Title

mountainClimber Identifies Alternative Transcription Start and Polyadenylation Sites in RNA-Seq

Journal

Cell Systems, 9(4)

ISSN

2405-4712

Authors

Cass, Ashley A
Xiao, Xinshu

Publication Date

2019-10-01

DOI

10.1016/j.cels.2019.07.011

Peer reviewed

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2020 October 23.

Published in final edited form as:

Cell Syst. 2019 October 23; 9(4): 393–400.e6. doi:10.1016/j.cels.2019.07.011.

mountainClimber identifies alternative transcription start and polyadenylation sites in RNA-seq

Ashley A Cass[1,2,†], **Xinshu Xiao**[1,2,3,4,5,*]

[1]Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, California, 90095, USA

[2]Department of Integrative Biology and Physiology, University of California Los Angeles, Los Angeles, California, 90095, USA

[3]Institute for Quantitative and Computational Biosciences, University of California Los Angeles, Los Angeles, California, 90095, USA

[4]Molecular Biology Institute, University of California Los Angeles, Los Angeles, California, 90095, USA

[5]Lead Contact

Summary

Alternative transcription start (ATS) and alternative polyadenylation (APA) create alternative RNA isoforms and modulate many aspects of RNA expression and protein production. However, ATS and APA remain difficult to detect in RNA sequencing (RNA-seq). Here, we developed mountainClimber, a *de novo* cumulative sum-based approach to identify ATS and APA as change points. Unlike many existing methods, mountainClimber runs on a single sample and identifies multiple ATS or APA sites anywhere in the transcript. We analyzed 2,342 GTEx samples (36 tissues, 215 individuals) and found that tissue type is the predominant driver of transcript end variations. 75% and 65% of genes exhibited differential APA and ATS across tissues respectively. In particular, testis displayed longer 5' untranslated regions (UTRs) and shorter 3' UTRs, often in genes related to testis-specific biology. Overall, we report the largest study of transcript ends across human tissues to our knowledge. mountainClimber is available at github.com/gxiaolab/mountainClimber.

Graphical Abstract

*Correspondence: gxxiao@ucla.edu, Xinshu Xiao, 610 Charles E. Young Drive South, UCLA, Terasaki Life Sciences Building 2000E, Los Angeles, CA 90095, Phone: (310) 206-6522. Fax: (310) 206-9184.

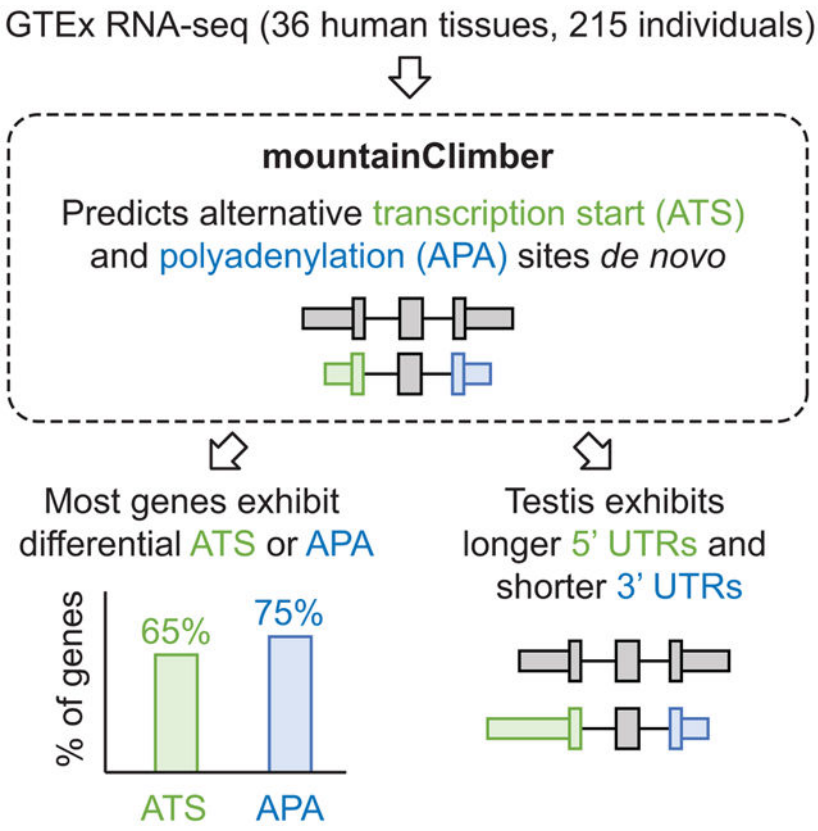
Author Contributions

A.A.C and X.X. designed the study and wrote the paper. A.A.C. implemented the software and performed all analyses.

†Present affiliation: Ambry Genetics Corporation, 15 Argonaut, Aliso Viejo, California, 92656, USA

Declaration of Interests

Ashley Cass is recently employed at Ambry Genetics Corporation as a Bioinformatics Scientist.



eTOC Blurb

mountainClimber identifies both ATS and APA sites *de novo* in RNA-seq. Unlike many existing methods, mountainClimber runs on a single sample and identifies multiple ATS or APA sites anywhere in the transcript. Analysis of 2,342 GTEx samples revealed that tissue type is the predominant driver of transcript end variations, and most genes exhibit ATS and/or APA. In particular, testis displayed longer 5' untranslated regions (UTRs) and shorter 3' UTRs, often in genes related to testis-specific biology.

Keywords

Alternative transcription start site; alternative polyadenylation; GTEx; human; tissues; change point; RNA-seq

Introduction

Alternative polyadenylation (APA) and alternative transcription start (ATS) are two important mechanisms that contribute to transcriptome diversity. Occurring in many mammalian genes, these processes may impose profound functional impact. APA, affecting >70% of mammalian genes, can influence many post-transcriptional aspects of the mRNA, such as mRNA stability, nuclear export and localization, and protein translation and localization (reviewed in Elkon et al., 2013; Mayr, 2018; Tian and Manley, 2017). In contrast, the primary impact of ATS, known to occur in 40-50% of mammalian genes, is the

modulation of protein production by altering open reading frames or sequence/structure motifs in the 5' UTR (Baek et al., 2007; Carninci et al., 2005; Leppek et al., 2018). Compared to the well-established tissue-specific nature of APA (Wang et al., 2008; Zhang et al., 2005), tissue-specific ATS is relatively less well studied.

To capture ATS and APA sites, specific experimental strategies have been developed, such as CAGE-seq (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014; Hon et al., 2017), PolyA-seq (Derti et al., 2012), 3'READS (Hoque et al., 2013), and 3'-seq (Lianoglou et al., 2013). However, these assays are not as accessible as the widely adopted RNA-seq methods that have generated numerous data sets. RNA-seq data, most of which are publicly available, represent valuable resources to study ATS and APA. Although several methods exist for identifying APA from RNA-seq (Arefeen et al., 2018; Kim et al., 2015b; Lu and Bushel, 2013; Shenker et al., 2015; Wang et al., 2014; Xia et al., 2014; Ye et al., 2018), no method has been developed specifically to capture both ATS and APA sites in RNA-seq data, to our best knowledge.

Here, we developed mountainClimber to enable *de novo* identification and analysis of transcription start sites (TSS) and polyadenylation (poly(A)) sites using RNA-seq data. This method has a number of unique features that distinguish it from existing methods, including its independence of known gene annotations, applicability to a single RNA-seq data set, and identification of multiple change points and different types of change points anywhere in the transcript. Additionally, mountainClimber is designed to analyze both the 5' and 3' ends of transcripts, allowing identification of ATS and APA simultaneously. We applied mountainClimber to the GTEx RNA-seq data (GTEx Consortium, 2015), which represents the largest study of tissue-specific ATS and APA in human to our knowledge. About 75% and 65% of tested genes had significantly differential APA and ATS in at least one pairwise tissue comparison, respectively. Testis exhibited widespread tissue-specific ATS, which often led to lengthening of the 5' UTRs. Testis-specific APA and ATS frequently affect the same set of genes, but these two mechanisms tend not to co-occur relative to the same tissue. Nevertheless, genes with such co-occurrence are involved in mitochondria-related function, and cell growth, division, and proliferation. Our study expands the repertoire of human tissue-specific APA and ATS sites.

Results

A *de novo* approach for change point detection in RNA-seq

mountainClimber was developed to identify change points while remaining robust to RNA-seq non-uniformity (Fig. 1 and Fig. S1A, STAR Methods). First, *de novo* transcription units (TUs) are identified by finding all continuous regions with RNA-seq reads (Fig. 1A). In each TU, mountainClimber leverages the RNA-seq non-uniformity by calculating the Cumulative Read Sum (CRS) and finding positions where the CRS significantly deviates from the null distribution (Fig. 1B). Fig. S2 illustrates two toy examples: a noisy trace without any change point, which generates the null CRS; another trace with one change point detectable as the elbow in the CRS. Such elbows in CRS are putative change points (Fig. 1C), which are further filtered to retain those with a significant change in read coverage (Fig. 1D and STAR Methods). Finally, the relative usage (RU) is calculated for each change point at the 5' and

3’ ends, such that the RU of all change points at each end sums to 1 (Fig. 1E). It should be noted that while we refer to this value as “relative usage”, it is possible that the observed RU is due to different stability of APA isoforms rather than different poly(A) site usage. Each 3’ end is labeled as DistalPolyA (3’ most poly(A) site), TandemAPA (APA within the last exon), IntronicAPA (APA within an intron), or ExonicAPA (APA within an exon). Similarly, each 5’ end is labeled as DistalTSS, TandemATSS, IntronicATSS, or ExonicATSS (Fig. 1F). In contrast to other methods, mountainClimber identifies intronic and exonic change points because change points are examined in the entire TU.

mountainClimber performance evaluation

To evaluate the performance of mountainClimber, we simulated tandemUTRs downloaded from MISO (Katz et al., 2010) with Flux Simulator (Griebel et al., 2012) (STAR Methods; Fig. S3A,B) and compared mountainClimber with IsoSCM (Shenker et al., 2015), because both methods identify 5’ and 3’ ends in a single sample in a *de novo* manner. As expected, recall and precision both increased with higher fold change at the 3’ end for both methods (Fig. 1G,H and Fig. S3C,D). mountainClimber outperformed IsoSCM in terms of recall, while the two methods yielded similar precision. This higher performance in recall was attributed to detection of TandemAPA change points rather than DistalPolyA, suggesting mountainClimber's advantage in identifying alternative change points (Fig. S3C). Additionally, mountainClimber’s performance is robust to the number of simulated APA sites (1, 2, or 3) per TU (Fig. S3E, F).

We also compared the performance of mountainClimber to that of DaPars (Xia et al., 2014). Since DaPars requires two conditions, we compared each simulated sample described above with a sample simulated with no change points (STAR Methods). mountainClimber, IsoSCM, and DaPars all achieved similar precision (Fig. S3G, I). However, mountainClimber outperformed the other two methods in terms of recall, especially for TandomAPA sites (Fig. S3H, J). Note that DaPars by default reports the annotated distal PolyA site from the input annotation file when a change point is detected, which explains its nearly perfect performance for DistalPolyA (STAR Methods, Fig. S3I, J).

We next evaluated performance using real RNA-seq data from MAQC Universal Human Reference RNA and Ambion Human Brain Reference RNA. Because we analyzed MAQC data as single samples, we did not evaluate DaPars’ performance on this dataset. Poly(A) sites of these samples were experimentally identified with PolyA-seq (Derti et al., 2012) and TSSs identified by FANTOM CAT (Hon et al., 2017) (STAR Methods), which will be assumed as the ground truth for evaluation purposes. To avoid ambiguity, TUs with at most one gene annotation were considered for downstream analysis (STAR Methods). As shown in Fig. 1I, mountainClimber achieved higher precision than IsoSCM regardless of the window size, w, around PolyA-seq sites used to define true positives. Notably, mountainClimber identified more sites around true PolyA-seq sites and poly(A) signal motifs A[A/T]TAAA than IsoSCM (Fig. 1J vs. K, S4A vs. S4B). As expected, most 3’ ends predicted by mountainClimber overlapped annotated poly(A) sites and 3’UTRs (Fig. S4C). In contrast to IsoSCM, mountainClimber also identifies 3’ ends located in introns and coding regions. mountainClimber’s precision is lower in introns and coding regions than in

3’ UTRs (Fig. S4C), which could be partly due to the limited accuracy or sensitivity of PolyA-seq in these regions.

Similarly, mountainClimber outperformed IsoSCM in terms of TSS precision (Fig. 1L) and predicted more change points with high fold changes near FANTAM CAT-predicted 5’ ends (Fig. 1M, N). This observation supports mountainClimber’s likely higher sensitivity in capturing *bona fide* ATS sites with biological significance.

Global analyses of alternative 3' and 5' transcript ends in human tissues

To investigate ATS and APA in humans, we analyzed 2,342 samples from 36 tissues and 215 individuals from GTEx (GTEx Consortium, 2015) after excluding low quality samples (STAR Methods and Table S1). To align thousands of samples in a reasonable amount of time and prioritize analysis of annotated genes, we used a simplified read mapping strategy (Fig. S1B) whose performance is similar to that used above (Fig. S4D, E).

A total of 5,786 TUs were identified in GTEx with at most one gene annotation and unambiguously inferred RNA strand (see STAR Methods for more details). We will refer to these TUs as genes hereafter. Most genes were predicted with one or two TSS or poly(A) sites and overlapped the expected Ensembl gene regions, supporting the validity of the predictions (Fig. S5). Additionally, some predicted poly(A) sites and TSSs were in coding exons and introns, which has been reported but no previous RNA-seq based methods were able to predict such sites (Elkon et al., 2013; Singh et al., 2018; Tian and Manley, 2017).

To quantify the total APA and ATS detected with increasing numbers of tissues, we calculated the average number of genes with APA and ATS sites with at least 10% RU in at least 10% of individuals (STAR Methods). As expected, APA and ATS were detected in increasingly more genes when more tissues were considered (Fig. 2A). Notably, 84% and 60% of TUs displayed APA and ATS respectively if all 36 tissues were considered (Fig. 2A). This observation, higher than the 75% and 50% reported previously (Carninci et al., 2005; Elkon et al., 2013; Mayr, 2018; Tian and Manley, 2017), was likely enabled by the usage of the largest dataset to date for analysis of ATS and APA.

Tissue type is the dominant driver of variation in transcript ends

mountainClimber’s analysis of single samples enables a systematic comparison of transcript ends across tissues and individuals. We calculated the weighted mean extension length (WMEL) for the 5' and 3' ends, where each segment length is weighted by its RU (STAR Methods). Although different types of alternative termini exist (Fig. S6A-C), 90% of genes exhibited TandemAPA and/or DistalPolyA and 82% exhibited TandemATSS and/or DistalTSS across samples (Fig. S6D-E and STAR Methods). Thus, we focused on analyses of these genes hereafter.

We used a linear mixed model to examine the variation of WMEL relative to tissue types or individuals (STAR Methods) (Mele et al., 2015). To enable a comparison encompassing most samples, we required the genes to have only TandemAPA or DistalPolyA in ≥90% of samples, resulting in 652 genes for the analysis of 5’ end and 829 for the 3’ end. This analysis showed predominant variation of WMEL across tissues, rather than individuals, for

both 5' and 3' ends (Fig. 2B, C). Interestingly, *HLA-C* was highly variable at both ends and is known to harbor many polymorphisms (reviewed in Parham, 2005). In summary, ATS and APA variation is mainly explained by tissue-related differences, but individual-specific variation does exist possibly reflecting genetic or functional diversity of certain genes.

Global comparisons of transcript ends across tissues

Since tissue type is the main driver of transcript termini variation, we next compared the WMEL values of the 5' and 3' ends across tissues. Testis had relatively low WMEL correlation with other tissues and had the longest average 5'-end length (Fig. S7A-B). It should be noted that testis RIN scores did not correlate with median WMEL, suggesting that longer 5' ends in testis are likely biological rather than technical artifact (Fig. S7C). At the 3' end, five tissues, whole blood, skeletal muscle, testis, cerebellum, and cerebellar hemisphere, were correlation outliers (Fig. S7D) and among those with the shortest or longest 3'-end length (Fig. S7E). Interestingly, the cerebellar regions were distinct from other brain regions in previous GTEx studies of RNA editing (Tan et al., 2017) and gene expression (Mele et al., 2015). These results again support the tissue-specific nature of transcript ends, with brain regions and testis demonstrating distinct patterns.

Tissue-specific alternative transcription start sites and polyadenylation sites

Next, we identified differential ATS and APA sites between pairwise tissues in 654 and 1,279 ATS and APA events that met our criteria for testing (STAR Methods). Of these, 424 / 654 (65%) and 960 / 1,279 (75%) of genes had significantly differential ATS and APA respectively (BH-corrected p-value ≤ 0.05 and absolute RU difference ≥ 0.05), consistent with the previous observations that 75% of genes have tissue-regulated APA (Wang et al., 2008) and 40-50% of genes utilize ATS sites (Baek et al., 2007; Carninci et al., 2005). Consistent with the results of WMEL correlation (Fig. S7A, D), testis had the most tissue-specific ATS events, whereas cerebellar hemisphere, cerebellum, and testis had the most tissue-specific APA sites (Fig. 2D, E). *APH1B* demonstrated a typical example of differential APA in a 3'UTR, where an annotated proximal poly(A) site is detected in testis, whereas a distal poly(A) site is detected in small intestine (Fig. 2F). *CPNE5* illustrates typical differential ATS in cortex vs. atrial appendage (Fig. 2G).

Atypical tissue-specific differential APA sites

Intronic polyadenylation is a well-established mechanism of gene regulation, first observed in 2007 and recently reported to be even more widespread (Berg et al., 2012; Oh et al., 2017; Singh et al., 2018; Tian et al., 2007). The ability to identify this type of poly(A) sites via RNA-seq is an advantage of our approach. A total of 12 genes exhibited significantly differential TandemAPA overlapping annotated introns in at least one pairwise tissue comparison (480 significant comparisons in total), with an example APA in *ABCF2* (Fig. 2H).

Another advantage of our approach is the ability to identify more than two change points. A total of 271 and 23 genes had two and three significantly differential change points in at least one pairwise tissue comparison respectively (with Fig. 2I showing one example).

Functional categories of genes with tissue-specific APA and ATS

We next tested whether ATS and APA tend to occur in tissue-specific or housekeeping genes. First, we observed that ATS/APA occurs in housekeeping genes (Eisenberg and Levanon, 2013) more often than expected (Fig. S8A). Additionally, the fraction of differentially expressed genes among APA and ATS genes was significantly less than the fraction of total genes tested (Fig. S8B and STAR Methods). Together, these results suggest that ATS/APA tends not to occur in ubiquitously expressed genes, consistent with previous observations (Lianoglou et al., 2013).

Gene Ontology analyses (Fig. S9A) revealed terms related to RNA splicing and processing for genes with longer 3’ ends in cerebellar regions (Fig. S9B-C), and cytoskeletal terms for genes with shorter 3’ ends in testis (Li et al., 2016) (Fig. S9D). On the other hand, genes with longer 3’ ends and those with longer 5’ ends in testis were both enriched with terms related to RNA splicing and processing (Fig. S9E-F).

Co-occurrence of alternative transcription start and polyadenylation sites in testis

Because tissue-specific ATS and APA were both abundant in testis and had similar functional terms, we next checked for their co-occurrence. As shown in Fig. S10A, there exists a significant overlap between genes with tissue-specific ATS and APA in any tissue compared with testis. However, co-occurrence of ATS and APA relative to a specific tissue was significantly avoided for all pairwise comparisons with testis (Fig. S10B). Together, these results suggest that the same set of genes in testis tend to be regulated at both the 5' and 3' end but avoided in the same tissue.

Nevertheless, although a minority, some genes did demonstrate a high frequency of co-occurrence of testis-specific ATS and APA in pairwise tissue comparisons. Such dual-regulation may reflect testis-specific function. Interestingly, three of these genes are nuclear-coded genes with mitochondria-related function while seven genes are related to cell growth, division, or proliferation (Fig. S10C-E and Fig. S11). Mitochondria structure, localization, and activity are regulated at different spermatogenesis stages (reviewed in Ramalho-Santos et al., 2009). APA and ATS in cell growth/division/proliferation may also be attributed to spermatogenesis. Thus, ATS and APA of these genes may contribute to spermatogenesis as well as characterize other testicular cell types in this heterogeneous tissue.

Discussion

Here, we presented mountainClimber, a *de novo* approach for identifying alternative 5’ and 3’ ends from RNA-seq. Application of mountainClimber to GTEx RNA-seq revealed that tissue type is the predominant driver of 5' and 3' end variations in ubiquitously expressed genes. Nevertheless, individual-specific 5' and 3' ends are important aspects of transcriptome diversity and should be examined in the future. Furthermore, the majority of genes exhibited ATS or APA across tissues. Both ATS and APA can impose regulation at the level of RNA stability, localization, and translation (Hinnebusch et al., 2016; Truitt and Ruggero, 2016). While further investigation of how ATS and APA relate to translational and other regulatory processes is important for an improved understanding of gene regulation in human tissues,

this question would be best addressed with matched mRNA and protein-level data, similar to a previous study in *Saccharomyces cerevisiae* (Cheng et al., 2018).

Our study demonstrated that mountainClimber is an effective method to simultaneously analyze both ends of transcripts in RNA-seq data. The nature of the "cumulative sum"-based method makes it robust to noisy read distributions. It should be noted that certain RNA-seq data sets may have suboptimal quality that leads to incomplete coverage of 5' or 3' ends, although this should not be a concern for GTEx data (Fig. S7B). If this is the case, interpretation of mountainClimber results should be carried out with caution. Additionally, it is possible that mountainClimber detects sense proximal RNAs or upstream antisense RNAs near the TSS when applied to non-strand-specific RNA-seq (Li et al., 2015). With the rapidly expanding amount of high-quality RNA-seq data (Xiang et al., 2018), we anticipate that mountainClimber will be a useful means to pinpoint the transcript ends and enable novel biological discoveries.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Xinshu Xiao (gxxiao@ucla.edu).

METHOD DETAILS

mountainClimber—mountainClimber consists of three major steps: defining *de novo* TUs in each sample, calling change points in each TU of each sample, and calculating relative usage (RU) of each change point. The pipeline is written in Python 2.7.2 and R 3.4.3, and relies on python modules pybedtools (Dale et al., 2011; Quinlan and Hall, 2010), scipy, numpy, peakutils, bisect, itertools, sklearn(Pedregosa et al., 2011), and pysam, and R packages ggplot2 (Wickham, 2009), reshape2 (Wickham, 2007), and dplyr.

***De novo* transcription unit identification.:** Given an input bedgraph file, consecutive windows of size w (default = 1000) are joined if at least p percent (default = 100%) of both windows have at least n average reads per base pair (bp) (default = 10). After merging consecutive windows, the ends of TUs are extended or trimmed until there are no zero-coverage bases. Split reads that align partially to one exon and partially to its neighboring exon can optionally be included as a bed file from intron start to intron end. After identifying TUs in each sample, overlapping TUs from all samples (for which comparisons will be carried out) are merged and annotated in order to have one universal set of TUs.

Change point identification.: Given a bedgraph file, bed file of split reads, and the *de novo* TUs, change points are called in each TU with length ≥ l (default = 1000) and average reads/bp in exons with length ≥ e (default = 10). If non-strand-specific RNA-seq is used, the strand is inferred by choosing the strand with the maximum number of supporting GT-AG splice site signals at exon-intron junctions. The Cumulative Read Sum (CRS) is defined as:

$$v_i = |x_i - x_{i-1}|$$

$$CRS_i = \frac{1}{CRS_l} \sum_{j=1}^{i} v_j$$

Where x_i is the total reads at position i in a TU of length l. The Kolmogorov-Smirnov (KS) test is used to test whether the CRS significantly deviates from the uniform distribution (the diagonal line in Fig. 1B). If the KS test p-value < t (default = 0.001), then proceed with calling change points. The elbows of the CRS are candidate change points. To identify them, the distance between the CRS and the diagonal line is calculated and denoised using a median filter with window size w. Then, the python peakutils module is used to call local maxima and minima in this distribution with parameters a and d, the normalized amplitude threshold and minimum distance between neighboring change points respectively. Since the CRS is a cumulative value, significant change points that follow a long segment of low coverage (e.g. a partially retained intron) do not appear to be local maxima or minima (Fig. 1C). This motivates the weighted CRS (wCRS), which effectively amplifies this signal at these positions:

$$w_i = \frac{v_i^2}{v_l}$$

$$wCVS_i = \frac{1}{wCVS_l} \sum_{j=1}^{i} | w_j - w_{j-1} |$$

On the other hand, change points in the first and last exon are more gradual due to the end-effect of the reads. Thus, we consider the union of candidate change points from the CRS and wCRS in the entire TU, and those from the CRS in the first and last exon. Finally, the change points are called again within a +/− w window in the non-denoised data to regain any resolution lost by denoising.

After calling all change points, five filters are imposed consecutively: (1) T-test p < t of the reads per base pair (bp) in the 2w bps before vs. after each of the predicted change points, (2) fold change of the first w bp of neighboring change points to be at least f (default = 1.5), (3) require that the segments in the first and last exon have strictly increasing and decreasing average coverage respectively, (4) distal segment expression > s (default = 1), (5) fold change of entire neighboring segments before vs. after is ≥ f.

To optimize parameters w, a, and d, mountainClimber considers different values (w = 100 up to min(l/100, 500) with step size 100; a = 0.05, 0.1, 0.15; d = 10, 50) and leverages the exon-intron junction information by choosing the combination that maximizes the total exon-intron junctions with at least n reads (default = 2) predicted within u bp (default = 10). After choosing the optimal parameters, terminal segments are removed if they have < z relative usage (default = 0.01), where relative usage is defined as the terminal segment coverage / maximum segment coverage in the TU, to remove transcriptional noise.

Finally, change points are labeled as follows: DistalTSS, DistalPolyA, TandemATSS, TandemAPA, Junction (if the change point is within u bp of an exon-exon junction with ≥ n reads; default u = 10; default n = 2), Exon, and Intron. The parameters selected for mountainClimber may affect these predictions. For example, a change point in an intron supported by two reads would be labeled Intron if n = 2, but would acquire one of the other labels if n > 2. If strand could not be inferred for nonstrand-specific RNA-seq, then change points are labeled: DistalLeft, DistalRight, TandemLeft, TandemRight, Exon, and Intron.

Relative usage calculation.: If a gene has three or fewer segments, only the distal 5' and 3' ends are reported. To calculate relative usage, the following change points are used to define 5' and 3' ends: (1) all change points labeled "Distal" and "Tandem" in the change point identification step described above, and (2) all change points identified before the first and after the last change point labeled "Junction". Ambiguous segments that could not be assigned to either 5' or 3' end were ignored. If a proximal segment has lower coverage than its neighboring distal segment (e.g. in introns) then the average reads/bp for that segment is set to 0. First, segments are ordered from lowest (k = 0) to highest coverage in each sample. RU is then calculated as follows:

$$RU_{k,s} = \begin{cases} \dfrac{\mu_{k,s}}{\max \mu_{k,s}}, & k = 0 \\ \dfrac{\mu_{k,s} - \mu_{k-1,s}}{\max \mu_{k,s}}, & k > 0 \end{cases}$$

$$RU_{\hat{k}} = \frac{1}{n} \sum_{s=0}^{n} RU_{\hat{k},s}$$

Where $\mu_{k,s}$ is the average reads/bp in each segment k and sample s. n is the total number of samples (excluding those with max $\mu_{k,s}$ = 0). After calculating $RU_{k,s}$ for each segment in each sample, segments are re-sorted back to distal to proximal order (indicated by $\hat{k}$), and the average is taken over all samples from each segment.

Comparison with annotation—mountainClimber change points were annotated with gene regions with the following priority: Junction > TSS > Poly(A) > CDS (coding exon) > 3'UTR > 5'UTR > Non-coding (exon in non-coding gene) > Intron > Intergenic (outside of annotated genes). Change points were labeled Junction, TSS, and Poly(A) if they were within 10 bp of an annotated TSS, poly(A) site, or exon-intron junction, respectively.

Mapping pipeline—There are two mapping pipeline options: (1) align to the genome prior to calling *de novo* TUs, or (2) align directly to the transcriptome +/−10kb (Fig. S1). In both pipelines, RSEM (Li and Dewey, 2011) was used to assign the most likely location for multi-mapped reads. While pipeline #1 yielded more (unannotated) TUs, pipeline #2 was significantly faster. Here, we briefly describe the major steps, with more details described in subsequent sections for simulations, MAQC, and GTEx mapping.

Genome alignment.: As recommended by RSEM, reads were aligned to the genome with up to 100 or 200 alignments. Bam files were converted to bedgraphs using bedtools genomecov, and TUs were called as described above. HISAT2 was run with the following parameters: --dta-cufflinks --mp 6,4 --no-softclip --no-mixed --no-discordant --add-chrname -k 100. STAR was run with the following parameters: --outSAMunmapped Within --outFilterType BySJout --outSAMattributes NH HI AS NM MD --outFilterMultimapNmax 200 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1 --runThreadN 8 --genomeLoad NoSharedMemory --outSAMtype BAM Unsorted --outSAMheaderHD @HD VN:1.4 SO:unsorted.

Prepare RSEM reference.: The RSEM reference was prepared from both *de novo* TUs and an existing annotation so that the aligner can better identify splice sites, as the *de novo* TUs do not contain splice site information.

Transcriptome alignment.: HISAT2 was used with the following parameters to allow up to 100 alignments per read and force HISAT2 to ignore indels in order to be compatible with RSEM: --mp 6,4 --no-softclip --no-unal --no-mixed --no-discordant --no-spliced-alignment --end-to-end --rdg 100000,100000 --rfg 100000,100000 -k 100. STAR was used with the following parameters: --outSAMunmapped Within -outFilterType BySJout –outSAMattributes NH HI AS NM MD –outFilterMultimapNmax 200 –outFilterMismatchNmax 999 –outFilterMismatchNoverLmax 0.04 –alignIntronMin 20 –alignIntronMax 1000000 –alignMatesGapMax 1000000 –alignSJoverhangMin 8 –alignSJDBoverhangMin 1 –sjdbScore 1 –runThreadN 8 –genomeLoad NoSharedMemory –outSAMtype BAM Unsorted –quantMode TranscriptomeSAM –outSAMheaderHD @HD VN:1.4 SO:unsorted.

RSEM.: After running rsem-calculate-expression, the alignment with maximum posterior probability was kept for each multi-mapped read.

Simulations—Flux Simulator (Griebel et al., 2012) was used with default parameters to simulate 100bp paired-end RNA-seq of annotated TandemUTRs downloaded from the MISO website (Katz et al., 2010). TandemUTR transcript isoforms were simulated with 1, 2, or 3 change points with varying distal / proximal expression. Simulated RNA-seq reads were aligned with pipeline #1 using HISAT2 and mm10 with Ensembl release 84. RSEM and mountainClimber were used as described above.

IsoSCM.: IsoSCM was run with default parameters. The median read coverage per segment reported in the coverage.gtf file was used to calculate the fold change at each change point. To assign true fold changes of annotated genes to IsoSCM's predictions, we overlapped the predictions with the true TandemUTR 3' ends +/−51bp.

DaPars.: In order to evaluate DaPars' performance on simulated data, we needed two sets of simulated samples. Thus, we compared each simulated sample described above with a sample simulated with no change points. In order to evaluate precision and recall at varying fold change cutoffs, we ran DaPars with relaxed criteria (FDR cutoff = 1, PDUI cutoff = 0,

and fold change cutoff = 0.1) and did not filter results based on significance. Fold change for DaPars proximal change points was calculated as (B_1_short_exp + B_1_long_exp + 1) / (B_1_long_exp + 1). Distal change points were defined using the Loci field, and fold change was either B_1_long_exp or A_1_long_exp if only detected in one sample, or the average of both if detected in both samples.

Performance evaluation.: A total of 14,377 TUs were called by mountainClimber, 13,289 (92%) of which overlapped annotated genes. mountainClimber identified 6,178 TUs with at most one gene annotation, while IsoSCM identified 5,070. Transcripts with at most one gene annotation (to avoid ambiguity) and at least 10 average reads per bp in exons were considered for precision and recall calculations. For precision and recall calculations, DistalPolyA and TandemAPA change points were considered for mountainClimber and 3p_exon change points were considered for IsoSCM. To compare varying degrees of difficulty in identifying 3' ends, we created equal sized bins of fold change of (proximal read coverage + 1) / (distal read coverage + 1) at true and predicted change points for recall and precision respectively. A predicted change point was a true positive if the closest true simulated change point was within 50bp. For recall, once a predicted change point was matched with a true change point, it could no longer be matched to another true change point. For precision, once a true change point was matched with a predicted change point, it could no longer be matched to another predicted change point.

MicroArray/Sequencing Quality Control (MAQC)—Ambion Human Brain Reference RNA (HBRR) and Universal Human Reference RNA (UHRR) total RNA sequencing (RNA-seq) data were downloaded from GEO accession GSE49712 (Rapaport et al., 2013). The following PolyA-seq sites were downloaded from the UCSC Genome Browser: MAQC UHR 1, UHR 2, Brain 1, Brain 2 (Derti et al., 2012). PolyA-seq sites with at least 1 RPM were considered. RNA-seq adapters were trimmed with cutadapt (Martin, 2011). Both mapping pipelines were run for comparison: pipeline #1 with STAR v2.5.2a (Dobin et al., 2013) and GENCODE v25, and pipeline #2 with HISAT2 v2.0.5 (Kim et al., 2015a) and Ensembl release 75, both aligned to hg19.

For pipeline #1, reads were mapped as described above, except split reads were not used in calling *de novo* TUs after genome alignment due to many split reads spanning multiple genes for unknown reasons. After genome alignment, *de novo* TU calls and RSEM analysis were carried out. Subsequently, *de novo* TUs were called a second time using junction reads from the transcriptome alignment since the junction reads from the genome alignment were problematic. These TUs were then used for calling change points. Pipeline #2 was used as described above.

IsoSCM was used as described above. Because we only considered TUs with at most one annotated gene for mountainClimber, we did the same for IsoSCM. To do so, we first merged all isoforms from the same transcript as identified by IsoSCM, then merged overlapping TUs across all samples and both strands.

To evaluate performance of mountainClimber and IsoSCM, TUs that satisfied the following requirements were used: (1) have at most one annotated gene, (2) at least 10 average reads

per bp, and (3) have inferred strand. These TUs were interrogated for overlap on the same strand with FANTOM CAT TSSs (Hon et al., 2017) and PolyA-seq poly(A) sites within 50, 100, 200, 300, 400, or 500 bp at both 5' and 3' ends. For predictions overlapping FANTOM CAT TSSs or PolyA-seq poly(A) sites within 300bp, we reported the positional precision binned by fold change at each position relative to the predicted change points. In this analysis, if a prediction was close to multiple FANTOM CAT or PolyA-seq sites, only the closest one was counted such that each prediction-to-FANTOM or -PolyA-seq site was a one-to-one match. Similarly, only the closest poly(A) signal (PAS) was plotted for each prediction (Fig. S3H,I). Precision was calculated as the total number of PolyA-seq poly(A) sites or FANTOM CAT sites (i.e. true positives) within +/−20bp of the prediction.

GTEx analysis—GTEx (GTEx Consortium, 2015) RNA-seq was downloaded through dbGaP under accession phs000424.v6.p1, except for subject phenotypes and sample attributes, which was downloaded under accession phs000424.v7. Gene-level RNA-seq read counts were downloaded under accession phe000020.v1.

Sample selection.: GTEx (GTEx Consortium, 2015) RNA-seq was downloaded through dbGaP accessions described above. A subset of GTEx samples were chosen as follows to maximize the number of tissues available per donor: individuals with less than 20 tissues (excluding cells) and tissues with less than 20 samples were excluded (esophagus, artery, skin sun-exposed, nerve – tibial, and minor salivary gland). Up to 100 individuals from each tissue were chosen, excluding those with different numbers of reads in readl and read2 and those from the same individual that were released multiple times (the most recent release was kept). Data sets with read length other than 76bp, median transcript integrity (TIN) score < 70 (calculated using tin.py from RSeQC (Wang et al., 2016)), and mapping rate < 50% were excluded, resulting in 2,342 samples from 36 tissues and 215 individuals (Table S1).

Mapping pipeline and mountainClimber.: Adapters were trimmed with cutadapt (Martin, 2011), adding -q 20,20 if FastQC returned "WARN" or "FAIL" for adapter content or per-base sequence quality, and –trim-n if FastQC returned "WARN" or "FAIL" for sequence content (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The mountainClimber mapping pipeline #2 was used (Fig. S1) with HISAT2 v2.0.5 (Kim et al., 2015a) and Ensembl release 75 aligned to hg19 standard chromosomes. mountainClimber default parameters were used except for: minimum expression -e was 5 average reads/bp.

A total of 20,292 TUs were identified across all GTEx samples, where 11,059 had at most one gene annotation and any predicted change points. Of these TUs, strand and therefore 5' and 3' ends were inferred for 5,786 TUs using splice site sequences. The other 5,273 TUs did not have inferred strand primarily due to lack of reads spanning exon junctions, among which only 346 (7%) were in annotated multiexon genes. For 28 genes, there were sufficient junction-spanning reads, but with non-canonical splice site sequences. Thus, our strand inference approach was successful for the majority of TUs corresponding to multi-exon genes.

Genome-wide APA and ATS calculation.: For each gene at each end, n tissues were randomly chosen (n = 1 to 36). The per-gene APA (or ATS) across tissues was calculated as the percentage of samples with more than one APA (or ATS) having at least 10% relative usage for each gene. The genome-wide APA (or ATS) measure was calculated as the percentage of genes with per-gene APA (or ATS) percentage of at least 10%. This process was carried out for 500 iterations.

Differential ATS and APA.: We used three major steps in the identification of significantly differential ATS and APA in pairwise tissues: (1) clustered change points across individuals and pairwise tissues, (2) calculated relative usage of clustered change points (described above), (3) tested for statistically significant differential ATS and APA.

First, change points were clustered across individuals and then across pairwise tissues using DBSCAN (sklearn.cluster). DBSCAN was chosen because it is not parameterized by the total number of clusters, but rather by the minimum number of points per cluster n and neighborhood size e. We used n = 50% of the number of individuals in the tissue of interest and e = the optimal window size from mountainClimber. For chrY, only males were clustered. The median position of each cluster was retained. Change point labels were prioritized as follows: Junction > DistalTSS > DistalPolyA > TandemATSS > TandemAPA > DistalLeft > DistalRight > Exon > Intron > TandemRight > TandemLeft.

Second, a minor modification was implemented in calculating the relative usage of each clustered segment; change points between the first and last exon were included if they were specific to one of the two tissues (e.g. an intronic poly(A) site present in one but not both tissues).

Third, differential ATS and APA were identified across pairwise tissues at both TU ends. The following criteria were used to identify genes eligible for differential testing: there were no ATS or APA identified and the distal segments were the same in both tissues, mean distal segment coverage was < d (default = 5) in both tissues, proximal coverage was < p (default = 0) in at least one tissue, or coverage was increasing from proximal to distal. If two clustered segments were tandem in at least one condition, then that change point was tested for differential usage.

To test for differential APA or ATS, four regions were considered – the segments before and after the change point in tissue l and in tissue2. APA and ATS were considered differential between two tissues if there was a significant difference in mean read counts of the distal segment. In order for the distal segments in the two tissues to be comparable, the distal segment coverage d_s in samples was scaled to the maximum proximal coverage p_s across all samples. Formally, each d_s was multiplied by λ_s:

$$\lambda_s = \frac{\max_s p_s + 1}{p_s + 1}$$

Because the number of replicates per condition is typically small in a given experiment, a standard t-test is underpowered to detect the difference between two means. Instead, we used

a data-driven estimation of expected variance across all tested distal segments for each pairwise tissue comparison (Yang et al., 2019). p-values were corrected with the Benjamini-Hochberg (BH) procedure. Change points with BH-corrected p-value <= 0.05 and absolute RU difference >= 0.05 were considered significantly differential.

Finally, the distal ends were labeled as follows: TandemTSS, TandemAPA, AFE (alternative first exon), and ALE (alternative last exon) when strand was inferred, and Tandem or AE (alternative exon) if strand was not inferred. AFE and ALE occur when the two tissues being tested have different 3' or 5' distal exons.

5' and 3' end grouping and WML.: After calculating relative usage, TUs were separated into four categories for 5' and 3' ends separately: (1) alternative first or last exon (AE) within a single sample, (2) AE across different samples, (3) the last segment was disrupted by an annotated intron, (4) tandem and distal UTRs (Fig. S6). To check for AE within and across samples, segments were clustered across all 2,342 samples with bedtools cluster. If multiple clusters were identified and any of the clusters contained ≥ 10 members, then the cluster with the maximum number of members was considered for grouping into either group 3 or 4. For group 3, only intronic regions with no exon overlap in any transcript isoform were considered for overlap with TUs. If the cumulative overlap with introns was at most 10bp or the entire last segment was contained within an intron, then the segment was assigned to group #4. Otherwise, the end was considered disrupted by an intron and placed in group #3.

For each end segment in groups 2, 3, and 4, the weighted mean length (WML) was defined as the weighted mean of each segment length, considering the RU of each segment as its weight. Finally, the weighted mean extension length (WMEL) was calculated by subtracting the minimum WML across all 2,342 samples from each WML (example illustrated in Fig. S6B).

Tissue and individual variabilities.: To assess the contribution of either tissue or individual variability to the observed variation in WMEL, we adopted an approach similar to (Méle et al., 2015). Briefly, a linear mixed model was used to model log2(WMEL) with tissue and subject modeled as random effects, and AGE, ETHNCTY and SEX as fixed effects. The lmer function from the lme4 package in R (Bates et al., 2015) was utilized as follows: lmer(log2(WMEL) ~ (1|tissue) + (1|subject) + AGE + ETHNCTY + SEX). To calculate the contribution of tissues and subjects to the observed variation, we divided their REML-estimated variance by the sum of the estimated variance of tissue + individual + residual.

Differential expression analysis.: We downloaded GTEx gene level read count data and used edgeR and limma (Ritchie et al., 2015; Robinson et al., 2010) to identify differentially expressed genes. Read counts for duplicated genes were summed across genomic locations. Lowly expressed genes were removed with the filterByExpr function with default parameters, resulting in 35,556 / 53,585 genes. All 2,342 samples were modeled together with functions model.matrix, voom, lmFit, and eBayes to assess differential expression. For each pairwise tissue comparison, we calculated the fraction of differentially expressed genes (p-value <= 0.05 and log2(fold change) >= 1). out of the total APA genes, ATS genes, and

total genes tested for differential expression (total genes tested was 35,556 for all pairwise comparisons).

Gene ontology analysis.: Gene ontology (GO) analysis was performed by testing for GO term enrichment of query genes compared to 10,000 sets of control genes with similar gene length and GC content as described previously (Lee et al., 2011). First, terms with less than three genes were excluded. Then, the enrichment p-value was calculated by fitting a normal distribution as described previously (Tran et al., 2019) and corrected with the Benjamini-Hochberg procedure. For each pairwise tissue comparison, genes with differential tandem ATS or APA were separated into two groups for both 5' and 3' ends: UTR lengthening and UTR shortening. The union of all genes that were significantly longer or shorter in each of the main five tissues of interest (whole blood, testis, skeletal muscle, cerebellum, and cerebellar hemisphere) compared to each of the other 35 other tissues were combined for GO analyses. The union of all tested genes in 3' and 5' ends were used as background, 1,122 and 546 total genes respectively.

DATA AND CODE AVAILABILITY

mountainClimber is available at github.com/gxiaolab/mountainClimber. We also provided the scripts used for downstream detection of differential ATS and APA usage.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the GTEx Project for generating the valuable data sets used in this study. This work was supported in part by grants from the National Institute of Health (U01HG009417, R01AG056476) and a National Institute of Health predoctoral training grant (T90DE022734). We would also like to thank the members of the Xiao laboratory and the UCLA Ribonomics of Gene Regulation group for their helpful discussions and comments on this work, especially Roberto Spreafico and Alexander Hoffmann.

References

Arefeen A, Liu J, Xiao X, and Jiang T (2018). TAPAS: tool for alternative polyadenylation site analysis. Bioinformatics 34, 2521–2529. [PubMed: 30052912]

Baek D, Davis C, Ewing B, Gordon D, and Green P (2007). Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. Genome Res. 17, 145–155. [PubMed: 17210929]

Bates D, Mächler M, Bolker B, and Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67, 1–48.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, et al. (2012). U1 snRNP determines mRNA length and regulates isoform expression. Cell 150, 53–64. [PubMed: 22770214]

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. (2005). The transcriptional landscape of the mammalian genome. Science 309, 1559–1563. [PubMed: 16141072]

Cheng Z, Otto GM, Powers EN, Keskin A, Mertins P, Carr SA, Jovanovic M, and Brar GA (2018). Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. Cell 172, 910–923.e16. [PubMed: 29474919]

Dale RK, Pedersen BS, and Quinlan AR (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27, 3423–3424. [PubMed: 21949271]

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl C. a., Johnson JM, and Babak T (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. 22, 1173–1183. [PubMed: 22454233]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Eisenberg E, and Levanon EY (2013). Human housekeeping genes, revisited. Trends Genet. 29, 569–574. [PubMed: 23810203]

Elkon R, Ugalde AP, and 6Agami R (2013). Alternative cleavage and polyadenylation: extent, regulation and function. Nat. Rev. Genet. 14, 496–506. [PubMed: 23774734]

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, et al. (2014). A promoter-level mammalian expression atlas. Nature 507, 462–470. [PubMed: 24670764]

Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, and Sammeth M (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 40, 10073–10083. [PubMed: 22962361]

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660. [PubMed: 25954001]

Hinnebusch AG, Ivanov IP, and Sonenberg N (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science 352, 1413–1416. [PubMed: 27313038]

Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. Nature 543, 199–204. [PubMed: 28241135]

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, and Tian B (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat. Methods 10, 133–139. [PubMed: 23241633]

Katz Y, Wang ET, Airoldi EM, and Burge CB (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods 7, 1009–1015. [PubMed: 21057496]

Kim D, Langmead B, and Salzberg SL (2015a). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360. [PubMed: 25751142]

Kim M, You B-H, and Nam J-W (2015b). Global estimation of the 3' untranslated region landscape using RNA sequencing. Methods 83, 111–117. [PubMed: 25899044]

Lee J-H, Gao C, Peng G, Greer C, Ren S, Wang Y, and Xiao X (2011). Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. Circ. Res. 109, 1332–1341. [PubMed: 22034492]

Leppek K, Das R, and Barna M (2018). Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat. Rev. Mol. Cell Biol. 19, 158–174. [PubMed: 29165424]

Li B, and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. [PubMed: 21816040]

Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, Park JY, Gunderson SI, Kalsotra A, Manley JL, et al. (2015). Systematic profiling of poly(a)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. PLoS Genet. 11, e1005166. [PubMed: 25906188]

Li W, Park JY, Zheng D, Hoque M, Yehia G, and Tian B (2016). Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. BMC Biol. 14, 6. [PubMed: 26801249]

Lianoglou S, Garg V, Yang JL, Leslie CS, and Mayr C (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 27, 2380–2396. [PubMed: 24145798]

Lu J, and Bushel PR (2013). Dynamic expression of 3' UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling. Gene 527, 616–623. [PubMed: 23845781]

Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17, 10–12.

Mayr C (2018). What Are 3' UTRs Doing? Cold Spring Harb. Perspect. Biol. a034728.

Melé0 M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. (2015). Human genomics. The human transcriptome across tissues and individuals. Science 348, 660–665. [PubMed: 25954002]

Oh J-M, Di C, Venters CC, Guo J, Arai C, So BR, Pinto AM, Zhang Z, Wan L, Younis I, et al. (2017). U1 snRNP telescripting regulates a size-function-stratified human genome. Nat. Struct. Mol. Biol. 24, 993–999. [PubMed: 28967884]

Parham P (2005). MHC class I molecules and KIRs in human history, health and survival. Nat. Rev. Immunol. 5, 201–214. [PubMed: 15719024]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. [PubMed: 20110278]

Ramalho-Santos J, Varum S, Amaral S, Mota PC, Sousa AP, and Amaral A (2009). Mitochondrial functionality in reproduction: from gonads and gametes to embryos and embryonic stem cells. Hum. Reprod. Update 15, 553–572. [PubMed: 19414527]

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, and Betel D (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 14, R95. [PubMed: 24020486]

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47. [PubMed: 25605792]

Robinson MD, McCarthy DJ, and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. [PubMed: 19910308]

Shenker S, Miura P, Sanfilippo P, and Lai EC (2015). IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. RNA 21, 14–27. [PubMed: 25406361]

Singh I, Lee S, Sperling AS, Samur MK, Tai Y, Fulciniti M, Munshi NC, Mayr C, and Leslie CS (2018). Widespread intronic polyadenylation diversifies immune cell transcriptomes. Nat. Commun. 9, 1716. [PubMed: 29712909]

Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, Liu KI, Zhang R, Ramaswami G, Ariyoshi K, et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature 550, 249–254. [PubMed: 29022589]

Tian B, and Manley JL (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. 18, 18–30. [PubMed: 27677860]

Tian B, Pan Z, and Lee JY (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res. 17, 156–165. [PubMed: 17210931]

Tran SS, Jun H-I, Bahn JH, Azghadi A, Ramaswami G, Van Nostrand EL, Nguyen TB, Hsiao Y-HE, Lee C, Pratt GA, et al. (2019). Widespread RNA editing dysregulation in brains from autistic individuals. Nat. Neurosci. 22, 25–36. [PubMed: 30559470]

Truitt ML, and Ruggero D (2016). New frontiers in translational control of the cancer genome. Nat. Rev. Cancer 16, 288–304. [PubMed: 27112207]

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, and Burge CB (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476. [PubMed: 18978772]

Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, Vedell PT, Barman P, Wang L, Weinshiboum R, et al. (2016). Measure transcript integrity using RNA-seq data. BMC Bioinformatics 17, 58. [PubMed: 26842848]

Wang W, Wei Z, and Li H (2014). A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. Bioinformatics 30, 2162–2170. [PubMed: 24728858]

Wickham H (2007). Reshaping Data with the reshape Package. J. Stat. Softw. 21, 1–20.

Wickham H (2009). ggplot2: Elegant Graphic for Data Analysis. Springer 1–210.

Xia Z, Donehower L. a., Cooper T. a., Neilson JR, Wheeler D. a., 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. [PubMed: 25409906]

Xiang Y, Ye Y, Zhang Z, and Han L (2018). Maximizing the Utility of Cancer Transcriptomic Data. Trends in Cancer 4, 823–837. [PubMed: 30470304]

Yang E-W, Bahn JH, Hsiao EY-H, Tan BX, Sun Y, Fu T, Zhou B, Van Nostrand EL, Pratt GA, Freese P, et al. (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. Nat. Commun. 10, 1338. [PubMed: 30902979]

Ye C, Long Y, Ji G, Li QQ, and Wu X (2018). APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. Bioinformatics 34, 1841–1849. [PubMed: 29360928]

Zhang H, Lee JY, and Tian B (2005). Biased alternative polyadenylation in human tissues. Genome Biol. 6, R100. [PubMed: 16356263]

Highlights

- mountainClimber identifies both ATS and APA sites *de novo* in RNA-seq
- The largest study of 5’ and 3’ transcript ends across human tissues to date
- 75% and 65% of genes exhibited differential APA and ATS in human tissues respectively
- Testis displayed longer 5’ UTRs and shorter 3’ UTRs, often in testis-specific genes

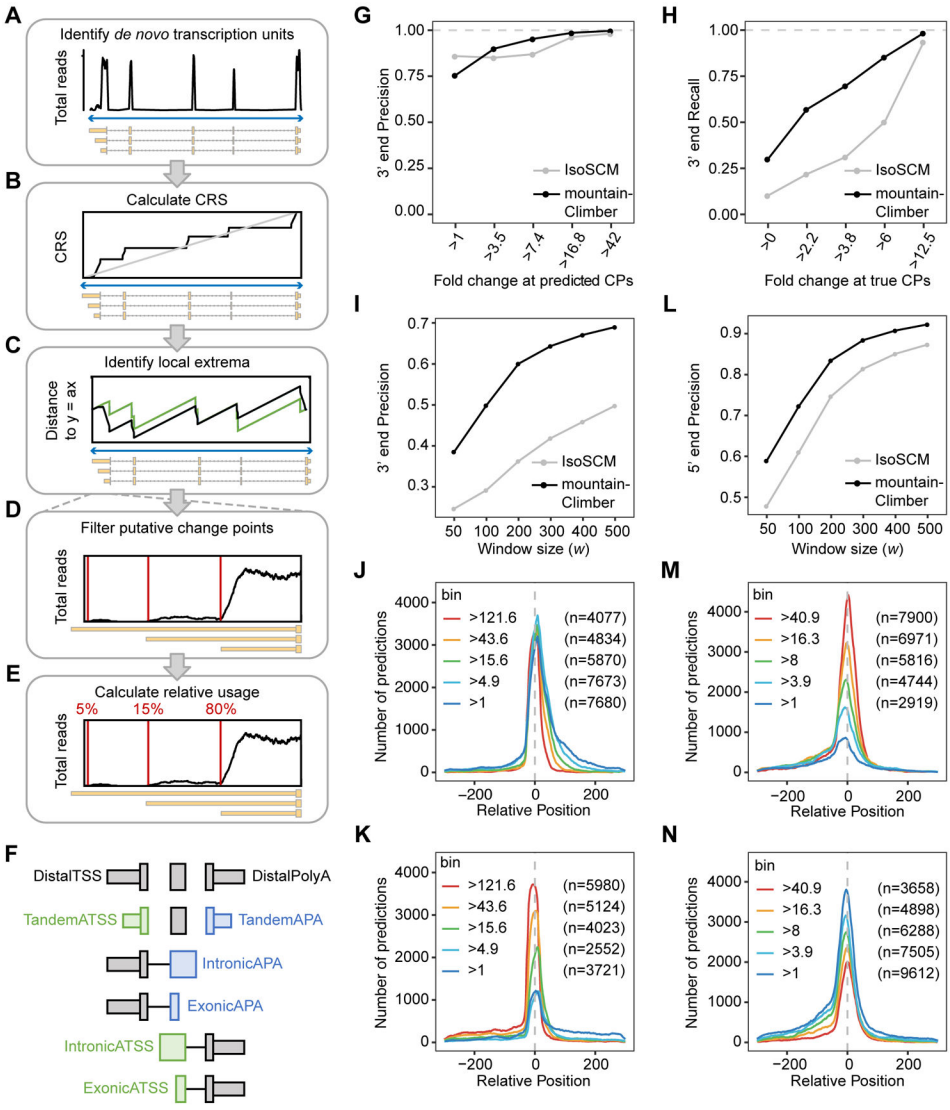


Figure 1. mountainClimber pipeline schematic and performance evaluation.

(**A**-**E**) The mountainClimber approach for identifying change points in each transcription unit (TU) in each sample. Example simulated RNA-seq is shown for *CDO1* (NM_033037), for which two change points are simulated. Simulated transcript isoform models are shown in yellow below each figure panel. (**A**) Identify *de novo* TUs. Poly(A)-selected RNA-seq is shown with the TU predicted shown in a blue line below. (**B**) Calculate cumulative read sum (CRS) as a function of position (black). The null distribution (diagonal line) is shown in grey. (**C**) Identify elbows in the CRS distribution by calculating the distance from the CRS (black) and weighted CRS (wCRS, green) to the diagonal line y = x. Note that wCRS is needed to observe elbows corresponding to both exon-intron junctions for each exon (black vs. green). (**D**) Filter putative change points by fold change and t-test. Dashed grey lines indicate zooming in to the last exon of the gene. Red lines indicate change points identified after filtering. (**E**) Calculate relative usage (RU) based on the average reads per bp in each segment at each end such that RUs sum to 1 at each transcript end. (**F**) The different types of

ATS and APA identified by mountainClimber. ATS and APA cases are colored green and blue respectively. (**G-H**) Performance on simulated RNA-seq and 3' ends. Fold change was calculated as the average reads/bp of proximal vs. distal segments. (**G**) Precision stratified by the fold change at predicted change points (CPs) (non-overlapping stratifications). (**H**) Recall stratified by fold change at true simulated CPs. (**I-N**) Performance on MAQC RNA-seq. (**I**) Precision for each window size w, where precision is calculated as the fraction of predicted change points that fell within w bp of any PolyA-seq site. (**J**) mountainClimber 3' predictions relative to PolyA-seq sites. Predicted 3' ends that are within 300bp of any PolyA-seq site (n = 30,134) were stratified by fold change into 5 bins. The x-axis indicates the position of the closest PolyA-seq site relative to each predicted poly(A) site, where positive (negative) values indicate the PolyA-seq site is downstream (upstream) of the prediction. The y-axis indicates the number of predictions at the corresponding position (x-axis) that have PolyA-seq support within +/−20bp. (**K**) Similar to (**J**), but for IsoSCM (n = 21,400). (**L**) Similar to (**I**), but for FANTOM CAT TSS. (**M**) Similar to (**J**), for the 5' end, comparing mountainClimber and FANTOM CAT sites (n = 28,350). (**N**) Similar to (**M**), but for IsoSCM (n = 31,961). For more details, see STAR Methods and Figures S1-S4.

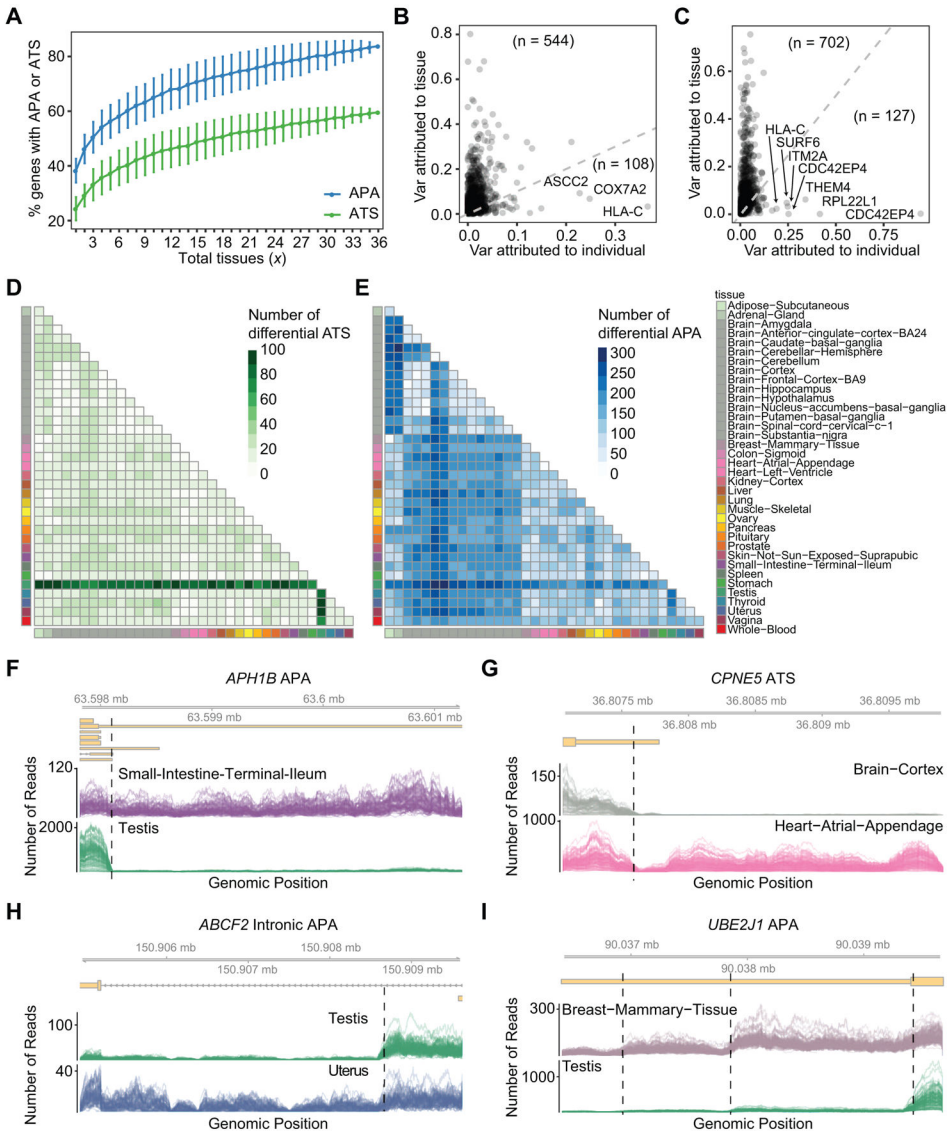


Figure 2. Landscape of alternative transcription start and polyadenylation sites in human tissues.

(**A**) Percentage of genes with APA (blue) or ATS (green) detected in *x* randomly chosen tissues (*x* = 1 to 36) across 500 iterations (STAR Methods). Mean and standard deviation are shown. (**B**) Variations in weighted mean extension length (WMEL) at the 5’ end attributed to individuals or tissues (STAR Methods). Numbers of genes below and above the dashed line y = x are shown. (**C**) Similar to (**B**), but for the 3’ end. (**D**) Number of significantly differential change points identified in each pairwise comparison in the 5’ end (BH-corrected p-value <= 0.05 and absolute RU difference >= 0.05). (**E**) Similar to (**D**), but for the 3’ end. (**F-I**) Examples of alternative 5’ and 3’ ends. The range shown contains the upstream and downstream segment of the differential change point(s) (i.e. the entire 5’ or 3’ end is not necessarily shown). Each line indicates the read counts for one individual at each nucleotide, and change points are indicated by black dashed lines. Genomic position is shown in grey (mb = megabase). Ensembl annotations are shown in yellow. (**F**) APA in

APH1B in testis vs. small intestine (p = 2.86e-251, RU difference = −0.524). (**G**) ATS in *CPNE5* in cortex vs. atrial appendage (p = 3.36e-256, RU difference = 0.60). (**H**) Intronic APA in *ABCF2* in testis vs. uterus (p = 2.47e-54, RU difference = 0.23). (**I**) Three APA change points in *UBE2J1* in breast vs. testis. See also Figures S5-S11.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
PolyA-seq data	Derti et al., 2012	UCSC Genome Browser hg19 tables: polyASeqSitesMaqcBrain1Fwd, polyASeqSitesMaqcBrain1Rev, polyASeqSitesMaqcBrain2Fwd, polyASeqSitesMaqcBrain2Rev, polyASeqSitesMaqcUhr1Fwd, polyASeqSitesMaqcUhr1Rev, polyASeqSitesMaqcUhr2Fwd, polyASeqSitesMaqcUhr2Rev
MAQC RNA-seq	Rapaport et al., 2013	GEO: GSE49712
MISO TandemUTRs	Katz et al., 2010	http://genes.mit.edu/burgelab/miso/annotations/miso_annotations_mm10_v1.zip
GTEx	GTEx Consortium, 2015	https://www.ncbi.nlm.nih.gov/gap/; RNA-seq phs000424.v6.p1; subject phenotypes and sample attributes phs000424.v7; gene-level RNA-seq read counts phe000020.v1
Software and Algorithms		
mountainClimber	Cass et al., 2019	https://github.com/gxiaolab/mountainClimber
IsoSCM	Shenker et al., 2015	https://github.com/shenkers/isoscm
DaPars	Xia et al., 2014	https://github.com/ZhengXia/dapars
RSEM	Li and Dewey, 2011	https://deweylab.github.io/RSEM/
HISAT2 v2.0.5	Kim et al., 2015a	https://ccb.jhu.edu/software/hisat2/index.shtml
STAR v2.5.2a	Dobin et al., 2013	https://github.com/alexdobin/STAR
Flux Simulator	Griebel et al., 2012	http://confluence.sammeth.net/display/SIM/Home
Python 2.7.2	Python Software Foundation	https://www.python.org/
R 3.4.3	The R Foundation	https://www.r-project.org/