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, b, *, 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

A R T I C L E   I N F O

Article history:
Received 28 February 2021
Accepted 14 August 2021
Available online 17 August 2021

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

A B S T R A C T

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.

© 2021 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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 emerged as a promising approach to tumor profiling with the potential to advance basic, translational, and clinical research. To accelerate the understand-
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 CRCC65 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; 6,463,573 spectra). Tissue information of the selected cancer 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 TMT data, 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 N [6] 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

| 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                          |

D. Hyung, Min-Jeong Baek, J. Lee et al. Computational and Structural Biotechnology Journal 19 (2021) 4759–4769
(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.
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. adj[ust]’ 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 onecogenes [34–35]. The SAAV, non-coding 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 line 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

---

**Table 2**

| Analysis Pipeline | # of cell lines | # of peptide | # of protein |
|-------------------|----------------|--------------|--------------|
| Platform          |                |              |              |
| - Proteomics      | 138            | –            | –            |
| - Phosphoproteomics | 58           | –            | –            |
| Mapping MS-GF*    | 138            | 113,467,727  |              |
| (step i)          | 58             | 21,652,328   |              |
| - Proteomics      | 138            | 4,502,111    |              |
| - Phosphoproteomics | 58         | 76,125       |              |
| Peptide identification and data interpretation with percolator (step ii) | 16,143 | 9,385 |
| Quantification for protein expression (step iii) | 13,862 | 7,727 |
| - Proteomics      | 124            | 4,129,728    |              |
| - Phosphoproteomics | 57         | 68,987       |              |
| Single Amino Acid Variation (step iv) | 11,171 | 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 |
| Quantitative comparison of CNA | 100 | 7,003 |
| Quantitative comparison of expression | 70 | 10,838 |
| Total number | 145 | 4,502,111 | 16,724 |
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
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 phospho-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 phosphopeptide 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...
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. 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.
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 encompasses 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 PI3K-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, Koo 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, Oelender 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/s41467-020-17364-9.
[3] Samaras P, Schmidt T, Frejno M, Gesualit S, Reinecke M, et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. Nucleic Acids Res 2020;48(D1):D1153–63. https://doi.org/10.1093/nar/gkaa974.
[4] Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskov D, 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.5b00110.10.1021/acs.jproteome.5b00110.
[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 next generation access point for perturbation-response signatures. Nucleic Acids Res 2019;47(D1):D442–50. https://doi.org/10.1093/nar/gky1023.
[6] Kim MS, Pinto SM, Getnet D, Nirujsog P, Manda SS, et al. A draft map of the human proteome. Nature 2014;509(7502):575–81. https://doi.org/10.1038/nature13207.
[7] Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509(7502):575–81. https://doi.org/10.1038/nature13207.
[8] Kim MS, Pinto SM, Getnet D, Nirujsog P, Manda SS, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons’ Data. Cell Syst. 2019;5(1):24–34.e10. https://doi.org/10.1016/j.cells.2019.06.006.
[9] 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(8):1237. https://doi.org/10.1001/jamaoncol.2018.2012.
[10] Rao MS, Langlois E, Dicks E, et al. CLUE: a resource for cancer-associated transcript fusions. Nucleic Acids Res 2018;46(1):D81–7. https://doi.org/10.1093/nar/gkx1084.
[11] Gao GF, Parker JS, Reynolds SM, Silva TC, Wang L-B, Zhou W, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genomics 2020;13(S5). https://doi.org/10.1186/s12920-020-0683-4.
[12] Kim D, Teng H, Wang X, Zhang B. Cancer neoantigen prioritization through search tool for proteomics. Nat Commun 2014;5:5277. https://doi.org/10.1038/ncomms6227.
Recruitment to Phosphotyrosine Sites. Cell 2019;179(2):543–560.e26. https://doi.org/10.1016/j.cell.2019.09.008.