RESEARCH ARTICLE

Alternative polyadenylation of single cells delineates cell types and serves as a prognostic marker in early stage breast cancer

Nayoung Kim[1,2], Woosung Chung[1], Hye Hyeon Eum[1,2], Hae-Ock Lee[1,2,3]*, Woong-Yang Park[1,2,3,4]*

1 Samsung Genome Institute, Samsung Medical Center, Seoul, South Korea, **2** Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Suwon, South Korea, **3** Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences &Technology, Sungkyunkwan University, Seoul, South Korea, **4** GENINUS Inc., Seoul, South Korea

* haeock.lee@samsung.com (HOL); woongyang.park@samsung.com (WYP)

Abstract

Alternative polyadenylation (APA) in 3' untranslated regions (3' UTR) plays an important role in regulating transcript abundance, localization, and interaction with microRNAs. Length-variation of 3'UTRs by APA contributes to efficient proliferation of cancer cells. In this study, we investigated APA in single cancer cells and tumor microenvironment cells to understand the physiological implication of APA in different cell types. We analyzed APA patterns and the expression level of genes from the 515 single-cell RNA sequencing (scRNA-seq) dataset from 11 breast cancer patients. Although the overall 3'UTR length of individual genes was distributed equally in tumor and non-tumor cells, we found a differential pattern of polyadenylation in gene sets between tumor and non-tumor cells. In addition, we found a differential pattern of APA across tumor types using scRNA-seq data from 3 glioblastoma patients and 1 renal cell carcinoma patients. In detail, 1,176 gene sets and 53 genes showed the distinct pattern of 3'UTR shortening and over-expression as signatures for five cell types including B lymphocytes, T lymphocytes, myeloid cells, stromal cells, and breast cancer cells. Functional categories of gene sets for cellular proliferation demonstrated concordant regulation of APA and gene expression specific to cell types. The expression of APA genes in breast cancer was significantly correlated with the clinical outcome of earlier stage breast cancer patients. We identified cell type-specific APA in single cells, which allows the identification of cell types based on 3'UTR length variation in combination with gene expression. Specifically, an immune-specific APA signature in breast cancer could be utilized as a prognostic marker of early stage breast cancer.

OPEN ACCESS

Citation: Kim N, Chung W, Eum HH, Lee H-O, Park W-Y (2019) Alternative polyadenylation of single cells delineates cell types and serves as a prognostic marker in early stage breast cancer. PLoS ONE 14(5): e0217196. https://doi.org/10.1371/journal.pone.0217196

Editor: Aamir Ahmad, University of South Alabama Mitchell Cancer Institute, UNITED STATES

Received: July 27, 2018

Accepted: May 8, 2019

Published: May 17, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This study was supported by grants from the National Research Foundation of Korea (NRF-2017R1A6A3A11027978 to N.K., NRF-2017M3C9A6044636 and NRF-2016R1A5A1011974 to H.L. and NRF-2017M3A9A7050803 to W.P.) and from the Ministry of Food and Drug Safety of Korea (16173MFDS004 to W.P.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. W.P. is employed by a commercial company, GENINUS Inc, who provided support in the form of salary for W.P., but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific role of this author is articulated in the 'author contributions' section.

Competing interests: W.P. is employed by a commercial company, GENINUS Inc. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Background

Alternative polyadenylation (APA) in 3' untranslated regions (3'UTR) is a major post-transcriptional mechanism, controlling gene expression by influencing transcript abundance, cellular localization, and interaction with microRNAs [1]. Recent studies have found that the change in 3'UTR length is tightly associated with the regulation of cell proliferation as well as differentiation during an immune response [2] and cancer growth [3]. The usage of shorter 3'UTRs via APA events is most common (91%) in cancer and occurs on a global scale [4]. There has been an increase in the trial to systematically detect APA events across diverse cancers including breast cancer [3, 5]. Especially, APA usages in specific genes and transcriptional signalings, such as PRELID1 [6], USP9X, SNX3, and YME1L1D [7], have been reported as a strong predictor of clinical outcomes in the breast cancer. Although the biological significance is widely accepted, its clinical application as a prognostic biomarker or therapeutic target is not fully evaluated. Thus, understanding the regulation of expression via APA events across diverse cell types may provide new insights into cancer therapeutics.

Recently, a range of algorithms has been developed to quantify relative changes in 3'UTR length using RNA sequencing data to infer APA events. There are two classes of analytical methods created for the identification of de novo 3'UTR sites. Algorithms such as Dynamic analyses of Alternative PolyAdenylation from RNA-Seq (DaPars) [4] and 3USS [8] were developed to identify the locations of novel 3'UTR sites, whereas predefined APA sites from public databases were utilized in Roar [9], MISO [10], and ChangePoints [11], etc. The combined use of those methods can provide an opportunity to identify novel and reliable 3'UTR APA events using large-scale RNA sequencing datasets.

Single-cell RNA sequencing is primarily used to explore intratumoral heterogeneity in gene expression. Detecting cell-to-cell variations in polyadenylation site usage has been suggested [12], but single-cell level analysis for APA events has rarely been explored on a large scale. Single-cell RNA sequencing datasets in diverse cancer types can be found in public repositories like JingleBells [13] and scRNASeqDB (https://bioinfo.uth.edu/scrnaseqdb/). Especially, full-length single-cell RNA sequencing data contains genome-wide reads that permit prediction of relative 3'UTR usage. The APAs inferred from those predictions allow comprehensive profiling of alternative 3'UTRs at the single-cell level.

In the present study, we analyzed a single-cell RNA sequencing dataset containing 515 cells from 11 breast cancer patients to profile comprehensive APA events across diverse cell types. We characterized 1,176 functional categories and 53 genes associated with shorter 3'UTR usage and over-expression specific to each cell type. Most of gene sets (94.7%) were unique, and only partial gene sets (5.3%) were recurrently identified in multiple cell types. Among the gene sets, those associated with proliferation marked distinct APA events strictly confined to specific cell types. Furthermore, the expression level of those gene signatures was significantly associated with patient survival. Finally, the APA profiles on 598 tumor cells covering breast cancer, glioblastoma, and renal cell carcinoma distinguished tumor samples belonging to the cancer type. Therefore, the large-scale single-cell analysis demonstrates that cell type and cancer type-specific transcriptional regulation is associated with APA signals.

Materials and methods

Data acquisition

The raw RNA-seq data for 515 cells from 11 breast cancer patients [14] and 34 cells from a renal cancer patient [15] were obtained from the NCBI Gene Expression Omnibus database under the accession codes GSE75688 and GSE73122. Raw RNA-seq data for 355 cells from 3 glioblastoma patients were downloaded from European Genome-phenome Archive (EGA) with accession code EGAS00001001880 [16]. The data were generated by C1 Single-Cell Auto Prep System (100–5760, Fluidigm, San Francisco, CA, USA) as full-length transcripts. Detailed information of clinical parameters, cell types, and samples for the acquired datasets were described in S1 Table. To estimate the usage of 3'UTR length, we generated the .bam file as an input for the method Roar using the 2-pass mode of STAR_2.4.0b (default parameters) [17]. We generated .bedgraph file as an input for the method DaPars from the .bam file using the 'genomCoverageBed' command (BEDtools v2.17.0) [18]. For the expression analysis in breast cancer, we extracted 34,942 genes for which expression values were present in at least one cell. Relative expression of each gene is represented by transcripts per million (TPM) using RSEM v1.2.17 (default parameters) [19]. As a normal reference, we obtained the raw RNA-seq data of normal breast, brain, and kidney tissues from the Body Map 2.0 project from ArrayExpress (Query ID: E-MTAB-513, available from http://www.ebi.ac.uk/arrayexpress). Sequential methods for read alignments and quantification were applied in accordance with the process of preparing single-cell RNA-seq data. For survival analysis, RNA-seq and clinical data from patients' Breast Invasive Carcinoma (BRCA) samples were obtained from The Cancer Genome Atlas (TCGA). This RNA-Seq data (Level 3) included 1,073 (updated on 2017) tumors, and the expression of each gene is represented as upper quartile normalized RSEM (RNA-Seq by Expectation Maximization) count estimate. And the RNA-seq BAM files for 110 BRCA/normal breast pairs were downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/). Subtypes of BRCA tumors were predicted using the R package 'genefu.'

Estimation of usage of 3'UTR length

APAs were estimated by length changes of 3'UTR using two complementary methods, DaPars (default parameters) [4] and Roar [9]. DaPars searches all regions of 3'UTR in the reference genome (hg19), while Roar uses the .gtf files generated from public APA databases of Poly-A_DB2 [20] using the given single-cell and bulk RNA-seq samples.

Transformation of gene-level data into gene sets

The APA estimates of 'change in Percentage of Distal polyA site Usage Index' (ΔPDUI, by DaPars) and 'Ratio of A Ratio' (roar, by Roar) were separately used as input data. Gene expression quantified as TPM was log2 transformed after plus 1. To assess the APA regulation and gene expression based on pathway activation, all gene symbols were matched with EntrezID and then applied to ssGSEA (using options in R packages called as 'GSVA') to calculate an enrichment score per gene set. A total of 5,917 Gene Ontology (GO) terms were collected for ssGSEA referring to a gene set database, MSigDB v6.0.

Selection of variable components

For the selection of variable components (genes and gene sets) for hierarchical clustering and dimensional reduction, we scored each gene and gene set for the variation level, defined as standard deviation (s.d.) across all single cells. Variable components were then identified as those with the variation level above the baseline, the mean and s.d. of all components. For gene-level analysis, we additionally filtered out genes estimated in less than 30% of all single cells.

Statistical analysis to select cell-type specific signatures

To compare gene expression and APA levels of various cell types, Pearson's correlation coefficient (PCC) was calculated in the scale of enrichment score for gene sets. To determine the statistical significance of PCC, we calculated p-values based on Fisher's Z transformation. To describe the enriched association of 3'UTR shortening and over-expression with a cell type, we calculated an odds ratio (OR) by quantifying a given cell population. All single-cells were classified into 4 groups–(A) cells showing 3'UTR shortening and over-expression on a cell type, (B) cells showing 3'UTR shortening and over-expression on the others, (C) cells not showing 3'UTR shortening and over-expression on a cell type, (D) cells not showing 3'UTR shortening and over-expression in the others. The median APA level and gene expression were used to determine whether a cell represented 3'UTR shortening and over-expression for each gene set and gene. Then, the Fisher exact test was used to determine the statistical significance of the agreement between individual query pairs [21]. For the selection of hits, we applied the cutoffs for PCC $>$ 0 (p-value $<$ 0.05) and OR $>$ 2 (p-value $<$ 0.01).

Selection of tumor-specific gene sets

To select gene sets showing significantly switched 3'UTR specific to each cancer type, delta and t-test p-value were calculated using APA prediction data transformed into the gene-set. The delta of a gene set is given by the difference between the average of single cells for a given tumor type and the others. The significance (p-value) was calculated by two-sided t-statistics.

Network-based clustering of gene sets

Clusters based on biological functions of gene sets were graphically organized into an interaction network using Cytoscape (v3.5.1) [22]. The edges in the network were determined by Jaccard index as intersection over union for the number of shared parent GO terms between two gene sets. The distances between nodes (gene sets) were defined using a force-directed layout.

Survival analysis

For the refined analysis of survival rate, we acknowledged patient survival if the time of death after diagnosis was longer than 10 years. The tumor samples were divided into two classes along 25th and 75th percentiles of expression for each target gene. Survival curves were fitted using a Kaplan–Meier formula in the R package 'OIsurv.' Additionally, we performed multivariate Cox regression to investigate the relative risks in the R package 'survival.' The regression model was constructed considering 10 events such as age, race, pathologic stage, tumor weight, the presence or absence of ER/PR/Her2 and an indication of drug/radiation regimen, and expression classes of each gene.

Results

Cellular heterogeneity of 3'UTR length changes

The availability of full-length single-cell RNA sequencing data from patient tumor tissues provides the opportunity to predict and compare the changes in 3'UTR length among diverse cell types. We applied two complementary methods to determine the shortening and lengthening of 3'UTRs. DaPars scans all regions of 3'UTR in genes to detect novel APA sites [4], while Roar focuses on the known APA sites of 3'UTR to improve the sensitivity [9]. We mainly utilized the results of 3'UTR switching estimated from DaPars.

In this study, we used full-length single-cell RNA sequencing data generated in breast cancer patients' tumor tissues to compare the APA regulations in diverse cell types [14].

Widespread shortening of 3'UTR is a global trend in cancer [3]. As expected, 3'UTR shortening was dominant in all cells derived from breast cancer patients' tumor tissues (Fig 1A and 1B, S1A Fig). To increase the resolution of pattering, overall APA patterns were compared at the level of gene sets using single sample GSEA (ssGSEA). Clustering based on APA profiles in gene sets showed subgroups divided into tumor and non-tumor cells (Fig 1C and 1D, S1B Fig). No differences were associated with cancer subtypes or sample batches.

Unique APA regulation among cell types

To identify the functional categories associated with diverse cell types, we utilized two metrics to determine the associations between 3'UTR usage and expression regulation in different cell types. Correlation coefficient reflects the global relationship of expression and 3'UTR length change in each cell type, while the odds ratio quantifies the specificity of 3'UTR shortening and expression of a gene set within each cell type (See methods for details). These two metrics allowed us to prioritize the cell type-specific gene sets for further study.

Consequently, we classified 1,176 gene sets specific to 5 main cell types of tumor, B lymphocytes, T lymphocytes, myeloid, and stromal cells (Fig 2A, S2 Table). Most of the hits were distinct among the cell types except for partial (5.3%) gene sets, which were recurrently selected in multiple cell types. These selected gene sets were clustered on the basis of their annotated GO terms in the network (Fig 2B, S2 Table). As a result, we observed dominant functional categories assigned to each cell type. Tumor cell-specific gene sets were widely distributed in diverse biological events such as apoptotic signaling, cell morphogenesis, and metabolic process. Previous work using deep sequencing of 3'UTRs of mRNA has shown that genes with switched APA sites are enriched in pathways including cell cycle, apoptosis, and metabolism in breast cancer cell lines [23]. Immune and stromal cell-specific gene sets were clustered together as functional categories of 'immune response' and 'response to stimulus'. In particular, the clustering pattern of gene sets specific to 4 cell types except for stromal cell was exhibited in an event of cell proliferation. General shortening in 3'UTR length is tightly related to states of cellular proliferation and dedifferentiation [1]. We further confirmed that 3'UTR shortening and gene expression for gene sets classified into cell proliferation showed correlation patterns restricted to the associated cell type (Fig 3). This result suggests that APA linked to expression regulation is highly dependent on unique cellular lineages. Thus, the understanding of APA at single-cell resolution is useful to recognize the difference in transcriptional regulation among diverse cells in cancers.

Distinct APA regulation among cancer types

Previous work has reported that regulation of APAs varies between tissues [24]. To assess APA variations among tissues and individual patients, we collected more public full-length single-cell RNA sequencing data generated from glioblastoma [16] and renal cell carcinoma [15] tissues. For the comparison of tumor cell types, we only used the APA-predicted data in 280 tumor cells from breast cancer patient' tumor tissues. Clustering based on 3'UTR length changes at the level of gene sets clearly distinguished tumor tissues (Fig 4A and 4B). We further investigated the gene sets contributing to tumor-specific clusters through differential APA regulation. We identified 739, 898, and 731 gene sets significantly ($p<0.01$) switched in breast cancer, glioblastoma, and renal cell carcinoma, respectively (S3 Table). Sorting through the delta, we compiled a list of the most differential gene sets (top10 hits) in each cancer and confirmed discriminative enriched patterns toward 3'UTR shortening (Fig 4C). The biologic functions highlighting the listed gene sets are known to play important roles in carcinogenesis for each cancer. For example, oxidoreductase family genes were reported as therapeutic targets in

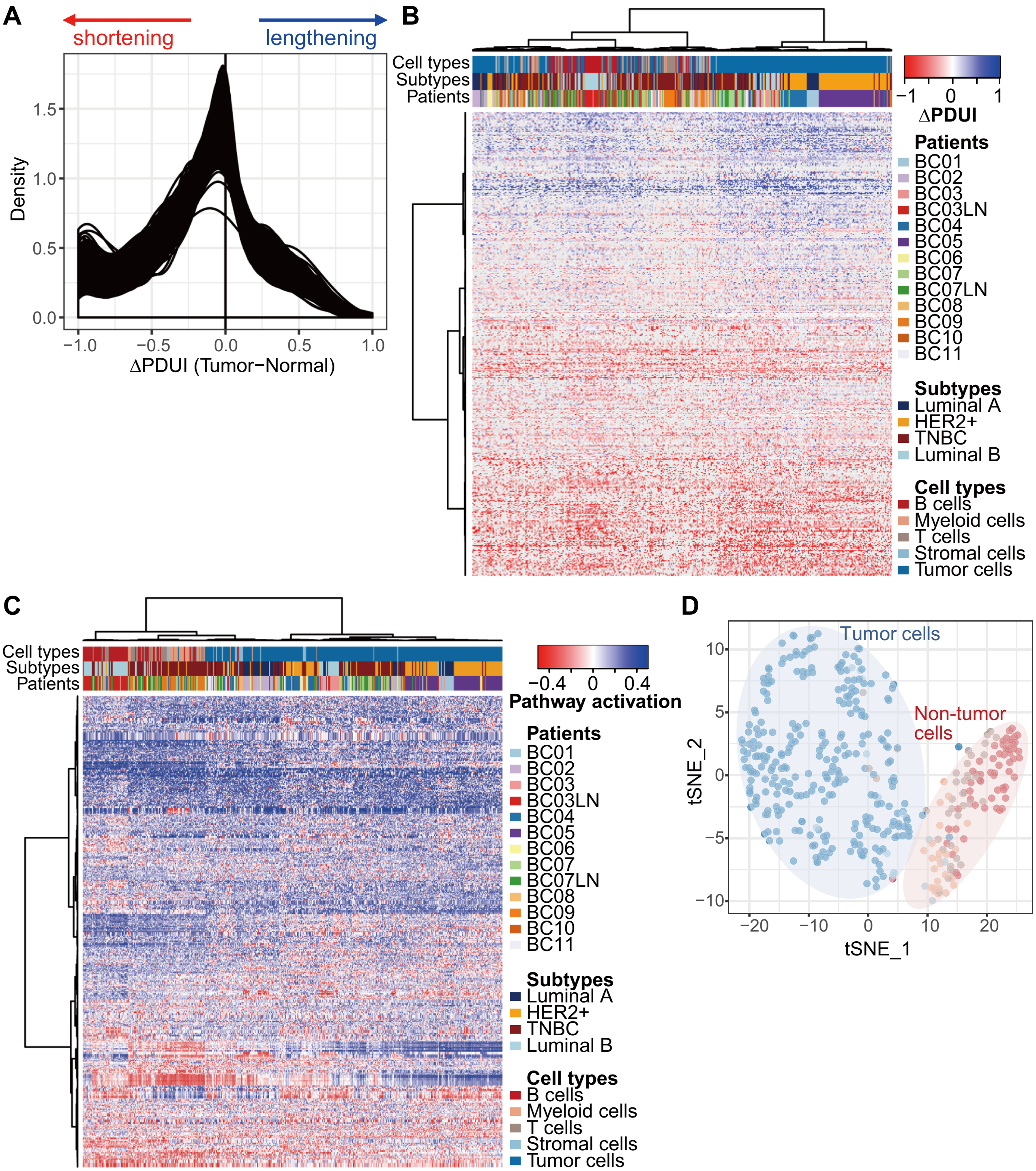


Fig 1. Different APA patterns between tumor and non-tumor cells. (A) The signals of APA (ΔPDUI calculated by DaPars) for all genes were shifted toward 3'UTR shortening in 515 single cells derived from 11 breast cancer patients (BC01-11). We followed the labels for the patients and cell types pre-defined to original datasets [14]. BC; Breast cancer, LN; Lymph nodes. Hierarchical clustering was performed using APA estimates (B) for genes and (C) gene sets on individual cells derived from breast cancer patients. Only 419 cells with at least 542 (10% of total genes) detected genes were selected for further analysis. A total of 1,452 genes and 3,353 gene sets showed the signals of APA on 419 cells. Among theme, 453 genes and 555 gene sets were selected as variable components for hierarchical clustering and dimensional reduction, respectively (see Methods). (D) Unsupervised tSNE on the gene set-level APAs separating 280 tumor and 139 non-tumor cells (immune and stromal cells) in breast cancer. Individual cells are colored for samples, consistent with the coloring in Fig 1B and 1C.

https://doi.org/10.1371/journal.pone.0217196.g001

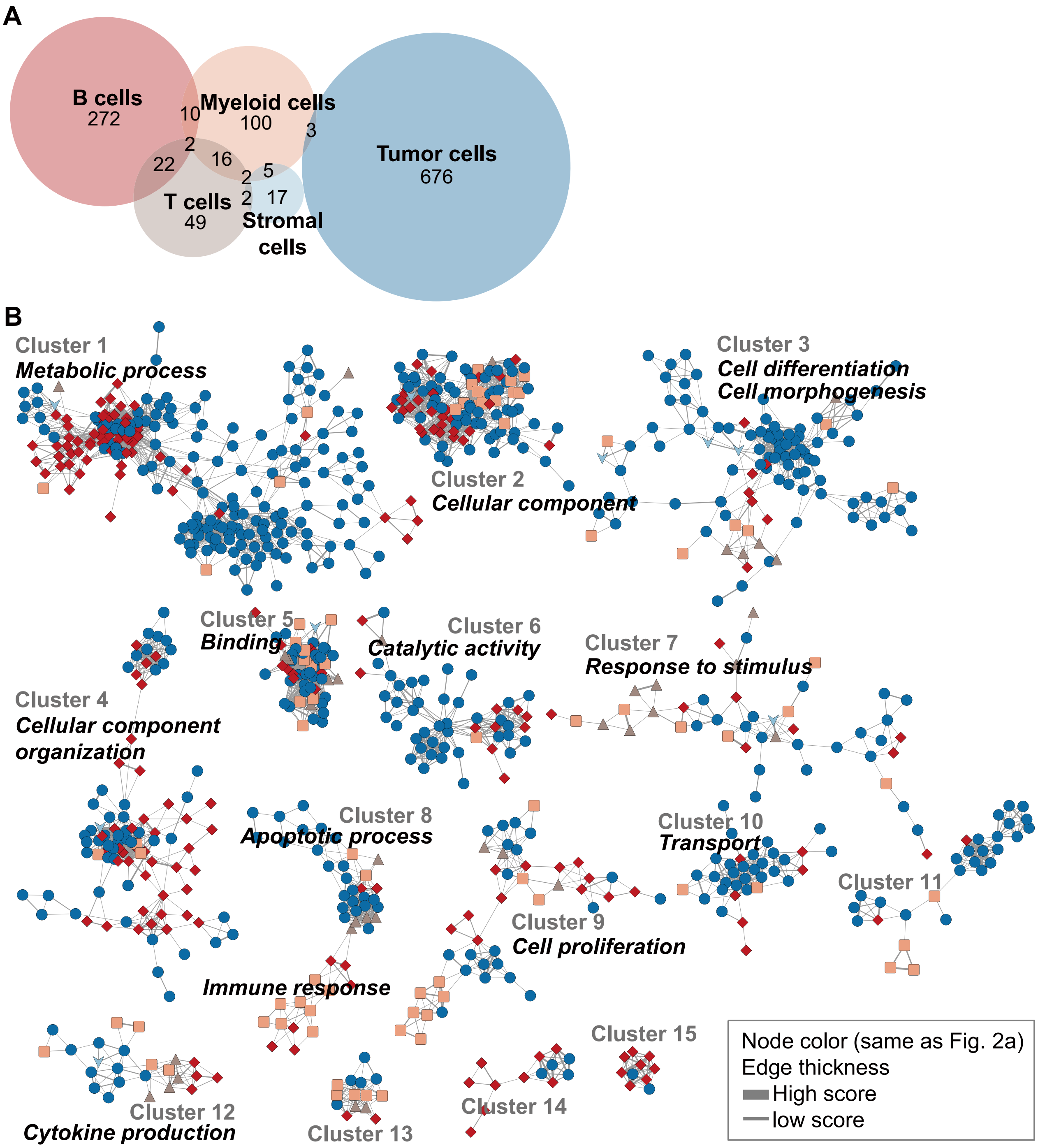


Fig 2. Cell type-specific functional categories associated with 3'UTR shortening and overexpression. (A) Venn diagram for 1,176 gene set hits specific to 5 cell types. Cell type specific gene sets were selected by correlation coefficient and odds ratio. (B) Network-based functional analysis of the selected gene sets. A total of 906 gene sets with sufficient interactions (> 0.5 Jaccard index and > 9 gene sets in a cluster) were selected for display. The cluster annotations of each gene set are available in S2 Table.

https://doi.org/10.1371/journal.pone.0217196.g002

brain tumor [25, 26]. A k63-linked ubiquitination is the major modulation inducing specific expression of calmodulin-like protein 5 in patients with primary breast cancer [27]. In addition, the role of the Arp2/3 complex was experimentally validated in the movements of kidney

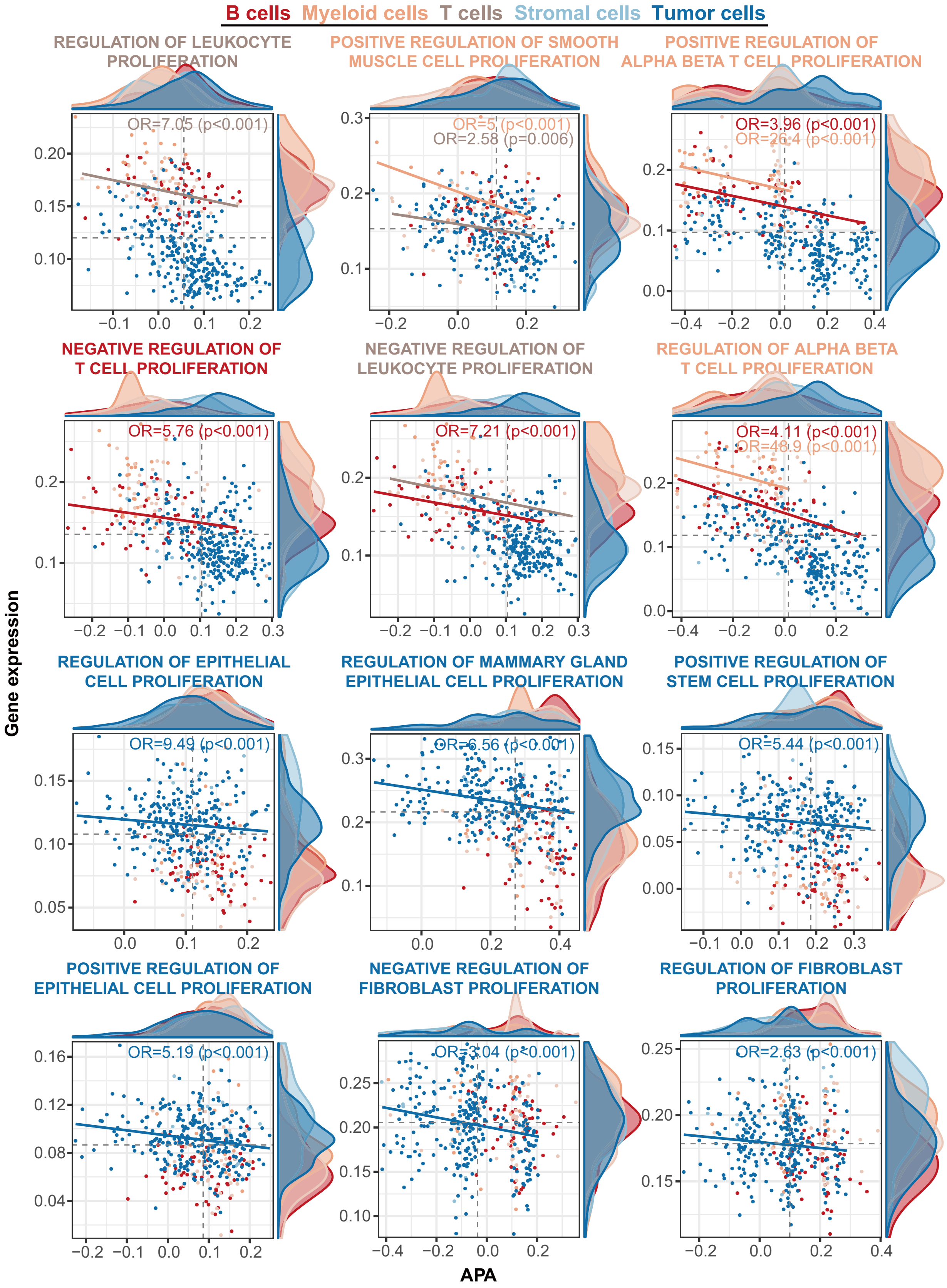

B cells
Myeloid cells
T cells
Stromal cells
Tumor cells
REGULATION OF LEUKOCYTE PROLIFERATION
OR=7.05 (p<0.001)
POSITIVE REGULATION OF SMOOTH MUSCLE CELL PROLIFERATION
OR=5 (p<0.001)
OR=2.58 (p=0.006)
POSITIVE REGULATION OF ALPHA BETA T CELL PROLIFERATION
OR=3.96 (p<0.001)
OR=26.4 (p<0.001)
NEGATIVE REGULATION OF T CELL PROLIFERATION
OR=5.76 (p<0.001)
NEGATIVE REGULATION OF LEUKOCYTE PROLIFERATION
OR=7.21 (p<0.001)
REGULATION OF ALPHA BETA T CELL PROLIFERATION
OR=4.11 (p<0.001)
OR=48.9 (p<0.001)
REGULATION OF EPITHELIAL CELL PROLIFERATION
OR=9.49 (p<0.001)
REGULATION OF MAMMARY GLAND EPITHELIAL CELL PROLIFERATION
OR=6.56 (p<0.001)
POSITIVE REGULATION OF STEM CELL PROLIFERATION
OR=5.44 (p<0.001)
POSITIVE REGULATION OF EPITHELIAL CELL PROLIFERATION
OR=5.19 (p<0.001)
NEGATIVE REGULATION OF FIBROBLAST PROLIFERATION
OR=3.04 (p<0.001)
REGULATION OF FIBROBLAST PROLIFERATION
OR=2.63 (p<0.001)
Gene expression
APA

Fig 3. Cell type-dependent 3' UTR shortening and expressional changes for genes associated with proliferation. The associations between 3'UTR switching and gene expression were compared for 12 gene sets classified into cellular proliferation. Dots in the plot represent 419 single-cells and are colored for each cell type. Odds ratio (OR) and significance are marked for specific cell types in the top right of the plot. The solid lines denote the fitted line constructed by generalized linear model (R function 'glm') and are colored for the matching cell types.

https://doi.org/10.1371/journal.pone.0217196.g003

cells [28]. These results suggest that the features of APA events are distinctively classified at the level of tissues, but not individuals.

Clinical relevance of APA-associated marker genes

To clarify the associations between gene markers and APA regulation, we re-analyzed single-cell sequencing data from breast cancer patients at the gene level. We found 53 genes associated with shorter 3'UTR usage and over-expression specific to each cell type (Fig 5A, S4 Table). Consistent with the gene set-level results, most of the hit genes were distinct among cell types except for 4 overlapping genes for immune cell types. To explore the clinical impact of those genes, we performed Kaplan-Meier survival analysis and examined the association of the expression changes and survival in breast cancer patients using TCGA RNA sequencing data. The tumor samples were annotated as 'high' and 'low' (25th and 75th percentiles, respectively) along with the expression signal of each gene. In this analysis using all 1,073 BRCA samples, 10 of 53 genes showed significantly ($p<0.05$) different survival rates between the two groups (Fig 5B and 5C). The examples of read density for four selected genes (*SET*, *HSP90AA1*, *YWHAZ*, and *RHOA)* is clearly lower in each single-cell on the portion after proximal APA site (S2 Fig). By changing sample windows along the cancer stages, we confirmed that expression levels for a total of 11 genes significantly ($p<0.05$) affected survival rate in early stage tumors. In stark contrast, only 2 genes affected survival under the cutoff p-value of 0.05 in late-stage tumors. There was no difference in survival for molecular subtypes of breast cancer. Additionally, to overcome the effects of confounding factors such as cancer stage, age, and race, we performed multivariate Cox regression analysis, which estimates survival probabilities using more than one variable for modeling. Among 11 genes denoted as significant from Kaplan-Meier survival analysis in tissue samples, we demonstrated that the expression levels for 10 genes were independent ($p<0.05$) factors affecting survival in BRCA patients.

Our data mining results suggest that cell type-specific genes showing 3'UTR shortening and over-expression harbor clinical relevance. The functions of the selected genes are tightly connected with the specific cell type (Fig 5C). For example, over-expression of *SET* (SET nuclear proto-oncogene), defined as a tumor cell-specific signature, has been demonstrated in 50–60% of breast cancer cases [29]. The study suggested inhibition of SET as a potential antitumor strategy in breast cancer [29, 30]. Heat Shock Protein 90 (*HSP90*) is a T lymphocyte-specific signature and is known as a regulator of LAT (Linker of Activated T cells), which induces T cell activation [31–33]. A recent study provided strong evidence that the combination of immunotherapy and HSP90 inhibitors augments T-cell mediated anti-tumor response [34]. B cell-derived 14.3.3 protein zeta/delta (*YWHAZ*), defined as a B-lymphocyte specific signature, plays an important role in T cell trafficking [35]. Finally, Ras homolog gene family, member A (*RHOA*) is a T and Myeloid cell specific signature and is pivotal for the functioning of T cells[36] and macrophages[37]. Taken together, these results show that APA-associated gene signatures have a biological and clinical impact and warrant further investigations as prognostic and therapeutic targets.

Discussion

Single-cell RNA sequencing provides an opportunity to investigate cell type specific connections among different genomic features. For instance, prior scRNA-seq studies of tumors have

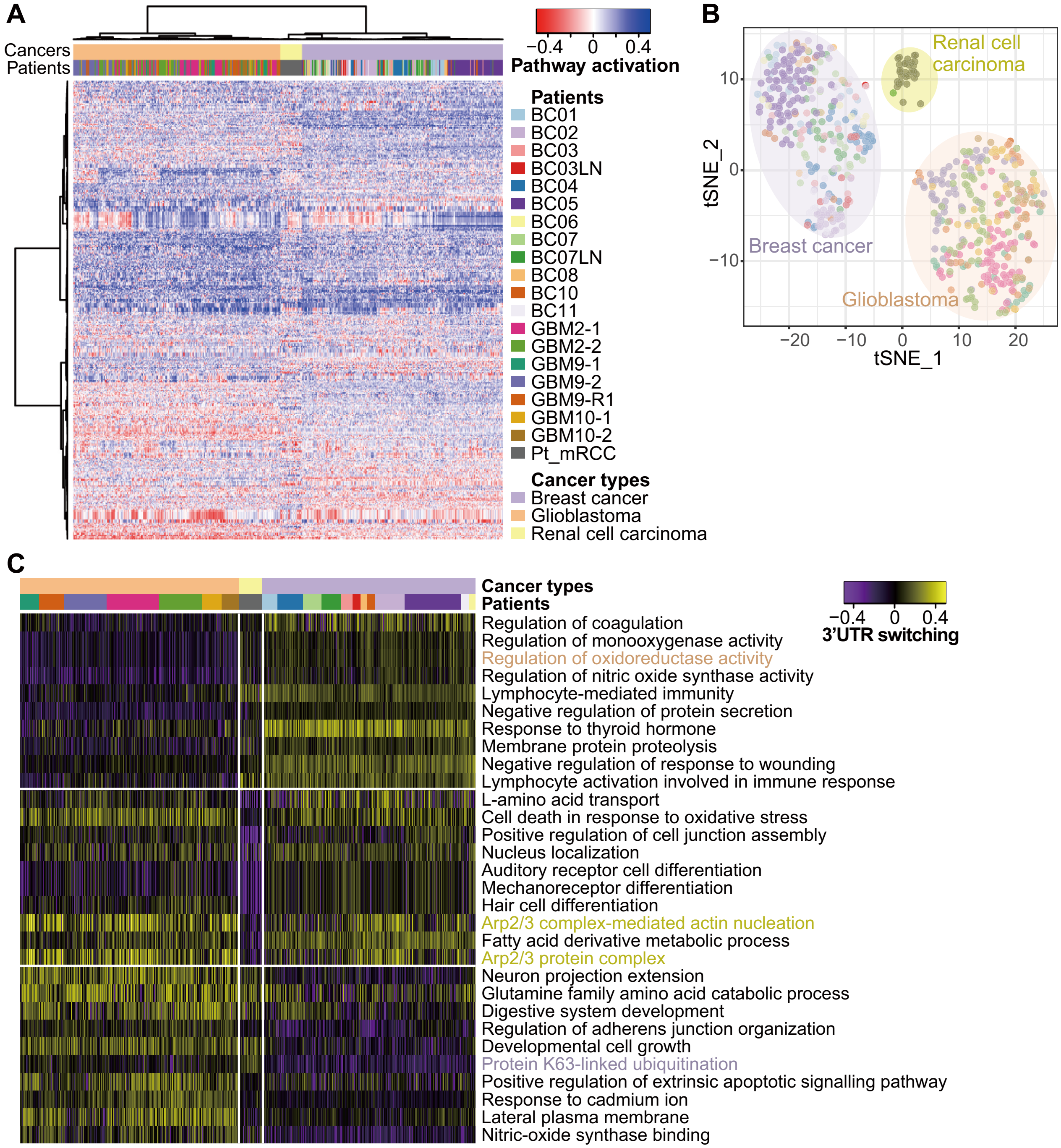
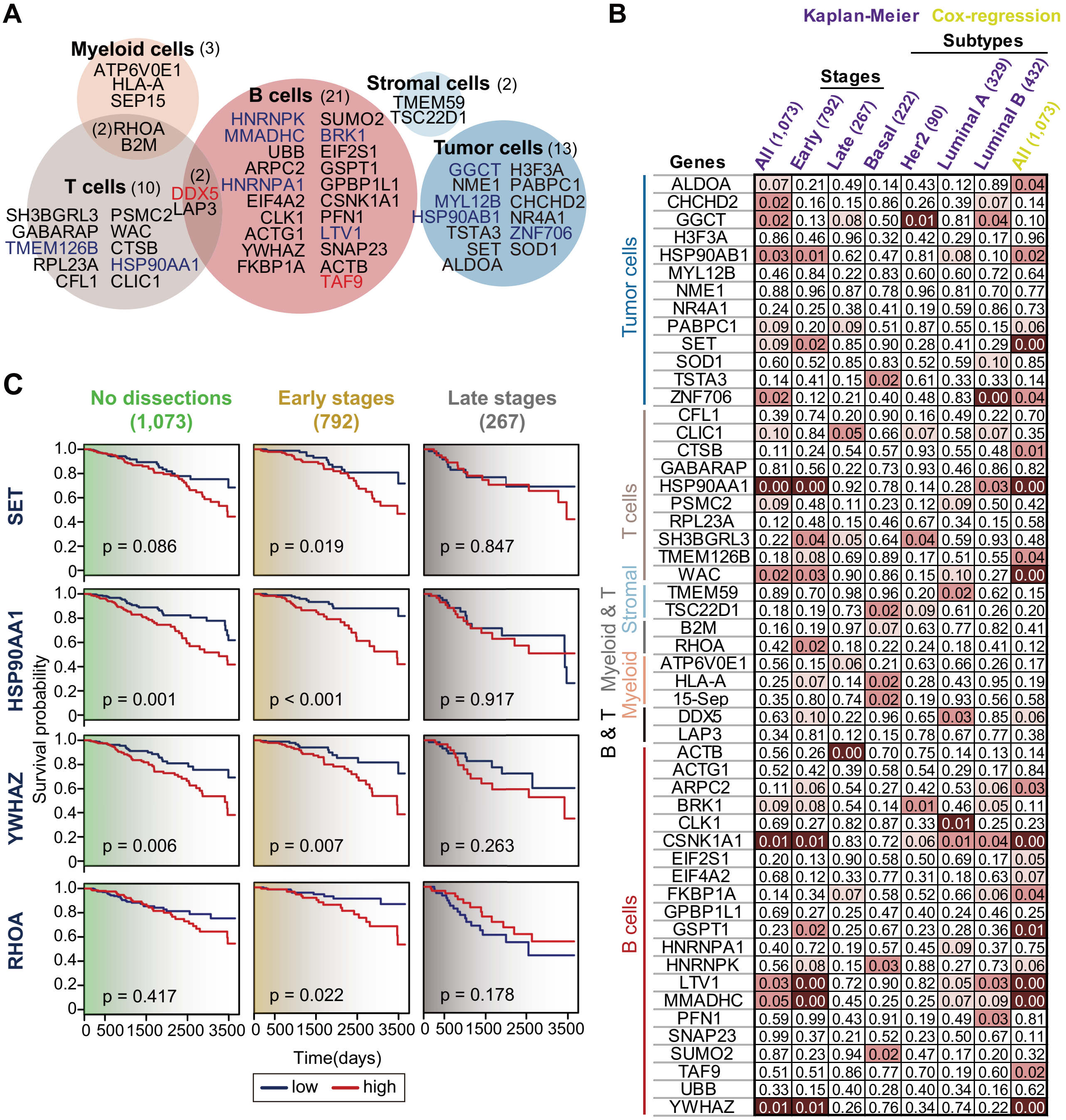


Fig 4. Heterogeneity of APA regulation among cancer types. (A) Hierarchical clustering using gene set-level APA estimates on 598 single-cells derived from 3 cancer types (280 tumor cells from 11 BC patients, 288 tumor cells from 3 GBM patients, and 30 tumor cells from 1 RCC patients). BC; Breast cancer, GBM; Glioblastoma, RCC; Renal cell carcinoma, T1,T2: Multiple regions in the brain, RLPS; Relapse. A total of 417 gene sets were used as variable components for hierarchical clustering and dimensional reduction (see Methods). (B) Unsupervised tSNE on the gene set-level APAs separating tumor cells along with cancer types. The dot colors represent patients which are the same as in Fig 4A. (C) APA map of top10 gene sets significantly (p<0.01) switched in each cancer type. The colors for cancer types and patients are the same as in Fig 4A.

https://doi.org/10.1371/journal.pone.0217196.g004

leveraged genetic features and polymorphism estimated from full-length transcriptomic data [38, 39]. Here we present a computational approach to explore transcriptional signatures



Genes	All (1,073)	Early (792)	Late (267)	Basal (222)	Her2 (90)	Luminal A (329)	Luminal B (432)	All (1,073)
ALDOA	0.07	0.21	0.49	0.14	0.43	0.12	0.89	0.04
CHCHD2	0.02	0.16	0.15	0.86	0.26	0.39	0.07	0.14
GGCT	0.02	0.13	0.08	0.50	0.01	0.81	0.04	0.10
H3F3A	0.86	0.46	0.96	0.32	0.42	0.29	0.17	0.96
HSP90AB1	0.03	0.01	0.62	0.47	0.81	0.08	0.10	0.02
MYL12B	0.46	0.84	0.22	0.83	0.60	0.60	0.72	0.64
NME1	0.88	0.96	0.87	0.78	0.96	0.81	0.70	0.77
NR4A1	0.24	0.25	0.38	0.41	0.19	0.59	0.86	0.73
PABPC1	0.09	0.20	0.09	0.51	0.87	0.55	0.15	0.06
SET	0.09	0.02	0.85	0.90	0.28	0.41	0.29	0.00
SOD1	0.60	0.52	0.85	0.83	0.52	0.59	0.10	0.85
TSTA3	0.14	0.41	0.15	0.02	0.61	0.33	0.33	0.14
ZNF706	0.02	0.12	0.21	0.40	0.48	0.83	0.00	0.04
CFL1	0.39	0.74	0.20	0.90	0.16	0.49	0.22	0.70
CLIC1	0.10	0.84	0.05	0.66	0.07	0.58	0.07	0.35
CTSB	0.11	0.24	0.54	0.57	0.93	0.55	0.48	0.01
GABARAP	0.81	0.56	0.22	0.73	0.93	0.46	0.86	0.82
HSP90AA1	0.00	0.00	0.92	0.78	0.14	0.28	0.03	0.00
PSMC2	0.09	0.48	0.11	0.23	0.12	0.09	0.50	0.42
RPL23A	0.12	0.48	0.15	0.46	0.67	0.34	0.15	0.58
SH3BGRL3	0.22	0.04	0.05	0.64	0.04	0.59	0.93	0.48
TMEM126B	0.18	0.08	0.69	0.89	0.17	0.51	0.55	0.04
WAC	0.02	0.03	0.90	0.86	0.15	0.10	0.27	0.00
TMEM59	0.89	0.70	0.98	0.96	0.20	0.02	0.62	0.15
TSC22D1	0.18	0.19	0.73	0.02	0.09	0.61	0.26	0.20
B2M	0.16	0.19	0.97	0.07	0.63	0.77	0.82	0.41
RHOA	0.42	0.02	0.18	0.22	0.24	0.18	0.41	0.12
ATP6V0E1	0.56	0.15	0.06	0.21	0.63	0.66	0.26	0.17
HLA-A	0.25	0.07	0.14	0.02	0.28	0.43	0.95	0.19
15-Sep	0.35	0.80	0.74	0.02	0.19	0.93	0.56	0.58
DDX5	0.63	0.10	0.22	0.96	0.65	0.03	0.85	0.06
LAP3	0.34	0.81	0.12	0.15	0.78	0.67	0.77	0.38
ACTB	0.56	0.26	0.00	0.70	0.75	0.14	0.13	0.14
ACTG1	0.52	0.42	0.39	0.58	0.54	0.29	0.17	0.84
ARPC2	0.11	0.06	0.54	0.27	0.42	0.53	0.06	0.03
BRK1	0.09	0.08	0.54	0.14	0.01	0.46	0.05	0.11
CLK1	0.69	0.27	0.82	0.87	0.33	0.01	0.25	0.23
CSNK1A1	0.01	0.01	0.83	0.72	0.06	0.01	0.04	0.00
EIF2S1	0.20	0.13	0.90	0.58	0.50	0.69	0.17	0.05
EIF4A2	0.68	0.12	0.33	0.77	0.31	0.18	0.63	0.07
FKBP1A	0.14	0.34	0.07	0.58	0.52	0.66	0.06	0.04
GPBP1L1	0.69	0.27	0.25	0.47	0.40	0.24	0.46	0.25
GSPT1	0.23	0.02	0.25	0.67	0.23	0.28	0.36	0.01
HNRNPA1	0.40	0.72	0.19	0.57	0.45	0.09	0.37	0.75
HNRNPK	0.56	0.08	0.15	0.03	0.88	0.27	0.73	0.06
LTV1	0.03	0.00	0.72	0.90	0.82	0.05	0.03	0.00
MMADHC	0.05	0.00	0.45	0.25	0.25	0.07	0.09	0.00
PFN1	0.59	0.99	0.43	0.91	0.19	0.49	0.03	0.81
SNAP23	0.99	0.37	0.21	0.52	0.23	0.50	0.67	0.11
SUMO2	0.87	0.23	0.94	0.02	0.47	0.17	0.20	0.32
TAF9	0.51	0.51	0.86	0.77	0.70	0.19	0.60	0.02
UBB	0.33	0.15	0.40	0.28	0.40	0.34	0.16	0.62
YWHAZ	0.01	0.01	0.26	0.76	0.34	0.74	0.22	0.00

Fig 5. Cell type specific genes with clinical associations. (A) Venn diagram for 53 gene hits specific to 5 cell types. Red-colored genes were concurrently identified in the results of Roar. Blue-colored genes are de novo markers not predefined in a public APA database. Number in the bracket indicates the number of gene hits in each cell type category. (B) Kaplan-Meier and Cox regression analyses for 53 genes in TCGA BRCA patients. Variables showing a significant difference in survival rate are highlighted in symbols (light red; p<0.01, red; p<0.05, and dark red; p<0.1). Number in the bracket indicates the number of tissue samples in each subset. (C) Kaplan-Meier survival plot for *SET*, *HSP90AA1*, *YWHAZ*, and *RHOA* in TCGA BRCA patients. The tumor samples were annotated as 'high' and 'low' (25th and 75th percentiles, respectively) groups for the expression signal of each gene. The p-value was determined by the Log-Rank test. The tumor tissue samples were categorized into early and late cancer stages (Stage I and II / III and IV) based on the 'pathologic_stage' information in the clinical dataset. Number in the bracket indicates the number of tissue samples in each subset.

https://doi.org/10.1371/journal.pone.0217196.g005

associated with alternative polyadenylations in 3'UTR in combination with gene expression regulations at single cell resolution. The single cell approach solves the complexity issues coming from bulk RNA sequencing which contains tumor, stromal, and immune components of tumor tissues. We exploited only high-depth and full-length scRNAseq, as massively parallel 3' or 5' scRNAseq data have sparse read coverage and densities which prevents an accurate polyadenylation analysis.

We recognize that even full-length scRNA-seq data has a limitation in assessing the full spectrum of 3'UTR switching for all genes within a cell. When we applied gene set level analysis, we could find transcriptional trends associated with APAs which largely separated tumor and non-tumor cells. For this, we combined complementary metrics reflecting cell-type labels inferred from gene expression data, gene expression levels, and change of 3'UTR lengths. The enrichment trends demonstrated discriminative biologic functions specific to cell types of tumor vs. their neighboring immune and stromal cells. Especially, signals of APA and gene expression for functional categories of 'the regulation of cell proliferation' were strikingly different among cell types (Fig 3), which confirmed the connection between APA and cellular proliferation in cancer and the immune system [2, 40].

Previous work suggested that a 3'UTR-based classifier could improve prognostic performance in triple-negative breast cancer patients [41]. To search for clinically relevant signatures in APA, we first selected 53 genes showing 3'UTR shortening and over-expression specific to each cell type (Fig 5A). Among these cell type-specific signatures, Kaplan-Meier and Cox regression analyses confirmed outstanding risk stratification by 11 genes for patients with BRCA tumors (Fig 5B and 5C). Although their 3'UTR shortening is clearly represented at the single-cell level (S2 Fig), bulk RNA-sequencing in TCGA BRCA failed to demonstrate the association of the change of APA signals and expression level of those genes (S3 Fig). As bulk RNA-sequencing represents the average of all cell categories in tumor tissues, the difference in read density between tumor and normal samples might have been mixed and burred. In addition, we cannot rule out the gene expression was controlled by other mechanisms such as nucleosome positioning, DNA binding regulatory proteins, and histone modifications. Nonetheless, profiling transcriptional trends by APAs combined with gene expression at single-cell resolution, provides a unique strategy to identify potential prognostic signatures.

Conclusions

We suggest a computational strategy to profile transcriptional regulation by alternative 3'UTR changes among individual cells, together with gene expression. Our study represents the first analysis of APA variations in combination with gene expression using scRNA-seq data and highlights the utility of APA-associated signatures for improving prognostic performance.

Supporting information

S1 Fig. Different APA patterns between tumor and non-tumor cells defined by Roar. Hierarchical clustering was performed using APA estimates (roar value calculated by Roar) (a) for genes and (b) gene sets on individual cells derived from breast cancer patients. Only 461 cells with at least 983 (10% of total genes) detected genes were selected for further analysis. A total of 3,262 genes and 4,754 gene sets showed the signals of APA on 461 cells. Among theme, 922 genes and 808 gene sets were selected as variable components for hierarchical clustering (see Methods).
(EPS)

S2 Fig. Example of read density and corresponding 3'UTR shortening. Sashimi plot produced with Integrative Genomics Viewer (IGV) of two alignments for representative single-cell and Bodymap samples for tumor and normal breast. We selected *SET*, *HSP90AA1*, *YWHAZ*, and *RHOA* for presentation, those genes with the strongest shortening specific to each cell type in breast cancer.
(EPS)

S3 Fig. Comparision between APA signals and gene expression in TCGA BRCA. No changes in APA signals along the expression level of target genes in BRCA tissue samples. The APA was predicted by DaPars algorithm from the TCGA BRCA versus matched normal breast tissue samples.
(EPS)

S1 Table. Overview of the collected single-cell RNA sequencing dataset.
(XLSX)

S2 Table. List of 1176 gene sets with significantly altered APA and concurrent expression specific for five cell types.
(XLSX)

S3 Table. Full list of significantly switched gene sets in each cancer type.
(XLSX)

S4 Table. Correlation and odds ratio results for 53 cell type-specific gene signatures.
(XLSX)

Acknowledgments

We thank research fellows in the Samsung Genome Institute for valuable comments on our work.

Author Contributions

Conceptualization: Hae-Ock Lee, Woong-Yang Park.

Data curation: Woosung Chung.

Formal analysis: Nayoung Kim.

Investigation: Nayoung Kim.

Methodology: Nayoung Kim.

Resources: Hye Hyeon Eum.

Supervision: Hae-Ock Lee.

Visualization: Nayoung Kim.

Writing – original draft: Nayoung Kim.

Writing – review & editing: Hae-Ock Lee, Woong-Yang Park.

References

1. Di Giammartino DC, Nishida K, Manley JL. Mechanisms and consequences of alternative polyadenylation. Molecular cell. 2011; 43(6):853–66. https://doi.org/10.1016/j.molcel.2011.08.017 PMID: 21925375; PubMed Central PMCID: PMC3194005.

2. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science. 2008; 320(5883):1643–7. https://doi.org/10.1126/science.1155390 PMID: 18566288; PubMed Central PMCID: PMC2587246.
3. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009; 138(4):673–84. https://doi.org/10.1016/j.cell.2009.06.016 PMID: 19703394; PubMed Central PMCID: PMC2819821.
4. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nature communications. 2014; 5:5274. https://doi.org/10.1038/ncomms6274 PMID: 25409906; PubMed Central PMCID: PMC4467577.
5. Begik O, Oyken M, Cinkilli Alican T, Can T, Erson-Bensan AE. Alternative Polyadenylation Patterns for Novel Gene Discovery and Classification in Cancer. Neoplasia. 2017; 19(7):574–82. https://doi.org/10.1016/j.neo.2017.04.008 PMID: 28624626; PubMed Central PMCID: PMC5476451.
6. Gillen AE, Brechbuhl HM, Yamamoto TM, Kline E, Pillai MM, Hesselberth JR, et al. Alternative Polyadenylation of PRELID1 Regulates Mitochondrial ROS Signaling and Cancer Outcomes. Molecular cancer research: MCR. 2017; 15(12):1741–51. https://doi.org/10.1158/1541-7786.MCR-17-0010 PMID: 28912168; PubMed Central PMCID: PMC5791535.
7. Akman HB, Oyken M, Tuncer T, Can T, Erson-Bensan AE. 3'UTR shortening and EGF signaling: implications for breast cancer. Human molecular genetics. 2015; 24(24):6910–20. https://doi.org/10.1093/hmg/ddv391 PMID: 26395459.
8. Le Pera L, Mazzapioda M, Tramontano A. 3USS: a web server for detecting alternative 3'UTRs from RNA-seq experiments. Bioinformatics. 2015; 31(11):1845–7. https://doi.org/10.1093/bioinformatics/btv035 PMID: 25617413; PubMed Central PMCID: PMC4443675.
9. Grassi E, Mariella E, Lembo A, Molineris I, Provero P. Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. BMC bioinformatics. 2016; 17(1):423. https://doi.org/10.1186/s12859-016-1254-8 PMID: 27756200; PubMed Central PMCID: PMC5069797.
10. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods. 2010; 7(12):1009–15. https://doi.org/10.1038/nmeth.1528 PMID: 21057496; PubMed Central PMCID: PMC3037023.
11. Wang W, Wei Z, Li H. A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. Bioinformatics. 2014; 30(15):2162–70. https://doi.org/10.1093/bioinformatics/btu189 PMID: 24728858; PubMed Central PMCID: PMC4103598.
12. Velten L, Anders S, Pekowska A, Jarvelin AI, Huber W, Pelechano V, et al. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. Molecular systems biology. 2015; 11(6):812. https://doi.org/10.15252/msb.20156198 PMID: 26040288; PubMed Central PMCID: PMC4501847.
13. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T. JingleBells: A Repository of Immune-Related Single-Cell RNA-Sequencing Datasets. Journal of immunology. 2017; 198(9):3375–9. https://doi.org/10.4049/jimmunol.1700272 PMID: 28416714.
14. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nature communications. 2017; 8:15081. https://doi.org/10.1038/ncomms15081 PMID: 28474673; PubMed Central PMCID: PMC5424158.
15. Kim KT, Lee HW, Lee HO, Song HJ, Jeong da E, Shin S, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome biology. 2016; 17:80. https://doi.org/10.1186/s13059-016-0945-9 PMID: 27139883; PubMed Central PMCID: PMC4852434.
16. Lee JK, Wang J, Sa JK, Ladewig E, Lee HO, Lee IH, et al. Spatiotemporal genomic architecture informs precision oncology in glioblastoma. Nature genetics. 2017; 49(4):594–9. https://doi.org/10.1038/ng.3806 PMID: 28263318; PubMed Central PMCID: PMC5627771.
17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886; PubMed Central PMCID: PMC3530905.
18. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics. 2014; 47:11 2 1–34. https://doi.org/10.1002/0471250953.bi1112s47 PMID: 25199790; PubMed Central PMCID: PMC4213956.
19. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011; 12:323. https://doi.org/10.1186/1471-2105-12-323 PMID: 21816040; PubMed Central PMCID: PMC3163565.
20. Lee JY, Yeh I, Park JY, Tian B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. Nucleic acids research. 2007; 35(Database issue):D165–8. https://doi.org/10.1093/nar/gkl870 PMID: 17202160; PubMed Central PMCID: PMC1899096.

21. Agresti A. Categorical data analysis. 2nd ed. ed. New York; Chichester: Wiley-Interscience; 2002.

22. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13 (11):2498–504. https://doi.org/10.1101/gr.1239303 PMID: 14597658; PubMed Central PMCID: PMC403769.

23. Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome research. 2011; 21(5):741–7. https://doi.org/10.1101/gr.115295.110 PMID: 21474764; PubMed Central PMCID: PMC3083091.

24. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221):470–6. https://doi.org/10.1038/nature07509 PMID: 18978772; PubMed Central PMCID: PMC2593745.

25. Valvona CJ, Fillmore HL, Nunn PB, Pilkington GJ. The Regulation and Function of Lactate Dehydrogenase A: Therapeutic Potential in Brain Tumor. Brain pathology. 2016; 26(1):3–17. https://doi.org/10.1111/bpa.12299 PMID: 26269128.

26. Okamura T, Kurisu K, Yamamoto W, Takano H, Nishiyama M. NADPH/quinone oxidoreductase is a priority target of glioblastoma chemotherapy. International journal of oncology. 2000; 16(2):295–303. PMID: 10639573.

27. Debald M, Schildberg FA, Linke A, Walgenbach K, Kuhn W, Hartmann G, et al. Specific expression of k63-linked ubiquitination of calmodulin-like protein 5 in breast cancer of premenopausal patients. Journal of cancer research and clinical oncology. 2013; 139(12):2125–32. https://doi.org/10.1007/s00432-013-1541-y PMID: 24146193.

28. Ilatovskaya DV, Chubinskiy-Nadezhdin V, Pavlov TS, Shuyskiy LS, Tomilin V, Palygin O, et al. Arp2/3 complex inhibitors adversely affect actin cytoskeleton remodeling in the cultured murine kidney collecting duct M-1 cells. Cell and tissue research. 2013; 354(3):783–92. https://doi.org/10.1007/s00441-013-1710-y PMID: 24036843; PubMed Central PMCID: PMC3850072.

29. Janghorban M, Farrell AS, Allen-Petersen BL, Pelz C, Daniel CJ, Oddo J, et al. Targeting c-MYC by antagonizing PP2A inhibitors in breast cancer. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(25):9157–62. https://doi.org/10.1073/pnas.1317630111 PMID: 24927563; PubMed Central PMCID: PMC4078832.

30. Li J, Yang XF, Ren XH, Meng XJ, Huang HY, Zhao QH, et al. Stable SET knockdown in breast cell carcinoma inhibits cell migration and invasion. Biochemical and biophysical research communications. 2014; 453(1):7–12. https://doi.org/10.1016/j.bbrc.2014.09.013 PMID: 25234598.

31. Zhang W, Sommers CL, Burshtyn DN, Stebbins CC, DeJarnette JB, Trible RP, et al. Essential role of LAT in T cell development. Immunity. 1999; 10(3):323–32. PMID: 10204488.

32. Aguado E, Martinez-Florensa M, Aparicio P. Activation of T lymphocytes and the role of the adapter LAT. Transplant immunology. 2006; 17(1):23–6. https://doi.org/10.1016/j.trim.2006.09.013 PMID: 17157209.

33. Hayashi K, Kamikawa Y. HSP90 is crucial for regulation of LAT expression in activated T cells. Molecular immunology. 2011; 48(6–7):941–6. https://doi.org/10.1016/j.molimm.2010.12.014 PMID: 21251717.

34. Mbofung RM, McKenzie JA, Malu S, Zhang M, Peng W, Liu C, et al. HSP90 inhibition enhances cancer immunotherapy by upregulating interferon response genes. Nature communications. 2017; 8(1):451. https://doi.org/10.1038/s41467-017-00449-z PMID: 28878208; PubMed Central PMCID: PMC5587668.

35. Chimen M, McGettrick HM, Apta B, Kuravi SJ, Yates CM, Kennedy A, et al. Homeostatic regulation of T cell trafficking by a B cell-derived peptide is impaired in autoimmune and chronic inflammatory disease. Nature medicine. 2015; 21(5):467–75. https://doi.org/10.1038/nm.3842 PMID: 25894827; PubMed Central PMCID: PMC4425550.

36. Manresa-Arraut A, Johansen FF, Brakebusch C, Issazadeh-Navikas S, Hasseldam H. RhoA Drives T-Cell Activation and Encephalitogenic Potential in an Animal Model of Multiple Sclerosis. Frontiers in immunology. 2018; 9:1235. https://doi.org/10.3389/fimmu.2018.01235 PMID: 29904389; PubMed Central PMCID: PMC5990621.

37. Liu Y, Kubiak JZ, Li XC, Ghobrial RM, Kloc M. Macrophages and RhoA Pathway in Transplanted Organs. Results and problems in cell differentiation. 2017; 62:365–76. https://doi.org/10.1007/978-3-319-54090-0_15 PMID: 28455717.

38. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344(6190):1396–401. https://doi.org/10.1126/science.1254257 PMID: 24925914; PubMed Central PMCID: PMC4123637.

39. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016; 352(6282):189–96. https://doi.org/10.1126/science.aad0501 PMID: 27124452; PubMed Central PMCID: PMC4944528.

40. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nature reviews Molecular cell biology. 2017; 18(1):18–30. https://doi.org/10.1038/nrm.2016.116 PMID: 27677860; PubMed Central PMCID: PMC5483950.

41. Wang L, Hu X, Wang P, Shao ZM. The 3'UTR signature defines a highly metastatic subgroup of triple-negative breast cancer. Oncotarget. 2016; 7(37):59834–44. https://doi.org/10.18632/oncotarget.10975 PMID: 27494850; PubMed Central PMCID: PMC5312352.