Published online 11 August 2020

doi: 10.1093/narcan/zcaa015

Pan-cancer proteogenomic analysis reveals long and circular noncoding RNAs encoding peptides

Ghofran Othoum[1,2], Emily Coonrod[1,2], Sidi Zhao[1,2], Ha X. Dang[1,2,3] and Christopher A. Maher[1,2,3,4,*]

[1]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO 63108, USA, [2]Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63108, USA, [3]Alvin J. Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63108, USA and [4]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO 63108, USA

Received May 06, 2020; Revised July 13, 2020; Editorial Decision July 23, 2020; Accepted July 24, 2020

ABSTRACT

Recent studies show that annotated long noncoding RNAs (lncRNAs) and circular RNAs (circRNAs) encode for stable, functional peptides that contribute to human development and disease. To systematically discover lncRNAs and circRNAs encoding peptides, we performed a comprehensive integrative analysis of mass spectrometry-based proteomic and transcriptomic sequencing data from >900 patients across nine cancer types. This enabled us to identify 19,871 novel peptides derived from 8,903 lncRNAs. Further, we exploited open reading frames overlapping the backspliced region of circRNAs to identify 3,238 peptides that are uniquely derived from 2,834 circRNAs and not their corresponding linear RNAs. Collectively, our pan-cancer proteogenomic analysis will serve as a resource for evaluating the coding potential of lncRNAs and circRNAs that could aid future mechanistic studies exploring their function in cancer.

INTRODUCTION

Long noncoding RNAs (lncRNAs) are a heterogeneous class of RNA molecules having >200 nucleotides with diverse regulatory mechanisms (1,2). A subset of lncRNAs has established oncogenic or tumor-suppressive roles in cancer (e.g. *HOTAIR*, *MALAT1* and *NEAT1*) (3–5) and has clinical utility as prognostic and predictive biomarkers (6,7). To date, annotating a lncRNA has relied on sequence-based features such as the presence and length of an open reading frame (ORF), as well as similarity to known proteins (8), to establish whether it lacks coding potential (1). However, an increasing number of studies are demonstrating functional roles for micropeptides encoded by short ORFs (sORFs) in well-characterized lncRNAs such as *HOXB-AS3* in colon cancer (9), *LINC00961* in muscle development (10) and the circular form of *LINC-PINT* in glioblastoma (11).

Circular RNAs (circRNAs) are RNA molecules of covalent continuous loops formed through noncanonical splicing methods such as backsplicing (12). Although circRNAs are abundantly present in the human transcriptome, their functions remain understudied (13). To date, one of the prevailing models is that circRNAs act as microRNA sponges (14). However, the discovery of novel functional peptides encoded by circRNAs in human development and disease offers an alternative mechanism of function. This can be exemplified by functional peptides resulting from the circular ORF (cORF), which can span the backsplice junction. Recent examples include a backspliced junction cORF between exons 3 and 4 of *FBXW7* in glioma (15), an 87-aa (amino acid)-long peptide not shared with linear *LINC-PINT* in glioblastoma (11), a backspliced junction cORF extending two *ZNF609* exons in myogenesis (16), and a backspliced junction cORF between exons 24 and 25 of *PPP1R12A* in colon cancer (17).

Despite numerous studies revealing the presence of translatable sORFs and cORFs that have a role in human development and disease (9–11,15–21), the computational tools used to predict whether a lncRNA or a circRNA can potentially translate into a protein have difficulty detecting sORFs (<100 codons) or unique cORFs that span backsplice junctions. The predominant experimental strategy for systematically identifying annotated noncoding RNAs that may encode small peptides is high-throughput sequencing of ribosome-protected fragments (22,23). However, this approach is limited since ribosome occupancy does not necessarily confirm active translation but may be a mere indicator of translation initiation (24). To address this gap, we discovered novel peptides encoded by lncRNAs and circRNAs through proteogenomic integration of 921 patients across nine cancer types using mass spectrometry-based proteomic data (MS/MS) from the Clinical Proteomic Tu-

[*]To whom correspondence should be addressed. Tel: +1 314 286 2856; Email: christophermaher@wustl.edu

mor Analysis Consortium (CPTAC) (25–32), with transcriptome sequencing data from The Cancer Genome Atlas (TCGA) and publicly available datasets (33). The identified peptides are available at the PepTransDB resource (Peptide Encoding Transcripts Database, https://www.maherlab.com/peptransdb).

MATERIALS AND METHODS

Data integration and proteogenomic search

To construct the database for the proteogenomic search, we used three-frame translation of annotated transcripts of lncRNAs from LNCipedia (34) and circRNAs from the pan-cancer circRNA compendium MiOncoCirc (35). Nonredundant translated ORFs from these transcripts and protein sequences from UniProt were included in the database (36). LncRNAs that match UniProt entries assigned as PE5 (uncertain) were filtered out to avoid redundancy in the database. Additionally, those overlapping protein-coding genes were kept as long as the predicted ORFs were not part of the coding sequence of the protein-coding gene. To allow the use of false discovery rate (FDR) filtering of identified peptide spectrum matches (PSMs), the target-decoy strategy was utilized, and the search database was appended with an equal number of decoy reversed sequences. Raw mass spectrometry files were downloaded from the CPTAC data portal (https://cptac-data-portal.georgetown.edu/datasets) for each respective cohort in mzML format. MS/MS spectra search was performed using the MSFragger search engine (37). We used semi-tryptic peptides with two allowed missed cleavage sites, precursor-ion mass tolerance of 20 ppm and allowed $^{12}C/^{13}C$ isotope errors. To identify putative sORFs and cORFs, we used sequences between start codons (AUG, CUG, UUG) and stop codons (UAG, UGA, UAA) in each of the forward translated frames with a minimum length of 100 nucleotides for lncRNAs and 200 nucleotides for circRNAs. The selected minimum threshold for ORF lengths was based on the lengths of ORFs in lncRNAs and circRNAs known to encode peptides.

For each cohort, the following post-translational modifications were specified depending on the protocol used for protein labeling: for TMT labeling, cysteine carbamidomethylation (+57.0215) and lysine TMT labeling (+229.1629) were specified as fixed modifications, while methionine oxidation (+15.9949), N-terminal protein acetylation (+42.0106) and TMT labeling of peptide N terminus and serine residues were specified as variable modifications; for iTRAQ labeling, cysteine carbamidomethylation (+57.0215), iTRAQ labeling of lysine (+144.10253) and peptide N terminus were specified as fixed modifications, while methionine oxidation (+15.9949) was specified as a variable modification; for label-free protocol, cysteine carbamidomethylation (+57.0215) was specified as a fixed modification, while methionine oxidation (+15.9949) was specified as a variable modification.

PSMs were processed and filtered using PeptideProphet as implemented in the Philosopher pipeline (https://github.com/Nesvilab/philosopher) (38) using high-mass accuracy binning and semi-parametric mixture modeling to assign posterior probabilities and to filter out PSMs found by chance. PeptideProphet results from each sample were then used to infer high-confidence protein groups for each cohort (protein inference), accounting for both unique and razor peptides. PSMs and proteins that did not meet the 1% FDR threshold were not included.

Sources used to identify high-confidence PSMs from lncRNA transcripts

Basewise PhastCons 100-way conservation scores for sORF coordinates were extracted from http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.phastCons100way.bw from the UCSC human genome database (39). Similarly, phylogenetic codon substitution frequency (PhyloCSF) scores for each frame were extracted from https://data.broadinstitute.org/compbio1/PhyloCSFtracks/hg38/latest/PhyloCSF+1.bw, https://data.broadinstitute.org/compbio1/PhyloCSFtracks/hg38/latest/PhyloCSF+2.bw, https://data.broadinstitute.org/compbio1/PhyloCSFtracks/hg38/latest/PhyloCSF+3.bw, https://data.broadinstitute.org/compbio1/PhyloCSFtracks/hg38/latest/PhyloCSF-1.bw, https://data.broadinstitute.org/compbio1/PhyloCSFtracks/hg38/latest/PhyloCSF-2.bw, https://data.broadinstitute.org/compbio1/PhyloCSFtracks/hg38/latest/PhyloCSF-3.bw, for frames 1,2,3,4,5,6 respectively. CPC2 (40) and CNIT (41) were used to predict the coding probability score for each transcript from LNCipedia (34).

LncRNA differential expression analysis

TCGA RNA-seq pre-aligned bam files were downloaded from the Cancer Genomics Hub (http://cghub.ucsc.edu/). To focus on lncRNAs deregulated in tumors, we performed differential expression analysis to compare gene expression between tumor and normal tissues using edgeR with the negative binomial model (42). Differentially expressed lncRNAs were identified as those with $\log_2$ fold change (FC) beyond ±2 in at least one of the following TCGA cohorts that have matched normal samples: HNSC (head–neck squamous cell carcinoma), LUAD (lung adenocarcinoma), KIRC (kidney renal clear cell carcinoma), LIHC (liver hepatocellular carcinoma), PRAD (prostate adenocarcinoma), BRCA (breast cancer), CRC (colorectal cancer) and STAD (stomach adenocarcinoma).

CircRNA mRNA sequence extraction and junction peptide identification

Using publicly available cDNA-Capture sequencing data from the pan-cancer circRNA compendium MiOncoCirc (35), terminal exon coordinates of all annotated circRNAs were used to extract the FASTA sequences using bedtools getfasta (43). Specifically, to extract the backsplice junction sequences, the mRNA sequence of the 3′ exon was concatenated to the mRNA sequence of the 5′ exon. For single-exon circRNAs, the exon mRNA sequence was repeated in order to cover the backsplice junction without truncating the exon. Peptides matching a UniProt coding sequence or any other ORF from the linear template were filtered out to identify peptides spanning backsplice junctions.

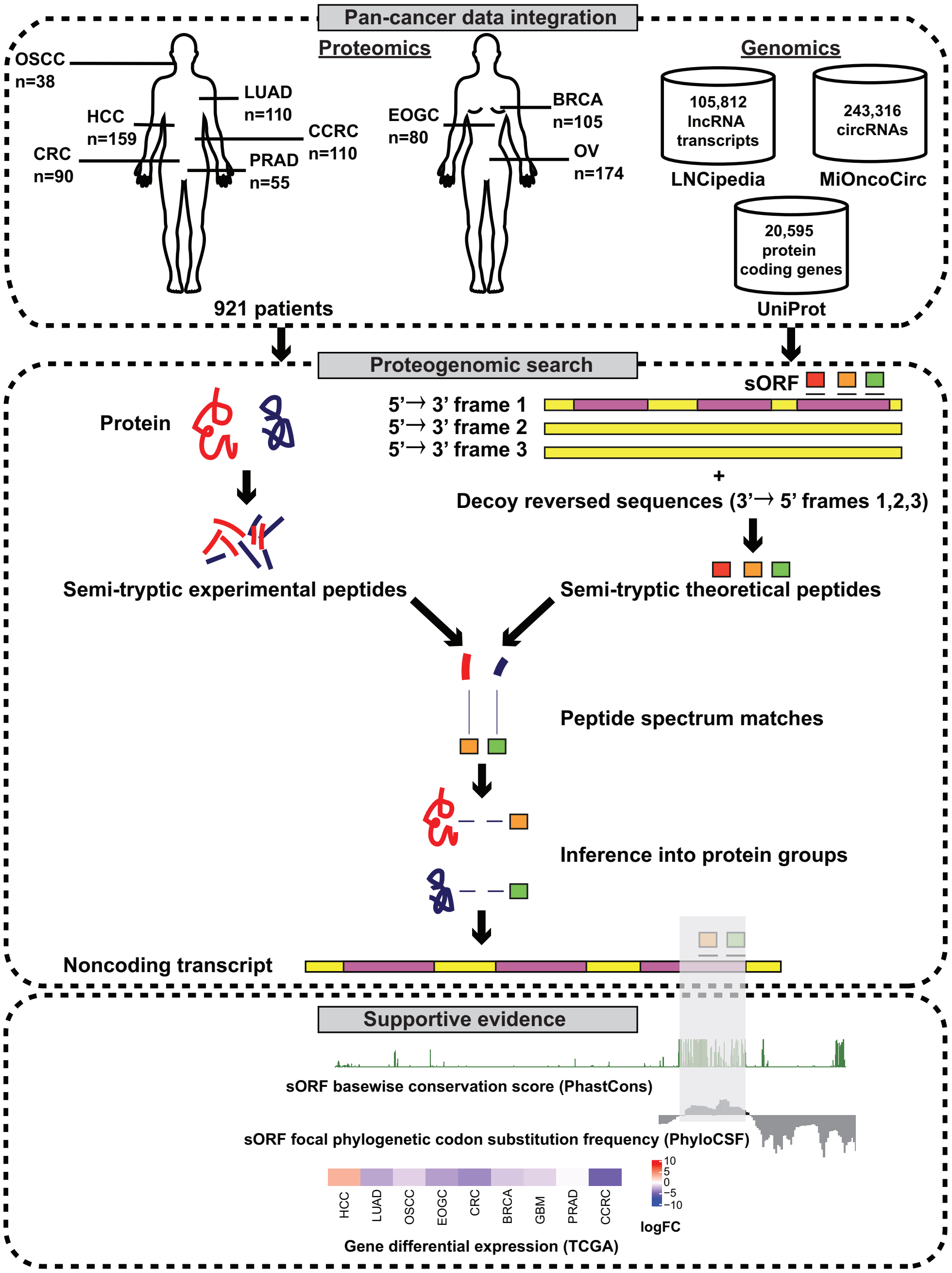
Pan-cancer data integration
Proteomics
Genomics
OSCC n=38
HCC n=159
CRC n=90
LUAD n=110
CCRC n=110
PRAD n=55
EOGC n=80
BRCA n=105
OV n=174
921 patients
105,812 lncRNA transcripts
LNCipedia
243,316 circRNAs
MiOncoCirc
20,595 protein coding genes
UniProt
Proteogenomic search
Protein
sORF
5'→ 3' frame 1
5'→ 3' frame 2
5'→ 3' frame 3
+
Decoy reversed sequences (3'→ 5' frames 1,2,3)
Semi-tryptic experimental peptides
Semi-tryptic theoretical peptides
Peptide spectrum matches
Inference into protein groups
Noncoding transcript
Supportive evidence
sORF basewise conservation score (PhastCons)
sORF focal phylogenetic codon substitution frequency (PhyloCSF)
HCC
LUAD
OSCC
EOGC
CRC
BRCA
GBM
PRAD
CCRC
Gene differential expression (TCGA)
10
5
0
−5
−10
logFC

Figure 1. Overview of lncRNAs with peptide support in predicted sORFs. Description of the workflow used to identify novel peptides across the following nine cancer cohorts: oral squamous cell carcinoma (OSCC), hepatocellular carcinoma (HCC), colorectal cancer (CRC), lung adenocarcinoma (LUAD), clear cell renal carcinoma (CCRC), prostate adenocarcinoma (PRAD), early-onset gastric cancer (EOGC), ovarian cancer (OV) and breast cancer (BRCA).

RESULTS

An integrated proteogenomic database for long and circular noncoding RNAs across nine cancer types

In order to systematically identify peptides encoded by noncoding transcripts in cancer patients, we used a bottom-up approach utilizing a comprehensive database of ORFs in lncRNA and circRNA transcripts. To ensure we correctly identified peptides uniquely mapping to lncRNAs and circRNAs, we included 20,595 protein sequences from UniProt (36). To reduce the database size while accurately estimating the FDR, we used annotated transcripts to predict ORFs in three-frame translated sequences, instead of using six-frame translation of the genomic sequences, as recommended previously (44). The database encompassed 620,578 canonical and noncanonical sORFs that were at least 100 nucleotides long from 105,812 lncRNA transcripts (54,150 genes) annotated in LNCipedia (34). We also included 130,526 ORFs from 243,316 circRNAs in MiOncoCirc (35) that were at least 200 nucleotides. The proteogenomic search was customized according to the parameters described to generate the proteomic data from each co-

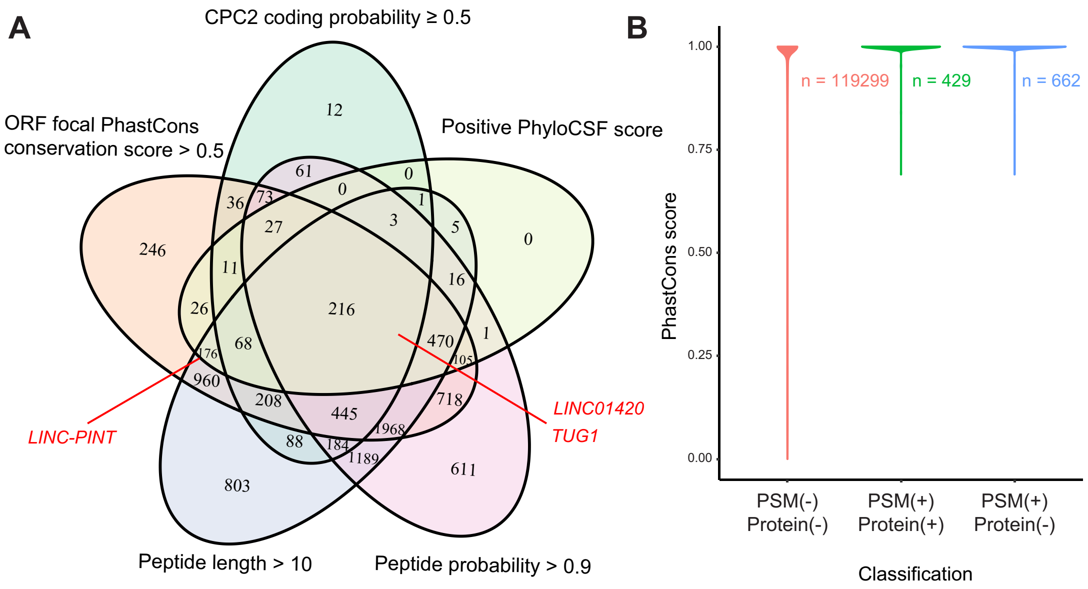

Figure 2. Consensus set of high-confidence peptide-encoding lncRNAs. **(A)** Venn diagram showing the overlap of genes with PSMs that have a maximum focal PhastCons conservation score for the encoding sORF >0.5, a peptide length >10, a positive PhyloCSF score, a CPC2 coding index ≥0.5 and a peptide probability score >0.9. LncRNAs previously identified to encode for peptides are shown in the Venn diagram. **(B)** Violin plot showing the distribution of PhastCons conservation scores of lncRNA transcripts with a high coding index and PhyloCSF scores that have peptide support or are inferred into protein groups, along with lncRNAs that lack peptide support.

hort, accounting for different post-translational modifications (Figure 1, Supplementary Table S1).

LncRNAs encoding peptides and sequence conservation

As an initial step, we identified all lncRNAs with at least one PSM in a predicted sORF independent of its relevance in cancer. The proteogenomic search allowed us to identify 10,589 lncRNA transcripts with PSM support (10% of all transcripts) corresponding to 8,903 genes (16.4% of lncRNAs) with a total of 19,871 peptides (Supplementary Tables S2 and S3).

Given the possibility that the identified peptides could include spurious matches due to the size of the proteogenomic database, we integrated additional evidence for translation from supporting datasets. Specifically, we required sORFs to be highly conserved (PhastCons score >0.5) and have a high PhyloCSF (45) score of the encoding sORF in three translated frames (>0). Additionally, only sORFs that produced peptides longer than 10 amino acids were retained. We finally considered the coding potential predicted using intrinsic sequence features, requiring a CPC2 coding index ≥0.5 (40). These filters led to a reduced set of 216 lncRNAs that were supported by proteomic samples as well as features related to the predicted sORF. The set encompassed previously well-characterized lncRNAs, such as *TUG1* in prostate cancer (19) and *LINC01420* in nasopharyngeal carcinoma (46) (Figure 2A). *LINC-PINT*, previously identified to encode a peptide in its circular form in glioblastoma (11), met all of the criteria except for the peptide probability and the coding index thresholds.

To evaluate whether our peptide-supported sORFs had greater sequence conservation, we compared PhastCons conservation scores between lncRNAs grouped by their level of experimental support. Upon considering all PSM and protein-supported lncRNAs, we found that lncRNAs supported by PSMs had greater sequence conservation compared to lncRNAs that lacked any peptide support (Supplementary Figure S1). Including the coding index and PhyloCSF filters confirmed greater conservation of PSM and protein-supported lncRNAs compared to the remaining unsupported lncRNAs (rank-sum test $P < 2.2e{-}16$) (Figure 2B).

Using protein inference to infer peptides into protein groups significantly reduced the number of candidates to 5,756 genes, 74 passing the aforementioned filters (conservation, PhyloCSF score, peptide length and coding index). Interestingly, among the top 50 lncRNAs inferred as protein groups was an *H19* isoform (LNCipedia ID: H19:11) that had a high coding potential and was supported by a 20-aa peptide and a 42-aa sORF in the BRCA and OSCC cohorts (Figure 3, Supplementary Figure S2). Although sORFs from lncRNAs inferred as proteins exhibit the highest conservation, exclusively considering lncRNAs that were inferred as proteins as the only MS/MS-supported transcripts is a limitation to the true landscape of lncRNAs encoding small proteins. For instance, considering the conservation and PhyloCSF scores of PSMs allowed the detection of peptide-encoding lncRNAs such as *lnc-EVX2-8* located at chr2:176200908–176201252 and *LINC00431* located at chr13:110983307–110990564. Both of these lncRNAs were not inferred as proteins, but their encoding sORFs had enriched PhyloCSF and conservation scores (Figure 4). The enriched conservation of the encoding sORF is also observed in known micropeptide-encoding genes, such as that encoding a peptide in *LINC01420* (18) (Supplementary Figure S3).

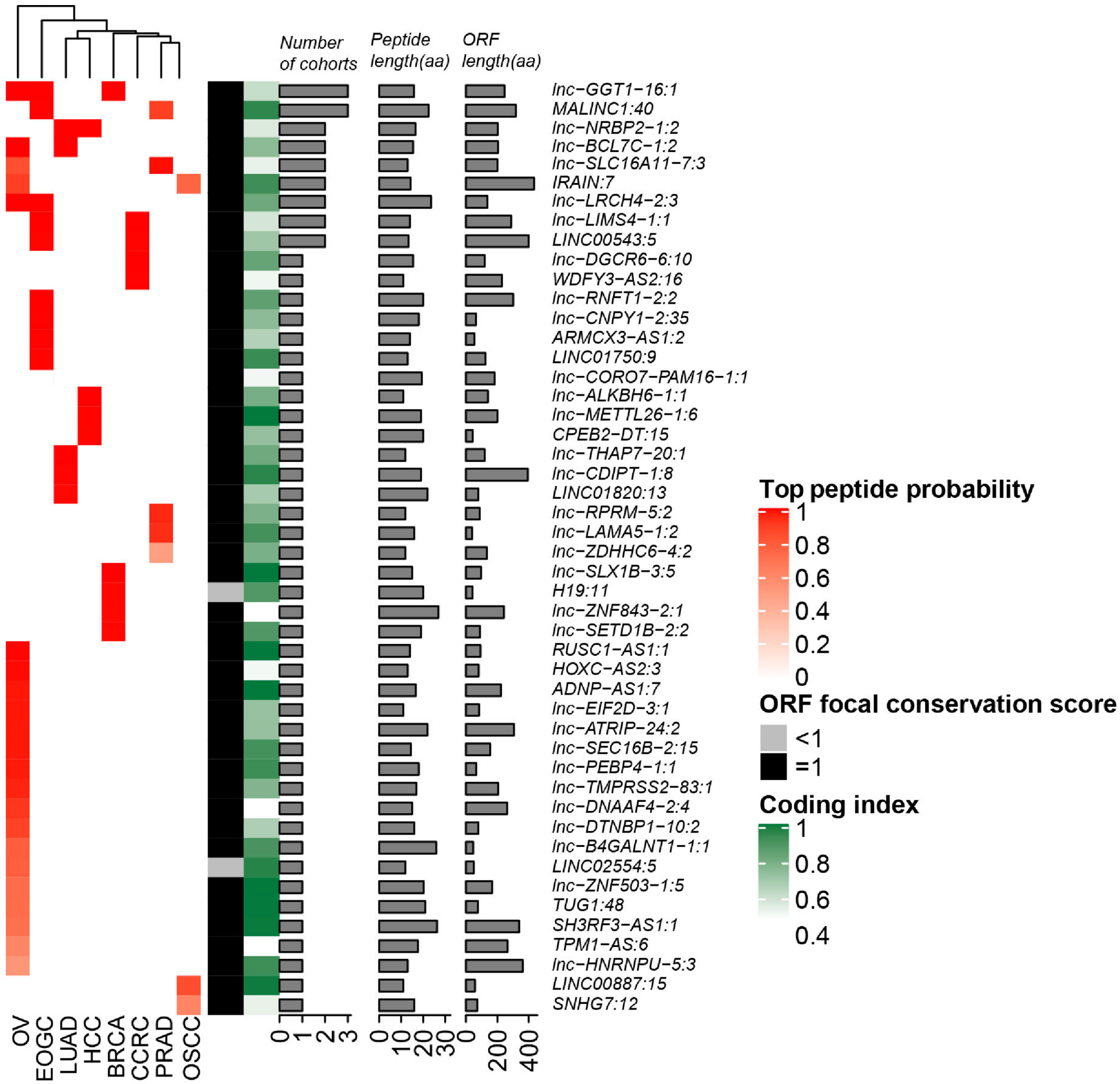

Figure 3. Average peptide probability of PSMs identified in transcripts inferred as proteins. From left to right, average peptide probability for the top peptide of the protein group, maximum focal 100-way PhastCons conservation score for encoding sORF, coding probability predicted using CPC2, frequency of each transcript in the nine cancer cohorts, length of top peptide for the protein group and length of the encoding sORF. Peptides are sorted based on their frequency in the cohorts.

LncRNAs with peptide support that are differentially expressed in cancer

To further implicate and prioritize noncoding RNAs encoding small peptides in cancer, we leveraged transcriptome data to identify differentially expressed lncRNAs. While *SPAR* (small regulatory polypeptide of amino acid response, encoded by *LINC00961*) was initially found to encode a functional peptide in muscle development (10), it was downregulated in BRCA and LUAD (log FC = −2.2 and −2.6, respectively). Further, concordant with previous reports implicating *HOXB-AS3* in suppressing colon cancer growth (9), its expression was downregulated in colon cancer (log FC = −2.6). Overall, of the genes showing high sequence conservation, a positive PhyloCSF score and high coding probability and with a long peptide (>10) in at least one cancer, 24 were differentially expressed in at least one of the nine cancers (log FC beyond ±2) (Figure 5).

Proteomic samples are enriched with peptides from cORFs extending junction exons

To identify ORFs emerging from circRNAs, we utilized a similar proteogenomic search approach for circRNAs utilizing the pan-cancer circRNA compendium MiOncoCirc (35). In total, there were 2,834 circRNAs with 3,238 peptides in at least one cancer type (Supplementary Table S4). To discover peptides unique to circRNAs, we filtered peptides that were shared with either the noncanonical ORFs (Figure 6A; type 2) or the coding sequence of the linear

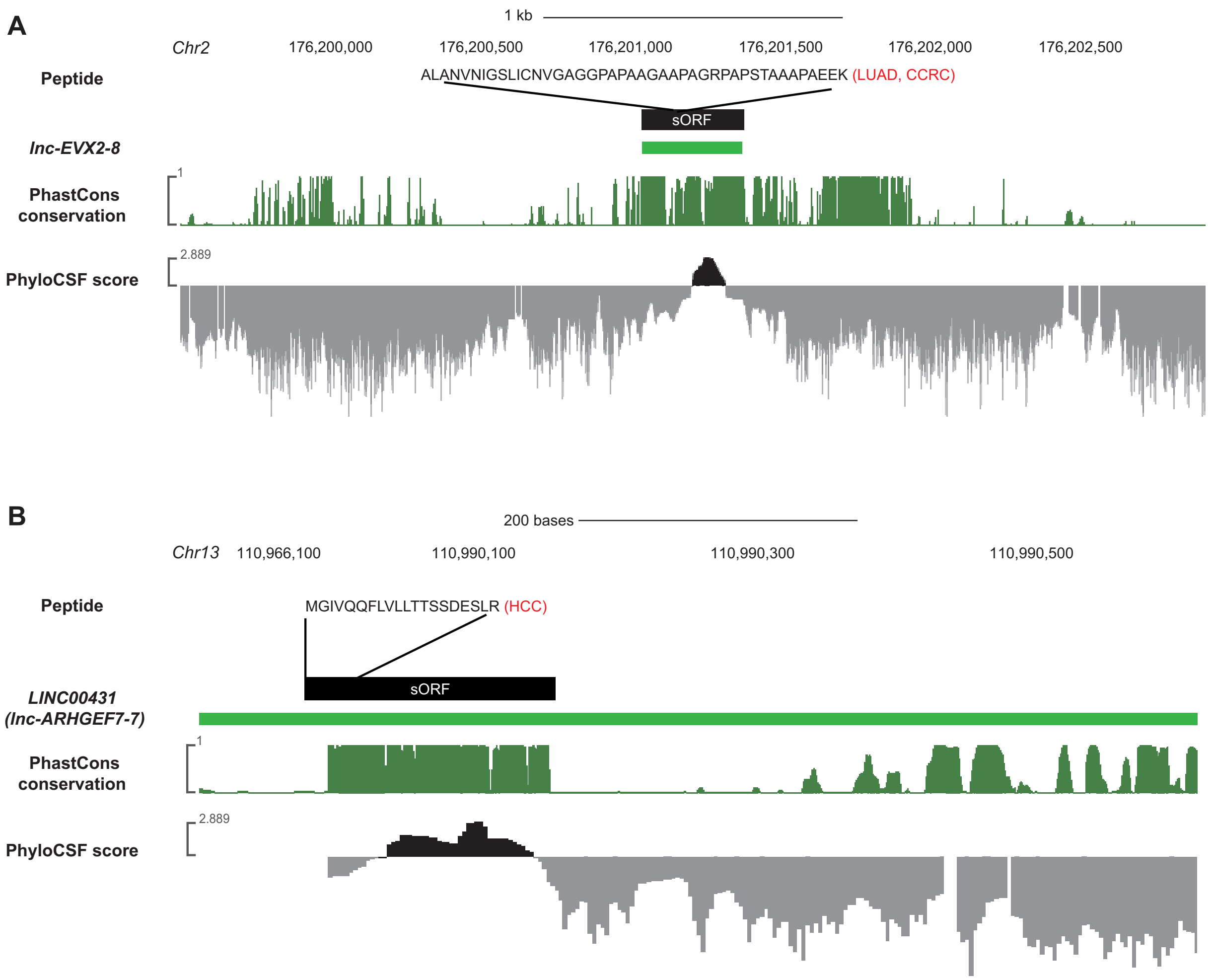

Figure 4. Focal conservation of peptide-supported sORF used to evaluate coding potential. UCSC browser views with 100-way PhastCons and PhyloCSF tracks showing peptides in **(A)** *lnc-EVX2-8* in LUAD and CCRC and **(B)** *LINC00431* (*lnc-ARHGEF7-7*) in HCC as exemplary cases of high focal conservation of the encoding sORF.

transcript (Figure 6A; type 1). This enabled us to focus on peptides that emerge either from a predicted cORF that spans the backsplice junction between two exons or from the circularization of a single exon (Figure 6A; type 3). This revealed 2,010 peptides spanning backsplice junctions in 1,964 circRNAs (Supplementary Table S5). To further prioritize peptides with strong support for both exons flanking the backsplice junction, we quantified the number of amino acids from each of the junction exons. Collectively, we found a total of 774 junction peptides where at least a third of the amino acids from the peptide map to each side of the junction.

In order to identify peptides covering most of the predicted cORF, we ranked the remaining peptides in descending order of length from each cancer, as well as the longest peptides shared between at least two cancers (a minimum of two cancers and a maximum of seven cancers) (Figure 6B). Several circRNAs showed exclusive or higher expression in the cohort at which their peptides were identified. For instance, *circCUBN*, supported by a 42-aa peptide in PRAD with 47.6% of the amino acids from the 3′ exon and 52.4% from the 5′ exon, was only expressed in prostate cancer (average of two mapped backsplice reads) but not expressed in any of the remaining cancer types (Figure 7A). Similarly, *circCOG5*, with a 7-aa peptide resulting from the circularization of exon 13 from the linear *COG5* template, was identified in seven cancers (BRCA, OSCC, EOGC, PRAD, LUAD, HCC, CCRC) except for ovarian cancer, a cohort that lacked peptide support for this circRNA (Figure 7B).

DISCUSSION

Most analysis of proteomic samples from cancer patients focused on the discovery of novel peptides from protein-coding genes, overlooking novel peptides from noncoding RNAs with potential roles in cancer. The robustness of our methods can be exemplified by our ability to confirm lncRNAs previously shown to encode functional proteins [a known sORF in *LINC01420* (18) and a novel sORF in *TUG1* (19)]. However, our analysis also identified many uncharacterized lncRNAs that are unique in each cancer cohort and potentially contribute to tumorigenesis via an encoded small protein, with ovarian cancer having the highest

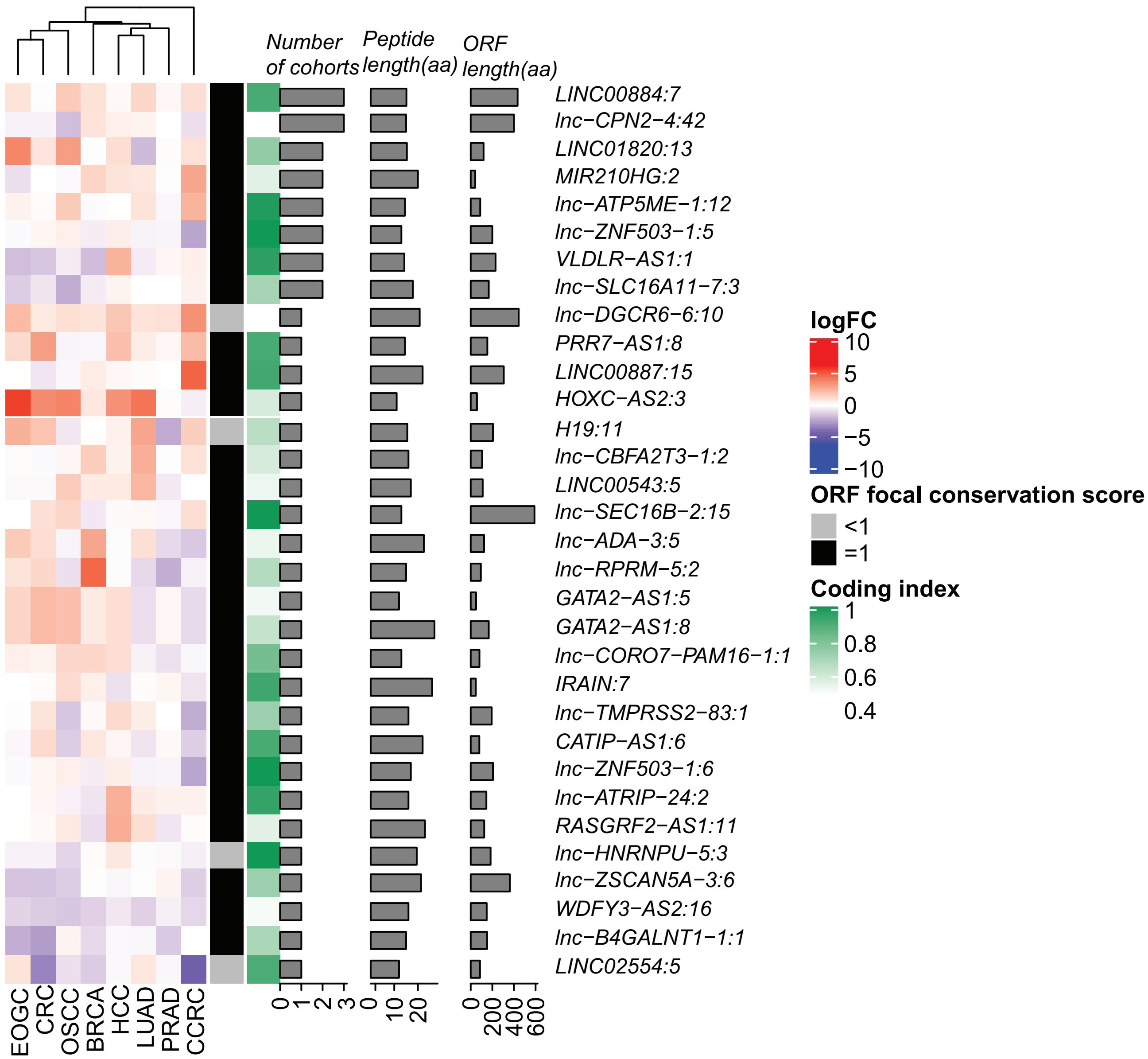

Figure 5. Conserved peptide-encoding RNAs that are differentially expressed in sampled cohorts. Heat map showing differentially expressed candidates. From left to right, log FC in the expression of the noncoding gene in primary relative to normal, maximum focal 100-way PhastCons conservation score for encoding sORF, coding probability, frequency of each transcript in the nine cancer cohorts, length of identified peptide for transcript and length of the encoding sORF. Peptides are sorted based on their frequency in the cohorts.

number of unique peptides (52.1% of all uniquely identified) and colorectal cancer having the lowest (1.5%). Examples of top-scoring lncRNAs (conservation score >0.9) that have uniquely identified peptides in each subtype include *lnc-HJURP-6* in Luminal B samples (117-aa-long ORF), *lnc-ZNF627-2* in HER2-enriched samples (444-aa-long ORF), *lnc-CIAO1-6* in Lumina A samples (118-aa-long ORF) and *AATBC* in Basal-like samples (52-aa-long ORF). There are also peptides uniquely identified in each cohort such as *lnc-PPME1-4* in PRAD (298-aa-long ORF), *lnc-PROC-5* in CCRC (57-aa-long ORF), *lnc-IL20RA-4* in LUAD (165-aa-long ORF), *lnc-BOLA2-2* in OSCC (294-aa-long ORF), *LINC02370* in OV (189-aa-long ORF), *lnc-NBPF14-1* in HCC (172-aa-long ORF), *lnc-ZNF716-2* in EOGC (251-aa-long ORF) and *LINC01579* in CRC (35-aa-long ORF). We did not observe an overall difference in the quantity of peptides between the four subtypes of breast cancer (1,220 peptides for Basal-like samples, 1,409 peptides for HER2-enriched samples, 1,419 peptides for Luminal A samples and 1,637 for BRCA-Luminal B samples). By performing a pan-cancer analysis, we were able to identify lncRNAs with peptide support across all nine cancer types (i.e. *lnc-PRSS1-2:1* with a 60-aa-long noncanonical ORF and *lnc-NUDCD2-8:1* with a 91-aa-long noncanonical ORF). The ability to reliably detect a small encoded peptide across cancer types further emphasizes their potential mechanistic role and potential contribution to tumorigenesis. It is also plausible that these broadly expressed noncoding RNAs en-

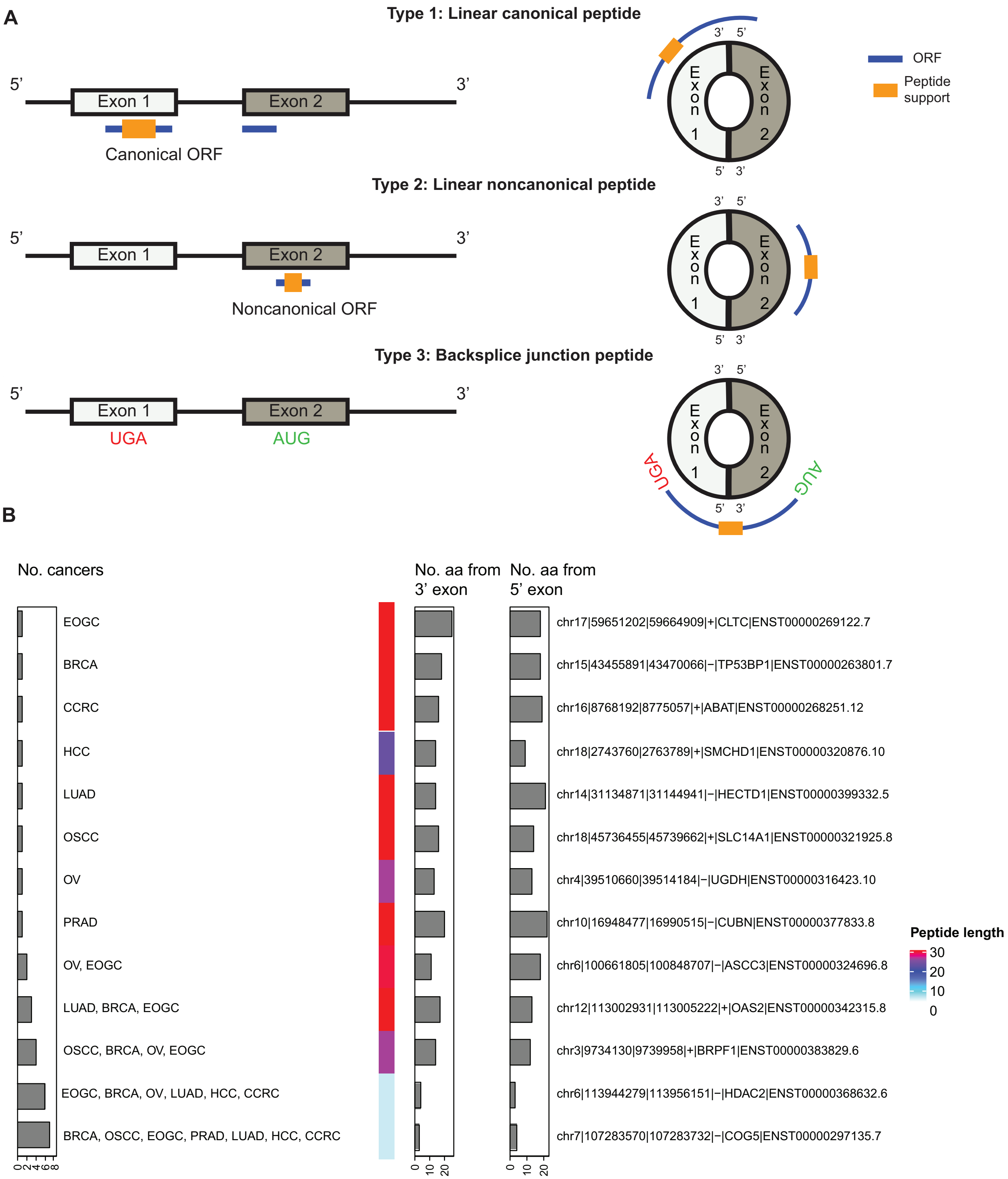

Figure 6. PSMs from cORFs spanning splice junction exons. **(A)** Classification of circRNAs: type 1, peptides that are shared with the coding sequence of the linear template of the gene; type 2, peptides in a noncanonical ORF in the linear template of the gene; and type 3, peptides that are unique to the backsplice junction sequence. **(B)** Longest backsplice junction peptides that have at least third of their lengths in the 3′ or 5′ exons. The longest peptide for each cohort and for different cohort combinations was selected, yielding a total of 13 peptides. From left to right, cancers in which the peptide was identified, peptide length, number of amino acids from 3′ exon and number of amino acids from 5′ exon.

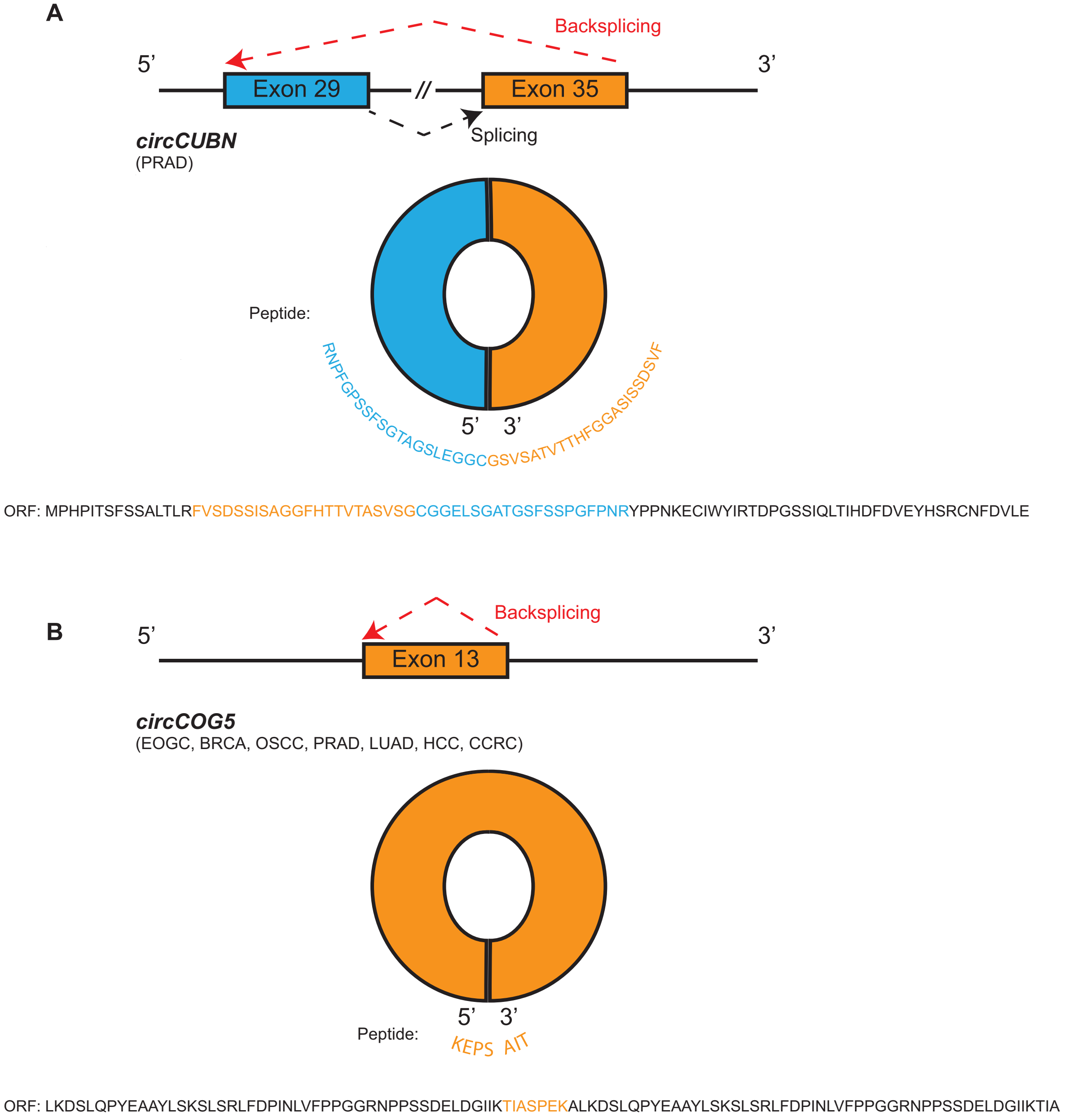
A
Backsplicing
5’
3’
Exon 29
//
Exon 35
Splicing
circCUBN
(PRAD)
Peptide:
RNPFGPSSFSGTAGSLEGGCGSVSATVTTHFGGASISSDSVF
5’
3’
ORF: MPHPITSFSSALTLRFVSDSSISAGGFHTTVTASVSGCGGELSGATGSFSSPGFPNRYPPNKECIWYIRTDPGSSIQLTIHDFDVEYHSRCNFDVLE
B
Backsplicing
5’
3’
Exon 13
circCOG5
(EOGC, BRCA, OSCC, PRAD, LUAD, HCC, CCRC)
5’
3’
Peptide:
KEPS AIT
ORF: LKDSLQPYEAAYLSKSLSRLFDPINLVFPPGGRNPPSSDELDGIIKTIASPEKALKDSLQPYEAAYLSKSLSRLFDPINLVFPPGGRNPPSSDELDGIIKTIA

Figure 7. Exemplary cases of cORFs extending splice junction regions. **(A)** *CircCUBN* has a 42-aa peptide in a 97-aa cORF identified in prostate cancer with 47.6% of the peptide from the 3′ exon and 52.4% from the 5′ exon. **(B)** *circCOG5* has a 7-aa peptide in a 103-aa cORF in seven cancers (EOGC, BRCA, OSCC, PRAD, LUAD, HCC, CCRC).

coding small proteins could have a similar regulatory role in additional cancer types beyond those in our study.

By using multiple conservative filters, we were able to nominate high-confidence noncoding RNAs that may encode small proteins. Notably, we also identified lncRNAs that were predicted to have sORFs, display sequence conservation and could encode for peptides but were not captured by the proteomic data. It is possible that the peptides were not sampled in the existing proteomic data, that these are tumor suppressors that have lost expression and are therefore under-represented or these noncoding RNAs function through small encoded proteins in a completely different context unrelated to cancer. As a result of applying stringent criteria, it is possible we are still under-representing the quantity of noncoding RNAs encoding small peptides. For instance, *PCAT14* (*Prostate Cancer Associated Transcript 14*), which was previously reported in prostate cancer patients (6) to encode a peptide corresponding to a previously reported HERV-K gag ORF (47), did not meet the PhyloCSF or the conservation filters but was supported by a high-confidence peptide. This highlights the importance of peptide support as evidence for the translation potential of a transcript compared to conservation score or coding index alone.

Our ability to identify encoded proteins within circRNAs has additional challenges relative to lncRNAs due to the shared sequence between the circRNA and linear transcript (Figure 6A; types 1 and 2). While it is not possible to discriminate whether a small protein encoded within these common sequences originated from the circRNA or linear transcript (or both), this can be addressed experimentally (11). Notably, we identified 3,238 peptides derived from 2,834 circRNAs corresponding to a noncanonical ORF (Figure 6A; type 2). Since these are not the canonical proteins, these candidates represent a rich resource of small proteins that may be functionally relevant and warrant further investigation independent of whether they originated from the circular or linear transcript. However, in our integrative analysis we focused on the peptides that correspond to the only region that is unique to the circRNA, the backspliced junction. Doing so allowed the identification of cancer-relevant junction peptides from circRNAs, such as *circCOG5* in seven cancers and *circHDAC2* in six cancers.

Despite recent findings showing lncRNAs and circRNAs functioning through encoded proteins, the detection of a peptide does not imply function. Therefore, we envision our results will serve as a resource for the community to confirm whether an lncRNA or a circRNA has any experimental evidence supporting its coding potential that may guide subsequent mechanistic studies. Further, as part of our analysis we revealed many novel lncRNAs that have peptide support but did not appear to be differentially expressed in the cancer types sampled. These findings have implications beyond cancer by revealing lncRNAs that may encode small proteins that function in human development and other diseases.

DATA AVAILABILITY

The identified peptides are available at the PepTransDB (https://www.maherlab.com/peptransdb). The proteogenomic pipeline and related scripts are available at GitHub (https://github.com/ChrisMaherLab/PepTransDB).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Cancer Online.

ACKNOWLEDGEMENTS

Part of the data used in this publication was generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium.

FUNDING

American Cancer Society Research Scholar Grant [to C.A.M.]; National Institutes of Health, National Cancer Institute [R01 CA203995 to C.A.M.]; National Institutes of Health, Molecular Oncology Training Grant [T32 CA113275-10 to S.Z.]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Batista,P.J. and Chang,H.Y. (2013) Long noncoding RNAs: cellular address codes in development and disease. *Cell*, **152**, 1298–1307.
2. Ransohoff,J.D., Wei,Y. and Khavari,P.A. (2018) The functions and unique features of long intergenic non-coding RNA. *Nat. Rev. Mol. Cell Biol.*, **19**, 143–157.
3. Huarte,M. (2015) The emerging role of lncRNAs in cancer. *Nat. Med.*, **21**, 1253–1261.
4. Silva-Fisher,J.M., Dang,H.X., White,N.M., Strand,M.S., Krasnick,B.A., Rozycki,E.B., Jeffers,G.G.L., Grossman,J.G., Highkin,M.K., Tang,C. *et al.* (2020) Long non-coding RNA RAMS11 promotes metastatic colorectal cancer progression. *Nat. Commun.*, **11**, 2156.
5. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.-C., Hung,T., Argani,P., Rinn,J.L. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
6. White,N.M., Zhao,S.G., Zhang,J., Rozycki,E.B., Dang,H.X., McFadden,S.D., Eteleeb,A.M., Alshalalfa,M., Vergara,I.A., Erho,N. *et al.* (2017) Multi-institutional analysis shows that low PCAT-14 expression associates with poor outcomes in prostate cancer. *Eur. Urol.*, **71**, 257–266.
7. Dang,H.X., White,N.M., Rozycki,E.B., Felsheim,B.M., Watson,M.A., Govindan,R., Luo,J. and Maher,C.A. (2020) Long non-coding RNA LCAL62/LINC00261 is associated with lung adenocarcinoma prognosis. *Heliyon*, **6**, e03521.
8. Sun,L., Luo,H., Bu,D., Zhao,G., Yu,K., Zhang,C., Liu,Y., Chen,R. and Zhao,Y. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.*, **41**, e166.
9. Huang,J.-Z., Chen,M., Chen,D., Gao,X.-C., Zhu,S., Huang,H., Hu,M., Zhu,H. and Yan,G.-R. (2017) A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol. Cell*, **68**, 171–184.
10. Matsumoto,A., Pasut,A., Matsumoto,M., Yamashita,R., Fung,J., Monteleone,E., Saghatelian,A., Nakayama,K.I., Clohessy,J.G. and Pandolfi,P.P. (2017) mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, **541**, 228–232.
11. Zhang,M., Zhao,K., Xu,X., Yang,Y., Yan,S., Wei,P., Liu,H., Xu,J., Xiao,F., Zhou,H. *et al.* (2018) A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat. Commun.*, **9**, 4475.
12. Kristensen,L.S., Andersen,M.S., Stagsted,L.V.W., Ebbesen,K.K., Hansen,T.B. and Kjems,J. (2019) The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.*, **20**, 675–691..
13. Santer,L., Bär,C. and Thum,T. (2019) Circular RNAs: a novel class of functional RNA molecules with a therapeutic perspective. *Mol. Ther.*, **27**. 1350–1363.
14. Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–461.
15. Yang,Y., Gao,X., Zhang,M., Yan,S., Sun,C., Xiao,F., Huang,N., Yang,X., Zhao,K., Zhou,H. *et al.* (2018) Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. *J. Natl. Cancer Inst.*, **110**, 304–315.
16. Legnini,I., Di Timoteo,G., Rossi,F., Morlando,M., Briganti,F., Sthandier,O., Fatica,A., Santini,T., Andronache,A., Wade,M. *et al.* (2017) Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Mol. Cell*, **66**, 22–37.
17. Zheng,X., Chen,L., Zhou,Y., Wang,Q., Zheng,Z., Xu,B., Wu,C., Zhou,Q., Hu,W., Wu,C. *et al.* (2019) A novel protein encoded by a circular RNA circPPP1R12A promotes tumor pathogenesis and metastasis of colon cancer via Hippo-YAP signaling. *Mol. Cancer*, **18**, 47.
18. D'Lima,N.G., Ma,J., Winkler,L., Chu,Q., Loh,K.H., Corpuz,E.O., Budnik,B.A., Lykke-Andersen,J., Saghatelian,A. and Slavoff,S.A. (2017) A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.*, **13**, 174–180.
19. Lewandowski,J.P., Dumbović,G., Watson,A.R., Hwang,T., Jacobs-Palmer,E., Chang,N., Much,C., Turner,K., Kirby,C., Schulz,J.F. *et al.* (2019) The Tug1 locus is essential for male fertility. bioRxiv doi: https://doi.org/10.1101/562066, 28 February 2019, preprint: not peer reviewed.
20. Liang,W.-C., Wong,C.-W., Liang,P.-P., Shi,M., Cao,Y., Rao,S.-T., Tsui,S.K.-W., Waye,M.M.-Y., Zhang,Q., Fu,W.-M. *et al.* (2019) Translation of the circular RNA circβ-catenin promotes liver cancer

cell growth through activation of the Wnt pathway. *Genome Biol.*, **20**, 84.

21. Pamudurti,N.R., Bartok,O., Jens,M., Ashwal-Fluss,R., Stottmeister,C., Ruhe,L., Hanan,M., Wyler,E., Perez-Hernandez,D., Ramberger,E. *et al.* (2017) Translation of circRNAs. *Mol. Cell*, **66**, 9–21.
22. Zeng,C., Fukunaga,T. and Hamada,M. (2018) Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. *BMC Genomics*, **19**, 414.
23. Bazin,J., Baerenfaller,K., Gosai,S.J., Gregory,B.D., Crespi,M. and Bailey-Serres,J. (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E10018–E10027.
24. Guttman,M., Russell,P., Ingolia,N.T., Weissman,J.S. and Lander,E.S. (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, **154**, 240–251.
25. Mertins,P., Mani,D.R., Ruggles,K.V., Gillette,M.A., Clauser,K.R., Wang,P., Wang,X., Qiao,J.W., Cao,S., Petralia,F. *et al.* (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, **534**, 55–62.
26. Mun,D.-G., Bhin,J., Kim,S., Kim,H., Jung,J.H., Jung,Y., Jang,Y.E., Park,J.M., Kim,H., Jung,Y. *et al.* (2019) Proteogenomic characterization of human early-onset gastric cancer. *Cancer Cell*, **35**, 111–124.
27. Clark,D.J., Dhanasekaran,S.M., Petralia,F., Pan,J., Song,X., Hu,Y., da Veiga Leprevost,F., Reva,B., Lih,T.-S.M., Chang,H.-Y. *et al.* (2019) Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, **179**, 964–983.
28. Zhang,H., Liu,T., Zhang,Z., Payne,S.H., Zhang,B., McDermott,J.E., Zhou,J.-Y., Petyuk,V.A., Chen,L., Ray,D. *et al.* (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, **166**, 755–765.
29. Chen,T.-W., Lee,C.-C., Liu,H., Wu,C.-S., Pickering,C.R., Huang,P.-J., Wang,J., Chang,I.Y.-F., Yeh,Y.-M., Chen,C.-D. *et al.* (2017) APOBEC3A is an oral cancer prognostic biomarker in Taiwanese carriers of an APOBEC deletion polymorphism. *Nat. Commun.*, **8**, 465.
30. Gao,Q., Zhu,H., Dong,L., Shi,W., Chen,R., Song,Z., Huang,C., Li,J., Dong,X., Zhou,Y. *et al.* (2019) Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell*, **179**, 1240.
31. Dou,Y., Kawaler,E.A., Cui Zhou,D., Gritsenko,M.A., Huang,C., Blumenberg,L., Karpova,A., Petyuk,V.A., Savage,S.R., Satpathy,S. *et al.* (2020) Proteogenomic characterization of endometrial carcinoma. *Cell*, **180**, 729–748.
32. Zhang,B., Wang,J., Wang,X., Zhu,J., Liu,Q., Shi,Z., Chambers,M.C., Zimmerman,L.J., Shaddox,K.F., Kim,S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.
33. Sinha,A., Huang,V., Livingstone,J., Wang,J., Fox,N.S., Kurganovs,N., Ignatchenko,V., Fritsch,K., Donmez,N., Heisler,L.E. *et al.* (2019) The proteogenomic landscape of curable prostate cancer. *Cancer Cell*, **35**, 414–427.
34. Volders,P.-J., Helsens,K., Wang,X., Menten,B., Martens,L., Gevaert,K., Vandesompele,J. and Mestdagh,P. (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*, **41**, D246–D251.
35. Vo,J.N., Cieslik,M., Zhang,Y., Shukla,S., Xiao,L., Zhang,Y., Wu,Y.-M., Dhanasekaran,S.M., Engelke,C.G., Cao,X. *et al.* (2019) The landscape of circular RNA in cancer. *Cell*, **176**, 869–881.
36. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
37. Kong,A.T., Leprevost,F.V., Avtonomov,D.M., Mellacheruvu,D. and Nesvizhskii,A.I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods*, **14**, 513–520.
38. Ma,K., Vitek,O. and Nesvizhskii,A.I. (2012) A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*, **13**, S1.
39. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.* (2019) The UCSC genome browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
40. Kang,Y.-J., Yang,D.-C., Kong,L., Hou,M., Meng,Y.-Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
41. Guo,J.-C.,, Fang,S.-S.,, Wu,Y.,, Zhang,J.-H.,, Chen,Y.,, Liu,J.,, Wu,B.,, Wu,J.-R.,, Li,E.-M., Xu,L.-Y. *et al.* (2019) CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Res.*, **47**, W516–W522.
42. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
43. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
44. Nesvizhskii,A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Methods*, **11**, 1114–1125.
45. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.
46. Yang,L., Tang,Y., He,Y., Wang,Y., Lian,Y., Xiong,F., Shi,L., Zhang,S., Gong,Z., Zhou,Y. *et al.* (2017) High expression of LINC01420 indicates an unfavorable prognosis and modulates cell migration and invasion in nasopharyngeal carcinoma. *J. Cancer*, **8**, 97–103.
47. Sayanjali,B. (2017) Genome-wide transcriptome analysis of prostate cancer tissue identified overexpression of specific members of the human endogenous retrovirus-K family. *Cancer Transl. Med.*, **3**, 1–12.