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

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.