Note: I just built the pipeline using a public dataset (N=1, same as the data from Hung lab) so that we can get the results as soon as the sequencing data comes out.

Methods

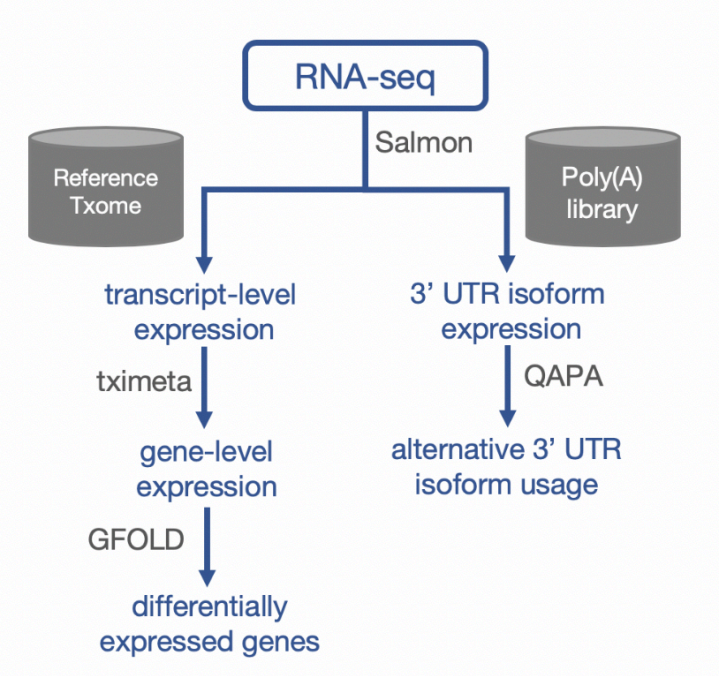

Figure 1. Workflow of the analysis. The input data is RNA-seq of cancer cells. Transcript expression levels are estimated by Salmon with reference transcriptome, gene-level values are aggregated by tximport. Differentially expressed genes are then identified by GFOLD. For 3' UTR isoforms, expression levels are estimated by Salmon using poly(A) site annotations as reference. QAPA is then used to estimate the usage of poly(A) sites.

Heck and neck cancer cell line

RNA-seq data of two samples of human head and neck cancer UMSCC47 cell line (GSE98805) treated with dimethyl sulfoxide (DMSO) and Aza/TSA, drugs that inhibit methylation, respectively were obtained from SRA (Sequence Read Archive).

Gene expression estimation

Single-end sequencing reads for the UMSCC47 cell line were downloaded from SRA. The FASTQ files of reads were processed by Salmon[1] (v0.15.0) under mapping-based mode using selective alignment with the reference transcriptome (*Homo sapiens*, GRCh38, Ensembl release 98).

Differential expression analysis

Transcript-level abundance estimates were summarized into gene-level values using tximeta[2] (v1.4.2). The gene-level read counts were put into GFOLD[3] (v1.1.4) to obtain generalized fold change for ranking differentially expressed genes. Genes with |GFOLD| > 0.1 were considered as differentially expressed.

3’ UTR isoform analysis

For 3’ UTR isoforms, a pre-compiled Poly(A) library for human (GRCh38) was first built using biomart[4] (release 98) gene metadata, GENCODE[5] gene prediction annotation, PolyASite database V2[6], and GENCODE poly(A) sites track. 3’ UTR expression levels were then estimated using Salmon[1] (v0.15.0) with the Poly(A) library as reference. The relative proportion of each isoform in a gene, measured as Poly(A) Usage (PAU) was calculated by QAPA[7] (Quantification of Alternative Polyadenylation, v1.3.0).

Genes that had a total expression value lower than 3 TPM in both samples were removed. To avoid overlapping non-strand specific RNA-seq reads, gene pairs whose distal 3’ UTRs had 3’ ends that were within 500 nt of each other were excluded. Genes with alternative UTR lengths of less than 100 nt were also excluded to reduce potentially noisy estimates.

Pathway enrichment analysis

Gene Ontology biological process (GOBP) terms were tested for selected genes under different thresholds, for example shortening genes (genes with Δ proximal PAU < -20) or up-regulated genes (GFOLD > 0.1), using goseq[8] (v1.38.0). Gene length data were processed from a GTF file (*Homo sapiens,* GRCh38, Ensembl release 98) and were manually provided to goseq for length bias correction.

MicroRNA target sites / 3’ UTR motif enrichment

To be done in the future.

Results

Genes with altered UTR lengths

Since APA is often regulated through the differential use of proximal poly(A) sites[9], the PAU of the proximal 3’ UTR isoform (denoted as PPAU) is calculated in order to assign different types of genes: 3’ UTR lengthening (ΔPPAU > 20), 3’ UTR shortening (ΔPPAU < –20), and no-change negative controls (–20 < ΔPPAU < 20).

By this definition, 238 (5.0%) and 347 (7.4%) genes lengthened and shortened respectively, while 4134 (87.6%) genes did not show a change in UTR length, indicating a global shortening of 3’ UTR. (Fig. 2A).

To explore the properties of 3’ UTR that lengthen, shorten, or don’t change, the lengths of the longest alternative UTR (aUTR) regions, i.e. the length of isoform with the most distal poly(A) site minus the length of isoform with the proximal poly(A) site, were compared (Fig. 2B). Kolmogorov-Smirnov tests suggested that shortening 3’ UTRs have significantly shorter aUTR lengths than those in the other two groups ($p < 0.05$ both, data not shown).

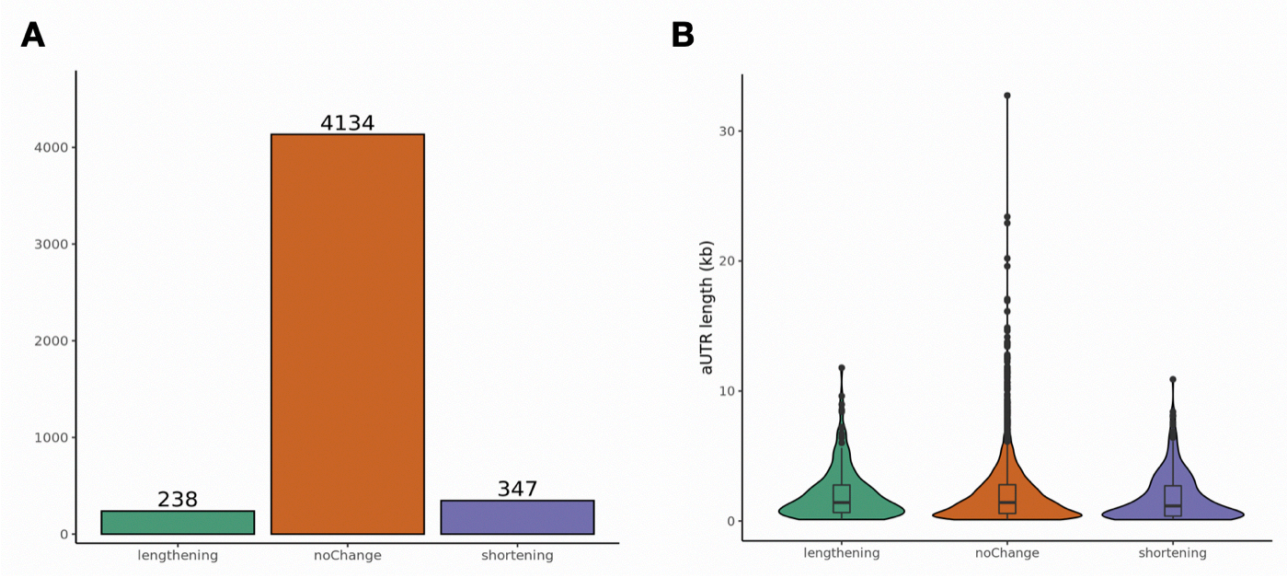

Figure 2. (A) Bar plot indicates the number of 3′ UTRs that lengthen (ΔPPAU > 20), shorten (ΔPPAU < −20), and do not change (|ΔPPAU| ≤ 20) where ΔPPAU is defined as the difference in PPAU between samples under two conditions. (B) Violin plot comparing the lengths of alternative 3′ UTR (aUTR) regions in lengthening, shortening, and non-changing 3′ UTRs.

Here we show the QAPA results of several genes relevant to immune checkpoint therapy (Table 1). *CD274* (PD-L1) has three alternative 3' UTR isoforms, and the result indicates that it was shortening after Aza/TSA treatment; however, the expression levels were too low to draw a strong conclusion. Both *PDCD1* (PD-1) and *CTLA4* have only one single isoform and no read was detected. This may be due to low sequencing depths.

Table 1. Alternative 3' UTR usage in selected genes relative to immune checkpoint therapy. Sample

Gene	APA isoform	UTR length	Sample 1 PAU	Sample 2 PAU	Sample 1 TPM	Sample 2 TPM
CD274	Proximal APA	619	62.953	30.516	0.329516	0.128644
CD274	Distal APA 1	2629	37.047	0	0.193917	0
CD274	Distal APA 2	2691	0	69.484	0	0.292919
PDCD1	Single APA	1177	NA	NA	0	0
CTLA4	Single APA	1153	NA	NA	0	0

1 – SRR5528013 (DMSO); sample 2 – SRR5528014 (Aza/TSA). TPM is transcripts per million.

Differentially expressed genes

GFOLD (generalized fold change) algorithm is especially useful when no biological replicate available. GFOLD generalizes the fold change by considering the posterior distribution of log fold change, such that each gene is assigned a reliable fold change.

The most up-regulated gene, *PPP1R10*, plays a role in many cellular processes including cell cycle progression, DNA repair, and apoptosis by regulating the activity of protein phosphatase 1. *GATAD2B* encodes a zinc finger protein transcriptional

repressor that is part of the methyl-CpG-binding protein-1 complex, which represses gene expression by deacetylating methylated nucleosomes (Table 2).

Table 2. (**A**) Table of top five down-regulated genes. (**B**) Table of top five up-regulated genes. Genes were ranked by |GFOLD| representing the variance information of the posterior distribution of fold change. Log2FC denotes log2 fold change.

A

Gene	GFOLD	log2FC
RPL39P3	-6.55	-9.28
FMC1-LUC7L2	-5.96	-8.70
AC009412.1	-5.74	-8.49
HSD17B8	-5.56	-8.30
ZNF623	-5.21	-6.96

B

Gene	GFOLD	log2FC
PPP1R10	6.65	9.39
GATAD2B	5.60	7.21
MARF1	5.36	8.10
AC006064.6	4.98	7.72
AC138969.2	4.74	7.49

The mRNA expression changes and APA changes were also compared (Fig. 3). 43 genes with shortening 3’ UTR displayed a significant increase in expression, but the trend was not so obvious compared with other situations, probably because the sample size used in this study was too small.

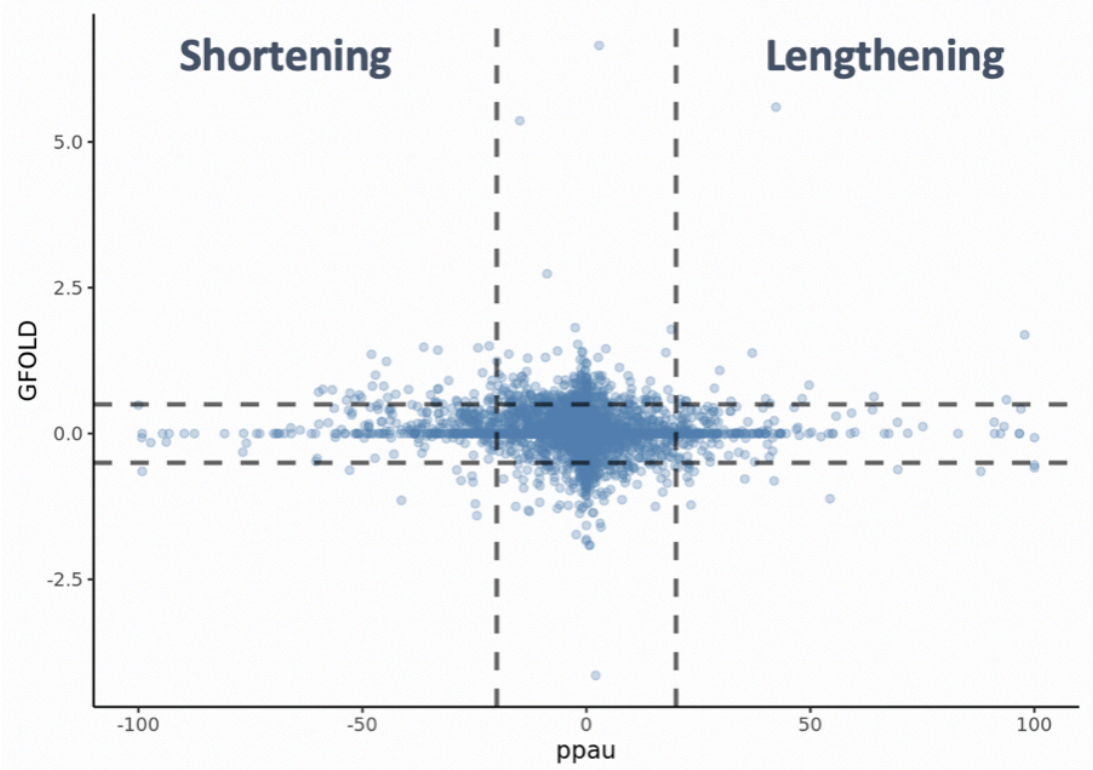

Figure 3. Comparison between mRNA expression changes (y-axis) and APA changes (x-axis). Lengthening 3′ UTRs are indicated on the right (ΔPPAU > 20), while shortening 3′ UTRs are on the left (ΔPPAU < – 20). Genes with significant differential up- or down-regulation are on the up and bottom, respectively (|GFOLD| > 0.5). Dashed horizontal lines indicate GFOLD thresholds, while dashed vertical lines indicate ΔPPAU thresholds.

Enriched pathways

Functional enrichment analysis was performed to identify enriched pathways on genes that were differentially expressed (Fig. 4) and on genes showing altered 3' UTR (Fig. 5).

The up-regulated genes are associated with metabolic processes, while terms related to protein targeting to membrane are significant among down-regulated genes. Double-strand break repair and chromatin silencing are enriched in lengthening genes, and interestingly, processes about methylation are enriched in shortening genes, which means the effects of drug Aza/TSA exerted through alternative polyadenylation (APA) events. This is consistent with the findings in cancers, stating that APA events could affect drug sensitivity, especially of drugs targeting chromatin modifiers[10].

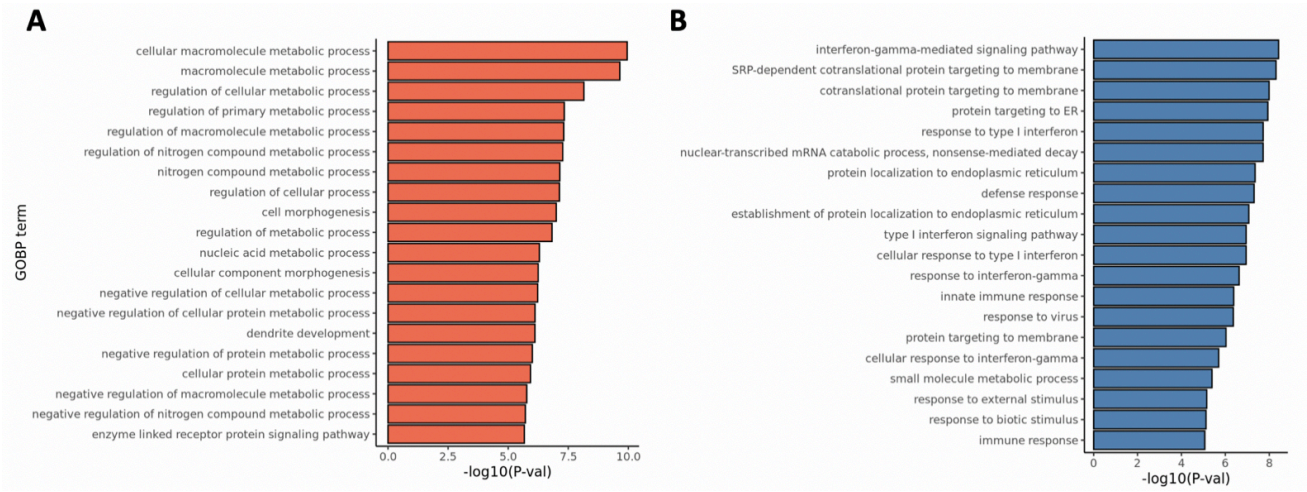

Figure 4. (**A**) Top 20 GO terms enriched in up-regulated genes. (**B**) Top 20 GO terms enriched in down-regulated genes.

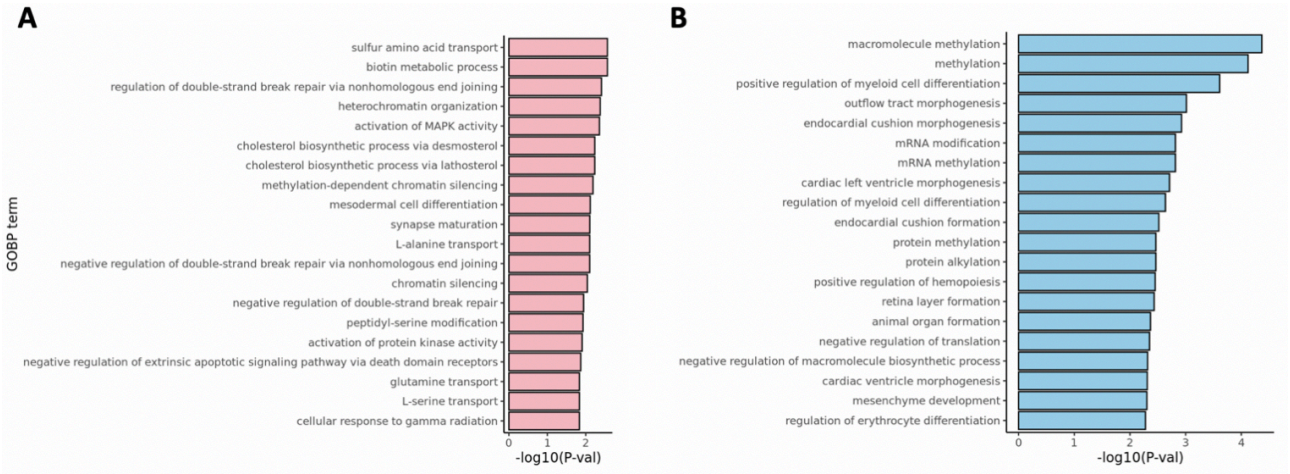

Figure 5. (**A**) Top 20 GO terms enriched in lengthening genes. (**B**) Top 20 GO terms enriched in shortening genes.

References

1. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* (2017). doi:10.1038/NMETH.4197
2. Love, M. I. *et al.* Tximeta: reference sequence checksums for provenance identification in RNA-seq. *bioRxiv* (2019). doi:10.1101/777888
3. Feng, J. *et al.* GFOLD: A generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts515
4. Smedley, D. *et al.* The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv350
5. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky955
6. Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz918
7. Ha, K. C. H., Blencowe, B. J. & Morris, Q. QAPA: A new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.* (2018). doi:10.1186/s13059-018-1414-4
8. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* (2010). doi:10.1186/gb-2010-11-2-r14
9. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* (2013). doi:10.1101/gad.229328.113
10. Xiang, Y. *et al.* Comprehensive characterization of alternative polyadenylation in human cancer. *J. Natl. Cancer Inst.* (2018). doi:10.1093/jnci/djx223