ScienceAdvances
AAAS

Manuscript Template

Title

- **Full title**: Deep top-down proteomics revealed significant proteoform-level differences between metastatic and non-metastatic colorectal cancer cells
- **Short title**: Top-down proteomics of colorectal cancer cells

Authors

Elijah N. McCool,[1,‡] Tian Xu,[1, ‡] Wenrong Chen,[2, ‡] Nicole C. Beller,[3] Scott M. Nolan,[1] Amanda B. Hummon,[3,4,*] Xiaowen Liu,[5,*] Liangliang Sun[1,*]

Affiliations

[1]Department of Chemistry, Michigan State University, 578 S Shaw Lane, East Lansing, Michigan 48824, United States

[2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Avenue, Indianapolis, Indiana 46202, United States

[3]Department of Chemistry and Biochemistry, The Ohio State University, 100 West 18th Avenue, Columbus, Ohio, United States 43210

[4]The Comprehensive Cancer Center, The Ohio State University, 500 West 12th Avenue, Columbus, Ohio 43210, United States

[5]Deming Department of Medicine, School of Medicine, Tulane University, 1441 Canal Street, New Orleans, LA, United States 70112

[‡] Those authors contributed equally to this work.

* Corresponding authors.

Amanda B. Hummon

Email: hummon.1@osu.edu

Phone: 1-614-688-2580

Xiaowen Liu

xwliu@tulane.edu

Phone: 1-504-988-9136

Liangliang Sun

Email: lsun@chemistry.msu.edu
Phone: 1-517-353-0498

Abstract

Understanding cancer metastasis at the proteoform level is crucial for discovering new protein biomarkers for cancer diagnosis and drug development. We present the first top-down proteomics (TDP) study of a pair of isogenic human non-metastatic and metastatic colorectal cancer (CRC) cell lines (SW480 and SW620). We identified 23,622 proteoforms of 2,332 proteins from the two cell lines, representing nearly 5-folds improvement in the number of proteoform identifications (IDs) compared to previous TDP datasets of human cancer cells. We revealed significant differences between the SW480 and SW620 cell lines regarding proteoform and single amino acid variants (SAAVs) profiles. Quantitative TDP unveiled differentially expressed proteoforms between the two cell lines and the corresponding genes had diversified functions and were closely related to cancer. Our study represents a pivotal advance in TDP towards the characterization of human proteome in a proteoform-specific manner, which will transform basic and translational biomedical research.

Teaser

Top-down proteomics of colorectal cancer cells provides proteoform-level knowledge about cancer metastasis.

MAIN TEXT

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and has a high mortality rate even with recent improvements in therapies.[1,2] CRC metastasis is the main cause of CRC-related death. New insights into the molecular mechanisms of CRC metastasis will undoubtedly be beneficial for developing more effective drugs.[3-5] Extensive studies have been completed with the goal of understanding CRC metastasis at the transcriptome level, generating tremendous information about the landscape of mRNA across different stages of CRC. [6,7] However, nucleic-acid–based measurements do not correlate well with protein abundance, which are the primary effectors of function in biology.[8] Quantitative bottom-up proteomics (BUP) studies of metastatic and non-metastatic CRC cell lines have discovered new protein regulators involved in CRC metastasis.[4,9,10] BUP usually provides limited information on the proteoforms, which represent all possible protein molecules derived from the same gene resulting from genetic variations, RNA alternative splicing, and protein post-translational modifications (PTMs).[11,12] Mass spectrometry (MS)-based top-down proteomics (TDP) directly measures intact proteoforms and provides opportunities to study functions of specific proteoforms.[13,14] Unfortunately, there is still no report in the literature about studying CRC metastasis using TDP, and this study will help to fill that gap.

Here, we performed the first deep TDP study of metastatic (SW620) and non-metastatic (SW480) human CRC cell lines, aiming to produce a comprehensive proteoform-level view of the two isogenic CRC cell lines and discover novel proteoform biomarkers of CRC metastasis. We employed four different capillary zone electrophoresis (CZE)-tandem MS (MS/MS) approaches, 1-D CZE-MS/MS, 2-D size exclusion chromatography (SEC)-CZE-MS/MS, 2-D reversed-phase liquid chromatography (RPLC)-CZE-MS/MS, and 3-D SEC-RPLC-CZE-MS/MS analyses of the two cell lines for proteoform identification (ID) and label-free quantification (LFQ), **Figure 1**. For 1-D CZE-MS/MS, each sample was analyzed by CZE-MS/MS in technical triplicate. For 2-D SEC-CZE-MS/MS, each sample was fractionated by SEC into 6 fractions, followed by CZE-MS/MS in technical triplicate. For 2-D RPLC-CZE-MS/MS, we fractionated each sample to 6 or 13 fractions by RPLC and analyzed each LC fraction by single-shot CZE-MS/MS (RPLC 13 fractions) or triplicate CZE-MS/MS measurements (RPLC 6 fractions). For 3-D SEC-RPLC-CZE-MS/MS, 52 LC fractions were collected for each sample, followed by CZE-MS/MS in technical triplicate. From 1-D separation to 3-D separations, the required amount of starting protein materials increased (from 100 μg to 2 mg) due to the unavoidable sample loss during sample collections and transfers. The TopPIC (version 1.4.0) software was used for data analysis,[15] and a 1% proteoform-level false discovery rate (FDR) was used to filter the database search results.

Results

Identification of over 23,000 proteoforms from CRC cells using CZE-MS/MS

One long-term goal of TDP is to characterize all the millions of proteoforms in the human body.[16,17] During the last decade, because of the improvement of proteoform sample preparation, LC and CZE separations, MS and MS/MS, 3,000-5,000 proteoforms corresponding to roughly 1,000 genes can be identified from one human cell line using LC-MS/MS-based platforms,[18-22] and up to 6,000 proteoform IDs corresponding to 850 genes have been reported from an *E. coli* sample using a CZE-MS/MS-based workflow.[23] Only one TDP study of a human cell line using CZE-MS/MS was reported with the identification of about 500 proteoforms.[24] Recently, the Kelleher group reported the identification of ~30,000 proteoforms of 1,690 human genes from 21 human cell types and plasma using RPLC-MS/MS-based strategies, representing a milestone in large-scale TDP.[21] On average, nearly 3,000 proteoforms were identified from one of the 21 human cell types.

In this work, we performed the first global TDP study of a pair of isogenic human non-metastatic and metastatic CRC cell lines (SW480 and SW620). Four different strategies were employed, **Figure 1**. We first compared the four different CZE-MS/MS strategies listed in **Figure 1B** in terms of the number and efficiency of proteoform IDs from the SW480 cells, **Figure 2A**. SEC-RPLC-CZE-MS/MS outperformed SEC-CZE-MS/MS, RPLC-CZE-MS/MS, and CZE-MS/MS in terms of the number of proteoform IDs due to better LC fractionation (2-D LC vs. 1-D or no LC) and much more CZE-MS/MS runs (52 *vs.* 6 and 13). In terms of the proteoform identification efficiency (the number of proteoform IDs per CZE-MS/MS run), the SEC-CZE-MS/MS (6 LC fractions) produced nearly 700 proteoform IDs per run, which is nearly 6-fold and 4-fold higher than those from SEC-RPLC-CZE-MS/MS and CZE-MS/MS, respectively. We drew two conclusions from the data. First, multi-dimensional separation is crucial for large-scale TDP analysis of human cell lysates due to their extremely high complexity. Second, SEC-CZE-MS/MS and RPLC-CZE-MS/MS under an optimized condition are powerful techniques for deep TDP of human cell lysates with high throughput.

In total, we collected over 400 MS raw files using the four CZE-MS/MS-based strategies and identified 23,622 proteoforms of 2,332 proteins from the SW480 and SW620 cell lines with a 1% proteoform-level FDR. The number of proteoform IDs from the CRC cells is about 5-8 fold higher than that reported in previous TDP studies of human cancer cells (23,622 vs. 3,000-5,000 proteoforms).[18-20] 17,316 and 14,504 proteoforms (on average 15,910 proteoforms) were identified from SW480 and SW620 cell lines, respectively, representing about 3-fold improvement in the number of proteoform IDs per human cell line compared to previous LC-MS/MS-based TDP datasets. The number of proteoform IDs is about 30-fold higher than previous human cell TDP datasets by CZE-MS/MS (~16,000 vs. ~500).[24] **Figure 2B** shows the number of proteoform IDs per complex sample using TDP in previous works and this study.[18-23] **Table S1** summarizes the details of those studies.

We need to point out that the nearly 16,000 proteoform IDs from SW480 or SW620 cells combine the results of four different CZE-MS/MS-based strategies and about 200 CZE-MS/MS runs. The previous literature studies typically employ one LC-MS/MS or CZE-MS/MS-based approach.[18-23] We also included the data of SW480 and SW620 cells from only SEC-CZE-MS/MS in **Figure 2B**. A total of 5,855 and 6,273 proteoforms (mean±standard deviation: 6,064±296) were identified from SW480 and SW620 cells, respectively, by SEC-CZE-MS/MS, via 18 CZE-MS/MS runs (6 SEC fractions × 3 CZE-MS/MS runs/fraction). The SEC-CZE-MS/MS produced significantly higher proteoform IDs (6,000 vs. 3,000-5,000) from a single human cell line than LC-MS/MS-based approaches in the literature with a drastically lower number of MS runs (18 vs. 40-800).

The data clearly demonstrate the power of our CZE-MS/MS-based TDP strategy for comprehensive characterization of proteoforms in complex proteome samples. We attribute the drastic improvement of proteoform IDs to the high separation efficiency of CZE for proteoforms, [25] high sensitivity of CZE-MS for proteoform detection, [25-27] and high orthogonality of LC and CZE for biomolecule separations. [23,28] The features of CZE-MS/MS for TDP have been systematically reviewed recently. [29,30] The list of identified proteoforms is shown in **Supplementary Material II**.

We further compared the proteoforms and proteins identified from the SW480 and SW620 cells using the SEC-CZE-MS/MS data. **Figure 2C** shows the heat map of proteoform overlaps among technical replicates of SW480 and SW620 cells. About 60-70% of proteoforms identified in one technical replicate of SW480 or SW620 cells were also identified in another replicate of the same cell line, indicating reasonable reproducibility of proteoform ID using SEC-CZE-MS/MS and the data-dependent acquisition mode. **Figure S1** shows base peak electropherograms of triplicate CZE-MS/MS measurements of the SW620 cell lysate (one SEC fraction), indicating good reproducibility of CZE-MS/MS for complex proteome samples regarding separation profile and base peak intensity. Interestingly, only about 40-50% of proteoforms identified in one replicate of SW480 cells (e.g., SW480_1) were identified in one replicate of SW620 cells (e.g., SW620_1). The proteoform overlaps in **Figure 2C** between the two cell lines are statistically significantly lower than that within each cell line (44±4% vs. 67±4%, $p<10^{-14}$, two-tailed student's t-test). The data clearly demonstrate that the pair of isogenic human non-metastatic (SW480) and metastatic (SW620) CRC cell lines have significantly different proteoform profiles. The two cell lines are also significantly different at the protein level, as demonstrated by the protein overlaps shown in **Figure S2.** The difference in protein overlaps between the two cell lines and within each cell line is statistically significant (69±8% vs. 83±3%, $p<10^{-6}$, two-tailed student's t-test).

TDP has some technical challenges for the identification of large proteoforms (i.e., >30 kDa). In this work, we focused on the characterization of proteoforms smaller than 30 kDa using a Thermo Q-Exactive HF mass spectrometer. **Figure S3** shows the mass distribution of identified proteoforms from SW480 and SW620 cells. The majority of identified proteoforms are 10 kDa or smaller, which is one main limitation of this study. It is worth noting that 1600-2200 proteoforms have masses larger than 10 kDa. **Figure 2D** shows the sequences and fragmentation patterns of two example proteoforms. Those two proteoforms were identified with high confidence and were also well characterized with N-terminal methionine removal and N-terminal acetylation.

Proteoforms of important genes in well-known CRC-related pathways

We further performed QIAGEN Ingenuity Pathway Analysis (IPA) analysis of the genes identified in this work by the four CZE-MS/MS-based strategies and determined several significantly enriched and well-known CRC-related pathways, including WNT/β-catenin Signaling (p-value: 10^{-3}), PI3K/AKT Signaling (p-value: 10^{-4}), mTOR Signaling (p-value: 10^{-14}), and ERK/MAPK Signaling pathways (p-value: 10^{-4}). [31,32] Those pathways play critical roles in CRC progression via regulating cell proliferation, apoptosis, survival and etc. We identified hundreds of proteoforms from dozens of genes for each pathway, **Figure 3A**. The lists of proteoforms are shown in **Supplementary Material II**. Comparable numbers of proteoforms were identified from SW480 and SW620 cells for PI3K/AKT Signaling, mTOR Signaling, and ERK/MAPK Signaling pathways. An obviously higher number of proteoforms was obtained from SW480 cells compared to SW620 cells for the WNT/β-catenin Signaling pathway (511 *vs.* 340). Combination of the data from SW480 and SW620 cells produced about 40% more proteoforms related to the four CRC pathways compared to one cell line alone, indicating the potential differences in proteoform profiles for the well-known CRC-related pathways between the non-metastatic and metastatic CRC cell lines. As shown in **Figure 3B**, the shared proteoforms between SW480 and SW620 cells for each pathway is only about 21%-38% of the total proteoforms identified from the two cell lines. The data suggest that proteoforms in those pathways could potentially play important roles in driving CRC progression and metastasis.

We highlighted some proteoforms of important genes (MARK2, SOX9, EIF4B, and EIF4EBP1) related to the WNT/β-catenin Signaling, mTOR Signaling, and PI3K/AKT Signaling pathways in **Table 1**. MARK2 plays vital roles in modulating directional cancer cell migration, which is crucial for cancer metastasis.[33] SOX9 is a high mobility group (HMG) box transcription factor and plays essential roles in regulating CRC progression.[34] Expression of SOX9 is closely associated with the 5-year overall survival rate of CRC patients.[34] EIF4B regulates cancer cell proliferation and has been reported as a potential target for developing anti-cancer therapies.[35] Phosphorylation of EIF4EBP1 has been reported as an important regulator of cancer progression.[36]

We identified some phosphorylated proteoforms of those genes, which are unique to either SW480 or SW620 cells, **Table 1**. For example, two phosphorylated proteoforms of MARK2 and Sox9 in the WNT/β-catenin Signaling were exclusively identified in the SW480 cells; two phosphorylated proteoforms of EIF4B in the mTOR Signaling pathway were identified solely in the SW620 cells. SW480 and SW620 cells have different phosphorylated proteoforms of EIF4EBP1 in the PI3K/AKT Signaling pathway. We further manually checked the intensities of those proteoforms in the SW480 and SW620 raw files by matching the m/z, charge state, and migration time information from the database search. The proteoform intensity data agree well with the database search results, **Table 1**. For example, the three phosphorylated proteoforms identified solely in SW620 cells have roughly 6-60-fold higher intensity in SW620 cells compared to SW480 cells. The extracted ion electropherograms (EIEs) of the two EIF4B phosphorylated proteforoms from triplicate CZE-MS/MS analyses are shown in **Figures S4** and **S5**. The data further suggests good reproducibility of proteoform measurements in terms of base peak proteoform intensity from technical triplicates (relative standard deviations (RSDs) ≤25%). Protein phosphorylation is well known for modulating cancer progression, including CRC. Although the roles of those four genes in regulating cancer progression have been well studied, the specific functions of those phosphorylated proteoforms of the genes have not been investigated. Here, for the first time, we documented the significant differences in protein phosphorylation of those genes between a non-metastatic and a metastatic CRC cell lines in a proteoform-specific manner. Those phosphorylated proteoforms could be central to the progression of CRC metastasis.

Proteoforms with PTMs and single amino acid variants

Protein PTMs modulate their biological function. For example, protein N-terminal acetylation influences the stability, folding, binding, and subcellular targeting of proteins.[37] Protein phosphorylation is well known for regulating cell signaling, gene expression, and differentiation.[38] Protein methylation plays important roles in modulating transcription.[39] All the data analyses in the following parts of the manuscript are based on the combined data from SEC-CZE-MS/MS, RPLC-CZE-MS/MS, and SEC-RPLC-CZE-MS/MS corresponding to 23,319 proteoforms (**Supplementary Material II**) unless specified otherwise.

This large-scale TDP study identified 4,872 proteoforms with N-terminal acetylation (+42 Da mass shift), 319 proteoforms with phosphorylation [+80 Da (single phosphorylation) or +160 Da (double phosphorylation) mass shift], 321 proteoforms with methylation (+14 Da mass shift), and 241 proteoforms with oxidation (+16 Da mass shift), **Figure 4A**. TDP is powerful for the characterization of combinations of various PTMs on proteoforms. Here we identified 54 proteoforms with two phosphorylation sites and 90 proteoforms with both acetylation and phosphorylation PTMs. **Figure 4B** shows the sequences and fragmentation patterns of 28 kDa heat- and acid-stable phosphoprotein (PDAP1) and Calmodulin-1 (CALM1) proteoforms with either two phosphorylation sites or the combination of N-terminal acetylation and one lysine trimethylation. Those PTMs of the two proteins agree with the literature data.[40, 41] Those two proteoforms were identified with high confidence and were well characterized in terms of PTMs. PDAP1 and CALM1 are both prognostic markers of cancer according to the Human Protein Atlas (https://www.proteinatlas.org/). However, the potential roles of those specific proteoforms of PDAP1 and CALM1 in cancer are still not clear. The capability of TDP for delineating those proteoforms opens the door of further investigating their potential functions in CRC.

One important value of TDP is its capability for delineation of various proteoforms from the same gene (proteoform family).[42] **Figure 4C** shows one example of *CALM1* proteoform family. CALM1 modulates many enzymes (kinases and phosphatases), ion channels, and many other proteins by calcium-binding. We identified 75 proteoforms of *CALM1*. Nearly 70% of those proteoforms start at the position 2 with the N-terminal methionine removal. Various truncated proteoforms, for example, with the starting positions around 40, 60, 80 and 120, were identified in a much lower frequency. The number of proteoform spectrum matches (PrSMs) can be used to roughly estimate the relative abundance of proteoforms.[21] For the *CALM1* proteoforms starting from position 2, about 90% of the corresponding PrSMs match to proteoforms covering the whole protein sequence (2-149), called intact proteoforms. The PrSMs corresponding to other Cterminally truncated proteoforms only account for 3% or lower. The intact proteoforms have various PTMs, including acetylation/trimethylation, oxidation, and phosphorylation. The intact proteoforms of *CALM1* with a 42-Da mass shift (acetylation/trimethylation) are the most abundant forms; intact proteoforms with additional oxidation (a 58-Da mass shift) or phosphorylation (a 122-Da mass shift) have much lower abundance according to the number of PrSMs of those proteoforms.

Cancers result from gene mutations, which produce proteoforms containing amino acid variants (AAVs). Although transcriptomic analysis can provide ample information about gene mutations and possible AAVs on proteins, it is valuable to detect proteoforms containing AAVs directly because gene expression can be regulated post-transcriptionally. BUP has been used for the identification of peptides containing single AAVs (SAAVs) from cancer cells.[43] The Kelleher group reported the identification of 10 proteoforms containing SAAVs from breast tumor xenografts in one TDP study.[44] Here we identified 111 proteoforms containing SAAVs of 82 genes from the SW480 and SW620 cell lines with a proteogenomic approach with a 5% proteoform-level FDR, representing one order of magnitude improvement in the number of identified proteoforms containing SAAVs compared to previous studies of cancer cells, **Figure 4D**. The SEC-CZE-MS/MS and RPLC-CZE-MS/MS (RPLC 6 fractions) data were used for the analysis. The transcriptomic variants based on the available RNA-Seq data were incorporated into the protein database for the identification of proteoforms containing SAAVs using TopPG, a recently developed bioinformatics tool.[45] We also manually inspected the MS/MS spectra of proteoforms containing the SAAV sites to ensure high-confidence IDs. Only 20% of the 111 proteoforms were identified from both cell lines, indicating potentially different SAAV profiles between the two cell lines, **Figure 4D**. To confirm the conclusion about SAAV proteoform profile differences, we further analyzed the SAAV containing proteoforms from 1-D CZE-MS/MS, **Figure S6**. Although the number of SAAV proteoforms from SW620 cells is about twice as many as that from SW480 cells, only half of the SW480 SAAV proteoforms are covered by the SW620 ones. Manual evaluation of some SAAV proteoforms exclusively identified from SW480 and SW620 cells in raw MS data supported the conclusion. **Figure S7** shows the EIEs of one TP53 proteoform containing SAAV from triplicate measurements of SW480 and SW620 cells. The TP53 proteoform was only identified in SW620 cells via MS/MS and its base peak intensity in SW620 cells was about 8-fold higher than that in SW480 cells (5.6±0.6E4 *vs.* 0.7±0.3E4).

Figure 4E shows the sequences and fragmentation patterns of two examples of proteoforms containing SAAVs. TP53 is an important tumor suppressor closely related to CRC development, and it is an essential member in WNT/β-catenin Signaling and PI3K/AKT Signaling pathways. We identified one TP53 proteoform containing an AAV at position 72 ($P \longrightarrow R$) due to the codon 72 polymorphism. Studies have shown the functional differences of the P72 and R72 proteoforms of TP53.[46,47] For example, the R72 proteoform does a markedly better job of inducing apoptosis compared to the P72 proteoform.[46] Another study indicated that the expression of P72 proteoform increased CRC metastasis, and that the R72 proteoform does not exist in the non-metastatic CRC cell line (SW480) based on the nucleic-acid data.[47] Interestingly, we only identified the R72 proteoform of TP53 in the SW620 cell line, not in the SW480 cell line, from the top-down MS data. *MSH6* is one of the DNA mismatch repair genes and its mutations play a crucial role in Lynch syndrome, which is an inherited form of CRC. We identified one MSH6 proteoform containing a SAAV due to polymorphism at position 39 ($G \longrightarrow E$). The G39E SAAV has been associated with an increased risk of CRC according to the nucleic-acid data.[48] We identified G39 proteoforms of MSH6 in both SW480 and SW620 cells, but identified the E39 proteoform only in the SW480 cells, not in the SW620 cells.

For the proteoforms containing SAAVs, we further performed QIAGEN Ingenuity Pathway Analysis (IPA) of the corresponding 82 genes. We revealed that 75 of those genes are associated with tumorigenesis of tissue (p-value: 0.0001), and three genes (MSH6, PITX1 and TP53) relate to the development of colon tumor (p-value: 0.002). Five of the genes related to tumorigenesis of tissue (AURKA, EIF5A, PFKFB3, POLE4, and TP53) are targets of cancer drugs. We further performed IPA network analysis and revealed that 17 out of the 82 genes are involved in a cancer-related network (network score 36), **Figure 4F**, suggesting their crucial roles in cancer and development. The 17 genes are highlighted in purple and those proteins belong to several different families, including enzyme (diamond shape, *LARS1*, *PARS1*, *ALDOA*, *MSH6*, and *PPIF*), phosphatase/kinase (triangle shape, *PGAM1*, *SET*, and *PFKFB3*), transcription regulator (oval shape, *TP53* and *PITX1*), and others (circle shape, *PSG1*, *SRP14*, *MAGEB2*, *MT1G*, *MT1H*, *MT1M*, and *ISG15*). Nine of those highlighted proteins have direct (solid line) or indirect (dotted line) interactions with TP53.

Quantitative TDP of metastatic and non-metastatic human CRC cell lines

We further carried out the first quantitative TDP study of a pair of metastatic (SW620) and non-metastatic (SW480) human CRC cell lines. The cell lysates of SW480 and SW620 cells were fractionated by SEC and each fraction was analyzed by CZE-MS/MS in technical triplicate. After database search with TopPIC, we identified roughly 4,000 proteoforms per replicate per cell line with a 1% proteoform-level FDR. The overall intensity distributions of identified proteoforms across technical triplicates and the two cell lines are consistent, **Figure S8**. We performed label-free quantification (LFQ) analysis using TopDiff (version 1.3.4), a tool in the TopPIC suite, which reported about 1,500 proteoforms with measured intensities in all the six samples (three replicates per cell line and two cell lines). The SEC-CZE-MS/MS system shows reasonably good reproducibility regarding the intensities of shared proteoforms, as evidenced by the strong linear correlations of proteoform intensities between technical replicates of SW480 or SW620 cells (Pearson correlation coefficients: 0.86-0.93), **Figure S9**. The Pearson correlation coefficients of proteoform intensity between SW480 and SW620 cells are statistically significantly lower than that between technical replicates of one cell line (0.71±0.01 vs. 0.90±0.03, $p<10^{-10}$, two-tailed student’s t-test), indicating significant differences between the two cell lines in terms of proteoform intensity. We used the Perseus software for further data analysis.[49] The two cell lines can be easily distinguished using the proteoform quantification profiles, **Figure 5A**. Two clusters of differentially expressed proteoforms across the six samples were revealed.

According to the volcano plot in **Figure 5B**, 460 proteoforms of 248 proteins showed statistically significant differences in abundance between the two cell lines (FDR<0.05). Specifically, 244 proteoforms of 152 proteins had higher abundance in the SW480 cell line and 216 proteoforms of 132 proteins had higher expression in the SW620 cell line. **Figure 5B** shows that one HMGN1 proteoform and one RBM8A proteoform have the most significant abundance changes between SW480 and SW620 cells. HMGN1 regulates gene expression and PTMs of core histones, affecting DNA repair and tumor progression.[50] It has been reported that RBM8A promotes tumor cell migration and invasion in the most common type of primary liver cancer.[51]

Comparing the overexpressed and underexpressed proteoforms in the two cell lines revealed that 36 genes (*e.g., DAP*, *CALM1*, *HDGF*, *JPT1*, and *NPM1*) have both overexpressed and underexpressed proteoforms in one cell line, suggesting that different proteoforms of the same gene had completely different expression patterns in the two cell lines. **Figure 5C** shows two differentially expressed proteoforms of *DAP* (Death-associated protein 1), one of those 36 genes. It has been reported that DAP modulates cell death and correlates with the clinical outcome of CRC patients.[52] Interestingly, we revealed that one phosphorylated proteoform of DAP (~7,607 Da, phosphorylation site S51 or T56) had a higher abundance in SW480 cells and another phosphorylated proteoform (~4,605 Da, phosphorylation site S51) showed higher expression in SW620 cells. Both the S51 and T56 are known to be phosphorylated according to PhosphoSitePlus, with S51 being the most common phosphorylation site of DAP. We noted that the differentially expressed proteoforms in this study include phosphorylated proteoforms of several important genes related to CRC, i.e., *RALY*,[53] *NPM1*,[54] *DAP*,[52] and *HDGF*,[55] **Table S2**. The functions of phosphorylated forms of those four proteins in modulating CRC development are still unclear. However, the differential expressions of those phosphorylated proteoforms in the metastatic and non-metastatic CRC cells suggest their potential roles in regulating CRC metastasis. We also manually checked the MS raw data of three of the phosphorylated proteoforms in **Table S2** (*NPM1*, *RALY*, and *HNRNPC*), and their EIEs are shown in **Figures S10**, **S11**, and **S12**. The results clearly indicate their significantly higher abundance in SW620 cells compared to SW480 cells, agreeing well with the data in **Table S2**.

We highlight several differentially expressed proteoforms of CALM1, JPT1 (HN1), and EPCAM. CALM-dependent systems play important roles in cancer metastasis.[56] JPT1 (HN1) promotes cancer metastasis via activating the NF-ҟB signaling pathway.[57] EPCAM is a human cell surface glycoprotein and plays crucial roles in tumor biology, especially CRC.[58] EPCAM has been recognized as an important therapeutic target for cancer. We discovered two CALM1 proteoforms having significantly higher abundance in SW620 cells compared to SW480 cells; one of them contains K116 trimethylation. We revealed one CALM1 proteoform showing higher abundance in SW480 cells and the proteoform carries N-terminal acetylation and a 58-Da mass shift between amino acid residues 73 and 89. The 58-Da mass shift can be explained as a trimethylation/acetylation plus oxidation. Three of JPT1 proteoforms have higher abundance in SW480 cells and one of them contains a 167-Da mass shift between the amino acid residues 66 and 89, where seven serine residues can be phosphorylated according to the PhosphoSitePlus database (https://www.phosphosite.org/). The 167-Da mass shift most likely represents a combination of phosphorylation and other PTMs. Interestingly, one JPT1 proteoform shows higher abundance in SW620 cells. We also observed two EPCAM proteoforms having higher abundance in SW480 cells.

We then performed IPA analyses of the genes of those differentially expressed proteoforms between SW480 and SW620 cells. Those genes are heavily involved in cancer-related diseases, for example, tumorigenesis of tissue and metastasis, **Figure 5D**. Five of those proteins (EIF4E, EPCAM, FKBP1A, GAA, and HSP90AB1) are drug targets. IPA network analyses revealed that 26 proteins (highlighted in purple) whose proteoforms showed higher abundance in SW480 compared to SW620 were involved in a cancer-related network (score 51), **Figure 5E**. Those proteins belong to several families, including enzyme (diamond shape, e.g., PARK7 and FKBP4), transcription regulator (oval shape, e.g., FUBP1), translation regulator (hexagon shape, e.g., CIRBP and EEF1A1), transporter (trapezium shape, e.g., SLC12A2 and LASP1), and others (circle shape, e.g., EPCAM and JPT1). Most of those proteins have direct (solid line) and indirect (dotted line) interactions with one another. We also carried out network analysis for the proteins whose proteoforms had higher expression in SW620 cells, and observed high-scores for cancer-related networks. **Figure 5F** shows one cancer-related network (score 54), and 26 of those proteins are involved in the network (highlighted in purple). Those proteins include several CRC-related important proteins, NPM1 (oval shape, transcription regulator, located in nucleus), DAP (transcription regulator, located in cytoplasm), and HDGF (square shape, growth factor, located in extracellular space). NPM1 is a crucial protein in the network and many of the highlighted proteins have direct interactions (solid line) with NPM1, for example, PARK7, VIM, and PPIA. NPM1 also has indirect interaction (dotted line) with the NFkB complex, which plays crucial roles in modulating DNA transcription and cell survival. Human NPM1 boosts the activation of NFkB according to Ingenuity relationships from the IPA analysis. Besides NPM1, several other highlighted proteins (e.g., HDGF and DAP) also have indirect interactions with the NFkB complex. For example, NFkB regulates the transcription of *HDGF*, and DAP deactivates the NFkB according to the IPA network analysis results. The IPA analysis also revealed that 13 proteoforms of three genes (EIF4B, EIF4E, EIF4EBP1) in the mTOR Signaling pathway had statistically significant differences in abundance between the SW480 and SW620 cells (**Supplementary Material II**).

Discussion

TDP is facing technical challenges for deep proteoform profiling of human cells. Although significant technical progresses have been achieved in LC-MS/MS-based TDP during the last two decades, the number of proteoform IDs per human cell line has been stabilized on the level of 3,000 for a decade.[18-22] Alternative strategies for deep TDP of human cells are needed. CZE-MS/MS has been recognized as one alternative strategy for TDP,[23,29,30,59] most likely due to the high separation efficiency of CZE and high sensitivity of CZE-MS for proteoform separation and detection. However, the performance of CZE-MS/MS for TDP profiling of human cell proteoforms is limited due to the extremely high sample complexity and limited sample loading capacity of CZE, which is evidenced by the 1D-CZE-MS/MS data of CRC cells in this work and our previous work with the identification of only hundreds of human proteoforms in one run.[24] In this study, we advanced TDP of human cells drastically in terms of the number of proteoform IDs per human cell line compared to previous LC-MS/MS-based studies (~16,000 vs. ~3,000) via coupling LC fractionations to CZE-MS/MS. This work represents an important progress in TDP, which aims to characterize the human proteome in a proteoform-specific manner (Human Proteoform Project).[16] We need to highlight that SEC-CZE-MS/MS and RPLC-CZE-MS/MS under optimized conditions will be powerful analytical techniques for deep TDP of human cells with high throughput, **Figure 2A**. CZE-MS/MS analyses of only six SEC fractions of a SW480 cell lysate produced about 4,000 proteoform IDs and roughly 700 proteoform IDs per CZE-MS/MS run. The data indicate that it is feasible now using LC-CZE-MS/MS (i.e., SEC-CZE-MS/MS) for deep TDP profiling of a large number of human cell types, which will potentially transform basic and translational biomedical research. The MS raw data have been deposited to the ProteomeXchange Consortium via the PRIDE [60] partner repository with the dataset identifier PXD029703.

TDP of metastatic and non-metastatic cells is crucial for discovering new protein biomarkers and providing a more accurate understanding of molecular mechanisms of cancer metastasis. According to the results from our qualitative and quantitative TDP of SW480 and SW620 cells, we had several conclusions about CRC metastasis. First, CRC cells have a significant transformation in proteoforms and SAAVs after metastasis, evidenced by obvious differences of proteoform and SAAV profiles between SW480 and SW620 cells. Second, different proteoforms from the same cancer-related gene (e.g., DAP, CALM1, HDGF, JPT1, RALY, and NPM1) may have potentially varied biological functions in modulating CRC metastasis, because they show opposite expression profiles between the SW480 and SW620 cells, **Figure 5B**. Some proteoforms of those genes have higher abundance in SW480 cells; some of their proteoforms show higher expression in SW620 cells. Third, PTMs (i.e., phosphorylation) of important cancer-related genes (i.e., DAP, HDGF, JPT1, RALY, NPM1, MARK2, SOX9, EIF4B, and EIF4EBP1) could play important roles in regulating CRC metastasis, evidenced by the significant abundance differences of phosphorylated proteoforms from those genes between the SW480 and SW620 cells. The differentially expressed proteoforms, especially those with PTMs, of important cancer-related genes could be novel proteoform biomarkers of CRC metastasis. Fourth, proteoforms of genes in well-known CRC-related pathways (WNT/β-catenin Signaling, PI3K/AKT Signaling, mTOR Signaling, and ERK/MAPK Signaling) are different between SW480 and SW620 cells, and those proteoforms could play vital roles in modulating CRC metastasis.

Our TDP strategies still have some technical limitations. One relates to the identification of large proteoforms. In this work, we focused on the characterization of proteoforms smaller than 30 kDa. CZE-MS/MS has much lower sample loading capacity compared to RPLC-MS/MS (nL vs. μL), resulting in a limited mass of protein materials that can be injected for measurements with CZE-MS/MS. This issue is particularly severe for the characterization of large proteoforms in a complex proteome sample because large proteoforms tend to have drastically lower signal-to-noise ratios than small proteoforms due to the much wider charge state distributions. Highly efficient size-based fractionation techniques must be employed to enrich large proteoforms before CZE-MS/MS. Additionally, more effort needs to be made to improve the sample loading capacity of CZE-MS/MS via investigating online sample stacking techniques or solid phase microextraction (SPME) methods. Another limitation relates to the extensive fragmentation of proteoforms for accurate localization of PTMs. The backbone cleavage coverage of proteoforms from commonly used collision-based fragmentation techniques (i.e., collision-induced dissociation (CID) and higher energy collision dissociation (HCD)) is limited. We expect that coupling our LC-CZE-MS/MS technique to a mass spectrometer with electron- or photon-based gas-phase fragmentation techniques (i.e., electron-capture dissociation (ECD),[61] electron-transfer dissociation (ETD),[62] and ultraviolet photodissociation (UVPD)[63]) will revolutionize TDP for the Human Proteoform Project.[16]

Materials and Methods

Materials and Reagents

MS-grade water, acetonitrile (ACN), methanol (MeOH), formic acid (FA) and HPLC-grade acetic acid (AA) were purchased from Fisher Scientific (Pittsburgh, PA). Ammonium bicarbonate (NH_4HCO_3), urea, dithiothreitol (DTT), iodoacetamide (IAA) and 3-(trimethoxysilyl)propyl methacrylate were from Sigma-Aldrich (St. Louis, MO). Hydrofluoric acid (HF, 48-51% solution in water) and acrylamide were purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail (EASYpacks) was from Roche (Indianapolis, IN).

Sample Preparation

SW480 (catalogue number CCL-228) and SW620 (catalogue number CCL-227) original cell lines were both purchased from ATCC (Manassas, VA) and were cultured in RPMI 1640 cell culture medium (Life Technologies Corporation, Grand Island, NY) supplemented with 10% fetal bovine serum (Thermo Scientific, Gaithersburg, MD) and 2mM L-glutamine (Invitrogen, San Diego, CA). The cells were incubated at 37°C with 5% CO2 and were passaged every 3-4 days. Both cell lines were last verified by Short Tandem Repeat (STR) sequencing in 2016 and were used within two months after resuscitation from frozen aliquots at -80°C.

Upon growing to confluency, cells were harvested and cleansed of remaining cell culture medium via subsequent washing with HPLC grade water (Fisher Scientific, Pittsburgh, PA) and centrifugation for 5-minute intervals at 15000 × g until supernatant was clear. Proteins were then extracted using mammalian cell lysis buffer. Cell lysis buffer consisted of 8 M urea, 50 mM Tris (pH 8.2), 1 mM β-glycerophosphate, 1 mM phenylmethylsulfonyl fluoride, 75 mM sodium chloride, 1 mM sodium fluoride, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate, and one protease inhibitor cocktail. The reagents for cell lysis buffer were purchased from Sigma-Aldrich and complete EDTA-free protease inhibitor cocktail tablet was purchased from Roche. Lysis buffer was added to the harvested cells which then underwent sonication on ice three times for 1-minute intervals at 15% amplitude. The resulting extracted proteins were then clarified of cellular debris by centrifugation at 15,000 rpm for 10 minutes. Proteins were quantified using a bicinchoninic acid (BCA) protein assay (Thermo Scientific Pierce, Rockford, IL) and then stored at -80°C until preparation for MS analysis.

SW480 and SW620 proteins were denatured at 37 °C for 30 minutes, reduced at 37 °C for 30 minutes using DTT, and then alkylated at room temperature in the dark for 20 minutes using IAA. The excess IAA were quenched by adding DTT and reacting for 5 min at room temperature.

For the experiment 1 (RPLC-CZE-MS/MS), 200 μg of proteins from SW480 and SW620 cells were reduced, alkylated, and acidified, followed by RPLC fractionation into 13 fractions and CZE-MS/MS. **For the experiment 2** (SEC-RPLC-CZE-MS/MS), 2 mg of proteins from SW480 and SW620 cells were reduced and alkylated before fractionated by SEC-RPLC and analyzed by CZE-MS/MS. **For the experiment 3** (RPLC-CZE-MS/MS), 420 μg of proteins from SW480 and SW620 cells were reduced and alkylated prior to fractionation by RPLC into 6 fractions and analyses by CZE-MS/MS. **For the experiment 4** (SEC-CZE-MS/MS), the samples were desalted after reduction and alkylation using a C4 trap column (4×10 mm, 3 μm particles, 300 Å pore size). Specifically, 500 μg of proteins from SW480 and SW620 cells was loaded onto the column and flushed with mobile phase A (2% (v/v) ACN, 0.1% FA) for 10 minutes at a flow rate of 1 mL/min. The proteins were eluted with mobile phase B (80% ACN, 0.1% FA) for 3 minutes at flow rate of 1 mL/min. The eluates were lyophilized with a speed vacuum and redissolved in 150 μL 0.1% formic acid (FA). Then proteins from SW480 and SW620 cells were fractionated by SEC into 6 fractions, followed by CZE-MS/MS analyses. **For the experiment 5** (1D-CZE-MS/MS), 100 μg of proteins from SW480 and SW620 cells were desalted using two methods. In one case, both samples were desalted by a C4 trap column as described in the experiment 4. In the other case, both samples were desalted by Amicon Ultra centrifugal filters with a molecular weight cutoff of 10 kDa. Desalting with centrifugal filter was performed by loading 100 μg of proteins onto the filter and washing the sample four times with 50 mM NH_4Ac at 14,000 × g. Finally, the sample was recovered in 30 μL of 50 mM NH_4Ac. The samples desalted with the C4 trap column and centrifugal filters were analyzed by 1D-CZE-MS/MS in technical triplicate.

Fractionation of the SW480 and SW620 proteome

All separations were performed on a 1260 Infinity II HPLC system from Agilent (Santa Clara, CA). Detection was performed using a UV-visible detector at a wavelength of 254 nm. Data was collected and analyzed using OpenLAB software. RPLC (C4, 2.1 × 250 mm, Sepax Technologies) and SEC (4.6 × 300 mm, 500 Å pores, Agilent) were performed offline (Agilent HPLC) for prefractionation. Fractions from SW620 and SW480 from experiment 1 (13 fractions × 2 samples), experiment 2 (84 fractions × 2 samples), experiment 3 (6 fractions × 2 samples), and experiment 4 (6 fractions × 2 samples) were analyzed by CZE-MS/MS, respectively.

In experiment 1, RPLC was used for sample fractionation with a 0.25 mL/min flow rate and gradient of 0-80% mobile phase (MP) B over 90 minutes (MPA: 2% ACN, 0.1% FA in water; MPB: 80% ACN, 0.1% FA in water). Fractions were collected from 15 to 22 minutes (fraction 1) and 22 to 70 minutes (12 fractions, 4 minutes per fraction). For experiment 2, both SEC and RPLC were used for fractionation prior to CZE-MS/MS. For SEC, the flow rate was 0.35 mL/min with a 0.05% TFA mobile phase. 2 mg of proteins in 800 µL solution was fractionated by SEC. Fractions were collected from 5-8 minutes (fraction 1) and 8-12.5 minutes (3 fractions, 1.5 minutes per fraction). One RPLC run was performed for each SEC fraction with a flow rate of 0.25 mL/min and gradient of 0-80% MPB (MPA: 2% ACN, 0.1% TFA in water; MPB: 10% IPA, 0.1% TFA in ACN) over 90 minutes with a 10-minutes equilibration with 100% MPA at the beginning of the separation. Fractions were collected from 20 to 25 minutes (fraction 1) and 25 to 65 minutes (20 fractions, 2 minutes per fraction). In experiment 3, RPLC fractionation was carried out using the same mobile phases as in experiment 1, and a 90-minute gradient was used with a 10-minute equilibration with 100% MPA at the beginning of the separation. Fractions were collected from 25 to 55 minutes (fraction 1), 50 to 70 minutes (4 fractions, 5 minutes per fraction), and 70 to 95 minutes (fraction 6). In experiment 4, SEC fractionation was performed with an Agilent Bio SEC-5 column (4.6 × 300 mm, 5 µm particles, 500 Å pore size). 220 µg of SW480 and SW620 proteins (1.5 mg/mL, 75 µL×2 injections) were loaded into the SEC column and separated isocratically at the flow rate of 0.3 mL/min with 0.1% FA as mobile phase. The first fraction is collected from 5.6 to 8.6 minutes. The second to the fifth fraction was from 8.6 to 14.6 minutes with 1.5 minutes per fraction. The final fraction was collected from 14.6 to 19.0 min. In the experiments 1-4, samples were dried down and redissolved in 50 mM NH_4HCO_3 (pH 8.0, ~2 mg/mL) for CZE-ESI-MS/MS.

CZE-MS/MS analysis

CZE separation was performed using a CESI 8000 Plus CE system (Beckman Coulter). A commercialized electrokinetically pumped sheath-flow CE-MS nanospray interface (CMP Scientific Corp) was applied for online coupling the CE system and mass spectrometer.[64,65] A glass emitter (orifice size: 20~30 µm) installed on the interface was filled with sheath buffer (0.2% FA, 10% methanol) to generate electrospray at voltage of 2-2.3 kV.

A 100 cm LPA coated fused silica capillary (50 µm i.d., 360 µm o.d.) was used for CZE separation in experiments 1, 2, 4 and 5, while a 70 cm LPA coated capillary (50 µm i.d., 360 µm o.d.) was employed for separation in experiment 3. The inner wall of the capillary was coated with LPA based on the procedure described in reference [66]. One end of the capillary was etched with HF to reduce the outer diameter of the capillary to about 70-80 µm based on the procedure described in reference [67]. (Caution: use appropriate safety procedures while handling hydrofluoric acid solutions)

In experiments 1, 2, 4 and 5, the capillary (100 cm) was loaded with 500 nL of sample. In experiment 3, the capillary (70 cm) was loaded with ~350 nL of sample. After sample loading, the capillaries were inserted into background electrolyte, containing 5% acetic acid (pH 2.4), and 30 kV voltage was applied at the sample injection end to carry out separations.

MS1 and MS2 data were collected on a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) under data-dependent acquisition (DDA) mode. The temperature of ion transfer tube was set to 320 °C and s-lens RF was 55. MS1 spectra were collected with following parameters: m/z range of 600-2000, mass resolution of 120,000 (at m/z 200), a microscan number of 3, AGC target value of 1E6, and maximum injection time of 100 ms. The top 5 most abundant precursor ions (charge state higher than 5, or charge state unassigned and intensity threshold 2E4) in the MS1 spectra were isolated with a window of 4 m/z and fragmented via HCD with NCE of 20%. The settings for MS2 spectra were resolution of 120,000 (at m/z 200), a microscan number of 3, AGC target value of 1E5, and maximum injection time of 200 ms. The dynamic exclusion was set to a duration of 30s and the isotopic peaks were excluded.

In experiments 2, 3, 4 and 5, each LC fraction was analyzed by CZE-MS/MS in triplicate. In experiment 1, each LC fraction was analyzed by a single CZE-MS/MS run. In total, 410 MS raw files with good protein signals were produced from experiments 1, 2, 3, and 4 for database search, including 26 MS raw files from experiment 1 (13 fractions × 2 samples), 312 MS raw files from experiment 2 (52 fractions × 2 samples × 3 replicates), 36 MS raw files from experiment 3 (6 fractions × 2 samples × 3 replicates), and 36 MS raw files from experiment 4 (6 fractions × 2 samples × 3 replicates). We need to note that we collected 84 fractions × 2 samples in the experiment 2. However, we only observed good protein signals from 52 LC fractions per sample. 12 MS RAW files were collected from the experiment 5 using CZE-MS/MS.

Data analysis for proteoform identification

All RAW files were analyzed with the TopPIC Suite (version 1.4.0) pipeline.[15,68] The RAW files were converted into mzML files with msconvert.[69] Then spectral deconvolution was performed with TopFD (version 1.4.0), which converts precursor and fragment isotope clusters into neutral monoisotopic masses and finds proteoform features by combining precursor isotope clusters with similar monoisotopic masses and close migration times in MS1 scans. The resulting mass spectra with monoisotopic neutral masses were stored in msalign files and the proteoform feature information was stored in text files. The human proteome database was downloaded from UniProt (UP000005640, 20350 entries, version October 23, 2019, only reviewed protein sequences were included) and concatenated with a random decoy database of the same size. Each msalign file was searched against the concatenated targe-decoy database using TopPIC (version 1.4.0). Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 1. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da. TopPIC reported a list of target and decoy proteoform-spectrum-matches (PrSMs) for each msalign file.

The proteoforms identified from all msalign files were merged and filtered with a proteoform-level FDR. First, the target and decoy PrSMs reported from all the msalign files were combined and filtered with a 5% spectrum-level FDR. The PrSMs were then clustered by grouping PrSMs into the same cluster if they were from the same protein and their precursor mass differences were not large than 2.2 Da. The PrSM with the best E-value was selected for each cluster and its proteoform was reported as the representative one for the cluster. The representative target and decoy proteoforms were finally filtered with a 1% proteoform-level FDR.

Proteoform quantification

There were 18 MS raw files from triplicate CZE-MS/MS analyses of the 6 SEC fractions for the SW480 or SW620 sample in experiment 4. The TopPIC suite pipeline reported a list of targe and decoy PrSM identifications for each raw file. Using the methods in the previous section, the PrSM identifications of the 36 MS raw files were merged and a list of proteoform identifications with a 1% proteoform-level FDR were reported. The abundance of a proteoform was computed as the sum of the proteoform abundances in the six SEC fractions, which were reported by TopFD. Proteoform identifications and their abundances were reported for each replicate using this method. Finally, TopDiff (version 1.4.0), a tool in TopPIC Suite, was used to match proteoform identifications across the three SW480 replicates and three SW620 replicates.

The quantitative results were further analyzed using Perseus software.[49] The intensities of each proteoform in triplicate CZE-MS/MS runs of SW480 and SW620 were normalized to the intensity of corresponding proteoform from the first run of SW480, converting proteoform intensity to proteoform ratio. Then, proteoform ratios of each run were divided by the corresponding median to make sure the ratios center at 1. After log2 transformation of all the data, the significantly differentially expressed proteoforms were determined by performing t-test analysis (FDR threshold: 0.05, S0: 1) using the Perseus software. The volcano plot [-log(p-value) vs. log2(fold change)] was generated.

Proteogenomic analysis

To generate sample-specific protein sequence databases with genetic variations for SW480 and SW620 cells, two RNA-Seq data sets (SRR8616059 for SW480 and SRR8615459 for SW620) [70] were downloaded from the Sequence Read Archive (SRA). The GATK pipeline [71] was employed to align short reads in the RNA-Seq data with the hg38 human genome to call single nucleotide variants (SNVs) and indels, which were further annotated using the gene-based annotation of ANNOVAR [72] (April 16, 2018). The annotated nonsynonymous SNVs and indels in exons were chosen for generating sample-specific protein sequence databases based on the basic annotation of the hg38 human genome in GENCODE [73]. Two sample-specific protein sequence databases were generated using TopPG [45] (version 1.0): one for SW480 cells and the other for SW620 cells. Each protein sequence database contained both reference protein sequences in the basic annotation of GENCODE and protein sequences with sample-specific variants. There were 74887 entries with 51485 reference sequences and 23402 sequences with variants in the database for SW480 cells and 75665 entries with 51432 reference sequences and 24233 sequences with sample-specific variants in the database for SW620 cells. The SW480 and SW620 mass spectra in experiments 3 and 4 were searched against their corresponding sample-specific database using TopPIC (version 1.4.0) with the same parameter setting in Section “Data analysis for proteoform identification”. Using the methods in Section “Data analysis for proteoform identification”, PrSMs identified in each cell line were combined and clustered, and proteoform identifications were filtered by a 5% proteoform-level FDR. Identifications with single amino acid variant (SAAV) sites were manually inspected. If a proteoform with SAAV sites contained no unexpected mass shifts or had at least three matched fragment ions between each SAAV site and the unexpected mass shift, it was reported as a confident proteoform identification with SAAV sites.

QIAGEN Ingenuity Pathway Analysis (IPA)

The cancer-related network analysis results shown in Figures 4F, 5E, and 5F were generated through the use of QIAGEN IPA (QIAGEN Inc., https://digitalinsights.qiagen.com/IPA).[74] Permissions have been granted by QIAGEN to use those copyrighted figures in this publication.

Statistical analysis

Data are presented as mean±standard deviations when available. For the statistical analysis of LFQ data of SW480 and SW620 cell lines, we performed both side t-test using the Perseus software [49] to determine the proteoforms with statistically significant abundance difference between the two cell lines with the following settings, S0=1 and FDR = 0.05.

References

1. M. Schmitt, F. R. Greten, The inflammatory pathogenesis of colorectal cancer. *Nat. Rev. Immunol.* **21**, 653-667 (2021).

2. S. K. Rehman, J. Haynes, E. Collignon, K. R. Brown, Y. Wang, A. M. L. Nixon, J. P. Bruce, J. A. Wintersinger, A. Singh Mer, E. B. L. Lo, C. Leung, E. Lima-Fernandes, N. M. Pedley, F. Soares, S. McGibbon, H. H. He, A. Pollet, T. J. Pugh, B. Haibe-Kains, Q. Morris, M. Ramalho-Santos, S. Goyal, J. Moffat, C. A. O'Brien, Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy. *Cell* **184**, 226-242 (2021).

3. S. D. Markowitz, M. M. Bertagnolli, Molecular Basis of Colorectal Cancer. *N. Engl. J. Med.* **361**, 2449-2460 (2009).

4. A. J. Schunter, X. Yue, A. B. Hummon, Phosphoproteomics of colon cancer metastasis: comparative mass spectrometric analysis of the isogenic primary and metastatic cell lines SW480 and SW620. *Anal. Bioanal. Chem.* **409**, 1749-1763 (2017).

5. B. Zhang, Clinical potential of mass spectrometry-based proteogenomics. *Nat. Rev. Clin. Oncol.* **16**, 256-268 (2019).

6. L. Xu, R. Wang, J. Ziegelbauer, W. W. Wu, R. F. Shen, H. Juhl, Y. Zhang, L. Pelosof, A. S. Rosenberg, Transcriptome analysis of human colorectal cancer biopsies reveals extensive expression correlations among genes related to cell proliferation, lipid metabolism, immune response and collagen catabolism. *Oncotarget* **8**, 74703-74719 (2017).

7. T. Huo, R. Canepa, A. Sura, F. Modave, Y. Gong, Colorectal cancer stages transcriptome analysis. *PLoS One* **12**, e0188697 (2017).

8. B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler, NCI CPTAC, Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387 (2014).

9. D. Besson, A. H. Pavageau, I. Valo, A. Bourreau, A. Bélanger, C. Eymerit-Morin, A. Moulière, A. Chassevent, M. Boisdron-Celle, A. Morel, J. Solassol, M. Campone, E. Gamelin, B. Barré, O. Coqueret, C. Guette, A quantitative proteomic approach of the different stages of colorectal cancer establishes OLFM4 as a new nonmetastatic tumor marker. *Mol. Cell. Proteomics* **10**, M111.009712 (2011).

10. D. Ghosh, H. Yu, X. F. Tan, T. K. Lim, R. M. Zubaidah, H. T. Tan, M. C. M. Chung, Q. Lin, Identification of Key Players for Colorectal Cancer Metastasis by iTRAQ Quantitative Proteomics Profiling of Isogenic SW480 and SW620 Cell Lines. *J. Proteome Res.* **10**, 4373-4387 (2011).

11. L. M. Smith, N. L. Kelleher, Consortium for Top Down Proteomics, Proteoform: a single term describing protein complexity. *Nat Methods* **10**, 186-187 (2013).

12. L. M. Smith, N. L. Kelleher, Proteoforms as the next proteomics currency. *Science* **359**, 1106-1107 (2018).

13. T. K. Toby, L. Fornelli, N. L. Kelleher, Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499-519 (2016).

14. I. Ntai, L. Fornelli, C. J. DeHart, J. E. Hutton, P. F. Doubleday, R. D. LeDuc, A. J. van Nispen, R. T. Fellers, G. Whiteley, E. S. Boja, H. Rodriguez, N. L. Kelleher, Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4140-4145 (2018).

15. Q. Kou, L. Xun, X. Liu, TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **32**, 3495-3497 (2016).

16. L. Smith, J. N. Agar, J. Chamot-Rooke, P. O. Danis, Y. Ge, J. A. Loo, L. Paša-Tolić, Y. O. Tsybin, N. L. Kelleher, Consortium for Top-Down Proteomics, The Human Proteoform Project: Defining the human proteome. *Sci. Adv.* **7**, eabk0734 (2021).

17. R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schlüter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlén, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, B, Zhang, How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206-214 (2018).

18. J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, C. Wu, S. M. M. Sweet, B. P. Early, N. Siuti, R. D. LeDuc, P. D. Compton, P. M. Thomas, N. L. Kelleher, Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254-258 (2011).

19. A. C. Catherman, K. R. Durbin, D. R. Ahlf, B. P. Early, R. T. Fellers, J. C. Tran, P. M. Thomas, N. L. Kelleher, Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics* **12**, 3465-3473 (2013).

20. L. C. Anderson, C. J. DeHart, N. K. Kaiser, R. T. Fellers, D. F. Smith, J. B. Greer, R. D. LeDuc, G. T. Blakney, P. M. Thomas, N. L. Kelleher, C. L. Hendrickson, Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *J. Proteome Res.* **16**, 1087-1096 (2017).

21. R. D. Melani, V. R. Gerbasi, L. C. Anderson, J. W. Sikora, T. K. Toby, J. E. Hutton, D. S. Butcher, F. Negrão, H. S. Seckler, K. Srzentić, L. Fornelli, J. M. Camarillo, R. D. LeDuc, A. J. Cesnik, E. Lundberg, J. B. Greer, R. T. Fellers, M. T. Robey, C. J. DeHart, E. Forte, C. L. Hendrickson, S. E. Abbatiello, P. M. Thomas, A. I. Kokaji, J. Levitsky, N. L. Kelleher, The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **375**, 411-418 (2022).

22. W. Cai, T. Tucholski, B. Chen, A. J. Alpert, S. McIlwain, T. Kohmoto, S. Jin, Y. Ge, Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Anal. Chem.* **89**, 5467-5475 (2017).

23. E. N. McCool, R. A. Lubeckyj, X. Shen, D. Chen, Q. Kou, X. Liu, L. Sun, Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome. *Anal. Chem.* **90**, 5529-5533 (2018).

24. Z. Yang, X. Shen, D. Chen, L. Sun, Toward a Universal Sample Preparation Method for Denaturing Top-Down Proteomics of Complex Proteomes. *J. Proteome Res.* **19**, 3315-3325 (2020).

25. R. A. Lubeckyj, A. R. Basharat, X. Shen, X. Liu, L. Sun, Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples. *J. Am. Soc. Mass Spectrom.* **30**, 1435-1445 (2019).

26. E. N. McCool, L. Sun, Comparing nanoflow reversed-phase liquid chromatography-tandem mass spectrometry and capillary zone electrophoresis-tandem mass spectrometry for top-down proteomics. *Se Pu* **37**, 878-886 (2019).

27. X. Han, Y. Wang, A. Aslanian, B. Fonslow, B. Graczyk, T. N. Davis, J. R. Yates 3rd, In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. *J. Proteome Res.* **13**, 6078-6086 (2014).

28. Z. Yang, X. Shen, D. Chen, L. Sun, Improved Nanoflow RPLC-CZE-MS/MS System with High Peak Capacity and Sensitivity for Nanogram Bottom-up Proteomics. *J. Proteome Res.* **18**, 4046-4054 (2019).

29. D. Chen, E. N. McCool, Z. Yang, X. Shen, R. A. Lubeckyj, T. Xu, Q. Wang, L. Sun, Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. *Mass Spectrom. Rev.* doi: 10.1002/mas.21714 (2021).

30. F. P. Gomes, J. R. Yates 3rd, Recent trends of capillary electrophoresis-mass spectrometry in proteomics research. *Mass Spectrom. Rev.* **38**, 445-460 (2019).

31. Z. Koveitypour, F. Panahi, M. Vakilian, M. Peymani, F. S. Forootan, M. H. N. Esfahani, K. Ghaedi, Signaling pathways involved in colorectal cancer progression. *Cell Biosci.* **9**, 97 (2019).

32. M. G. Francipane, E. Lagasse, mTOR pathway in colorectal cancer: an update. *Oncotarget* **5**, 49-66 (2014).

33. A. M. Pasapera, S. M. Heissler, M. Eto, Y. Nishimura, R. S. Fischer, H. R. Thiam, C. M. Waterman, MARK2 regulates directed cell migration through modulation of myosin II contractility and focal adhesion organization. *Curr Biol.* **32**, 2704-2718 (2022).

34. B. Lü, Y. Fang, J. Xu, L. Wang, F. Xu, E. Xu, Q. Huang, M. Lai, Analysis of SOX9 expression in colorectal cancer. *Am. J. Clin. Pathol.* **130**, 897-904 (2008).

35. D. Shahbazian, A. Parsyan, E. Petroulakis, J. Hershey, N. Sonenberg, eIF4B controls survival and proliferation and is regulated by proto-oncogenic signaling pathways. *Cell Cycle* **9**, 4106-9 (2010).

36. Y. Chen, J. Wang, H. Fan, J. Xie, L. Xu, B. Zhou, Phosphorylated 4E-BP1 is associated with tumor progression and adverse prognosis in colorectal cancer. *Neoplasma.* **64**, 787-794 (2017).

37. R. Ree, S. Varland, T. Arnesen, Spotlight on protein N-terminal acetylation. *Exp. Mol. Med.* **50**, 1-13 (2018).

38. D. E. Kalume, H. Molina, A. Pandey, Tackling the phosphoproteome: tools and strategies. *Curr. Opin. Chem. Biol.* **7**, 64-69 (2003).

39. D. Y. Lee, C. Teyssier, B. D. Strahl, M. R. Stallcup, Role of protein methylation in regulation of transcription. *Endocr. Rev.* **26**, 147-170 (2005).

40. H. Zhou, S. Di Palma, C. Preisinger, M. Peng, A. N. Polat, A. J. Heck, S. Mohammed, Toward a comprehensive characterization of a human cancer cell phosphoproteome. *J. Proteome Res.* **12**, 260-271(2013).

41. T. Sasagawa, L. H. Ericsson, K. A. Walsh, W. E. Schreiber, E. H. Fischer, K. Titani, Complete amino acid sequence of human brain calmodulin. *Biochemistry* **21**, 2565-2569(1982).

42. Y. Dai, K. E. Buxton, L. V. Schaffer, R. M. Miller, R. J. Millikin, M. Scalf, B. L. Frey, M. R. Shortreed, L. M. Smith, Constructing Human Proteoform Families Using Intact-Mass and Top-Down Proteomics with a Multi-Protease Global Post-Translational Modification Discovery Database. *J. Proteome Res.* **18**, 3671-3680 (2019).

43. Z. Tan, J. Zhu, P. M. Stemmer, L. Sun, Z. Yang, K. Schultz, M. J. Gaffrey, A. J. Cesnik, X. Yi, X. Hao, M. R. Shortreed, T. Shi, D. M. Lubman, Comprehensive Detection of Single Amino Acid Variants and Evaluation of Their Deleterious Potential in a PANC-1 Cell Line. *J. Proteome Res.* **19**, 1635-1646 (2020).

44. I. Ntai, R. D. LeDuc, R. T. Fellers, P. Erdmann-Gilmore, S. R. Davies, J. Rumsey, B. P. Early, P. M. Thomas, S. Li, P. D. Compton, M. J. C. Ellis, K. V. Ruggles, D. Fenyö, E. S. Boja, H. Rodriguez, R. R. Townsend, N. L. Kelleher, Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol. Cell. Proteomics* **15**, 45-46 (2016).

45. W. Chen, X. Liu, Proteoform Identification by Combining RNA-Seq and Top-Down Mass Spectrometry. *J. Proteome Res.* **20**, 261-269 (2021).

46. B. Jeong, W. Hu, V. Belyi, R. Rabadan, A. J. Levine, Differential levels of transcription of p53-regulated genes by the arginine/proline polymorphism: p53 with arginine at codon 72 favors apoptosis. *FASEB J.* **24**, 1347-1353 (2010).

47. V. R. Katkoori, U. Manne, L. S. Chaturvedi, M. D. Basson, P. Haan, D. Coffey, H. L. Bumpers, Functional consequence of the p53 codon 72 polymorphism in colorectal cancer. *Oncotarget.* **8**, 76574-76586 (2017).

48. P. Zelga, K. Przybyłowska-Sygut, M. Zelga, A. Dziki, I. Majsterek, Polymorphism of Gly39Glu (c.116G>A) hMSH6 is associated with sporadic colorectal cancer development in the Polish population: Preliminary results. *Adv. Clin. Exp. Med.* **26**, 1425-1429 (2017).

49. S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, J. Cox, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731-740 (2016).

50. Y. Postnikov, M. Bustin, Regulation of chromatin structure and function by HMGN proteins. *Biochim. Biophys. Acta* **1799**, 62-8 (2010).

51. R. Liang, Y. Lin, J. Ye, X. Yan, Z. Liu, Y. Li, X. Luo, H. Ye, High expression of RBM8A predicts poor patient prognosis and promotes tumor progression in hepatocellular carcinoma. *Oncol. Rep.* **37**, 2167-2176 (2017).

52. Y. Jia, L. Ye, K. Ji, A. Toms, M. L. Davies, F. Ruge, J. Ji, R. Hargest, W. G. Jiang, Death associated protein 1 is correlated with the clinical outcome of patients with colorectal cancer and has a role in the regulation of cell death. *Oncol. Rep.* **31**, 175-182 (2014).

53. L. Sun, A. Wan, Z. Zhou, D. Chen, H. Liang, C. Liu, S. Yan, Y. Niu, Z. Lin, S. Zhan, S. Wang, X. Bu, W. He, X. Lu, A. Xu, G. Wan, RNA-binding protein RALY reprogrammes

mitochondrial metabolism via mediating miRNA processing in colorectal cancer. *Gut* **70**, 1698-1712 (2021).

54. S. Grisendi, C. Mecucci, B. Falini, P. P. Pandolfi, Nucleophosmin and cancer. *Nat. Rev. Cancer* **6**, 493-505 (2006).

55. B. Sun, X. Gu, Z. Chen, J. Xiang, MiR-610 inhibits cell proliferation and invasion in colorectal cancer by repressing hepatoma-derived growth factor. *Am. J. Cancer Res.* **5**, 3635-3644 (2015).

56. A. Villalobo, M. W. Berchtold, The Role of Calmodulin in Tumor Cell Migration, Invasiveness, and Metastasis. *Int. J. Mol. Sci.* **21**, 765 (2020).

57. J. Chen, J. Qiu, F. Li, X. Jiang, X. Sun, L. Zheng, W. Zhang, H. Li, H. Wu, Y. Ouyang, X. Chen, C. Lin, L. Song, Y. Zhang. HN1 promotes tumor associated lymphangiogenesis and lymph node metastasis via NF-κB signaling activation in cervical carcinoma. *Biochem. Biophys. Res. Commun.* **530**, 87-94 (2020).

58. A. Armstrong, S. L. Eck, EpCAM: A new therapeutic target for an old cancer antigen. *Cancer Biol. Ther.* **2**, 320-6 (2003).

59. A. M. Belov, R. Viner, M. R. Santos, D. M. Horn, M. Bern, B. L. Karger, A. R. Ivanov, Analysis of Proteins, Protein Complexes, and Organellar Proteomes Using Sheathless Capillary Zone Electrophoresis - Native Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **28**, 2614-2634 (2017).

60. Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A. F. Jarnuczak, T. Ternent, A. Brazma, J. A. Vizcaíno, The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**(D1), D442-D450 (2019).

61. Y. Ge, B. G. Lawhorn, M. ElNaggar, E. Strauss, J.-H. Park, T. P. Begley, F. W. McLafferty, Top Down Characterization of Larger Proteins (45 kDa) by Electron Capture Dissociation Mass Spectrometry. *J. Am. Chem. Soc.* **124**, 672−678 (2002).

62. N. M. Riley, M. S. Westphall, J. J. Coon, Activated Ion Electron Transfer Dissociation for Improved Fragmentation of Intact Proteins. *Anal. Chem.* **87**, 7109-16 (2015).

63. J. B. Shaw, W. Li, D. D. Holden, Y. Zhang, J. Griep-Raming, R. T. Fellers, B. P. Early, P. M. Thomas, N. L. Kelleher, J. S. Brodbelt, Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation, *J. Am. Chem. Soc.* **135**, 12646–12651 (2013).

64. R. Wojcik, O. O. Dada, M. Sadilek, N. J. Dovichi, Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Commun. Mass Spectrom.* **24**, 2554-2560 (2010).

65. L. Sun, G. Zhu, Z. Zhang, S. Mou, N. J. Dovichi, Third-Generation Electrokinetically Pumped Sheath-Flow Nanospray Interface with Improved Stability and Sensitivity for Automated Capillary Zone Electrophoresis-Mass Spectrometry Analysis of Complex Proteome Digests. *J. Proteome Res.* **14**, 2312-2321 (2015).

66. G. Zhu, L. Sun, N. J. Dovichi, Thermally-initiated free radical polymerization for reproducible production of stable linear polyacrylamide coated capillaries, and their application to proteomic analysis using capillary zone electrophoresis-mass spectrometry. *Talanta* **146**, 839-843(2016).

67. L. Sun, G. Zhu, Y. Zhao, X. Yan, S. Mou, N. J. Dovichi, Ultrasensitive and Fast Bottom-up Analysis of Femtogram Amounts of Complex Proteome Digests. *Angew. Chem. Int. Ed.* **52**, 13661-13664 (2013).

68. X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna, P. A. Pevzner, Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. *Mol. Cell. Proteomics* **9**, 2772– 2782 (2010).

69. D. Kessner, M. Chambers, R. Burke, D. Agus, P. Mallick, ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534– 2536 (2008).

70. M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald 3rd, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet, B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, W. R. Sellers, Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503-508 (2019).

71. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

72. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids Res.* **38**, e164-e164 (2010).

73. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760-1774 (2012).

74. A. Krämer, J. Green, Jr. J. Pollard, S. Tugendreich, Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–30 (2014).

Acknowledgments

Funding: The work was funded by National Cancer Institute (NCI) through the grant R01CA247863 (Sun, Hummon, and Liu). We also thank the support from National Institute of General Medical Sciences (NIGMS) through grants R01GM125991 (Sun and Liu) and R01GM118470 (Liu and Sun). Sun also thanks the support from the National Science Foundation (CAREER Award, Grant DBI1846913). We thank MSU AgBioResearch and Michigan State University for the access to QIAGEN Ingenuity Pathway Analysis (IPA) platform.

Author contributions: E.N.M. performed the experiments for proteoform identifications using RPLC-CZE-MS/MS and SEC-RPLC-CZE-MS/MS. T.X. performed the experiment for proteoform identification and/or quantification using SEC-CZE-MS/MS and 1D-CZE-MS/MS. W.C. carried out all the database search using TopPIC for proteoform ID and quantification. E.N.M., T.X., and W.C. worked together for data analysis and made the first draft of the manuscript. N.C.B. did all the cell culture and initial sample preparation of SW480 and SW620 cells. S.M.N. performed the LC fractionations. A.B.H., X.L., and

L.S. conceived the original idea. X.L. supervised the database search part of the project. L.S. supervised the project. All authors provided comments and contributed to the final manuscript.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The MS raw data have been deposited to the ProteomeXchange Consortium via the PRIDE [60] partner repository with the dataset identifier PXD029703.

Table 1. Selected proteoforms of important genes related to WNT/β-catenin Signaling, mTOR Signaling, and PI3K/AKT Signaling pathways.

Gene	Pathway	Proteoform	SW480	SW620	Proteoform intensities (SW480/SW620)*
MARK2	WNT/β-catenin Signaling	M.(S)[Acetyl]SARTPLPTLNERDTEQPTLGHLDSK(PSSKSNMIRGRNSAT)[mass shift: 96 Da, phospho and oxidation]SADEQPHIGNY.R	×	ND	4.8E5/2.8E4
SOX9	WNT/β-catenin Signaling	R.SQYDYTDHQNSSSYYSHAAGQGTGLYSTFTYMNPAQRPMYTPIADTSGV(PSIPQTHS)[mass shift: 78 Da, phospho]PQHWEQPVYTQLTRP.	×	ND	3.0E5/4.6E4
EIF4B	mTOR Signaling	M.AASAKKKNK(KGKTISLTDFL)[mass shift: 122 Da, phospho and acetylation/trimethylation]AEDGGTGGGSTYVSKPVSWADETDDLEGDVSTTWHSNDDDVYRAPPIDRSILPTAPR.A	ND	×	7.5E4/4.4E5
EIF4B	mTOR Signaling	M.(A)[Acetyl]ASAKKKNKKGKTISLTDFLAEDGG(T)[mass shift: 80 Da, phospho]GGGSTYVSKPVSWADETDDLEGDVSTTWHSNDDDVYRAPPIDR.S	ND	×	5.0E4/3.1E5
EIF4EBP1	PI3K/AKT Signaling	.MSGGSS(C)[Carbamidomethylation]SQTPSRAIPAT(RRVVLGDGVQLPPGDYSTT)[mass shift: 81 Da, phospho]PGGTLFSTTPGGTRIIYDRKFLME(C)[Carbamidomethylation]RNSPVTKTPPRDLPTIPGVTSPSSDEPPMEASQSHLRNSPEDKRAGGEESQFEMDI.	ND	×	6.0E4/3.5E6
EIF4EBP1	PI3K/AKT Signaling	K.TPPRDLPTIPGVTS(PSSDEPPMEASQSHLRNS)[mass shift: 81Da, phospho]PEDKRAGGEESQFEMDI.	×	ND	1.5E5/5.0E4

“x” suggests that the proteoform is identified in the sample. “ND” indicates that the proteoform is not identified in the sample. * The proteoform intensities were observed by manually checking the raw data based on the migration time, charge, and m/z of proteoforms in the database search results. The average intensity of the identified charge state of each proteoform across the proteoform peak is shown in the table.

Figure 1. Schematic of the experimental design. (A) Schematic design of the TDP study of metastatic (SW620) and non-metastatic (SW480) CRC cells using CZE-ESI-MS/MS and LC-CZE-ESI-MS/MS for proteoform identification and label-free quantification. (B) Four CZE-MS/MS-based strategies in this work with the amounts of protein starting materials.

Figure 2. Summary of proteoform identification results of this study. (A) Proteoform IDs from SW480 cells using different CZE-MS/MS-based strategies. The error bars represent the standard deviations of the number of proteoform IDs from technical triplicates. (B) The number of proteoform and protein IDs per complex proteome sample using RPLC- or CZE-MS/MS-based TDP strategies. The data of studies 5, 6 and 7 are shown as mean ± standard deviations from various proteome samples. For example, the mean and standard deviations of proteoform and protein counts from SW480 and SW620 cells are shown in the studies 6 and 7. (C) Heat map of proteoform overlaps from technical triplicates of SW480 and SW620 cells using SEC-CZE-MS/MS. Each number in the figure represents a ratio between the number of shared proteoforms in two conditions (e.g., SW480_1 (x-axis) and SW620_1 (y-axis)) and the total number of identified proteoforms in one of the two conditions listed on the y-axis (e.g., SW620_1). For example, the proteoform overlap between SW480_1 (x-axis) and SW620_1 (y-axis) is 0.4, which indicates the ratio between the number of shared proteoforms in those two conditions and the total number of identified proteoforms in SW620_1. (D) Sequences and fragmentation patterns of identified example proteoforms in the study.

Figure 3. Summary of proteoforms from genes involved in well-known CRC-related pathways. (A) The number of proteoforms and genes in four CRC-related pathways identified from SW480 and SW620 cells. (B) Overlaps of identified and pathway-related proteoforms between SW480 and SW620 cells.

Figure 4. Analyses of the identified proteoforms from CRC cells with PTMs and single amino acid variants (SAAVs). (A) Proteoforms with various PTMs, including N-terminal acetylation, phosphorylation, methylation, and oxidation. (B) Sequences and fragmentation patterns of two proteoforms, one proteoform of PDAP1 with two phosphorylation sites and one proteoform of CALM1 with N-terminal acetylation and one lysine trimethylation. (C) Summary of all the identified proteoforms of calmodulin-1 (CALM1) regarding starting positions, relative abundance based on the number of PrSMs, and PTMs. (D) The number of proteoforms containing SAAVs identified from the SW480 and SW620 cells and the overlap of those proteoforms. The SEC-CZE-MS/MS and RPLC-CZE-MS/MS (RPLC 6 fractions) data were used for the analysis. The error bars in the figure represent the standard deviations of proteoforms from triplicate measurements. (E) Sequences and fragmentation patterns of two proteoforms containing SAAVs. (F) SAAVs containing proteoforms correspond to many genes (highlighted in purple) that are involved in a cancer related network according to the IPA analysis. The diamond, triangle, oval, and circle shapes represent proteins belonging to enzyme, phosphatase/kinase, transcription regulator and others, respectively. The solid and dotted lines represent direct and indirect interactions. Copyright permission has been granted by QIAGEN for using the network data.

Figure 5. Summary of the LFQ data of SW480 and SW620 cells from SEC-CZE-MS/MS in technical triplicate. (A) Heat map and cluster analysis of the quantified proteoforms (~1500 proteoforms) regarding LFQ intensities. A Z-score normalization was employed. The red color represents high intensity and the blue color indicates low intensity. (B) Volcano plot showing differentially expressed proteoforms between the two cell lines. The quantified proteoforms (~1500) were used for the analysis. Red dots and blue dots represent proteoforms having statistically significantly higher abundance in SW480 and in SW620, respectively. Gene names of

some differentially expressed proteoforms are labeled. The Perseus software was used for generating the heat map in (A) and Volcano plot in (B) with the following settings (S0=1 and FDR = 0.05).[49] (C) Sequences and fragmentation patterns of two phosphorylated proteoforms of the gene DAP. One has higher abundance in SW480 cells and the other has higher expression in SW620 cells. (D) An IPA analysis reported some cancer related diseases that are related to the differentially expressed genes in the two cell lines. Proteoforms with higher abundance in SW480 cells (E) or higher abundance in SW620 cells (F) correspond to genes that are involved in cancer-related networks with high scores. Those genes are highlighted in purple. The diamond, oval, hexagon, trapezium, square, and circle shapes represent enzyme, transcription regulator, translation regulator, transporter, growth factor, and others. The solid and dotted lines represent direct and indirect interactions. Copyright permission has been granted by QIAGEN for using the network data.

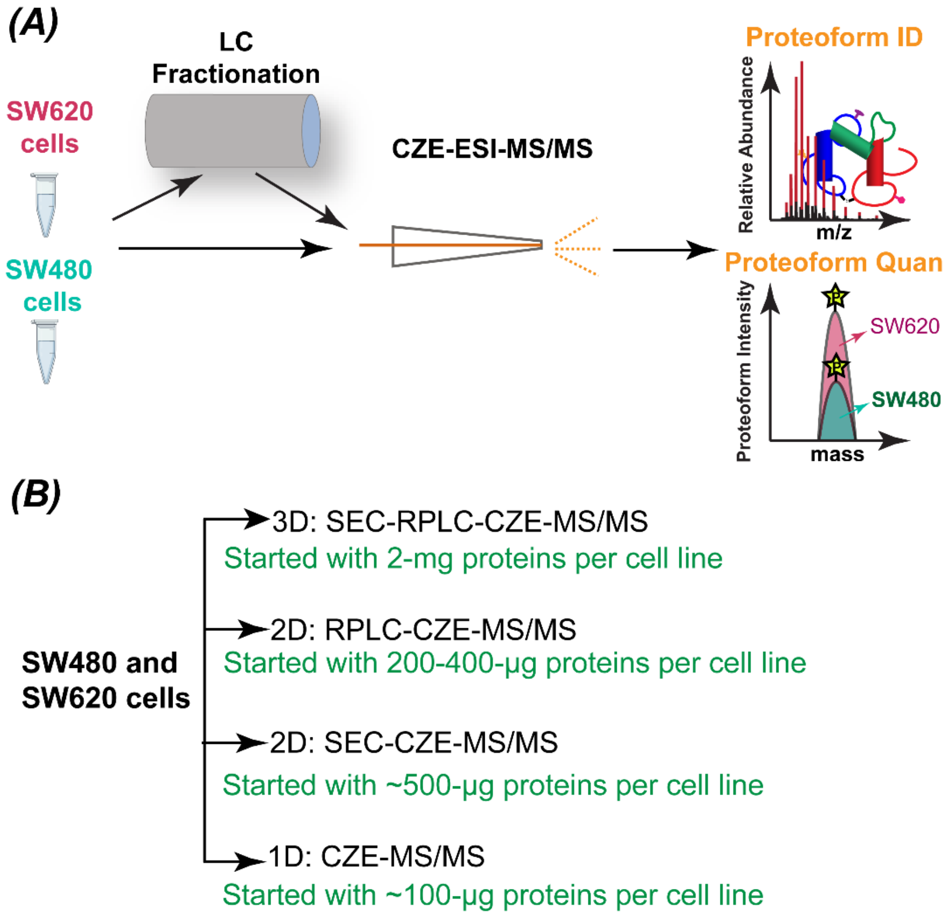
(A)
LC Fractionation
SW620 cells
SW480 cells
CZE-ESI-MS/MS
Proteoform ID
Relative Abundance
m/z
Proteoform Quan
Proteoform Intensity
SW620
SW480
mass
(B)
SW480 and SW620 cells
3D: SEC-RPLC-CZE-MS/MS
Started with 2-mg proteins per cell line
2D: RPLC-CZE-MS/MS
Started with 200-400-µg proteins per cell line
2D: SEC-CZE-MS/MS
Started with ~500-µg proteins per cell line
1D: CZE-MS/MS
Started with ~100-µg proteins per cell line

Figure 1

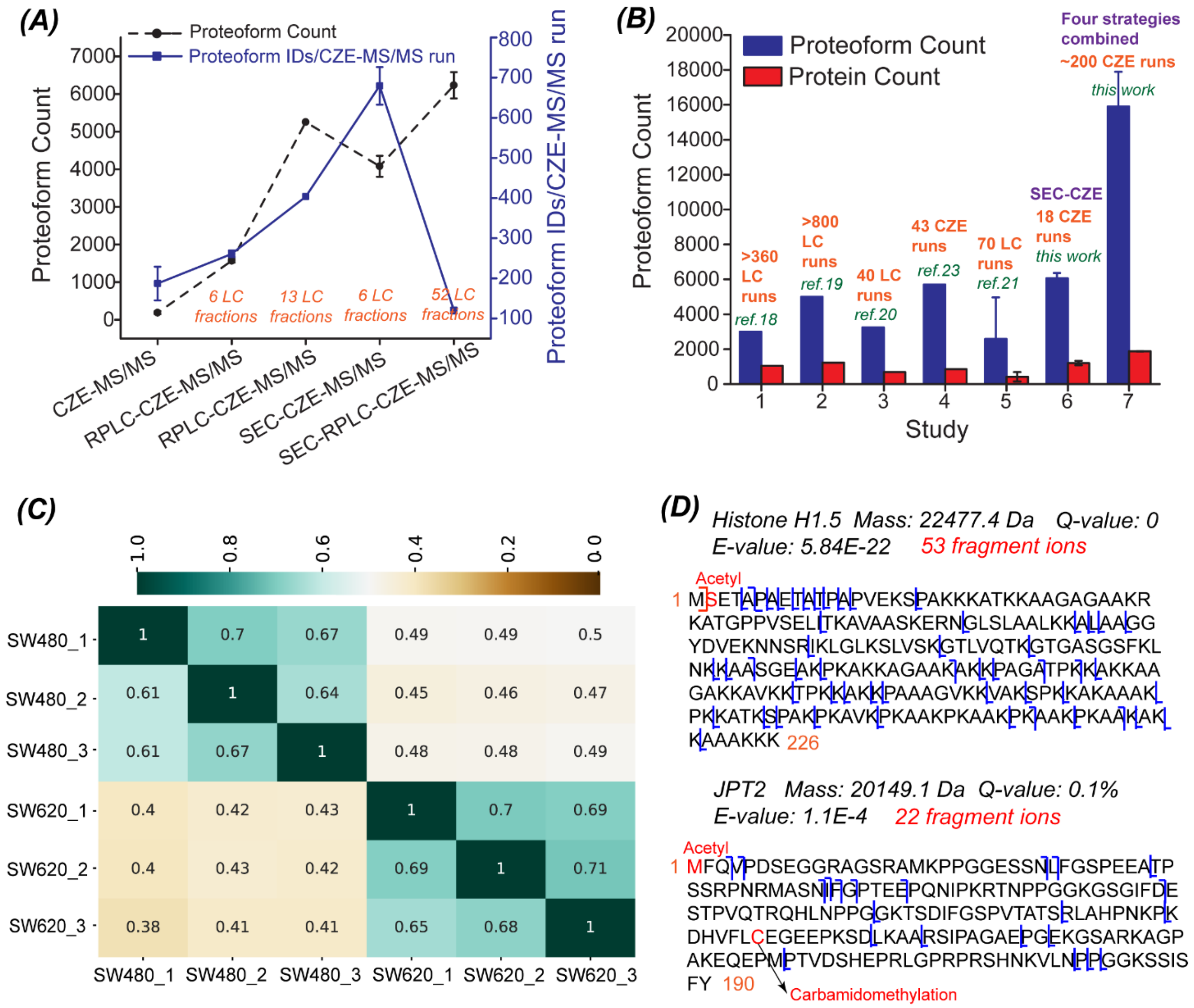
(A)
Proteoform Count
Proteoform IDs/CZE-MS/MS run
Proteoform Count
0
1000
2000
3000
4000
5000
6000
7000
Proteoform IDs/CZE-MS/MS run
100
200
300
400
500
600
700
800
6 LC fractions
13 LC fractions
6 LC fractions
52 LC fractions
CZE-MS/MS
RPLC-CZE-MS/MS
RPLC-CZE-MS/MS
SEC-CZE-MS/MS
SEC-RPLC-CZE-MS/MS
(B)
Proteoform Count
Protein Count
Proteoform Count
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
>360 LC runs
ref.18
>800 LC runs
ref.19
40 LC runs
ref.20
43 CZE runs
ref.23
70 LC runs
ref.21
SEC-CZE
18 CZE runs
this work
Four strategies combined
~200 CZE runs
this work
1
2
3
4
5
6
7
Study
(C)
1.0
0.8
0.6
0.4
0.2
0.0
SW480_1 1 0.7 0.67 0.49 0.49 0.5
SW480_2 0.61 1 0.64 0.45 0.46 0.47
SW480_3 0.61 0.67 1 0.48 0.48 0.49
SW620_1 0.4 0.42 0.43 1 0.7 0.69
SW620_2 0.4 0.43 0.42 0.69 1 0.71
SW620_3 0.38 0.41 0.41 0.65 0.68 1
SW480_1 SW480_2 SW480_3 SW620_1 SW620_2 SW620_3
(D)
Histone H1.5 Mass: 22477.4 Da Q-value: 0
E-value: 5.84E-22 53 fragment ions
Acetyl
1 MSETAPAETATPAPVEKSPAKKKATKKAAGAGAAKR
KATGPPVSELITKAVAASKERNGLSLAALKKALAAGG
YDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKL
NKKAASGEAKPKAKKAGAAKAKKPAGATPKKAKKAA
GAKKAVKKTPKKAKKPAAAGVKKVAKSPKKAKAAAK
PKKATKSPAKPKAVKPKAAKPKAAKPKAAKPKAAKAK
KAAAKKK 226
JPT2 Mass: 20149.1 Da Q-value: 0.1%
E-value: 1.1E-4 22 fragment ions
Acetyl
1 MFQVPDSEGGRAGSRAMKPPGGESSNLFGSPEEATP
SSRPNRMASNIFGPTEEPQNIPKRTNPPGGKGSGIFDE
STPVQTRQHLNPPGGKTSDIFGSPVTATSRLAHPNKPK
DHVFLCEGEEPKSDLKAARSIPAGAEPGEKGSARKAGP
AKEQEPMPTVDSHEPRLGPRPRSHNKVLNPPGGKSSIS
FY 190
Carbamidomethylation

Figure 2

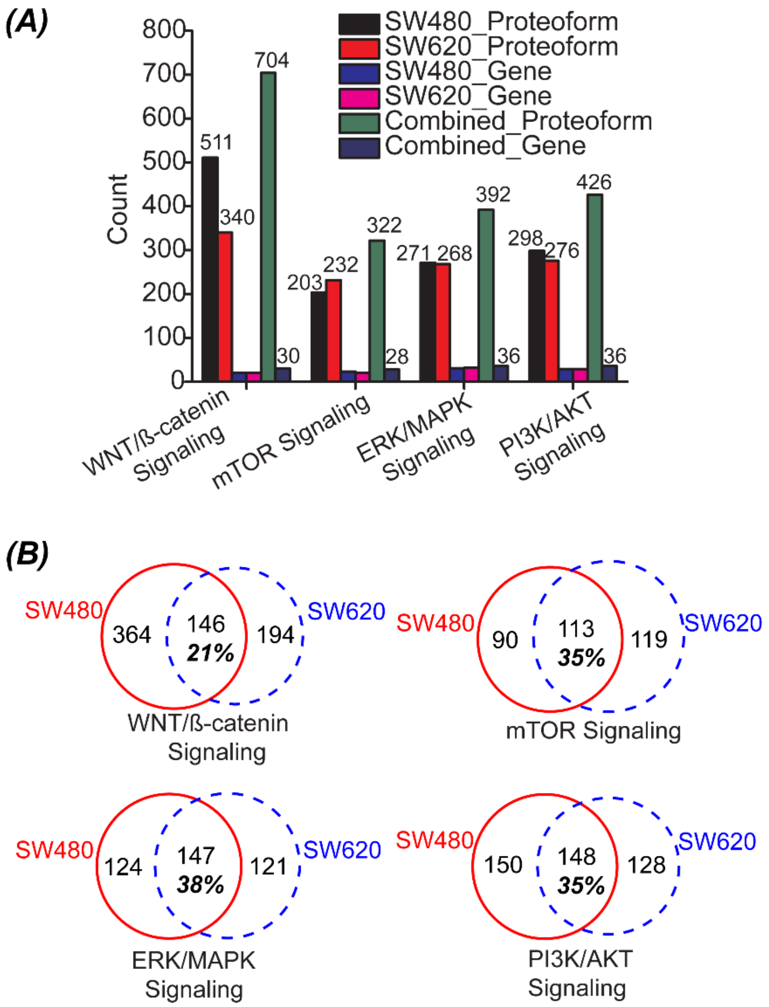
(A)
Count
800
700
600
500
400
300
200
100
0
SW480_Proteoform
SW620_Proteoform
SW480_Gene
SW620_Gene
Combined_Proteoform
Combined_Gene
511
340
704
30
203
232
322
28
271
268
392
36
298
276
426
36
WNT/ß-catenin Signaling
mTOR Signaling
ERK/MAPK Signaling
PI3K/AKT Signaling
(B)
SW480
364
146
21%
194
SW620
WNT/ß-catenin Signaling
SW480
90
113
35%
119
SW620
mTOR Signaling
SW480
124
147
38%
121
SW620
ERK/MAPK Signaling
SW480
150
148
35%
128
SW620
PI3K/AKT Signaling

Figure 3

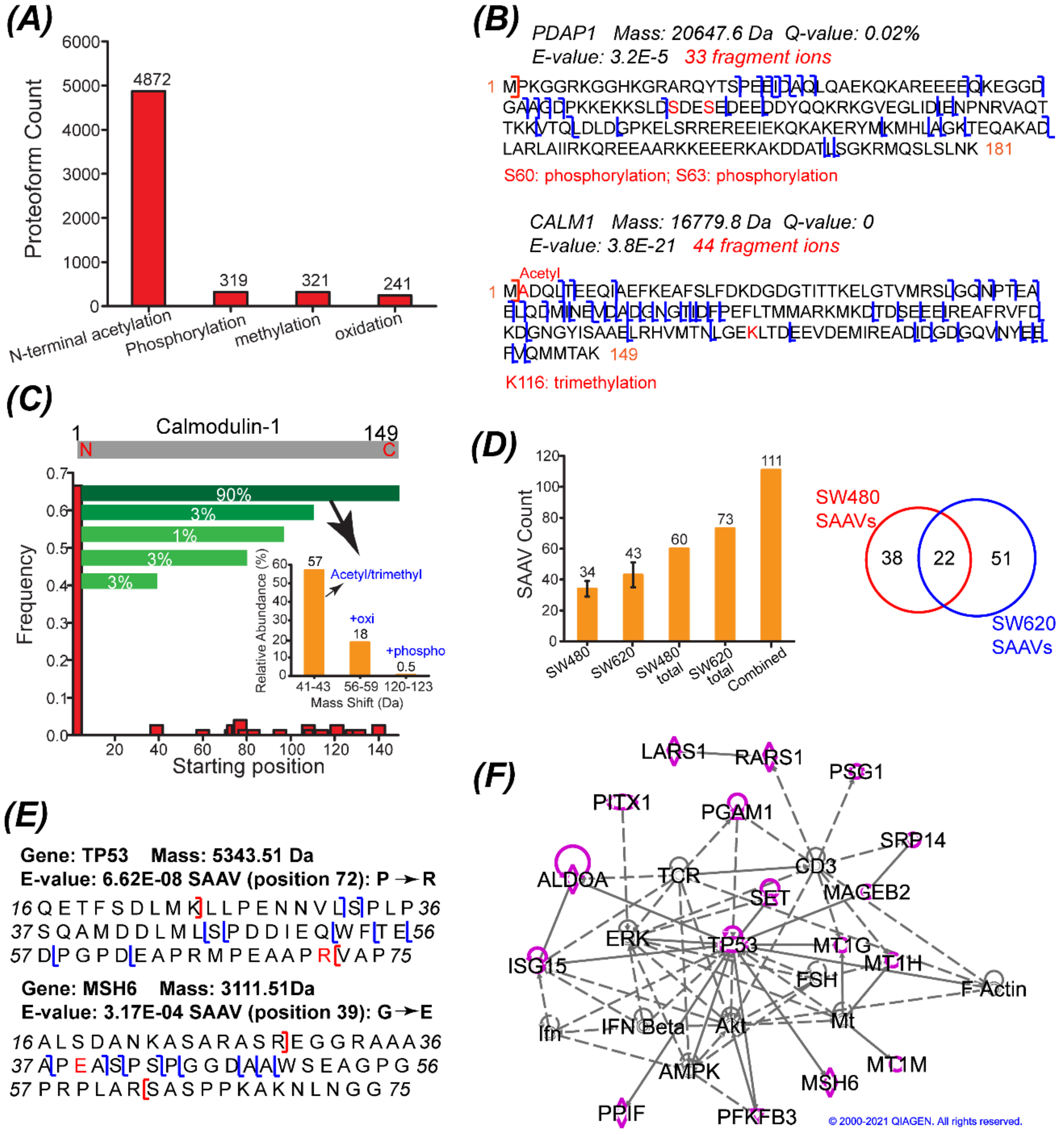
(A)
Proteoform Count
6000
5000
4000
3000
2000
1000
0
4872
319
321
241
N-terminal acetylation
Phosphorylation
methylation
oxidation
(B)
PDAP1 Mass: 20647.6 Da Q-value: 0.02%
E-value: 3.2E-5 33 fragment ions
1 MPKGGRKGGHKGRARQYTSPEEIDAQLQAEKQKAREEEEQKEGGD
GAAGDPKKEKKSLDSDESEDEEDDYQQKRKGVEGLIDIENPNRVAQT
TKKVTQLDLDGPKELSRREREEIEKQKAKERYMKMHLAGKTEQAKAD
LARLAIIRKQREEAARKKEEERKAKDDATLSGKRMQSLSLNK 181
S60: phosphorylation; S63: phosphorylation
CALM1 Mass: 16779.8 Da Q-value: 0
E-value: 3.8E-21 44 fragment ions
Acetyl
1 MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEA
ELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFD
KDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEE
FVQMMTAK 149
K116: trimethylation
(C)
Calmodulin-1
1
149
N
C
Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
90%
3%
1%
3%
3%
Relative Abundance (%)
60
50
40
30
20
10
0
57
Acetyl/trimethyl
+oxi
18
+phospho
0.5
41-43
56-59
120-123
Mass Shift (Da)
20
40
60
80
100
120
140
Starting position
(D)
SAAV Count
120
100
80
60
40
20
0
34
43
60
73
111
SW480
SW620
SW480 total
SW620 total
Combined
SW480 SAAVs
38
22
51
SW620 SAAVs
(E)
Gene: TP53 Mass: 5343.51 Da
E-value: 6.62E-08 SAAV (position 72): P → R
16 Q E T F S D L M K L L P E N N V L S P L P 36
37 S Q A M D D L M L S P D D I E Q W F T E 56
57 D P G P D E A P R M P E A A P R V A P 75
Gene: MSH6 Mass: 3111.51Da
E-value: 3.17E-04 SAAV (position 39): G → E
16 A L S D A N K A S A R A S R E G G R A A A 36
37 A P E A S P S P G G D A A W S E A G P G 56
57 P R P L A R S A S P P K A K N L N G G 75
(F)
LARS1
RARS1
PSG1
PITX1
PGAM1
SRP14
ALDOA
TCR
CD3
SET
MAGEB2
ERK
TP53
MT1G
ISG15
MT1H
FSH
F-Actin
Ifn
IFN Beta
Akt
Mt
AMPK
MT1M
MSH6
PPIF
PFKFB3
© 2000-2021 QIAGEN. All rights reserved.

Figure 4

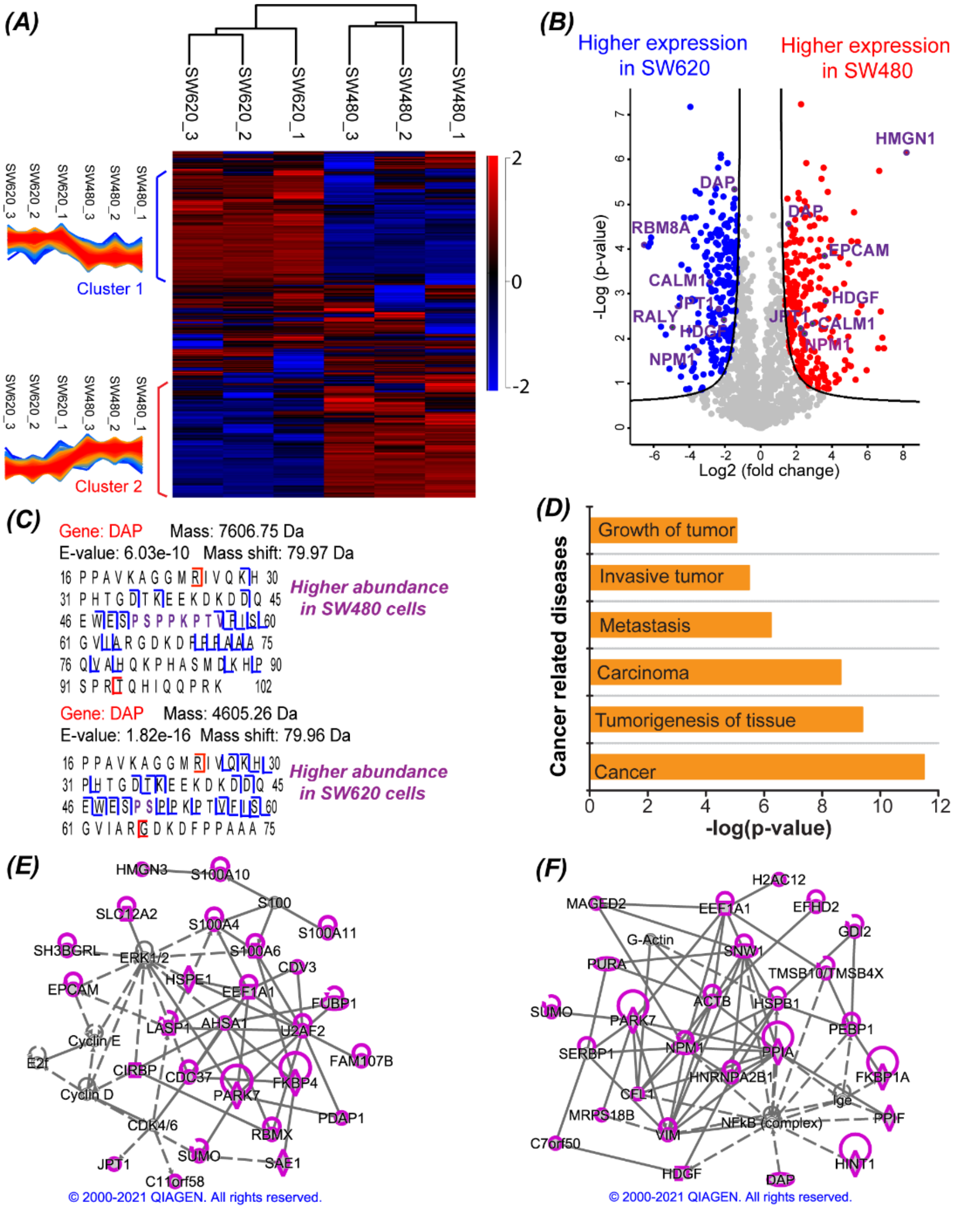
(A)
SW620_3
SW620_2
SW620_1
SW480_3
SW480_2
SW480_1
2
0
-2
Cluster 1
Cluster 2
(B)
Higher expression in SW620
Higher expression in SW480
-Log (p-value)
Log2 (fold change)
HMGN1
DAP
RBM8A
EPCAM
CALM1
HDGF
JPT1
RALY
NPM1
(C)
Gene: DAP Mass: 7606.75 Da
E-value: 6.03e-10 Mass shift: 79.97 Da
16 P P A V K A G G M R I V Q K H 30
31 P H T G D T K E E K D K D D Q 45
46 E W E S P S P P K P T V F L S 60
61 G V L A R G D K D F P P A A A 75
76 Q L V A H Q K P H A S M D K H P 90
91 S P R T Q H I Q Q P R K 102
Higher abundance in SW480 cells
Gene: DAP Mass: 4605.26 Da
E-value: 1.82e-16 Mass shift: 79.96 Da
16 P P A V K A G G M R I V Q K H 30
31 P H T G D T K E E K D K D D Q 45
46 E W E S P S P P K P T V F L S 60
61 G V I A R G D K D F P P A A A 75
Higher abundance in SW620 cells
(D)
Cancer related diseases
Growth of tumor
Invasive tumor
Metastasis
Carcinoma
Tumorigenesis of tissue
Cancer
-log(p-value)
(E)
HMGN3
S100A10
S100
SLC12A2
S100A4
S100A11
SH3BGRL
ERK1/2
S100A6
CDV3
EPCAM
HSPE1
EEF1A1
FUBP1
Cyclin E
LASP1
AHSA1
U2AF2
E2f
FAM107B
CIRBP
CDC37
FKBP4
Cyclin D
PARK7
PDAP1
CDK4/6
RBMX
JPT1
SUMO
SAE1
C11orf58
© 2000-2021 QIAGEN. All rights reserved.
(F)
H2AC12
MAGED2
EEF1A1
EFHD2
G-Actin
GDI2
SNW1
PURA
TMSB10/TMSB4X
SUMO
ACTB
HSPB1
PARK7
PEBP1
SERBP1
NPM1
PPIA
FKBP1A
HNRNPA2B1
CFL1
Ige
MRPS18B
NFkB (complex)
PPIF
C7orf50
VIM
HINT1
HDGF
DAP
© 2000-2021 QIAGEN. All rights reserved.

Figure 5