peptides in noncoding regions (class-specific FDR 1%) in 137 tumor cell lines, of which 45,936 were supported by two or more peptides. All the information was summarized in a table including cancer and cell line



Cell line	Gene name	Peptide	Amino change	Base change	Genomic position	Type
BT-20	CBX3	FTDADNTWEPEENLDC PELNEAFLNSQK	p.I73N	c.218T>A	g.chr7:26248063	loss_of_binding_function
BT-20	CCT5	AFADALEVISMALSENS GMNPIQTMTEVR	p.P459S	c.1375C>T	g.chr5:10263303	loss_of_binding_function
BT-20	CEBPZ	NLIDSSLFVSAEEFGHL LDENMGSK	p.N971I	c.2912A>T	g.chr2:37430124	loss_of_binding_function
BT-20	CES4A	MMDIVQDATFVYATLQT AHYHRDAGLPVYLYEFE HHACG	p.R356C	c.1066C>T	g.chr16:67040338	loss_of_binding_function
BT-20	CHCHD2	AAPRPAPVAQPTAAAP PSAVGSSAAAPR	p.P35T	c.103C>A	g.chr7:56172116	loss_of_binding_function
BT-20	CHORDC1	FQEHIIQAPKPVEAIQR PSPDEPMTNLELK	p.K107Q	c.319A>C	g.chr11:89947196	loss_of_binding_function
BT-20	CLIC4	FLDGNEMTLSDCNLLPK	p.A187S	c.559G>T	g.chr1:25166494	loss_of_binding_function
BT-20	CLTC	HSIMDFAMPYFIQVMKE YLTK	p.N1590S	c.4769A>G	g.chr17:57763111	gain_of_Phospho_site, gain_of_binding_site
BT-20	COPS5	VNAEAAAYEYMAAYIE NAK	p.Q110E	c.328C>G	g.chr8:67971496	loss_of_binding_function
BT-20	CUL3	EDGSEVGVGGAQVTG SDTR	p.N586D	c.1756A>G	g.chr2:225360635	loss_of_binding_function

Fig. 2. Overview of cell lines comprehensive characterization and user interface of PEN. **A.** Integration between nucleotide alteration and novel peptide (top left, sky blue color represents SAAV; red, SAAV frequency of the selected cell lines; and blue, SAAV frequency in TCGA cohort for selected cell lines). **B.** Quantitative comparison between expression and genomic alteration (top right, green represents cis/trans expression between RNA and CNA; blue, cis/trans expression between protein and CNA; pink, the centromere region; and white, not-identified cis/trans expression). **C.** mRNA-protein abundance (bottom left, the x-axis represents mean centered protein expression and the y-axis mean centered mRNA expression; all expressions were normalized by mean centered normalization method [22]. **D.** Predicted phosphorylation-associated single nucleotide variations (bottom right, in the peptide column, blue are novel phosphorylation sites, yellow are known phosphorylation sites, pink are SAAV sites, and red are known phosphorylation sites contain SAAV). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the genomic location is visualized with link to UCSC genome browser (Fig. 3B).

The fusion peptides were found among 48,522 fusion sequences from the TCGA fusion database [37]. Oncogenic chimeric fusions have been exploited for their enormous potential as cancer biomarkers [38]. Potential fusion peptides that are specific to various human cell lines were identified, including 17,999 fusion peptides in 95 tumor cell lines. The fusion peptides were categorized into three different classes: Tier 1 was mapped to the fusion breakpoint, tier 2 to the 3′ fusion sequence, and tier 3 to the 5′ fusion sequence (Fig. 3C). The mapping status of the fusion peptide was visualized in a separate window. The fusion peptide for *CLTC/VMP1* was identified in bone, breast, and colon cancer cell lines in PEN. This gene has been implicated in gene fusion events in breast cancer [39]. *VMP1* is a recurrent fusion transcript involved in 30% of breast cancers [40]. The novel fusion peptide for *AXL/MBIP* was identified in blood, breast, and colon cell lines in PEN. While *AXL/MBIP* fusion has been reported in the large-scale sequencing of lung cancer samples [41], PEN contains only one lung cancer cell line (A-539), and this fusion peptide has not been identified in lung cancer cell lines. These results showed that PEN provides consistent results for cancer biology and has some limitations on protein coverage and data-specific prediction for novel peptides.

3.5. Integration between phosphorylation and genomics

The SAAVs of post-translational modifications (PTMs) can have a significant impact on the structure, function, and location of proteins in cells [42]. PEN provides a large-scale functional interpretation of PTM-associated genomic variations. The gain and loss of the

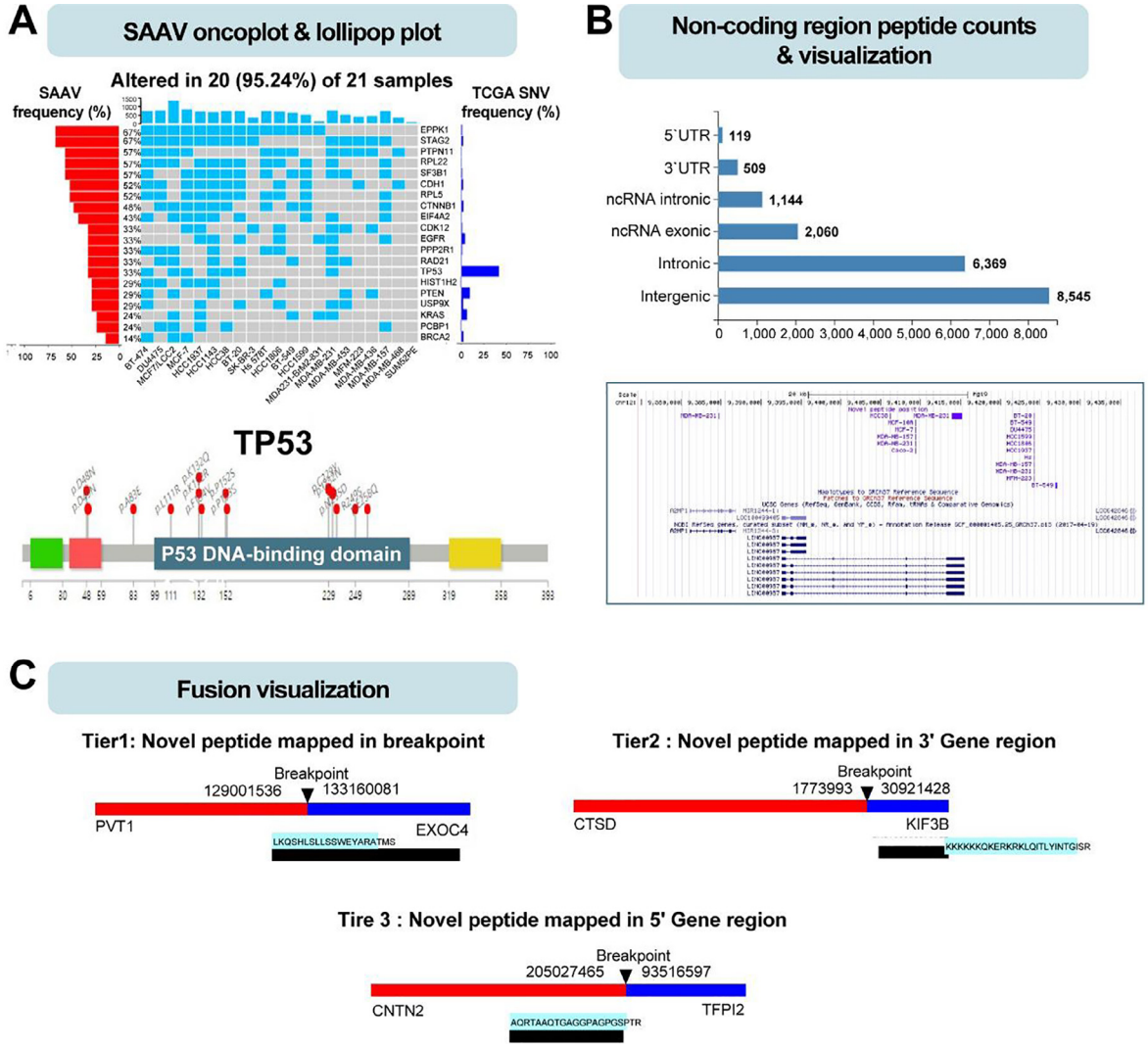


Fig. 3. Novel peptide identification in human cancer cell lines. **A.** Heatmap of Single Amino Acid Variation (SAAV) in the selected cancer types (upper panel). In the SAAV oncoplot, the columns show the cell lines and rows show genes. Detailed information of the mutation profile of a selected gene that will show in a lollipop plot and data table after clicking in the gene name (lower panel). **B.** Novel peptides for the selected cancer type or cell line (upper panel). Detailed list of peptide sequences and non-coding gene names provided for six categories. Users can compare peptide information with UCSC links. **C.** Three tiers of fusion peptides. The red and blue colors are 5′ and 3′ fusion partners, respectively. Black denotes the mapping peptide. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

phosphorylation site (STY) in a target protein may be important features for predicting cancer-causing mutations and may represent a molecular cause of tumors for a number of inherited and somatic mutations. We expect to obtain a comprehensive perspective of cellular processes and their interplay by integrating the information about phosphopeptides, SAAV, and known phosphorylation sites in proteins. We selected the proteomics and phosphor-enrichment proteomics data of one cell line and integrated the SAAV peptides (Fig. 4A). SAAV peptides containing gain PTM sites were identified in both platform (proteomics and phosphoproteomics). However, SAAV peptide containing known loss PTM site were identified in proteomic platform, but no SAAV phosphorpeptide was identified through phosphoproteomics. The known PTM sites were obtained from the UniProt PTM list. To identify SAAV peptides positions in a protein, all SAAV peptides were mapped to protein sequence using BLAST and allowing one mismatch and E-value < 0.05 (Fig. 4A). We identified 7,138 phosphopeptides containing SAAVs from 37 studies and 57 cell lines. Moreover, 5,462 PTM sites changed by somatic mutation; 4,987 PTM sites created by somatic mutation, and 475 PTM with loss of function were also identified. Phosphopeptides were categorized into four different classes (Fig. 4B): (i) Loss of binding function mutations (amino acids changed near known phosphorylation sites) were classified as known phosphorylation sites that were identified with SAAV peptides in proteomics, but without identified phosphopeptides in phosphoproteomics. (ii) Loss of phosphorylation site mutations (amino acid changed in a known phosphorylation site) were classified as those in which an SAAV peptide is identified and the known phosphorylation site changed with other amino acids. (iii) Gain of binding site mutations (novel phosphorylation site predicted near amino acid change) were classified as those in which SAAV phosphopeptides were identified and novel phosphorylation sites were predicted. (iv) Gain of phosphorylation site mutations (a novel phosphorylation site created with amino acid change) were classified as those where an SAAV phosphopeptide was identified and the mutation site was predicted to be a novel phosphorylation site.

Literature on the topic provides evidence supporting our prediction. In this sense, using PEN we observed an SAAV phosphopeptide of EGFR mutated at S1019L and P1170S, and predicted a novel phosphorylation site in that phosphopeptide. Accordingly, Lundby et al. proposed that S1019L and P1170S are related to the

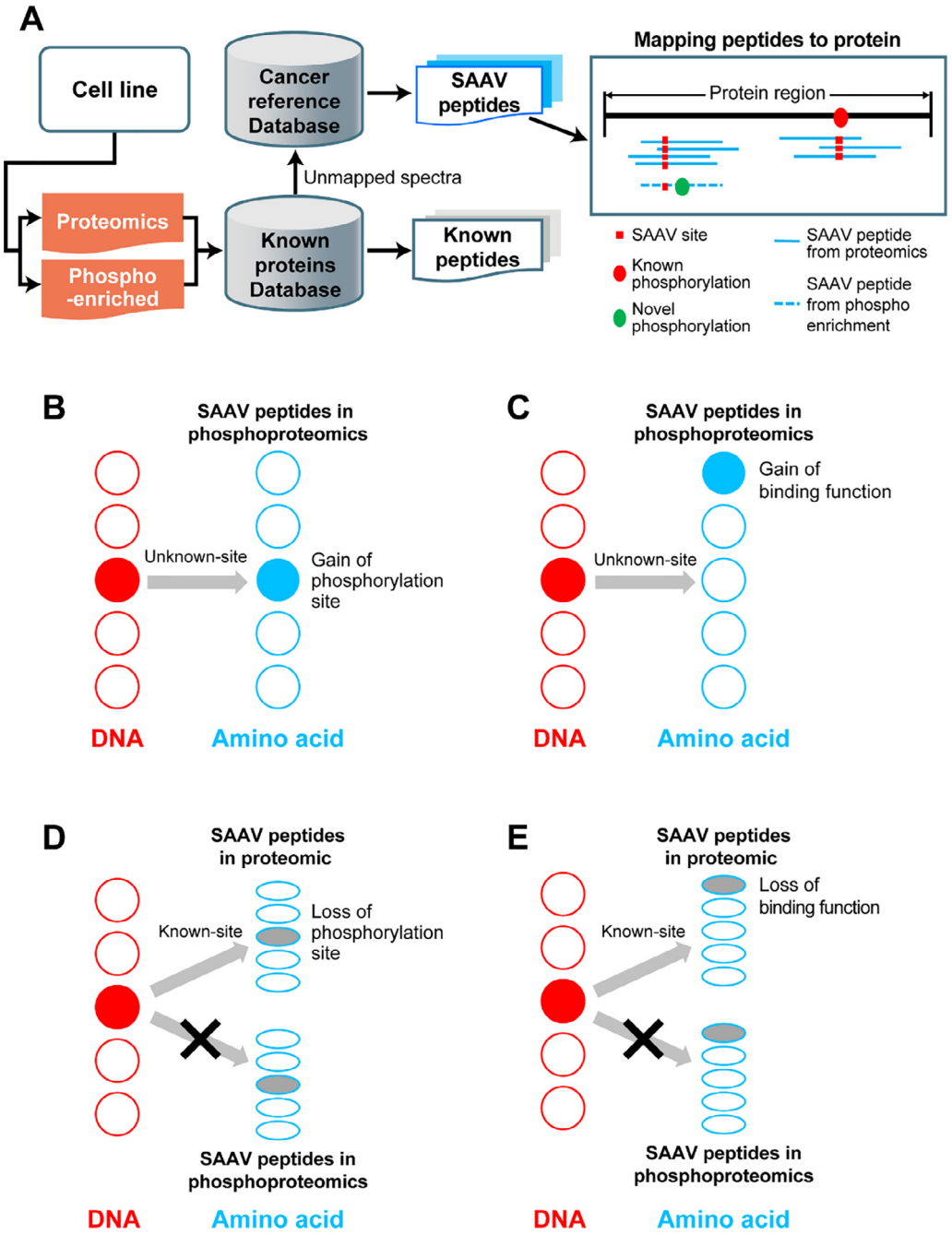


Fig. 4. Integration schema between genomic alterations and phosphopeptides. **A** Flower chart of predicted phosphorylation sites containing SAAV. The phosphoproteome and proteome data of a specific cell line was selected. All SAAV peptides were mapped to protein sequence using BLAST and allowing one mismatch. The novel phosphorylation sites were categorized into four groups, **B** gain of phosphorylation site, **C** gain of binding function, **D** loss of phosphorylation site, and **E** loss of binding function.

phosphorylation of *EGFR* in a different manner and oncogenic signaling is activated in cancer [43]. Furthermore, the N-terminal phosphorylation of *CTNNB1* at S33, S37, and S44 has been reported to affect phosphorylation of D32N [44]. PEN predicted SAAV phosphopeptides with the same effects as human cancer cell lines. PEN demonstrates that functional relationships are encrypted in patterns of co-regulated or anti-regulated phosphorylation site variations.

3.6. Quantitative comparison of protein versus transcript expression

Because of the extensive characterization of the human cancer cell lines; the 10,838 proteins quantified across the 70 cancer cell lines were quantitatively compared with mRNA expression patterns. From the expression pattern in each cell line, we grouped consistent and inconsistent expressions of mRNA and protein. The mRNA and protein expression data were normalized using the mean centered method. This is accomplished by calculating the mean of intensities of each sample, and then scaling the data so that the mean in all samples match. A center normalized protein with a value of '0', represents the mean protein expression of the sample. If the center normalized protein expression is '1' or '-1', that protein expression is 2 time higher or lower than the mean expression level. The first capital letter designates the expression of protein and the second capital letter the expression of mRNA (U: up-expression, D: down-expression, UU: up-expression of both

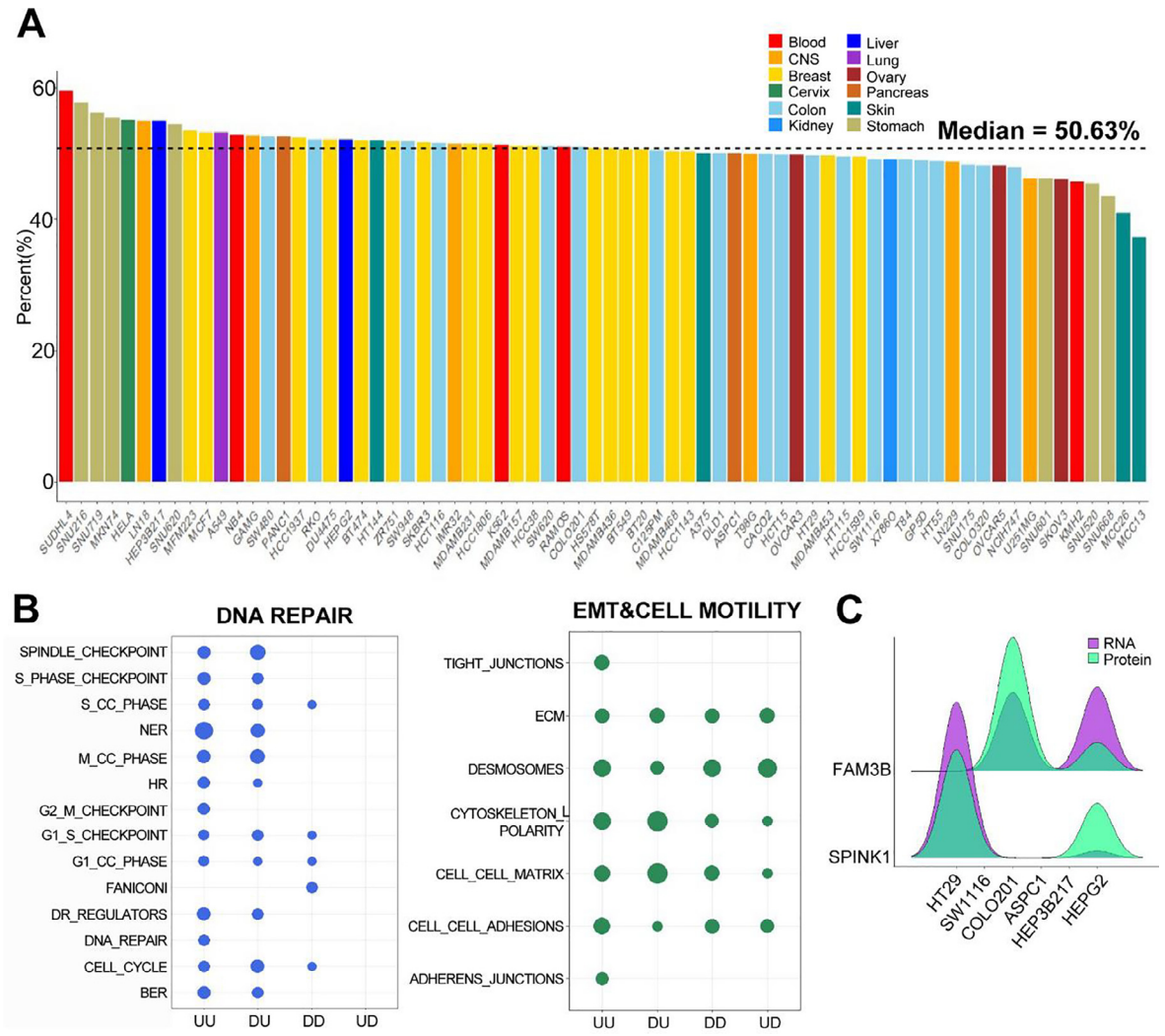


Fig. 5. Consistent and inconsistent protein expression of human cancer cell lines. **A.** Percentage of consistent proteins across 90 cell lines. The colors in the bar represent the cancer type. **B.** Bubble plot of the pathways enriched in consistent and inconsistent genes, at 0.05*p*-value. The bubble size represent the amount of cell lines. **C.** Cumulative expression pattern of two highly consistent genes. Colorectal (*HT29*, *SW1116*, and *COLO201)*, pancreatic (*ASPC1*), and liver (*HEP3B217* and *HEPG2*) cancer.

protein and mRNA, UD: up-expression in protein but down-expression in mRNA, DU: down-expression in protein but up-expression in mRNA, DD: down-expression of both protein and mRNA). The biological function of the proteins in each group was represented using cancer-related signaling in the Atlas of Cancer Signaling Networks (ACSN) [45]. PEN covers 1,719 proteins in the ACSN, a manually curated pathway database presenting published biochemical reactions involved in cancer [45]. Enrichment tests were performed using GSA, a gene-set-based quantification method. When mapping the P value for the enrichment test, PEN provided distinct cancer pathway patterns of particular cell lines with mRNA-protein abundance. Users can search for consistent and inconsistent genes in PEN and emphasize interesting pathway-associated genes in the mRNA-protein expression plot.

The proteins in the groups with consistent mRNA-protein expression (UU and DD) comprised approximately 50% (UU: 46% and DD: 4%) of all cell lines (Fig. 5A). The gene set analysis (GSA) approach revealed substantial diversity in pathway enrichment. Some tissue specificities were revealed, with particular cell line cancer-related signaling pathways showing distinct patterns. For the UU group, DNA repair was significantly enriched in 72% of the cell lines. However, the ENT-cell-motility-related pathway in the consistent and inconsistent groups was enriched across cell lines. (Fig. 5B) PEN provides comprehensive cancer-related signaling pathway patterns and provides cell line-specific functions in cancer.

Moreover, consistent or inconsistent genes show different mRNA-protein abundance patterns across cancer types. The highly consistent genes, including *SPINK1* and *FAM3B*, showed UU expression in colorectal cancer, but not in pancreatic and liver cancer (Fig. 5C). According to the literature and the Human Protein Atlas (HPA) search, these genes have important roles in colorectal cancer and validate intestinal tissue-specific expression in HPA [46–47]. These results indicate that PEN is a useful tool for preliminary cancer biology research.

3.7. Case study: Drug resistance and putative target estimation

PEN provides comprehensive information of expression, variation, and function to study drug response and the molecular features of cancer cell lines. For example, by using a SAAV pattern it would be possible to infer which cell lines are EGFR tyrosine kinase inhibitor (TKI) resistant and to estimate putative targets for the creation of a therapeutic strategy. EGFR is involved in cell growth and may also be found highly expressed in certain types of cancer [48]. In this sense, the search results obtained with PEN showed

that EGFR protein and mRNA level are highly expressed in breast adenocarcinoma. However, this elevated EGFR expression is not associated with CNA in breast adenocarcinoma. PEN encompass five cancer cell lines of breast adenocarcinoma, three of them (MDA-MB453, SK-BR-3, and SKBR3-AZDRc) contain a wild type EGFR, and two (MDA-MB-231 and MDA-MB-468) a mutant EGFR. The resistance of EGFR TKI can occurred through mutation of EGFR or KRAS [49]. Among these cell lines, only MDA-MB-231 has both mutations. Furthermore, we find that there was a functional association between EGFR TKI resistance and the protein-mRNA expression of MDA-MB-231. In addition, the PIK3-AKT-mTOR pathway, which is activated in many cancer types and in tumors resistant to agents targeting upstream TKI [50], is enriched in MDA-MB-231, but not in MDA-MB-468. Through this search, we can estimate resistant cell lines and identify driver molecular features for EGFR TKI. For example, to overcome EGFR TKI resistance, we performed a search, obtained a kinase enrichment result in breast adenocarcinoma, and found that PDGFRalpha/beta presents high activity. PDGFRalpha/beta is closely related to tumor development and a combination of EGFR and PDGFR inhibitors have been shown to overcome EDGR TKI resistance [48]. PEN proposed comprehensive information for cancer biology and tumor association molecular features.

4. Discussion and conclusions

An important application of proteogenomics is to provide comprehensive information for mRNA-protein expression and identify cancer-specific protein variations. The goal of PEN is to provide an extensive proteogenomics data set for cancer biology, functional understanding of RNA-protein abundance and to identify novel peptides containing genomic alterations.

The approximately two hundred million recorded MS/MS spectra yielded 4,502,111 matches to known and custom database peptide sequences in this two-stage strategy. We were able to identify 59% of the known proteins, while we found that 9% possessed cancer-specific variations. Thus, a proportion of these peptides may not be accessible using known proteins sequences. However, by using cancer-specific variation, we were able to achieve an exceptionally high individual protein coverage in cancer cells. Therefore, we presume that a large fraction of the novel peptides is highly specific to individual cell lines or it is only expressed under tumor conditions. In addition, we constructed the analysis pipeline based on the CPTAC common data analysis pipeline [4]. The CPTAC Common Data Analysis Platform (CDAP) analysis steps included peak-picking and quantitative data extraction, database searching, gene-based protein parsimony, false discovery rate based filtering, and detail options for tumor samples. To update the PEN database and integrate the proteogenomics data set between PEN and CPTAC, we used the same parameters and cutoff options than for CDAP. If the user want to change the parameters of the peptide identification step, they can download the data set of each step from the PEN web page and adjust the parameters. Foremost, accurate control of the FDR of novel identifications is crucial and has to be correctly accounted [23]. Consistent with previous studies [24–26], two-stage FDR control was more stringent than the separate FDR control. In order to strengthen quality control for novel peptides of fusion gene, we used BLAST to remove the peptides mapping known protein sequence.

In quantitative comparison of expression versus CNA, the number of *trans*-associated genes was higher than the number of *cis*-associated genes. This event was also observed in other proteogenomics study of cancer [51]. Furthermore, many oncogenic genes will be present in associated signaling pathway and several downstream genes will be effected by oncogenic gene expression. From these results, we can conclude that the high *trans*-associated genes number obtained is between the expected values.

Similar studies have proposed some of the greatest insights for cancer proteogenomics [7,10,16,1–5]. Nevertheless, a limitation of those studies is that the database were designed for open community resources or provided cross section for cancer biology. We still lack a comprehensive and integrated database for functional inference of mRNA-protein abundance, phosphorylation containing SAAV, and novel peptides for non-coding and fusion proteins. PEN can provide new insights into the proteogenomics landscape of cancer cells, allowing more detailed studies of oncogenesis and cancer treatment.

However, it is important to be aware that PEN have certain limitations including i) low protein coverage, the proteomics analysis does not achieve the required dynamic range of detection, highlighting a limitation of the current proteomics technologies for comprehensive proteome coverage and biomarker discovery [52]. We identified average 1,121 human proteins. ii) False positive in novel peptide identification. A main limitation of quantitation by proteomics is its low sensitivity [53]. We attempted to identify high quality novel peptide with an *in silico* method, but perfect elimination of false positive was unclear, however putative novel peptides for cancer was supported. Thus, PEN is limited to the validation of peptides with novel peptides (SAAV and fusion). We believe that it will become particularly useful in proteogenomics studies aiming to study cancer biology and to identify cancer-specific mutation at the protein level. iii) Data-specific interpretation. Because of its dynamic protein coverage, the data obtained may be regarded as 'missing' if it has not been obtained under the same identification conditions. Data-specific bias can be drawn from the missing data. Regular updates of proteogenomics data are critical to overcome this limitation, and are also particularly important as several omics studies are currently in progress, including the Depmap portal. We are planning to update the data annually and to expand the scope to pharmaco-proteogenomics, by including detailed drug response information available from Depmap.

In conclusion, our advanced PEN database provides a comprehensive view of the proteogenomics of cancer models and insights into the functional understanding of multi-omics data in human cancer models. Due to the construction of a proteogenomics data set and the implementation of several novel information, PEN can be described as an integrated resource of up-to-date information from genomic alterations to the effects on translation. These new data and functions would be valuable for understanding oncogenic and molecular functions in cancer cell lines.

5. Availability of data and material

PEN is freely accessible at http://combio.snu.ac.kr/pen.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant (NRF-2014M3C9A4064815 and NRF-2017R1D1A1B03031126) and the National Cancer Center Grant (2110030 and 2110510).

CRediT authorship contribution statement

Daejin Hyung: Software, Visualization, Methodology, Writing - original draft. **Min-Jeong Baek:** Data curation, Writing - original draft, Validation. **Jongkeun Lee:** Software, Visualization, Methodology. **Juyeon Cho:** Data curation, Validation. **Hyoun Sook Kim:** Data curation, Validation. **Charny Park:** Conceptualization, Investi-

gation, Writing - original draft, Writing - review & editing, Supervision. **Soo Young Cho:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by KREONET.

References

[1] Guo T, Luna A, Rajapakse VN, Koh CC, Wu Z, et al. Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. iScience 2019;21:664–80. https://doi.org/10.1016/j.isci.2019.10.059.
[2] Frejno M, Meng C, Ruprecht B, Oellerich T, Scheich S, et al. Proteome activity landscapes of tumor cell lines determine drug responses. Nat Commun 2020;11(1):3639. https://doi.org/10.1038/s41467-020-17336-9.
[3] Samaras P, Schmidt T, Frejno M, Gessulat S, Reinecke M, et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. Nucleic Acids Res 2020;248(D1):D1153–63. https://doi.org/10.1093/nar/gkz974.
[4] Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskoi DV, et al. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. J Proteome Res 2016;15(3):1023–32. https://doi.org/10.1021/acs.jproteome.5b0109110.1021/acs.jproteome.5b01091.s002.
[5] Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 2019;47(D1):D442–50. https://doi.org/10.1093/nar/gky1106.
[6] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2020;30(1):207–10. https://doi.org/10.1093/nar/30.1.207.
[7] Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic Acids Res 2020;48(D1):D1145–52. https://doi.org/10.1093/nar/gkz984.
[8] Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. A draft map of the human proteome. Nature 2014;509(7502):575–81. https://doi.org/10.1093/nar/gkz984.
[9] Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509(7502):582–7. https://doi.org/10.1038/nature13319.
[10] Stathias V, Turner J, Koleti A, Vidovic D, Cooper D, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res 2020;48(D1):D431–9. https://doi.org/10.1093/nar/gkz1023.
[11] Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534(7605):55–62. https://doi.org/10.1038/nature18003.
[12] Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell 2020;183(5):1436–56 e31. https://doi.org/10.1016/j.cell.2020.10.036.
[13] Li Y, Wang X, Cho J-H, Shaw TI, Wu Z, Bai B, et al. JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. J Proteome Res 2016;15(7):2309–20. https://doi.org/10.1021/acs.jproteome.6b0034410.1021/acs.jproteome.6b00344.s00110.1021/acs.jproteome.6b00344.s002.
[14] Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun 2018;9(1):903. https://doi.org/10.1038/s41467-018-03311-y.
[15] Li Y, Wang G, Tan X, Ouyang J, Zhang M, Song X, et al. ProGeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection. BMC Med Genomics 2020;13(S5). https://doi.org/10.1186/s12920-020-0683-4.
[16] Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 2019;569(7757):503–8. https://doi.org/10.1038/s41586-019-1186-3.
[17] Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 2013;29(24):3235–7. https://doi.org/10.1093/bioinformatics/btt543.
[18] Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 2014;5:5277. https://doi.org/10.1038/ncomms6277.
[19] Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007;4(11):923–5. https://doi.org/10.1038/nmeth1113.
[20] Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics 2010;10(6):1150–9. https://doi.org/10.1002/pmic.200900375.
[21] Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER, Kalocsay M, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell 2020;180(2):387–402.e16. https://doi.org/10.1016/j.cell.2019.12.023.
[22] Nusinow DP, Gygi SP. A Guide to the Quantitative Proteomic Profiles of the Cancer Cell Line Encyclopedia. bioRxiv https://doi.org/10.1101/2020.02.03.932384.
[23] Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun 2020;11(1):1759. https://doi.org/10.1038/s41467-020-15456-w.
[24] Woo S, Cha SW, Na S, Guest C, Liu T, et al. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics 2014;14(23–24):2719–30. https://doi.org/10.1002/pmic.201400206.
[25] Li H, Park J, Kim H, Hwang K-B, Paek E. Systematic Comparison of False-Discovery-Rate-Controlling Strategies for Proteogenomic Search Using Spike-in Experiments. J Proteome Res 2017;16(6):2231–9. https://doi.org/10.1021/acs.jproteome.7b0003310.1021/acs.jproteome.7b00033.s001.
[26] Ivanov MV, Lobas AA, Karpov DS, Moshkovskii SA, Gorshkov MV. Comparison of False Discovery Rate Control Strategies for Variant Peptide Identifications in Shotgun Proteogenomics. J Proteome Res 2017;16(5):1936–43. https://doi.org/10.1021/acs.jproteome.6b01014.
[27] Taus T, Kocher T, Pichler P, Paschke C, Schmidt A, et al. Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res 2011;10(12):5354–62. https://doi.org/10.1021/pr200611n.
[28] Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. Brief Funct Genomics 2015;14(5):352–7. https://doi.org/10.1093/bfgp/elv017.
[29] Shao X, Lv N, Liao J, Long J, Xue R, Ai Ni, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet 2019;20(1). https://doi.org/10.1186/s12881-019-0909-5.
[30] Gao GF, Parker JS, Reynolds SM, Silva TC, Wang L-B, Zhou W, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. Cell Syst. 2019;9(1):24–34.e10. https://doi.org/10.1016/j.cels.2019.06.006.
[31] Goodman AM, Piccioni D, Kato S, Boichard A, Wang H-Y, Frampton G, et al. Prevalence of PDL1 Amplification and Preliminary Response to Immune Checkpoint Blockade in Solid Tumors. JAMA Oncol. 2018;4(9):1237. https://doi.org/10.1001/jamaoncol.2018.1701.
[32] Joo M, Kang YK, Kim MR, Lee HK, Jang JJ. Cyclin D1 overexpression in hepatocellular carcinoma. Liver 2001;21(2):89–95. https://doi.org/10.1034/j.1600-0676.2001.021002089.x.
[33] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 2012;28(10):1353–8. https://doi.org/10.1093/bioinformatics/bts163.
[34] Othoum G, Coonrod E, Zhao S, Dang HX, Maher CA. Pan-cancer proteogenomic analysis reveals long and circular noncoding RNAs encoding peptides. NAR. Cancer 2020;2(3). https://doi.org/10.1093/narcan/zcaa015.
[35] Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure J-P, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med 2018;10(470):eaau5516. https://doi.org/10.1126/scitranslmed.aau5516.
[36] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42(D1):D980–5. https://doi.org/10.1093/nar/gkt1113.
[37] Hu X, Wang Q, Tang M, Barthel F, Amin S, et al. TumorFusions: an integrative resource for cancer-associated transcript fusions. Nucleic Acids Res. 2018;46(D1):D1144–9. https://doi.org/10.1093/nar/gkx1018.
[38] Conlon KP, Basrur V, Rolland D, Wolfe T, Nesvizhskii AI, MacCoss MJ, et al. Fusion peptides from oncogenic chimeric proteins as putative specific biomarkers of cancer. Mol Cell Proteomics 2013;12(10):2714–23. https://doi.org/10.1074/mcp.M113.029926.
[39] Giacomini CP, Sun S, Varma S, Shain AH, Giacomini MM, Balagtas J, et al. Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. PLoS Genet 2013;9(4):e1003464. https://doi.org/10.1371/journal.pgen.1003464.
[40] Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, Vardy LA, et al. Transcriptional consequences of genomic structural aberrations in breast cancer. Genome Res 2011;21(5):676–87. https://doi.org/10.1101/gr.113225.110.
[41] Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res 2012;22(11):2109–19. https://doi.org/10.1101/gr.145144.112.
[42] Cesnik AJ, Shortreed MR, Sheynkman GM, Frey BL, Smith LM. Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. J Proteome Res 2016;15(3):800–8. https://doi.org/10.1021/acs.jproteome.5b0081710.1021/acs.jproteome.5b00817.s00110.1021/acs.jproteome.5b00817.s002.
[43] Lundby A, Franciosa G, Emdal KB, Refsgaard JC, Gnosa SP, Bekker-Jensen DB, et al. Oncogenic Mutations Rewire Signaling Pathways by Switching Protein

Recruitment to Phosphotyrosine Sites. Cell 2019;179(2):543–560.e26. https://doi.org/10.1016/j.cell.2019.09.008.
[44] Al-Fageeh M, Li Q, Dashwood WM, Myzak MC, Dashwood RH. Phosphorylation and ubiquitination of oncogenic mutants of beta-catenin containing substitutions at Asp32. Oncogene 2004;23(28):4839–46. https://doi.org/10.1038/sj.onc.1207634.
[45] Kuperstein I, Bonnet E, Nguyen HA, Cohen D, Viara E, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis 2015;4:e160. https://doi.org/10.1038/oncsis.2015.19.
[46] Tiwari R, Pandey SK, Goel S, Bhatia V, Shukla S, et al. SPINK1 promotes colorectal cancer progression by downregulating Metallothioneins expression. Oncogenesis 2015;4:e162. https://doi.org/10.1038/oncsis.2015.23.
[47] Li Z, Mou H, Wang T, Xue J, Deng Bo, Qian L, et al. A non-secretory form of FAM3B promotes invasion and metastasis of human colon cancer cells by upregulating Slug expression. Cancer Lett 2013;328(2):278–84. https://doi.org/10.1016/j.canlet.2012.09.026.
[48] Huang L, Fu L. Mechanisms of resistance to EGFR tyrosine kinase inhibitors. Acta Pharm Sin B 2015;5(5):390–401. https://doi.org/10.1016/j.apsb.2015.07.001.
[49] Guerrab AE, Bamdad M, Kwiatkowski F, Bignon Y-J, Penault-Llorca F, Aubel C. Anti-EGFR monoclonal antibodies and EGFR tyrosine kinase inhibitors as combination therapy for triple-negative breast cancer. Oncotarget. 2016;7(45):73618–37. https://doi.org/10.18632/oncotarget.v7i4510.18632/oncotarget.12037.
[50] Papadimitrakopoulou V. Development of PI3K/AKT/mTOR pathway inhibitors and their application in personalized therapy for non-small-cell lung cancer. J Thorac Oncol. 2012;7(8):1315–26. https://doi.org/10.1097/JTO.0b013e31825493eb.
[51] Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, et al. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell 2020;182(1):200–25 e35. https://doi.org/10.1016/j.cell.2020.06.013.
[52] Angel TE, Aryal UK, Hengel SM, Baker ES, Kelly RT, et al. Mass spectrometry-based proteomics: existing capabilities and future directions. Chem Soc Rev 2012;41(10):3912–28. https://doi.org/10.1039/c2cs15331a.
[53] Mulvey C, Thur B, Crawford M, Godovac-Zimmermann J. How Many proteins are Missed in Quantitative proteomics Based on Ms/Ms sequencing Methods? Proteomics Insights 2010;3:61–6. https://doi.org/10.4137/PRI.S5882.