NEOPLASIA
www.neoplasia.com

Volume 19 Number xx Month 2017 pp. 574–582 574

Alternative Polyadenylation Patterns for Novel Gene Discovery and Classification in Cancer

Oguzhan Begik[*], Merve Oyken[*], Tuna Cinkilli Alican[*], Tolga Can[†,‡] and Ayse Elif Erson-Bensan[*,‡]

[*]Department of Biological Sciences, M.E.T.U., Ankara, 06800, Turkey; [†]Department of Computer Engineering, M.E.T.U., Ankara, 06800, Turkey; [‡]Cancer Systems Biology Laboratory (CanSyL), M.E.T.U., Ankara, 06800, Turkey

Abstract

Certain aspects of diagnosis, prognosis, and treatment of cancer patients are still important challenges to be addressed. Therefore, we propose a pipeline to uncover patterns of alternative polyadenylation (APA), a hidden complexity in cancer transcriptomes, to further accelerate efforts to discover novel cancer genes and pathways. Here, we analyzed expression data for 1045 cancer patients and found a significant shift in usage of poly(A) signals in common tumor types (breast, colon, lung, prostate, gastric, and ovarian) compared to normal tissues. Using machine-learning techniques, we further defined specific subsets of APA events to efficiently classify cancer types. Furthermore, APA patterns were associated with altered protein levels in patients, revealed by antibody-based profiling data, suggesting functional significance. Overall, our study offers a computational approach for use of APA in novel gene discovery and classification in common tumor types, with important implications in basic research, biomarker discovery, and precision medicine approaches.

Neoplasia (2017) 19, 574–582

Introduction

Despite the flow of new information provided by genome and transcriptome sequencing studies, certain aspects of diagnosis, prognosis, and treatment of cancer patients are still important challenges to be addressed. Therefore, a better understanding of the complexity of cancer necessitates characterization of "less obvious but potentially important" changes that we generally fail to detect or consider to be noise in conventional experimental setups. From this perspective, gene expression studies face a key bottleneck; conventional methods are generally not tailored to detect nor quantify 3′ isoforms generated by alternative polyadenylation (APA) [1]. This may negatively impact our ability to discover cancer-related genes and comprehensively understand critical molecular mechanisms underlying disease progression.

APA isoforms are formed as a result of endonucleolytic cleavage of the nascent RNA at alternative poly(A) sites [2]. APA is tightly regulated and is responsive to proliferative, tissue-specific, or developmental cues [3]. APA-generated short or long 3′ untranslated region (UTR) isoforms harbor different *cis*-elements where microRNAs (miRNAs) and/or RNA-binding proteins bind [4]. Consequently, APA isoforms have different stability, localization, and translation efficiency, all of which significantly modulate protein levels and/or activity. Considering that majority of human genes have multiple poly(A) sites in their 3′-ends [5], APA constitutes an important but less understood layer of complexity in gene expression regulation. Recently, deregulation of APA has gained increasing interest in cancer research because APA emerges as a novel mechanism to activate oncogenes, generally by 3′UTR shortening and loss of repressive *cis*-elements. For example, 3′UTR shortening of *CCND1* (Cyclin D1) mRNA prevents the miRNA-mediated repression and causes further increase in *CCND1* levels, which correlate with decreased overall survival of patients [6]. Insulin-like growth factor 2 mRNA binding protein 1 (*IGF2BP1*) also goes through a shortening

Abbreviations: APA, alternative polyadenylation; poly(A), polyadenylation; miRNAs, microRNAs; 3′UTR, 3′ untranslated region; ER, estrogen receptor; SLR, short to long ratio; SAM, statistical analysis of microarrays; CfsSubsetEval, correlation-based feature selection subset evaluation; BFL, best first list; PCA, principle component analysis; IHC, immunohistochemistry; mRNA, messenger RNA

Address all correspondence to: Ayse Elif Erson-Bensan, Department of Biological Sciences, M.E.T.U., Ankara, 06800, Turkey.
E-mail: erson@metu.edu.tr

Received 19 January 2017; Revised 19 April 2017; Accepted 24 April 2017

© 2017 The Authors. Published by Elsevier Inc. on behalf of Neoplasia Press, Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

http://dx.doi.org/10.1016/j.neo.2017.04.008

of 3′UTR, and this shorter isoform is associated with profound oncogenic transformation [7]. In addition, our group reported hormone-responsive APA, where estrogen treatment resulted with upregulation and 3′UTR shortening of cell division cycle 6 (*CDC6*), a major regulator of DNA replication, in breast cancer cells. Thus, as a result of APA, the short *CDC6* isoform was linked to higher CDC6 protein levels and increased S-phase entry [8].

While numerous cases of 3′UTR shortening have been linked to increased protein levels and oncogene activation [7], consequences of 3′ UTR shortening on protein levels and functions may be complex. It turns out that 3′UTR shortening may also lead to changes in secondary structure of the mRNA, exposing hidden *cis*-elements, this time leading to decreased protein levels [9]. In addition, 3′UTR isoforms can have different functions as scaffolds to tether RNA-binding proteins that alter the localization and even function of the translated protein [10]. Alternatively, in cases where proximal poly(A) signals are within introns or coding exons, truncated proteins can be generated with potentially different and/or opposing functions (reviewed in [11]). Hence, APA can contribute to the oncogenic phenotype through various mechanisms [7].

Despite these potential impacts, APA-generated isoforms are generally undetected simply because we do not look for them in conventional gene expression analyses. Here, we report a meta-analysis pipeline for APA isoform discovery to improve cancer-related gene discovery efforts. Identification of cancer specific APA isoforms is likely to have important implications in basic cancer research and biomarker discovery fields. We anticipate our proposed comprehensive approach to be applicable to other malignancies where expression datasets are available.

Materials and Methods

Datasets

For the discovery datasets, breast, colon, lung, ovarian, and prostate cancer patient data (GSE2109), as part of Expression Project for Oncology from National Center for Biotechnology Information Gene Expression Omnibus (GEO), were utilized. Data for gastric cancer and normal samples (GSE29272) were obtained from GEO.

Breast cancer patient data included 318 patients: 69 patients (21.7%) were diagnosed as estrogen receptor negative (ER–), 146 (45.9%) were ER+, and 12 patients (3.8%) were diagnosed with triple-negative breast cancer. Colon cancer patient data included 249 cancer samples: 203 patients (81.5%) were diagnosed with adenocarcinoma, 30 (12%) with mucinous carcinoma, 15 (6%) with carcinoma arising in a villous adenoma, and 1 (0.4%) with signet ring cell carcinoma. Gastric cancer patient data included 134 patients: 62 patients (46%) were diagnosed with cardia adenocarcinoma, and 72 (54%) were diagnosed with noncardia adenocarcinoma. Lung cancer patient data had 105 samples: 32 patients (31%) were diagnosed with squamous cell carcinoma, 29 (28%) with lung adenocarcinoma, and 13 (13%) with bronchioloalveolar carcinoma. Ovarian cancer patient data included 166 samples: 28 patients (16.9%) were diagnosed with papillary serous carcinoma, 27 (16.3%) with papillary serous adenocarcinoma, and 15 patients (9%) with endometrioid cancer. Prostate cancer patient data had 73 samples: 63 patients (86%) were diagnosed as acinar type adenocarcinoma and 10 (14%) as adenocarcinoma-NOS.

Detection and Quantification of APA Events

APADetect tool [12] was used to detect and quantify APA events in common cancers. CEL files of Human Genome U133A (HGU133A, GPL96) and U133 Plus 2.0 arrays (HGU133Plus2, GPL570) were analyzed to identify intensities of probes that were grouped based on poly(A) site locations extracted from PolyA_DB [13]. For each transcript, mean signal intensities of proximal and distal probe sets were calculated. The ratio of proximal probe set mean to the distal probe set was called the "short to long" ratio (SLR). SLR values of cancer samples were compared to those of corresponding normal tissue samples. Next, SLR values were further subjected to significance analysis of microarrays (SAM) [14], as implemented by the TM4 Multiple Array Viewer tool [15], for statistical significance after log normalization. A fold change filter further eliminated APA events below a determined threshold (SLR >1.5 for shortening events or SLR <0.66 for lengthening events). SLR values reported in at least 85% of the samples were included in the subsequent analysis and classification pipeline.

Feature Selection

Correlation-based feature selection subset evaluation (CfsSubsetEval) method was used to avoid overfitting and "curse of dimensionality" problems [16,17], as implemented in WEKA data mining software [18]. CfsSubsetEval assessed the performance of a subset of attributes (i.e., SLR values) based on predictive ability and redundancy. The subset space of all the attributes were searched using the BestFirst algorithm with default parameters in WEKA [19]. The attributes were evaluated using 10-fold cross validation. To increase specificity and sensitivity, we selected SLR values that were listed as best attributes in at least 5 of the 10 cross-validations. This group of APA events was identified as best first list (BFL) (Supplementary Tables 1, 2). Heatmap illustration of APA events in BFL was done with a hierarchical clustering implemented in Multiple Array Viewer tool [15]. For the hierarchical clustering based on Pearson correlation coefficient, average linkage-based gene tree with optimized gene leaf order was used as parameter. For a distance-based comparison of the samples, we constructed color-coded gene distance matrices for normal and cancer samples using Pearson correlation coefficient (Supplementary Figure 1).

Random Forest

Random forest classifiers [20,21] were trained using SLR values in BFL. Both the selection of features and training of random forest classifiers were conducted using only the discovery (i.e., training) datasets. The classification accuracy was assessed in independent validation datasets. Use of random forest classifiers was also important for error balancing which can be critical for cancer studies as the number of control samples is usually smaller than the number of cancer samples. Confusion matrix for cancer type analysis was constructed as an output of random forest analysis.

Principle Component Analysis (PCA)

PCA [22] was performed to visualize the SLR-based separation between samples in a lower dimensional space. PCA, as implemented in WEKA, was used with default parameters. Dimensionality reduction was accomplished by choosing the top two principle components in the normal versus cancer separation and top three principle components in the cancer classification. PCA results were then visualized using GraphPad Prism 6 software and Gnuplot (http://gnuplot.sourceforge.net).

Ontology and Network Analysis

Significant APA events (SLRs <0.66 or >1.5) were analyzed by Gene Set Enrichment Analysis (GSEA) (http://www.broadinstitute.org/gsea/index.jsp) [23] and Molecular Signature Database [23]. Network database STRING (http://string-db.org) [24] was used to find potential networks in the APA-regulated transcript lists.

Immunohistochemistry (IHC) Data

Antibody-based protein profiling using IHC data from Human Protein Atlas database (http://www.proteinatlas.org) [25] was used to evaluate the protein levels of significant BFL genes. Staining intensities in normal tissues and cancer samples were presented as pathology-based annotation of protein expression levels (low, medium, high).

Results

Microarray data can be retrospectively analyzed to discover differential APA isoforms using algorithms based on probe intensities and poly(A) site location information. Earlier, we have developed and successfully implemented a meta-analysis tool for APA isoform discovery: "APADetect" [8,12] (described in "Materials and Methods"). To decipher deregulated APA profiles, we compiled a pool of cancer patient expression data representing six common cancer types (breast, colon, gastric, lung, ovary, and prostate cancers) [26] from the Expression Project for Oncology and National Center for Biotechnology Information GEO databases (Supplementary Table 3). Given that APA is tissue specific [27], independent datasets for corresponding normal tissues were included in the APA detection pipeline (Supplementary Table 4). Overall, we analyzed expression data for a total of 1045 cancer patients (318 for breast, 249 for colon, 134 for gastric, 105 for lung, 166 for ovarian, and 73 for prostate cancers) compared to 479 corresponding normal tissue samples (81 for breast, 81 for colon, 134 for gastric, 104 for lung, 38 for ovary, and 41 for prostate). For each cancer/tissue type, CEL files were analyzed by APADetect tool followed by SAM analysis (Figure 1*A*).

We identified a shift toward proximal poly(A) site selection in breast (158 of 222, %71), gastric (58 of 105, %55), and prostate cancers (96 of 104, 92%) compared to their corresponding normal tissues, whereas distal poly(A) selection was more prominent in colon (142 of 225, 63%), lung (197 of 216, 91%), and ovarian cancers (172 of 225, 76%) (Figure 1*B*).

Network and ontology analyses of significant APA events (SLR values >1.5 or <0.66, n = 458) were investigated for biological significance with STRING and GSEA tools [23,24,28] (Figure 2, *A* and *B*). Cancer-specific selection of proximal or distal poly(A) site was mostly enriched for RNA-related process (i.e., polyA RNA binding, RNA binding) and pathways that may be associated with the proliferative state of cells, suggesting APA events to potentially modulate a diverse network of downstream events (Figure 2, *A* and *B*).

Using SLR Values for Normal Versus Cancer Discrimination

Following the discovery of global APA profiles, we investigated whether subgroups of identifier APA events can discriminate normal and cancer samples. Therefore, we used feature selection and classification methods available in WEKA data mining software [18]. CfsSubsetEval and random forest methods (see "Materials and Methods" for details) showed that normal and different cancer samples were distinguished using a total of 63 unique attributes (BFL) (Supplementary Tables 1 and 2). In particular, the number of attributes with classifier potential was 16 for breast, 9 for colon, 11 for gastric, 14 for lung, 11 for ovarian, and 13 for prostate. Several APA events (e.g., *TOP2A*, *BGN*, *RPL13*) were redundant in more than one cancer type. Heat maps illustrating color-coded levels of increased or decreased SLR values, representing 3′UTR shortening or lengthening, in cancer patients compared to normal tissues are shown in Figure 3*A*. Next, we applied PCA using SLR values of transcripts in BFL and produced visualizations in which key APA events between normal tissues and cancer types were represented by a point in the plane formed by two principal axes. In the discovery set, the first two principle components explained 41%, 56%, 68%, 72%, 71%, and 69% of the total variance in breast, colon, gastric, lung, ovarian, and prostate cancer samples, respectively (Figure 3*B*, Supplementary Table 5). Gene distance matrices for normal and cancer samples using Pearson correlation coefficient are given in Supplementary Figure 1.

To further test the discriminating power of these APA events, random forest classifiers trained in discovery datasets were then applied to an independent validation set (Supplementary Table 3). The true-positive rate for distinguishing cancer samples from their corresponding normal tissues was quite significant using these classifiers. The lowest true-positive rate was 0.82 (F value, 0.81) in gastric cancer, and the highest was 1.00 in prostate cancers (Supplementary Table 6). For the validation set, PCA showed that the first two principle components explained 45%, 58%, 68%, 61%, 56%, and 74% of the total variance in breast, colon, gastric, lung, ovary, and prostate samples, respectively (Figure 3*C*, Supplementary Table 7).

Considering that APA is tissue specific [3,29], we also tested whether we can discriminate individual cancer types from a mixed pool of other cancers. Therefore, an additional CfsSubsetEval was performed to find the best attribute list for cancer classification. The feature selection on the discovery set identified 21 best first attributes for cancer discrimination (Supplementary Table 8). These attributes were then used to build a classifier model by random forest for the validation set of cancer patients. Our classifier model distinguished all six cancer types from each other. The lowest F value (0.777) was for lung cancer, where 12 of 19 (63%) of patients were correctly identified as lung cancer. F values for breast, colon, gastric, ovarian, and prostate cancers were; 0.95, 0.96, 0.95, 0.96, and 1.00, respectively (Figure 4*A*). PCA showed cancer type discrimination represented by three principal axes (Figure 4*B*, Supplementary Movie 1), suggesting good generalization performance of the classifiers. Normal tissue samples were also successfully distinguished using another classifier of 54 APA events (Supplementary Figure 2, Supplementary Tables 8 and 9).

Significance of APA Events in Cancer

Given the classifier power of APA events, a test group in BFL was individually investigated in detail for biological relevance. First, the genomic positions of active poly(A) sites in relation to gene structures were examined (Supplementary Table 2). Majority of alternative poly(A) site usage cases that have classifier abilities for common cancer types were in 3′UTRs (47 of 63); 11 of APA events resulted with activation of proximal poly(A) sites in introns and 5 poly(A) sites in coding exons (Figure 5*A*). Among these APA events, highest and lowest SLR values were plotted for cancer patients compared to normal tissue samples. SLR values indicative of APA in patient groups were indeed all statistically significant compared to normal tissue controls (Figure 5***A***) ($P < .0001$).

Next, because proximal poly(A) site usage is generally associated with higher protein levels due to loss of repressive *cis*-elements, we asked whether APA isoform variation has any potential effects on resulting protein levels in cancer patients. Therefore, protein levels of the test group of APA events were investigated in the Human Protein Atlas (http://www.proteinatlas.org/cancer) database [25] where protein expression data are derived from antibody-based protein profiling using IHC in normal tissue and cancer patient samples. For

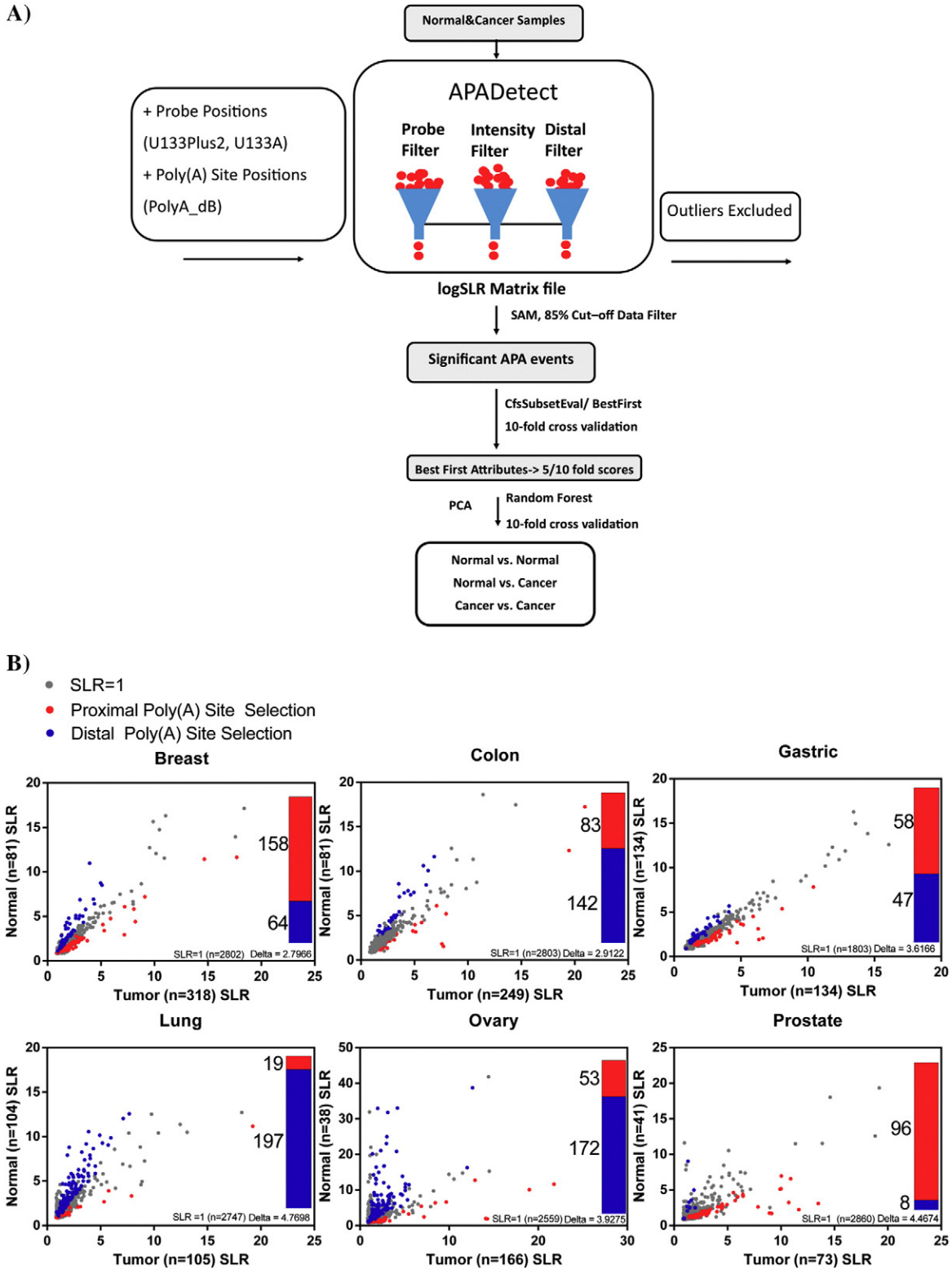


Figure 1. (A) Study workflow. CEL files of microarray data for normal (479) and cancer samples (1045) were processed with the previously developed APADetect tool [12]. U133Plus2 and U133A platforms were simultaneously analyzed through APADetect, increasing potential sample numbers. Poly(A) genomic position information was extracted from PolyA_DB. Probe intensities grouped by poly(A) site positions were processed through probe, intensity, and distal filters, which exclude outliers. LogSLR matrix file is the output file where individual APA events were assigned an SLR. APA events, which were detected in at least 85% of samples, were then run through SAM. To identify the attributes that have the best ability to discriminate between normal and cancer samples, correlation-based feature selection was compiled with BestFirst algorithm and 10-fold cross validation. Significant attributes were used in random forest and principle component analyses for classifying cancers compared to normal tissues and specific cancer types among other cancers. (B) Volcano plots for all APA events detected in six different cancer types compared to corresponding normal tissues. *X* and *Y* axes represent SLR of APA events in cancer and normal samples. Red dots represent proximal poly(A) usage, blue dots represent distal poly(A) usage, and gray dots represent insignificant SLR values in cancer patients compared to normal tissue. Bar graphs show the number of APA events.

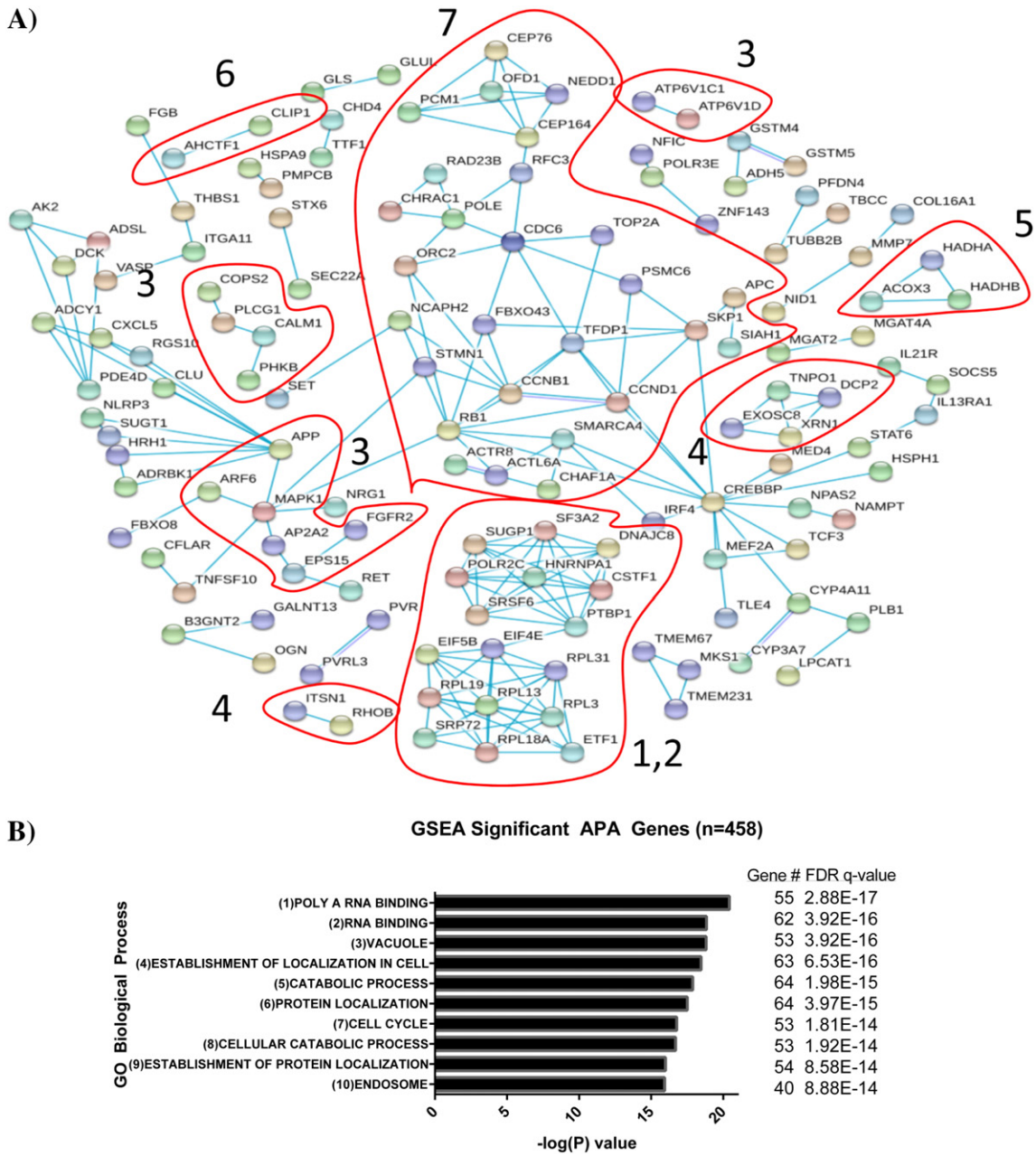


Figure 2. Network and ontology enrichment analysis for significant APA events. (A) STRING analysis illustrated multiple clusters, numbered according to GO Biological Process classes shown below. Unconnected proteins are excluded from the illustration. (B) GSEA analysis indicates enriched biological processes with highest log(*P*) values. FDR *q* values indicate the false-discovery rate for each biological process.

each cancer type, high SLRs in patients were found to be generally associated with varying degrees of high protein staining compared to normal tissues (Figure 5***B***). For example, increased proximal polyadenylation pattern of topoisomerase II alpha (*TOP2A*) was in agreement with high level of staining in breast and lung cancers. Another significant case of APA was detected for ribosomal protein L13 (*RPL13*) in breast, lung, and ovarian cancers where IHC data also suggested medium to high level of staining in all three types of cancer patients compared to normal tissue. Likewise, WD and tetratricopeptide repeats 1 (*WDTC1*) with 3′UTR shortening in lung cancers had high level of IHC staining in lung cancer patients.

We then investigated staining patterns of proteins whose transcripts did not appear to be regulated by APA (SLR = 1). A randomly selected group of (SLR = 1) transcripts had similar protein staining patterns in normal and in cancer samples, further strengthening our approach (Supplementary Figure 3).

Interestingly, high-level protein staining was also detected for some cases [e.g., SET nuclear proto-oncogene (SET)], for which 3′ UTR lengthening was observed in cancer cells, suggesting longer isoforms to enhance translation (Supplementary Figure 3*A*). This is consistent with a previous observation on 3′UTR shortening to potentiate translational repression due to altered secondary structure of the mRNA, exposing miRNA binding sites only on

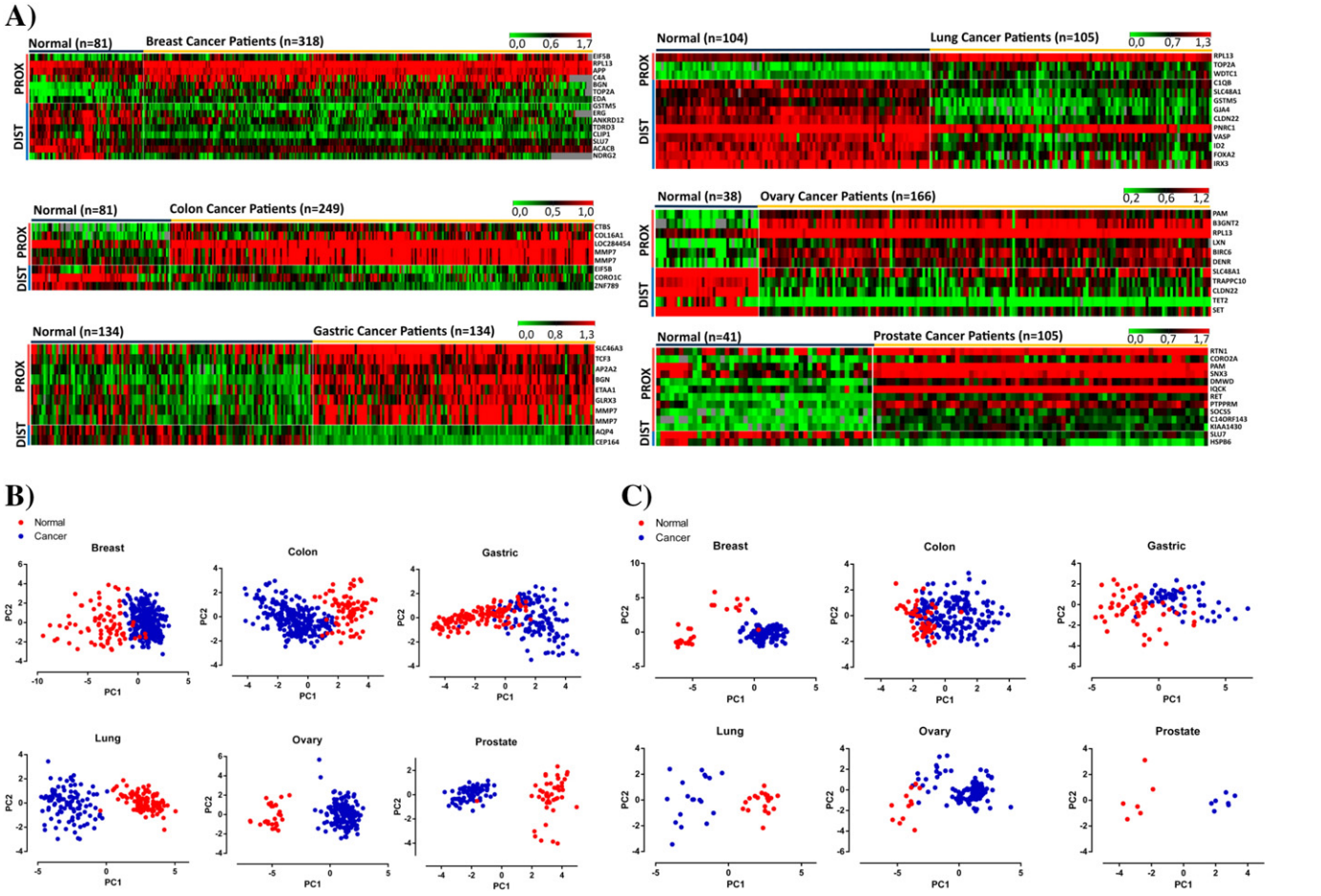


Figure 3. (A) Characteristic signatures of alternative polyadenylation cases listed in BFL in different cancer types compared to normal tissues. Heatmap key was adjusted in each sample set for optimal visualization and does not reflect the absolute SLR values (Supplementary Table 1). Red and green colors indicate increased or decreased SLR values (>1.5 fold or <0.5) in cancers compared to control tissues. Red color signifies use of proximal poly(A) sites, whereas green color signifies use of distal poly(A) sites. APA events are classified as proximal (Prox) or distal (Dist) poly(A) site usage. Each column represents a patient, and each row represents proximal or distal poly(A) site usage. Unsupervised clustering differentiated APA events. (B) Two-dimensional PCA of normal and cancer samples for discovery set using BFL. (C) PCA in independent validation sets using APA events in BFL.

the short mRNA [9]. On the other hand, 3′UTR lengthening was detected in our analysis for proline rich nuclear receptor coactivator 1 (PNRC1), which is a nuclear receptor coactivator that also interacts with GRB2 (an adapter protein involved in growth factor/Ras-mediated signaling pathways) and suppresses GRB2-mediated Ras/MAP-kinase activation. Therefore, PNRC is a potential tumor suppressor candidate, and in fact, its expression is reported to be low in breast cancers [30]. Other cases of distal poly(A) site usage and corresponding protein levels in cancer samples are shown in Supplementary Figure 3*A*.

We identified another group of APA events where active polyadenylation sites reside in introns or coding exons [e.g., SLU7 homolog, splicing factor (*SLU7*)] (Supplementary Figure 4*A*). Selection of these sites suggests APA to be coupled to alternative splicing and lead to production of truncated proteins. In the case of *SLU7*, activation of the exonic poly(A) site Hs.435342.1.9 leads to a ~70 aa truncation at the C-terminus of the protein. Interestingly, SLU7 itself is a known splicing factor that binds to the C13orf25 primary transcript in which the polycistronic oncomiR miR-17-92 resides [31]. Hence, alterations of the protein length may have functional significance. Another case, a more redundant APA case, was for *TOP2A* which had increased selection of a proximal poly(A) site within Exon 26, Hs.156346.1.29 (Supplementary Figure 4*B*). Activation of this coding sequence poly(A) site leads to 384 amino acid truncation at the C-terminus. Therefore, exonic and intronic poly(A) activation cases are of great interest to investigate whether protein activities are altered, specifically when these proteins are being considered as drug targets. It is also important to be aware of these isoforms during the choice of antibodies for detection purposes.

Discussion

In cancer cells, deregulated APA can be an important source of isoform diversity which may be overlooked in conventional gene expression analyses. For example, sequencing is becoming the standard method for transcriptome analysis, but random priming and differential PCR amplification lead to read depletion near 3′ ends, negatively affecting APA isoform discovery [32] (also reviewed in [11]). Our proposed approach takes advantage of microarray platforms where probe sets are generally designed from 3′UTRs, making it possible to identify APA events via altered signal intensities near poly(A) sites. A drawback to microarray data use is that only

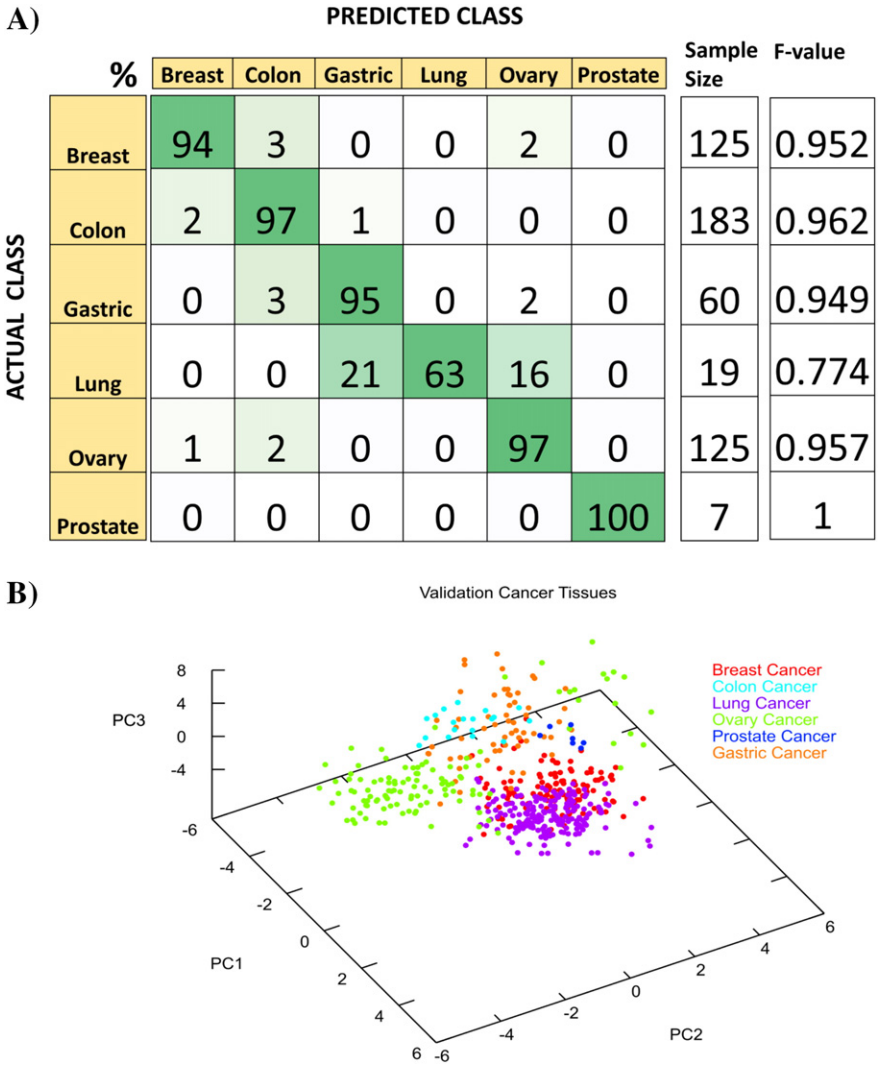


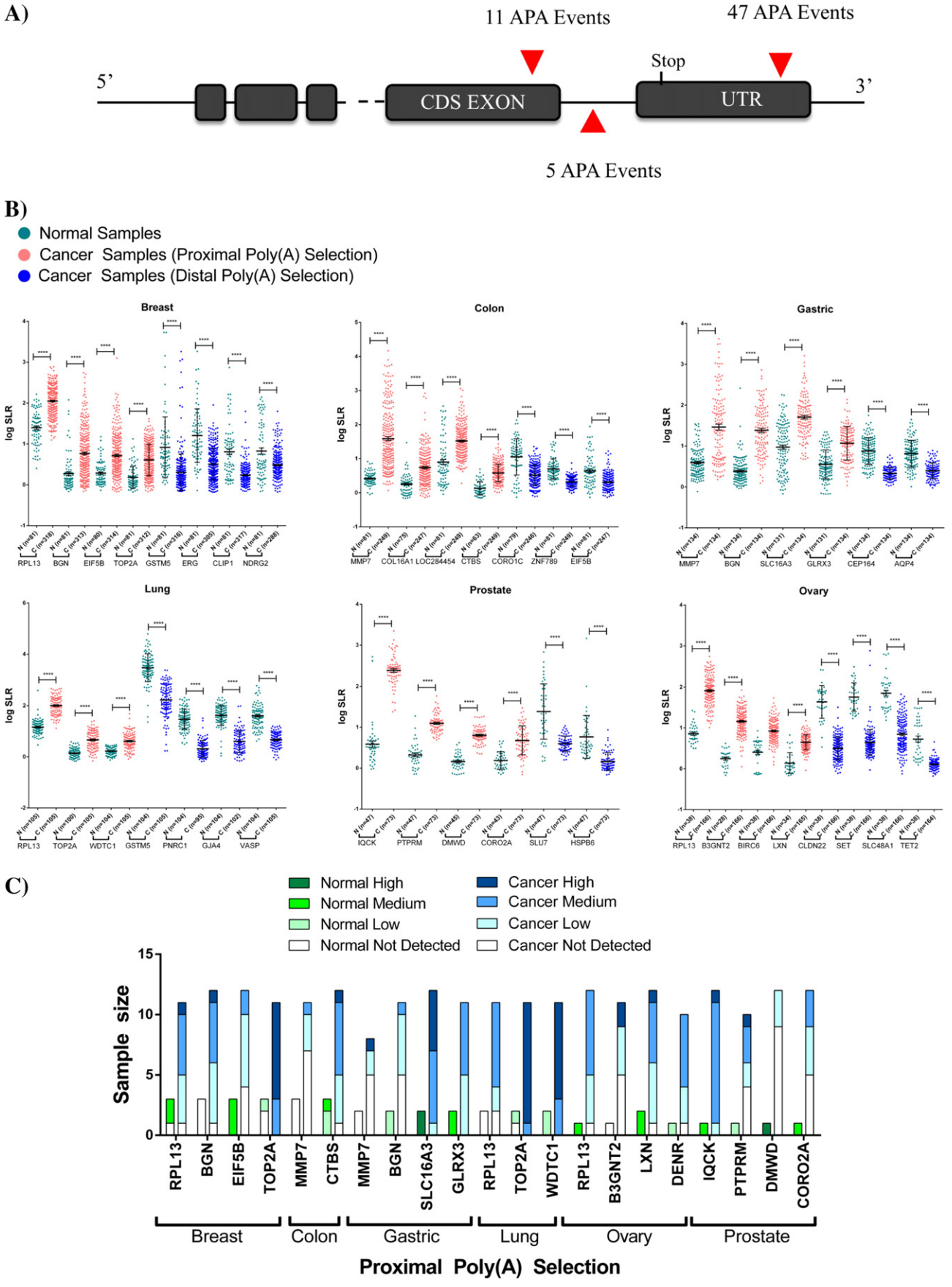
Figure 4. Classification of individual cancer types from a mixed pool of other cancers. (A) Our classifier model consisting of 21 APA events distinguished all 6 cancer types from each other. Confusion matrix demonstrates the performance of the classification model where percentage of correct calls, number of patients, and *F* values are indicated. (B) PCA showed cancer type discrimination represented by three principal axes (Supplementary Movie 1).

transcripts with probe sets divided by poly(A) sites are informative. Despite this limitation, an important advantage of our probe-based approach is rapid and simultaneous analysis of publicly available datasets. In addition, our pipeline presented here offers distinct and reproducible results to reveal APA-generated diversity in cancer cells [8,12].

Here, we show cancer-specific APA patterns. Our results indicate altered use of proximal or distal poly(A) sites in cancer transcriptome compared to normal tissue. While some of the APA-regulated genes we identified have already been implicated in tumorigenesis (e.g., *TOP2A*, *EIF5B*), a consequential link between APA and how these isoforms may contribute to tumorigenesis is not currently known. In addition, for certain oncogene activation or tumor suppressor inactivation cases, APA may provide insight into how protein levels/functions are altered in cancer cells. For example, shortened or lengthened 3′UTRs may lose or retain binding sites for *trans*-factors, which may impact the level and/or subcellular localization of the resulting protein. Alternatively, activation of intronic or coding sequence poly(A) sites may generate isoforms that differ at the 3′ends with different coding potentials.

We have to note the possibility that there may be other more proximal or distal poly(A) sites on transcripts; however, either they are not activated in that specific tissue or the probe distribution on microarrays does not allow detection of such poly(A) sites. Therefore, it is possible that APA isoforms may be more widespread in these and other cancer types. Indeed, an earlier work by Xia et al. presented APA isoform diversity in 358 TCGA pan-cancer tumor/normal pairs in a different set of tumor types (bladder urothelial carcinoma, head and neck squamous cell carcinoma, lung squamous cell carcinoma, lung adenocarcinoma, breast invasive carcinoma, kidney renal clear cell carcinoma, uterine corpus endometriosis carcinoma) using an APA-specific algorithm (DaPars) to analyze RNA-seq data [33].

In summary, our approach and results have two significant implications: 1) revealing APA isoform variation in cancer cells may help the discovery of as of yet unknown but potentially important novel cancer-related genes/pathways, and 2) tissue- and cancer type–specific APA may provide novel targets for diagnostic and prognostic purposes, which could help optimize patient outcomes using precision medicine approaches.


A)
11 APA Events
47 APA Events
Stop
5'
CDS EXON
UTR
3'
5 APA Events
B)
Normal Samples
Cancer Samples (Proximal Poly(A) Selection)
Cancer Samples (Distal Poly(A) Selection)
Breast
Colon
Gastric
Lung
Prostate
Ovary
log SLR
RPL13 BGN EIF5B TOP2A GSTM5 ERG CLIP1 NDRG2
MMP7 COL16A1 LOC284454 CTBS CORO1C ZNF789 EIF5B
MMP7 BGN SLC16A3 GLRX3 CEP164 AQP4
RPL13 TOP2A WDTC1 GSTM5 PNRC1 GJA4 VASP
IQCK PTPRM DMWD CORO2A SLU7 HSPB6
RPL13 B3GNT2 BIRC6 LXN CLDN22 SET SLC48A1 TET2
C)
Normal High
Normal Medium
Normal Low
Normal Not Detected
Cancer High
Cancer Medium
Cancer Low
Cancer Not Detected
Sample size
15
10
5
0
RPL13 BGN EIF5B TOP2A MMP7 CTBS MMP7 BGN SLC16A3 GLRX3 RPL13 TOP2A WDTC1 RPL13 B3GNT2 LXN DENR IQCK PTPRM DMWD CORO2A
Breast
Colon
Gastric
Lung
Ovary
Prostate
Proximal Poly(A) Selection

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neo.2017.04.008.

Funding

APA work in our laboratory is funded by METU internal funds and TUBITAK grants (112S478, 114Z884).

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgements

Authors would like to thank Dr. M. Muyan for discussions and critical reading of the manuscript.

References

[1] Mishra PJ, Banerjee D, and Bertino JR (2008). MiRSNPs or MiR-polymorphisms, new players in microRNA mediated regulation of the cell: Introducing microRNA pharmacogenomics. *Cell Cycle* **7**, 853–858.

[2] Tian B and Manley JL (2017). Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**, 18–30.

[3] Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, and Bartel DP (2012). Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**, 2054–2066.

[4] Erson-Bensan AE and Can T (2016). Alternative polyadenylation: another foe in cancer. *Mol Cancer Res* **14**, 507–517.

[5] Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, and Babak T (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**, 1173–1183.

[6] Rosenwald A, Wright G, Wiestner A, Chan WC, Connors JM, Campo E, Gascoyne RD, Grogan TM, Muller-Hermelink HK, and Smeland EB, et al (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185–197.

[7] Mayr C and Bartel DP (2009). Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684.

[8] Akman BH, Can T, and Erson-Bensan AE (2012). Estrogen-induced upregulation and 3′-UTR shortening of CDC6. *Nucleic Acids Res* **40**, 10679–10688.

[9] Hoffman Y, Bublik DR, Ugalde AP, Elkon R, Biniashvili T, Agami R, Oren M, and Pilpel Y (2016). 3′UTR shortening potentiates microRNA-based repression of pro-differentiation genes in proliferating human cells. *PLoS Genet* **12**, e1005879.

[10] Berkovits BD and Mayr C (2015). Alternative 3′UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367.

[11] Erson-Bensan AE (2016). Alternative polyadenylation and RNA-binding proteins. *J Mol Endocrinol* **57**, F29–F34.

[12] Gelsi-Boyer V, Orsetti B, Cervera N, Finetti P, Sircoulomb F, Rouge C, Lasorsa L, Letessier A, Ginestier C, and Monville F, et al (2005). Comprehensive profiling of 8p11-12 amplification in breast cancer. *Mol Cancer Res* **3**, 655–667.

[13] Zhang H, Hu J, Recce M, and Tian B (2005). PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res* **33**, D116–D120.

[14] Tusher VG, Tibshirani R, and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116–5121.

[15] Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, and Thiagarajan M, et al (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378.

[16] Johnson S, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, Labourier E, Reinert K, Brown D, and Slack F (2005). RAS is regulated by the let-7 microRNA family. *Cell* **120**, 635–647.

[17] Hall AM (1999). Correlation-Based Feature Selection for Machine Learning Department of Computer Science, vol. Doctor of Philosophy. The University of Waikato; 1999 178.

[18] Chendrimada T, Finn K, Ji X, Baillat D, Gregory R, Liebhaber S, Pasquinelli A, and Shiekhattar R (2007). MicroRNA silencing through RISC recruitment of eIF6. *Nature* **447**, 823–828.

[19] Hall MA (2000). Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.; 2000

[20] Kovalchuk O, Tryndyak V, Montgomery B, Boyko A, Kutanzi K, Zemp F, Warbritton A, Latendresse J, Kovalchuk I, and Beland F, et al (2007). Estrogen-induced rat breast carcinogenesis is characterized by alterations in DNA methylation, histone modifications and aberrant microRNA expression. *Cell Cycle* **6**, 2010–2018.

[21] Lee Y, Kim M, Han J, Yeom K, Lee S, Baek S, and Kim V (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23**, 4051–4060.

[22] Jolliffe IT (2002). Principal Component Analysis. Second ed: Springer; 2012.

[23] Frankel L, Christoffersen N, Jacobsen A, Lindow M, Krogh A, and Lund A (2008). Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *J Biol Chem* **283**, 1026–1033.

[24] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, and Tsafou KP, et al (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447–D452.

[25] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, and Asplund A, et al (2015). Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419.

[26] Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, Muller RU, Straka E, Su L, and Burki EA, et al (2008). A SNP in a let-7 microRNA complementary site in the KRAS 3′ untranslated region increases non–small cell lung cancer risk. *Cancer Res* **68**, 8535–8540.

[27] Miura P, Shenker S, Andreu-Agullo C, Westholm JO, and Lai EC (2013). Widespread and extensive lengthening of 3′UTRs in the mammalian brain. *Genome Res* **23**, 812–825.

[28] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, and Laurila E, et al (2003). PGC-1alpha–responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267–273.

[29] O'Donnell K, Wentzel E, Zeller K, and Dang C (2005). Mendell J. c-Myc–regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839–843.

[30] Zhou D, Chen B, Ye JJ, and Chen S (2004). A novel crosstalk mechanism between nuclear receptor-mediated and growth factor/Ras-mediated pathways through PNRC-Grb2 interaction. *Oncogene* **23**, 5394–5404.

[31] Urtasun R, Elizalde M, Azkona M, Latasa MU, García-Irigoyen O, Uriarte I, Fernández-Barrena MG, Vicent S, Alonso MM, and Muntané J, et al (2016). Splicing regulator SLU7 preserves survival of hepatocellular carcinoma cells and other solid tumors via oncogenic miR-17-92 cluster expression. *Oncogene* **35**(36), 4719–4729.

[32] Zheng D, Liu X, and Tian B (2016). 3'READS+, a sensitive and accurate method for 3' end sequencing of polyadenylated RNA. *RNA* **22**, 1631–1639.

[33] Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, and Li W (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**, 5274.

Figure 5. (A). BFL APA events and exemplary log SLR values for proximal and top distal poly(A) site usage cases plotted for each cancer type. Unpaired *t* test with Welch's correction was used to compare group means. **** indicates statistical significance ($P < .0001$). (B) Protein levels detected by IHC in cancer and normal samples are shown. Data for antibody-based protein profiling was extracted from the Human Protein Atlas [25]. Sample size indicates the number of cancer and normal samples scored for staining intensities (low, medium, high) as pathology-based annotation of protein expression levels.