Supplemental Online Material

Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression

Steve Lianoglou, Vidur Garg, Julie L. Yang, Christina S. Leslie, Christine Mayr
Supplemental Online Methods

3'-seq computational analyses

Read alignment. The reads from each experiment were preprocessed to remove low-quality 3' ends (Phred score < 30), 3' adapters and 3' poly(A) runs. These reads were aligned to the human genome (hg19) using BWA (version 0.5.9) resulting in 145,323,294 aligned reads (97,936,320 uniquely mapping). All aligned reads were pooled together in order to identify peaks (corresponding to cleavage events) across the genome. Peaks were identified by convolving the read coverage signal with the second derivative of a Gaussian filter. The locations where the convolved signal crosses zero correspond to rising and falling edges in the original signal, and these are used as boundaries for the peaks. Peaks with fewer than five supporting reads were removed. Each peak was annotated with information about the genomic locus where it landed, i.e. the gene name and location (intron, CDS, 3'UTR, 5'UTR). Peaks landing in intergenic regions further than 5 kb downstream and 1 kb upstream from annotated genes were removed from downstream analysis in this study. The expression of each peak in each experiment was quantified by counting the number of uniquely mapping reads landing inside peak boundaries within each experiment and further normalized by sequencing depth (counts per million).

Removing internally primed peaks. Internally primed peaks were flagged by locating an A-rich stretch (5/6 A's) in a 22 nt window surrounding the cleavage event (10 nt upstream and 12 nt downstream). If no known pA signals (Tian et al. 2005) were found within 10-45 nt upstream from the cleavage event, or if the cleavage event was not located at the end of a previously annotated 3'UTR, these peaks were removed. When we compared the number and locations of the peaks obtained by 3'-seq to the location of the bands obtained by Northern blots we found concordance of the two methods in 94% of cases. In only one out of 136 samples tested the discrepancy observed was due to counting an internal priming event in the 3'-seq data which was not present in the Northern blot analysis. This shows that the frequency of false-positive events due to internal priming is low.

Removing low usage cleavage events. After removing artifacts, we identified 189,200 cleavage events that landed in known genic regions of the genome. A large majority of these peaks were low coverage events – 129,960 of these were never expressed above 3 TPM in any of the experiments where they were observed and likely represented either transcriptional or sequencing noise. In order to focus the analysis on robustly expressed ApA events, we
required that a cleavage event satisfy both a minimal expression threshold and a *usage* threshold corresponding to the fraction of transcripts for the gene ending at the event; that is, we wanted to exclude the situation where a weak and seldom used ApA signal generates an event that passes the expression threshold simply because the gene itself is highly expressed. We therefore kept cleavage events that accounted for ≥ 20% of the 3'UTR reads in ≥ 1 experiment. If an event was lower expressed but was expressed by > 5% of the 3'UTR reads in the majority of samples (≥ 75%) we also kept the event. These filtering steps reduced the atlas to 41,893 genic peaks. The remaining peaks were further clustered together if they were within 50 nt of each other, in order to focus on ApA events that substantially changed 3'UTR length, resulting in an atlas of 36,752 highly confident, functional cleavage and polyadenylation events. We removed converging genes if they were less than 1 kb apart and considered for our statistical analysis only the 13,354 RefSeq-annotated protein-coding genes we found expressed in our samples higher than 5.5 TPM, calculated from all the reads that fall into 3'UTRs. If a gene in the RefSeq database has several annotated mRNA transcripts, we only include the 3'UTR of the most 3' located terminal exon in our analyses.

**Identification of significant changes in observed 3'UTR isoform expression between samples.** We wanted to identify changes in 3'UTR isoform expression ratios between samples in a statistically rigorous manner that accounted for biological variability between replicates and was robust to changes in mRNA abundance between conditions. To do so, we repurposed a recently published method named DEXSeq that was originally developed to identify differential exon usage between conditions using RNA-seq data (Anders et al. 2012).

The reads aligning to each cleavage site are assumed to follow a negative binomial (NB) distribution, and a GLM is used to describe the systematic sources of variation as follows. The count of reads aligning to cleavage site \( l \) of gene \( i \) in sample \( j \) is modeled as \( NB(s_j \mu_{ijl}, \alpha_{il}) \), where \( s_j \) is a scaling factor to account for library sequencing depth, \( \alpha_{il} \) is the dispersion of the negative binomial distribution for this counting bin, and the scaling-corrected expected count \( \mu_{ijl} \) is decomposed into four factors:

\[
\log \mu_{ijl} = \beta_i^G + \beta_i^P + \beta^C_{il} + \beta^PC_{il},
\]

where:

- \( \beta_i^G \) represents the baseline expression of gene \( i \)
- \( \beta_i^P \) is the logarithm of the expected fraction of reads mapping to gene \( i \) and assigned to cleavage site \( l \) (up to an additive constant)
\[ \beta_{i\rho j}^C \] is the logarithm of the fold change in overall expression of gene \( i \) under the experimental condition of sample \( j \), denoted as \( \rho_j \).

\[ \beta_{i\rho j}^{PC} \] is the effect that condition \( \rho_j \) has on the fraction of reads assigned to cleavage site \( l \).

We use this model as a basis to test if the last term, \( \beta_{i\rho j}^{PC} \), is non-zero, which would suggest that the relative abundance of the cleavage site depends on the condition the gene is expressed in. To do so, we fit reduced and full versions of the previously described model for each gene on a cleavage-site-by-cleavage-site basis. Using the fits from the reduced and saturated models, we then perform a likelihood-ratio-test to obtain a p-value used to reject the null hypothesis that the value for the interaction coefficient for this cleavage site is zero. When the null is rejected, we assert that the condition the gene is expressed in has an effect on the relative expression of the given cleavage site under test.

The reduced model has the interaction coefficient between the cleavage site and the condition removed:

\[
\log \mu_{ijl} = \beta_i^C + \beta_l^P + \beta_{i\rho j}^C
\]

The full model includes a term for the interaction coefficient only for the cleavage site under test:

\[
\log \mu_{ijl} = \beta_i^C + \beta_l^P + \beta_{i\rho j}^C + \beta_{i\rho j}^{PC} \delta_{ll'}
\]

where \( \delta_{ll'} \) is the Kronecker delta function which evaluates to 1 only when \( l = l' \), otherwise it is 0.

**Applying the statistical framework in the absence of replicates.** Our GLM approach requires that the NB distribution is re-parameterized for each cleavage site with the appropriate value of its dispersion estimate given the mean expression of the cleavage event under test. The estimate of the dispersion as a function of the mean is normally made using replicates for each condition. However, in our tissue atlas, we did not have replicates for most tissues, so instead we used three pairs of biological replicate samples that we did have – from our isogenic transformation systems (including naïve B cells) – to estimate the dispersion as a function of the mean and assert that the same relationship would hold in the human tissue data.
Specifically, we estimated the mean-dispersion relationship by analyzing two pairs of replicate conditions against each other and calculated the fit for the dispersion function. We further computed an inflated dispersion function by using an estimated 0.55 regression quantile for the fit (Figure 1). We then used the remaining pair of replicates to check whether the estimated mean-dispersion function gave p-values for replicate-to-replicate variation that were distributed according to the null hypothesis (namely, p-values between replicates should be uniformly distributed). Each panel in our Figure (2) shows the results when plotting the p-value given from the GLM (y-axis) versus the expected uniform values of the p-value (x-axis) when using an inflated fit of the dispersion as it is estimated from different sets of replicate batches indicated in the panel title.

We chose the mean-dispersion function estimated from the third panel from the left as it best recapitulated the uniform distribution of p-values for the set of experiments (MCF10A, blue dots) that were held out from the data used for fitting (B cells and MCF10AR).

**Identifying significant changes in gene expression from 3'-seq data.** In addition to using 3'-seq as a means to identify cleavage sites and their expression, we can sum the reads from all of the cleavage sites located in the 3'UTR of a gene to obtain mRNA abundance levels. The data summarized in this way is then used as "normal" RNA-seq data and analyzed with DESeq to identify differentially expressed genes between conditions of interest (Anders and Huber...
Genes with evidence of differential expression at an FDR-adjusted $p < 0.05$ were used for downstream analysis.

**Visualization of the 3'UTR isoform expression pattern by SUI or UI.** The UI (UTR index) was defined as the fraction of reads mapping to a specific pA site out of all the reads mapping to the 3'UTR, whereas the SUI (short UTR index) defines the fraction of reads mapping to the *first* pA site out of all the reads mapping to the 3'UTR. Thus, the SUI is only used when we focus our investigation on the first functional pA site in the 3'UTR (Fig. 4, Fig. 5). To define differences in ApA isoform expression pattern we use the UI (Fig. 2, Fig. 3). In order to report the number of genes that show differential 3'UTR isoform expression between samples, we count every gene only once. If a gene has several pA sites we only report the UI of the pA site that shows the most significant difference in isoform expression. But 90% of the significant changes in ApA isoform expression occur at the first pA site. Using Northern blots to confirm the 3'UTR isoform expression pattern, we use the SUI because all the genes that were analyzed showed their change in 3'UTR isoform abundance at this site.

**pAM genes versus NpAM genes.** All the genes that showed a significant difference in 3'UTR isoform expression across the seven tissue samples (FDR-adjusted $p < 0.1$) were called pAM genes and the remaining multi-UTR genes were called NpAM genes. However, genes with FDR-adjusted $p$-values between 0.1 and 0.3 often showed a convincing difference in ApA isoform expression on the Northern blots (Supplemental Fig. 2C), therefore, for Fig. 4G and 4H only NpAM genes with very high FDR-adjusted $p > 0.6$) are shown. In the rest of the manuscript NpAM genes have FDR-adjusted $p > 0.1$.

**Conservation analysis**

The conservation of the 3'UTR regions surrounding the proximal pA sites of genes was determined using the phastCons scores calculated from 46 different species from the UCSC genome browser database. Calculation of conservation for the miRNA analysis used conservation in 5 species. The five-way species alignment (consisting of human hg19, mouse mm9, rat rn5, dog canFam3, and chicken galGal2) was extracted from the 46-way multiple alignment and a conservation sequence was computed such that a nucleotide was considered conserved if the human nucleotide was found to be conserved in 3 out of the other 4 species. Conservation scores were calculated based on nucleotide by nucleotide conservation in at least 4 out of 5 species as described above. The conservation score for a gene’s 3'UTR was the fraction of nucleotides in the human sequence that were conserved in 3 out of 4 other species.
miRNA/target analysis

The set of miRNAs tested included the ten highest expressed miRNAs in each of our tissue samples (Supplemental Table 9) (Eisenberg et al. 2007; Landgraf et al. 2007; Jima et al. 2010) as well as the 87 broadly conserved seed families, according to Targetscan (http://www.targetscan.org). The miRNAs that were expressed in our tissues but were not broadly conserved were miR-140-3p, miR-142-5p, miR-202-1, miR-320 and miR-423.

We first examined whether the distal 3'UTR regions of pAM genes are enriched for conserved miRNA seed matches relative to their proximal 3'UTR regions. The background set was defined as proximal 3'UTR regions of those genes. Then we compared the distal and proximal 3'UTRs of NpAM genes. Finally, we compared the distal 3'UTR regions of pAM genes to the 3'UTRs of single-UTR genes.

A binomial exact test was performed on the number of the miRNA hits in the 3'UTRs of a foreground test set compared to a background set of genes. Here we consider sliding a 7-mer window in the conserved regions of the 3'UTRs, a binomial trial would be each instance of the sliding window in the conserved sequence of the 3'UTRs and a success would be whether or not that 7-mer is an exact match of the miRNA seed in our list of 92 miRNAs.

As a control, we also computed enrichment statistics for 10,000 randomly generated seeds, generated according to the nucleotide distribution in the broadly conserved set. We performed the same binomial exact test for these random seeds. Empirical p-values were calculated based on the binomial test p-values of the random seeds, and these were then corrected to control for false discoveries using the Benjamini-Hochberg procedure. The cutoffs adjusted p-values at the 10% FDR and 25% FDR levels were mapped back to the binomial p-values that are shown in our plots (Fig. 5G). In this way, we calculated the empirical Benjamini-Hochberg adjusted cutoffs in our plots.

Ubiquitous or tissue-specific miRNA expression was determined using published data (Lu et al. 2005; Wienholds et al. 2005; Landgraf et al. 2007).

Gene ontology (GO) analysis was performed using DAVID (Huang da et al. 2009). We report background-corrected nominal p-values obtained by Fisher’s exact test. The background for the transformation experiments consisted of the union of genes expressed in either sample (Fig. 2I, 2J). The background for enriched GO terms in ubiquitously expressed pAM genes were all ubiquitously expressed genes across the seven tissue samples. The enrichment of the tissue-specific GO terms in genes with tissue-specific expression of the shorter 3'UTR isoform was
done without correction because the background correction was already performed in the previous analysis (Fig. 3C). The background for GO analysis for Fig. 5A consisted of all ubiquitously transcribed genes.

**Statistical analyses**

All the Mann-Whitney tests performed are two-sided. For all the statistical tests that compare either two groups or up to seven conditions (Chi square, Mann-Whitney, Kruskal-Wallis, Fisher’s exact, KS test), we regard the cut-off $p< 0.05$ as significant. When we perform genome-wide analyses assessing differences in mRNA abundance levels or 3'UTR isoform levels, we use the GLM and report FDR-corrected $p$-values to correct for multiple testing. For mRNA abundance we use an FDR-adjusted $p< 0.05$ and for 3'UTR isoform levels we use an FDR-adjusted $p< 0.1$, because of the somewhat higher variation detected.

**Supplemental online references**


Supplemental Figures
S1A. Genomic distribution of reads obtained by 3'-seq.
S1B. Nucleotide distribution surrounding pA cleavage sites.
S1C. Confirmation of extended 3'UTRs detected by 3'-seq using Northern blot analysis.
S1D. Different filtering thresholds show that the location of 3' cleavage events is largely independent of the tissue-context.
S1E. Northern blots of the genes shown in Figure S1D.
S1F. Northern blots of the genes shown in Figure 1B.
S1G. Correlation of mRNA abundance measured by 3'-seq with published RNA-seq data.
S1H. Correlation of mRNA abundance and 3'UTR isoform expression ratios of replicates using 3'-seq or Northern blot analyses.
S1I. Comparison of different 3' end sequencing methods with Northern blot analysis for NUDT21.
S1J. As in S1I, but for PTEN.

S2A. Western blots of MCF10A cells stably expressing activated HRAS.
S2B. HRAS-induced oncogenic transformation of MCF10A cells.
S2C. HRAS-induced increase in expression of shorter 3'UTR isoforms for RAC1, PAFAH1B1, RALA.
S2D. Increase in protein expression of RALA in MCF10A expressing activated HRAS.

S3. Differentiation is associated with changes towards expression of shorter as well as longer 3'UTR isoforms.
S3A. Testis vs ES cells.
S3B. Skeletal muscle vs ES cells.
S3C. Breast vs ES cells.
S3D. Ovary vs ES cells.

S4A. Contingency tables for Figure 4A.
S4B. Widely transcribed genes use the proximal pA sites significantly less than tissue-restricted genes.
S4C. The increase in the number of 3'UTR isoforms in widely expressed genes is biological.
S4D. As in S4B, but for multi-UTR genes with two 3'UTR isoforms.
S4E. Distribution of SUI of tissue-restricted and ubiquitously transcribed multi-UTR genes across the seven tissue samples.
S4F. Distribution of SUI of tissue-restricted and ubiquitously transcribed multi-UTR genes across the seven cell line samples.

S5A. Cumulative distributions of conservation scores in the 3'UTRs of pAM and NpAM genes.
S5B. Cumulative distributions of conservation scores in ubiquitously transcribed single-UTR, pAM and NpAM genes. The proximal and distal 3'UTRs are shown separately for multi-UTR genes.

Supplemental Tables
Table S1. 3'-seq read counts for all samples.
Table S2. Comparison of 3'UTR isoforms detected by 3'-seq and Northern blot analysis after applying different thresholds for filtering (see Fig. S1D, E).
Table S3. GO terms for changes in mRNA levels and 3'UTR isoform ratios in B cells +/- EBV.
Table S4. GO terms for changes in mRNA levels and 3'UTR isoform ratios in MCF10A +/- HRAS.
Table S5. GO terms ubiquitously transcribed genes with a tissue-specific 3'UTR isoform expression pattern.
Table S6. GO terms for ubiquitously transcribed single-UTR genes.
Table S7. GO terms for ubiquitously transcribed pAM genes.
Table S8. GO terms for ubiquitously transcribed NpAM genes.
Table S9. Ten highest expressed miRNAs expressed in the seven tissue samples.
Table S10. miRNA expression pattern for miRNAs that preferentially target pAM genes and single-UTR genes.